Published in final edited form as:
Nat Biotechnol. 2022 May ; 40(5): 741–750. doi:10.1038/s41587-021-01136-7.

Partitioning RNAs by length improves transcriptome reconstruction from short-read RNA-seq data

Francisca Rojas Ringeling[1,13], **Shounak Chakraborty**[1,13], **Caroline Vissers**[2], **Derek Reiman**[3], **Akshay M. Patel**[1], **Ki-Heon Lee**[4], **Ari Hong**[5,6], **Chan-Woo Park**[4], **Tim Reska**[1], **Julien Gagneur**[7,8,9], **Hyeshik Chang**[5,6,10], **Maria L. Spletter**[11], **Ki-Jun Yoon**[4], **Guo-li Ming**[12], **Hongjun Song**[12], **Stefan Canzar**[1,✉]

[1]Gene Center, Ludwig-Maximilians-Universität München, Munich, Germany.

[2]Department of Biochemistry & Biophysics, University of California, San Francisco, San Francisco, CA, USA.

[3]Department of Biomedical Engineering, University of Illinois at Chicago, Chicago, IL, USA.

[4]Department of Biological Sciences, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea.

[5]Center for RNA Research, Institute for Basic Science (IBS), Seoul, Republic of Korea.

[6]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Republic of Korea.

[✉]**Correspondence and requests for materials** should be addressed to Stefan Canzar. canzar@genzentrum.lmu.de.
[13]These authors contributed equally: Francisca Rojas Ringeling, Shounak Chakraborty.

Author contributions
F.R.R. and S. Canzar conceived the original idea. S. Chakraborty and S. Canzar designed the kallisto-ls, StringTie-ls and Trintity-ls methods, with input from F.R.R. S. Chakraborty implemented the kallisto-ls, StringTie-ls and Trinity-ls methods and conducted all experiments and evaluations on simulated data. F.R.R. and S. Canzar contributed to the evaluation on simulated data. D.R. performed the initial computational study on simulated data (not shown). F.R.R. performed differential splicing analysis in NPCs, with input from S. Canzar. F.R.R. developed and performed Ladder-seq experiments, with input from H.S., G.M., M.L.S. and J.G. C.V. conducted RT–qPCR experiments. K.-J.Y. bred mice and harvested NPCs. K.-J.Y., H.C., K.-H.L., A.H. and C.-W.P. performed ONT long-read sequencing experiments. A.M.P., S. Chakraborty, F.R.R. and T.R. contributed to analysis of ONT long-read datasets. F.R.R. and S. Canzar wrote the manuscript, with input from all authors. S. Canzar supervised the entire project. All authors read and approved the final manuscript.

Competing interests
The authors declare no competing interests.

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

Code availability
The kallisto-ls, StringTie-ls and Trinity-ls programs and workflows are available at:

- https://github.com/canzarlab/ladderseq_quant,
- https://github.com/canzarlab/ladderseq_assembly and
- https://github.com/canzarlab/ladderseq_denovo,

respectively. The results of our benchmark studies can be reproduced via a Snakefile[33], available at https://github.com/canzarlab/ladderseq_benchmark.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41587-021-01136-7.

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

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-021-01136-7.

[7]Department of Informatics, Technical University of Munich, Garching, Germany.

[8]Institute of Human Genetics, Technical University of Munich, Munich, Germany.

[9]Institute of Computational Biology, Helmholtz Zentrum München, Neuherberg, Germany.

[10]School of Biological Sciences, Seoul National University, Seoul, Republic of Korea.

[11]Biomedical Center, Department of Physiological Chemistry, Ludwig-Maximilians-Universität München, Martinsried-Planegg, Germany.

[12]Department of Neuroscience and Mahoney Institute for Neurosciences, University of Pennsylvania, Philadelphia, PA, USA.

Abstract

The accuracy of methods for assembling transcripts from short-read RNA sequencing data is limited by the lack of long-range information. Here we introduce Ladder-seq, an approach that separates transcripts according to their lengths before sequencing and uses the additional information to improve the quantification and assembly of transcripts. Using simulated data, we show that a kallisto algorithm extended to process Ladder-seq data quantifies transcripts of complex genes with substantially higher accuracy than conventional kallisto. For reference-based assembly, a tailored scheme based on the StringTie2 algorithm reconstructs a single transcript with 30.8% higher precision than its conventional counterpart and is more than 30% more sensitive for complex genes. For de novo assembly, a similar scheme based on the Trinity algorithm correctly assembles 78% more transcripts than conventional Trinity while improving precision by 78%. In experimental data, Ladder-seq reveals 40% more genes harboring isoform switches compared to conventional RNA sequencing and unveils widespread changes in isoform usage upon m^{6}A depletion by *Mettl14* knockout.

Short-read RNA sequencing (RNA-seq) is the most widely used assay for transcriptome profiling, and many computational methods have been developed to identify and quantify transcripts from the produced sequence read data. Transcript quantification methods assign reads to known species-specific transcripts to obtain a quantitative measurement for their relative expression, and the assembly of transcript sequences can reveal novel types of RNA molecules. In contrast to the reference-based assembly that builds full-length transcripts from reads ordered by a prior alignment to a reference genome, the de novo assembly approach reconstructs transcripts based on the sequence overlap of reads alone and can be applied to species for which no or just a highly fragmented reference genome is available.

Despite many methodological advances, the accuracy of transcript-level inference methods developed over the last decade is severely limited by the lack of long-range information contained in each individual short read. They perform particularly poorly in the detection and quantification of lowly expressed transcripts and transcripts from complex genes[1–3]that share large parts of their sequences due to alternative splicing. Multi-sample approaches, such as the recently introduced PsiCLASS[4], try to address these limitations by assembling transcripts simultaneously across multiple RNA-seq samples. On the other hand, third-generation technologies, such as those marketed by Pacific Biosciences or Oxford Nanopore

Technologies (ONT), are able to read full-length transcripts but at a lower throughput, a higher error rate and a higher cost per base[5].

Here we propose Ladder-seq, a new variant of the RNA-seq protocol that effectively breaks gene complexity by separating mRNAs according to their lengths into a small number of bands before their fragmentation. This experimental deconvolution can guide an algorithm to assemble or assign reads to transcripts only of a correct length. We extend and tailor state-of-the-art RNA-seq analysis methods for quantification, reference-based assembly and de novo assembly to use the extra layer of information introduced in Ladder-seq to detect and quantify transcripts at an unprecedented level of accuracy and reveal transcripts that are invisible to conventional RNA-seq approaches.

More accurate transcript-level estimates from Ladder-seq will facilitate downstream differential analysis, which we exploited in a study of epitranscriptomic regulation of splicing in mouse neural progenitor cells (NPCs). m^6A is the most abundant internal modification of mRNA in eukaryotic cells[6] and is involved in multiple aspects of mRNA biology. Here we reveal a critical role of m^6A methylation in NPCs as a regulator of alternative splicing, which is highly prevalent in the nervous system[7–9] and has been associated with neurological disorders such as autism.

Results

Generation of Ladder-seq libraries of mouse NPCs.

We generated Ladder-seq datasets from *Mettl14* wild-type (WT) and knockout (KO) mouse NPCs (Methods). *Mettl14* encodes for a methyltransferase necessary for m^6A methylation of mRNA. Four independent replicates were prepared per genotype.

Compared to conventional RNA-seq, in Ladder-seq, mRNAs are separated by their lengths into a small number of bands before their fragmentation (Fig. 1a). To achieve mRNA separation by transcript length, we performed denaturing gel electrophoresis. After electrophoresis, each sample was cut into seven bands guided by a single-stranded RNA ladder running on the same gel. This effectively reduced gene complexity in our dataset (Extended Data Fig. 1) by partitioning transcripts expressed per gene into different subgroups. We denote the size of each subgroup as its effective complexity. mRNAs were effectively separated into seven distinct length ranges with a certain degree of overlap between consecutive bands (Fig. 1b and Supplementary Table 1). mRNA from each band of each sample was extracted from the agarose gel, and equal volumes per band were used for cDNA library construction. Each band from each sample was given a unique barcode to track the originating band (per sample) of each read.

Correlation analyses of transcript expressions show high technical reproducibility of our Ladder-seq protocol ($r = 0.96–0.98$; Supplementary Fig. 1). Furthermore, transcript expression levels were well-correlated between each of the four WT Ladder-seq samples and three conventional RNA-seq reference datasets (without length separation) from WT NPCs ($r = 0.81–0.82$; Supplementary Fig. 2 and Supplementary Table 2), despite using different experimental batches. Pearson correlation coefficients of our Ladder-seq samples were similar to those of five public RNA-seq samples of mouse NPCs[10,11] (Supplementary Tables 2 and 3), which holds also when correlation was stratified by transcript length ranges that follow the location of cuts used in our experiments (Supplementary Fig. 3). Transcripts with low correlation did not differ significantly in length from highly correlated tanscripts (Supplementary Fig. 4). The total number of detected annotated transcripts is highly similar between Ladder-seq and conventional RNA-seq (Supplementary Figs. 5 and 6), and the detection rate increased with transcript length, as previously reported[12] (Supplementary Figs. 7 and 8).

The separation of mRNAs by length from NPCs on an agarose denaturing gel introduces separation errors that result in the spread of molecules of the same transcript species across different bands. Even though transcript length is the main determinant for mRNA migration[13], residual secondary structure formation plays a role in determining the migration pattern of transcripts (migration errors). This might vary between molecules and can even occur under denaturing conditions. We apply a histogram-based method to estimate a discrete density function according to which reads obtained from transcripts of a given length distribute across bands (Methods). We rely on reads that map uniquely to annotated transcripts with high confidence. The ‘in silico gel’ in Fig. 1c (and Extended Data Fig. 2) confirms the migration of transcripts according to their annotated length.

Transcript quantification—kallisto-ls.

Reads that map to a unique genomic position often cannot be assigned unambiguously to one of a gene’s transcripts, because alternatively spliced isoforms might overlap in genomic coordinates. Transcript quantification methods, therefore, use a statistical model of RNA-seq to probabilistically assign reads to transcripts. We have extended this statistical model to our new protocol, Ladder-seq, and implemented an expectation maximization (EM) algorithm that infers transcript abundances that can best explain the observed reads and their (inexact) separation into bands. The read’s band contains transcripts of a specific length range and, thus, provides valuable information when trying to probabilistically resolve its assignment ambiguity between transcripts of different lengths (Fig. 2a). Based on estimated migration patterns of transcripts, we adjust the probability of obtaining a read in a given band from a specific transcript by the probability of seeing a transcript of the same length in the corresponding band (Methods).

We extend the EM implementation in one of the most widely used software tools, kallisto[14], to quantify transcripts based on pseudo-alignments of Ladder-seq reads. To assess the advantages of our Ladder-seq-tailored EM implementation, kallisto-ls, over conventional kallisto, we compared their performance on simulated Ladder-seq samples and matching RNA-seq samples, respectively (Extended Data Fig. 3). As in the original benchmark in ref. [14], we simulated 30 million and 75 million 2 × 75-bp paired-end reads. From each simulated RNA-seq sample, we derive a matching Ladder-seq sample by introducing an in silico length separation. We assign each read randomly to one of a fixed number of bands (here, seven), where the random assignment follows a distribution that reflects migration patterns estimated from our mouse NPC data.

We measure quantification accuracy by mean absolute relative difference (MARD) and Pearson correlation (Methods), the same metrics used in a benchmark of transcript quantification methods[1]. kallisto-ls makes use of the additional length information contained in the Ladder-seq data to quantify transcripts more accurately than conventional kallisto (Fig. 2b and Extended Data Fig. 4). In fact, kallisto-ls is able to quantify transcripts of genes expressing ten isoforms as accurately (in terms of MARD) as conventional kallisto is able to quantify merely two expressed isoforms.

To evaluate the effect that a more precise length separation has on the accuracy of Ladder-seq, we mimic an idealized version of the Ladder-seq protocol, which perfectly separates transcripts by length without any migration errors. To this end, the same set of reads is partitioned into the same number of bands deterministically according to the length of the originating transcript. Figure 2b (and Extended Data Fig. 4) shows that a more accurate length separation can improve quantification accuracy even further, yielding a reduction in MARD of more than 31% for genes expressing four transcripts.

Reference-based transcript assembly—StringTie-ls.

Current methods for reference-based assembly represent reads connecting neighboring exons by a graph structure, such as the splicing graph[15], and infer transcripts as paths through this graph. However, the space of possible candidate transcripts that can be obtained by combining locally connected exons in paths through the graph can grow exponentially, and smoothing the local coverage along transcripts cannot unambiguously point to a single best subset of transcripts[16].

Here, we propose a computational framework (Fig. 3a) that enables conventional RNA-seq assembly methods to exploit the extra layer of information provided by Ladder-seq to reduce the ambiguity of combining distant splicing events into transcript isoforms. In this scheme, a separate splicing graph is built from reads in each band, and transcript length constraints aid in breaking (too-long) erroneous fusions and in eliminating (too-short) transcript fragments. Length constraints are derived from distributions of transcript lengths across bands, which are estimated using a histogram-based method (Methods). We use kallisto-ls to assign reads to assembled transcripts according to our statistical model of Ladder-seq.

We chose StringTie2 (ref. [17]) as the presumably most accurate RNA-seq assembly method[17,18] to illustrate the benefit of our Ladder-seq-tailored assembly approach (StringTie-ls) over its conventional RNA-seq counterpart. We simulated additional Ladder-seq samples that mimic an improved length separation step by gradually reducing the degree of migration errors (Methods).

Figure 3b (and Extended Data Fig. 5) shows that StringTie-ls is able to correctly reconstruct a much larger fraction of expressed transcripts than conventional StringTie2, and, as expected, this improvement in sensitivity increases with gene complexity. For genes expressing four transcripts, StringTie-ls detects 16% more transcripts than conventional StringTie2, and this improvement increases to 31.1% and 35.2% for complex genes expressing seven and ten transcripts, respectively. The sensitivity gap between these two technologies widens with a more accurate length separation of transcripts, reaching an improvement of 25.2% for genes expressing four transcripts and 49.2% and 58.7% for genes of complexity 7 and 10, respectively, in the most optimistic scenario. At the same time, StringTie-ls assembles transcripts with higher precision across all complexity classes. StringTie-ls benefits considerably from the additional length information that allows it to detect too-short transcript fragments. For genes expressing single transcripts, for example, StringTie-ls recognizes 699 of 824 false-positive assemblies from conventional StringTie2 as being too short and eliminates them, improving precision by 30.8%.

In addition, we compared transcripts assembled from our Ladder-seq NPC samples to transcripts identified from long reads generated by ONT. We performed ONT long-read native RNA (ONT-RNA) and direct cDNA (ONT-cDNA) sequencing of WT and *Mettl14* KO mouse NPCs. Expression levels were well-correlated between ONT and Ladder-seq samples (Supplementary Fig. 9 and Supplementary Tables 11 and 12) and consistent with previously reported correlations between ONT and RNA-seq data[12,19].

Third-generation sequencing technologies, such as those from ONT and Pacific Biosciences, can produce reads longer than 10,000 bp, which, in principle, can capture full-length transcripts. The lower sequencing depth and the higher error rate, however, result in an incomplete transcriptome reconstruction that will also include false transcripts. Nevertheless, a transcript assembled from short reads is likely to be truly expressed if it can be independently identified in the long-read data. Conventional StringTie2 missed many long-read transcripts successfully recovered by StringTie-ls, in both conditions and compared to both native RNA and cDNA libraries (Supplementary Tables 13 and 14). The large number of transcripts assembled only from short reads that matched an annotated transcript can be attributed to the incompleteness of the long-read transcriptomes.

De novo transcript assembly—Trinity-ls.

To study the transcriptome of species for which no or just a highly fragmented reference genome is available, or in samples with a substantially altered genomic sequence, transcripts need to be assembled de novo. Omitting the read mapping step that arranges reads in order leaves the sequence overlap of reads as the only source of information to be used by methods for this most challenging transcript-level inference task. Most methods, including one of the most widely used methods, Trinity[20], stitch together k-mers, subsequences of k nucleotides, to transcript sequences by traversing paths in so-called de Bruijn graphs. No part of these data connects subpaths at longer distances, which can cause erroneous fusions of isoforms or paralogs, especially in complex genes with a large number of alternative splicing events[21].

Here, we follow a similar strategy as in the reference-based assembly (Fig. 3a) to access the additional layer of information provided by Ladder-seq to guide the de novo assembly of full-length transcripts by Trinity. We use Trinity to compute length-constrained paths in de Bruijn graphs representing k-mer connectivity rather than paths in splicing graphs. Again, we quantify assembled transcripts by probabilistically assigning reads using our statistical model of Ladder-seq, taking into account estimated migration errors.

Figure 4 (and Extended Data Fig. 6) shows an enormous performance gain of Trinity-ls over conventional Trinity on our simulated data, in terms of both sensitivity and precision. In total, Trinity-ls correctly recovers an additional 4,072 (78%) transcripts compared to Trinity while, at the same time, increasing precision equally by 78%. A more accurate separation of transcripts by length further boosts the performance of Trinity-ls, approaching an additional 163% of correctly discovered transcripts and a 3.9-fold increase in precision when transcripts are perfectly separated by their lengths.

Ladder-seq improves differential analysis of transcriptomes.

We evaluated the effect of a more accurate reconstruction of transcriptomes on differential analysis between two biological conditions. We used Ladder-seq to profile the transcriptome of WT and *Mettl14* KO mouse NPCs. To assess transcript usage under these conditions, we first assembled transcripts using StringTie-ls on each sample to identify novel transcripts that are expressed consistently across replicates of the same genotype. We quantified annotated (Ensembl release 95) and newly reconstructed transcripts using kallisto-ls and compared their expression between conditions to detect their differential usage. For comparison with conventional RNA-seq, we ran the same computational pipeline replacing the Ladder-seq-tailored methods, kallisto-ls and StringTie-ls, by their conventional counterparts, which ignore the separation of reads into bands (Extended Data Fig. 7a).

Ladder-seq identified 40% more genes harboring switching isoforms in *Mettl14* KO compared to conventional RNA-seq (Extended Data Fig. 7b and Supplementary Table 19). Taking gene complexity—that is, the number of expressed transcripts per gene—as a measure of difficulty in assembling transcripts, genes identified as switching exclusively by Ladder-seq appear to be particularly difficult to reconstruct by the conventional pipeline without the additional length separation (Fig. 5a). In contrast, Ladder-seq breaks down gene complexity, effectively reducing the number of transcripts that need to be reconstructed in an individual band. This effective complexity is considerably lower in all three categories of genes identified as switching (Fig. 5a), including genes identified as switching only by the conventional pipeline.

Ladder-seq uncovers otherwise buried transcripts that are not identified by conventional RNA-seq. This is exemplified by the isoform switch in gene *Pi4k2a*, which is only identified by our method (Fig. 5b,c). StringTie-ls uncovered a shorter transcript that is absent from the Ensembl release 95, but it does appear in the later release 98 version (ENSMUST00000235932) and is also present in the ONT long-read data (TCONS_00005143 in Supplementary Tables 20 and 21), confirming that what Ladder-seq assembled is indeed accurate. In addition, we confirmed this isoform switch with reverse transcription quantitative polymerase chain reaction (RT–qPCR) (Fig. 5d). Additional illustrative examples of isoform switches uncovered only by Ladder-seq are shown in Extended Data Fig. 7c–f.

Ladder-seq makes use of estimated probability distributions, which describe how mRNA molecules migrated through the denaturing gel. We used Jensen–Shannon divergence (JSD) to compare these estimated migration patterns of transcripts to distributions of reads across bands assigned to them by conventional kallisto or by kallisto-ls. JSD values for kallisto-ls were consistently low for all identified switching genes, which is to be expected given that kallisto-ls makes explicit use of these distributions to guide the assignment of reads. On the other hand, JSD values for conventional kallisto were highest for those genes identified as switching only by conventional RNA-seq (Fig. 5e). These large JSD values are likely an indication of erroneous assignments of reads by conventional kallisto, because JSD values also increase with the difficulty of the quantification task (Fig. 5f). More generally, we observed that the more conventional kallisto differs from kallisto-ls, the more its assigned read band distribution deviates from the estimated distribution, resulting in larger JSD values (Extended Data Fig. 7g).

Finally, we used ONT long reads of WT and KO NPCs to validate novel transcripts involved in isoform switches. Of all 499 novel switching isoforms detected exclusively by Ladder-seq, 206 (41.3%) were identified from ONT-cDNA or ONT-RNA long-read data by FLAIR or assembled by StringTie2 or were contained in a recently published ONT long-read mouse NPC transcriptome[22]. Only 18 of 97 (18.6%) novel switching isoforms reported only by conventional RNA-seq were confirmed by long-read sequencing.

Mettl14 KO leads to isoform switches in m^{6}A methylated genes.

We next set out to delineate the characteristics of isoform switches and their relationship to m^{6}A methylation. To assess whether m^{6}A is associated with isoform switches in *Mettl14* KO, we identified m^{6}A-tagged genes in a public m^{6}A RNA IP and sequencing dataset from mouse NPCs[23]. We found that switching genes are significantly enriched for m^{6}A methylated genes ($P = 2.36 \times 10^{-19}$) (Fig. 6a). These genes are enriched for Gene Ontology (GO) terms related to transcriptional regulation, neurogenesis and synaptic signaling (Extended Data Fig. 8a).

To investigate the involvement of m^{6}A methylation in isoform switching, we explored a potential spatial proximity between m^{6}A and alternative splicing. We assessed whether exonic segments[24,25] bounding differentially spliced regions are enriched for m^{6}A methylation (Methods). We found a significant enrichment of m^{6}A within these segments ($P = 8.6 \times 10^{-39}$) (Fig. 6b). This enrichment persists when normalizing for segment length ($P = 1.09 \times 10^{-5}$), which accounts for a possible bias toward longer exons[26,27]. Illustrative examples of m^{6}A methylation within a differentially spliced exonic segment are shown for neurogenesis-related genes *Fbxl5* (ref. [28]) and *Ptprz1* (ref. [29]) (Fig. 6c and Extended Data Fig. 8b).

We then studied the consequences of isoform switches on functional protein domains. We found 295 genes with loss of functional domains in the upregulated isoform in the KO. GO analysis of these genes shows enrichment for terms related to neuronal function, such as glutamatergic synaptic transmission, synapse organization and GABA secretion (Extended Data Fig. 8c and example in Fig. 6d).

Although other types of splicing events were balanced between WT and *Mettl14* KO, upregulated isoforms in KO had significantly more intron retention losses than gains (Fig. 6e and Extended Data Fig. 8d). Again, these genes were enriched for m^{6}A methylated genes ($P = 1.6 \times 10^{-6}$). GO analysis revealed enrichment for terms unrelated to neuronal functions but rather associated with pluripotency, such as DNA repair and gamete generation (Extended Data Fig. 8e). We found enrichment for nonsense-mediated decay (NMD) insensitive isoforms as well as for shorter 3′ untranslated region (UTR) (Fig. 6f), both hallmarks of decreased regulation of gene expression[30,31]. Finally, we validated a selection of identified isoform switches by RT–qPCR (Fig. 5d and Supplementary Table 22).

Long-read sequencing confirms many Ladder-seq transcripts.

We next compared the Ladder-seq-inferred WT transcriptome of mouse NPCs (Extended Data Fig. 7a) with transcripts identified by FLAIR[32] from our ONT-cDNA and ONT-RNA long reads. We found that 63.3% of ONT-cDNA transcripts were contained in at least one WT Ladder-seq transcriptome with relative expression of at least 0.1 transcripts per million (TPM). Of those, a larger fraction of transcripts was independently assembled by StringTie2 from the ONT-cDNA data or contained in a recently published ONT long-read mouse NPC transcriptome (Dong et al.[22]), compared to those reported only by ONT-cDNA (Extended Data Fig. 9a). The substantially lower validation rate suggests that a larger fraction of transcripts missing in the Ladder-seq transcriptomes were falsely inferred by FLAIR from ONT-cDNA reads and, similarly, from our ONT-RNA data (Supplementary Fig. 10a). As expected[12], Ladder-seq detected more annotated genes and transcripts than could be mapped from the ONT libraries (Supplementary Figs. 11 and 12). Nevertheless, 71.1% of transcripts reconstructed by Ladder-seq with relative abundance of at least 1 TPM were identified by FLAIR or assembled by StringTie2 in the ONT-cDNA dataset or were contained in Dong et al. (Extended Data Fig. 10). This overlapping set of transcripts showed higher expression levels than the remaining set of transcripts (Extended Data Fig. 9b), suggesting the limited sequencing depth of the ONT dataset as one possible explanation for their absence in the long-read transcriptome[12]. This was consistently observed in the ONT-RNA data (Extended Data Fig. 10 and Supplementary Fig. 10b). A more likely explanation for the low abundance of transcripts reported only by FLAIR (Supplementary Fig. 13) is a higher rate of incorrectly inferred sequences among them, as suggested by their low validation rate and low fraction of annotated transcripts (2.7% of FLAIR-only transcripts (TPM ≥1) compared to 69% of Ladder-seq-only transcripts (TPM ≥1)). Of transcripts upregulated in WT or KO as part of an isoform switch in our Ladder-seq analysis, 57.8% were identified by FLAIR or assembled by StringTie2 in our WT and KO ONT-cDNA datasets. Again, overlapping switching transcripts were higher expressed than uniquely identified ones (Extended Data Fig. 9b and Supplementary Fig. 10b).

For five of the six isoform switches validated by RT–qPCR (Fig. 5d), the two participating isoforms were identified by at least one of the two methods (StringTie2 or FLAIR) in the ONT-cDNA dataset (Supplementary Table 20). The single switch for which both methods independently detected both isoforms was formed by the two highest expressed transcripts. In contrast, the only isoform missed by both methods was the lowest expressed among all 12 transcripts. Overall, the two methods disagreed on the presence of six of 12 validated switching isoforms, which underlines the non-trivial nature of the computational task of inferring high-confidence transcripts from long reads. As expected, the lower sequencing depth in the ONT-RNA dataset resulted in a smaller number of confirmed isoforms (Supplementary Table 21).

Discussion

In this work, we introduced Ladder-seq, a combined experimental–computational approach that substantially improves the accuracy with which the set of expressed transcripts can be inferred from short RNA-seq reads. The experimental separation of transcripts by their lengths provides an additional layer of information that can be used by computational analysis methods to detect and quantify transcripts that cannot be distinguished based on short-read data alone. We showed that a more accurate reconstruction of the transcriptome benefits its subsequent comparison and, in our experiments, revealed isoform switches of differentially methylated transcript isoforms that are invisible to conventional RNA-seq approaches.

Our computational framework for reference-based and de novo assembly of transcripts from Ladder-seq reads employs the previously developed methods StringTie2 and Trinity without any internal modifications. We, therefore, provide a Snakemake-based[33] workflow template that allows users to implement the same framework based on other methods that have originally been developed for the analysis of conventional RNA-seq data. This will make many computational methods that have been developed over the last decade instantly available for the analysis of Ladder-seq datasets. On the other hand, we expect algorithms that are tailored to the specifics of Ladder-seq to even further improve the accuracy of reconstructed transcriptomes.

On the experimental side, the Ladder-seq protocol involves a denaturing gel electrophoresis to achieve length separation of mRNAs. In our proof-of-principle experiment, we separated transcripts into seven bands. In principle, a larger number of cuts could further reduce the effective complexity transcriptome-wide (Supplementary Fig. 14) or of a subset of genes of interest and, thus, simplify the computational task of inferring their expressed transcripts. On the other hand, fewer cuts might be sufficient to achieve a similar improvement over conventional RNA-seq for species with a less complex transcriptome. In our repository, we, therefore, provide R code that can guide the selection of the number and approximate location of cuts. We used a gel-based approach to separate transcripts because of its relative simplicity and low cost. However, the separation of mRNAs by their lengths could be achieved using other technologies, including solid-phase reversible immobilization beads[34], capillary electrophoresis[35] and ion-pair reversed-phase high-performance liquid chromatography[36]. These methods will vary in degrees of accuracy in separating mRNAs, costs and level of involvement for the experimentalist. As we showed with our simulated data experiments, a higher accuracy in the separation step will yield a greater advantage in transcriptome reconstruction.

High accuracy of Ladder-seq transcriptomes of mouse NPCs was confirmed by comparison with transcripts inferred from ONT long reads. Although the overlap between the two technologies was large, many transcripts were uniquely inferred from long reads. Their substantially lower validation rate, however, suggests the presence of a larger fraction of false transcripts. Alternatively, the low expression of transcripts uniquely identified by Ladder-seq indicates the limited sequencing depth of ONT as a possible reason for their absence in the long-read dataset.

Both differences between long-read sequencing and Ladder-seq are expected. Even though long-read technology greatly simplifies many analytical challenges that occur in short-read assembly, experimental challenges and higher error rate of long reads motivated the development of different computational strategies to extract high-confidence, full-length transcripts. Different approaches and filtering criteria can yield substantially different results[22], as observed in our own experiments using StringTie2 and FLAIR. In addition, long-read sequencers have much lower throughput and, thus, detect a much smaller fraction of genes and transcripts as contained in short-read libraries. The lower sequencing depth renders the statistical comparison of transcript abundances between conditions as performed in our study infeasible. Current studies, therefore, combine long reads with high-throughput short-read (Ilumina) sequencing[37] and limit the differential analysis to fold-change calculations[38]. Ladder-seq improves this limitation by combining the high throughput of short-read RNA-seq with the ability to reveal transcript isoforms that are invisible to conventional RNA-seq. However, if a large number of overlapping transcripts expressed by a complex gene have similar lengths, Ladder-seq will not offer any benefit over conventional RNA-seq in resolving such intrinsically difficult expression patterns from short reads.

In our Ladder-seq experiment on mouse NPCs, we explored the consequences of the deletion of m^{6}A writer protein *Mettl14* on isoform usage. Ladder-seq identified a large number of genes with isoform switches. We showed that differentially spliced exonic segments of a transcript tend to lie close to a methylation site. This result suggests a direct involvement of m^{6}A in alternative splicing in NPCs, possibly through interaction of m^{6}A readers with the splicing machinery, as it has been reported for other cell types and organisms[39–42]. Which nuclear m^{6}A reader is active in NPCs remains to be determined. An intriguing finding of our study is the enrichment for intron retention losses in *Mettl14* KO NPCs in non-neuronal genes related to DNA repair and gamete generation. Intron retentions are known to act as regulators of gene expression during normal development[43], and previous work reported progressive intron retention gains in genes related to cell cycle, pluripotency and DNA repair during the process of differentiation from mouse embryonic stem cells to neurons[44]. Expression of these genes is under tighter control as differentiation progresses. Intron retention losses in *Mettl14* KO NPCs suggest that they are in a lesser state of differentiation compared to WT NPCs, which fits with our previous finding of delayed differentiation of radial glial cells in *Mettl14* KO mice[45]. To our knowledge, this is the first in-depth analysis of m^{6}A-mediated alternative splicing in NPCs, and it highlights the diversity of m^{6}A function within a single cell type. It further extends the role of m^{6}A in NPCs from mediating mRNA degradation[45] to regulating isoform usage, which is known to be especially important in the brain.

Ladder-seq—the concerted advancement of the RNA-seq protocol and its computational methods—will allow research facilities to study the composition and dynamics of the transcriptome at an unprecedented level of accuracy based on a technology that has been established for over a decade.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-021-01136-7.

Methods

Estimating mRNA migration.

The accurate assembly and quantification of transcripts from Ladder-seq reads requires the computational estimation of transcript migration errors across bands. To estimate the migration pattern of a transcript of length ℓ through the agarose gel across k bands, we introduce probability mass function $f(x)$ over discrete random variable $x \in [k]: = \{1, \ldots, k\}$, which indicates the band to which transcripts of length ℓ migrate. If we observe reads sampled from transcripts of length ℓ in bands $X_1, \ldots, X_n \in [k]$, then we simply count how often reads are obtained in a given band and take the relative frequency as density estimate:

$$\hat{f}(i) = \frac{\sum_{j=1}^{n} \mathbb{1}(X_j = i)}{n},$$

where $\mathbb{1}$ is the indicator function that takes value 1 if its argument evaluates to true and 0 otherwise. To obtain reads for which we can infer the originating transcript with high confidence, we select reads that uniquely map to a single annotated transcript. More precisely, we run the kallisto pseudo-alignment step and select all reads that are compatible with only a single transcript according to the NH tag.

In addition, we account for potentially incomplete transcript annotations that might cause reads sampled from unannotated transcripts (of different length) to negatively affect our migration estimate of a transcript (length) that it was wrongly assigned to. To this end, we assemble transcripts using StringTie2 from reads pooled across bands and aligned using STAR. We augment the transcript annotation with novel transcripts before running the kallisto pseudo-alignment to obtain a more conservative selection of uniquely mapping reads. We do not consider reads mapping (uniquely) to newly assembled transcripts.

We further restrict observations to reads that uniquely map to protein-coding transcripts (Ensembl release 95), which are typically annotated more accurately, and which we were able to confirm to be expressed through the StringTie2 assembly on the intron chain level. We required a minimum of 50 reads to uniquely map to a transcript of length, at most, 8,000 bp to be considered in our estimation. The resulting set of reads, along with their band of origin (identified by a barcode), constitute observations $X_1, \ldots, X_n$ for the length of the transcript that they uniquely align to.

If no (high-quality) transcript catalog is available based on which uniquely mapping reads can be identified—for example, in de novo assembly or, if the species is poorly studied—mapping reads to synthetic RNA spike-in controls of varying lengths[49] can be used to similarly estimate transcript migration error. Alternatively, a supplementary sample of a well-annotated organism—for example, mRNA from a human cell line—can be run in parallel on the same gel and allocated a small fraction of sequencing output.

Because transcripts of similar length show similar migration patterns through the gel[13], we combine reads uniquely mapping to transcripts of a certain length range to more reliably estimate $f(x)$ based on a larger number of reads. Starting from the shortest transcripts, we greedily define transcript length ranges as the shortest possible length intervals longer than 100 bp that contain at least 50 different transcript species to which at least a total of 700,000 reads map uniquely. For each of these length ranges, we estimate one probability mass function $f(x)$ as described above. The resulting length ranges are listed in Supplementary Table 23.

Simulation.

We extend the widely used RNA-seq simulator RSEM[50] by an additional in silico length separation step, which includes the introduction of migration errors to simulate data with characteristics similar to that generated by our novel Ladder-seq protocol. Because the effectiveness of the experimental deconvolution of reads into different bands by Ladder-seq depends on the differences in lengths of expressed, overlapping transcripts, we simulated reads from a transcriptome using abundances and error profiles learned from a real dataset. Following the approach in ref. [14], we simulated 30 million and 75 million 2 × 75-bp paired-end reads from transcripts whose abundances were estimated by RSEM from sample NA12716_7 of the Genetic European Variation in Health and Disease (GEUVADIS)[51]. Given the RNA-seq reads produced by the simulator, we generate a matching Ladder-seq sample by assigning each read randomly to one of a fixed number of bands to introduce in silico length separation. This random assignment follows the probability mass function estimated from our NPC Ladder-seq sample KO 1, given the length of the transcript that originates the read (provided by the simulator). We use seven bands to reflect the specifics of our NPC Ladder-seq samples. See Extended Data Fig. 3 for an overview of the benchmark strategy.

To show how a more accurate experimental separation of transcripts by length can benefit transcript-level inference from Ladder-seq, we additionally simulated three Ladder-seq experiments that introduce gradually decreasing levels of migration errors. For every transcript length range for which we have estimated probability mass function $f(x)$ from our NPC Ladder-seq sample, we halve the relative frequency of reads in every band as we move further away from its mode and normalize all values to sum up to 1. More precisely, for bands numbered consecutively from 1 to k, let m denote the band that contains the mode of $\hat{f}(x)$ estimated for a given length range. Then,

$$f^1(i) = \frac{\hat{f}(i)/2^{|i-m|}}{\sum_{j=1}^{k} \hat{f}(j)/2^{|i-m|}} \tag{1}$$

Similarly, $f^2(x)$ and $f^3(x)$ are obtained by replacing $\hat{f}$ in (1) by $f^1(x)$ and $f^2(x)$, respectively. By randomly assigning simulated reads according to probability mass functions $f^i(x)$, $i = 1,\ \ 2,\ \ 3$, instead of $\hat{f}(x)$, we obtain three additional Ladder-seq datasets with reduced levels of migration errors.

Finally, we simulated a most-optimistic Ladder-seq experiment that is able to perfectly separate transcripts by length, without introducing any migration error. This leads to a degenerate probability mass function for each length range implied by the seven in silico cuts in which the read band is a constant random variable that takes only a single value, the correct band corresponding to that length range.

Evaluation.

We used the same metrics as in a benchmark of transcript quantification methods[1] to measure the accuracy of kallisto and kallisto-ls estimates of transcript expression. MARD denotes the arithmetic mean of absolute relative differences, calculated as $|i - j|/(i + j)$ for estimated and ground truth counts i and j, respectively. We excluded transcripts with zero estimates by both methods—that is, if $i + j = 0$. Pearson correlation was calculated between $\log_2$ transformed TPM values, after adding 0.1 TPM.

Consistent with previous studies[17,52], the accuracy of reference-based and de novo assemblies is evaluated using sensitivity defined as TP/(TP+FN) and precision defined as TP/(TP+FP), where true positives (TPs) denote correctly assembled transcripts; false negatives (FNs) denote true transcripts missing in the assembly; and false positives (FPs) denote wrongly assembled transcripts. We considered a transcript truly expressed if reads sampled by RSEM in the 30-million-reads dataset fully cover the transcript and if it was estimated by RSEM to be expressed in GEUVADIS sample NA12716_7 with at least 0.1 TPM. An identical ground truth transcriptome facilitates comparison of sensitivity and precision values between different sequencing depths and between reference-based and de novo assemblies. As in refs. [53,54], we used GffCompare[55] to compare transcripts assembled by StringTie2 or StringTie-ls to truly expressed transcripts. GffCompare defines an assembled transcript as correct if it shares the same sequence of introns with a true transcript. In the de novo assembly benchmark, correct assemblies by Trinity and Trinity-ls needed to be identified through an alignment of their sequences, which we computed using BLAT[56]. Applying commonly used criteria[57,58], we require the sequences to align with 95% identity and, at most, 1% insertion and deletion rate and apply transcript coverage cutoffs of 80%, 85%, 90% and 95%.

Transcript quantification by kallisto-ls.

After estimating migration patterns in a Ladder-seq sample using the histogram-based method described above, kallisto-ls uses an EM algorithm similar to that of kallisto to infer maximum likelihood estimates of transcript abundances in our statistical model of Ladder-seq. kallisto is based on the following likelihood function[14] of RNA-seq:

$$L(\alpha) \propto \prod_{e \in E} \left(\sum_{t \in e} \frac{\alpha_t}{l_t} \right)^{c_e} \tag{2}$$

It counts the number of fragments c_e that cannot be distinguished by the set of transcripts e that they are compatible with and are, thus, considered equivalent. l_t denotes the effective length[59] of transcript t, and parameters α_t denote the probability of obtaining a fragment from a transcript t.

In Ladder-seq, we observe fragments that originate from transcripts in different bands. The probability of obtaining a fragment from a transcript t in band b, then, is $\alpha_t\beta_{tb}$, where β_{tb} denotes the fraction of transcript t in band b, which we precompute in $\hat{f}(b)$ for each range of transcript lengths as described above. If we split equivalence class counts c_e between k different bands, that is

$$c_e = \sum_{b=1}^{k} c_{eb},$$

then the likelihood function for Ladder-seq becomes:

$$L(\alpha) \propto \prod_{e \in E} \prod_{b=1}^{k} \left(\sum_{t \in e} \frac{\alpha_t\beta_{tb}}{l_t} \right)^{c_{eb}} \tag{3}$$

The observed data likelihood remains a concave function under this adjustment (see next section), provided we precompute the extent of migration errors. We can, thus, extend the EM algorithm implemented in kallisto to find the values of α that maximize likelihood (3). The EM algorithm alternates between fractionally assigning fragments to transcripts in different bands based on current parameter estimates and recalculating parameters from these fragment assignments. Consistent with the original kallisto implementation, the EM algorithm terminates if $\alpha_t N$ has changed by less than 1% compared to the previous iteration for every transcript t with $\alpha_t N > 0.01$, where N is the total number of fragments.

Proof of concavity of Ladder-seq likelihood.—The log-likelihood function of Ladder-seq is:

$$\ln(L(\alpha)) = \sum_{e \in E} \sum_{b=1}^{k} c_{eb} \ln\left(\sum_{t \in e} \frac{\alpha_t\beta_{tb}}{l_t} \right). \tag{4}$$

For arbitrary but fixed $e \in E$ and $b \in [k]$, we define

$$f(\alpha) = c_{eb}\ln\left(\sum_{t \in e} \frac{\alpha_t \beta_{tb}}{l_t}\right). \tag{5}$$

Analogous to ref. [60], we prove in the following that $f(\alpha)$ is concave, from which it follows that ln $(L(\alpha))$ is concave too. Let $H(\alpha)$ represent the Hessian matrix of function $f(\alpha)$:

$$H_{jk}(\alpha) = \frac{\partial^2 c_{eb}\ln\left(\sum_{t \in e} \frac{\alpha_t \beta_{tb}}{l_t}\right)}{\partial \alpha_j \partial \alpha_k} \tag{6}$$

$$= -c_{eb}\frac{\beta_{jb}\beta_{kb}}{l_j l_k}\frac{1}{\left(\sum_{t \in e} \frac{\alpha_t \beta_{tb}}{l_t}\right)^2} \tag{7}$$

Then, we can rewrite $H(\alpha) = -z(\alpha)x^T x$, where

$$z(\alpha) = \frac{c_{eb}}{\left(\sum_{t \in e} \frac{a_t \beta_{tb}}{l_t}\right)^2} \text{ and} \tag{8}$$

$$x = \left(\frac{\beta_{1b}}{l_1}, \frac{\beta_{2b}}{l_2}, \frac{\beta_{3b}}{l_3}, \ldots, \frac{\beta_{|e|b}}{l_{|e|}}\right). \tag{9}$$

Because $z(\alpha) > 0$, we have for all $y = (y_1, y_2, \ldots, y_{|e|})$:

$$yH(\alpha)y^T = y\left(-z(\alpha)x^T x\right)y^T \tag{10}$$

$$= -z(\alpha)\left(yx^T\right)\left(xy^T\right) \tag{11}$$

$$= -z(\alpha)\left(yx^T\right)^2 \tag{12}$$

$$\leq 0 \tag{13}$$

Thus, $H(\alpha)$ is negative semi-definite, and $f(\alpha)$ is concave.

Reference-based transcript assembly by StringTie-ls.

Reads from all bands are aligned to the reference genome sequence using a short-read aligner, such as STAR[47]. For every band, and every union of two consecutive bands, we assemble transcripts using StringTie2 with default options. We additionally pool reads from neighboring bands to recover potentially low-expressed transcripts that migrated close to the boundary between two bands.

StringTie-ls estimates migration patterns in a Ladder-seq sample using the histogram-based approach described above. It uses these estimates to identify too-short transcript fragments and too-long transcript fusions. More precisely, for a transcript t of length ℓ assembled in the j-th band, we look up the probability mass function $f(x)$ corresponding to the length range that contains ℓ to determine the most likely band b_i to which a transcript of length ℓ would have migrated to. If $j \neq i$ and $j \neq i + 1$, we remove t. Note that band b_{i+1} corresponds to the next longer range of transcripts but can also contain slightly shorter transcripts from band b_i due to secondary structure effects. Similarly, if t was assembled in the combination of bands j and $j + 1$, we discard t if $j < i$ or $j > i + 2$. To account for potential overlap with longer UTRs, we do not remove too-long transcripts assembled in a band $i + 2\ldots7$ if they are sufficiently highly expressed (>1 TPM), if they contain a unique intron and if their first or last exon is longer than 500 bp.

The individual assemblies are subsequently merged using the GffCompare tool, which computes the union of all intron chains. In other words, transcripts that imply the same sequence of introns as a transcript assembled in a different band are discarded. We further eliminate single-exon transcripts that are identified as redundant by the merge mode of StringTie2 as well as transcript fragments that are fully contained in other transcripts with compatible intron chains. These transcripts most likely constitute transcript fragments that were only partially assembled from reads obtained from transcripts that migrated to a different band. We retain, however, transcripts with identical (partial) intron chains if they start or end within an intron of the containing transcript, unless a very small overhang of, at most, 2 bases indicates noisy read alignments. Finally, we quantify assembled transcripts using our statistical model of Ladder-seq implemented in kallisto-ls and report transcripts estimated to be expressed with at least 0.1 TPM.

De novo transcript assembly by Trinity-ls.

The Ladder-seq-based de novo assembly follows a very similar scheme as applied in the reference-based assembly. We run Trinity on the reads from each band separately using default parameters. In contrast to the reference-based assembly, we do not pool reads from neighboring bands because the absence of a reference genome makes it more difficult to subsequently detect and remove false-positive transcripts. After estimating migration patterns from Ladder-seq data using the histogram-based method described above, Trinity-ls applies length constraints to assembled transcripts following the same strategy as in the reference-based approach. It then concatenates the individual assemblies, because the absence of a reference genome does not allow detection of potential redundancy with respect to the exon–intron structure of transcripts. Again, Trinity-ls quantifies assembled transcripts using our statistical model of Ladder-seq implemented in kallisto-ls and applies an expression threshold of 0.1 TPM.

Animals.

All animal procedures used in this study were performed in accordance with the protocol approved by the Institutional Animal Care and Use Committee of Johns Hopkins University School of Medicine.

Mettl14 conditional KO mice contain a deletion of exons 7, 8 and 9 in the developing mouse nervous system starting at embryonic day (E) 11.5. Deletion was achieved using the *Nestin-Cre*;*Mettl14*$^{f/f}$ cKO model[45].

Mettl14 floxed mice were crossed with *Nestin-Cre* mice and maintained in C57BL/6J background before all experiments. E14.5 embryos were collected (three *Nestin-Cre*$^{+/+}$;*Mettl14*$^{f/f}$ or three *Nestin-Cre*$^{-/+}$;*Mettl14*$^{f/f}$, respectively) to isolate NPCs from the forebrains. Mice were bred and maintained under specific pathogen-free conditions and kept at an ambient temperature of 21 °C and humidity of 40–60% under a 12-h light/dark cycle with standard chow diet.

Primary mouse NPCs.

Mouse NPCs were isolated from *Mettl14* WT and cKO mouse embryonic cortices and cultured in Neurobasal medium (Gibco BRL) containing 20 ng ml^{-1} of FGF2, 20 ng ml^{-1} of EGF, 5 mg ml^{-1} of heparin, 2% B27 (v/v, Gibco BRL), GlutaMAX (Invitrogen) and penicillin–streptomycin (Invitrogen) on culture dishes precoated with Matrigel matrix (2%, Corning).

Generation of Ladder-seq libraries from mouse NPCs.

Total RNA fraction was isolated from cultured NPC samples with TRIzol reagent (Thermo Fisher Scientific) followed by total RNA extraction using RNA Clean and Concentrator-25 (Zymo Research). mRNA was isolated from total RNA with Dynabeads mRNA Purification Kit (Thermo Fisher Scientific).

Next, 5 μg of mRNA per sample was loaded in each well of a denaturing agarose gel (MOPS/2% formaldehyde). NEB single-stranded RNA ladder was loaded in the wells flanking the samples for guidance in the gel slicing step. Gel electrophoresis was run at 100 V for 55 min in chilled 1× MOPS buffer. Gel was stained with SYBR Gold (Thermo Fisher Scientific) and visualized under ultraviolet light for slicing. Each sample was sliced into seven fractions (bands) by slicing the gel at the following approximate lengths: 1,000 bp, 1,500 bp, 2,000 bp, 3,000 bp, 4,000 bp and 6,000 bp. mRNA was extracted from gel slices using the Zymoclean Gel RNA Recovery Kit (Zymo Research) with gel incubation at room temperature. RNA-seq libraries were prepared with the NEBNext Ultra II RNA Library Prep Kit for Illumina, and each band of each sample used a unique index PCR primer (NEBNext Multiplex Oligos for Illumina). Libraries were multiplexed 1:1 for sequencing in the NextSeq 500 (Illumina), yielding approximately 100 million 2 × 75-bp paired-end reads per sample.

Nanopore direct cDNA sequencing.

For nanopore direct cDNA sequencing, two biological replicates per genotype (WT and *Mettl14* KO) were prepared from mouse NPCs. Total RNA was extracted from cultured NPC pellets with TRIzol reagent (Invitrogen), treated with DNase I (Takara) and cleaned up using RNeasy MinElute (Qiagen). Each 1.5 μg of purified total RNA with 0.1 μl of RCS from direct RNA-seq kit (SQK-RNA002) and 0.1 ng of the SIRV set 3 (Lexogen) control was prepared as a sequencing library following manufacturer instructions (SQK-DCS109, ONT), with some modifications as follows. A mixture of 1 μl each of RNase T1 (1 U μl^{-1}, Invitrogen) and RNase A (10 mg ml^{-1}, Thermo Fisher Scientific) was treated to degrade RNA after reverse transcription. Second-strand synthesis was carried out with 10 U of DreamTaq Hot Start DNA Polymerase (5 U μl^{-1}, Thermo Fisher Scientific) with 5 μl of 10× DreamTaq buffer and 2 μl of dNTP Mix (10 mM each, Thermo Fisher Scientific) by incubating the mixture at 95 °C for 90 s, 50 °C for 30 s and 72 °C for 20 min. The libraries were sequenced in parallel with four R9.4.1 flowcells (FLO-MIN106D, ONT) and separate MinION Mk1b devices (controlled by MinKNOW 4.1.2, ONT). The basecalls were produced offline using guppy 4.5.2 with ONT's high-accuracy model, yielding approximately 5.8 million and 4.9 million reads in WT and KO NPCs, respectively.

Nanopore native RNA sequencing.

For nanopore native RNA sequencing, two biological replicates per genotype (WT and *Mettl14* KO) were prepared from mouse NPCs. Total RNA was extracted using TRIzol from cell pellets following the manufacturer's protocol (Invitrogen). After DNase I treatment (Takara), the reaction was cleaned up using RNeasy MinElute (Qiagen). Each 4 μg of the purified total RNA was prepared as a sequencing library for direct RNA-seq by the standard kit (SQK-RNA002, ONT). The libraries were loaded onto R9.4.1 flowcells (FLO-MIN106D, ONT) and sequenced using four MinION Mk1b devices separately in parallel (MinKNOW 4.1.2, ONT). Acquired squiggles were basecalled offline using guppy 4.4.1 with the 'res_rna2' flipflop model (ONT), yielding approximately 2.1 million and 1.8 million filtered high-quality reads in WT and KO NPCs, respectively.

Reconstruction of WT and *Mettl14* NPC transcriptomes.

We used StringTie-ls to reconstruct novel transcripts expressed in WT and *Mettl14* NPCs but employ Ladder-seq replicates and the well-annotated mouse reference transcriptome (Ensembl release 95) to obtain high-quality transcriptomes for the two conditions. More specifically, we assemble transcripts using StringTie-ls in each sample independently and consider all transcripts that do not match any annotated transcript in their sequence of introns as novel. Among these novel transcript structures, we retain those that occur in at least three of the four replicates per genoypte and merge the two resulting sets of transcripts to a high-confidence set of novel transcripts across genotypes. We add these novel transcripts to the mouse reference transcriptome and use this catalog of transcripts in all subsequent analyses of NPC samples. We apply the same procedure when comparing the outcomes to the conventional RNA-seq analysis but replace StringTie-ls by conventional StringTie2. Depending on the quality of the reference annotation, a stepwise addition of novel isoforms as in AIDE[5] can help prioritize annotated transcripts in the subsequent quantification.

To compute the number and rate of detected annotated transcripts (Ensembl release 95) in a Ladder-seq or RNA-seq dataset, we quantified transcript abundance using conventional kallisto, pooling reads from different bands in Ladder-seq. A transcript was considered detected if its estimated count was at least 1.

Differential isoform usage analysis.

Abundance estimates per sample were obtained with kallisto-ls. The R Bioconductor package IsoformSwitchAnalyzeR[61] was used for differential isoform usage (DIU) analysis. Identification of differentially used isoforms across all genes with IsoformSwitchAnalyzeR is done through DEXseq[62], which is a statistical method originally developed for differential exon usage based on the likelihood ratio test that has since been shown to adequately control for false discovery rate (FDR) in the setting of DIU. Analysis of consequences of isoform switches was performed through IsoformSwitchAnalyzeR with the function analyzeSwitchConsequences. This function allows the addition of input data from CPAT[63] for analysis of coding potential and from PfamScan[64] for protein domain annotation.

Analysis of published m^{6}A sequencing from mouse NPCs.

We built a set of high-confidence m^{6}A peaks from a publicly available dataset of m^{6}A sequencing in mouse NPCs[23]. BED files containing peaks called by MACS2 (ref. [65]) from two replicates with two input samples each were downloaded from the National Center of Biotechnology Information’s Gene Expression Omnibus (GSE104686). Using BEDTools intersect[66], we identified peaks that were reproducible in both replicates with both input controls. We then annotated these high-confidence peaks using the annotatePeaks.pl program from the Homer suite[67] to identify the genes harboring m^{6}A methylation.

Analysis of m^{6}A enrichment at differentially spliced regions.

Pairs of switching isoforms from m^{6}A methylated genes were partitioned into minimal exonic segments that are bounded by splice sites, transcription start sites or transcription end sites of the two involved transcripts. These segments represent the largest exonic fragments that are entirely contained in one or both of the two transcripts. A segment bounds a differentially splice region if it is part of only one of the two transcripts, if it is not the first or last segment of that transcript and if it is adjacent to a segment that is contained in both transcripts. We take into account the length of segments in Fisher’s exact test by distinguishing individual bases that can lie within or outside of bounding segments and that can be methylated or not.

GO enrichment analysis.

All GO enrichment analyses were performed using the R Bioconductor package TopGO[68]. Only genes passing the pre-filtering step for differential isoform usage (TPM >1) were considered for the gene universe.

Splicing analysis.

Alternative splicing analysis was performed using the IsoformSwitchAnalyzeR R Bioconductor package[61] with the functions extractSplicingSummary, which summarizes the types of alternative splicing events occurring in each isoform switch, and extractSplicingEnrichment, which identifies the uneven usage of a particular alternative splicing type in one of the conditions assayed.

Processing of ONT long-read libraries.

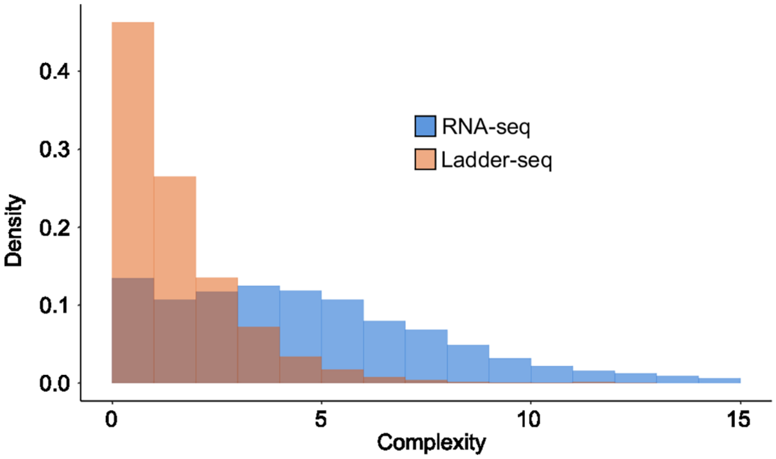
ONT reads were aligned to the Ensembl mouse genome assembly GRCm38 using minimap2 version 2.17-r941. Following recommendations at https://github.com/lh3/minimap2, we used option `-ax splice` to allow spliced alignments and provided splice junctions extracted from the corresponding Ensembl release 95 transcriptome annotation with parameter `–junc-bed`. In the alignment of native RNA reads, we additionally used options `-k14 -uf` as recommended. We used FLAIR version 1.5.1 (ref. [32]) to identify transcripts and StringTie2 (ref. [17]) to assemble transcripts from ONT reads. We ran FLAIR with default settings on pooled reads from both WT and KO replicates and extracted condition-specific transcripts that had an estimated count of at least 1 in at least one of the two replicates per condition. FLAIR uses minimap2 internally to align reads using options `-ax splice -t 8 –secondary=no` and corrects misaligned splice sites using the Ensemble 95 annotation. It groups corrected reads with identical intron chains while comparing TSS/TSE with a window size of 100 bp, collapsing them to representative transcripts. It retains transcripts with at least three aligned reads with a minimum MAPQ of 1. StringTie2 was run with the `-L` option (for long reads) on each of two BAM files generated, respectively, from pooled replicates of two conditions. GffCompare version 0.10.4 was used to compare transcripts between ONT datasets and with transcripts assembled in Ladder-seq. Transcripts were considered identical if they shared the same sequence of introns.

To quantify expression and to compute the number and rate of detected annotated transcripts (Ensemble release 95) in an ONT dataset, we followed the strategy proposed in ref. [12]. We aligned reads to the mouse cDNA sequences from Ensembl GRCm38.95 using minimap2 with options `-ax map-ont` and quantified their expression using salmon version 1.2.1 with options `-l A` and `–noErrorModel`. A transcript was considered detected if its estimated count was at least 1.

RT–qPCR analysis.

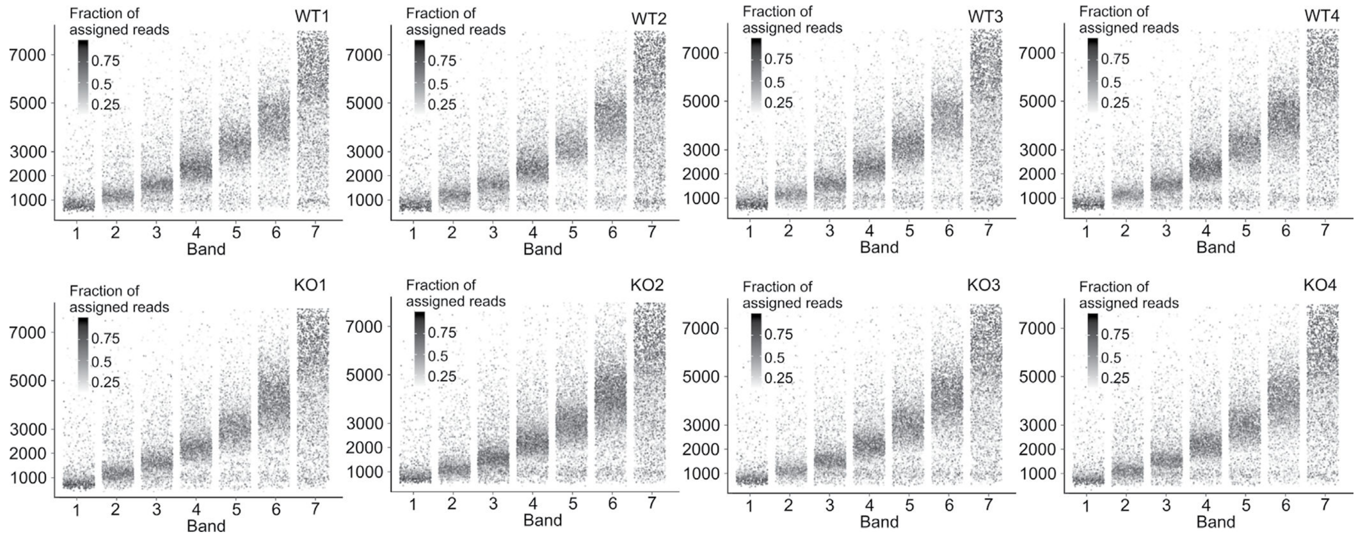
For relative isoform expression analysis, total RNA was isolated from biological triplicate WT and *Mettl14* KO NPC samples using the RNeasy Plus Mini Kit (Qiagen) and treated with DNAseI. Equal amounts of total RNA from each sample were then reverse transcribed using SMARTScribe Reverse Transcriptase (Takara). Relative isomer expression was measured by quantitative real-time PCR on a 7500 Real-Time PCR system (Applied Biosystems) by adding SYBR Green to cDNA and using custom primers unique to each isomer of interest (Supplementary Table 22). For each gene, three isomers were tested: one common isomer identified in both WT and *Mettl14* KO NPC RNA-seq data and two distinct isomers with differential expression between WT and *Mettl14* KO NPC RNA-seq data. All samples were tested in triplicate and normalized to β-actin. The expression of the differentially expressed isomers was normalized to the expression level of the shared isomer, which consistently showed no significant difference between WT and KO NPCs.

Extended Data



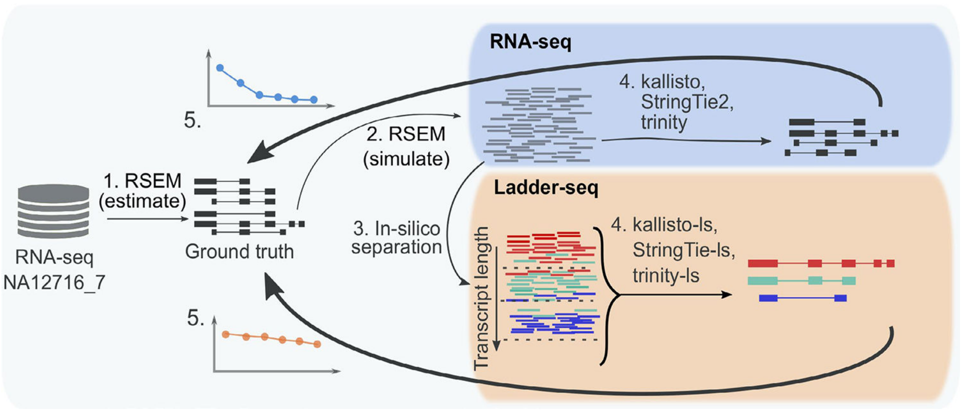
Extended Data Fig. 1 |. Reduced (effective) gene complexity in Ladder-seq.

We estimate transcript expression in *Mettl14* KO sample 1 using kallisto on Ladder-seq reads pooled across bands and show the histogram of gene complexity measured as the number of expressed transcripts. In Ladder-seq, we partition the set of expressed transcripts into 7 bands and count the number of transcripts contained in each band according to their annotated length (plus 200 nt average poly(A) tail size[46]), assuming cuts at 1000 bp, 1500 bp, 2000 bp, 3000 bp, 4000 bp and 6000 bp. The resulting histogram of effective gene complexity shows an increased fraction of gene bands with low complexity.



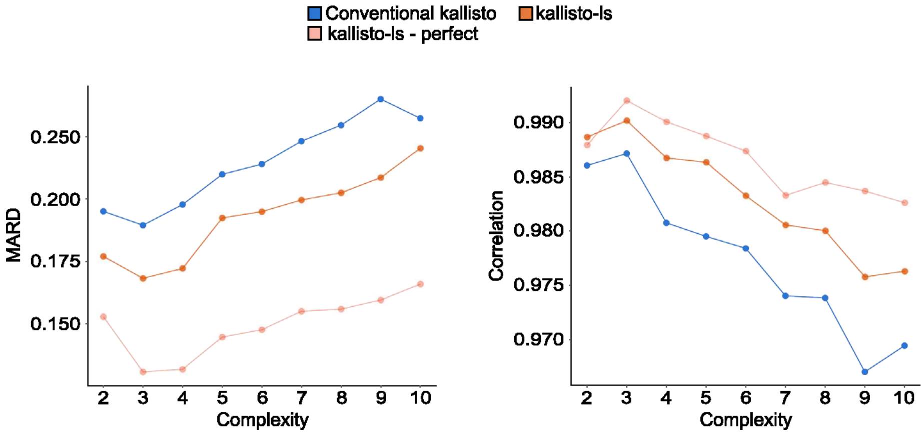
Extended Data Fig. 2 |. In silico gel for WT and Ko NPC samples.

For every annotated transcript the intensity of a point with y- coordinate equal to its annotated length (plus 200 nt average poly(A) tail size) shows the fraction of reads obtained from each band (x-axis) that can be assigned unambiguously to it. As expected, each band contains predominantly reads from transcripts of a distinct length range.



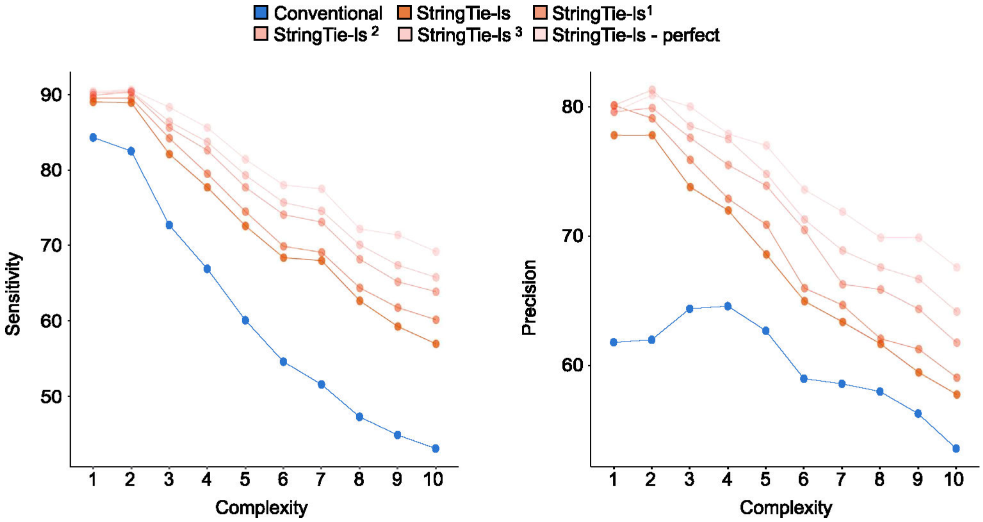
Extended Data Fig. 3 |. Overview of the benchmark strategy.

1. The ground truth transcriptome including abundances and error profile is calculated by RSEM from GEUVADIS sample NA12716_7. 2. Reads are simulated from the ground truth transcriptome by RSEM to obtain RNA-seq samples of different sequencing depths. 3. A matching Ladder-seq sample is obtained by separating reads in silico according to probability mass functions estimated from our NPC Ladder-seq sample (and variants thereof). 4. Transcripts are quantified and assembled by our Ladder-seq tailored transcript analysis methods kallisto-ls, StringTie-ls, and Trinity-ls from the Ladder-seq sample, while their conventional counterparts are run on the corresponding RNA-seq sample. 5. The results are compared to the ground truth to evaluate and compare their accuracy.



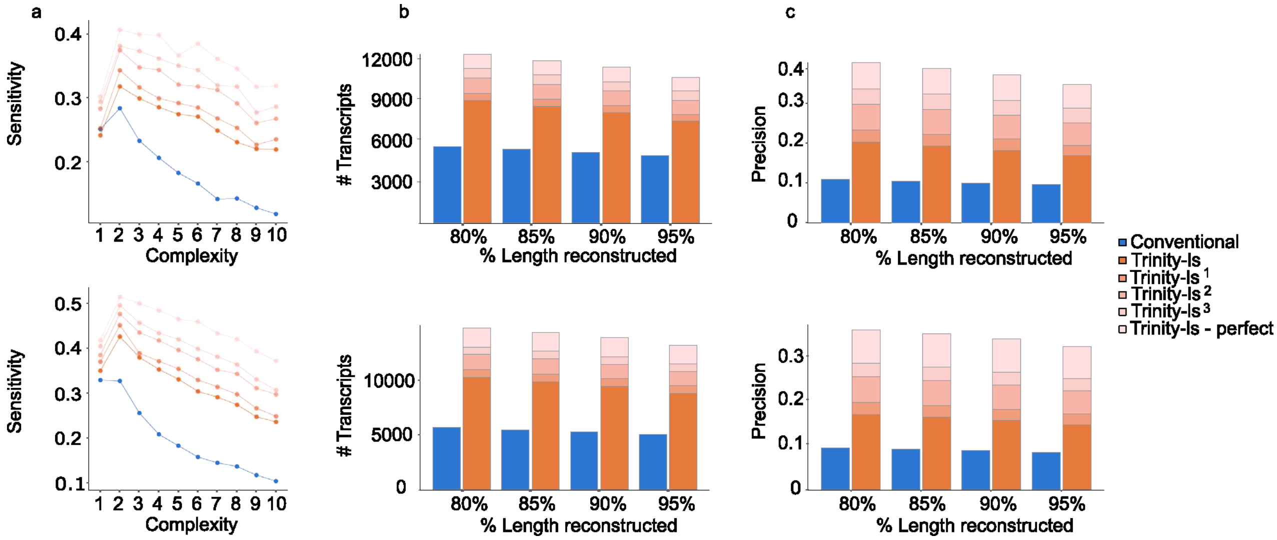
Extended Data Fig. 4 |. Quantification accuracy of kallisto-ls on 75 million simulated reads.

Mean values across 20 simulations are reported. Pearson correlation of estimated and ground truth abundance in log2 transformed transcripts per million (TPM) and mean absolute relative difference (MARD) are shown as a function of gene complexity, that is the number of transcripts expressed by a gene. For ease of visualization, we omit genes expressing a single transcript, many of which are estimated to be lowly expressed in GEUVADIS sample NA12716_7 by RSEM.



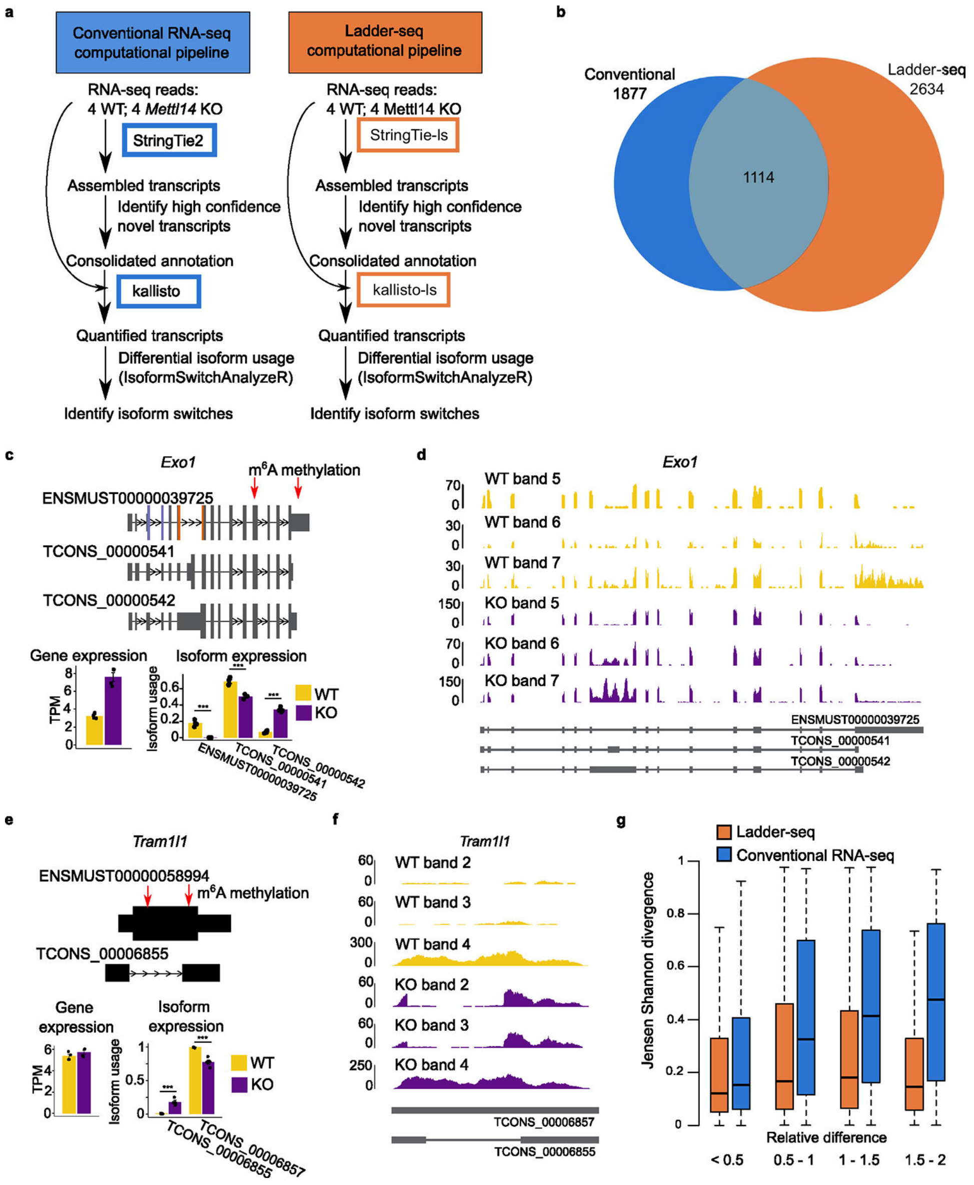
Extended Data Fig. 5 |. Accuracy of transcript assembly from 75 million simulated reads.

RNA-seq and Ladder-seq reads were aligned identically to the reference genome (GRCh38) using STAR. Sensitivity and precision of StringTie-ls and its conventional counterpart StringTie2 are shown as a function of gene complexity measured as the number of expressed transcripts. Sensitivity and precision are calculated with respect to the same set of ground truth transcripts as in the smaller 30 million read pairs data set.



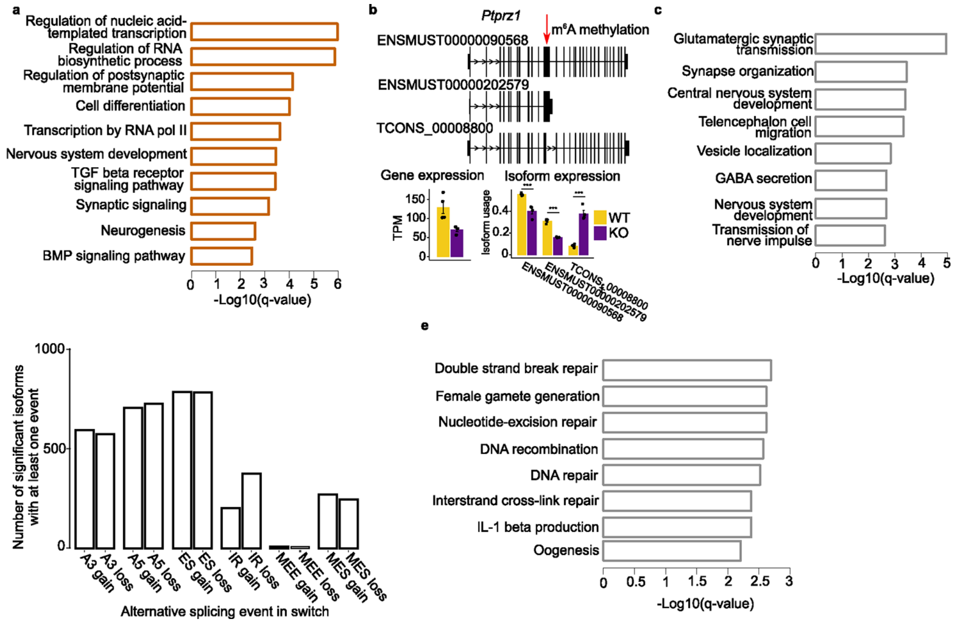
Extended Data Fig. 6 |. Accuracy of *de novo* transcript assembly from 30 million (top row) and 75 million (bottom row) simulated reads.

(a) Sensitivity of Trinity-ls and its conventional counterpart Trinity at 90% transcript length cut-off is shown as a function of gene complexity measured as the number of expressed transcripts. (b) Total number of correctly assembled transcripts at different transcript length cut-offs. (c) Precision at different transcript length cut-offs.



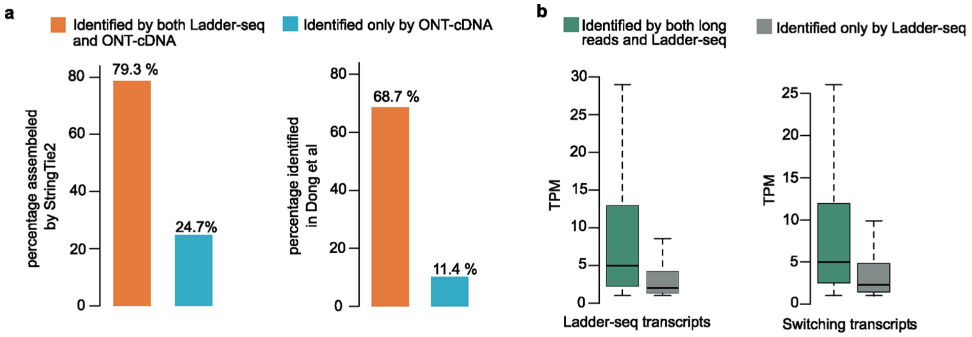
Extended Data Fig. 7 |. Ladder-seq improves differential analysis of reconstructed transcriptomes.

(a) Computational pipeline for differential isoform usage analysis with conventional RNA-seq and Ladder-seq. Reads were aligned using STAR aligner prior to transcript assembly for both pipelines. (b) Venn diagram showing overlap between switching genes identified by Ladder-seq and conventional RNA-seq. (c and e) Isoform switches identified only by Ladder- seq in genes *Exo1* and *Tram1l1* (between n=4 WT and n=4 KO samples). Red arrows show location of m^{6}A methylation. TCONS_00000541 and TCONS_00000542 are novel isoforms of *Exo1* detected only by Ladder-seq. TCONS_00006855 is a novel isoform of *Tram1l1* that was assembled by both methods, but conventional RNA-seq failed to identify the isoform switch. Without length information, conventional RNA-seq reads in KO bands 2 and 3 were predominantly assigned to the annotated transcript in band 4. Barplots represent mean ± SEM; ***FDR corrected p<0.001. (d and f) Coverage plots for switching genes *Exo1* and *Tram1l1* showing separation of reads from transcripts of different lengths. (g) Jensen Shannon divergence for Ladder-seq and conventional RNA-seq for all identified transcripts grouped by relative difference in abundance estimation by both methods (n=18761 for <0.5, n=12722 for 0.5–1, n=7918 for 1–1.5, n=6292 for 1.5–2 relative difference). Relative difference is defined as the absolute difference in estimated transcript abundance (in TPM) divided by the average of the two. Boxplot definition: Bottom and top of the box correspond to lower and upper quartiles of the data, bar is the median and whiskers are median ± 1.5x interquartile range.



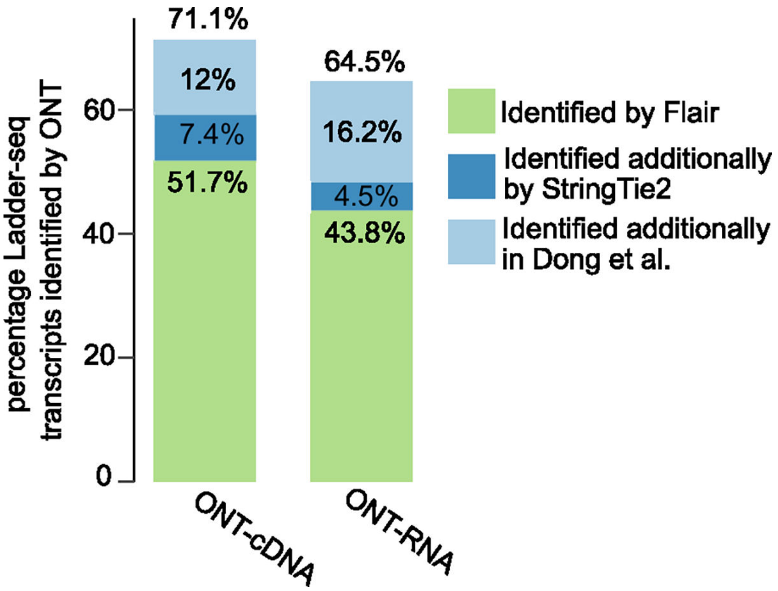
Extended Data Fig. 8 |. *Mettl14* Ko leads to isoform switches in m^{6}A methylated genes.

(a) Gene Ontology for m^{6}A methylated genes containing isoform switches. (b) Isoform switch in *Ptprz1* (between n=4 WT and n=4 KO samples). Red arrow shows location of m^{6}A methylation. Barplots represent mean ± SEM; ***FDR corrected p<0.001. (c) Gene Ontology analysis for genes with loss of protein domains in KO NPCs. (d) Splicing analysis: Number of gains and losses of each splicing event in KO NPCs. A3: Alternative 3’ acceptor site; A5: Alternative 5’ acceptor site; ES: Exon skipping; IR: Intron retention; MEE: Mutually exclusive exon; MES: Multiple exon skipping. (e) Gene Ontology enrichment analysis of genes with intron retention loss in KO NPCs.



Extended Data Fig. 9 |. Comparison of Ladder-seq and oNT-cDNA sequencing on mouse NPCs.

(a) Orange bars show validation by StringTie2 (left panel) or by an independent ONT dataset (Dong et al.) (right panel) of transcripts found by both Ladder-seq and ONT-cDNA while light blue bars show validation values for transcripts reported only by ONT-cDNA. In comparison, 32.5% of transcripts uniquely identified by Ladder-seq (average TPM ≥ 1) were also identified in the dataset by Dong et al. (b) Boxplots showing expression levels (TPM) for transcripts identified both by long- reads and Ladder-seq (green boxes) and for transcripts identified only by Ladder-seq (grey boxes). Left panel shows values for all Ladder-seq transcripts with TPM higher than 1 (n=6169 identified only by Ladder-seq, n=15099 by both). Right panel shows values for Ladder-seq switching transcripts with TPM higher than 1 (n=905 identified only by Ladder-seq, n=2012 by both). Boxplot definition: Bottom and top of the box correspond to lower and upper quartiles of the data, bar is the median and whiskers are median ± 1.5x interquartile range.



Extended Data Fig. 10 |. Cumulative percentage of Ladder-seq transcripts identified by long-read sequencing.

Bars show percentage of Ladder-seq transcripts identified by FLAIR (green), plus those additionally identified by StringTie2 (blue), plus transcripts additionally found in a recently published long-read mouse NPC transcriptome (light blue).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We would like to thank A. A. Kha for helpful discussions and the members of the Canzar laboratory for comments and suggestions. We thank P. J. Shaw and M. Halic for constructive feedback on mRNA length separation. We thank S. Esser for excellent technical assistance. S. Chakraborty was supported by a Deutsche Forschungsgemeinschaft (DFG) fellowship through the Graduate School of Quantitative Biosciences Munich. F.R.R. was supported, in part, by the Bavarian Gender Equality Grant. This work was funded by the TTIC Research fund (to S. Canzar) and by grants from the National Institutes of Health (R35NS097370 to G.M. and R35NS116843 to H.S), the Simons Foundation (575050 to H.S.), the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) SFB-TRR 338/1 2021-452881907 (to S. Canzar), DFG Sachbeihilfe (417912216 to M.L.S.), Deutsche Gesellschaft für Muskelkranke e.V. Forschungsförderung (awarded in 2019 to M.L.S.), the Institute for Basic Science (IBS-R008-D1 to A.H. and H.C.), the National Research Foundation of Korea (2019R1C1C1006600 and 2020M3A9E4039670 to K.-J.Y. and 2021R1C1C1010758 to A.H. and H.C.), funded by the Ministry of Science, and by ICT of Korea, Young Investigator Grant, from the Suh Kyungbae Foundation (to K.-J.Y.).

Data availability

Ladder-seq raw sequencing data from WT and *Mettl14* KO mouse NPCs, conventional RNA-seq data from WT mouse NPCs and ONT long-read sequencing from WT and Mettl14 KO mouse NPCs are available in the Gene Expression Omnibus (GSE158985). m^{6}A peaks from two replicates of m^{6}A sequencing in mouse NPCs were downloaded from the Gene Expression Omnibus (GSE104686).

References

1. Zhang C, Zhang B, Lin L-L & Zhao S Evaluation and comparison of computational tools for RNA-seq isoform quantification. BMC Genomics 18, 583 (2017). [PubMed: 28784092]
2. Teng M et al. A benchmark for RNA-seq quantification pipelines. Genome Biol. 17, 74 (2016). [PubMed: 27107712]
3. Aguiar D et al. Bayesian nonparametric discovery of isoforms and individual specific quantification. Nat. Commun. 9, 1681 (2018). [PubMed: 29703885]
4. Song L, Sabunciyan S, Yang G & Florea L A multi-sample approach increases the accuracy of transcript assembly. Nat. Commun. 10, 5000 (2019). [PubMed: 31676772]
5. Li WV et al. AIDE: annotation-assisted isoform discovery with high precision. Genome Res. 29, 2056–2072 (2019). [PubMed: 31694868]
6. Desrosiers RC, Friderici KH & Rottman FM Characterization of novikoff hepatoma mRNA methylation and heterogeneity in the methylated 5′ terminus. Biochemistry 14, 4367–4374 (1975). [PubMed: 169893]
7. Barbosa-Morais NL et al. The evolutionary landscape of alternative splicing in vertebrate species. Science 338, 1587–1593 (2012). [PubMed: 23258890]
8. Jelen N, Ule J, Živin M & Darnell RB Evolution of nova-dependent splicing regulation in the brain. PLoS Genetics 3, e173 (2007). [PubMed: 17937501]
9. Merkin J, Russell C, Chen P & Burge CB Evolutionary dynamics of gene and isoform regulation in mammalian tissues. Science 338, 1593–1599 (2012). [PubMed: 23258891]
10. Tardaguila M et al. SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. Genome Res. 28, 396–411 (2018). [PubMed: 29440222]
11. Chen K et al. Genome-wide binding and mechanistic analyses of Smchd1-mediated epigenetic regulation. Proc. Natl Acad. Sci. USA 112, E3535–E3544 (2015). [PubMed: 26091879]

12. Soneson C et al. A comprehensive examination of Nanopore native RNA sequencing for characterization of complex transcriptomes. Nat. Commun. 10, 3359 (2019). [PubMed: 31366910]
13. Hurowitz EH & Brown PO Genome-wide analysis of mRNA lengths in *Saccharomyces cerevisiae*. Genome Biol. 5, R2 (2003). [PubMed: 14709174]
14. Bray NL, Pimentel H, Melsted P & Pachter L Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. 34, 525 (2016). [PubMed: 27043002]
15. Heber S, Alekseyev M, Sze S-H, Tang H & Pevzner PA Splicing graphs and EST assembly problem. Bioinformatics 18, S181–S188 (2002).
16. Pachter L Models for transcript quantification from RNA-seq. Preprint at https://arxiv.org/abs/1104.3889 (2011).
17. Kovaka S et al. Transcriptome assembly from long-read RNA-seq alignments with Stringtie2. Genome Biol. 20, 278 (2019). [PubMed: 31842956]
18. Ebrahim Sahraeian SM et al. Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis. Nat. Commun. 8, 59 (2017). [PubMed: 28680106]
19. Glinos DA et al. Transcriptome variation in human tissues revealed by long-read sequencing. Preprint at https://www.biorxiv.org/content/10.1101/2021.01.22.427687v1 (2021).
20. Grabherr MG et al. Full-length transcriptome assembly from RNA-seq data without a reference genome. Nat. Biotechnol. 29, 644–652 (2011). [PubMed: 21572440]
21. Chang Z, Wang Z & Li G The impacts of read length and transcriptome complexity for de novo assembly: a simulation study. PLoS ONE 9, 1–8 (2014).
22. Dong X et al. The long and the short of it: unlocking nanopore long-read RNA sequencing data with short-read differential expression analysis tools. NAR Genom. Bioinform. 3, lqab028 (2021). [PubMed: 33937765]
23. Wang Y et al. N^6-methyladenosine RNA modification regulates embryonic neural stem cell self-renewal through histone modifications. Nat. Neurosci. 21, 195–206 (2018). [PubMed: 29335608]
24. Canzar S, Andreotti S, Weese D, Reinert K & Klau GW CIDANE: comprehensive isoform discovery and abundance estimation. Genome Biol. 17, 16 (2016). [PubMed: 26831908]
25. Alqassem I, Sonthalia Y, Klitzke-Feser E, Shim H & Canzar S McSplicer: a probabilistic model for estimating splice site usage from RNA-seq data. Bioinformatics 37, 2004–2011 (2021). [PubMed: 33515239]
26. Batista PJ et al. m^6a RNA modification controls cell fate transition in mammalian embryonic stem cells. Cell Stem Cell 15, 707–719 (2014). [PubMed: 25456834]
27. Ke S et al. A majority of m^6a residues are in the last exons, allowing the potential for 3′ UTR regulation. Genes Dev. 29, 2037–2053 (2015). [PubMed: 26404942]
28. Yamauchi T, Nishiyama M, Moroishi T, Kawamura A & Nakayama KI FBXL5 inactivation in mouse brain induces aberrant proliferation of neural stem progenitor cells. Mol. Cell. Biol. 37, e00470–16 (2017). [PubMed: 28069738]
29. Kuboyama K, Fujikawa A, Suzuki R & Noda M Inactivation of protein tyrosine phosphatase receptor type Z by pleiotrophin promotes remyelination through activation of differentiation of oligodendrocyte precursor cells. J. Neurosci. 35, 12162–12171 (2015). [PubMed: 26338327]
30. Kurosaki T, Popp MW & Maquat LE Quality and quantity control of gene expression by nonsense-mediated mRNA decay. Nat. Rev. Mol. Cell Biol. 20, 406–420 (2019). [PubMed: 30992545]
31. Lianoglou S, Garg V, Yang JL, Leslie CS & Mayr C Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. Genes Dev. 27, 2380–2396 (2013). [PubMed: 24145798]
32. Tang AD et al. Full-length transcript characterization of *SF3B1* mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. Nat. Commun. 11, 1438 (2020). [PubMed: 32188845]
33. Köster J & Rahmann S Snakemake—a scalable bioinformatics workflow engine. Bioinformatics 28, 2520–2522 (2012). [PubMed: 22908215]
34. DeAngelis MM, Wang DG & Hawkins TL Solid-phase reversible immobilization for the isolation of PCR products. Nucleic Acids Res. 23, 4742–4743 (1995). [PubMed: 8524672]

35. Sobczak K & Krzyzosiak WJ RNA structure analysis assisted by capillary electrophoresis. Nucleic Acids Res. 30, e124 (2002). [PubMed: 12434006]

36. Azarani A & Hecker KH RNA analysis by ion-pair reversed-phase high performance liquid chromatography. Nucleic Acids Res. 29, e7 (2001). [PubMed: 11139637]

37. Wang Y et al. High-resolution profile of transcriptomes reveals a role of alternative splicing for modulating response to nitrogen in maize. BMC Genomics 21, 353 (2020). [PubMed: 32393171]

38. Li R et al. Direct full-length RNA sequencing reveals unexpected transcriptome complexity during *Caenorhabditis elegans* development. Genome Res. 30, 287–298 (2020). [PubMed: 32024662]

39. Haussmann IU et al. m^{6}a potentiates *Sxl* alternative pre-mRNA splicing for robust *Drosophila* sex determination. Nature 540, 301–304 (2016). [PubMed: 27919081]

40. Bartosovic M et al. N^6-methyladenosine demethylase FTO targets pre-mRNAs and regulates alternative splicing and 3′-end processing. Nucleic Acids Res. 45, 11356–11370 (2017). [PubMed: 28977517]

41. Xiao W et al. Nuclear m^{6}a reader YTHDC1 regulates mRNA splicing. Mol. Cell 61, 507–519 (2016). [PubMed: 26876937]

42. Zhou KI et al. Regulation of co-transcriptional pre-mRNA splicing by m^{6}a through the low-complexity protein hnRNPG. Mol. Cell 76, 70–81 (2019). [PubMed: 31445886]

43. Jacob AG & Smith CWJ Intron retention as a component of regulated gene expression programs. Hum. Genet. 136, 1043–1057 (2017). [PubMed: 28391524]

44. Braunschweig U et al. Widespread intron retention in mammals functionally tunes transcriptomes. Genome Res. 24, 1774–1786 (2014). [PubMed: 25258385]

45. Yoon K-J et al. Temporal control of mammalian cortical neurogenesis by m^{6}a methylation. Cell 171, 877–889 (2017). [PubMed: 28965759]

46. Eckmann CR, Rammelt C & Wahle E Control of poly(A) tail length. Wiley Interdiscip. Rev. RNA 2, 348–361 (2011). [PubMed: 21957022]

47. Dobin A et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21 (2013). [PubMed: 23104886]

48. Conforti L et al. Kif1Bβ isoform is enriched in motor neurons but does not change in a mouse model of amyotrophic lateral sclerosis. J. Neurosci. Res. 71, 732–739 (2003). [PubMed: 12584731]

References

49. Jiang L et al. Synthetic spike-in standards for RNA-seq experiments. Genome Res. 21, 1543–1551 (2011). [PubMed: 21816910]

50. Li B & Dewey CN RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. BMC Bioinformatics 12, 323 (2011). [PubMed: 21816040]

51. Lappalainen T et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501, 506–511 (2013). [PubMed: 24037378]

52. Chang Z et al. Bridger: a new framework for de novo transcriptome assembly using RNA-seq data. Genome Biol. 16, 30 (2015). [PubMed: 25723335]

53. Pertea M et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. 33, 290–295 (2015). [PubMed: 25690850]

54. Shao M & Kingsford C Accurate assembly of transcripts through phase-preserving graph decomposition. Nat. Biotechnol. 35, 1167–1169 (2017). [PubMed: 29131147]

55. Pertea M, Kim D, Pertea GM, Leek JT & Salzberg SL Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat. Protocols 11, 1650 (2016). [PubMed: 27560171]

56. Kent WJ BLAT—the BLAST-like alignment tool. Genome Res. 12, 656–664 (2002). [PubMed: 11932250]

57. Xie Y et al. SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. Bioinformatics 30, 1660–1666 (2014). [PubMed: 24532719]

58. Liu J, Yu T, Mu Z & Li G TransLiG: a de novo transcriptome assembler that uses line graph iteration. Genome Biol. 20, 81 (2019). [PubMed: 31014374]

59. Roberts A & Pachter L Streaming fragment assignment for real-time analysis of sequencing experiments. Nat. Methods 10, 71–73 (2013). [PubMed: 23160280]

60. Li B, Ruotti V, Stewart RM, Thomson JA & Dewey CN RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics 26, 493–500 (2009). [PubMed: 20022975]

61. Vitting-Seerup K & Sandelin A The landscape of isoform switches in human cancers. Mol. Cancer Res. 15, 1206–1220 (2017). [PubMed: 28584021]

62. Anders S, Reyes A & Huber W Detecting differential usage of exons from RNA-seq data. Genome Res. 22, 2008–2017 (2012). [PubMed: 22722343]

63. Park HJ et al. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucleic Acids Res. 41, e74 (2013). [PubMed: 23335781]

64. Finn RD et al. Pfam: the protein families database. Nucleic Acids Res. 42, D222–D230 (2014). [PubMed: 24288371]

65. Zhang Y et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 9, 1–9 (2008).

66. Quinlan AR & Hall IM BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842 (2010). [PubMed: 20110278]

67. Heinz S et al. Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. Mol. Cell 38, 576–589 (2010). [PubMed: 20513432]

68. Alexa A, Rahnenführer J & Lengauer T Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics 22, 1600–1607 (2006). [PubMed: 16606683]

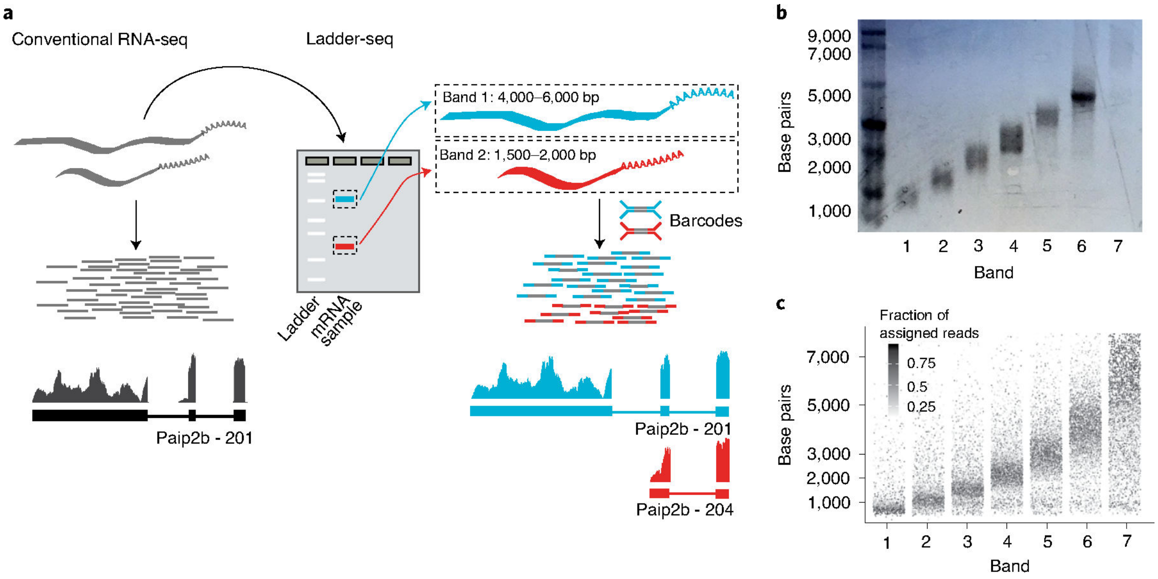


Fig. 1 |. Ladder-seq uses mRNA length information to aid transcriptome reconstruction.

a, Ladder-seq uses a denaturing agarose gel to separate mRNA by length into discrete bands before library preparation and sequencing. Each band contains transcripts of a certain length range that depends on the location of cuts through the gel. The originating band of the resulting reads is tracked using barcodes. In our dataset of mouse NPCs, Ladder-seq reveals transcript Paip2b-204 that contains intronic sequence of transcript Paip2b-201. **b**, Assessment of length separation by denaturing gel electrophoresis. Length-separated mRNA was run on a new denaturing agarose gel with each band loaded into a separate lane. This assay was conducted once. **c**, In silico gel. For every annotated transcript, the intensity of a point with y coordinate equal to its annotated length (plus 200-nt average poly(A) tail size[46]) shows the fraction of reads obtained from each band (x axis) that can be assigned unambiguously to it.

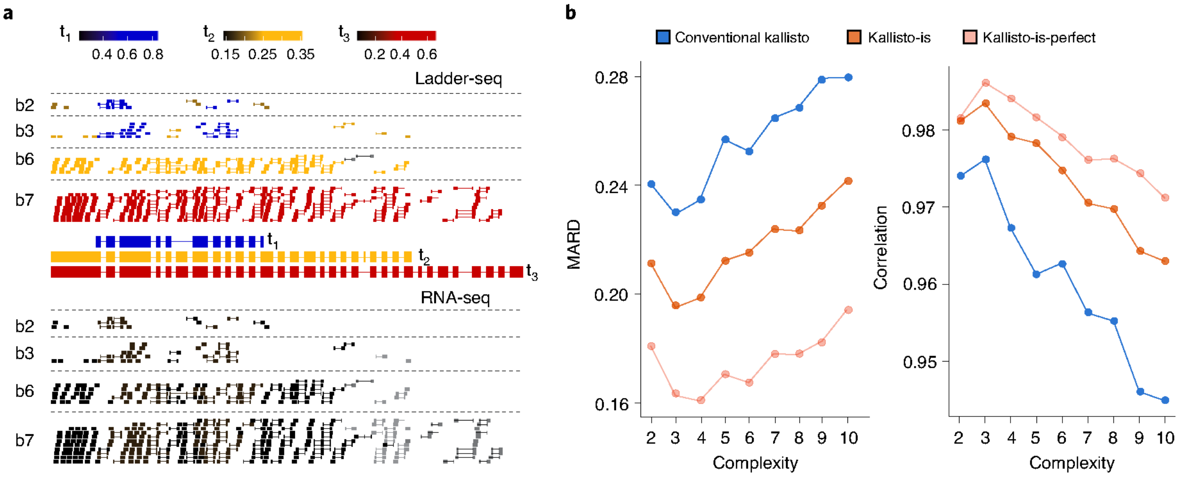


Fig. 2 |. Reduced read assignment ambiguity in Ladder-seq improves transcript quantification.

a, This illustrative example shows reads that were sampled in bands 2, 3, 6 and 7 in our genome-wide simulation study from three transcripts (t_1 = ENST00000519483, t_2 = ENST00000524124 and t_3 = ENST00000357668; not all transcripts shown). The color of each read indicates the transcript to which the read is dominantly assigned after the first E-step of the EM algorithm in the original kallisto implementation based on conventional RNA-seq data (bottom) and in our extension of the algorithm to Ladder-seq (top). More precisely, we color every read according to the additional fraction that is assigned to the transcript of maximal assignment. The original algorithm fractionally assigns each read equally to every transcript that it overlaps (normalized by length), leading to indistinguishable black reads. Our adaptation of the algorithm uses the partitioning of reads into bands to hint at the read's originating transcript, shown by matching read and transcript colors. Based on the migration patterns estimated from the length of the three transcripts, our EM algorithm assigns larger read fractions to transcripts that are expected to occur more abundantly in the read's band (Methods). This length-based deconvolution allows the EM algorithm to ultimately quantify transcript abundances more accurately. In this example, our Ladder-seq-specific EM algorithm estimates 17, 257 and 67 counts (rounded) for transcripts t_1, t_2 and t_3 respectively, which closely match their true expression of 15, 250 and 83 counts, respectively. In contrast, original kallisto fails to detect expression of t_3 (zero counts) and overestimates expression of t_2 (334 counts) from highly ambiguous RNA-seq reads. It estimates four counts for t_1. **b**, Quantification accuracy of kallisto-ls compared to conventional kallisto on 30 million simulated reads. Mean values over 20 repeated simulations are reported. Pearson correlation of estimated and ground truth abundance in $\log_2$ transformed TPM and mean absolute relative difference are shown as a function of gene complexity—that is, the number of transcripts expressed by a gene. Genes expressing a single transcript (omitted) or two transcripts were estimated to be lowly expressed by RSEM (Supplementary Table 4), making their quantification less accurate (Supplementary Tables 5 and 6).

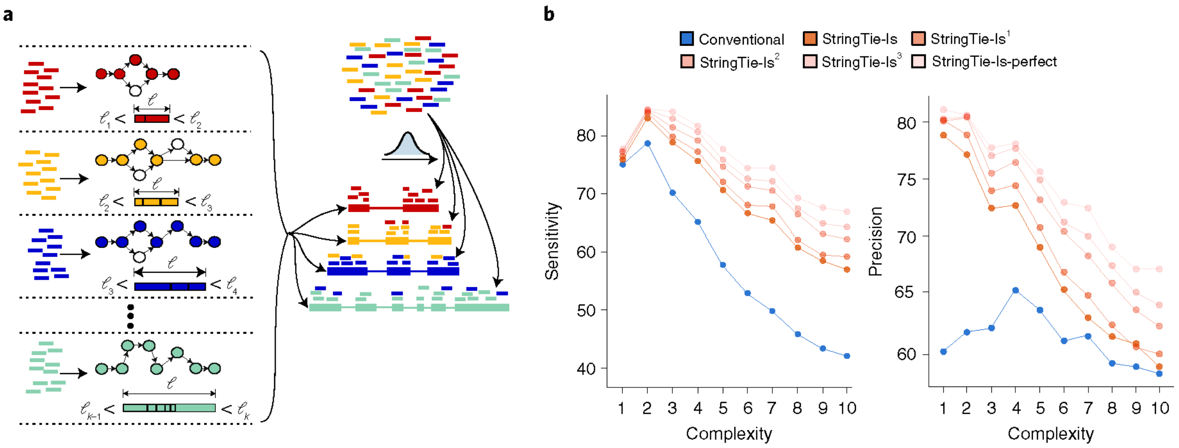


Fig. 3 |. Ladder-seq-based transcript assembly.

a, Overview of the proposed computational framework. For each band, a graph is constructed that captures connectivity information contained in read alignments. Reference-based assembly methods, such as StringTie2, use variants of splicing graphs to capture connectivity of exonic segments in expressed transcripts evidenced by spliced alignments of reads. Transcript sequences are then assembled by traversing paths through these graphs according to some optimization criteria, such as maximum flow for StringTie2. In contrast to conventional RNA-seq, where truly expressed transcripts need to be identified among a large number of possible paths through a single graph per locus, Ladder-seq limits the search for expressed transcripts to paths in smaller graphs that are constructed for each band separately. In addition, reads in different bands are obtained from transcripts of a certain length range, imposing length constraints that can further direct the search for the correct paths. After having inferred the best possible set of transcripts satisfying given length constraints in each band independently, we integrate them to a refined set of transcripts by assigning reads to them according to our statistical model of Ladder-seq, which relies on previously estimated migration patterns through the gel. **b**, Accuracy of transcript assembly from 30 million simulated RNA-seq and matching Ladder-seq reads. Reads were aligned to the reference genome using STAR[47]. Sensitivity (left) and precision (right) of StringTie-ls and its conventional counterpart StringTie2 are shown as a function of gene complexity defined as the number of expressed transcripts. The lower ground truth expression of some genes with complexity 1 (Supplementary Table 4) makes them detectable with lower sensitivity than transcripts of genes with complexity 2. StringTie-ls^i denotes the result of StringTie-ls on the simulated Ladder-seq dataset to which i-fold error reduction was applied (Methods), starting from the migration error estimated from the NPC sample (StringTie-ls). StringTie-ls-perfect represents the results of StringTie-ls on the most optimistic Ladder-seq experiment in which transcripts perfectly separate by length, without any migration error. All results are listed in Supplementary Tables 7–10.

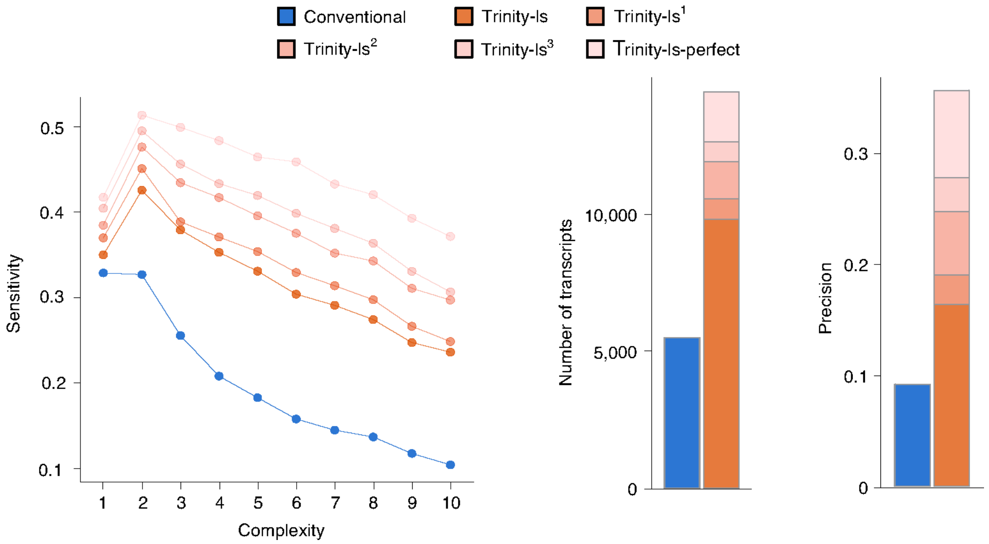


Fig. 4 |. Accuracy of de novo transcript assembly from 75 million simulated RNA-seq and matching Ladder-seq reads.

Trinity-ls^i denotes the results of Trinity-ls on the simulated Ladder-seq dataset to which *i*-fold error reduction was applied (Methods), starting from the migration error estimated from the NPC sample (Trinity-ls). Trinity-ls-perfect represents the results of Trinity-ls on the most optimistic Ladder-seq experiment in which transcripts perfectly separate by length, without any migration error. A transcript is correctly assembled if its BLAT alignment to a true transcript covers at least 90% of the full transcript length. Left, Sensitivity of Trinity-ls and its conventional counterpart Trinity is shown as a function of gene complexity defined as the number of expressed transcripts. The low expression of some genes expressing a single transcript (Supplementary Table 4) makes them more difficult to assemble than transcripts of genes with a higher complexity. Middle, Total number of correctly assembled transcripts. Right, Overall precision. Assembled transcript fragments cannot be assigned unambiguously to individual genes. All results are listed in Supplementary Tables 15–18.

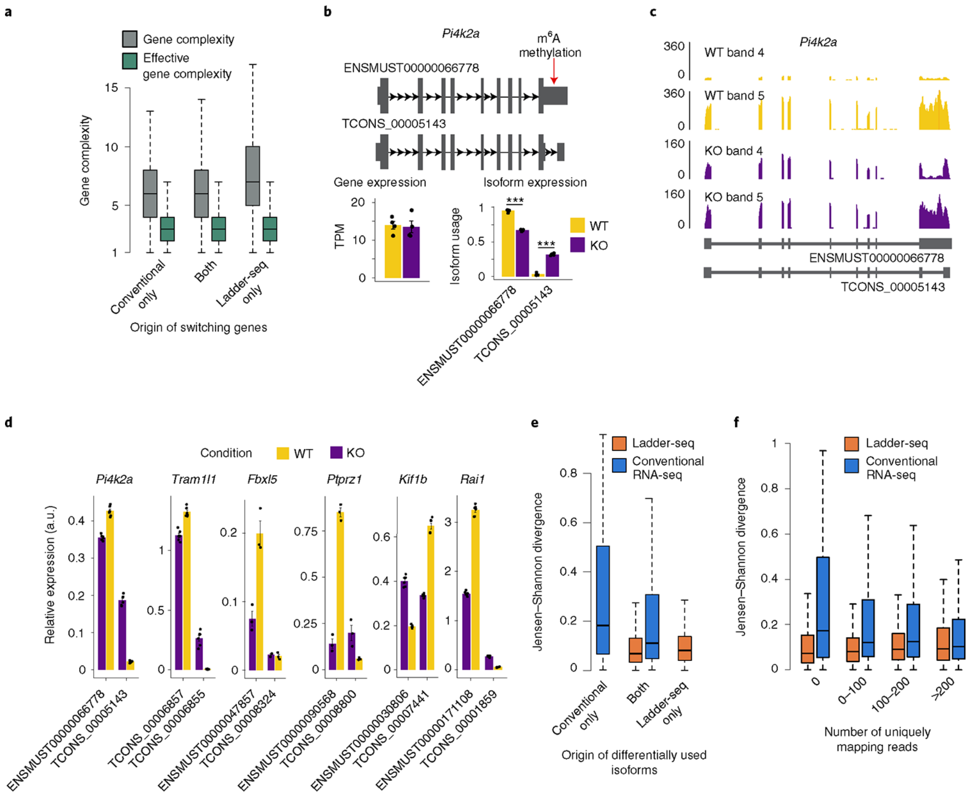


Fig. 5 |. Ladder-seq improves differential analysis of reconstructed transcriptomes.

a, Gene complexity and effective gene complexity for switching genes identified by only one of the two or both methods (*n* = 755 by conventional only, *n* = 1,512 by Ladder-seq only and *n* = 1,123 by both). The effective complexity is defined as the number of expressed transcripts contained in a single band. **b**, Isoform switch identified only by Ladder-seq. *Pi4k2a* expresses mostly the annotated ENSMUST00000066778 transcript in WT (*n* = 4), whereas KO (*n* = 4) also expresses a shorter unannotated transcript (TCONS_00005143) in which a normally m^{6}A-tagged exonic region is spliced out. The red arrow shows the location of m^{6}A methylation. Overall gene expression level is unchanged between WT and KO. Bar plots represent mean ± s.e.m.; ***FDR-corrected $P < 0.001$. **c**, Coverage plot for bands 4 and 5 of *Pi4k2a* showing how reads from the shorter unannotated transcript TCONS_00005143 are separated from reads belonging to the longer ENSMUST00000066778. **d**, Relative quantification of isoform expression with RT–qPCR. Three biological replicates were tested per genotype. Each sample was tested in triplicate and normalized to β-actin. Expression levels of each differentially expressed isoform were normalized to the expression of a common isoform identified in both WT and *Mettl14* KO, which consistently showed no significant difference between WT and KO NPCs. Bars represent mean values; error bars represent the s.e.m. **e**, JSD between estimated and assigned read band distributions for differentially used isoforms identified by only one of the two or both methods (*n* = 729 by conventional only, *n* = 1337 by Ladder-seq only and *n* = 1,745 by both). **f**, JSD between estimated and assigned read band distributions for all identified transcripts by Ladder-seq and conventional RNA-seq grouped by number of available uniquely mapping reads (Ladder-seq: *n* = 14,024 for 0, *n* = 1,266 for 0–100, *n* = 1,392 for 100–200 and *n*

= 6,206 for >200 uniquely mapping reads; Conventional: *n* = 17,568 for 0, *n* = 1,446 for 0–100, *n* = 1,483 for 100–200 and *n* = 6,248 for >200 uniquely mapping reads). Box plot definition: bottom and top of the box correspond to lower and upper quartiles of the data; bar is the median; and whiskers are median ±1.5× interquartile range. a.u., arbitrary units.

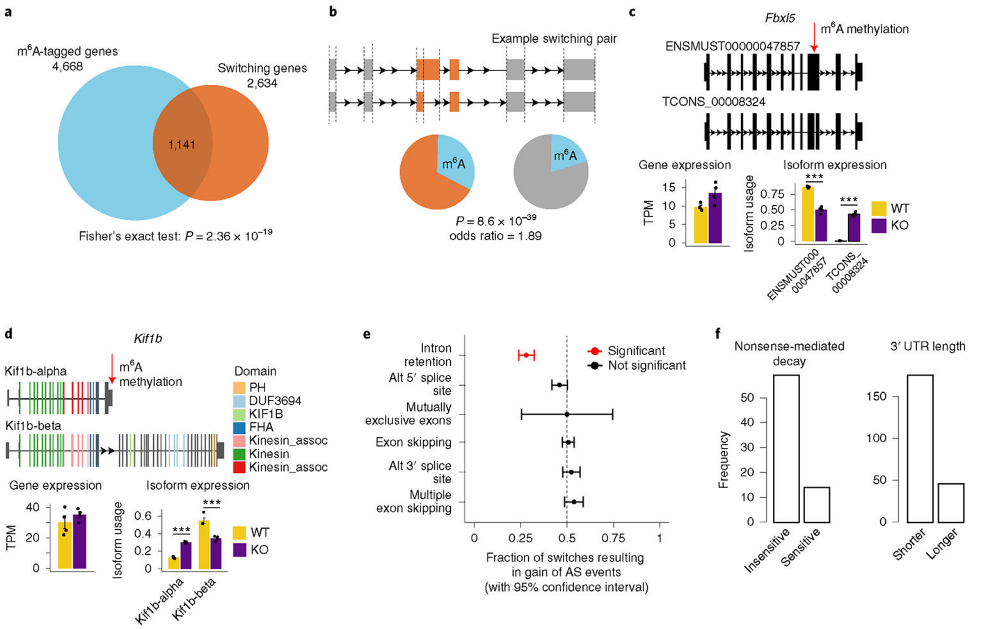


Fig. 6 |. *Mettl14* Ko leads to isoform switches in m^{6}A methylated genes and leads to loss of protein domains and loss of intron retentions.

a, Venn diagram showing overlap between switching genes and m^{6}A methylated genes. *P* value from two-sided Fisher's exact test. **b**, Enrichment of m^{6}A methylation within exonic segments bounding differentially spliced regions. In this example, both the differentially spliced exonic region and the two segments flanking it (in orange) are considered. Pie charts show percentage of exonic segments overlapping m^{6}A peaks. *P* value and odds ratio from two-sided Fisher's exact test. **c**, **d**, Isoform switch in genes *Fbxl5* (**c**) and *Kif1b* (**d**) (between *n* = 4 WT and *n* = 4 KO samples). Red arrows show location of m^{6}A methylation. Bar plots represent mean ± s.e.m; ***FDR-corrected $P < 0.001$. Isoform switch in *Kif1b* leads to upregulation of shorter Kif1b-alpha isoform, which lacks multiple domains and is expressed in non-neuronal tissues. The longer Kif1b-beta is the neuronal isoform and is responsible for the transport of synaptic vesicle precursors[48]. Overall gene expression level is unchanged between WT and KO. **e**, Splicing enrichment analysis. Proportion of isoform switches (*n* = 2,634) resulting in gain of each splicing event in KO NPCs (*n* = 482 intron retentions, *n* = 573 alt 5′ splice sites, *n* = 12 mutually exclusive exons, *n* = 926 exon skippings, *n* = 483 alt 3′ splice sites and *n* = 396 multiple exon skippings). Test of equal proportions was used to identify proportions significantly different from 0.5. Points indicate fraction of switches resulting in gain of splicing event, and bars represent 95% confidence intervals. **f**, Number of intron retention losses resulting in NMD sensitive or insensitive isoforms (left) and shorter or longer 3′ UTR length (right).