**Multicellular, IVT-derived, unmodified human transcriptome for nanopore direct RNA analysis**

Caroline A. McCormick[1‡], Stuart Akeson[1‡], Sepideh Tavakoli[1], Dylan Bloch[1], Isabel N. Klink[1], Miten Jain[1,2#], Sara H. Rouhanifard[1#]

[1]*Dept. of Bioengineering, Northeastern University, Boston, MA*

[2]*Dept. of Physics, Northeastern University, Boston, MA*

[‡]*These Authors Contributed Equally,* [#]*Co-corresponding Authors*

## Abstract

Nanopore direct RNA sequencing (DRS) enables measurements of native RNA modifications. Modification-free transcripts are an important control for DRS. Additionally, it is advantageous to have canonical transcripts from multiple cell lines to better account for human transcriptome variation. Here we generated and analyzed Nanopore DRS datasets for five human cell lines using *in vitro* transcribed (IVT) RNA. We compared performance statistics amongst biological replicates. We also documented nucleotide and ionic current level variation across cell lines. These data will serve as a resource to the community for RNA modification analysis.

## Introduction

Post-transcriptional chemical modifications can elicit structural[1] and functional[2] changes in RNAs within coding[3] and non-coding[4] sequences. Aberrant RNA modifications have been linked to human diseases, including mitochondrial encephalomyopathy, lactic acidosis and stroke-like episodes (MELAS)[5], acute myeloid leukemia (AML)[6], heart failure[7], and amyotrophic lateral sclerosis (ALS)[8]. Next-generation sequencing studies can unravel the biological function of some mRNA modifications[9]. However, identifying the precise location of mRNA modifications remains a challenge[10].

Nanopore Direct RNA sequencing (DRS) has emerged as a method for analyzing native RNA strands based on ionic current disruptions during translocation. With appropriate negative controls, deviations in the ionic current disruptions may be attributed to RNA modifications[11], such as pseudouridine(Ψ)[12], $N^6$-methyladenosine (m6A)[13], and inosine (I)[14]. These negative controls include but are not limited to enzymatic knockouts of RNA modification machinery, synthetic RNA controls, expected distribution pore models for computational analysis, and genomic and transcriptomic templates for *in vitro* transcription (IVT)-based negative controls.

Knockouts and knockdowns of RNA modifying enzymes (such as pseudouridine synthases and methyltransferases) can be generated using CRISPR[15] or siRNAs[16]. These strategies have been used to identify putative sites of modification[17]. Alternatively, solid-phase synthesis and IVT can be used to create synthetic RNA oligonucleotides with and without modifications. Previously, these strategies have been used in conjunction with DRS[18] to analyze modification-related differences in Nanopore ionic current[19,20,12,21]. However, longer RNA oligonucleotides have higher costs and lower fidelity and characterizing every possible RNA signature quickly becomes procedurally implausible and cost prohibitive.

A computational alternative is the use of a kmer lookup table-based pore model. These are precomputed through the assessment of observed nanopore data and made available by Oxford Nanopore Technologies. A significant deviation in ionic current distribution from a pore model at a given location is suggestive of modification[22], assuming all other confounding factors, such as surrounding sequence context, nucleotide variation, and satellite modifications, are mitigated[22]. This strategy for RNA modification detection has been successful for several modifications,

including m6a[22]. However, pore models depend on a set of ionic curated data specific to the version of nanopores and chemistry. The data required to produce new pore models regularly is impractical for noncommercial entities and the scope of the kmers (5mer for the current pore models) is unable to capture the effect of long-range sequence context. This limits the power of the pore model to accurately predict ionic current since the surrounding sequence context impacts the ionic current corresponding to a kmer[19]. Further, the methods used to construct the current pore models are not publicly available, including the nature of the training data. This can limit the scope of conclusions and confidence in modification prediction.

IVT-derived, unmodified transcriptomes[19,23] are an attractive option for analyzing modifications using DRS[12,19]. To generate these libraries, polyadenylated RNA strands are reverse transcribed to cDNA, and then *in vitro* transcribed into RNA using canonical nucleotides. This process maintains the general sequence context and diversity of the initial poly(A) RNA sample. However, IVT RNA derived from a single cell line may not comprehensively capture the landscape of nucleotide variation and expressed genes in the human transcriptome creating a need for matched IVT controls for every cell line. By expanding to multiple human cell lines, a more complete transcriptome can be represented. We analyzed the effect of multiple biological replicates compared to multiple experiments from a single replicate. We documented and cataloged putative single nucleotide variants (SNV) for all five of our cell lines. Finally, we assembled ionic current level alignments for each IVT dataset. This publicly available, pan-human, IVT DRS dataset will be a resource to the direct RNA analysis community, circumventing the need for expensive IVT library preparation and sequencing for human samples.

## Results

We generated and analyzed DRS data for IVT RNA from three biological replicates of cultured HeLa cells[12]. The throughput for these experiments ranged between 1.01 and 1.86 million strand reads per flow cell. All replicates were base called using Guppy (v6.4.2); reads falling below the quality threshold of 7 were ignored for all downstream analyses. These reads were aligned to a reference set composed of gencode v38 transcripts and ONT RCS sequence (spike-in) using minimap2[24] (-ax map-ont), yielding 807K, 1.30M, and 1.75M. Median alignment identities for the three replicates ranged between 89.96 - 91.36% (**Figure 1a**).

We compared the represented genes across biological replicates, we pooled the aligned reads and ranked observed genes by the total number of aligned reads. We randomly chose 25 genes from each quartile and determined if they were present in each biological replicate (**Figure 1b**). We noted differences among replicates in quartile 1, which includes genes with the lowest number of reads (1-3 reads per gene). This is attributable to the variation in sequencing throughput per replicate. These differences substantially decrease in quartile 2 and are absent in the other quartiles that include genes with higher numbers of reads (25-34,905 reads per gene). This suggests that biological replicates are largely consistent for most of the genes observed.

We observed 17,570 unique genes that had an alignment from at least one DRS read from the pooled biological replicates. We iteratively filtered the number of genes observed in the data based on read count cutoff (**Figure 1c**; **Supplementary Table 1**). As the read count cutoff increased, the percentage of genes (out of the total observed genes) decreased. For all read count cutoffs, the combined IVT HeLa DRS data set out-competes any individual replicate suggesting that total read count rather than the number of biological replicates is more important for maximizing coverage.

To investigate further the relationship between the depth and sequencing throughput, we determined the total number of genes represented by each replicate (**Supplementary Table 2**). To that end, we calculated the breakdown of biological replicate overlap with variable read count cutoffs ranging from 1-10 (**Figure 1d**). At a read count cutoff of ten, the number of unique genes captured by all three replicates decreased to 10,853.

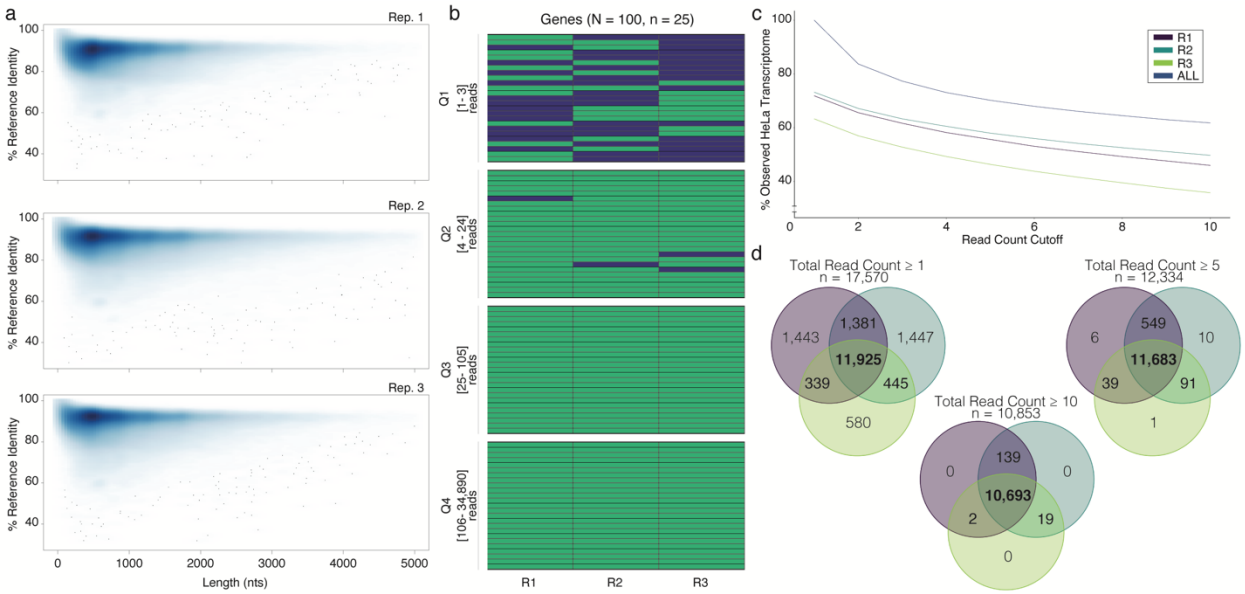


**Figure 1. Alignment performance comparison among biological replicates for HeLa cells. a,** Average percent identity of aligned HeLa reads against read length generated with Nanoplot[25]. **b,** Heat map of HeLa replicate gene coverage for randomly selected genes with at least 1 read in one of the replicates. HeLa transcriptome separated into quartiles by depth of coverage (number of total reads). 25 genes randomly selected per quartile (n = 25, N =100). Green indicates coverage of a gene, blue indicates absence. **c,** Percent of observed HeLa transcriptome coverage against total HeLa minimum read count cut-off. The ALL curve is the sum of read counts across replicates for each gene. **d,** Venn Diagram breakdown of unique genes by HeLa IVT replicate at HeLa total minimum read count cutoff.

We assessed the variability in SNV identification across the three biological HeLa IVT replicates. We only considered locations where all three replicates had a minimum of 10 reads to avoid ambiguous SNVs, and only considered SNVs that accounted for 30% or higher of total reads aligning to a position to filter out IVT or sequencing-related errors (**Supplementary Table 4**). We then performed a set intersection between SNVs for each pair of biological replicates (**Supplementary Table 5**) and calculated the corresponding dice similarity coefficient (**Figure 2a**). The dice similarity coefficient measures the overlap between two binary datasets. The similarity was highest at the 70% occurrence threshold. At the 70% occurrence threshold, the biological replicate similarity ranged from 53.9%-78.0%.

We pooled the aligned reads from the three HeLa replicates into a single data set to catalog SNVs. Like our strategy above, we removed all positions with less than 10 aligned reads and only considered SNVs with over 30% occurrence. This analysis identified 62,708 putative variant positions (**Supplementary Table 3**). Of those 62,708 putative variants, 24,879 were documented in Ensembl[26]. We assessed the remaining variants to identify SNVs occurring in low confidence 9-mers. These 9-mers were identified using a nucleotide quality-based analysis from an independent biological DRS dataset (see Methods). Among the 37,829 remaining variants, 8,930

occurred in low-confidence 9-mers. The remaining 28,899 variants were categorized as a combination of putative SNVs and systematic errors due to sample preparation and sequencing (**Figure 2b**). The number of putative SNVs decreases at higher thresholds, e.g., at a 50% variant occurrence threshold we identified 3,200 putative SNVs (**Supplementary Table 3**). We performed sanger sequencing on 5 positions with reported SNVs at an 80% occurrence threshold. Sanger sequencing confirmed four of five predicted SNV positions as genomic variants (**Figure 2c, d**).

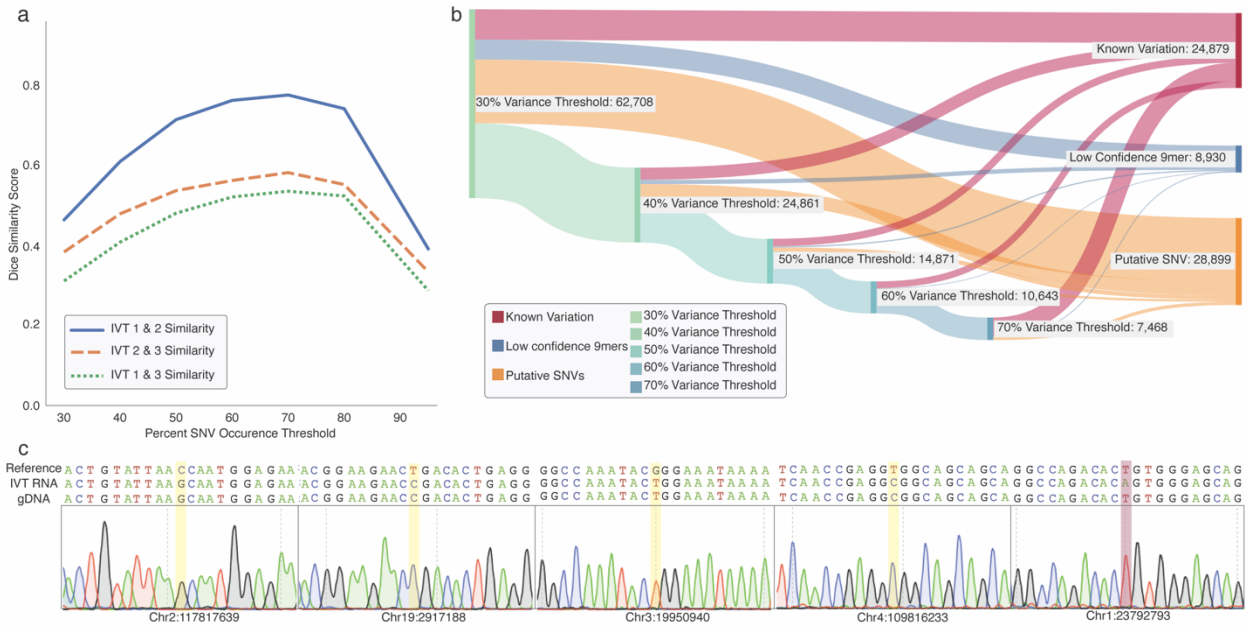


**Figure 2. SNV analysis for HeLa cells**. **a.** Dice Similarity Score between SNVs for positions with at least 10 reads across all HeLa replicate datasets. Similarity scores are measured over multiple SNV occurrence thresholds. **b.** SNV identification and binning over a series of occurrence thresholds for Pooled HeLa replicates. **c.** Sanger sequencing results for five positions with observed SNVs (occurrence 80%+). Yellow bars show agreement, red bars show disagreement.

Next, we wanted to determine the minimum number of reads covering 75% of the 17,570 unique genes in our combined IVT HeLa dataset. We randomly subsampled an increasing number of aligned reads from the combined IVT HeLa dataset (ranging from 100k to 1M) at intervals of 100k (**Figure 3a**). We repeated this subsampling 100 times per interval and noted that 800,000 aligned reads achieved our 75% threshold.

Our combined IVT DRS HeLa dataset only captures a subsection of the human transcriptome. To expand the scope of this dataset, we generated DRS data for IVT RNA derived from four additional cell lines: i) SH-SY5Y, a neuroblastoma; ii) HepG2, a hepatoma; iii) NTERA, an epithelial-like cell line derived from testes; and iv) A549, alveolar basal epithelial cells. The throughput for these data ranged from 1.3 million to 2.4 million aligned reads per cell line (**Supplementary Table 6**).

We estimated the relative number of unique genes observed for each cell line by subsampling up to 1,000,000 reads (**Figure 3a**). Next, for each of the 5 cell lines, we tested how well the remaining 4 cell lines captured the observed gene level diversity (**see methods**). NTERA had the lowest proportion of observed genes represented in the sample population at ~88% (**Figure 3b**). This shows, for our dataset, that a combination of different cell lines captures most of the observed gene level diversity for a target cell line. We also computed the pairwise dice similarity

between each cell line's observed genes. This analysis revealed that HeLa and A549 were the most similar, while NTERA and HepG2 were the least similar (**Supplemental Figure 1)**.

Next, we tested how well our IVT data captured the gene level diversity of a publicly available IVT dataset (GM12878 B-lymphocyte cells)[19]. We sampled reads equally from all 5 cell lines and computed a set of observed genes. We computed the intersection of this set with a set of genes derived using 150,000 reads from the GM12878 IVT data (**Figure 3c**). When using all cell lines, this represented ~90.98% of the GM12878 genes. We repeated this analysis using the most similar (HeLa & A549) and least similar (NTERA & HepG2) cell line combinations in our data. These subsamples represented ~87.74% and ~90.36% of the observed GM12878 genes respectively. We additionally examined the number of alignments in our data for the 10 most observed genes based on the GM12878 IVT dataset (**Figure 3d**). We observed that different cell line combinations, regardless of similarity, produced higher numbers of reads for at least one gene. This indicates that a more specific IVT strategy will yield higher read counts for a target gene.

Finally, we performed SNV analysis on each of the 4 additional cell lines. From all 5 cell lines at an SNV occurrence threshold of 80% we identified 26,239 total SNVs. We also used nanopolish eventalign[19] to generate ionic current alignments for the entire IVT dataset. Figure 3e shows ionic current distributions for an example of five overlapping 5-mers. Cell lines with low coverage have poorly defined distributions in comparison to cell lines with higher throughput.

In summary, we have generated an IVT-derived and unmodified human transcriptome for nanopore direct RNA analysis by combining reads from multiple human cell lines. These data provide a critical negative control that can be used for comparative analyses (such as RNA modifications) across human cell lines. We anticipate that these publicly available data will contribute towards democratizing direct RNA analysis and serve as a resource to the community.

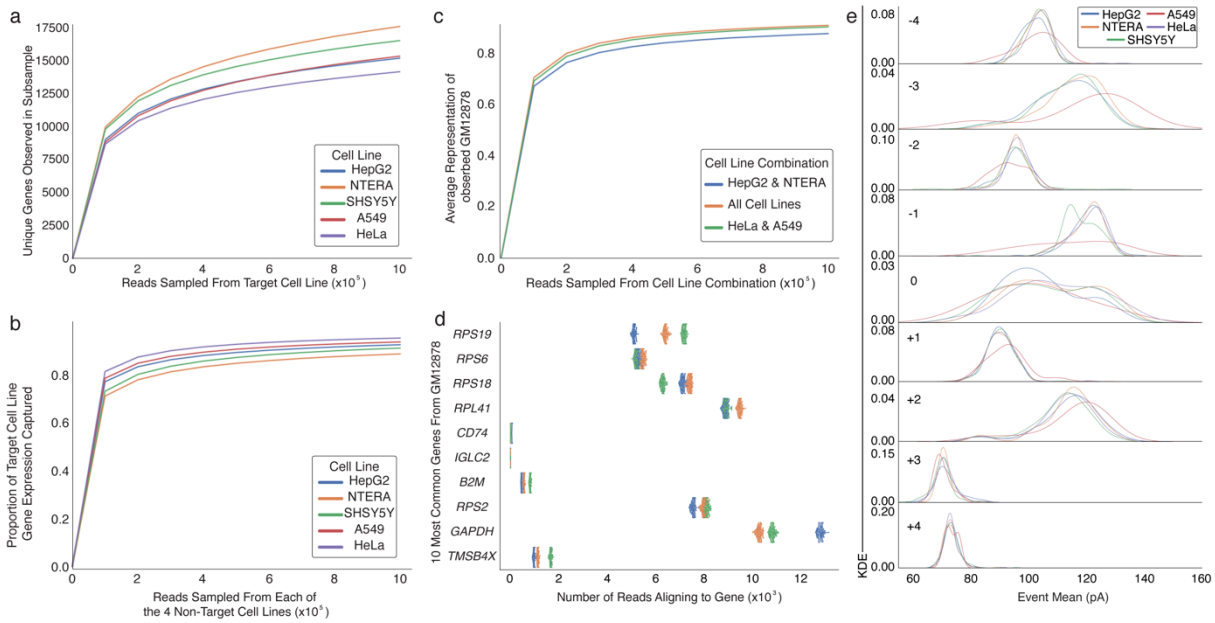


**Figure 3.** a. Unique genes identified for each of the 5 cell lines at different rates of subsampling. b. Target cell line gene representation as a function of reads sampled from four sample population cell lines. Each cell line was used as the target cell line, the line on the graph corresponds to the representation for the target cell line listed in the legend. c. Proportion of GM12878 genes represented in subsamples from cell line combinations. d. Number of reads aligning to the top 10 observed genes from GM12878 in a 1 million read subsample from each of the cell line combinations. e. Ionic Current distributions for MCM5 gene (chr22:35424407 +/- 4) across all 5 cell lines.

**Code Availability**
All code used in this work is publicly available at github.com/RouhanifardLab/PanHumanIVT

**Data Availability**
The data that support this study are available from the corresponding author upon reasonable request. Sequences were aligned to genome version hg38.p10. All FASTQ files and Fast5 raw data generated in this work have been made publicly available in NIH NCBI SRA under the BioProject accession: PRJNA947135

**Online Methods**

**Accessing Publicly Available Data Sets**
HeLa IVT replicates 1 and 2 were constructed from the fast5 files of two independent HeLa IVT datasets publicly available through NIH NCBI SRA BioProject accession: PRJNA777450.
A list of URLs for downloading the fast5 files from a GM12878l IVT dataset can be found at https://s3.amazonaws.com/nanopore-human-wgs/rna/links/NA12878-IVT-RNA_All.files.txt.
Only the fast5 files marked with UCSC were used for our analysis.

**Cell culture**
HeLa, HepG2, A549, and NTERA-2 cells were cultured in Dulbecco's modified Eagle's medium (Gibco, 10566024) as a base; SH-SY5Y cells were cultured in a base of 1:1 EMEM:F12. All media was supplemented with 10% Fetal Bovine Serum (FB12999102, FisherScientific) and 1% Penicillin-Streptomycin (Lonza,17602E). Cells were cultured at 37°C with 5% $CO_2$ in 10 cm tissue culture dishes until confluent.

**Total RNA extraction and Poly(A) selection**
Total RNA extraction from cells and Poly(A) selection was performed using the protocol outlined previously[12]. Briefly, six 10 cm cell culture dishes with confluent cells were washed with ice-cold PBS and lysed with TRIzol (Invitrogen,15596026) at room temperature and transferred to an RNAse-free microcentrifuge tube. Chloroform was added to separate the total RNA in the aqueous supernatant from the organic phase containing DNA and cell debris below following centrifugation. The aqueous supernatant was then transferred to a fresh RNAse-free microcentrifuge tube, and an equal volume of 70% absolute ethanol was added. PureLink RNA Mini Kit (Invitrogen, 12183025) was used to purify the extracted total RNA in accordance with the Manufacturer's protocol. Total RNA concentration was measured using the Qubit™ RNA High Sensitivity (HS) assay (Thermo Fisher, Q32852).

Poly(A) selection was performed using NEBNext Poly(A) mRNA Magnetic Isolation Module (NEB, E7490L) according to the Manufacturer's protocol. The isolated Poly(A) selected RNA was eluted from the beads using Tris buffer. The poly(A) selected RNA concentration was measured using the same Qubit™ assay listed above.

***In vitro* transcription and polyadenylation**
The protocol for IVT, capping, and polyadenylation is described previously[12]. Briefly, cDNA-PCR Sequencing Kit (SQK-PCS109) kit facilitated the reverse transcription (RT) and strand switching (SS). VN and Strand-Switching primers were added to 100 ng of poly(A) selected RNA from the

abovementioned step. cDNA was produced by Maxima H Minus Reverse Transcriptase (Thermo Scientific, EP0751). Using a thermocycler, the reaction protocol is as follows: RT and SS for 90 minutes at 42°C (1 cycle), Heat inactivation for 5 mins at 85°C (1 cycle) and hold at 4°C (∞). PCR amplification was performed using LongAmp Taq 2X Master Mix (NEB, M0287S) and the Nanopore_T7_IVT_Forward and Reverse primers. The thermocycling conditions are as follows: initial denaturation for 30 seconds at 95°C (1 cycle), denaturation for 15 seconds at 95°C (11 cycles), annealing for 15 seconds at 62°C (11 cycles), extension for 15 minutes at 65°C (11 cycles), a final extension for 15 minutes at 65°C (1 cycle), and then an indefinite hold at 4°C. The PCR products were treated with Exonuclease 1 (NEB, M0293S) to digest single-stranded products. Purification of the resulting product was performed using Sera-Mag Select beads (Cytiva, 29343045) according to the Manufacturer's protocol. IVT was performed on the purified PCR product using a HiScribe T7 High yield RNA Synthesis Kit (NEB, E2040S) and purified using a Monarch RNA Cleanup Kit (NEB, T2040S), both according to the Manufacturer's protocols. Polyadenylation was performed using E. *coli* Poly(A) Polymerase (NEB, M0276S) using 12 µg of input RNA, and EDTA was added to halt the reaction. The product was purified again using a Monarch RNA Cleanup Kit with a final elution volume of 12 µL. Concentration was taken using the Qubit™ RNA HS assay. All primers are listed in **Supplementary Data 8**.

**Sequencing, Basecalling, and Alignment Procedure**

Each cell line was sequenced individually using Nanopore Direct RNA sequencing (DRS). DRS runs were basecalled with Guppy v6.4.2 using the high accuracy model using the default basecalling quality score filter of Q ≥7[27]. Basecalled reads were aligned with minimap2 to the GRCh38.p10 reference genome and Gencode.v38 for downstream analysis.

**Genomic DNA Extraction and Sanger Sequencing**

To analyze putative SNVs, we performed sanger sequencing on HeLa genomic DNA (gDNA). gDNA extraction was performed using a Monarch Genomic DNA Purification Kit (NEB, T3010S) following the Manufacturer's protocol for cultured cells with an input of 5e6 HeLa cells.

PCR primers to amplify ~200 nt regions surrounding the SNVs were designed using Primer-BLAST with default settings and checking primer specificity against the *Homo Sapiens* genome. These primers are listed in **Supplementary Data 8**. The PCR reaction was set up with Q5 polymerase (NEB, M0491L) using the Manufacturer's protocol. Thermocycling conditions were as follows: initial denaturation at 98 °C for 30 seconds, 25 cycles of 98 °C for 10 seconds then 63 °C for 20 seconds and 72 °C for 15 seconds, final extension at 72 °C for 2 minutes, and holding at 10 °C. PCR products were purified using a Monarch PCR & DNA Cleanup Kit (NEB, T1030S) following the Manufacturer's protocol. Concentration of eluted DNA was determined using a Nanodrop. To confirm specific amplification, the purified PCR products were imaged on a 2% agarose TBE gel. Samples were sent to Quintara Biosciences for SimpliSeq™ Sanger sequencing.

**Variant Analysis Methods**

To identify the scope of single nucleotide variations (SNVs) present in the sequenced IVT data we followed an align, pileup, and compare strategy. We used the alignments to the GRCh38.p10 reference genome.

minimap2 (-ax splice)

A positional pileup for each IVT replicate was created with pysamstats.

pysamstats (--t variation)

The initial variant calling process documented every location with a minimum of 10x coverage in the pooled cell line data set with at least one called nucleotide differing from the reference. Subsequent filtering removed variants based on the number of occurrences of a given variant in relation to the number of canonical bases and deletions present at a location. We applied a filter of 30%, 40%, 50%, 60%, 70%, 80%, and 95% sequentially to the data.

Each set of variants was broken down into three distinct bins; variants previously documented in Ensembl, variants that occurred in low confidence 9mers, and novel variants for the IVT replicate set. Low confidence 9mers were computed by averaging the phred score across all 9mers in an independent biological DRS dataset and selecting the lowest quartile of mean phred scores as low confidence 9mers.

The final data set consisted of 3 compartments for each variant presence threshold; a set of potentially novel IVT variants, variants with known documentation in Ensembl, and variants that occurred in low confidence 9mers.

**Biological Replicate SNV Comparison**

We found the triple intersection of positions with at least 10 reads in each biological replicate. We found all of the SNVs in those positions and calculated dice similarity between each combination of replicates across a series of occurrence thresholds.

The Dice Similarity Coefficient (DSC) was chosen for the similarity measure since it considers both positive and negative matches towards the similarity of two sets. Matches either 1-1 or 0-0 were considered as true positives (TP), mismatches 0-1 were categorized as false positives (FP) and mismatches 1-0 were categorized as false negatives (FN). Mathematically speaking false positives and false negatives are treated identically in this calculation of similarity.

$$DSC = \frac{2TP}{2TP + FP + FN}.$$

**Estimating Unique Genes as a Function of Read Count**

For each cell line we subsampled between 0 and 1,000,000 reads in increments of 100,000 and created a set of genes observed in those reads. We repeated our subsampling experiment 100 times and averaged the cardinalities of gene sets across all 100 repetitions at every value in the range of subsampled reads. The unique number of genes at each interval was calculated as the cardinality of the observed gene set.

**Round Robin Gene Coverage Saturation**

To calculate the degree to which 4 of our cell lines (representative population) could approximate the observed gene set of the 5th, we subsampled 1,000,000 reads from the 5th cell line (target cell line) to create a representation of which genes were covered by that cell line. For each of the 4 other cell lines we iteratively subsampled between 0 and 1,000,000 reads in increments of 100,000, for a combined total of 0 to 4,000,000 reads, and created a set of observed genes from those reads.

At each iteration we divided the cardinality of the union of our representative population with the target population by the cardinality of our target population to yield a proportion covered value. We repeated our sampling of the representative population 100 times and averaged the results for each sample interval. We repeated this process with each cell line acting as the target cell line. In total the iterative process was performed 500 times, yielding 5 saturation curves.

**Comparison of Cell Lines to Publically Available GM12878 IVT Data**

We used an IVT data set of ~200,000 reads for GM12878 B-lymphocyte cells[19] as an independent comparison point. We performed a unique gene saturation and gene coverage subsampling experiment to measure the effectiveness of our IVT data set captures GM12878 data set. The GM12878 dataset was basecalled following the same basecalling procedure detailed above, then was aligned to gencode.v38 and filtered to only include primary alignments.

**Unique Gene Saturation for GM12878 Sample by Cell Line Combination**

We subsampled 150,000 reads from the GM12878 alignment and created a set of observed "target" genes. We selected three combinations of our cell lines, most similar cell lines (HeLa and A549), least similar cell lines (NTERA and HepG2), and all cell lines to test against the GM12878 target genes. We subsampled 0 to 1,000,000 reads in increments of 100,000 from each combination of cell lines. For each read increment the number of reads was split evenly between each of component cell lines, i.e., 300,000 reads sampled from NTERA and HepG2 would mean that 150,000 reads were sampled from each. We divided the cardinality of the intersection of our sampled set and the GM12878 set by the cardinality of the GM12878 set to obtain the gene saturation at every read sample increment. This process was repeated 100 times and the average was calculated for each read increment.

**GM12878 10 Most Observed Gene Alignments by Cell Line Combination**

We determined the 10 most observed genes in our GM12878 dataset by sampling 150,000 reads and selecting the 10 genes with the highest number of alignment hits. We sampled 1,000,000 reads from each of our cell line combinations (All cell lines, NTERA & Hep2G, and HeLa & A549) and determined the number of reads aligning to each of our 10 most observed GM12878 genes. This process was repeated 100 times for each cell line combination.

**Statistical analysis**

All experiments were performed in multiple independent experiments, as indicated in figure legends. All statistics and tests are described fully in the text or figure legend.

**Acknowledgments.** S.H.R. acknowledges support from NIH 5R01HG011087-02 and support through an Opportunity Fund by the Technology Development Coordinating Center at Jackson Laboratories (NHGRI federal award no. U24HG011735).

**Author Contributions**

C.A.M. and S.H.R. conceived of the research. C.A.M. and S.A. designed the experiments. C.A.M., S.T., D.B., and I.N.K. performed experiments and sequencing runs. C.A.M., S.A., D.B., and I.N.K. analyzed the data with guidance from S.H.R., and M.J. C.A.M. and S.A. wrote the paper with guidance from S.H.R. and M.J.

## REFERENCES


1. Liu, N. *et al.* N6-methyladenosine-dependent RNA structural switches regulate RNA–protein interactions. *Nature* vol. 518 560–564 Preprint at https://doi.org/10.1038/nature14234 (2015).
2. Barbieri, I. *et al.* Promoter-bound METTL3 maintains myeloid leukaemia via m6A-dependent translation control. *Nature* **552**, 126 (2017).
3. Nachtergaele, S. & He, C. Chemical Modifications in the Life of an mRNA Transcript. *Annual Review of Genetics* vol. 52 349–372 Preprint at https://doi.org/10.1146/annurev-genet-120417-031522 (2018).
4. Sloan, K. E. *et al.* Tuning the ribosome: The influence of rRNA modification on eukaryotic ribosome biogenesis and function. *RNA Biol.* **14**, 1138–1152 (2017).
5. Kirino, Y. *et al.* Codon-specific translational defect caused by a wobble modification deficiency in mutant tRNA from a human mitochondrial disease. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 15070–15075 (2004).
6. Weng, H. *et al.* METTL14 Inhibits Hematopoietic Stem/Progenitor Differentiation and Promotes Leukemogenesis via mRNA mA Modification. *Cell Stem Cell* **22**, 191–205.e9 (2018).
7. Jain, M. *et al.* RNA editing of Filamin A pre-mRNA regulates vascular contraction and diastolic blood pressure. *EMBO J.* **37**, (2018).
8. McMillan, M. *et al.* RNA methylation influences TDP43 binding and disease pathogenesis in models of amyotrophic lateral sclerosis and frontotemporal dementia. *Mol. Cell* **83**, 219–236.e7 (2023).
9. Meyer, K. D. *et al.* Comprehensive Analysis of mRNA Methylation Reveals Enrichment in 3′ UTRs and near Stop Codons. *Cell* vol. 149 1635–1646 Preprint at https://doi.org/10.1016/j.cell.2012.05.003 (2012).
10. McIntyre, A. B. R. *et al.* Limits in the detection of m6A changes using MeRIP/m6A-seq. *Sci. Rep.* **10**, 1–15 (2020).
11. Smith, A. M., Jain, M., Mulroney, L., Garalde, D. R. & Akeson, M. Reading canonical and modified nucleobases in 16S ribosomal RNA using nanopore native RNA sequencing. *PLoS One* **14**,

e0216709 (2019).

12. Tavakoli, S. *et al.* Semi-quantitative detection of pseudouridine modifications and type I/II hypermodifications in human mRNAs using direct long-read sequencing. *Nat. Commun.* **14**, 334 (2023).
13. Liu, H. *et al.* Accurate detection of m6A RNA modifications in native RNA sequences. Preprint at https://doi.org/10.1101/525741.
14. Nguyen, T. A. *et al.* Direct identification of A-to-I editing sites with nanopore native RNA sequencing. *Nat. Methods* **19**, 833–844 (2022).
15. Batista, P. J. *et al.* m6A RNA modification controls cell fate transition in mammalian embryonic stem cells. *Cell Stem Cell* **15**, 707 (2014).
16. Dai, Q. *et al.* Quantitative sequencing using BID-seq uncovers abundant pseudouridines in mammalian mRNA at base resolution. *Nat. Biotechnol.* 1–11 (2022).
17. Begik, O. *et al.* Quantitative profiling of pseudouridylation dynamics in native RNAs with nanopore sequencing. *Nat. Biotechnol.* **39**, 1278–1291 (2021).
18. CURLCAKE. https://cb.csail.mit.edu/cb/curlcake/.
19. Workman, R. E. *et al.* Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nat. Methods* **16**, 1297–1305 (2019).
20. Fleming, A. M., Mathewson, N. J., Howpay Manage, S. A. & Burrows, C. J. Nanopore Dwell Time Analysis Permits Sequencing and Conformational Assignment of Pseudouridine in SARS-CoV-2. *ACS Cent. Sci.* (2021) doi:10.1021/acscentsci.1c00788.
21. Gamper, H. *et al.* Synthesis of Long RNA with a Site-Specific Modification by Enzymatic Splint Ligation. Preprint at https://doi.org/10.1101/2022.09.17.508400.
22. Leger, A. *et al.* RNA modifications detection by comparative Nanopore direct RNA sequencing. *Nat. Commun.* **12**, 7198 (2021).
23. Zhang, Z. *et al.* Systematic calibration of epitranscriptomic maps using a synthetic modification-free RNA library. *Nat. Methods* **18**, 1213–1222 (2021).

24. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

25. De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M. & Van Broeckhoven, C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* **34**, 2666–2669 (2018).

26. Cunningham, F. *et al.* Ensembl 2022. *Nucleic Acids Res.* **50**, D988–D995 (2022).

27. Wick, R. R., Judd, L. M. & Holt, K. E. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol.* **20**, 129 (2019).