The case for using Mapped Exonic Non-Duplicate (MEND) read counts in RNA-Seq experiments: examples from pediatric cancer datasets

Authors: Holly C. Beale, Jacquelyn M. Roger, Matthew A. Cattle, Liam T. McKay, Drew Thomson, Katrina Learned, A. Geoffrey Lyle, Ellen T. Kephart, Rob Currie, Du Linh Lam, Lauren Sanders, Jacob Pfeil, John Vivian, Isabel Bjork, Sofie R. Salama, David Haussler, Olena M. Vaske

Abstract

RNA sequencing (RNA-Seq) has been broadly adopted by biologists of many specialties for quantification of gene expression. Since the accuracy of measurements is dependent on the amount of sequencing performed, conscientious authors often report the number of reads generated. However, some types of reads are not informative for measuring the expression of coding genes. Unmapped and non-exonic reads do not contribute to gene expression quantification. Without molecular barcodes, duplicate reads that reflect gene expression cannot be distinguished from technical artifacts; they are uninformative for determining the degree to which RNA-Seq reflects the biology of the original sample. We surveyed bulk RNA-Seq datasets from 2179 individual tumors (2018 pediatric/adolescent/young adult; 66 adult; 95 unknown) from 48 cohorts, sequenced to a total depth of 0.2-668 million reads (median ($\tilde{x}$) 61 million; interquartile range (IQR) 53 million) to determine the fractions of uninformative reads. Unmapped reads constitute 1-77% of all reads ($\tilde{x}$ 3%; IQR 3%); duplicate reads constitute 3-100% of mapped reads ($\tilde{x}$ 27%; IQR 30%); and non-exonic reads constitute 4-97% ($\tilde{x}$ 25%; IQR 21%) of mapped, non-duplicate reads. Informative reads--Mapped, Exonic, Non-duplicate (MEND) reads--constitute 0-79% of total reads ($\tilde{x}$ 50%; IQR 31%). Further, we find that MEND reads and Mapped Non-Duplicate (MND) read counts have 0.22 and 0.23 Pearson correlations to the number of expressed genes expressed above 1 Transcripts Per Million, while total and mapped reads have corresponding correlations of -0.05 and -0.04. The correlation of MEND read counts to measured genes increases when only highly expressed genes are considered, while the correlation with MND reads decreases. Since the fraction of uninformative reads vary, we propose using only definitively informative reads, MEND reads, for the purposes of asserting the accuracy of gene expression measured in a bulk RNA-Seq experiment. We provide a Docker image containing 1) the existing required tools (RSeQC, sambamba and samblaster) and 2) a custom script. Together, these tools take a BAM file containing aligned reads and report MEND counts. We recommend that all results, sensitivity studies and depth recommendations use MEND units.

Introduction

Assessing the accuracy and reproducibility of gene expression results obtained from the analysis of RNA-Seq data has been a priority since the development of the assay. Seminal studies showed the relationship between the amount of sequence data generated during an experiment (depth of sequencing) and the reproducibility of the resulting gene expression measurements (Marioni et al., 2008; Mortazavi et al., 2008). However, RNA-Seq data is not homogenous. Of the tens of millions of sequences (reads) in a typical dataset (the RNA-Seq data generated from one biological sample), some reads cannot be mapped back to the reference transcriptome. Others map to genome regions outside of exons or have been duplicated during the library construction process or sequencing. Nearly all methods for quantifying gene expression in bulk RNA-Seq data count reads that align to exons in a gene; thus, unmapped and non-exonic reads do not contribute to measurements and are uninformative regarding the accuracy of the experiment (Bray et al., 2016; Li and Dewey, 2011). Therefore, considering the total number of reads as a proxy for RNA-Seq gene expression accuracy can result in inflated accuracy estimates.

Duplicate reads may be due to either highly abundant transcripts or technical error. The process of preparing RNA-Seq libraries involves PCR amplification. This step can generate duplicated identical or nearly identical reads. While the original read represents gene expression in the experimental system, the artifactual duplicate reads do not. However, not all duplicate reads are artifactual. Each gene has a finite number of unique read sequences that can be generated from it, and a very highly expressed gene can contain identical reads that reflect gene expression rather than a technical artifact (Klepikova et al., 2017).

Here we analyze 2179 bulk, paired end, polyA-selected RNA-Seq datasets to characterize the read types present in the datasets and evaluate what fraction of commonly reported data is unequivocally relevant to the accuracy of gene expression measurements. We compare the correlation of total reads and MEND reads to the number of measured genes.

Methods

We surveyed 2179 publicly available bulk RNA-Seq datasets. Accession numbers, clinical data and read counts for each dataset are in Table S1. Of the 2179 datasets, 2018 were from pediatric/adolescent/young adult cancer tumors, 66 were from adult cancer tumors, and 95 were from cancer tumors of individuals with unknown ages, where adults are defined as being over 30 years of age. Of the 1692 datasets with reported gender of the patient, 42% were female and 58% were male. Of the 602 datasets with reported race of the patient, 27 patients were Asian, 70 were Black/African American, 3 were Native Hawaiian or Other Pacific Islander, 494 were White and 7 were Other without further definition. None were American Indian or Alaskan Native. Of 861 datasets with reported results of the patient's Hispanic or Latino identity, 128 were Hispanic or Latino. The datasets came from five repositories (Table S2).

All libraries were prepared with polyA selection. All data were generated via paired-end Illumina sequencing technology. The median sequence length is 101 bases (Figure 1B). The TOIL RNA-Seq pipeline was run as previously described (Vivian et al., 2017). Briefly, adapters were removed with CutAdapt v1.9. Reads were then aligned with STAR v2.4.2a with indices based on GRCh38 and gencode v23. RSEM v1.2.25 was used to quantify gene expression. The source code of the pipeline is available at https://github.com/UCSC-Treehouse/pipelines.

Mapped, Exonic, Non-Duplicate (MEND) reads were quantified as previously described (Vaske et al., 2019). Briefly, duplicates were marked with Samblaster v0.1.22 and the RSeQC v2.7.10 tool read_distribution.py quantified exonic read and tag counts. The script parseReadDist.R estimated the number of MEND reads by counting tags in CDS exons, 5' UTR exons and 3' UTR exons and multiplying by reads per tag. The process for estimating MEND read counts is available as a stand-alone docker image at https://hub.docker.com/r/hbeale/treehouse_bam_qc/ and the source code is at https://github.com/UCSC-Treehouse/mend_qc.

Since a pair of reads provides information about two nearby sequences in a single transcript, read counts are reported in pairs. For example, 20 reads means that there are 20 pairs of reads. Datasets are assigned to cohorts based on project accession (for EGA and SRA datasets), by disease sub-study for NCI Therapeutically Applicable Research to Generate Effective Treatments (TARGET), or by disease for datasets in the St Jude Cloud. Cohorts were assigned IDs in descending order of size. Cohort assignments are intended to approximate a typical sequencing project performed by one research group at one sequencing center.

Results

Data sources and dataset characteristics

In the course of gathering pediatric cancer RNA-Seq datasets for our comparative RNA-Seq compendium (Vaske et al., 2019), we performed read type analysis on 2179 datasets from 48 cohorts of cancer patients. Original data sources are reported for each dataset (Table S1); repositories and cohort information is aggregated in Tables S2 and S3. The cohorts range in size from 3 to 337 datasets (Figure 1A); the median number of datasets in a cohort is 24.5. The source tumors represent a variety of hematologic and solid malignancies (Table 1).

Disease	n	percent
Acute lymphoblastic leukemia	680	31.20%
Acute myeloid leukemia	221	10.10%
Medulloblastoma	201	9.20%

Glioma	193	8.90%
Osteosarcoma	157	7.20%
Acute megakaryoblastic leukemia	103	4.70%
Ependymoma	98	4.50%
Ewing sarcoma	70	3.20%
Rhabdoid tumor	65	3.00%
Rhabdomyosarcoma	53	2.40%
Lymphoma	49	2.20%
Embryonal rhabdomyosarcoma	42	1.90%
Alveolar rhabdomyosarcoma	40	1.80%
Glioblastoma multiforme	29	1.30%
Choroid plexus carcinoma	25	1.10%
Synovial sarcoma	22	1.00%
Other	131	6.00%

Table 1

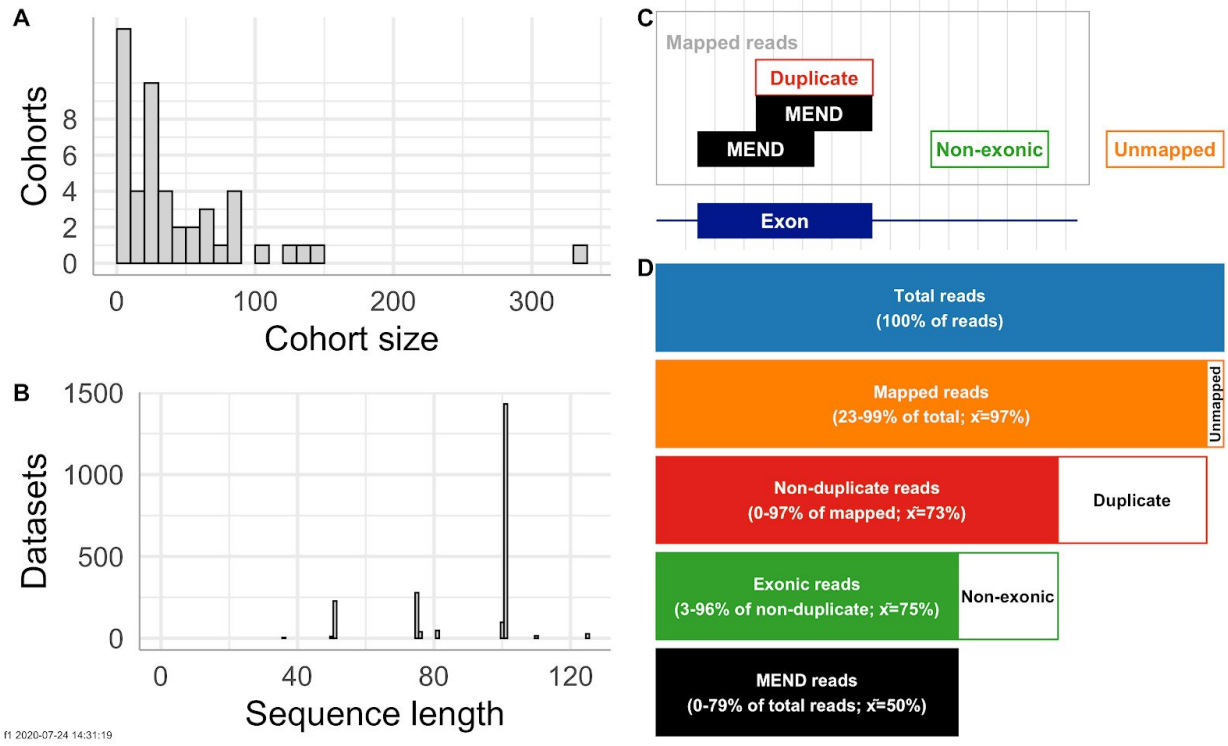


Figure 1

Caption for Figure 1: RNA-Seq datasets consist of 4 main types of sequencing reads, which were measured across datasets from 48 cohorts with a variety of read lengths. A. Distribution of number of datasets per cohort. B. Distribution of length of paired end reads in this study. C. Simplified schematic illustrating read types. The X axis (blue) is a genomic sequence containing an exon. The other boxes each represent one sequencing read. Two of five reads are MEND reads. Other reads do not map to the genome (Unmapped; orange border), map to a non-exonic region of the genome (Non-exonic; green border), or are duplicates of other reads (Duplicate; red border). The MEND reads (black) fit none of these categories and are considered informative for determining the accuracy of gene expression quantification. D. Schematic illustrating read type quantification. Bars representing uninformative reads are white with a colored border. For each informative fraction, the range and median (x˜) are reported.

Read types in RNA-Seq data

We interrogated the read types present in our RNA-Seq datasets in our gene expression quantification pipeline (Fig 1C). We obtained the number of total and mapped reads from the aligner log. We marked duplicates in the aligned BAM file, and counted them, along with exonic reads, using RSeQC. Duplicate reads are reported as a fraction of mapped reads, and exonic reads are reported as a fraction of non-duplicate reads.

Most RNA-Seq datasets contain a small percentage of unmapped reads. In the data from 2179 datasets, 75% of datasets have fewer than 6% unmapped reads (Fig 2A). The distribution is left-skewed with a long right tail. The value of excluding these from read counts is apparent, as these reads do not correspond to any known expressed gene; in 77 datasets, more than 25% of reads are unmapped. Including those reads in any measure of the sensitivity of gene expression measurement would misguide the researcher.

The percentage of mapped reads that are duplicate reads ("percent duplicates") is more varied. 426 datasets have more than 50% duplicates (Fig 2A). Some cohorts are characterized by high duplicate fractions (Fig 2B). For example, 72 of the 127 datasets in Cohort 4 have more than 98% duplicates. All 72 have a total sequencing depth above 170 million reads.
However, even cohorts with generally low duplicate fractions can contain anomalous datasets; of the 41 cohorts with a median of less than 50% duplicates, 26 contain at least one dataset with more than 50% duplicates.

If duplicate reads were only a function of genes being especially deeply sequenced, we would expect sequencing depth to explain most of the variability in the fraction of duplicate reads. The

total sequencing depth has a 0.58 Pearson correlation with the fraction of duplicate reads, explaining 34% of the variability (Supplemental Figure 1). The majority of the variability in the fraction of duplicate reads is independent of read depth. Consequently, the fraction of duplicate reads cannot be inferred from the total read depth.

Like percent of duplicates, the percent of non-exonic reads among all mapped, non-duplicate reads ("percent non-exonic") has a broad distribution compared to other read type fractions, with an IQR of 21%. 330 datasets have a fraction of non-exonic reads above 50%.

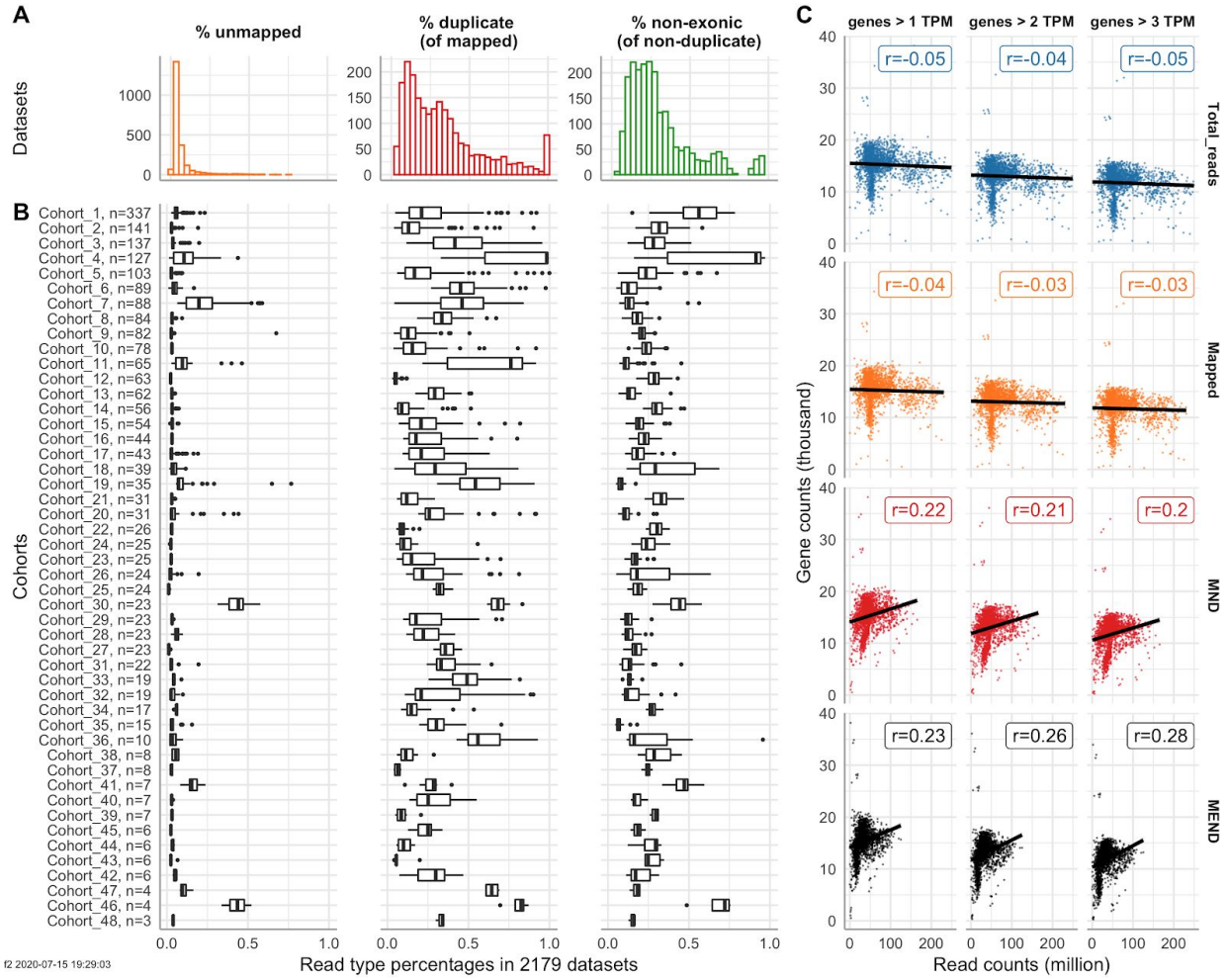


Figure 2

Caption for Figure 2: Read type fractions vary within and between cohorts, and MEND counts correlate best with measured gene counts. A. The percent distribution of different uninformative read types observed in 2179 datasets. B. The percentage of read types observed in cohorts, annotated with the number of datasets in the cohort. C: Relationships between the number of genes expressed in a dataset to the number of reads of different types. Only the 1996 datasets

with more than 100 measured genes and fewer than 250 million total reads were included. Correlations to the number of genes expressed above 1, 2 and 3 TPM are shown.

Correlation of read counts with the number of genes measured

If total or all mapped read counts were informative about the sensitivity of gene expression measurements, we would expect them to correlate with the number of expressed genes. When calculated using all 2179 datasets, total and mapped reads were inversely correlated (Pearson r=-0.4 for both) with the number of genes with expression above 1 Transcripts Per Million (TPM) (Supplemental Figure 2). We recalculated the correlations after excluding 1) the 78 datasets with fewer than 100 measured genes (all of which had more than 170 million total reads) and 2) the 105 datasets sequenced to more than 250 million total reads (such deep sequencing is usually intended for detecting rare events rather than measuring gene-level expression). In the resulting data from 1996 datasets, total and mapped reads are not correlated with the number of genes with expression above 1 TPM (Pearson r = -0.05 and -.0.04, respectively; Figure 2C). MND and MEND read counts are correlated at Pearson r = 0.22 and 0.23, respectively. When genes with higher expression are counted, the correlation of MEND counts to measured genes increases, while the correlation for MND decreases.

Conclusion

Researchers want to know that their data is sufficient for the measurements they're making. Here we show that, for the purpose of determining whether an RNA-Seq dataset is sufficient for accurately measuring expression of known genes, the fraction of relevant content of an RNA-Seq dataset (percent of MEND reads) varies substantially within and between cohorts. We confirm the relevance of MEND read counts to gene expression measurements by demonstrating that MEND read counts are correlated to the number of measured genes and total read counts are not.

There are several reasons why a survey of this breadth has not been previously performed. Obtaining and processing clinical datasets from multiple sources is an intensive effort. Tumor datasets are usually controlled access, and obtaining the 48 cohorts we report on here required multiple legal agreements (Learned et al., 2019). Analyzing read types requires genome-aligned reads; the files containing genome-aligned reads are large and are not generated when using the much faster pipelines that quantify gene expression via pseudoalignment. Large RNA-Seq cohorts such as GTEX and TCGA use consistent methods and exclude datasets that fail their stringent and consistent quality control (GTEx Consortium, 2017; Hoadley et al., 2018). They lack the kind of heterogeneity observed in our cohorts. In short, generating this data for more than 2000 datasets is time-consuming, expensive, and requires staff with a variety of expertises. We performed the analysis reported here in order to ensure the validity of the measurements we

include in the RNA-Seq compendium (https://treehousegenomics.ucsc.edu/public-data/) that we use for comparative analysis (Vaske et al., 2019).

Measuring the number of MEND reads in a dataset is specific to the alignment parameters and gene model. We use Gencode v23, which is inclusive, defining more than 60,000 genes. By default, the aligner we use, STAR, defines reads that map to as many as 20 positions as mappable. If we changed our pipeline, asking STAR to exclude reads mapping to more than 2 positions and using a more conservative gene model with 30,000 genes, the same dataset would have fewer MEND reads due to the loss of reads that map too much or map only to regions newly defined as non-exonic.

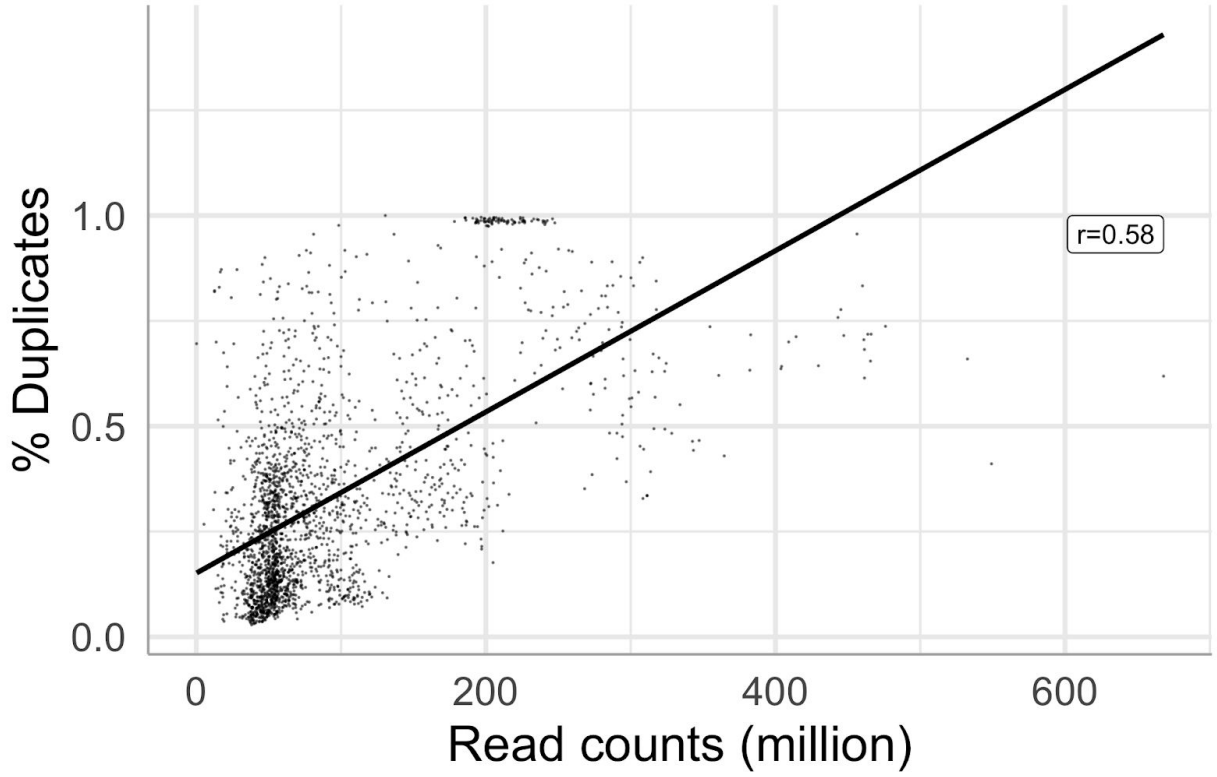
People planning RNA-Seq experiments look for guidance on how much sequencing their experiment requires. For comparing gene expression measurements between datasets, ENCODE recommends a minimum of 30 million mapped reads (ENCODE Project Consortium, 2011); the GEUVADIS consortium study had a minimum goal of 20 million reads (’t Hoen et al., 2013). However, of the 2078 datasets in this study with more than 30 million mapped reads, 16% contain fewer than 25% informative (MEND) reads. We speculate that these guidelines were not intended to include those datasets, some of which measure fewer than 100 genes.

Based on these results, we recommend that 1) publications reporting the results of an RNA-Seq study should report the number of MEND reads present in each dataset; 2) sensitivity studies should include read type fractions and report on the relationship between MEND reads and the measured outcome; and 3) sequencing depth recommendations should be based on MEND reads.

Acknowledgements

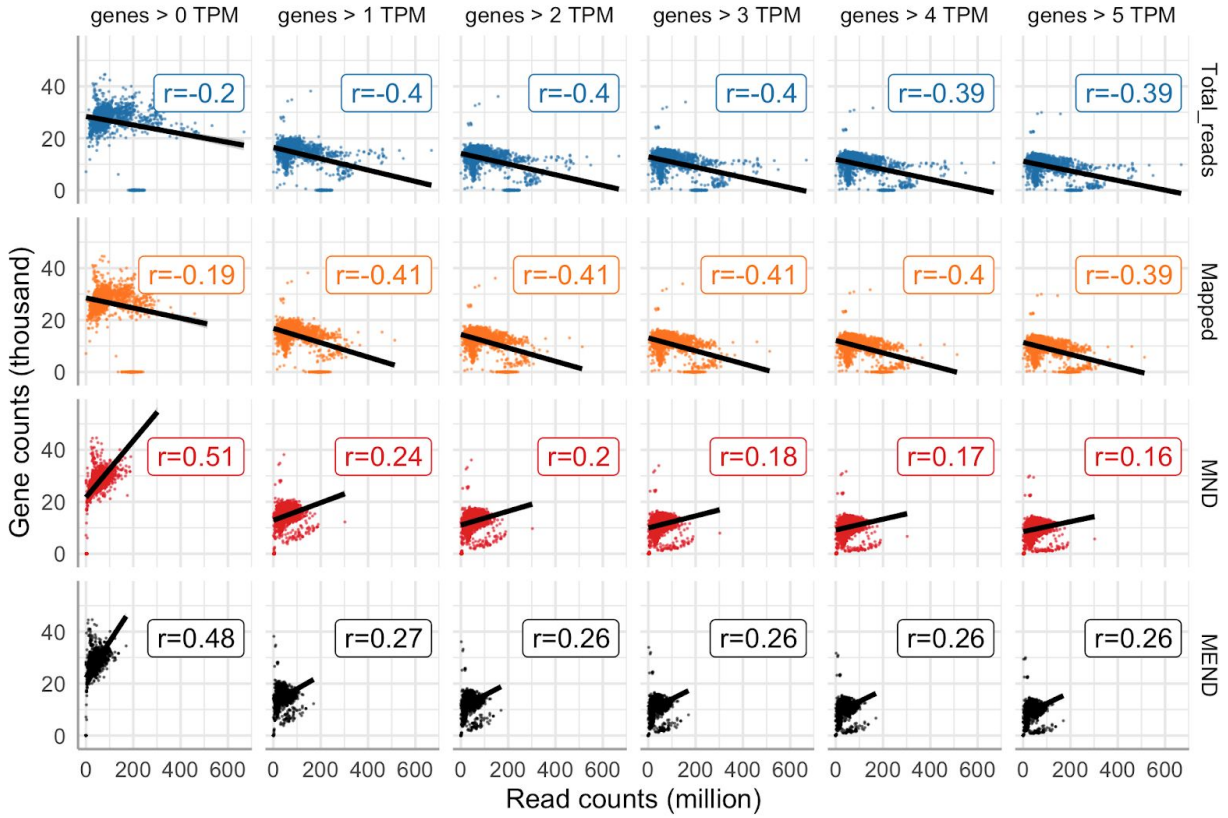
We acknowledge the work of all our colleagues at the Genomics Institute; the Computational Genomics Lab has provided an invaluable base for this work, allowing us to analyze large data sets relevant to pediatric cancer research. We thank Alejandro Sweet-Cordero and Alex G. Lee for valuable feedback on MEND analysis. We thank the many researchers who have shared their sequence data: https://treehousegenomics.soe.ucsc.edu/public-data/acknowledgments.html. Finally, we honor and thank all the children and adults who consented to donate their data to advance research in pediatric cancer.

Supplemental figures



Supplemental Figure 1

Caption for Figure S1. The percent of duplicate reads increases with the total number of reads in a dataset, but accounts for less than half of the variability. The Pearson correlation (r) between the values is 0.58 and explains 34% of the variability in the data (r^2=0.336); n = 2179.



Supplemental Figure 2

Caption for Figure S2. Of the four read types, MEND reads have the highest correlation to genes expressed above 1 TPM. The number of genes expressed (Y axis) above the threshold value (0, 1, 2, 3, 4, and 5 TPM, grouped by columns) are plotted against read counts (X axis). The type of reads counted (Total, Mapped, MND and MEND) are grouped by rows. The pearson correlation (r) is shown for each combination of read type and gene threshold. All 2179 datasets are included in each plot.

Table S1

https://docs.google.com/spreadsheets/d/1awKt5e3wYMWiliMwida1-8HLYrhUaOWP1GK26uV5ags/edit?usp=sharing

The accession numbers in Table S1 are the definitive sources of the RNA sequencing data. The DOI links to citations are provided for convenience. They may reflect the citation provided by the data provider, a citation we identified referring to the RNA sequencing data, or a citation we identified referring to the patient whose tumor was sequenced.

Table S2

Name	Abbreviation	URL
Short Read Archive	SRA	https://www.ncbi.nlm.nih.gov/sra
European Genome-phenome Archive	EGA	https://www.ebi.ac.uk/ega/home
St. Jude Cloud	SJC	https://www.stjude.cloud/
Database of Genotypes and Phenotypes	dbGaP	https://www.ncbi.nlm.nih.gov/gap/
Cavatica	Cavatica	https://cavatica.squarespace.com/

Table S3

cohort_code	cohort_name	source_repository	repository_cohort_accession	n_in_cohort
Cohort_1	TARGET-10	dbGaP	phs000218	337
Cohort_2	SJC_BALL	SJC	SJC-DS-1001	141
Cohort_3	TARGET-20	dbGaP	phs000218	137
Cohort_4	EGAD00001003279	EGA	EGAD00001003279	127
Cohort_5	SJC_AMLM	SJC	SJC-DS-1001	103
Cohort_6	phs000673	dbGaP	phs000673	89
Cohort_7	TARGET-40	dbGaP	phs000218	88
Cohort_8	phs000720	dbGaP	phs000720	84
Cohort_9	SJC_HGG	SJC	SJC-DS-1001	82
Cohort_10	SJC_EPD	SJC	SJC-DS-1001	78
Cohort_11	TARGET-52	dbGaP	phs000218	65
Cohort_12	EGAD00001001098	EGA	EGAD00001001098	63
Cohort_13	phs000768	dbGaP	phs000768	62
Cohort_14	SJC_ETV	SJC	SJC-DS-1001	56
Cohort_15	SJC_LGG	SJC	SJC-DS-1001	54
Cohort_16	SJC_CBF	SJC	SJC-DS-1001	44
Cohort_17	SJC_RHB	SJC	SJC-DS-1001	43
Cohort_18	EGAD00001001620	EGA	EGAD00001001620	39
Cohort_19	phs000699	dbGaP	phs000699	35
Cohort_20	CBTTC	Cavatica	CBTTC	31
Cohort_21	SJC_ERG	SJC	SJC-DS-1001	31
Cohort_22	SJC_PHALL	SJC	SJC-DS-1001	26
Cohort_23	EGAD00001001666	EGA	EGAD00001001666	25
Cohort_24	SJC_CPC	SJC	SJC-DS-1001	25
Cohort_26	phs000900	dbGaP	phs000900	24
Cohort_25	EGAD00001000648	EGA	EGAD00001000648	24
Cohort_30	TARGET-21	dbGaP	phs000218	23
Cohort_27	EGAD00001000356	EGA	EGAD00001000356	23
Cohort_28	EGAD00001000617	EGA	EGAD00001000617	23

Cohort_29	SJC_OS	SJC	SJC-DS-1001	23
Cohort_31	EGAD00001001927	EGA	EGAD00001001927	22
Cohort_32	EGAD00001002680	EGA	EGAD00001002680	19
Cohort_33	SRP126664	SRA	SRP126664	19
Cohort_34	EGAD00001000158	EGA	EGAD00001000158	17
Cohort_35	SRP092501	SRA	SRP092501	15
Cohort_36	EGAD00001000826	EGA	EGAD00001000826	10
Cohort_37	SJC_E	SJC	SJC-DS-1001	8
Cohort_38	SJC_MB	SJC	SJC-DS-1001	8
Cohort_41	TARGET-30	dbGaP	phs000218	7
Cohort_39	SJC_HYPO	SJC	SJC-DS-1001	7
Cohort_40	SJC_MEL	SJC	SJC-DS-1001	7
Cohort_42	EGAD00001000328	EGA	EGAD00001000328	6
Cohort_44	SJC_Other	SJC	SJC-DS-1001	6
Cohort_43	SJC_INF	SJC	SJC-DS-1001	6
Cohort_45	SJC_WLM	SJC	SJC-DS-1001	6
Cohort_47	TARGET-50	dbGaP	phs000218	4
Cohort_46	SRP006575	SRA	SRP006575	4
Cohort_48	SRP040454	SRA	SRP040454	3

Bibliography

Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. *34*, 525–527.

ENCODE Project Consortium (2011). Encode Standards, Guidelines and Best Practices for RNA-Seq.

GTEx Consortium (2017). Genetic effects on gene expression across human tissues. Nature *550*, 204–213.

Hoadley, K.A., Yau, C., Hinoue, T., Wolf, D.M., Lazar, A.J., Drill, E., Shen, R., Taylor, A.M., Cherniack, A.D., Thorsson, V., et al. (2018). Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. Cell *173*, 291-304.e6.

't Hoen, P.A.C., Friedländer, M.R., Almlöf, J., Sammeth, M., Pulyakhina, I., Anvar, S.Y., Laros, J.F.J., Buermans, H.P.J., Karlberg, O., Brännvall, M., et al. (2013). Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. Nat. Biotechnol. *31*, 1015–1022.

Klepikova, A.V., Kasianov, A.S., Chesnokov, M.S., Lazarevich, N.L., Penin, A.A., and Logacheva, M. (2017). Effect of method of deduplication on estimation of differential gene expression using RNA-seq. PeerJ *5*, e3091.

Learned, K., Durbin, A., Currie, R., Kephart, E.T., Beale, H.C., Sanders, L.M., Pfeil, J., Goldstein, T.C., Salama, S.R., Haussler, D., et al. (2019). Barriers to accessing public cancer genomic data. Sci. Data *6*, 98.

Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*, 323.

Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M., and Gilad, Y. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. *18*, 1509–1517.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods *5*, 621–628.

Vaske, O.M., Bjork, I., Salama, S.R., Beale, H., Shah, A.T., Sanders, L., Pfeil, J., Lam, D.L., Learned, K., Durbin, A., et al. (2019). Comparative Tumor RNA Sequencing Analysis for Difficult-to-Treat Pediatric and Young Adult Patients With Cancer. JAMA Netw. Open *2*, e1913968–e1913968.

Vivian, J., Rao, A.A., Nothaft, F.A., Ketchum, C., Armstrong, J., Novak, A., Pfeil, J., Narkizian, J., Deran, A.D., Musselman-Brown, A., et al. (2017). Toil enables reproducible, open source, big biomedical data analyses. Nat. Biotechnol. *35*, 314.