Title: Framework for determining accuracy of RNA sequencing data for gene expression profiling of single samples

Authors: Holly C. Beale[1,2], Jacquelyn M. Roger[3], Matthew A. Cattle[3], Liam T. McKay[3], Katrina Learned[2], A. Geoffrey Lyle[1,2], Ellen T. Kephart[2], Rob Currie[2], Du Linh Lam[2], Lauren Sanders[2,3], Jacob Pfeil[2,3], John Vivian[2,3], Isabel Bjork[2,†], Sofie R. Salama[2,3,4,†], David Haussler[2,3,4,†], Olena M. Vaske[1,2,†]

Institutions:

1 Department of Molecular, Cell and Developmental Biology, UC Santa Cruz, Santa Cruz, CA, USA

2 Genomics Institute, UC Santa Cruz, Santa Cruz, CA, USA

3 Department of Biomolecular Engineering, School of Engineering, UC Santa Cruz, Santa Cruz, CA, USA

4 Howard Hughes Medical Institute

† These senior authors contributed equally

Corresponding authors:

Holly C Beale (hcbeale@ucsc.edu)

Olena M Vaske (olena@ucsc.edu)

Abstract

Background

The clinical value of identifying aberrant gene expression in tumors is becoming increasingly evident. In order for multi-gene expression analysis to achieve wider adoption and eventually be developed as a Clinical Laboratory Improvement Amendments (CLIA)-approved test, the input sample requirements, sensitivity, specificity and reference ranges must be quantified.

Methods

We analyzed paired-end Illumina RNA sequencing (RNA-Seq) data from 1088 tumor samples from 29 projects. We categorized reads based on where and how well they map to the genome, as well as their PCR duplicate status. We subsampled 5 deeply sequenced samples, identified exceptionally highly expressed genes and samples with similar gene expression profiles.

Results

We addressed variability in RNA-Seq dataset composition by defining reference ranges for four types of reads found in sequencing data: unmapped (0-13%); mapped duplicate (2-66%); mapped non exonic (0-26%) and mapped, exonic, non-duplicate (MEND, 27-76%). With 20 million MEND reads, we detected over-expressed genes ("up-outlier" genes) with a median sensitivity of 96.1% and specificity of 99.8%; sample similarity had 96.6% sensitivity and 100.0% specificity.

Conclusions

This strategy for measuring RNA-Seq data content and identifying thresholds could be applied to a clinical test of a single sample, specifying minimum inputs and defining the sensitivity and specificity. We estimate that a sample sequenced to the depth of 70 million total reads will typically have sufficient data for accurate gene expression analysis.

Introduction

The role of gene expression profiling in diagnosis, prognosis and treatment selection is rapidly expanding (1,2). Diseases with few clinically relevant mutational profiles, like most pediatric cancers, have the potential to particularly benefit from this expansion (3).

The expression of protein coding genes can be measured across the whole transcriptome by sequencing messenger RNA (RNA-Seq). Relative measurements generated by this method have been shown by the FDA to be accurate and reproducible across platforms and sites (4).

However, to gain widespread adoption in the clinic and obtain CLIA approval, the parameters required for an accurate and reproducible result of an RNA-Seq experiment must be defined. The accuracy of RNA-Seq measurements depends on the depth of sequencing, measured in reads (5,6). Typically, tens of millions of paired-end reads are generated in a single RNA-Seq experiment. Each read is computationally assigned to its transcript of origin, and the expression level of that transcript is calculated based on the number of corresponding reads.

If too few reads are generated, the resulting quantifications may be inaccurate due to undersampling the transcriptome. However, even a deeply sequenced sample can be inaccurate. For instance, reads that cannot be mapped to a transcript do not contribute to its quantification; nor do reads mapped to non-exonic parts of the genome. Additionally, reads with identical sequences (duplicates) can result from technical artifacts or real biological abundance of the transcript. Consequently, to accurately measure gene expression, not only must the number of reads be sufficient, but the counted reads must contribute to the quantification and represent the biology of the original sample.

Previous studies of sufficient sequencing depth have measured input in total or mapped reads and have focused on average performance across many genes or sets of samples. Here we show that an alternative to counting the total number of reads is required for predicting accuracy in RNA-Seq experiments. We describe a method for measuring mapped, exonic, non-duplicate (MEND) reads, and we identify relevant MEND thresholds for two precision medicine assays, gene expression outlier and molecular similarity analyses. We also use our analysis to define the number of total reads in a sequencing project required to achieve sufficient MEND reads.

Methods

Tumor RNA-Seq data was obtained as previously described (2). All analyzed data was generated with Illumina sequencers from paired end libraries generated with polyA selection.

Quantifying gene expression in RNA-Seq data

RNA-Seq data is processed using the RNA-Seq CGL pipeline (7). Specifically, adapters are removed with CutAdapt v1.9 (8). FastQC v0.11.5 is used to obtain quality metrics on

each FASTQ input file (9). Reads are then aligned by STAR v 2.4.2a using indices generated from the human reference genome GRCh38 and the human gene models Gencode 23 (10). RSEM 1.2.25 is used to quantify gene expression (11).

Subsampling RNA-Seq fastq data

Subset generation is performed using the seqtk 1.3-r106 sample command (https://github.com/lh3/seqtk) using the FASTQ-formatted sequences from the top level (parent) samples as input. Random seeds are set to support reproducibility (Supplemental Table 2).

MEND pipeline

The genome-aligned reads generated by STAR in the RNA-Seq CGL pipeline are sorted by name using sambamba v0.6.1 (12), piped to Samblaster 0.1.22 for duplicate marking (13) and then sorted by coordinate by sambamba. The tool read_distribution.py from RSeQC v2.7.10 identifies exonic reads and quantifies them as tags (14). The pipeline, which combines these tools with a scripted calculation described below, is freely available at https://github.com/UCSC-Treehouse/mend_qc.

MEND analysis

Here and throughout the work the number of reads is reported in pairs. One million reads consists of two million sequences, one from each end of a million nucleotide fragments.

We gather 5 statistics from the RNA-Seq and MEND pipelines. The total number of reads is obtained from the FastQC output variable "total sequences." The counts of uniquely mapped reads and multiply mapped reads are extracted from the STAR-generated log ("Log.final.out ") where they are reported as "Uniquely mapped reads number" and "Number of reads mapped to multiple loci" respectively. These numbers are added to determine the number of mapped reads. The number of non-duplicate reads is obtained from the output of RSeQC's read_distribution tool where it is reported as "Total Reads" (the tool ignores reads that have been marked as duplicates). The number of mapped, exonic, non duplicate (MEND) reads are also derived from the output of RSeQC's read_distribution tool. We 1) calculate the number of reads per tag (from the top table of the read_distribution.py output, "Total Reads"/"Total Tags"); 2) calculate the number of exonic tags (from the bottom table, the sum of the Tag_count column for the rows "CDS_Exons," "5'UTR_Exons" and "3'UTR_Exons"). 3) multiply and halve the two previously calculated numbers: "number of exonic tags" * "number of reads per tag" /2. This calculation is performed by the authors' script parseReadDist.R (https://github.com/UCSC-Treehouse/mend_qc).

From these pipeline milestones, the read type composition of an RNA-Seq sample is inferred. The number of "Not mapped" reads is the difference between FastQC's total sequences and mapped reads. The number of duplicate reads is the difference between mapped reads and mapped, non-duplicate (MND) reads. The number of non-exonic reads is the difference between the MND and MEND read counts. MEND reads are counted as described above. Together, the four types of reads (unmapped, duplicates, non exonic and MEND) account for the total sequencing depth. As a further QC measure, the number of multi-mapped reads is divided by the sum of the mapped reads to determine the fraction of mapped reads that are multi-mapped.

Outlier genes

To detect gene expression outliers in an index sample, the expression in the sample is compared to expression in a cohort for each gene independently. Here the cohort was the Treehouse public compendium v9 (https://treehousegenomics.soe.ucsc.edu/explore-our-data/ & https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE129326), composed of 11,454 Illumina polyA RNA-Seq samples that were processed with the RNA-Seq CGL pipeline. For each gene, the cohort's outlier threshold is the third quartile + (1.5 * interquartile range) (2) of expression of that gene in all samples the cohort. If a gene in the index sample is expressed above the threshold, and if the expression is also among the 5% most of expressed genes within the sample, the gene is considered an over-expression

outlier in the index sample. 58,580 genes are measured; 2920 genes are among the 5% most of expressed genes any given sample.

Correlated samples

To identify samples correlated to an index sample, the Spearman correlation of the expression of genes in the index sample and each of the 11,454 samples in the compendium is calculated. Only genes expressed in more than 20% of all samples and among the 80% most variable subset of the remaining genes are used in the calculation. A correlated sample is defined as any sample with 1) the same diagnosis and 2) a correlation above a background correlation level (estimated as the 95th percentile of all correlations in the compendium, here 0.875).

Additional computation

The RNA-Seq CGL and MEND pipelines were run on OpenStack (https://www.openstack.org/) instances with 64GB of memory and 12 VCPUs. Analysis was performed and figures were generated using the software program R versions 3.4.4 and 3.5.2 (15), the RStudio IDE (RStudio Team, 2016), and the tidyverse, knitr, cowplot, gridGraphics, viridis and janitor packages (16–21). The code necessary to reproduce the analysis reported in this paper is in https://github.com/UCSC-Treehouse/mend_qc_publication.

Results

Individual gene measurements

To determine the required depth of sequence for a single RNA-Seq tumor sample, we first measured the effect of sequencing depth on gene expression. We measured gene expression in five samples with at least 80 million paired end reads (Table 1; Supplemental Table 2). We compared those measurements to the gene expression obtained when using only a subset of the original input data (Figures 1 and S1). The 13 subsets of each parent sample (the original five samples) ranged in depth from 1 million reads to over 90 million reads (Supplemental Table 3). With 1 million reads, the values of the measured cancer genes ranged from 65 to 578% of the values measured at the maximum depth; with 5 million reads, the range was 0 to 133%; with 10 million reads, the range was 88 to 120%.

Genes with similar expression

To assess how variable gene expression is at a given depth, we examined three sets of genes that span the range of expression observed for typical cancer genes in Figure 1A. For each parent sample, we looked at the value of each gene at the greatest depth it was sequenced (we considered this the final gene expression value) (Figure 1B and

S1B). The first bin contained genes that had a final value between 5.48 and 5.52 log2(TPM+1); the second between 3.48 and 3.52 log2(TPM+1) and the third between 1.48 and 1.52 log2(TPM+1). In order to make the comparisons equivalent, each boxplot was limited to 73 measurements, the number of genes in the smallest set. Consistent with previously published results (6, 22), measurements taken at greater depths were more consistent than measurements at lower depths within each expression bin. At the same depth, genes with low expression had more variability than genes with high expression.

Composition of RNA-Seq datasets

We next assessed composition of 1088 RNA-Seq datasets. We sequentially applied four measures that estimated the numbers of total reads, mapped reads, non-duplicate reads and exonic reads. With these numbers we determined the composition of the dataset with respect to four types of reads: unmapped; duplicate; non-exonic; and mapped, exonic, non-duplicate (MEND). Figure 2A shows the types of reads present in 55 representative datasets from the available cohort of 1088 samples. Unmapped reads (dark grey), and non-exonic reads (light grey) are not used for gene expression quantification and therefore these reads are subsequently excluded from read counts when determining the relationship between 1) read depth and 2) sensitivity and specificity. Duplicate reads (light green) can include technical artifacts as well as reads that represent true transcript abundance. Removal of duplicate reads reduces false

positives but also causes false negatives (23). We include duplicate reads in gene expression quantification. However, we exclude them from the tally of reads that are definitively informative. The dark green reads in Figure 2A are MEND reads and the samples are ordered by descending total read count. The number of reads of each type are not consistently predictable based on the total read count. We therefore chose MEND as a read count measure that consistently represented biological signal in an RNA-Seq dataset. We subsequently assess the sensitivity and specificity of RNA-Seq analysis assays based on the number of MEND reads rather than total reads.

Reference ranges of read types

Reference range, defining which assay outputs are statistically normal is required by the CLIA framework (https://www.cms.gov/Regulations-and-Guidance/Legislation/CLIA/Downloads/6064bk.pdf). Considering only the 834 of the 1088 samples with more than 20 million MEND reads (a threshold discussed below), we sought to define the typical composition of a tumor RNA-Seq sample. We calculated the mean and standard deviation (Fig 2B) of the fraction represented by each read type. We defined the reference range of a pediatric RNA-Seq tumor sample as the mean +/- 2 standard deviations (sd) for each read type: unmapped reads, duplicate (mapped) reads, and non-exonic (mapped, non-duplicate) reads. We also observed variability in the fraction of subtypes of mapped reads (Fig S2A). We defined a reference range for the subset of mapped reads mapped to multiple

positions (multi-mapped) rather than single positions (uniquely mapped) (Fig S2B). We then retained 690 of the 1088 samples that had read type composition values within the reference ranges. Because the median number of MEND reads in these samples was 37.4 million, we analyzed the effects of MEND read depth on samples with up to 48 million MEND reads.

Effect of sequencing depth on outlier analysis

Outlier analysis identifies genes with exceptionally high or low expression in single RNA-Seq datasets(2). This method compares an individual gene measurement to the expression value of the same gene in a cohort. The up outlier threshold for a gene is the 75th percentile of expression in the cohort plus 1.5 times the interquartile range. An up-outlier is a gene that both exceeds the up-outlier threshold for a cohort and is among the 5% of most highly expressed genes within the sample. We performed outlier analysis on the 65 subsamples from the 5 parent samples (Table 1 and Supplemental Table 1). We defined a true up outlier as one that was present in at least 3 of the 4 deepest subsets, 36, 40, 44 or 48 million MEND reads. Figure 3A shows the outlier calls in the 13 subsets of sample TH_Eval_019. False negative and false positive calls were more common at lower depths but were not entirely eliminated at high depths.

We combined the results from the 5 parent samples and evaluated the fractions of the accurate outlier calls. The trends observed in the parent sample S5 (Figure 3A) were

also observed in the other parent samples (Figure 3B): False calls are common at low coverage, decrease substantially, and continue at a low frequency even at high coverage. We calculated the sensitivity and specificity at each depth for each parent sample (Figure 3C). Table 2 reports the median sensitivity and specificity at each depth. Sensitivity increases sharply at 0-12 million MEND reads, then increases at a slower rate at 13-28 million reads. The highest four depths (36, 40, 44, and 48 million MEND reads) contribute to the definition of a true outlier which potentially confounds those sensitivity measurements. The 32 million MEND read data points have sensitivity and specificity similar to those at higher depths. To optimize sensitivity and specificity, we propose that the threshold for outlier analysis should be between 20 and 32 million MEND reads.

Molecular similarity analysis

We determined the effect of decreasing coverage on a second individual RNA-Seq analysis assay, gene expression correlation, which is used to characterize the similarity of a patient's sample to a disease cohort (Figure 4). A sample that was found to be correlated in 3 of the 4 deepest subsamples was considered to be a sample that was truly correlated to the parent. In contrast to outlier detection, false negatives were more common at low depths. Where 87 percent of outliers were detectable with 1 million MEND reads (the dark blue portion of the left-most blue bar, Figure 1B), only 2 percent of correlated samples were detectable with 1 million MEND reads (the dark purple

portion of the left-most purple bar, Figure 4B). Also in contrast to outlier detection, false positive correlated samples (light orange, Figure 4B) were much rarer than false positive outliers (light red, Figure 3B), and occured at high depths, rather than the low depths at which most false positive outliers were found.

Specificity and sensitivity were calculated for the 59 correlated background samples from four parent samples (Figure 4C, Table 2). Sensitivity increased sharply from 1 to 12 million MEND reads and slowly between 12 and 40 million MEND reads. The highest four depths (36, 40, 44, and 48 million MEND reads) contributed to the definition of a true correlation which potentially confounds those sensitivity measurements. The sensitivity and specificity at depths between 20 and 32 million MEND were identical.

Specifying the optimal number of total reads for a sequencing project

We revisited the original question of how much sequencing data is needed for gene expression outlier and molecular similarity assays. We considered each MEND read subset depth as a candidate threshold (e.g. 1 million MEND reads, 4 million MEND reads, etc). Taking the 690 samples that are within the read composition reference ranges described in Figure 2, we rescaled read counts for each to estimate the total number of reads required for that sample to meet the MEND read threshold. There is a degree of imprecision in the prediction because the fraction of duplicate reads is

depth-dependent. In rescaling, the fraction of each read type remains consistent and independent of depth. Table 2 contains the 95th percentile of the 690 required total read counts predicted for each threshold.

The two RNA-Seq analysis assays described here, outlier and molecular similarity analysis, both rely on having representative comparison cohorts. A threshold of 20 million MEND reads captures most of the achievable specificity and sensitivity while retaining sufficiently large comparison cohorts. By analyzing the predictions of total reads required for 20 million MEND reads, we see that 99.7% of samples with 70 million total reads will contain 20 million MEND reads, provided the samples are within the read composition reference ranges described in Figure 2 . Predictions for 38 samples are shown in Figure 2c.

Computing requirements for MEND pipeline

Based on the processing times of 65 subsamples on computers with 64GB of memory and 12 VCPUs, the expression quantification pipeline has a minimum duration of 28.4 minutes, and processing each million total reads requires, on average, an additional 6.7 minutes. The MEND QC pipeline has a minimum duration of 1.6 minutes for every RNA-Seq dataset plus 2.1 minutes per million total reads present in the data. For a sample with 70 million total reads, the expression pipeline would be predicted to take

497 minutes (8.3 hours). Adding MEND QC would increase the duration by 149 minutes (2.5 hours), increasing the total time by 30 percent relative to running only the STAR/RSEM expression pipeline.

Discussion

Here we showed that gene expression measurements increase in consistency with depth of sequencing and that the biologically relevant fraction of sequence varies greatly between samples. We have described a method for measuring MEND reads and demonstrated how to use it to identify specificity, sensitivity and appropriate input thresholds for two clinically relevant RNA-Seq assays, identification of over-expressed genes in a sample and identification of samples with similar gene expression profiles. We define reference ranges for read types in tumor RNA-Seq data. For 99.7% of samples within reference range for all four read types, a sample with 70 million total reads has at least 20 million MEND reads.

Measurements of RNA degradation, concentration, and purity are critical upstream quality measurements that are used to determine whether a sample should be sequenced (24). However, a sample with adequate upstream quality will not necessarily yield reproducible data. MEND analysis is a downstream assay that can detect the consequences of upstream problems with RNA input as well as problems caused by cross-species contamination, PCR overamplification and insufficient sequencing depth. MEND results can be used to determine how input amount affects the accuracy of

RNA-Seq-based assays. We showed that MEND reads are the most appropriate read type for defining quality standards for gene expression application of RNA-Seq data. We suggest that upstream quality measures, such as RNA Integrity Number (RIN), and MEND quality measures should be used in concert (25).

FastQC, RSeQC and other tools can interrogate the quality of RNA-Seq data (26). Like the upstream assays, many RNA-Seq quality tools are specific and useful for troubleshooting. MEND analysis uses logs generated by STAR and output from FastQC and RSeQC to quantify read types. With reference ranges defined by read type counts of representative cohorts, MEND analysis can be applied to single samples to provide a broad determination about whether a sample is sufficient for measuring gene expression. MEND analysis was developed as part of the Treehouse project (treehousegenomics.ucsc.edu), which relies on comparisons of single RNA-Seq datasets to large tumor cohorts. We needed a tool that allowed us to determine which available samples were appropriate for these cohorts.

Previous studies have demonstrated the dependence of RNA-Seq accuracy on sequencing depth (5,6,22,27). We took two steps to expand this observation . First, we defined input thresholds using the relevant read types (MEND). Secondly, we focused on clinically relevant analyses of the RNA-Seq data that can be applied to a single patient.

RNA-Seq is a powerful technology that has great clinical potential. However, performance parameters of this technology for single samples need to be rigorously studied and defined before the full potential of this technology is realized. Here we proposed the MEND framework that can be used to establish accuracy, precision, sensitivity, specificity, reference range of gene expression applications of single Illumina RNA-Seq experiments. We hope that this framework will pave the way for more RNA-Seq applications in the clinic and will contribute to the development of RNA-Seq gene expression profiling as clinical assays for individual patients.

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 would also like to thank the American Association of Cancer Researchers (AACR) for accepting our abstract on this topic at the 2017 annual meeting. Finally, we honor and thank all the children and adults who consented to donate their data to advance research in pediatric cancer.

References

1. Cardoso F, van't Veer LJ, Bogaerts J, Slaets L, Viale G, Delaloge S, et al. 70-Gene signature as an aid to treatment decisions in early-stage breast cancer. N Engl J Med 2016;375:717–29.

2. Newton Y, Rassekh SR, Deyell RJ, Shen Y, Jones MR, Dunham C, et al. Comparative RNA-Sequencing analysis benefits a pediatric patient with relapsed cancer. JCO Precis Oncol 2018;1–16.

3. Byron SA, Van Keuren-Jensen KR, Engelthaler DM, Carpten JD, Craig DW. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. Nat Rev Genet 2016;17:257–71.

4. Xu J, Gong B, Wu L, Thakkar S, Hong H, Tong W. Comprehensive assessments of RNA-seq by the SEQC consortium: FDA-led efforts advance precision medicine. Pharmaceutics 2016;8:8.

5. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res 2008;18:1509–17.

6. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 2008;5:621–8.

7. Vivian J, Rao AA, Nothaft FA, Ketchum C, Armstrong J, Novak A, et al. Toil enables reproducible, open source, big biomedical data analyses. Nat Biotechnol 2017;35:314.

8. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J 2011;17

9. Andrews S. FastQC: a quality control tool for high throughput sequence data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (Accessed May 2019)

10. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013;29:15–21.

11. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 2011;12:323.

12. Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: fast processing of NGS alignment formats. Bioinformatics 2015;31:2032–34.

13. Faust GG, Hall IM. SAMBLASTER: fast duplicate marking and structural variant read extraction. Bioinformatics 2014;30:2503–05.

14. Wang L, Wang S, Li W. RSeQC: quality control of RNA-seq experiments. Bioinformatics 2012;28:2184–85.

15. R Core Team. R: A language and environment for statistical computing. https://www.R-project.org/ (Accessed May 2019)

16. Firke S. janitor: Simple tools for examining and cleaning dirty data. https://CRAN.R-project.org/package=janitor (Accessed May 2019)

17. Garnier S. viridis: Default color maps from “matplotlib”. https://CRAN.R-project.org/package=viridis (Accessed May 2019)

18. Murrell P, Wen Z. gridGraphics: Redraw base graphics using “grid” graphics. https://CRAN.R-project.org/package=gridGraphics (Accessed May 2019)

19. Wickham H. tidyverse: Easily install and load “tidyverse” packages. https://CRAN.R-project.org/package=tidyverse (Accessed May 2019)

20. Wilke CO. cowplot: Streamlined plot theme and plot annotations for “ggplot2”. https://CRAN.R-project.org/package=cowplot (Accessed May 2019)

21. Xie Y. knitr: A comprehensive tool for reproducible research in R. In: Stodden V, Leisch F, Peng RD, editors. Implement Reprod Comput Res. Chapman and Hall/CRC; 2014.

22. Wang Y, Ghaffari N, Johnson CD, Braga-Neto UM, Wang H, Chen R, et al. Evaluation of the coverage and depth of transcriptome by RNA-Seq in chickens. BMC Bioinformatics 2011;12:S5.

23. Klepikova AV, Kasianov AS, Chesnokov MS, Lazarevich NL, Penin AA, Logacheva M. Effect of method of deduplication on estimation of differential gene expression using RNA-seq. PeerJ 2017;5:e3091.

24. Paszkiewicz KH, Farbos A, O’Neill P, Moore K. Quality control on the frontier. Front Genet 2014;5.

25. Schroeder A, Mueller O, Stocker S, Salowsky R, Leiber M, Gassmann M, et al. The RIN: an RNA integrity number for assigning integrity values to RNA measurements. BMC Mol Biol 2006;7:3.

26. ’t Hoen PAC, Friedländer MR, Almlöf J, Sammeth M, Pulyakhina I, Anvar SY, et al. Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. Nat Biotechnol 2013;31:1015–22.

27. Tarazona S, García-Alcalde F, Dopazo J, Ferrer A, Conesa A. Differential expression in RNA-seq: A matter of depth. Genome Res 2011;21:2213–23.

Tables

Table 1. RNA-Seq datasets used in the study.

Sample	Source	Disease	Sequence Length (per read end)	Percent Duplicates
S1	SRA: SRP040454	embryonal rhabdomyosarcoma	75	28
S2	doi: 10.24370/SD_BHJXBDQK	glioma (astrocytoma)	100	26
S3	EGA: EGAD00001000158	medulloblastoma	100	11
S4	dbGap: phs000673.v2.p1	undifferentiated spindle cell sarcoma	100	38
S5	EGA: EGAD00001001927	primitive neuroectodermal tumor of the central nervous system	101	26

Table 2. Sensitivity and specificity at MEND depths

MEND reads (million)	Outlier		Correlation		Approximate total reads needed (millions)
	Median sensitivity	Median specificity	Sensitivity	Specificity	
1	0.8649	0.9822	0.0172	1.0000	3.3
4	0.9442	0.9917	0.5690	1.0000	13.1
8	0.9571	0.9942	0.7759	1.0000	26.3
12	0.9742	0.9964	0.9483	1.0000	39.4
16	0.9805	0.9953	0.9483	1.0000	52.6
20	0.9610	0.9982	0.9655	1.0000	65.7
24	0.9798	0.9975	0.9655	1.0000	78.8
28	0.9805	0.9974	0.9655	1.0000	92.0
32	0.9871	0.9975	0.9655	1.0000	105.1
36	0.9910	0.9986	0.9828	1.0000	118.3
40	0.9914	0.9989	1.0000	1.0000	131.4

44	1.0000	0.9985	1.0000	1.0000	144.5
48	0.9957	0.9989	1.0000	0.9920	157.7

The approximate total reads needed is the 95th percentile of all total read counts predicted by rescaling. For each of the 690 samples with read types in the reference ranges, for each MEND threshold (column 1), we retained the fractional composition of the RNA-Seq sample, multiplying the number of reads for each read type by the factor required to have the desired number of MEND reads. The resulting read counts were summed to determine the number of total reads required to reach that threshold.

Figure Captions

Figure 1. Reproducibility of gene expression measurement increases with depth of sequence.

Caption: Gene expression (y-axis) is plotted against total reads used for the measurement. Data is shown for two parent samples (S1 and S2), see also Supplemental Figure 1A. Points show individual gene expression at various depths. B. Similarly expressed genes describe a range of measured expression. Each boxplot represents 73 gene measurements. Genes with expression 5.48-5.52, 3.48-3.52, or 1.48-1.52 log2(TPM+1) (horizontal blue lines) at the highest depth are grouped.

Figure 2. Total number of reads in a sample does not predict the number of MEND reads.

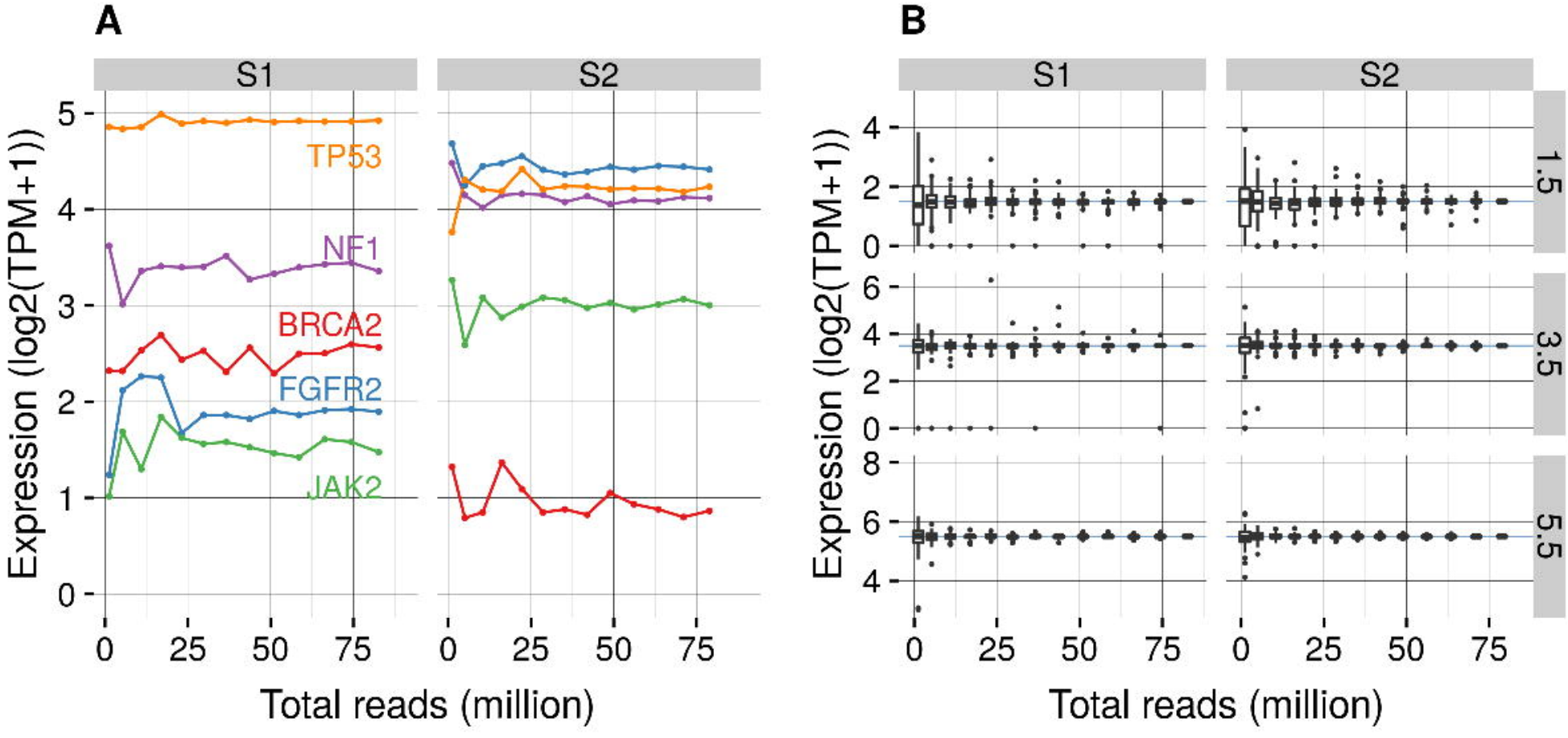
Caption: A. Four read types are present in fifty-five representative RNA-Seq datasets, including mapped, exonic, non-duplicate (MEND) reads. Asterisks indicate parent samples from Table 1. Samples with text annotations NM (not mapped), D (duplicates), NE (non-exonic), MM (multi-mapped), or M (MEND) have read type ratios outside the limits (Figs 2B, S2B). B. The typical range of each read type fraction ($\mu \pm 2\sigma$) are shown for the 834 samples with more than 20 million MEND reads. C. Read types in 38 of the 55 representative samples were rescaled to 20 million MEND reads.

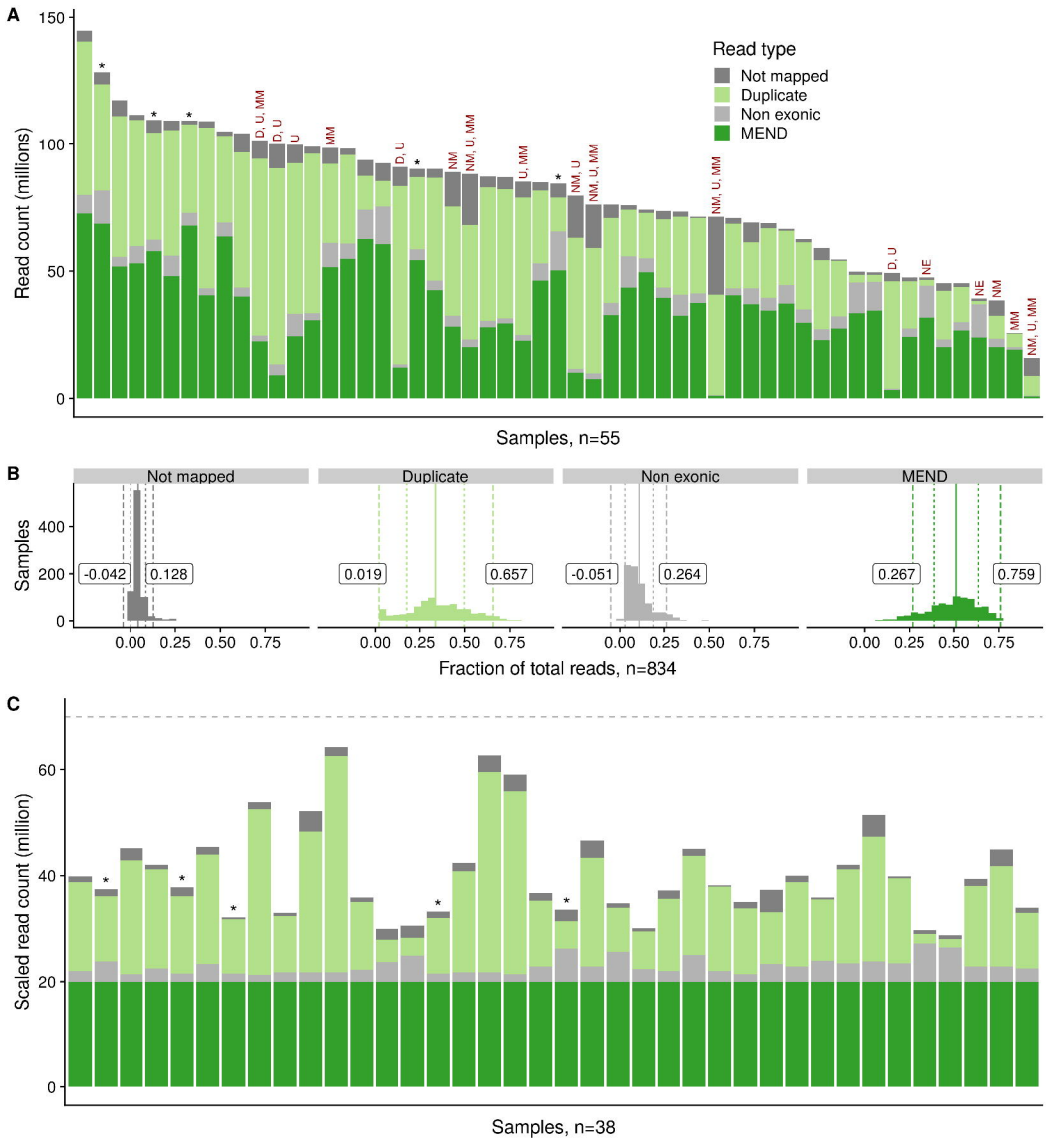
Figure 3. Identification of the number of MEND reads required for accurate outlier analysis.

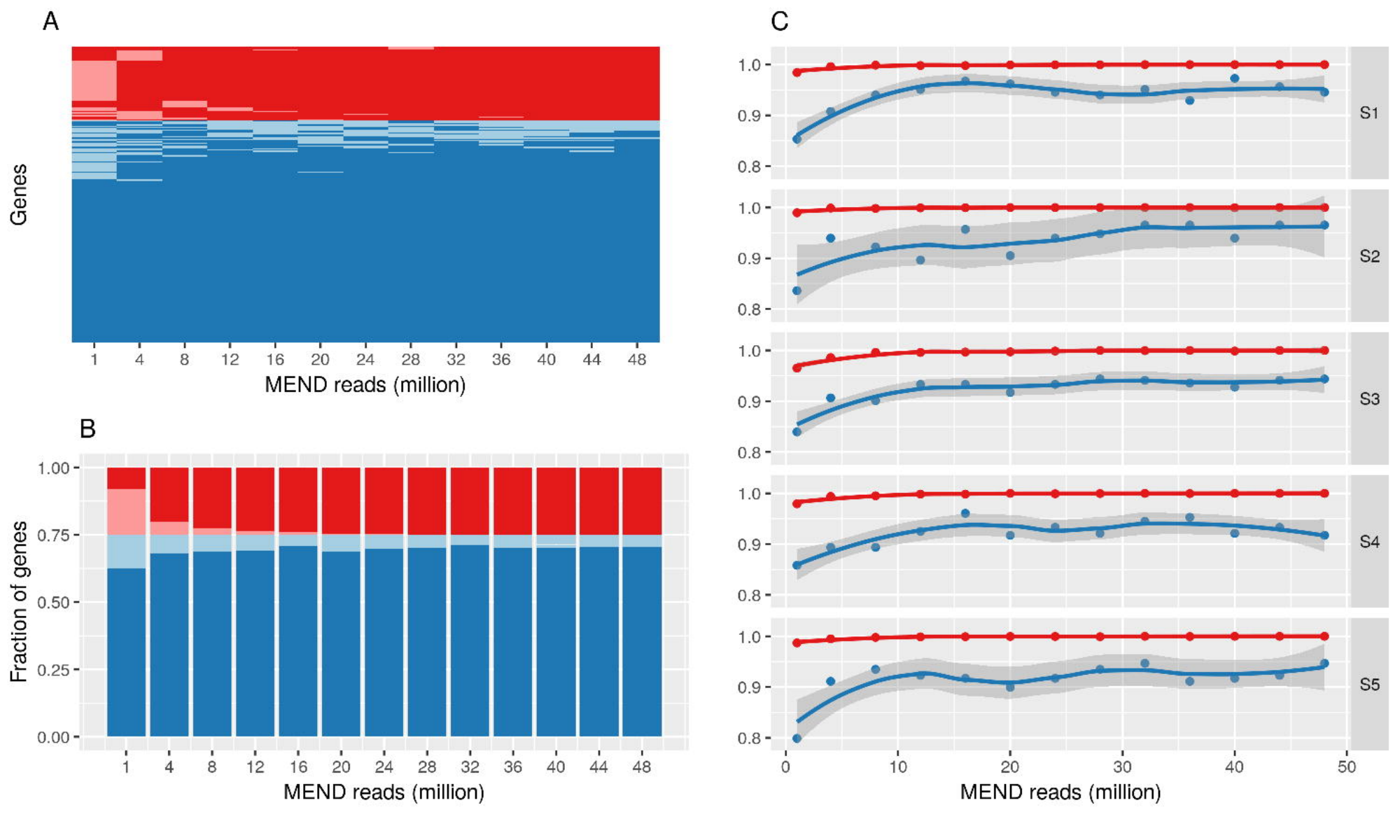
Caption: The outlier status of S5 (A) and the frequency in all 5 samples (Table 1) (B) of genes (y-axis) at increasing depths (x-axis) is dark blue for true positive, dark red for true negative, light blue for false negative, and light red for false positive. **C.** Sensitivity (blue) and specificity (red) at increasing depths (x-axis) for each parent sample. A locally weighted regression estimate (line) and confidence interval (CI) is depicted for sensitivity and specificity. The CI for specificity is narrower than the plotted line.

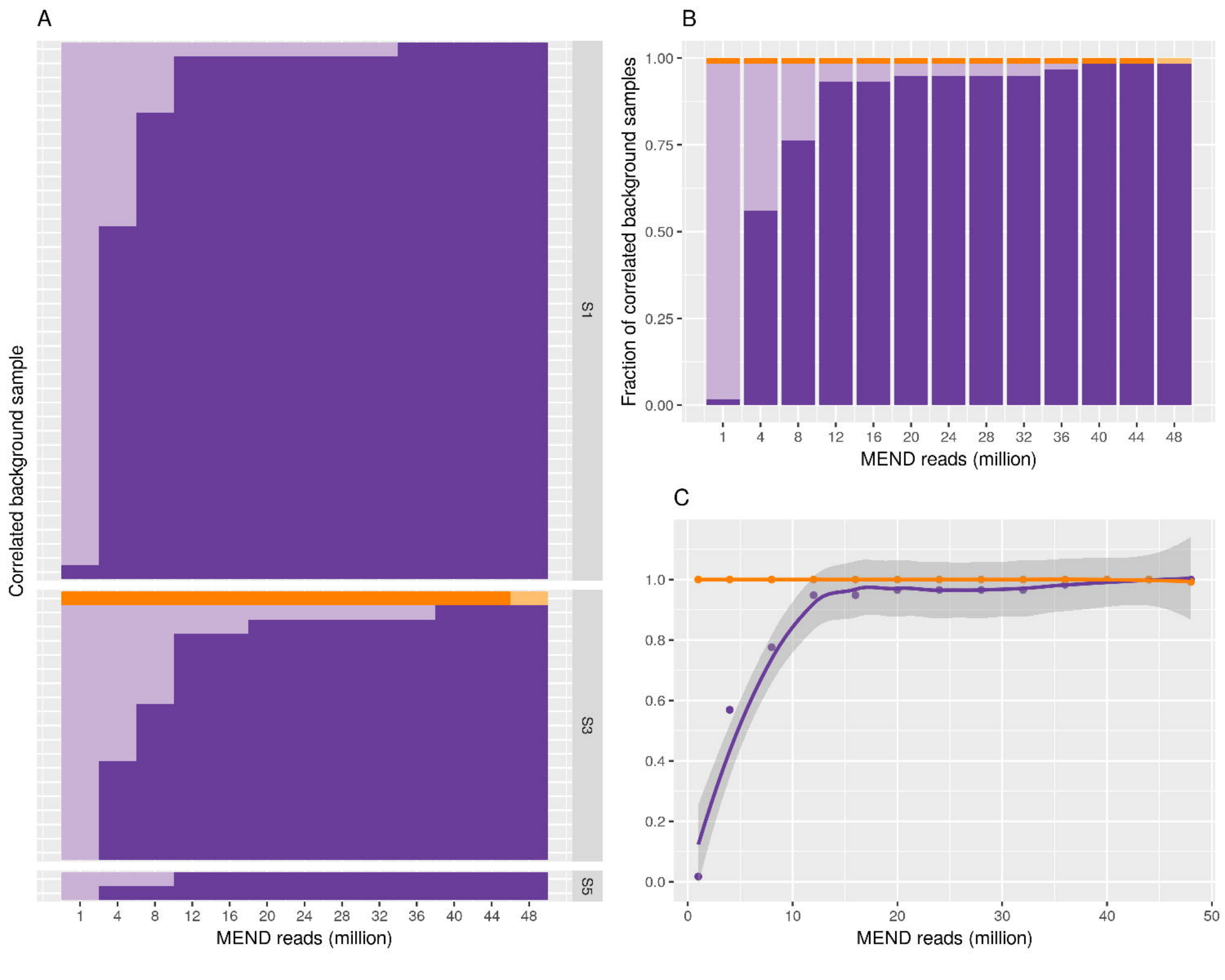
Fig. 4. Definition of MEND threshold for sample similarity analysis.

Caption: A. The correlation status of cohort samples (y-axis) for each parent sample at increasing depths (x-axis) is dark purple for true positive, dark orange for true negative, light purple for false negative, and light orange for false positive. **B.** Frequency of calls in A, colored as in A. **C.** Sensitivity (purple) and specificity (orange) are plotted for each parent sample. A locally weighted regression estimate (line) and confidence interval (CI) is depicted for sensitivity and specificity. The CI for specificity is narrower than the plotted line.


A
S1
S2
TP53
NF1
BRCA2
FGFR2
JAK2
Expression (log2(TPM+1))
Total reads (million)
B
S1
S2
1.5
3.5
5.5
Expression (log2(TPM+1))
Total reads (million)


A
Read count (millions)
150
100
50
0
Read type
Not mapped
Duplicate
Non exonic
MEND
D, U, MM
D, U
U
MM
D, U
NM
NM, U, MM
U, MM
NM, U
NM, U, MM
NM, U, MM
D, U
NE
NE
NM
MM
NM, U, MM
Samples, n=55
B
Not mapped
Duplicate
Non exonic
MEND
Samples
400
200
0
-0.042
0.128
0.019
0.657
-0.051
0.264
0.267
0.759
0.00
0.25
0.50
0.75
Fraction of total reads, n=834
C
Scaled read count (million)
60
40
20
0
Samples, n=38


A
Genes
1
4
8
12
16
20
24
28
32
36
40
44
48
MEND reads (million)
B
Fraction of genes
1.00
0.75
0.50
0.25
0.00
1
4
8
12
16
20
24
28
32
36
40
44
48
MEND reads (million)
C
1.0
0.9
0.8
S1
S2
S3
S4
S5
0
10
20
30
40
50
MEND reads (million)


A
Correlated background sample
S1
S3
S5
1
4
8
12
16
20
24
28
32
36
40
44
48
MEND reads (million)
B
Fraction of correlated background samples
1.00
0.75
0.50
0.25
0.00
1
4
8
12
16
20
24
28
32
36
40
44
48
MEND reads (million)
C
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
40
50
MEND reads (million)