Supplementary Methods

Data Overview: We analyzed the publicly available SRA data, the latest (v6) release of GTEx samples, and the latest release of TCGA. The SRA data consisted of 49,848 publicly available samples spanning over 146 terabases of reads. These reads were downloaded and analyzed, resulting in a final set of48,558 samples that could be fully downloaded and processed (1,290 samples could be downloaded only partially and were excluded, see Supplementary Methods). The GTEx data consisted of 9,662 samples spanning nearly 900 billion reads, 550 individuals and the K562 cell line, and 32 tissues. These reads were downloaded and analyzed, resulting in a final set of 9,479 samples that could be fully downloaded and processed (183 could be downloaded only partially; see Supplementary Methods). The TCGA data consisted of 11,350 samples spanning over 1.6 trillion reads, 10,040 individuals, and 33 cancer types. With the exception of reads from one sample for which the raw data was formatted incorrectly (see below), all TCGA RNA-seq samples were fully downloaded and processed.

Alignment. GTEx and public SRA samples were selected by searching the SRA website as discussed in *selecting GTEx, SRA and TCGA samples* subsection. Samples that could not be downloaded using fastq-dump were eliminated as described in incomplete and invalid input data. Samples were aligned in a spliced fashion to the hg38 assembly of the human genome using Rail-RNA. Alignments were performed in batches on computer clusters rented from the Amazon Web Services Elastic MapReduce service. The alignment pipeline was divided into two phases, where the first phase ("preprocessing") downloads and reformats the data and the second phase performs spliced alignment. Outputs of the pipeline include, for each sample, a junction coverage file (similar to a TopHat "junctions.bed" file) and a BigWig file[37] containing a genome-wide coverage vector. Further details are presented in *alignment with Rail-RNA* section below and the Rail-RNA study[21]. Gene and exon counts were compiled using the BigWig files output by Rail-RNA and the Gencode v25 annotation[38]. For exon counts, we first obtained a set of non-overlapping "unioned" exons. Gene and exon counts were compiled into per-project tables and RangedSummarizedExperiment objects. We expanded the tables with several metadata columns containing, for example, read count, paired-end status, GEO accession, and the tissue type as predicted by the SHARQ beta resource (http://www.cs.cmu.edu/~ckingsf/sharq/). For GTEx and TCGA we included metadata provided by the respective projects as detailed in the counting genes and exons.

Use cases. All R code used for the analyses performing the use cases is available from the website: http://leekgroup.github.io/recount-analyses/ including HTML version of the following PDF files (Supplementary Code 2).

- Use case 1. GTEx comparison, Supplementary Note 1: http://leekgroup.github.io/recount-analyses/example_gtex/compare_with_GTEx_reproducible.pdf
- Use case 2. Meta-analysis, Supplementary Note 2: http://leekgroup.github.io/recount-analyses/example_meta/meta_analysis.pdf
- Use case 3. Multi-level differential expression analyses gene/exon, Supplementary Note 3: http://leekgroup.github. io/recount-analyses/example_de/recount_SRP032789.pdf and annotation agnostic, Supplementary Note 4: http://leekgroup.github.io/recount-

analyses/example_de/recount_DER_SRP032789.pdf and validation of results in a second study, Supplementary Note 5: http://leekgroup.github.io/recount-analyses/example_de/recount_SRP019936.pdf

Selecting GTEx samples. On November 21, 2015, we queried the SRA website at http://www.ncbi.nlm.nih.gov/sra for RNA-seq samples in the GTEx project (accession: SRP012682). Precise search terms were (SRP012682) AND "strategy rna seq"[Properties]. A screenshot is available at https://github.com/nellore/runs/raw/master/gtex/SRA_GTEx_search_screenshot_6.37.16_PM_ET_11.21.2015.png (Supplementary Code 3). We used the "send to file" function to download search results to a table, available at https://github.com/nellore/runs/raw/master/gtex/SraRunInfo.csv. The table lists 9,795 run accessions, some of which were mmPCR-seq. We restricted our attention to the 9,662 mRNA-seq samples in the v6 release of the GTEx consortium.

Selecting SRA samples. Public SRA samples were selected from SRA as follows. On February 3, 2016, we queried the SRA website at http://www.ncbi.nlm.nih.gov/sra for publicly available human RNA-seq samples. Precise search terms were "platform illumina"[Properties] AND "strategy rna seq"[Properties] AND "human"[Organism] AND "cluster public"[Properties] AND "biomol rna"[Properties]. A screenshot is available at https://github.com/nellore/ runs/raw/master/sra/v2/hg38/SRA_RNA-seq_search_screenshot_3.33.19_PM_ET_02.03.2016. png. We used the "send to file" function to download the search results to a table, available at https://github.com/nellore/runs/raw/master/sra/v2/hg38/SraRunInfo.csv. The table lists 50,186 run accessions, each with metadata fields including layout (SINGLE or PAIRED) and number of reads (i.e., read pairs for paired-end samples).

Selecting TCGA samples. On September 29, 2016, we used the SPARQL query interface of Seven Bridges Genomics Cancer Genomics Cloud (CGC)[39] to retrieve storage paths of the 11,350 RNA-seq samples in TCGA. The script https://github.com/nellore/runs/blob/master/tcga/ tcga_file_list.py reproduces this query, and the resulting table is available at https://raw.githubusercontent.com/nellore/runs/master/tcga/tcga_file_list.tsv.

Alignment with Rail-RNA. We used Rail-RNA v0.2.3[21] to align all GTEx and SRA samples we could download to the hg38 assembly. We used Rail-RNA v0.2.4b to align TCGA samples; v0.2.4b differs from v0.2.3 only by adding preprocessing capabilities that accommodate the TAR archives in which raw TCGA RNA-seq data on CGC are stored. Alignment was performed using the Amazon Web Services Elastic MapReduce commercial cloud computing service. Spot instances allow users to bid for excess computing capacity. If the fluctuating market price drops below a users's bid, the instances could be lost, halting the computation. So saving money by bidding for spot instances comes with risk, and rather than aligning all samples in one batch, we distributed this risk by dividing alignment up into batches. For GTEx, we randomly divided the set of 9,662 samples up into 30 batches, each with 322 or 323 samples; for other SRA samples, we randomly divided the set of samples into 100 batches, each with 501 or 502 samples; for TCGA samples, we randomly divided the set of 11,350 samples up into 30 batches, each with 378 or 379 samples. Analysis of each batch was itself divided into a preprocessing job flow and an alignment job flow. A preprocessing job flow writes preprocessed reads to Amazon's cloud storage service S3

after either (a) using SRA Tools fastq-dump (https://github.com/ncbi/sra-tools) to download compressed reads from the National Center for Biotechnology Information server (for samples stored on SRA) or (b) downloading TAR archives storing raw reads via temporary links to Amazon's cloud storage service S3 produced by the CGC API (for TCGA samples). Preprocessing and alignment job flows was run on clusters of m3.xlarge instances, c3.2xlarge instances, or c3.8xlarge instances. Each m3.xlarge instance has 4 Intel Xeon E5-2680 v2 (Ivy Bridge) processing cores and 15 GB of RAM; each c3.2xlarge instance has 8 Intel Xeon E5-2680 v2 (Ivy Bridge) processing cores and 15 GB of RAM; each c3.8xlarge instance has 32 Intel Xeon E5-2680 v2 (Ivy Bridge) processing cores and 60 GB of RAM. Our GTEx alignment runs may be reproduced by following instructions at https://github.com/nellore/runs/blob/master/gtex/README.md. Our SRA alignment runs may be reproduced by following instructions at https://github.com/nellore/runs/blob/master/sra/v2/README.md. Our TCGA alignment runs may be reproduced by following instructions at https://github.com/nellore/runs/blob/master/tcga/README.md. Alignment of GTEx data is described in more detail in the supplementary material of [22].

Incomplete and invalid input data. For the analysis of the public SRA samples, some samples could not be downloaded, due to failures of the fastq-dump software. The issue persisted even when we attempted to restart the fastq-dump process. Samples exhibiting this issue can be excluded from analysis. These samples are listed here: https://github.com/nellore/runs/raw/master/sra/v2/hg38/NOTES. That file also describes two other preprocessing issues that led us to exclude a few other samples from analysis: (1) sequence input encoded in a manner we did not recognize; (2) miscellaneous errors reported by fastq-dump. In addition, for each of a small number of both GTEx and public SRA samples, fastq-dump would return success (an exit code of 0) when an error had occurred and the number of reads output by the tool disagreed with the number of reads from SRA metadata given in the SraRunInfo.csv file for that sample. So for each sample, we compared the number of reads ingested by Rail-RNA with the number of reads reported in SraRunInfo.csv. Sometimes, the sample was listed as paired-end when the number of mates (i.e., twice the number of read pairs listed in SRARunInfo.csv) was exactly twice the number of reads Rail-RNA processed; other times, the sample was listed as single-end but the number of reads Rail-RNA processed was greater than the number of reads listed. In these cases, we made note of our suspicion that the library layout listed on SRA was incorrect. The table https://github.com/nellore/runs/raw/master/gtex/incomplete.tsv (respectively, https://github.com/nellore/runs/raw/master/sra/v2/incomplete.tsv) lists all GTEx (respectively, public SRA) samples that either could not be downloaded or were incompletely downloaded, and it includes these suspected layout misreports in a column. This table was generated by the script https://github.com/nellore/runs/raw/master/gtex/incomplete.py (respectively, https://github.com/nellore/runs/raw/master/sra/v2/incomplete.py) for GTEx (respectively, public SRA) samples. Finally, the numbers of records in the two FASTQ files for the paired-end TCGA sample corresponding to the TAR archive "TCGA-AB-2909-03A-01T-074413 rnaseq fastq.tar" were mismatched, so we did not align it. Details are available at https://github.com/nellore/runs/ blob/master/tcga/NOTES.

Alignment cost. Costs to run preprocessing/alignment job flows on Amazon Elastic MapReduce as well as store data on Amazon S3 and transfer results back to our local cluster are broken down by day and Amazon Web Services service in three CSV files: for

our GTEx runs, see https://github.com/nellore/runs/blob/master/gtex/costs.csv; for our runs on other SRA samples, see https://github.com/nellore/runs/blob/master/sra/v2/costs.csv; and for our runs on TCGA samples, see https://github.com/nellore/runs/blob/master/tcga/costs.csv. The total cost of our GTEx runs was US $28,368.15, the total cost of our runs on other SRA samples was US $82,343.47, and the total cost of our TCGA runs was US $30,913.24. For a fine-grained picture of cluster activity in time during our GTEx runs, see Section 3.1 of the Supplementary Material of [22].

Counting genes and exons. When creating the count tables, genes and exons (features) were determined using the Gencode v25 (CHR regions) annotation; specifically ftp://ftp.sanger.ac.uk/pub/gencode/Gencode_human/release_25/gencode.v25.annotation.gff3.gz. For each gene, we first obtained a set of non-overlapping exonic intervals by taking the "union" of the gene's annotated exons. We call these "unioned exons". We summed per-base coverage across each unioned exon using bwtool[40] version 1.0. We obtained gene-level counts by summing per-base coverage totals of its constituent unioned exons. For each SRA project and each feature type we created a RangedSummarizedExperiment object[41] containing columns for:

• SRA study id

• SRA sample id

• SRA experiment id

• SRA run id

• read counts as reported by SRA

• number of reads downloaded from SRA and subsequently aligned with Rail-RNA

• proportion of reads reported by SRA that aligned

• whether the sample was paired-end or not

• whether SRA likely misreported the paired-end label (as explained above)

• mapped read count by Rail-RNA

• coverage area under the curve (AUC), i.e. total number of aligned bases not including soft-clipped bases

• tissue type as predicted by SHARQ (http://www.cs.cmu.edu/~ckingsf/sharq/)

• cell type as predicted by SHARQ

• biosample submission date

• biosample publication date

• biosample update date

• average read length as reported by SRA

• GEO accession id

• sample title as reported by GEO

• sample characteristics as reported by GEO

• name of the coverage BigWig file

For GTEx we included the metadata from the file "GTEx Data V6 Annotations SampleAttributesDS.txt" available from http://www.gtexportal.org/home/datasets. For TCGA, we consolidated metadata from three sources: a query via the Genomic Data Commons (GDC) API (https://gdc.cancer.gov/developers), queries via the CGC API[39], and queries using TCGAbiolinks[42]. We merged the information from all three sources and found metadata information for 11,285 samples (including the sample that we later discarded). For each sample we normalized the coverage to 40 million 100 base-pair reads using the AUC information. We then computed the base-level coverage sum for each project using the sample normalized coverage with bwtool[40] version 1.0 and divided the coverage sum by the number of samples in the project resulting in the project mean coverage. We then stored the mean coverage for each SRA project in a BigWig file[37]. The mean coverage can then be used to identify expressed regions with derfinder[23] via the *recount* Bioconductor package. The code and log files for creating these files as well as the *recount2* website are available at https://github.com/leekgroup/recount-website (Supplementary Code 4).

Exon-exon junction output processing. We postprocessed Rail-RNA's cross-sample junction TSVs to obtain two gzip-compressed files for each project on SRA: [SRA project accession number].junction id with transcripts.bed.gz, a file in BED format that contains junction coordinates, assigns each junction a unique identifier in the name column, and also lists which tran- scripts in GENCODE v24[38] contain the junction or its corresponding donor and acceptor site; and [SRA project accession].junction coverage.tsv.gz, which for each junction identifier gives the number of reads that map across the junction in each sample in the project. This postprocessing was performed by two scripts run in succession: https://github.com/nellore/runs/blob/master/sra/v2/hg38/junctions_by_project.py and https://github.com/nellore/runs/blob/master/sra/v2/hg38/add_knowngene.py. For each project, we created a RangedSummarizedExperiment using the sample metadata as in the exon and gene cases while annotating the exon-exon junctions with the:

• exon-exon junction id

• GENCODE v24 transcript id

• transcript name, gene id and gene symbol based on Gencode v25

• class: annotated, exon skip, alternative end, fusion, novel

• proposed gene id and gene symbol based on Gencode v25

The code is available at https://github.com/leekgroup/recount-website.

Meta-analysis use case. We downloaded processed data from the *recount2* website using the *recount* Bioconductor package. We selected colon samples labeled as controls from studies SRP029880 (a study of colorectal cancer[25], n=19) and SRP042228 (a study of Crohn's disease[26], n=41). Whole blood samples labeled as controls were taken from SRP059039 (a study of virus-caused diarrhea, unpublished, n=24), SRP059172 (a study of blood biomarkers for brucellosis, unpublished, n=47) and SRP062966 (a study of lupus, unpublished, n=18). We then compared these control blood to control colon using limma-voom[17] to identify genes that were differentially expressed. We then selected samples from the GTEx project, processed and downloaded from *recount2* project SRP012682[24]. We performed three different differential expression analyses: (1) comparing whole blood to colon, (2) comparing whole blood to lung, and (3) comparing different batches of expression data. For both the meta-analysis and the GTEx analyses we get rankings of the genes for differential expression. We then made concordance at the top plots comparing the rankings of the analysis from SRA[43]. In the SRA analysis we compared control blood samples to control colon examples. When we compare the ranking of genes for differential expression in this analysis to the ranking for the same analysis in GTEx we see strong concordance. Comparison of blood to lung shows worse concordance of differential expression results. Finally, as a negative control we plotted concordance between differential expression for colon versus blood in SRA against differential expression between batches in GTEx. As expected, we observed limited concordance in this case.

GTEx use case. We downloaded the processed GTEx v6 gene expression data from http://www.gtexportal.org ("GTEx Analysis v6 RNA-seq RNA-SeQCv1.1.8 gene reads.gct.gz") and compared the provided gene counts to those generated by *recount2* as described above. This involved creating a consensus dataset of genes and samples because the GTEx-processed data were quantified using GENCODE v19[38] (gencode.v19.genes.patched contigs.gtf) and the *recount2* data were quantified using Gencode v25. We restricted our consideration to the 8,551 RNA-seq samples for which gene counts were available from http://www.gtexportal.org. Next, we matched genes between datasets using the Gencode gene IDs and GTEx Ensembl Gene IDs, which resulted in a common set of 51,491 genes in both the GTEx- and *recount2*-processed data (18,998 protein-coding genes). We then computed Pearson correlations for each gene between the two processing approaches. Finally, we performed a differential expression analysis comparing colon and whole blood samples using limma and voom [17] in both processed datasets, adjusting for RNA integrity number (RIN).

Multi-level use case with breast cancer data. Principal component analysis (PCA) was run to identify any sample outliers. All samples were retained for downstream analyses. The 26,742 genes with an average normalized count greater than 5 across samples were included for downstream analysis. Similarly, 259,626 exons and 45,933 exon-exon junctions with reads overlapping these genes were also included for differential expression (DE) analysis. We further included analysis of data at the level of expressed regions to highlight the utility of summarizing expression in an annotation-agnostic manner. At the expressed-region level, 130,518 regions with an average normalized read count greater than 5 across samples were included for DE analysis using derfinder[23]. DE between the two cancer subtypes was carried out using limma and voom[17]. For multiple comparison correction, we calculated q-values from the observed p-values after estimating the proportion of differentially expressed genes, exons, or exon-exon junctions in the experiment [44]. Features

with calculated q-values smaller than 0.05 between different tissue types were declared statistically significant. To compare the results across these levels of data, we used Simes' rule[45] to calculate a gene-based p-value for all exons, exon-exon junctions, and differentially expressed regions that overlap genes. We then computed rank-based concordance between the gene and either the exon, junction, or expressed region level results. In the replication dataset, count data were filtered as before and PCA was utilized to identify global sample outliers. One TNBC tumor sample was identified as an outlier and removed from analysis. IHW[30] uses a set of independent weights for each gene in one study to weight the hypothesis tests in the second study to increase power to detect differences. Here, for each gene the absolute value of the test statistic from study SRP032789 were used as weights for DE analysis in study SRP019936. This resembles using an empirical prior and treating the second study as a validation of the first.

GitHub repository versions. The GitHub repositories that contain code used in this manuscript can be viewed at the precise version that was used at the time of publication. The full code for these repositories can be downloaded via GitHub or explored interactively via the following links:

Supplementary Code 1: https://github.com/leekgroup/recount-analyses/tree/074aa9a03228a4c52ecb765230911741d1ce7f02

Supplementary Code 2: https://github.com/leekgroup/recount-website/tree/39b3c0befd2021d0b99be9d46a16b05d30ad9ce5

Supplementary Code 3: https://github.com/nellore/runs/tree/09131db68125db26c0150ce59065f7366f40ef84

Supplementary Code 4: https://github.com/leekgroup/recount-contributions/tree/8deaa718432afc80d6a1543b4a0ac1d8bd8fc66a

Supplementary References

17 Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* **15**, R29, doi:10.1186/gb-2014-15-2-r29 (2014).

21 Nellore, A. *et al.* Rail-RNA: scalable analysis of RNA-seq splicing and coverage. *Bioinformatics*, doi:10.1093/bioinformatics/btw575 (2016).

22 Nellore, A., Wilks, C., Hansen, K. D., Leek, J. T. & Langmead, B. Rail-dbGaP: analyzing dbGaP-protected data in the cloud with Amazon Elastic MapReduce. *Bioinformatics* **32**, 2551-2553, doi:10.1093/bioinformatics/btw177 (2016).

23 Collado-Torres, L. *et al.* Flexible expressed region analysis for RNA-seq with derfinder. *Nucleic Acids Res* **45**, e9, doi:10.1093/nar/gkw852 (2017).

24 Consortium, G. & others. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348**, 648-660 (2015).

25 Kim, S. K. *et al.* A nineteen gene-based risk score classifier predicts prognosis of colorectal cancer patients. *Mol Oncol* **8**, 1653-1666, doi:10.1016/j.molonc.2014.06.016 (2014).

26 Haberman, Y. *et al.* Pediatric Crohn disease patients exhibit specific ileal transcriptome and microbiome signature. *J Clin Invest* **124**, 3617-3633, doi:10.1172/JCI75436 (2014).

30 Ignatiadis, N., Klaus, B., Zaugg, J. B. & Huber, W. Data-driven hypothesis weighting

increases detection power in genome-scale multiple testing. *Nat Methods* **13**, 577-580, doi:10.1038/nmeth.3885 (2016).
37 Kent, W. J., Zweig, A. S., Barber, G., Hinrichs, A. S. & Karolchik, D. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics* **26**, 2204-2207, doi:10.1093/bioinformatics/btq351 (2010).
38 Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22**, 1760-1774, doi:10.1101/gr.135350.111 (2012).
39 *Cancer Genomics Cloud*.
40 Pohl, A. & Beato, M. bwtool: a tool for bigWig files. *Bioinformatics* **30**, 1618-1619, doi:10.1093/bioinformatics/btu056 (2014).
41 SummarizedExperiment: SummarizedExperiment container (2016).
42 Colaprico, A. *et al.* TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res* **44**, e71, doi:10.1093/nar/gkv1507 (2016).
43 Irizarry, R. A. *et al.* Multiple-laboratory comparison of microarray platforms. *Nat Methods* **2**, 345-350, doi:10.1038/nmeth756 (2005).
44 Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 9440-9445, doi:10.1073/pnas.1530509100 (2003).
45 Simes, R. J. An Improved Bonferroni Procedure for Multiple Tests of Significance. *Biometrika* **73**, 751-754, doi:Doi 10.2307/2336545 (1986).