RSEM

RNA-SEQ BY EXPECTATION–MAXIMIZATION

RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome

Bo Li and Colin N Dewey ✉

Abstract

Background

RNA-Seq is revolutionizing the way transcript abundances are measured. A key challenge in transcript quantification from RNA-Seq data is the handling of reads that map to multiple genes or isoforms. This issue is particularly important for quantification with de novo transcriptome assemblies in the absence of sequenced genomes, as it is difficult to determine which transcripts are isoforms of the same gene. A second significant issue is the design of RNA-Seq experiments, in terms of the number of reads, read length, and whether reads come from one or both ends of cDNA fragments.

PROCEDURE

▸ First, align a set of RNAseq reads to the reference transcripts (using Bowtie).

▸ The resulting alignments are used to estimate abundances and their credibility intervals.

▸ Done with a Trinity wrapper script:

> ▸ `align_and_estimate_abundance.pl`
>
> ▸ must be done separately for each replicate/treatment!

RSEM
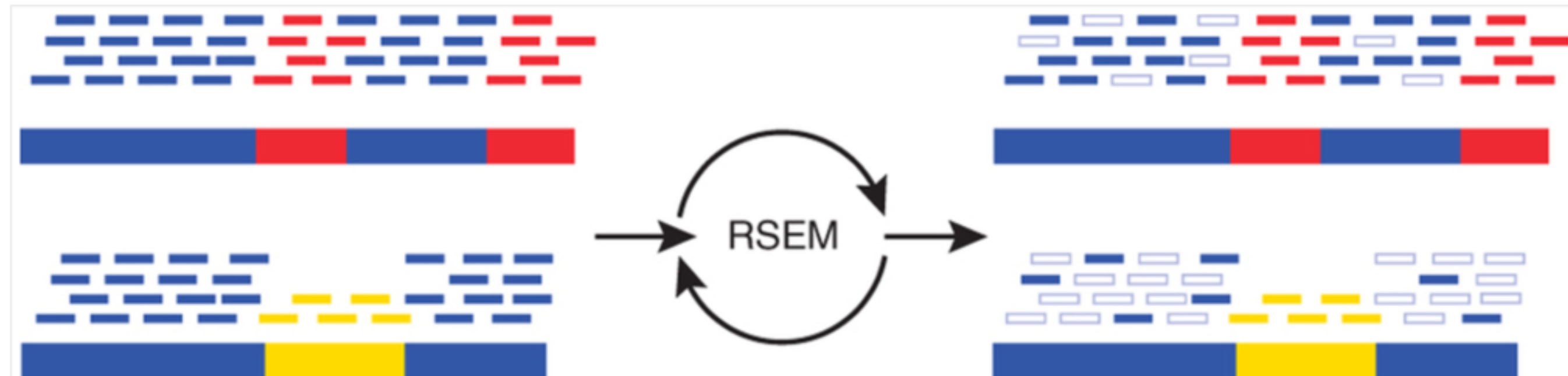
An illustrative example of abundance estimation for two transcripts with shared (blue) and unique (red, yellow) sequences. To estimate transcript abundances, RNA-seq reads (short bars) are first aligned to the transcript sequences (long bars, bottom). Unique regions of isoforms will capture uniquely mapping RNA-seq reads (red and yellow short bars), and shared sequences between isoforms will capture multiply-mapping reads (blue short bars). An expectation maximization algorithm, implemented in the RSEM software, estimates the most likely relative abundances of the transcripts and then fractionally assigns reads to the isoforms based on these abundances. The assignments of reads to isoforms resulting from iterations of expectation maximization are illustrated as filled short bars (right), and eliminated assignments are shown as hollow bars. Note that assignments of multiply-mapped reads are in fact performed fractionally according to a maximum likelihood estimate. Thus, in this example, a higher fraction of each read is assigned to the more highly expressed top isoform than to the bottom isoform.
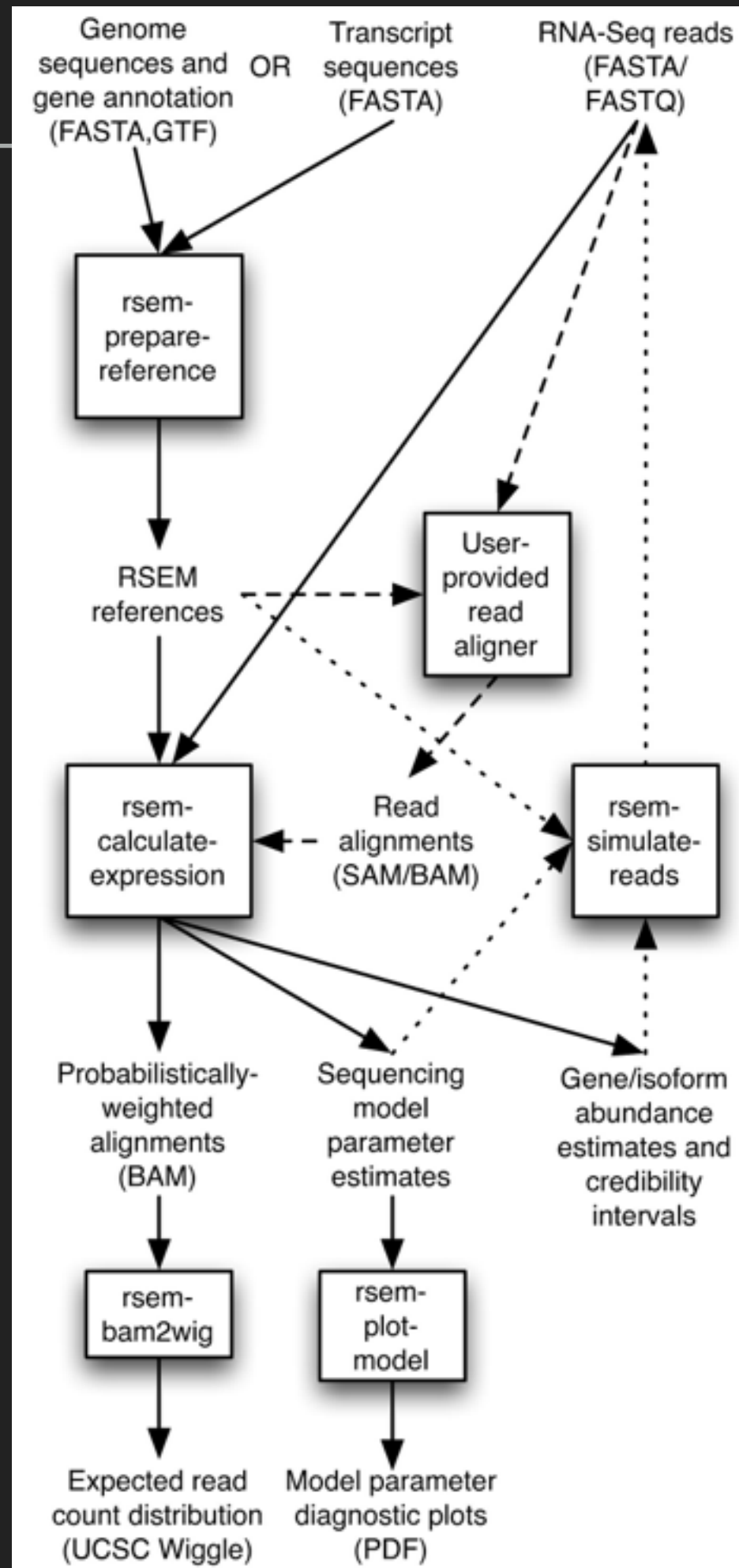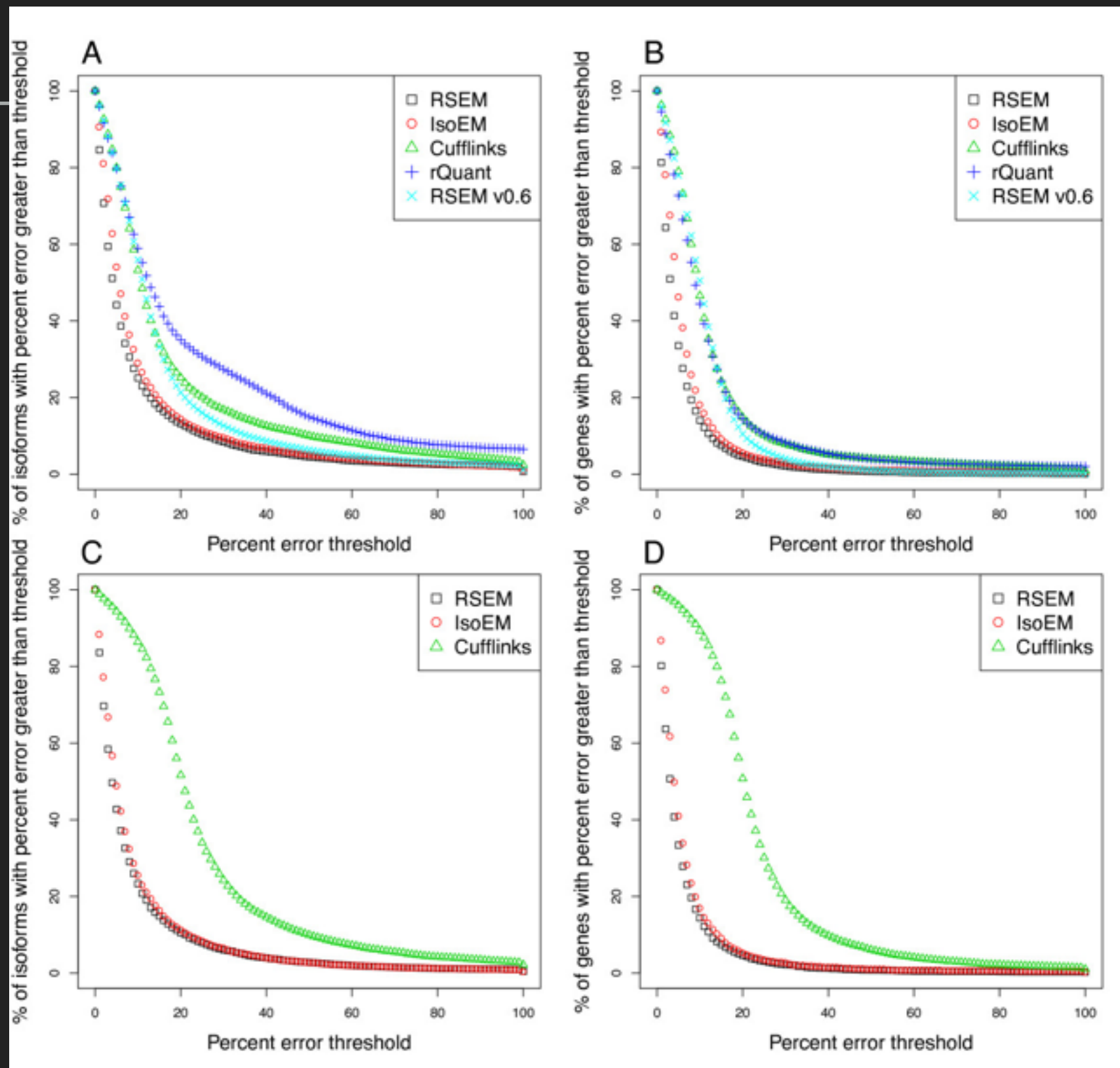
▸ Accuracy of four RNA-Seq quantification methods. The percent error distributions of estimates from RSEM, IsoEM, Cufflinks, and rQuant on simulated RNA-Seq data. The error distributions of global isoform and gene estimates from PE data are shown in (A) and (B), respectively. Global isoform and gene estimate error distributions for SE data are shown in (C) and (D), respectively.

OUTPUT

▸ The primary output generated by RSEM is the file containing the expression values for each of the transcripts.

gene_id	transcript_id(s)	length	effective_length	expected_count	TPM	FPKM
TRINITY_DN100_c0_g1	TRINITY_DN100_c0_g1_i1	253.00	238.93	0.00	0.00	0.00
TRINITY_DN103_c0_g1	TRINITY_DN103_c0_g1_i1	524.00	509.93	0.00	0.00	0.00
TRINITY_DN103_c1_g1	TRINITY_DN103_c1_g1_i1	152.00	138.05	0.00	0.00	0.00
TRINITY_DN104_c0_g1	TRINITY_DN104_c0_g1_i1	174.00	160.01	0.00	0.00	0.00
TRINITY_DN105_c0_g1	TRINITY_DN105_c0_g1_i1	221.00	206.95	0.00	0.00	0.00
TRINITY_DN105_c1_g1	TRINITY_DN105_c1_g1_i1	238.00	223.93	1.00	1139.60	2872.63
TRINITY_DN107_c0_g1	TRINITY_DN107_c0_g1_i1	161.00	147.03	0.00	0.00	0.00
TRINITY_DN108_c0_g1	TRINITY_DN108_c0_g1_i1	190.00	175.99	0.00	0.00	0.00
TRINITY_DN108_c1_g1	TRINITY_DN108_c1_g1_i1	195.00	180.98	0.00	0.00	0.00

OUTPUT

```
transcript_id          gene_id            length  effective_length  expected_count  TPM      FPKM      IsoPct
TRINITY_DN0_c0_g1_i1    TRINITY_DN0_c0_g1  328     198.75            29.00           9093.16  43883.19  100.00
TRINITY_DN0_c0_g2_i1    TRINITY_DN0_c0_g2  329     199.75            0.00            0.10     0.48      100.00
```

▸ expected_count: number of RNA-Seq fragments predicted to be derived from that transcript

▸ FPKM: fragments per kilobase of cDNA per million fragments mapped

▸ TPM: transcripts per million

▸ The next step is to take expression estimates and normalize across samples.

A scaling normalization method for differential expression analysis of RNA-seq data

Mark D Robinson ✉ and Alicia Oshlack ✉

Abstract

The fine detail provided by sequencing-based transcriptome surveys suggests that RNA-seq is likely to become the platform of choice for interrogating steady state RNA. In order to discover biologically important changes in expression, we show that normalization continues to be an essential step in the analysis. We outline a simple and effective method for performing normalization and show dramatically improved results for inferring differential expression in simulated and publicly available data sets.

TMM

▸ The trimmed mean of M-values normalization method: The total RNA production, S_k , cannot be estimated directly, since we do not know the expression levels and true lengths of every gene. However, the relative RNA production of two samples, $f_k = S_k / S_{k'}$, essentially a global fold change, can more easily be determined. We propose an empirical strategy that equates the overall expression levels of genes between samples under the assumption that the majority of them are not DE. One simple yet robust way to estimate the ratio of RNA production uses a weighted trimmed mean of the log expression ratios (trimmed mean of M values (TMM)).
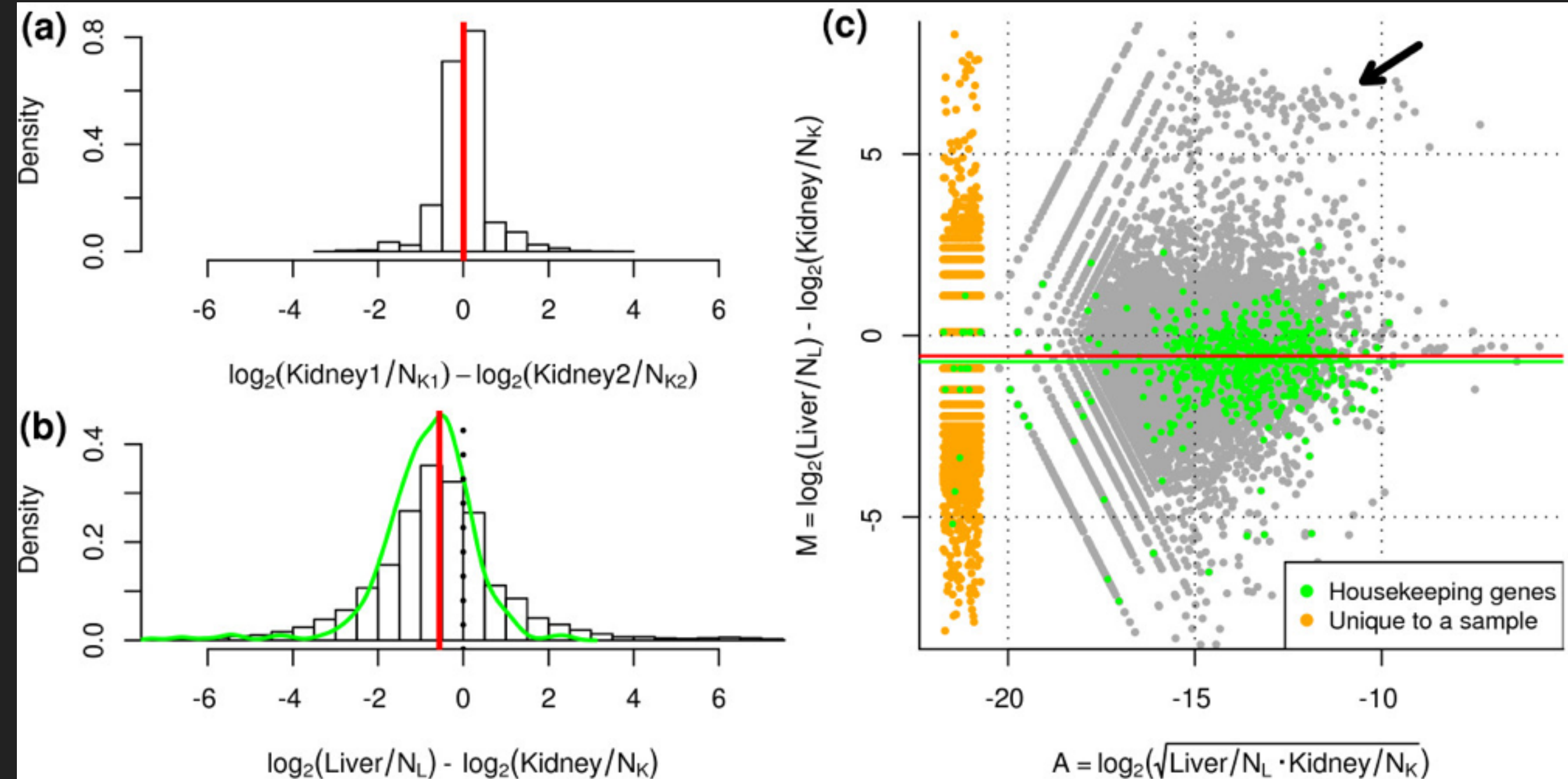
▸ Normalization is required for RNA-seq data. Data from [6] comparing log ratios of (a) technical replicates and (b) liver versus kidney expression levels, after adjusting for the total number of reads in each sample. The green line shows the smoothed distribution of log-fold-changes of the housekeeping genes. (c) An M versus A plot comparing liver and kidney shows a clear offset from zero. Green points indicate 545 housekeeping genes, while the green line signifies the median log-ratio of the housekeeping genes. The red line shows the estimated TMM normalization factor. The smear of orange points highlights the genes that were observed in only one of the liver or kidney tissues. The black arrow highlights the set of prominent genes that are largely attributable for the overall bias in log-fold-changes.

PROCEDURE

▸ abundance_estimates_to_matrix.pl

▸ Trinity_trans.counts.matrix

GSNO_SRR1582648	GSNO_SRR1582646	GSNO_SRR1582647	wt_SRR1582649	wt_SRR1582651	wt_SRR1582650	
TRINITY_DN270_c0_g1_i1	0.00	0.00	0.00	1.00	0.00	1.00
TRINITY_DN286_c0_g1_i1	38.00	36.00	39.00	5.00	7.00	8.00
TRINITY_DN472_c0_g1_i1	2.00	0.00	0.00	0.00	0.00	0.00
TRINITY_DN269_c0_g1_i1	0.00	0.00	0.00	0.00	0.00	0.00
TRINITY_DN378_c0_g1_i1	0.00	0.00	0.00	0.00	0.00	0.00
TRINITY_DN258_c0_g1_i1	0.00	0.00	0.00	0.00	0.00	0.00
TRINITY_DN392_c0_g1_i1	0.00	0.00	2.00	0.00	0.00	0.00
TRINITY_DN407_c0_g1_i1	0.00	0.00	0.00	0.00	0.00	0.00
TRINITY_DN328_c0_g1_i1	0.00	0.00	0.00	0.00	0.00	0.00
TRINITY_DN61_c0_g1_i1	0.00	0.00	0.00	0.00	0.00	0.00
TRINITY_DN221_c0_g1_i1	0.00	0.00	0.00	0.00	0.00	0.00
TRINITY_DN260_c1_g1_i1	3.00	4.00	7.00	8.00	8.00	5.00
TRINITY_DN357_c0_g1_i1	0.00	0.00	0.00	0.00	0.00	0.00
TRINITY_DN63_c1_g1_i1	19.00	26.00	33.00	3.00	4.00	3.00
TRINITY_DN401_c0_g1_i1	2.00	3.00	4.00	3.00	6.00	2.00
TRINITY_DN637_c0_g1_i1	0.00	0.00	0.00	1.00	2.00	1.00
TRINITY_DN288_c0_g1_i1	0.00	0.00	1.00	0.00	0.00	0.00
TRINITY_DN625_c0_g1_i1	0.00	0.00	1.00	2.00	3.00	1.00
TRINITY_DN97_c0_g1_i1	0.00	0.00	0.00	0.00	0.00	0.00

TMM NORMALIZED COUNTS

	GSNO_SRR1582648	GSNO_SRR1582646	GSNO_SRR1582647	wt_SRR1582649	wt_SRR1582651	wt_SRR1582650
TRINITY_DN270_c0_g1_i1	0.000	0.000	0.000	1382.140	0.000	1599.862
TRINITY_DN286_c0_g1_i1	19831.723	15748.460	16547.800	2036.568	3442.755	3763.621
TRINITY_DN472_c0_g1_i1	3139.756	0.000	0.000	0.000	0.000	0.000
TRINITY_DN269_c0_g1_i1	0.000	0.000	0.000	0.000	0.000	0.000
TRINITY_DN378_c0_g1_i1	0.000	0.000	0.000	0.000	0.000	0.000
TRINITY_DN258_c0_g1_i1	0.000	0.000	0.000	0.000	0.000	0.000
TRINITY_DN392_c0_g1_i1	0.000	0.000	2511.305	0.000	0.000	0.000
TRINITY_DN407_c0_g1_i1	0.000	0.000	0.000	0.000	0.000	0.000
TRINITY_DN328_c0_g1_i1	0.000	0.000	0.000	0.000	0.000	0.000
TRINITY_DN61_c0_g1_i1	0.000	0.000	0.000	0.000	0.000	0.000
TRINITY_DN221_c0_g1_i1	0.000	0.000	0.000	0.000	0.000	0.000
TRINITY_DN260_c1_g1_i1	1701.427	1902.282	3228.057	3540.731	4288.504	2556.256
TRINITY_DN357_c0_g1_i1	0.000	0.000	0.000	0.000	0.000	0.000
TRINITY_DN63_c1_g1_i1	11708.546	13440.787	16537.510	1442.567	2337.418	1666.528
TRINITY_DN401_c0_g1_i1	1288.233	1621.389	2095.379	1507.745	3671.449	1161.284
TRINITY_DN637_c0_g1_i1	0.000	0.000	0.000	1155.640	2970.275	1337.249
TRINITY_DN288_c0_g1_i1	0.000	0.000	243.398	0.000	0.000	0.000
TRINITY_DN625_c0_g1_i1	0.000	0.000	1048.765	2006.101	3821.984	1160.229
TRINITY_DN97_c0_g1_i1	0.000	0.000	0.000	0.000	0.000	0.000

EXN50

▸ ExN50: Calculation of N50 that is limited to the top most highly expressed transcripts that represent 50% of the total normalized expression data.

HANDS-ON

▸ Go back to http://github.com/SmithsonianWorkshops/ RNA-seq and follow the "5a_Transcript quantification.md" tutorial

POST-COURSE SURVEY

▸ Please fill out the pre-course survey:

http://bit.ly/2kRTGN4