Nature Methods

Computational methods for transcriptome annotation and quantification using RNA-seq

Manuel Garber, Manfred G Grabherr, Mitchell Guttman & Cole Trapnell

Supplementary figure 1	Comparison between seed and Burrows-Wheeler alignments
Supplementary figure 2	Comparison between seed and Burrows-Wheeler alignments when aligning to a reference genome of a different species
Supplementary figure 3	Fraction of fully reconstructed transcript per expression quantile by three transcriptome reconstruction methods
Suuplementary figure 4	
Supplementary figure 5	
Supplementary table 1	Comparison between methods in different categories for read mapping and transcriptome reconstruction

Nature Methods: doi.10.1038/nmeth.1613

	Category	Mouse transcriptome alignment			Rat transcriptome alignment	
		CPU Hours	Memory[1]	Aligned paired reads	CPU Hours	Aligned paired reads
Stampy	Unspliced seed aligner	110	67 Mb	126,466,017	110	124,542,236
BWA	Unspliced B-W aligner	8	500 Mb	108,073,744	18[2]	83,263,812

	Category	Mouse genome alignment			
		CPU Hours	Memory	Spliced reads	Unique spliced reads
GSNAP	Seed spliced aligner	340	5.5 G	18,502,068[3]	15,436,727
TopHat	Exon First spliced aligner	44	11 G	12,468,695[3]	10,420,126

	Category	Reconstruction of the mouse ES transcriptome					
		CPU Hours	Total Memory	Genes fully reconstructed	Mean Number of isoforms per reconstruction	Mean fragments per known annotation	Number of fragments predicted
Cufflinks	Genome guided reconstruction method	10	1.4 G	5,994	1.2	1.4	159,856
Scripture	Genome guided reconstruction method	16	3.5 G	6,221	1.6	1.3	61,922
Trans-Abyss	Genome independent reconstruction method	650	120 G[4]	3,330	4.7	2.6	3,117,238

1. Similar memory usage was required for the alignment to the rat transcriptome.
2. Raised gap extend and minimum mismatch parameters to increase sensitivity
3. Alignment accuracy was not evaluated as part of this test, so the biological relevance of spliced sites discovered by TopHat and GSNAP could not be verified. This table is intended to illustrate general computational resource requirements of spliced aligners only.
4. It is possible to subdivide an Abyss reconstruction process into smaller units which can be run each independently using as little as 16 gigabytes of RAM.

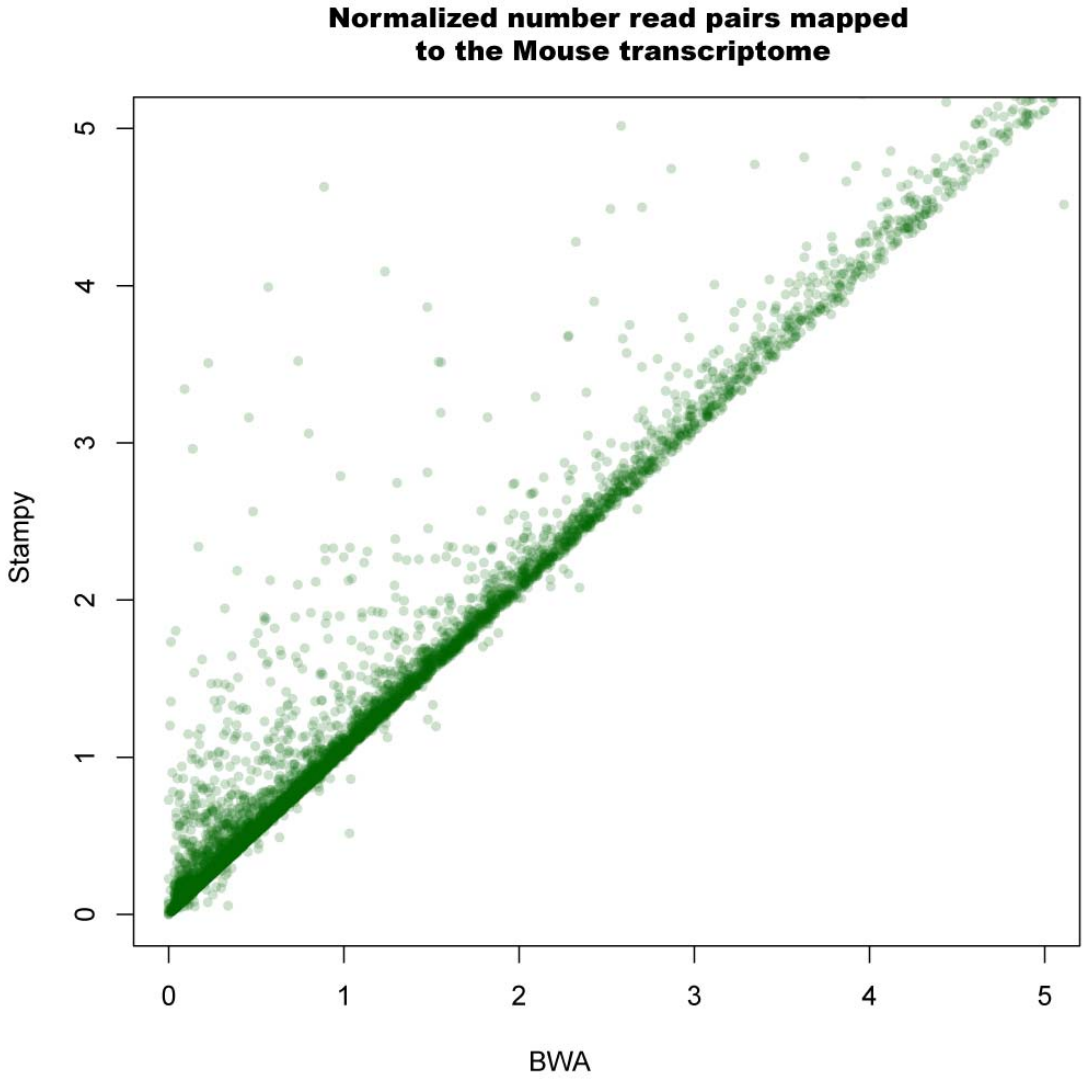
Supplementary table 1. Comparison between methods in different categories for read mapping and transcriptome reconstruction.

The table shows three comparisons showing results of aligning and reconstruction transcripts using a previously published mouse embryonic stem cell dataset consisting of 58 million paired-end reads[1].

Top: comparison of unspliced Seed (Stampy) and Burrows-wheeler (BWA) aligners for mapping reads to both the mouse and rat transcriptome consisting of 8,557 genes expressed in mES that have a rat ortholog.

Middle: comparison of compute resources required by Seed-Extend (GSNAP) and Exon first (Tophat) aligners when mapping the same number of reads against the same reference genome.

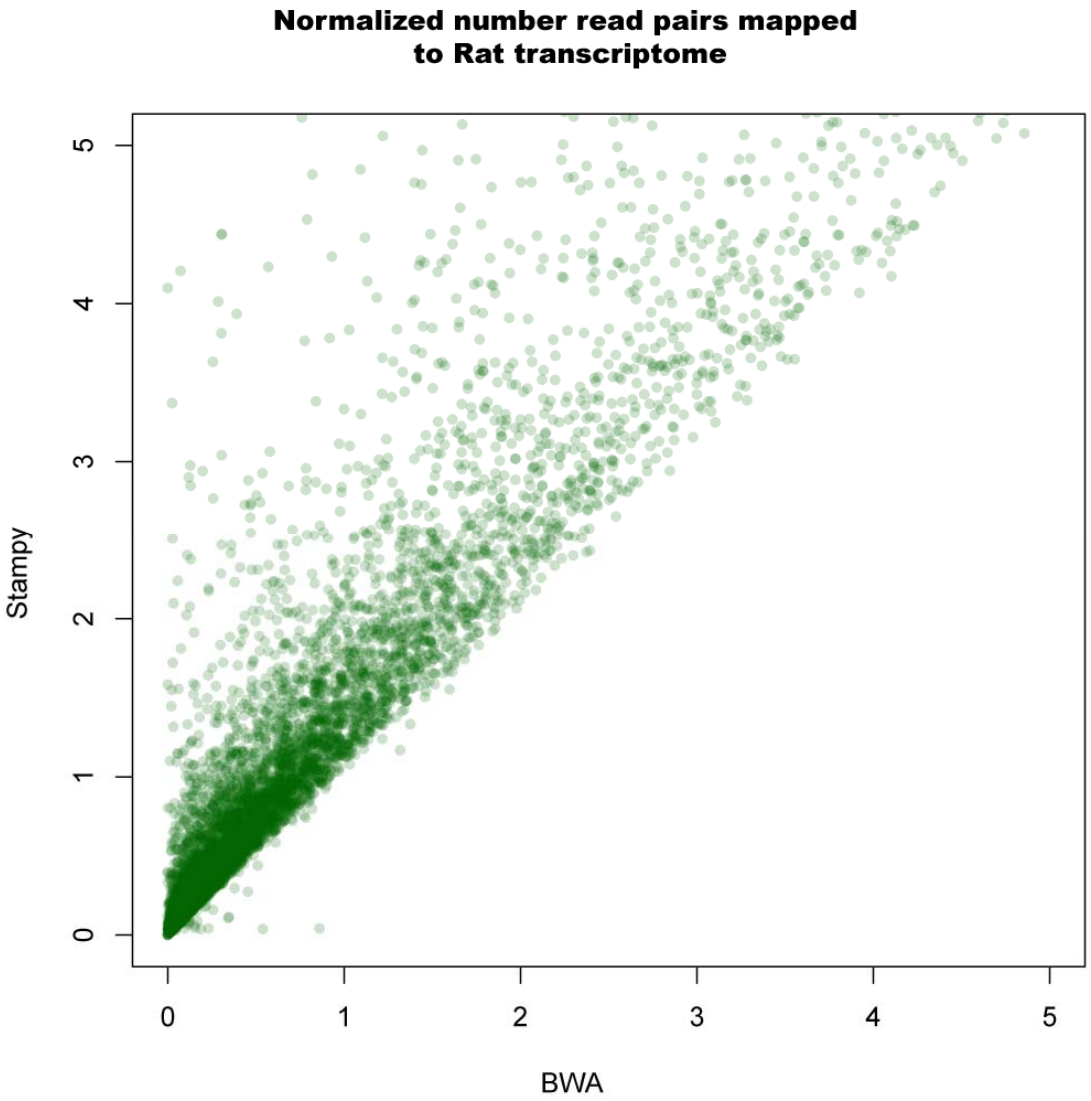
Bottom: comparison of three transcriptome reconstruction methods run with default parameters. Cufflinks and Scripture where applied to the TopHat alignments. A gene is called fully reconstructed when there exists a transcript overlapping the know 5' and 3' exons with the known internal structure. Fragments per annotation were computed by first taking the union model for each reconstruction, following by counting the number of union models within Refseq annotations



Supplementary figure 1. Comparison between seed and Burrows-Wheeler alignments.

RNA-Seq reads were aligned to the mouse sequence of 8,557 genes with an orthologous annotation in rat. Each point in the plot shows the length normalized count of paired end reads aligned to a gene using BWA (x-axis) and Stampy (y-axis).

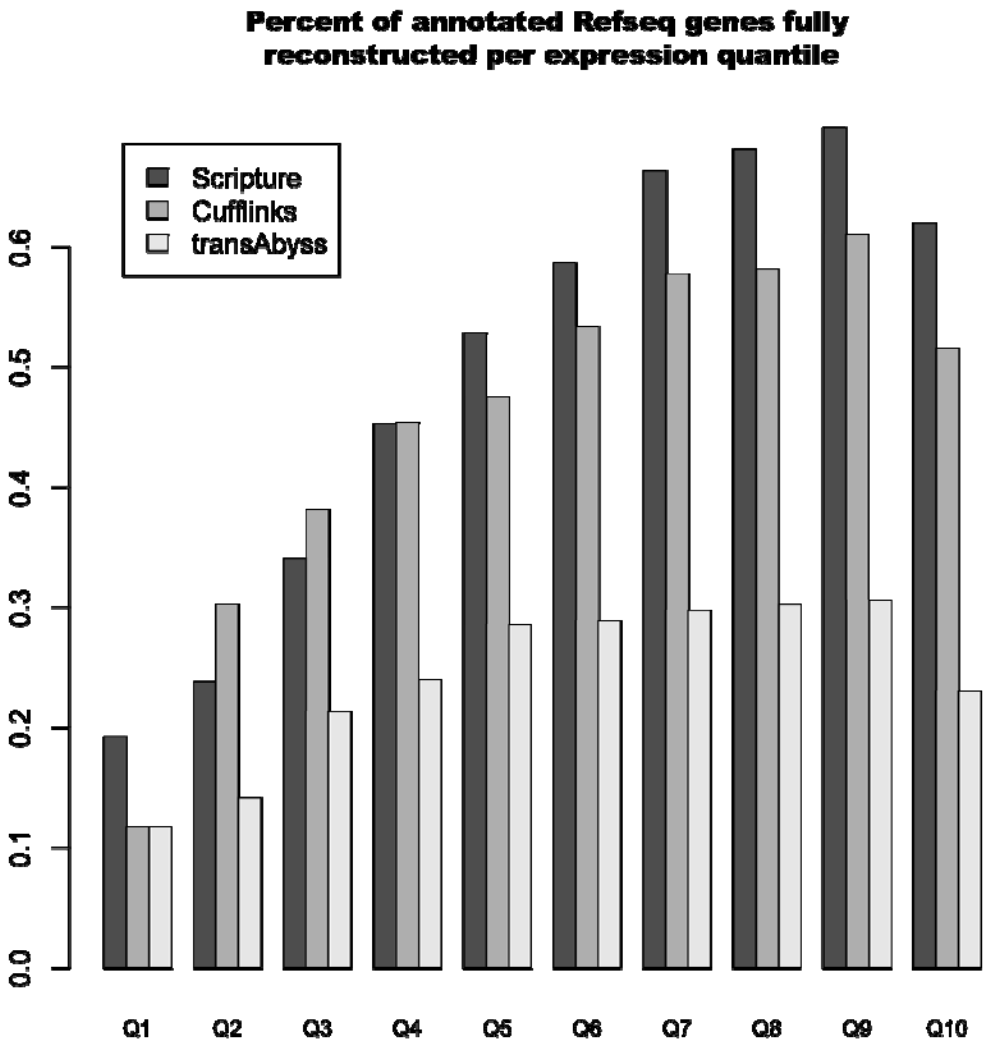
Nature Methods: doi.10.1038/nmeth.1613



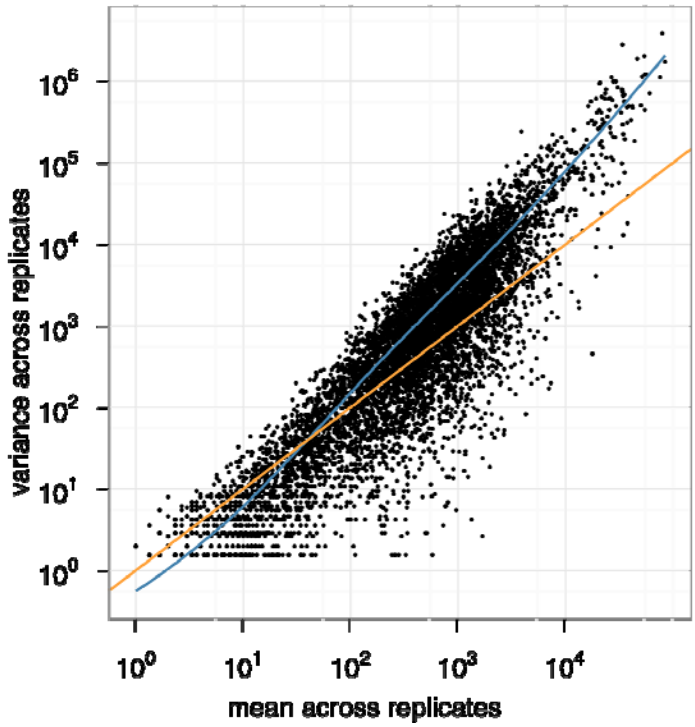
Supplementary figure 2. Comparison between seed and Burrows-Wheeler alignments when aligning to a reference genome of a different species

RNA-Seq reads were aligned to the rat sequence of 8,557 genes with an orthologous annotation in mouse. Each point in the plot shows the length normalized count of paired end reads aligned to a gene using BWA (x-axis) and Stampy (y-axis).

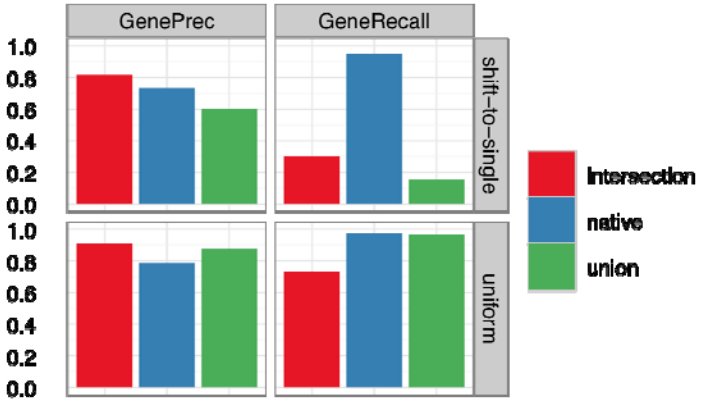
Nature Methods: doi.10.1038/nmeth.1613



Supplementary figure 3. Fraction of fully reconstructed transcript per expression quantile by three transcriptome reconstruction methods. Cufflinks and Scripture where applied to the TopHat alignments. A gene is called fully reconstructed when there exists a transcript overlapping the know 5' and 3' exons with the known internal structure. Fragments per annotation were computed by first taking the union model for each reconstruction, following by counting the number of union models within Refseq annotations.



Supplementary figure 4. Simulation of RNA-Seq. 50bp paired end reads were generated from FlyBase transcripts according to an expression profile calculated from real RNA-Seq data from S2 cells produced by modENCODE. Variability in the fragment count for each gene was modeled by a negative binomial distribution parameterized by a gamma function (blue line). The gamma function was fit through the observed variability in the real S2 data across the dynamic range of expression. The data is clearly overdispersed with respect to the Poisson distribution (orange line).



Supplementary figure 5. Performance of quantification methods in simulated data. Top squares show the precision (left) and recall (right) in detecting expressed genes in the presence of isoform switching via differential splicing. Bottom squares show similar accuracy when all isoforms of perturbed genes were increased uniformly by 2-fold. In each scenario, three read counting schemes were evaluated. The “native” transcript

expression method involves counting reads, inferring individual transcript abundances, and summing these to calculate overall gene expressed in each condition before testing for significant differences. Two computationally simpler schemes approximate gene expression by counting reads in constitutive exons only ("exon intersection") or in all exons ("exon union") and then normalizing counts by exonic length to calculate gene expression. Accuracy of the three schemes is similar in the uniform perturbation (i.e. when there is no isoform-level switching). This is expected, because a change in gene expression is reflected by an equal change in the count obtained by any of the three methods. However, in the presence of isoform switching, a change in overall gene expression is not necessarily reflected in the raw number of reads originating from the gene. To arrive at the correct expression value for the gene, Cuffdiff must infer individual isoform isoform abundances.

Supplementary Methods

Simulated data for differential analysis.

We simulated two hypothetical sequencing experiments in which 100 randomly selected multi-isoform genes were differentially expressed between two conditions. We used a custom simulator (similar to methods described previously[2]) to generate 20 million 50b paired-end reads of the *Drosophila melanogaster* transcriptome (FlyBase v5.12) for two replicates of each condition. We ran Cufflinks 1.0.0 on reads from a modENCODE RNA-Seq library from S2 cells (SRX003834) to calculate the S2 unperturbed expression profile. We used the unperturbed profile to generate the reads for the control condition in each experiment. *In silico* library size and fragment length distribution were similar to SR003834. Variability in fragment counts from each gene across replicates was modeled by the negative binomial distribution, with parameters chosen from the empirical variance model fitted by Cuffdiff on the real S2 data (Supplementary Figure 3). This model has been proposed to accommodate the overdispersion in fragment counts observed across biological replicates[3,4]. We perturbed the S2 profile in two ways, to benchmark expression call accuracy in presence and absence of alternative splicing. In the first experiment we perturbed the expression of 100 genes by uniformly increasing all of their alternative isoforms' abundance by two-fold, and then simulated sequencing both the perturbed and control condition. In the second simulation, we perturbed gene expression by redistributing all reads from all isoforms to originate from the shortest isoform. This perturbation maximizes the impact on overall expression due to redistribution of counts *without* altering the total number of reads that originate from the gene. In selecting genes for perturbation, we required that the overall increase in gene expression from this perturbation would be at least two fold.

We then called differentially expressed genes with Cuffdiff using the three read count methods by applying cuffdiff on 1) the RefSeq annotated transcripts (Isoform expression method), 2) constituent exons (intersection method) and 3) on pseudotranscripts resulting from merging all exons within a gene (union method). Supplementary figure 4 shows the results of comparing each of these three quantification methods' performance under the two simulations. Cuffdiff was run with FPKM upper-quartile normalization enabled and

an FDR threshold of 5%. To avoid confounding the benchmarking analysis with read mapping issues, we used a perfect mapping of reads to the genome.

1. Guttman, M. et al. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol* **28**, 503-10 (2010).
2. Trapnell, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**, 511-5 (2010).
3. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol* **11**, R106 (2010).
4. Robinson, M.D., McCarthy, D.J. & Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-40 (2009).