Supplementary Information for Grabherr et al., 2011

Nature Biotechnology: doi:10.1038/nbt.1883

SUPPLEMENTARY NOTE

Assembly of the fission yeast transcriptome

Inchworm assembled the yeast data set into 811,364 contigs with length at least 48 bases (2*(k-1), k=25, see above). Only 8,234 of the contigs are at least 350 bases long (approximately the mean insert size in our RNA-Seq library) and those comprise 13.4 Mb of total sequence. At this stage, 15% (660 of 4265) of the Inchworm-reconstructed, Oracle-matching, transcripts were recovered as falsely fused into single contigs. These mostly correspond to adjacent genes that overlap in their untranslated regions (UTRs), a common phenomenon in yeasts[1, 2]. By examining the clustering of read mate-pairings, 375 of the 660 falsely-fused transcripts were automatically teased apart into individual full-length transcripts (see above). Chrysalis grouped all contigs into 23,607 components and built a set of de Bruijn graphs, with a total of 24M unique k-mer nodes. After filtering and analyzing the graph, Butterfly outputs 27,841 linear contigs longer than 100 bases, grouped into a final set of 23,232 components.

Assembly of the mouse transcriptome

First, Inchworm assembled the reads into ~1.9M contigs (43 Mb resides in 32,466 sequences >= 350 bp), containing 7,346 annotated full-length transcripts. **Second**, Chrysalis pooled the contigs into 156,211 components. **Finally**, Butterfly reported 179,340 contigs (48,497 of length greater than 350bp), residing in 151,115 remaining components, fully capturing the 8,185 transcripts at 7,749 loci at full length.

SUPPLEMENTARY METHODS

Yeast strains and growth conditions.

Cultures were grown in the following rich medium: Yeast extract (1.5%), Peptone (1%), Dextrose (2%), SC Amino Acid mix (Sunrise Science) 2 grams per liter, Adenine 100 mg/L, Tryptophan 100 mg/L, Uracil 100 mg/L, at 200 RPM in an New Brunswick Scientific air-shaker.

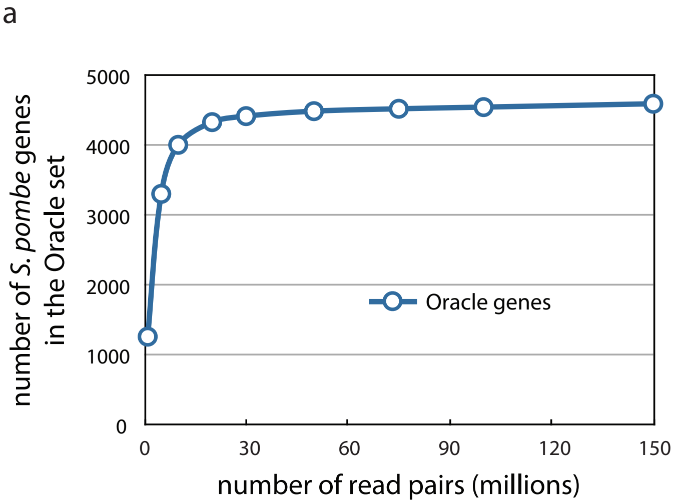
For glucose depletion (mid-log, diauxic shift, and stationary phase samples), overnight cultures were grown to saturation in 3 ml rich medium. From the 3 ml overnight cultures, 300 ml of rich media was inoculated at the OD_{600} corresponding to $1x10^6$ cell/ml and grown in New Brunswick Scientific shaking water baths. Culture density was monitored by OD_{600}. Glucose levels were monitored using the YSI 2700 Select Bioanalyzer. Cells were harvested at mid-log, diauxic shift (defined as the timepoint when glucose is depleted from the medium), and when growth plateaus by quenching them in 60% liquid methanol at -40°C that was later removed by centrifugation at -9°C and stored overnight at -80°C. Harvested cells were later washed in RNAse-free water and archived in RNAlater (Ambion) for future preparations.

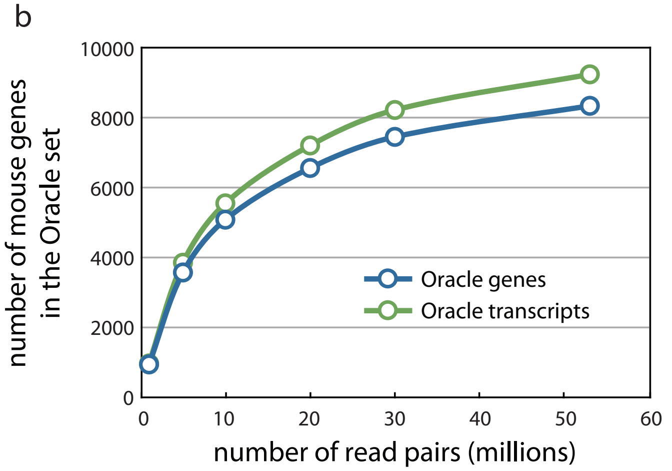
For heat shock, overnight cultures were grown in 650ml of media at 22°C to between $3x10^7$ and $1x10^8$ cell/ml OD_{600} = 1.0. The overnight culture was split into two 300ml cultures and cells from each were collected by removing the media via vacuum filtration (Millipore). The cell-containing filters were re-suspended in pre-warmed media to either control (22°C) or heat-shock temperatures (37°C). Density measurements were taken approximately one minute after cells were re-suspended to ensure that concentrations did not change during the transfer from overnight media. 60ml of culture were harvested at 15 minutes after re-suspension by quenching them in 60% liquid methanol at -40°C that was later removed by centrifugation at -9°C and stored overnight at -80°C. Harvested cells were later washed in RNAse-free water and archived in RNAlater (Ambion) for future preparations. Cells were also harvested from cultures just before treatment for use as controls.

Mouse dendritic cell isolation and tissue culture

6-8 weeks female C57BL/6J mice were obtained from the Jackson Laboratories. Bone Marrow DCs were collected from femora and tibiae and plated on non-tissue culture treated plastic dishes in RPMI medium (Gibco Invitrogen) supplemented with 10% FBS, L-glutamin, penicillin/streptomycin, MEM non-essential amino acids, HEPES, sodium pyruvate, β-mercaptoethanol, and GM-CSF (15 ng/mL; Peprotech). At day 5, floating CD11c+ cells were collected and sorted on MACS columns using the CD11c (N418) MicroBeads kit (Myltenyi Biotec). CD11c+ cells where replated at a concentration of 10^6 cells/ml and collected 12 hours post sorting.

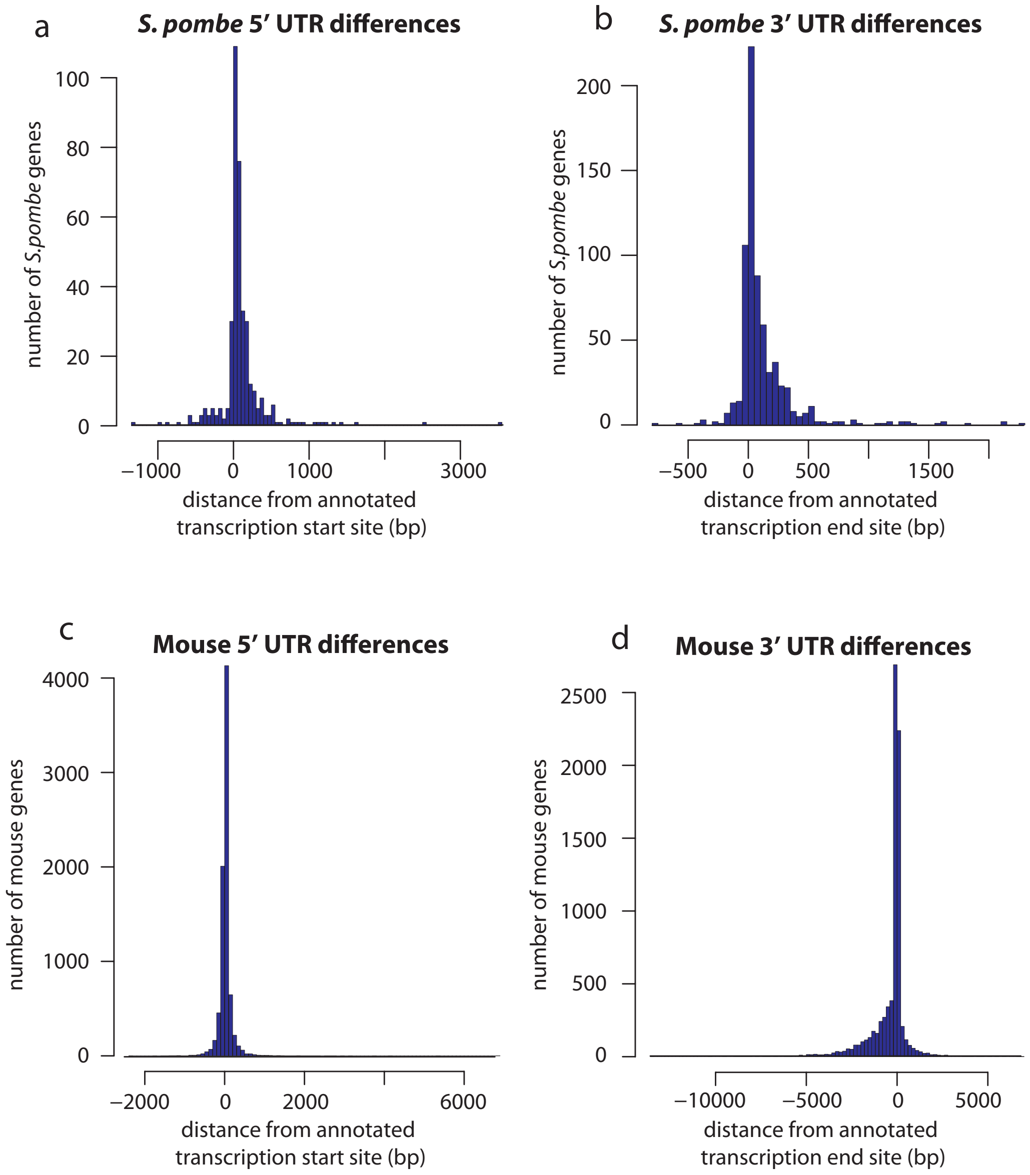
SUPPLEMENTARY FIGURES AND LEGENDS





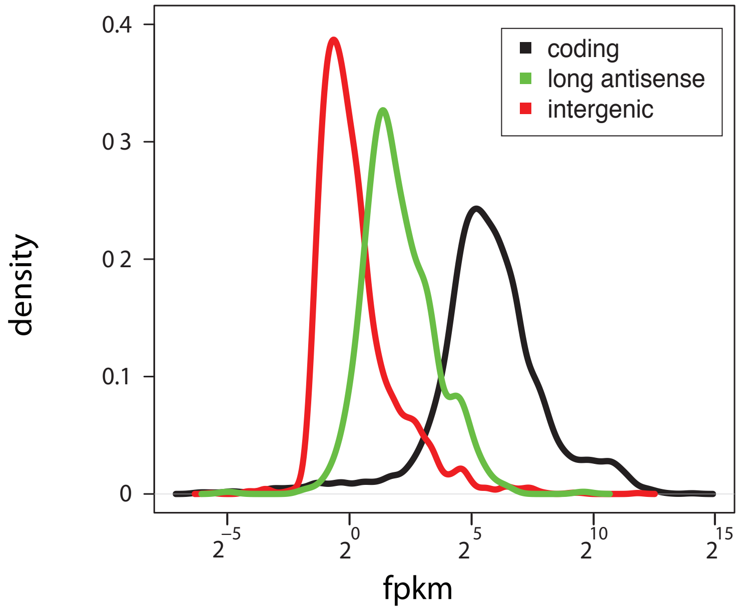
Supplementary Figure 1. Impact of the number of reads on the oracle set.

Shown are the numbers of *S. pombe* genes (**a,** blue) or mouse genes (**b**, blue) or transcripts **(b**, green) that are captured by the Oracle set at different numbers of input read pairs (x axis). The oracle set begins to saturate at 25M read pairs (or 50M reads) for the *S. pombe* RNA-Seq data (**a**), but is likely not saturated with the entire set of 53M read pairs on the mouse data set (**b**).



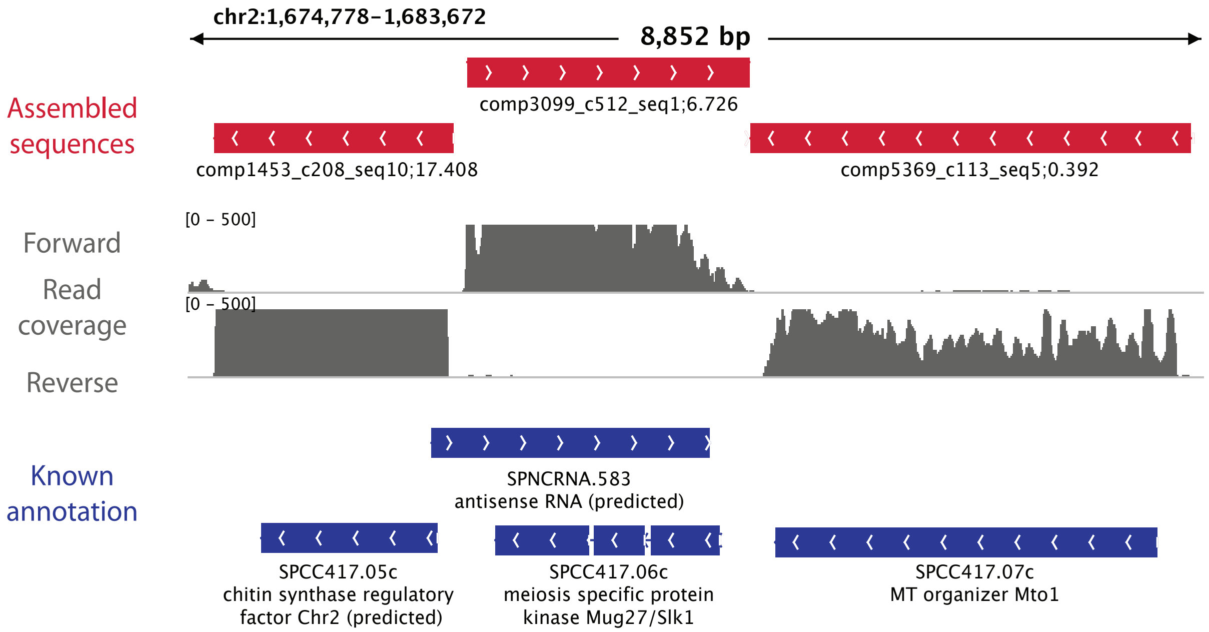
Supplementary Figure 2. UTR differences between Trinity transcripts and the annotated reference.

Shown are the distributions of changes in UTR length between Trinity transcripts and the annotated reference at the 5'UTR (a,c) and 3'UTR (b,d) of *S. pombe* (**a,b**) and mouse (**c,d**).



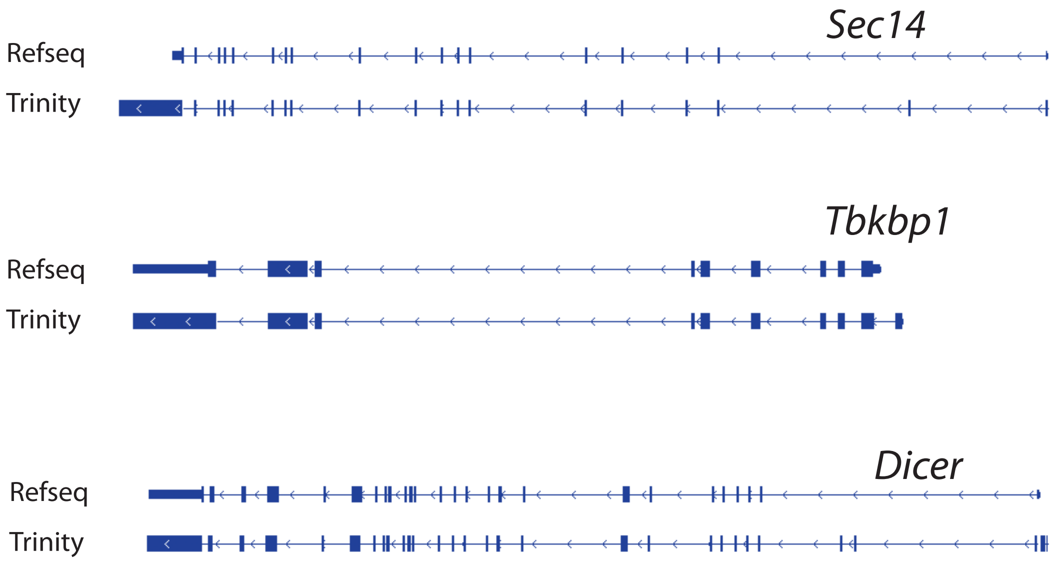
Supplementary Figure 3. Distribution of expression levels for protein-coding and antisense transcripts.

Shown are the distributions of expression levels (FPKM) for coding (blue), long antisense (green), and intergenic (red) Trinity-assembled transcripts in *S. pombe*.



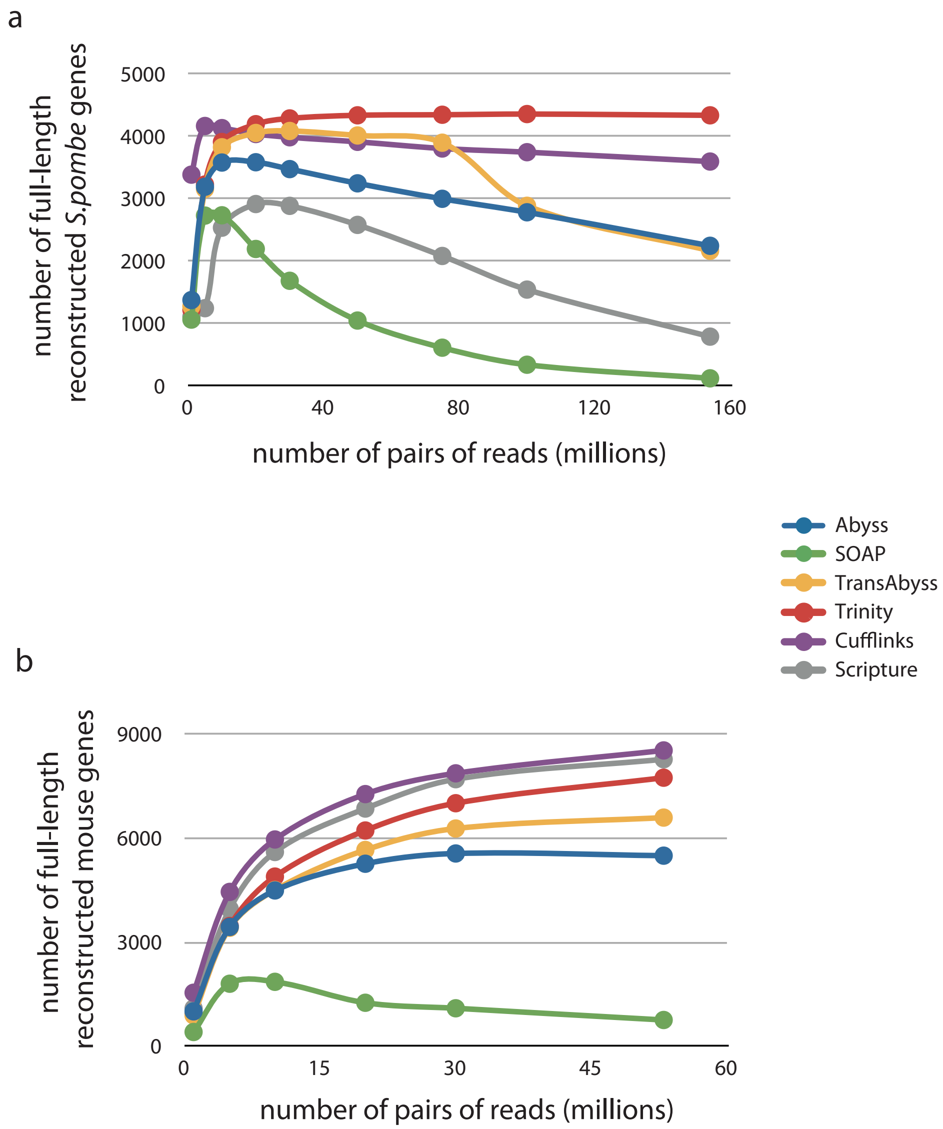
Supplementary Figure 4. Trinity identifies antisense transcription in yeast.

Shown are examples of Trinity assemblies (red) along with the corresponding annotated transcripts (blue) and coverage of underlying reads (green) all aligned to the *S. pombe* genome (for graphical clarity; no alignments were used to generate the assemblies). Trinity's assembly of comp3099 corresponds to the predicted antisense transcript SPNCRNA.583.



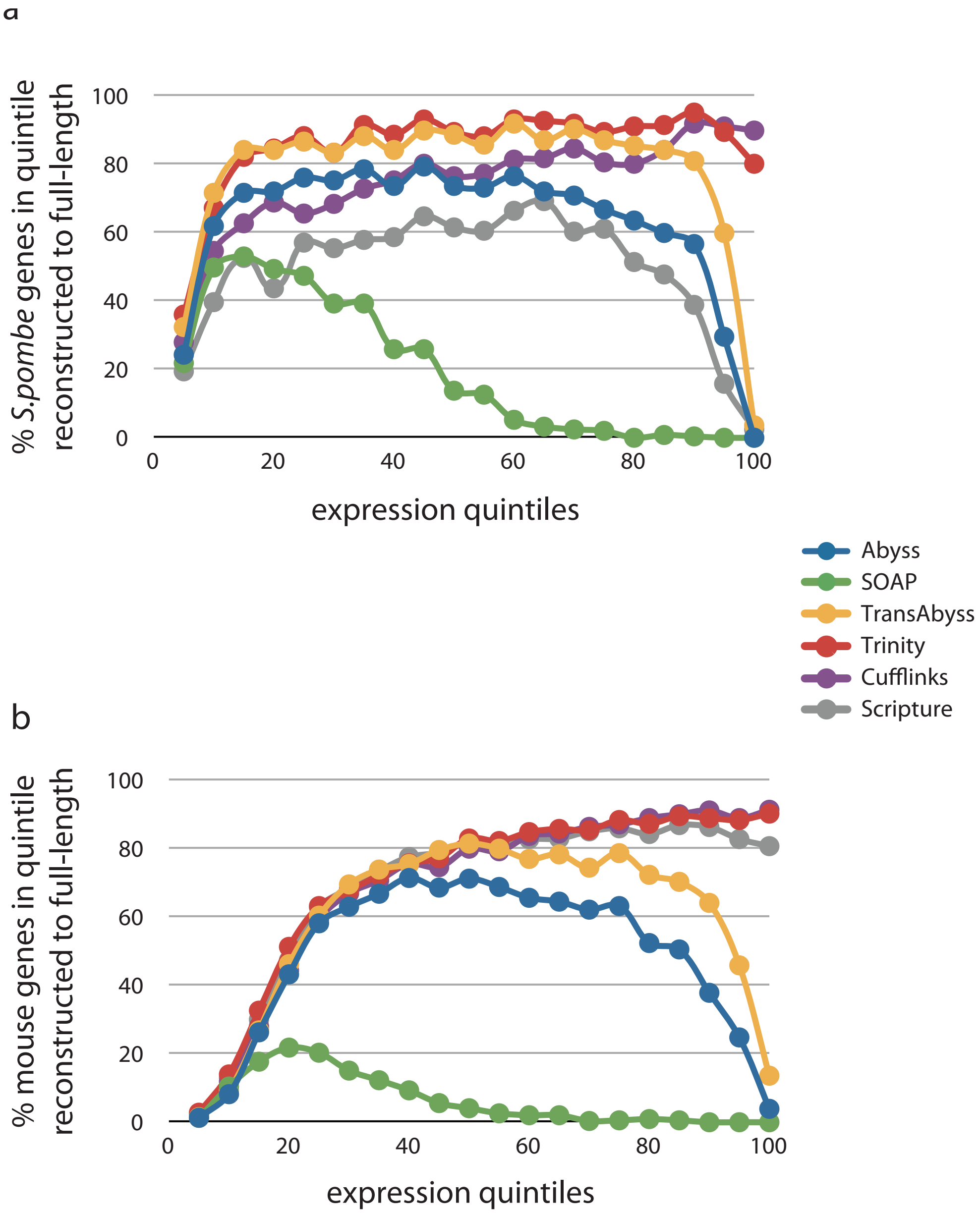
Supplementary Figure 5. Examples for UTR exon additions in mouse.

Shown are examples of Trinity assemblies (bottom) and the corresponding reference annotation (top) for (**a**) *Sec14* (one extra internal UTR exon), (**b**) *Tbkbp1* (one extra UTR exon at the 5' end), and (**c**) *Dicer* (multiple internal and 5' end UTR exons).



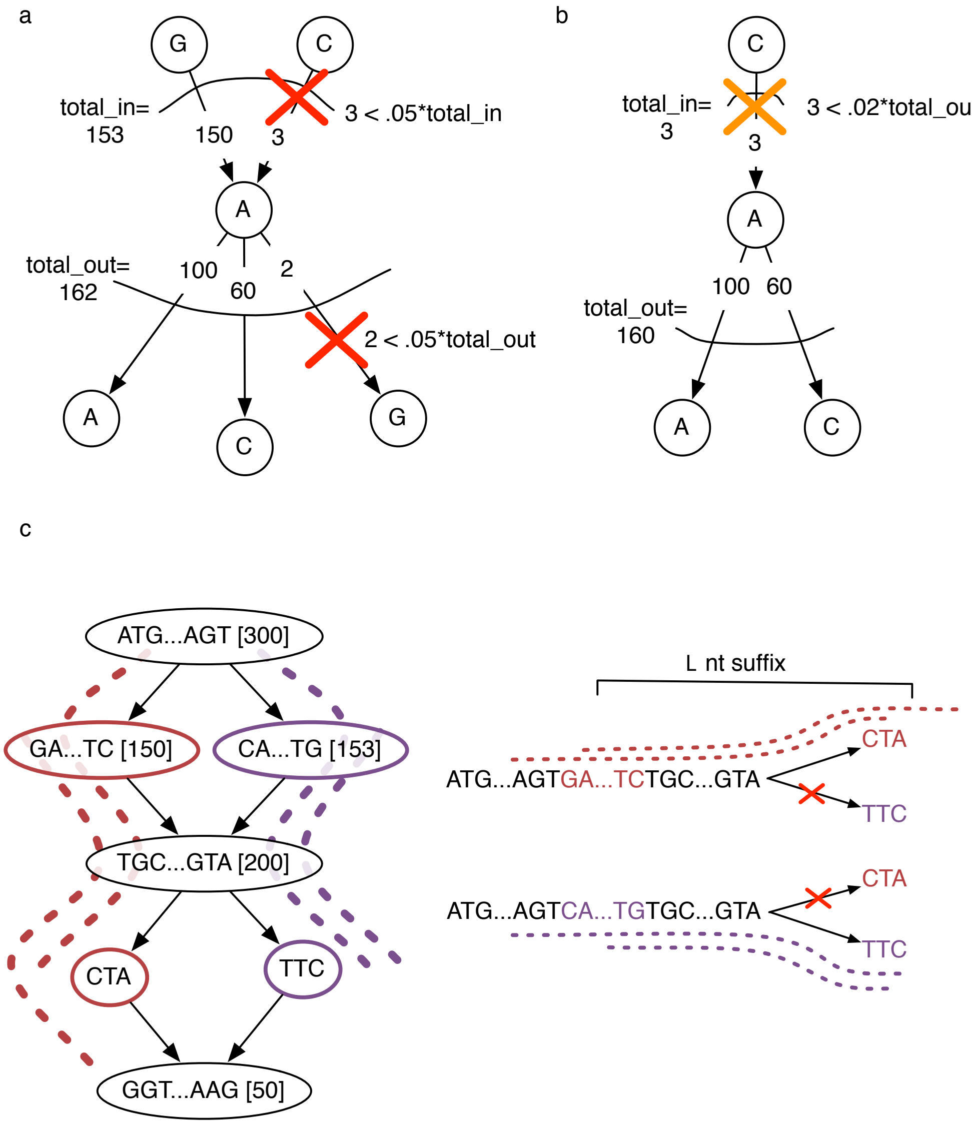
Supplementary Figure 6. The number of full-length transcripts reconstructed by each method at different numbers of input reads.

Shown are the number of annotated full-length transcripts (Y axis) reconstructed at different input read numbers (X axis) for each of Trinity (red), TransAbyss (yellow), Abyss (blue), SOAPdenovo (green), Scripture (purple) and Cufflinks (grey) in yeast (**a**) and mouse (**b**).



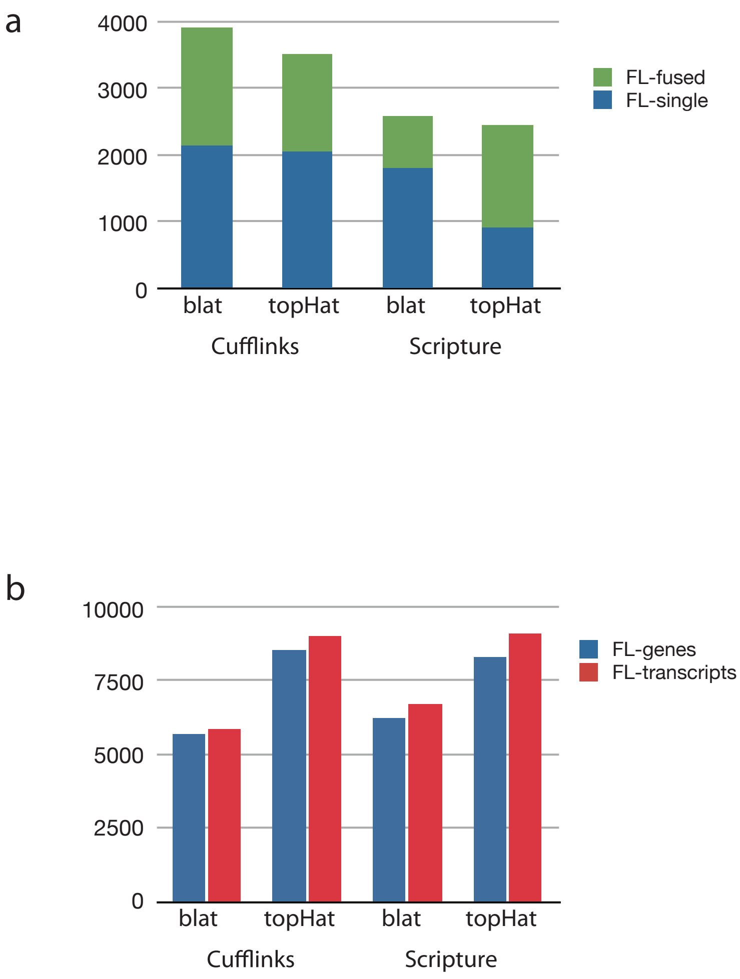
Supplementary Figure 7. The number of full-length transcripts reconstructed by each method at different expression levels.

Shown are the numbers of full-length Oracle transcripts (Y axis) reconstructed at different expression quintiles (X axis) by each of Trinity (red), TransAbyss (yellow), Abyss (blue), SOAPdenovo (green), Scripture (purple) and Cufflinks (grey) in yeast (**a**) and mouse (**b**).



Supplementary Figure 8. Butterfly edge pruning and path finding.

(a,b) Shown are the two cases where we would remove an edge (respectively, see **Methods**). **(c)** Illustrates the progress of the path finding process. On the left we see the compacted graph, each node shows the beginning and end of its sequence, and its length in square brackets. On the right we examine 2 possible extensions for the two paths that reached node (TGC…GTA), and show that we have L-suffix support only for the red and purple paths, and not their chimera paths.



Supplementary Figure 9. The effect of the choice of alignment program on mapping-first transcriptome reconstruction.

Shown are the numbers of annotated full-length transcripts (blue) and full-length fused transcripts (green) reconstructed by Cufflinks and Scripture, based on Blat and the latest version of Tophat in yeast (**a**) and mouse (**b**).

SUPPLEMENTARY TABLES

Supplementary Table 1. Comparison of sensitivity of different methods on the S. pombe transcriptome.

Listed are the number of full-length (FL) genes, the percentage of false fusions, the total number of contigs, the number of contigs that could be mapped to the genome, the number of genes hat overlap with mapped contigs, and the average number of contigs per gene.

	Scripture (blat)	Cufflinks (blat)	ABySS	Trans-ABySS	SOAP-denovo	Trinity
FL genes	2585	3913	3248	4015	1049	4338
% falsely fused genes	30	45	36	27	26	5
Total contigs	14909	4605	6343	39178	12392	27841
Contigs mapped	11714	3258	4601	31974	5456	7057
Genes captured	3838	4182	4533	4871	3400	4874
Average contig coverage/ gene	4.37	1.07	1.06	5.08	1.01	1.37

Supplementary Table 2. Comparison of sensitivity of different methods on the mouse transcriptome.

Listed are the number of full-length (FL) genes, the percentage of false fusions, the total number of contigs, the number of contigs that could be mapped to the genome, the number of genes hat overlap with mapped contigs, and the average number of contigs per gene.

	Scripture (tophat)	Cufflinks (tophat)	ABySS	Trans-ABySS	SOAP-denovo	Trinity
FL transcripts	9086	9010	5561	7025	761	8185
FL genes	8293	8536	5500	6598	760	7749
Total contigs	300148	31121	46783	203085	145518	179340
Contigs mapped	119515	19342	17427	111309	34816	31706
Genes captured	10432	10806	9879	10685	10035	11334
Average contig coverage / gene	12.0	1.65	1.25	5.93	1.12	2.05

Nature Biotechnology: doi:10.1038/nbt.1883

Supplementary Table 3. Base error stats for Trinity transcripts.

Listed are the number of aligned bases, matches, mismatches, insertions and deletions.

	S. pombe	Mouse
# Full-length Trinity Transcripts	4230	8178
# aligned bases	8942895	21400061
# matching bases	8942241	21397375
# mismatches	654	2686
Mismatch rate	7.31e-05	1.26e-04
# genome inserted bases	299	1551
Genome inserted base rate	3.34e-05	7.25e-05
# transcript inserted bases	528	2875
Transcript inserted base rate	5.90e-05	1.34e-04

References

1. Wilhelm, B.T. et al. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* **453**, 1239-1243 (2008).
2. Yassour, M. et al. Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. *Proc Natl Acad Sci USA* **106**, 3264-3269 (2009).