Effect of *de novo* transcriptome assembly on quality of read mapping and transcript quantification

Supplementary File 1

Table S1

Parameter settings used when invoking the Flux Simulator

	Yeast	Dog	Mouse
NB_MOLECULES	5,000,000	35,000,000	85,000,000
FRAGMENTATION	YES	YES	YES
FRAG_SUBTRATE	RNA	RNA	RNA
FRAG_METHOD	UR	UR	UR
RTRASCRIPTION	YES	YES	YES
RT_PRIMER	RH	RH	RH
FILTERING	YES	YES	YES
PCR_PROBABILITY	0.05	0.05	0.05
ERR_FILE	76	76	76
PAIRED_END	YES	YES	YES
READ_LENGTH	150	150	150
READ_NUMBER	3,000,000	21,000,000	50,000,000

Table S2

Sequence Information of RNA-Seq Reads Used in This Study

	Experimental data		
	Yeast	Dog	Mouse
No. of fragments	5,449,060	20,480,445	43,145,380
Max read length	101	50	76
Sequencing depth	118.0	33.2	44.5
Insert size *mean*[*]	204.6	150.5	289.8
Insert size *SD*[*]	96.3	53.9	78.1
Strand specificity	Non-stranded	Non-stranded	RF-stranded
	Simulated data		
	Yeast	Dog	Mouse
No. of fragments	1,212,904	8,464,764	20,209,629
Max Read length	150	150	150
Sequencing depth	33.5	32.1	33.1
Insert size *mean*[*]	193.8	194.0	194.1
Insert size *SD*[*]	30.2	30.1	30.1
Strand specificity	Non-stranded	Non-stranded	Non-stranded

[*] The mean and standard deviation of the insert sizes were estimated using Burrows-Wheeler Aligner.

Table S3

Definitions of Contig Categories

Categories	Definitions
Full-length	1. Only **one** transcript assigned to the contig in interest. 2. The corresponding transcript only assigned to the contig in interest. 3. Both the contig and its corresponding transcript are **unique**. 4. Both *recovery* and *accuracy* of the global alignment (*contig*, *transcript*) ≥ 0.90 5. $-10\% \leq$ *difference in length* of (*contig*, *transcript*) $\leq 10\%$
Incompleteness	1. Only **one** transcript assigned to the contig in interest. 2. The corresponding transcript only assigned to the contig in interest. 3. Both the contig and its corresponding transcript are **unique**. 4. The *accuracy* of the global alignment (*contig*, *transcript*) ≥ 0.90 5. *Difference in length* of (*contig*, *transcript*) $< -10\%$
Over-extended	1. Only **one** transcript assigned to the contig in interest. 2. The corresponding transcript only assigned to the contig in interest. 3. Both the contig and its corresponding transcript are **unique**. 4. The *recovery* of the global alignment (*contig*, *transcript*) ≥ 0.90 5. *Difference in length* of (*contig*, *transcript*) $> 10\%$
Family-collapse	1. All the transcripts that assigned to the contig in interest are in the same connected component. 2. All the corresponding transcript only assigned to the contig in interest. 3. The contig is **unique**. 4. All the transcripts that assigned to the contig are **not unique**. 5. There exists **at least one** global alignment (*contig*, *transcript*) that match the following criteria: 1. Both *recovery* and *accuracy* ≥ 0.90 2. $-10\% \leq$ *difference in length* $\leq 10\%$
Duplication	1. Only **one** transcript assigned to the contig in interest. 2. The corresponding transcript is assigned to multiple contigs which are in the same connected component. 3. The contig is **not unique**. 4. The corresponding transcript is **unique**. 5. There exists **at least one** global alignment (*contig*, *transcript*) that match the following criteria: 1. Both *recovery* and *accuracy* ≥ 0.90 2. $-10\% \leq$ *difference in length* $\leq 10\%$

Table S4

Biological Variability of Reference Transcripts.

	Yeast	Dog	Mouse
No. of genes	5,107	18,045	21,510
No. of transcripts	5,107	23,078	56,706
No. of genes w/ single transcript	5,107	13,713	8,071
No. of transcripts per gene	1.000	1.279	2.636
Max number of transcripts for a gene	1	7	47
No. of unique transcripts	4,806	15,299	16,134
No. of connected components	4,893	18,881	27,940
Connected components size *mean*	1.044	1.222	2.030
Connected components size *max*	62	26	163
Maximum length	14,733	105,543	123,179
Total length	8,512,860	62,014,091	143,577,796
Average length	1,666.90	2,687.15	2,531.97
N50	1,944	3,776	3,523

Table S5

Basic Statistics for Assembled Contigs

Dataset		Assembly	No. of contigs	No. of unique contigs	No. of connected component	Connected component size *mean*	Connected component size *max*	Contigs N50
Simulated	Yeast	rnaSPades	5,620	5,550	5,582	1.007	4	1,575
		TransABySS	5,443	5,264	5,325	1.022	9	1,652
		Trinity	5,440	5,268	5,335	1.020	7	1,711
	Dog	rnaSPades	12,189	11,736	11,953	1.020	7	2,797
		TransABySS	14,239	11,030	12,177	1.169	30	2,687
		Trinity	14,922	11,729	12,872	1.159	42	3,108
	Mouse	rnaSPades	12,592	10,866	11,689	1.077	6	3,148
		TransABySS	17,162	10,055	12,260	1.400	22	2,756
		Trinity	16,320	10,276	12,423	1.314	24	3,295
Experimental	Yeast	rnaSPades	4,776	4,434	4,588	1.041	12	2,652
		TransABySS	5,890	3,776	4,579	1.286	9	2,632
		Trinity	5,506	4,143	4,704	1.170	22	2,634
	Dog	rnaSPades	22,397	20,027	21,136	1.060	11	2,403
		TransABySS	21,359	14,846	17,493	1.221	13	2,181
		Trinity	21,607	17,652	19,278	1.121	21	2,141
	Mouse	rnaSPades	24,522	18,064	20,792	1.179	18	3,186
		TransABySS	37,840	15,511	21,995	1.720	24	2,413
		Trinity	39,060	17,009	23,347	1.673	44	3,051

Table S6

Number of Contigs for Contig Categories

Dataset		Assembly	Full-length	Over-extension	Incomp-leteness	Family-collapse	Duplication
Simulated	Yeast	rnaSPades	2,831	11	564	52	15
		TransABySS	3,175	5	482	43	69
		Trinity	3,306	5	382	51	34
	Dog	rnaSPades	3,546	70	827	1,469	22
		TransABySS	3,340	36	868	1,056	460
		Trinity	3,402	61	944	1,090	288
	Mouse	rnaSPades	1,638	98	253	3,076	14
		TransABySS	1,492	71	293	2,288	339
		Trinity	1,597	92	287	2,431	142
Experimental	Yeast	rnaSPades	689	1,151	132	7	13
		TransABySS	657	868	153	5	229
		Trinity	525	1,067	141	5	119
	Dog	rnaSPades	441	1,385	2,077	209	120
		TransABySS	320	804	1,832	145	348
		Trinity	368	973	1,963	149	148
	Mouse	rnaSPades	1,544	231	585	2,353	626
		TransABySS	972	84	545	988	2,813
		Trinity	1,176	111	585	1,103	1,733