De novo assembly and analysis of RNA-seq data

Gordon Robertson[1], Jacqueline Schein[1], Readman Chiu[1], Richard Corbett[1], Matthew Field[1], Shaun D Jackman[1], Karen Mungall[1], Sam Lee[2], Hisanaga Mark Okada[1], Jenny Q Qian[1], Malachi Griffith[1], Anthony Raymond[1], Nina Thiessen[1], Timothee Cezard[1,4], Yaron S Butterfield[1], Richard Newsome[1], Simon K Chan[1], Rong She[1], Richard Varhol[1], Baljit Kamoh[1], Anna-Liisa Prabhu[1], Angela Tam[1], YongJun Zhao[1], Richard A Moore[1], Martin Hirst[1], Marco A Marra[1,3], Steven J M Jones[1,3], Pamela A Hoodless[2,3] & Inanc Birol[1]

We describe Trans-ABySS, a *de novo* short-read transcriptome assembly and analysis pipeline that addresses variation in local read densities by assembling read substrings with varying stringencies and then merging the resulting contigs before analysis. Analyzing 7.4 gigabases of 50-base-pair paired-end Illumina reads from an adult mouse liver poly(A) RNA library, we identified known, new and alternative structures in expressed transcripts, and achieved high sensitivity and specificity relative to reference-based assembly methods.

Current methods for sequencing transcriptomes using short-read technologies (RNA-seq) generate millions of short sequence reads. These reads are associated with transcript models after mapping the reads to a reference genome, facilitated by extending the genome sequence to include sequences for junctions between annotated exons[1], as in alternative expression analysis by sequencing (ALEXA-seq)[2], or by using a spliced-read alignment algorithm[3–7]. Recently, two methods have been reported that use the latter approach to generate gene models and predict isoforms from aligned reads[8,9]. Read alignments, however, are subject to bias resulting from reads mapping to multiple locations[10,11] and mismatches caused by genome variation[12–14]. When *de novo* assembly is applied to short-read transcriptome data, longer assembled contigs, rather than reads, are aligned to a reference genome, and contig alignments can be compared to transcript annotations to identify new transcripts and new transcript structures. Such an approach has the advantage of requiring no prior knowledge of exon-exon junctions; in addition, *de novo* assembly can be used when a reference genome is unavailable or is poorly annotated[15].

The parameters required for an effective assembly depend on the depth of coverage. For de Bruijn graph assemblers such as ABySS[16,17], which process each read into a set of overlapping substrings (k-mers) of length k base pairs (bp), the most important parameter is the k-mer length. Whole-genome shotgun assembly sequencing libraries attempt to provide a uniform representation of the genome. For these libraries it is reasonable to identify and work with the assembly that corresponds to an optimal k value. In non-normalized transcriptome shotgun libraries, however, individual transcripts differ widely in expression and thus present a wide range of sequence representations to an assembler. A single k value is therefore unlikely to yield an optimal overall assembly (**Supplementary Note 1**).

Recently, we applied the ABySS short-read assembler to human transcriptome data[18]. Based on an assembly for a single k value, in this preliminary analysis we had identified contig structures and alignments that are consistent with alternative isoforms, and thus suggested that ABySS could be effective for transcriptome analysis. However, to make *de novo* assembly practical for characterizing annotated and new transcript structures, we anticipated that it would be necessary to assemble at different k values to address variable transcript expression and multiple expressed isoforms. Here we describe the result of work to address these issues, Trans-ABySS, a method and pipeline for assembly and analysis of non-normalized short-read transcriptome data.

To develop the approach and assess its performance, we generated 147.1 million (7.36 gigabases (Gb)) quality-filtered 50-bp Illumina paired-end reads from a transcriptome library constructed from adult mouse liver poly(A) RNA. We chose this model organism because it has a well-annotated transcriptome and is considered genetically uniform. We assembled the reads using ABySS v1.1.1 (**Supplementary Fig. 1** and **Supplementary Note 1**).

To assess assembly performance as a function of k values, we compared assemblies of the 147.1 million reads for k values ranging from 26 to 50 bp. The number of contigs generated in an assembly ranged from 2.57 million ($k = 26$ bp) to 0.14 million ($k = 50$ bp). Each assembly was dominated by shorter contigs; the fraction of contigs shorter than 100 bp varied from 94.2% for $k = 26$ bp to 31.9% for $k = 50$ bp. Assembly N50 values, the contig lengths for which 50% of the sequence in an assembly is in contigs of this size or larger, were highest for intermediate k values, with a maximum of 1,458 bp at $k = 39$ bp (**Supplementary Fig. 2**). We focused our analysis on the contigs that were at least $2(k - 1)$ bp because $k - 1$ bp is the required k-mer overlap for extending contigs during assembly, and $2(k - 1)$ bp is the expected length for contigs that represent alternative splice junctions. Thus, we considered this contig size range to be the

[1]Canada's Michael Smith Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, Canada. [2]Terry Fox Laboratory, British Columbia Cancer Agency, Vancouver, Canada. [3]Department of Medical Genetics, University of British Columbia, Vancouver, Canada. [4]Present address: University of Edinburgh, Edinburgh, UK. Correspondence should be addressed to I.B. (ibirol@bcgsc.ca).

RECEIVED 18 JUNE; ACCEPTED 13 SEPTEMBER; PUBLISHED ONLINE 10 OCTOBER 2010; DOI:10.1038/NMETH.1517

most informative for each assembly[19] (**Supplementary Fig. 1c**, **Supplementary Note 1** and Online Methods).

To allow comparing *de novo* assemblies to methods based on ungapped read alignments, we used the Burrows-Wheeler aligner[20] to map the reads to the mouse reference genome and known exon-exon junctions. Of 136.7 million aligned reads (6.83 Gb), we retained 118.7 million (5.93 Gb) that had a mapping quality of at least 10 (such reads mapped to unique genomic locations with alignments whose probabilities of being incorrect were ≤ 0.1)[20]. Of the retained reads, 77% aligned to exons and exon-exon junctions, indicating that the dataset was of high quality (**Supplementary Table 1** and **Supplementary Fig. 3**).

We used the exonerate aligner[19] to map the contigs to the reference genome and assessed transcript representation by aligned contigs as a function of normalized read depth from Burrows-Wheeler aligner read alignments to Ensembl[21] transcript models. Of the 34,400 transcripts that had an average read coverage of at least 20×, 72% were represented by a single contig whose alignment overlapped at least 80% of the transcript's total exon length (**Supplementary Fig. 4**). When we assessed transcript coverage from individual assemblies, we observed that transcripts with lower read depths were represented more effectively with lower k values, whereas those with higher read depth were represented more effectively with higher k values (**Fig. 1a**). Hence, assembly across a range of k values is required to recover contigs that represent transcripts with very different expression levels. We retained assemblies for k values of 26–50 bp. Because this generated a large number of contigs, 9.50 million of which had lengths $(L) \geq (2k - 2)$ bp, before analysis we merged contigs from independent assemblies, achieving a smaller nonredundant set that contained 1.20 million contigs with $L \geq (2k - 2)$ bp (**Fig. 1b** and **Supplementary Fig. 5**). Finally, we removed contigs that were unlikely to convey information on transcript structure by filtering the merged contig set, without referring to the mouse genome or transcript annotations (Online Methods); eliminating shorter contigs that were unlikely to represent alternative splice junctions or that were likely to represent genomic or intronic sequence generated a filtered set that contained 0.76 million contigs with $L \geq (2k - 2)$ bp.

We then assessed the alignments of filtered contigs relative to annotated transcript structures. Because contig alignment errors can generate false positive new transcript structures, we also aligned the contigs using BLAST-like alignment tool (BLAT)[22] and retained only candidate transcript structures that were supported by both aligners. Of the 762,486 contigs longer than $(2k - 2)$ bp, 578,140 (76%) had at least one alignment block overlapping an exon from at least one of University of California Santa Cruz (UCSC) genome browser[23], RefSeq[24], Ensembl[21] or AceView[25] transcript models, representing a total of 16,204 annotated genes.

We compared our *de novo* assembly results to reference-based assembly, using Cufflinks[9] and Scripture[8], two recent methods that use the output of the gapped read aligner TopHat[6] to reconstruct gene models. Considering Ensembl transcripts with a fractional exon coverage by a single contig of at least 0.8, Cufflinks was more effective in generating exon coverage for read coverage depths below ~10×, Trans-ABySS was more effective above this depth, and Scripture generated lower exon coverage than either of these methods (**Fig. 2a**, **Supplementary Fig. 6**, **Supplementary Table 2** and **Supplementary Note 1**). Cufflinks and Trans-ABySS reported comparable total numbers of transcripts (11,516 and 11,253 transcripts, respectively) but Scripture reported fewer (9,214 transcripts).

We used a large set of annotated exon-intron boundaries to compare the sensitivity and specificity of TopHat read alignments, Cufflinks and Scripture contigs, and Trans-ABySS contig alignments (**Fig. 2b**). Our reference was the 298,893 nonredundant annotated introns generated by pooling all UCSC genome browser, RefSeq, Ensembl and AceView transcript models. Tophat's specificity was highest (0.940) for introns with 15–40 gapped reads of support. Scripture showed relatively low specificity. Both Cufflinks (0.948) and Trans-ABySS (0.945, 0.959 with splice-site filtering; **Supplementary Fig. 7**) were more specific than Tophat. Sensitivities for Cufflinks and Trans-ABySS were similar to Tophat's for introns with two and one gapped read(s) of support, respectively. Note that introns reported as nonspecific may include both false positives and bona fide introns that were not included in the reference intron set. These results indicate that when variable transcript expression levels and multiple expressed isoforms are addressed, *de novo* assembly offers a high sensitivity and specificity for identifying transcript structures.

We then assessed the new structures identified by Trans-ABySS contig alignments. Comparing contig alignments to four sets of transcript models (**Supplementary Fig. 8** and Online Methods),

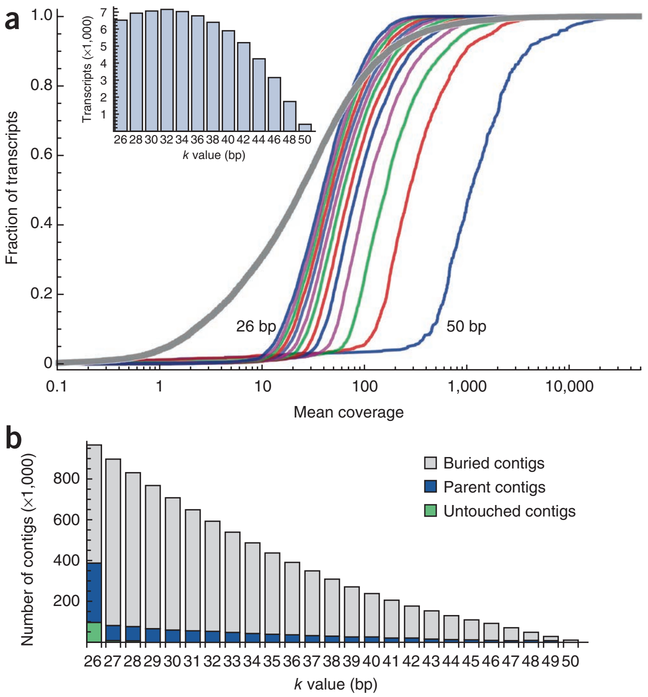

Figure 1 | Representation of transcripts and contigs across assemblies. (**a**) Distributions of normalized mean transcript coverage from read-to-genome alignments and assembly k-mer length, for unmerged contigs from assemblies for every other k value between 26 and 50 bp (left to right, with the curve for each k value in a different color). Results are shown for all Ensembl v54 mouse transcripts (gray), and for contigs that cover at least 80% of the transcript's total exon length. Inset, distribution of transcripts for each each k value. (**b**) Result of contig merging for main contigs from assemblies with k values of 26–50 bp. 'Buried' contigs are those with an exact sequence match within a longer 'parent' contig from another assembly. 'Untouched' contigs have no sequence match in another assembly.

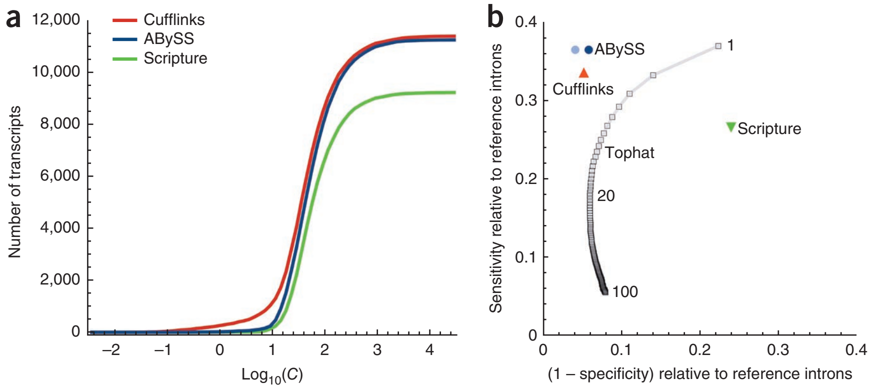

Figure 2 | Performance comparisons between ABySS and reference-based transcriptome analysis tools. (**a**) Number of Ensembl v54 transcripts reconstructed to 80% by ABySS, Cufflinks and Scripture by a single contig as a function of mean read coverage, *C*. (**b**) Intron-level sensitivity and specificity of Trans-ABySS, Cufflinks, Scripture and TopHat, relative to all 298,893 nonredundant introns from UCSC genome browser, RefSeq, Ensembl and AceView transcript models. Tophat split-read alignments are shown as a curve for intron support levels ranging from 1 to 100 reads. Alignments of Trans-ABySS *de novo* contigs, and reference-based Tophat Cufflinks and Scripture generated contigs, are represented as points, each of which represents a set of contigs. Two points are shown for ABySS: non-reference-based filtering with (light blue) or without (dark blue) contig-level splice-site filtering.

we identified 866 candidate new events corresponding to annotated loci (**Supplementary Table 3**). These consisted of 94 new exons, 117 skipped exons, 56 new introns, 293 retained introns, 184 alternative 3′ or 5′ exon splicing events, and 122 new untranslated regions (UTRs). Representative examples of some of these event types are shown in **Supplementary Figures 9–13**. We noted that contig alignments could predict new events that were too short to be detected by ungapped alignments of the 50-bp reads (**Supplementary Fig. 9**). Retained introns were the most frequent new structure. The majority of the genes with such candidates were highly expressed (**Supplementary Fig. 14**), suggesting that some of these events may represent incomplete splicing[26], with the unprocessed introns having sufficient read representation to assemble into contigs. The pipeline supports filtering such structures using the mean read coverage for the flanking exons and the ratio of the mean coverage of the flanking exons to that for the predicted retained intron (**Supplementary Fig. 15**).

We developed a contig-based method for detecting polyadenylation sites using reads that aligned at junctions of a transcript and a poly(A) tail[27], and mate-pair reads with one mate within a transcript's poly(A) tail (**Supplementary Fig. 16** and **Supplementary Note 1**). The method can be used with or without an annotated reference genome. We confirmed known polyadenylation start sites in 1,299 annotated transcripts, and inferred 84 new polyadenylation sites that corresponded to 56 new short 3′ UTRs and 32 new long 3′ UTRs (**Supplementary Table 4** and **Supplementary Fig. 17**).

We anticipated that normal mouse liver tissue would have no fusion genes but note that the pipeline includes an algorithm for detecting such chimeric transcripts (**Supplementary Fig. 18** and **Supplementary Note 1**). We used this algorithm to identify validated translocations in short-read transcriptome data for human non-Hodgkin lymphomas (data not shown).

The pipeline also includes a gene-level expression metric based on reads aligned to contigs that can be used with or without an annotated reference genome. To assess this metric, we showed that gene-level expression estimates for aligned Trans-ABySS contigs were comparable to those from ALEXA-seq[2] read alignments (**Supplementary Fig. 19** and **Supplementary Note 1**). For the 8,190 genes with fractional contig-to-exon coverage of at least 0.8, Trans-ABySS and ALEXA-seq expression estimates had a Pearson's correlation coefficient of 0.921.

We note that identifying new structures efficiently requires accurate alignments of *de novo* contigs to a genome assembly, and improvements in this area will depend on developing more effective alignment approaches. Several other issues remain challenging for both *de novo* and reference-based methods, such as repetitive regions, nonuniform local read densities and complete isoform reconstruction (**Supplementary Note 1**).

For species that lack reference genome sequences, or whose genomes are poorly annotated, *de novo* short-read transcriptome assembly may be a practical alternative to conventional expressed sequence tag–based approaches and to methods that depend on short-read alignments. For example, with sequencing technologies becoming less expensive and more widespread, the method described may have an important impact in evolutionary developmental biology. When an annotated genome sequence is available, the pipeline can be used to detect events that are not annotated as well as events that are not represented by the reference genome, as in tumor transcriptomes.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturemethods/.

Accession codes. Short Read Archive: SRA012213.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

Funding for this work was provided in part by Genome Canada, Genome British Columbia, Michael Smith Foundation for Health Research and the Canadian Institute of Health Research (CIHR), including the CIHR Bioinformatics Training Program for Health Research. We thank S. Morrissy and G. Taylor for insightful discussions, A. He for technical assistance, A. Fejes for assistance with coverage bias calculations, and A. Tuin and N. Watkins (DNA Software) for assistance with primer design.

AUTHOR CONTRIBUTIONS

G.R. and J.S. wrote the paper. J.S., G.R. and K.M. reviewed predictions and recommended analysis methods. G.R. coordinated analysis and validation. B.K., A.-L.P. and A.T. constructed libraries under the supervision of YJ.Z. S.L. generated biological material and performed RT-PCR validation. R.A.M. supervised sequencing activities. Y.S.B., T.C., R. Corbett, R. Chiu, M.F., M.G., J.Q.Q., R.N., H.M.O., N.T., R.V., S.K.C. and R.S. developed analysis methods and code and performed analyses. R. Corbett and R. Chiu performed comparisons with reference-based methods. S.D.J. develops and maintains ABySS and generated the ABySS assemblies. A.R. contributed algorithms and code for ABySS. M.A.M., S.J.M.J. and P.A.H. directed research. S.J.M.J. suggested analysis methods. YJ.Z. and M.H. developed the WTSS protocol. J.S. supervised activities. P.A.H. supervised validation. I.B. developed ABySS and Trans-ABySS and directed bioinformatics work.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturemethods/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Pepke, S., Wold, B. & Mortazavi, A. *Nat. Methods* **6**, S22–S32 (2009).
2. Griffith, M. *et al. Nat. Methods* **7**, 843–847 (2010).
3. Ameur, A. *et al. Genome Biol.* **11**, R34 (2010).
4. Au, K.F. *et al. Nucleic Acids Res.* **38**, 4570–4578 (2010).
5. De Bona, F. *et al. Bioinformatics* **24**, i174–i180 (2008).
6. Trapnell, C., Pachter, L. & Salzberg, S.L. *Bioinformatics* **25**, 1105–1111 (2009).
7. Wu, T.D. & Nacu, S. *Bioinformatics* **26**, 873–881 (2010).
8. Guttman, M. *et al. Nat. Biotechnol.* **28**, 503–510 (2010).
9. Trapnell, C. *et al. Nat. Biotechnol.* **28**, 511–515 (2010).
10. Li, B. *et al. Bioinformatics* **26**, 493–500 (2010).
11. Li, J., Jiang, H. & Wong, W.H. *Genome Biol.* **11**, R50 (2010).
12. Krawitz, P. *et al. Bioinformatics* **26**, 722–729 (2010).
13. Cartwright, R.A. *Mol. Biol. Evol.* **26**, 473–480 (2009).
14. Degner, J.F. *et al. Bioinformatics* **25**, 3207–3212 (2009).
15. Birzele, F. *et al. Nucleic Acids Res.* **38**, 3999–4010 (2010).
16. Simpson, J.T. *et al. Genome Res.* **19**, 1117–1123 (2009).
17. Flicek, P. & Birney, E. *Nat. Methods* **6** (Suppl.), S6–S12 (2009).
18. Birol, I. *et al. Bioinformatics* **25**, 2872–2877 (2009).
19. Slater, G.S. & Birney, E. *BMC Bioinformatics* **6**, 31 (2005).
20. Li, H. & Durbin, R. *Bioinformatics* **25**, 1754–1760 (2009).
21. Hubbard, T.J. *et al. Nucleic Acids Res.* **37**, D690–D697 (2009).
22. Kent, W.J. *Genome Res.* **12**, 656–664 (2002).
23. Hsu, F. *et al. Bioinformatics* **22**, 1036–1046 (2006).
24. Pruitt, K.D., Tatusova, T. & Maglott, D.R. *Nucleic Acids Res.* **35**, D61–D65 (2007).
25. Thierry-Mieg, D. & Thierry-Mieg, J. *Genome Biol.* **7** (Suppl.), 11–14 (2006).
26. Melamud, E. & Moult, J. *Nucleic Acids Res.* **37**, 4873–4886 (2009).
27. Nagalakshmi, U. *et al. Science* **320**, 1344–1349 (2008).

ONLINE METHODS

Sample preparation. For adult RNA, a C57BL/6J female (5 months old, nonpregnant) was killed and the liver perfused with PBS to eliminate blood cells from the liver. Once the redness in the liver disappeared and the tissue became pale, the liver was manually dissected and homogenized. One to two grams of the homogenate was put into Trizol (Invitrogen) for RNA extraction.

Library construction and sequencing. For whole-transcriptome shotgun sequencing (WTSS), polyadenylated RNA was purified from 10 μg of DNase I (Invitrogen) treated total RNA using the MACS mRNA Isolation kit (Miltenyi Biotec). Double-stranded cDNA was synthesized from the purified poly(A)$^+$ RNA using a Superscript Double-Stranded cDNA Synthesis kit (Invitrogen) and random hexamer primers (Invitrogen) at a concentration of 5 μM. The resulting cDNA was sheared using a Bioruptor UCD-200 (Diagenode) with output setting of H (high, 200 W) for 50 min in pulses of 30 s interspersed with 30 s of cooling and was then size fractionated using 8% PAGE. The 190–210-bp DNA fraction was excised, eluted overnight at 4 °C in 300 μl of elution buffer (5:1, LoTE buffer (3 mM Tris-HCl (pH 7.5), 0.2 mM EDTA)–7.5 M ammonium acetate) and purified using a QIAquick purification kit (Qiagen). The sequencing library was prepared following the Illumina Genome Analyzer (GA) paired-end library protocol with 10 cycles of PCR amplification. PCR products were purified on Qiaquick MinElute columns (Qiagen) and assessed and quantified using an Agilent DNA 1000 series II assay and Qubit fluorometer (Invitrogen), respectively. The resulting libraries were sequenced on an Illumina Genome Analyzer II following the manufacturer's instructions. WTSS libraries were paired-end sequenced to 50 cycles.

Read processing. Transcriptome reads were generated using two versions of Illumina's GA basecalling pipeline. Two lanes were processed with v1.9.5, which used alignment-dependent quality scores and no 'chastity' filtering. One of these lanes offered only single-end rather than paired-end data. Six lanes of data were processed with a later version, v1.3.2, which used alignment-independent quality scoring and 'chastity' filtering. Read pairs that had the first six bases in common were removed ('shadow read filtering').

Read alignment used 147.06 M shadow- and chastity-filtered reads from the seven paired-end lanes. ABySS assembly used these, as well as 6.39 M shadow- and chastity-filtered reads from the single-end lane, for a total of 153.44 M reads. An additional 51.94 M reads were available that were shadow-filtered but failed chastity filtering. ABySS used the mate-pair information in these reads in scaffolding contigs at its paired-end stage (see below) but did not use k-mers from these read sequences for extending single-end contigs. In total, including this second group of reads, ABySS accessed 205.38 M reads.

ABySS assembly. The ABySS assembly process consists of single-end and paired-end stages[16,28] (**Supplementary Fig. 1**, **Supplementary Fig. 20** and **Supplementary Note 1**). The single-end stage is based on a de Bruijn graph construct, in which, given a parameter k, reads are transformed into tiled k-mers, represented as nodes, and $(k - 1)$-base overlaps between them, as directed edges. Allelic differences, repeat sequences with minor variations and recurrent coincident base-calling read errors form 'bubbles' along this graph. These are 'popped' by removing the variant with lower coverage from the graph, and both variants are recorded, with the $(k - 1)$ bases of contextual sequence on either end. After error removal in the k-mer space, unambiguous 'walks' along the graph define single-end contigs. In the paired-end stage, mate pairs that align within individual single-end contigs define the empirical distribution of mate pair distances. Mate pairs in which the reads align to different single-end contigs and the empirical distribution are then used to estimate intercontig distances, and contigs that can be unambiguously merged (that is, those whose measured distances on the de Bruijn graph are consistent with the estimated distances) are merged to form paired-end contigs. The total numbers of k-mers assembled into each contig is recorded as a surrogate for the contig's read-coverage depth.

Contig merges are made only when certain criteria for contig length and number of supporting mate pairs are satisfied; for the work reported here, a minimum of 10 mate pairs were required, and the list of merged contigs had to start and end with contigs that were at least 100 bp long. These parameters affect the potential junction contigs that will be in the set of $L < (2k - 2)$ contigs.

Contig assembly and processing. We used ABySS v1.1.1 to generate assemblies for $26 \leq k \leq 50$, with parameters $E = 0$ and $n = 10$.

To merge the contig lists from all k-mer assemblies, we considered 'main' contigs that were at least $(2k - 2)$ bp long (**Supplementary Fig. 1** and **Supplementary Note 1**). We paired k-mer assemblies with adjacent k values (that is, k_i with k_{i+1}, k_{i+2} with k_{i+3} and so on) and used BLAT[22] v34 to align the contigs from one paired assembly ('a') against the contigs of its paired assembly ('b') (**Supplementary Fig. 2**). We then repeated the alignments in the opposite direction, aligning 'main' contigs from 'b' against those from 'a'. Considering results from both directions, we discarded contigs from one assembly that were completely subsumed or 'buried', as exact full-length matches, within contigs from the other assembly. We continued the pairwise comparisons in a hierarchical manner until we obtained a single set of contigs (**Fig. 1b**). We also removed 'untouched' contigs, retaining only 'parent' contigs.

We applied two types of nonreference-based filtering to merged contigs (that is, contig filtering that did not require a reference genome or transcript annotations).

First, we identified main contigs that were $(2k - 2)$ bp long and aligned those to the rest of main contigs. We removed any that failed to align to a single main contig in two exact-match blocks. These were unlikely to be junction contigs that represented real alternative assembly paths. Similarly, we identified single-end junction contigs that had been used to create extended junction contigs and removed any that did not align to a single main contig in two exact-match blocks. We note that this filtering may be somewhat too stringent because some of the rejected contigs may have been valid contigs but did not align with an exact sequence match. Such contigs may have different contig lengths owing to, for example, differences in homopolymer run lengths or other sequencing artifacts.

Second, we removed 'island' single-end contigs (that is, contigs that were not included in paired-end contigs) that had no adjacent contigs in the de Bruijn graph and were shorter than a threshold of 150 bp.

doi:10.1038/nmeth.1517

Such contigs are likely to represent intergenic or intronic regions and so were unlikely to be informative of transcript structure.

We also assessed an additional filtering step that would be available for species that have a reference genome sequence, by retaining only contigs whose alignments to the genome contained at least one implied intron in which we could recognize either a canonical (GT-AG) or noncanonical (GC-AG or AT-AC) acceptor-donor pair. For the mouse genome, more than 98.5% of splice sites are accounted for by these three splicing signals[29]. In our filtered contig set, 97.9% of alignments had at least one intron with recognizable donor and acceptor sites (**Supplementary Fig. 17**).

Aligning contigs to the reference genome. We aligned contigs to the reference mm9 genome using exonerate[19] v2.2 in est2genome mode. For each contig, we considered only the highest-scoring alignment(s). We retained only those contigs in which the highest-scoring alignment was unique and for which at least 90% of the contig length matched the genome. We used custom Python scripts to parse the reported alignments to identify single-nucleotide variations and insertion deletions and to parse the GFF output files into custom tracks for manual review in the UCSC Genome Browser[30].

To provide support for candidate sequence variants or new transcript events identified from the exonerate alignments, we also aligned the contigs with BLAT[22] v34. To be considered for further analysis, any candidate event was required to be present in both exonerate and BLAT alignments.

Identifying new transcripts and transcript events. The genomic alignment for each contig consisted of one or more alignment blocks along the length of the contig. We compared the genome coordinates of all alignment blocks for each contig with the genomic coordinates for exons in each transcript model in UCSC genome browser gene, RefSeq, Ensembl and AceView annotations, which we obtained from the UCSC mouse mm9 genome browser. Because contig ends do not necessarily reflect transcript ends, the two terminal contig alignment coordinates (that is, the outer edges of the first and last alignment block) were excluded; we considered only alignment coordinates between the terminal coordinates.

We assigned a 'full match' for a contig alignment to an annotated transcript when (i) coordinates of the inside edges of the outer (terminal) alignment blocks matched coordinates of the transcript's exons, and (ii) the coordinates of all internal alignment blocks also 'matched' all of the exons in the transcript model internal to the exons identified in case i (**Supplementary Fig. 8**). Edges between an alignment block and transcript exon were considered 'matched' when they had the same coordinate, or, for potential new splice site events, when they were located after the previous matching pair or before the next matching pair. Potential exon-skipping events were identified when neighboring alignment blocks skipped intervening exons in all transcript models from all four annotation systems considered. Potential new exons were identified when one or more extra alignment blocks were present between neighboring exons, considering all transcript models. Potential introns, or deletions within an exon, were identified by gap(s) identified in neighboring alignment blocks whose outermost edges matched (see above) the terminal edges of a single exon. Potential new UTRs were identified when extra alignment blocks extended beyond the first or last exon of all transcript models but did not match exons from overlapping transcripts. Potential intron retention events were identified when intervening intron(s) from all transcript models at the same location were captured in a single alignment block. A multiblock alignment that did not match any of the transcript models represented a potential new transcript. Contig alignments were required to have at least three alignment blocks for exon skipping and new transcript events.

We identified candidate new events independently in exonerate and BLAT contig alignments. To identify a high-confidence subset of new events, we required that coordinates of such an event agree between the two aligners. However, we required this only for the part(s) of the contig alignment that represented (or marked) the new event and allowed the alignments to differ away from the new event. For the case of a new transcript, though, the entire BLAT and exonerate alignments were required to be identical.

We distinguished 5′ versus 3′ UTRs and alternative splicing by inferring transcript strand from donor and acceptor splice sites reported by exonerate for a multiblock alignment. We reported splice site sequences for such events as new splice sites, new exons, new introns and new UTRs. Open reading frames for all contigs containing potential new events were predicted using BioPython (http://www.biopython.org/) to evaluate whether the new events would lead to premature stops in translation.

For the retained intron ranking metric analysis (**Supplementary Fig. 15**), coordinates of mouse retained intron records from the ASTD[31] v1.1 database were converted from mm8 to mm9 with the UCSC genome browser liftOver utility (http://genome.ucsc.edu/cgi-bin/hgLiftOver/). A custom Perl script determined the average coverage of each retained intron and its flanking exons from a wiggle-format file.

Quantifying gene-level expression. We compared gene-level expression results for Trans-ABySS to two methods that align reads to a reference genome that has been extended with exon-exon junctions from annotated transcripts: ALEXA-seq[2] and a whole-transcriptome shotgun sequencing (WTSS) pipeline (unpublished data; **Supplementary Fig. 21** and **Supplementary Note 1**).

The pipeline includes a general method for determining a contig-based expression metric for gene loci, which can be used with or without an annotated genome sequence. For the comparison reported here, we used the NCBI37/mm9 mouse reference genome and Ensembl v54 transcript annotations. We considered reads aligned to all contigs whose alignment blocks on the reference genome overlapped with exons in transcript model annotations, as follows.

For each Ensembl v54 mouse gene, we identified all filtered contigs whose alignment blocks overlapped exons for each transcript for the gene. For each gene locus, considering all transcripts for that gene, we generated a list of unique identifiers for these contigs, then coordinates for a union set of all alignment blocks and the total length of alignment blocks in this union block set.

We aligned reads to each filtered contig using Bowtie[32]. We required exact matches, but allowed multimapping, to accommodate contigs whose alignment blocks overlapped. We then identified all reads on each contig whose blocks overlapped union exons for a gene and resolved overlapping contig alignments by extracting a list of unique read identifiers for all reads associated with the gene's contigs.

We transformed such a read identifier list into a gene-level expression metric by dividing the total length of reads in the list

doi:10.1038/nmeth.1517

by the total length of the union set of alignment blocks. For each normalized coverage value, we also calculated an expression score by dividing by the sum of the number of million reads aligned to all union alignment block sets.

To compare gene-level expression between Trans-ABySS and ALEXA-seq, and Trans-ABySS and the WTSS pipeline, we considered only the subset of 8,190 Ensembl genes for which the gene-level fractional coverage of union exons by union contig alignment blocks was at least 0.8. We calculated a gene-level fractional coverage as the ratio of the length of the union contig alignment block set to the length of the union exon set. Because Trans-ABySS contig alignment blocks can extend outside of the extents of exons in a union exon set for a gene (owing to, for example, retained introns, 3′ UTRs that are longer than annotated for Ensembl transcripts or new exons), the union alignment block length used was for the intersection of the union alignment block set with the union exon set.

Note that although the results reported here used only Ensembl transcript annotations, the method is general and can use other sets of transcript annotations.

Other methods. Detailed information on *de novo* transcriptome assembly, issues for *de novo* and reference-based transcriptome assembly, comparing *de novo* and reference-based assembly, detecting new polyadenylation sites, identifying fusion genes, quantifying gene-level expression, validating new transcripts and transcript events, the WTSS aligned-read pipeline and generating splice graph visualizations is available in **Supplementary Note 1**.

Availability. Trans-ABySS pipeline scripts are available at http://www.bcgsc.ca/platform/bioinfo/software/.

28. Jackman, S.D. & Birol, I. *Genome Biol.* **11**, 202 (2010).
29. Sheth, N. *et al. Nucleic Acids Res.* **34**, 3955–3967 (2006).
30. Rhead, B. *et al. Nucleic Acids Res.* **38** Database issue, D613–D619 (2010).
31. Koscielny, G. *et al. Genomics* **93**, 213–220 (2009).
32. Trapnell, C. & Salzberg, S.L. *Nat. Biotechnol.* **27**, 455–457 (2009).

doi:10.1038/nmeth.1517