

## Sequence analysis

**De novo transcriptome assembly with ABySS**

Inanç Birol<sup>1,\*</sup>, Shaun D. Jackman<sup>1</sup>, Cydney B. Nielsen<sup>1</sup>, Jenny Q. Qian<sup>1</sup>, Richard Varhol<sup>1</sup>, Greg Stazyk<sup>1</sup>, Ryan D. Morin<sup>1</sup>, Yongjun Zhao<sup>1</sup>, Martin Hirst<sup>1</sup>, Jacqueline E. Schein<sup>1</sup>, Doug E. Horsman<sup>2</sup>, Joseph M. Connors<sup>2</sup>, Randy D. Gascoyne<sup>2</sup>, Marco A. Marra<sup>1</sup> and Steven J. M. Jones<sup>1</sup>

<sup>1</sup>Genome Sciences Centre, 100-570 W 7th Avenue, Vancouver, BC V5Z 4S6 and <sup>2</sup>British Columbia Cancer Agency, 600 West 10th Avenue, Vancouver, BC V5Z 4E6, Canada

Received on April 27, 2009; revised on June 5, 2009; accepted on June 9, 2009

Advance Access publication June 15, 2009

Associate Editor: Alex Bateman

**ABSTRACT**

**Motivation:** Whole transcriptome shotgun sequencing data from non-normalized samples offer unique opportunities to study the metabolic states of organisms. One can deduce gene expression levels using sequence coverage as a surrogate, identify coding changes or discover novel isoforms or transcripts. Especially for discovery of novel events, *de novo* assembly of transcriptomes is desirable.

**Results:** Transcriptome from tumor tissue of a patient with follicular lymphoma was sequenced with 36 base pair (bp) single- and paired-end reads on the Illumina Genome Analyzer II platform. We assembled ~194 million reads using ABySS into 66 921 contigs 100 bp or longer, with a maximum contig length of 10 951 bp, representing over 30 million base pairs of unique transcriptome sequence, or roughly 1% of the genome.

**Availability and Implementation:** Source code and binaries of ABySS are freely available for download at <http://www.bcgsc.ca/platform/bioinfo/software/abyss>. Assembler tool is implemented in C++. The parallel version uses Open MPI. ABySS-Explorer tool is implemented in Java using the Java universal network/graph framework.

**Contact:** ibirol@bcgsc.ca

**1 INTRODUCTION**

Second-generation sequencing technologies are increasingly being employed for characterizing genomes (Bentley *et al.*, 2008; Dohm *et al.* 2007; Farrer *et al.*, 2009; Hernandez *et al.* 2008; Kozarewa *et al.*, 2009; Ossowski *et al.* 2008; Salzberg *et al.*, 2008; Warren *et al.*, 2009) and transcriptomes (Fullwood *et al.*, 2009; Jackson *et al.*, 2009; Morin *et al.*, 2008; Wang *et al.* 2008; Yassour *et al.*, 2009). Expanding read lengths, protocols for paired end reads, and the ability to sequence fragments with larger insert sizes have all enabled *de novo* assemblies of genomes. In contrast, analysis of transcriptome sequence data has mainly relied on alignment of reads to reference sequence data sets (Fullwood *et al.*, 2009; Morin *et al.*, 2008; Wang *et al.* 2008; Yassour *et al.*, 2009). Although powerful, analysis methods based on read alignments are limited

by their inability to detect structural alterations not present in the reference sequence data, especially when the read lengths are short.

Recently there has been an effort to develop a tool for transcriptome assembly using short read technologies based on simulated data (Jackson *et al.*, 2009), but it is not yet demonstrated to be applicable to experimental data. Here, we present a *de novo* assembly approach for transcriptome analysis using the ABySS assembler tool (Simpson *et al.*, 2009), which works on experimental data, and we show that transcriptome assembly yields interesting biological insights. ABySS was developed initially for *de novo* assembly of genomes, with a special emphasis on large genomes, and we previously demonstrated its capacity by assembling the human genome using 36–42 bp short reads.

The analysis of a transcriptome assembly is substantially different from the analysis of a genome assembly. For instance, whereas genome sequence coverage levels can be distributed randomly or fluctuate as a consequence of repeat content, transcriptome coverage levels are additionally highly dependent on gene expression levels. Similarly, whereas contig growth ambiguities in a genome assembly represent unresolved repeat structures or alleles, in a transcriptome assembly these ambiguities may also correspond to variations in isoforms and gene families, thus harboring useful and important information. Due to these variations, as well as the abundance of small transcripts, the contiguity of a transcriptome assembly is low. Thus, again unlike a genome assembly, contiguity of an assembly is not indicative of its quality.

Despite these challenges, a transcriptome assembly is desirable as it may facilitate resolution of isoforms by detecting interesting features such as alternative splicing events, as well as discovery of novel transcripts. Using sequence coverage as a surrogate, it will also enable the measurement of exon-, transcript- and variant-level degrees of expression. Of course, analysis of the assembled contigs still requires comparison to a reference genome and/or transcriptome in resequencing or comparative genomics studies. Thus, tools developed for nucleotide and structural variation detection based on alignments are still relevant, and transcriptome assembly in effect enables such tools to work with longer sequences.

In this work, we also report on our assembly visualization tool, ABySS-Explorer (Nielsen *et al.*, 2009), which uses the output from ABySS and enables manual inspection and refinement of assemblies.

\*To whom correspondence should be addressed.

Furthermore, it aids incorporation of additional data, such as paired-end sequencing data with longer inserts.

We used ABySS to assemble the transcriptome of a follicular lymphoma tumor sample, and ABySS-Explorer to visually inspect our contigs. In this article we illustrate our analysis methods, and share our encouraging preliminary results.

## 2 METHODS

### 2.1 Patient sample

The transcriptome data belongs to a patient who presented at 44 years of age with bulky stage II A intra-abdominal follicular, grade 1 non-Hodgkin lymphoma based on an inguinal lymph node biopsy. The staging bone marrow biopsy revealed no lymphoma. Initial treatment consisted of eight cycles of CVP-R (cyclophosphamide, vincristine, prednisone and rituximab) chemotherapy and produced a partial response. However, within 3 months symptomatic progression of lymphoma was evident within the abdomen and a repeat inguinal lymph node biopsy revealed residual grade 1 follicular lymphoma. We obtained informed consent from the patient, approved by the Research Ethics Board, and material from this biopsy was subjected to genomic analyses including whole transcriptome shotgun sequencing (WTSS).

### 2.2 Library construction and sequencing

RNA was extracted from the tumour biopsy sample using AllPrep DNA/RNA Mini Kit (Qiagen, USA) and DNaseI (Invitrogen, USA) treated following the manufacturer's protocol. We generated three WTSS libraries, one from amplified complementary DNA (cDNA), another from the same amplified cDNA with normalization, and the last from unamplified cDNA, as follows.

**2.2.1 WTSS-lite and normalized WTSS-lite libraries** Double-stranded amplified cDNA was generated from 200 ng RNA by template-switching cDNA synthesis kit (Clontech, USA) using Superscript Reverse Transcriptase (Invitrogen, USA), followed by amplification using Advantage 2 PCR kit (Clontech, USA) in 20 cycle reactions. Custom biotinylated PCR primers containing MmeI recognition sequences were used to facilitate the removal of primer sequences from cDNA template for WTSS-Lite library construction.

Normalized cDNA was generated from 1.2 µg of the above amplified cDNA using Trimmer cDNA Normalization Kit (Evrogen, Russia) followed by amplification using the same biotinylated PCR primers, in a single 15 cycle reaction with Advantage 2 Polymerase (Clontech, USA). The normalized and amplified cDNA pool generated with the 1/2× duplex-specific nuclease (DSN) enzyme dilution was chosen to be the template for WTSS-Lite normalized library construction.

For both WTSS-Lite and normalized WTSS-Lite libraries, the removal of amplification oligonucleotide templates from the cDNA ends was accomplished by the binding to M-280 Streptavidin beads (Invitrogen, USA), followed by MmeI digestion. The supernatant of digest was purified and prepared for library construction as follows: roughly 500 ng of cDNA template was sonicated for 5 min using a Sonic Dismembrator 550 (cup horn, Fisher Scientific, Canada), and size fractionated using 8% PAGE gel. The 100–300 bp size fraction was excised for library construction according to the Genomic DNA Sample Prep Kit protocol (Illumina, USA), using 10 cycles of PCR and purified using a Spin-X Filter Tube (Fisher Scientific) and ethanol precipitation. The library DNA quality was assessed and quantified using an Agilent DNA 1000 series II assay and Nanodrop 1000 spectrophotometer (Nanodrop, USA) and diluted to 10 nM.

**2.2.2 Unamplified WTSS library** We used 5 µg RNA to purify polyA+RNA fraction using the MACS mRNA Isolation Kit (Miltenyi Biotec, Germany). Double-stranded cDNA was synthesized from the

purified polyA+RNA using Superscript Double-Stranded cDNA Synthesis kit (Invitrogen, USA) and random hexamer primers (Invitrogen, USA) at a concentration of 5 µM. The cDNA was sheared and library was constructed by following the same Genomic DNA Sample Prep Kit protocol (Illumina, USA).

**2.2.3 Sequencing** Derived cDNA libraries were used to generate clusters on the Illumina cluster station and sequenced on the Illumina Genome Analyzer II platform following the manufacturer's instructions. We ran seven lanes each of the two amplified libraries to generate 36 bp single end tag (SET) reads, and seven lanes of the unamplified library to generate 36 bp paired end tag (PET) reads.

### 2.3 Assembly method

We assembled the reads using ABySS (Simpson *et al.*, 2009). The ABySS algorithm is based on a de Bruijn di-graph representation of sequence neighborhoods (de Bruijn, 1946), where a sequence read is decomposed into tiled sub-reads of length  $k$  ( $k$ -mers) and sequences sharing  $k - 1$  bases are connected by directed edges. This approach was introduced to DNA sequence assembly by Pevzner *et al.* (2001) and was followed by others (Butler *et al.*, 2008; Chaisson and Pevzner, 2008; Jackson *et al.*, 2009; Zerbino and Birney, 2008). Although memory requirements for implementing de Bruijn graphs scale linearly with the underlying sequence, ABySS uses a distributed representation that relaxes these memory and computation time restrictions (Simpson *et al.*, 2009).

A de Bruijn graph captures the adjacency information between sequences of a uniform length  $k$ , defined by an overlap between the last and the first  $k - 1$  characters of two adjacent  $k$ -mers. ABySS starts by cataloging  $k$ -mers in a given set of reads and establishes their adjacency, represented in a distributed data format. The resulting graph is then inspected to identify potential sequencing errors and small-scale sequence variation.

When a sequence has a read error, it alters all the  $k$ -mers that span it, which form branches in the graph. However, since such errors are stochastic in nature, their rate of observation is substantially lower than that of correct sequences. Hence, they can be discerned using coverage information, and branches with low coverage can be culled to increase the quality and contiguity of an assembly. This is especially true for genomic sequences. For transcriptomes, however, sequence coverage depth is a function of the transcript expression level it represents. Therefore, such culling needs to be performed with extra care. Accordingly, in the assembly stage, we applied trimming for those (false) branches when the absolute coverage levels were below a threshold of 2-fold ( $2 \times$ ). In the analysis stage, we evaluated assembly branches using the local coverage information, as well as contig lengths. For instance, in a neighborhood where a contig  $C_1$  branches into contig  $C_2$  and  $C_3$ , with coverage levels of  $x_1$ ,  $x_2$  and  $x_3$ , and contig lengths  $l_1$ ,  $l_2$  and  $l_3$ , respectively, if  $x_1$  and  $x_2$  are significantly higher than  $x_3$ , and  $l_3$  is shorter than a threshold, then we assume that  $C_3$  is a false branch.

Some repeat read errors and small-scale sequence variation between approximate repeats or alleles result in some of the branches merging back to the trunk of the de Bruijn graph. We call such structures 'bubbles', and remove them during the assembly. Since they may represent real albeit alternative sequence at that location, we preserve the information they carry by recording them in a special log file, along with the variant we leave in the assembly contig and their coverage levels. These log entries are later used to postulate effects of allelic variations on expression levels.

After the false branches are culled and bubbles removed, unambiguously linear paths along the de Bruijn graph are connected to form the single end tag assembly (SET) contigs. The branching information is also recorded for the subsequent paired end tag assembly (PET) stage and further analysis. At the SET stage, every  $k$ -mer represented in the assembly contigs has a unique occurrence. Using that information, we apply a streamlined read-to-assembly alignment routine. We use the aligned read pairs to (i) infer read distance distributions between pairs in libraries that form our read set, and (ii) identify contigs that are in a certain neighborhood defined by these distributions.



**Fig. 1.** Excerpt from an ABySS-Explorer view, where edges represent contigs and nodes represent common  $k-1$ -mers between adjacent contigs. The labels correspond to SET contig IDs. Contig length and coverage are indicated by the length and the thickness of the edges, respectively. Arrows and edge arc shape indicate the direction of contigs and the polarity of the nodes distinguish reverse complements of common  $k-1$ -mers between adjacent contigs.

The adjacency and the neighborhood information are used by the PET routine to merge SET contigs connected by pairs unambiguously, while keeping the list of the merged SET contigs (or the *pedigree* information) in the FASTA header for backtracking.

The adjacency, neighborhood and pedigree information, along with the contig coverage information are also used by our assembly visualization tool, ABySS-Explorer. Figure 1 shows the ABySS-Explorer representation of some SET contigs in a neighborhood. Note that both the edges and the nodes are polarized in accordance with the directionality of the contigs and the  $k-1$ -mer overlaps between them, respectively. In the interactive view, when a user double-clicks on a contig, its direction and node connection polarizations flip to reflect the reverse complement. Paired end tags often resolve paths along SET contigs and they are subsequently merged in the PET stage. We indicate such merged contigs by dark gray paths in the viewer, when one of the contigs contributing to a merge is selected.

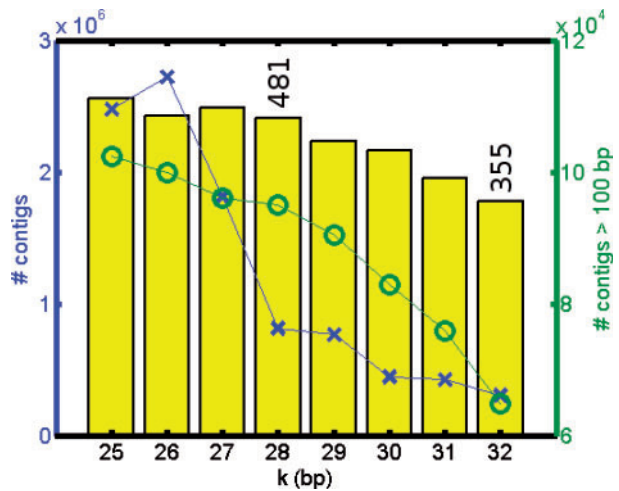
The ABySS-Explorer representation encodes additional information including contig coverage (indicated by the edge thickness) and contig length (indicated by the edge lengths). A wave representation is used to indicate contig length such that a single oscillation corresponds to a user-defined number of nucleotides. A long contig results in packed oscillations, which obscure the arrowhead indicating its direction. To resolve this ambiguity, the *envelope* of the oscillations outlines a leaf-like shape with the thicker stem of the shape marking the start of the contig and the thinner tip pointing to its end. For example, contig 383 936 is a 627 bp long contig, which is much longer than the shortest contig 297 333 (29 bp), but its direction is still evident from its shape, with the thinner tip pointing to the right.

We performed a parameter search for assembly optimization by varying  $k$  values in the range between 25 bp and 32 bp for the SET stage. Figure 2 shows some key statistics of our assemblies as a function of  $k$ . We picked the best assembly to be that for  $k=28$ , as the number of contigs drops significantly between  $k=27$  and  $k=28$ , while the number of contigs 100 bp or longer do not increase, which indicate a substantial improvement in contiguity. Beyond  $k=28$ , the number of contigs in both categories keeps decreasing, but so does the assembly N50.

### 3 RESULTS

Each assembly with a distinct  $k$ -value is performed on a cluster of 20 compute nodes, each with 2 GB of memory. Since the number of possible  $k$ -mers changes with the parameter  $k$ , the runtime of the SET stage varies as a function of  $k$ . For the optimal assembly parameter of  $k=28$ , the first stage of the assembly takes about 123 min for 194 million reads. The runtime of the second stage highly depends on the results of the first stage, most importantly, on the number of contigs generated. In this case, it takes about 7.6 h to assemble 812 300 SET contigs into 764 365 PET contigs on a single workstation. Key statistics on this optimal assembly are presented in Table 1.

With this assembly, we reconstructed over 30 million bases of sequence. Following alignment of these PET contigs to the reference human genome, hg18, using BLAT (Kent, 2002) we observed that



**Fig. 2.** Comparison of SET assemblies with  $k$  in [25, 32]. The blue crosses show the number of contigs (left axis), the green circles show the number of those that are 100 bp or longer (right axis). The bars indicate the assembly N50 on an arbitrary scale. The N50 values of  $k=28$  and  $k=32$  assemblies are as indicated.

**Table 1.** Assembly statistics of the follicular lymphoma transcriptome for the SET and PET stages with  $k=28$

$k=28$ assembly	SET	PET
Number of contigs	812 300	764 365
Number of contigs $\geq 100$ bp	95 080	66 921
N50 (bp)	481	1114
Max (bp)	7386	10 951
Total (Mb)	29	30

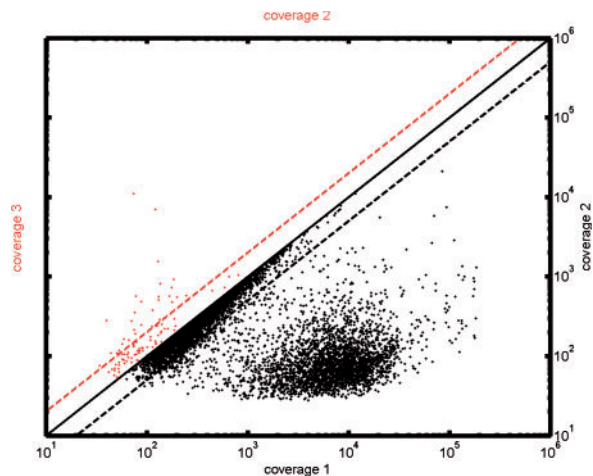
Assembly N50 values and total reconstruction are reported for contigs of length 100 bp or longer.

92.5% of bases are in contigs that overlap a gene, as defined by the UCSC gene list (Hsu *et al.*, 2006). The remaining 7.5% of bases are entirely within intronic or intergenic regions. Only 1317 bases, or a mere 0.004% were found to not align to hg18, and these primarily aligned to bacterial genome sequences, suggesting sample contamination.

In the SET stage, we report 15 831 bubbles, of which 15 651 (98.9%) contain two variants, and the remaining 180 (1.1%) contain three variants. We aligned these bubble sequences to hg18 using BLAT, requiring 95% or better identity. Out of 15 831 branches, 6832 (43.2%) contain at least two variants that align uniquely to the same coordinates, indicating heterozygous (mostly single) nucleotide variants. A small fraction of the total (93) but over half of the three variant branches hit uniquely to the same coordinate, suggesting a mixed cell population, typical for tumor tissue samples.

Using the fold coverage of the bubble branches as a surrogate for their expression levels, we analyzed the relative expression levels between alleles. When we compared the primary allele (highest coverage) to the secondary allele (second highest coverage), we found that the expression level difference was less than 1-fold for 3362 (49.2%) of the branches. Comparing the expression level change between the secondary and the tertiary alleles where





**Fig. 3.** Coverage level comparison between the primary, secondary and tertiary branches in assembly bubbles. The lower (upper) triangular region shows a comparison between the primary and secondary (secondary and tertiary) coverage levels, represented by black (red) markers.

available, we found that similar expression was more common. For 63 (67.7%) of them change in the expression levels was less than 1-fold.

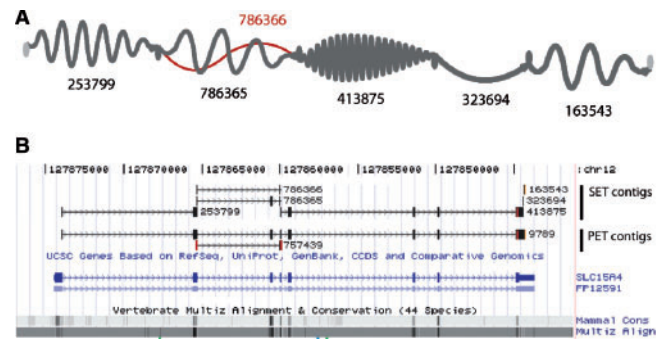
The scatter plot in Figure 3 shows a comparison between the coverage levels of the alleles. The lower-right triangle portion of the plot (below the solid line) shows the comparison between the coverage levels of the first and the second alleles (black markers), and the upper-left triangle portion of the plot shows the same for the second and the third alleles (red markers). In this plot, we can easily see two distinct populations of expression pairs. A large fraction of pairs cluster near equal coverage line (solid line in Fig. 3), and a group of coverage level pairs indicating significant (over 1-fold) expression level differences (below and above the dashed black and red lines for the lower and upper triangular regions, respectively).

When we investigate the coding status of the reported bubbles in coding regions, and co-relate that with the expression level differences, we observe that those with significant differences are more likely to have non-synonymous changes (78.0%) compared to those with less than one-fold difference (40.7%).

At the SET stage, alternative splicing events, as well as genomic events such as heterozygous indels and sequence similarity may result in the assembly of two parallel contigs that lie between two neighboring contigs. These two parallel contigs indicate that two alternative sequences were observed between these neighboring contigs. The shorter of the two parallel contigs will be composed of exactly  $2(k-1)$  bp and represents the junction of the two neighboring contigs, with  $k-1$  bp coming from each neighbor. The longer contig represents the additional sequence, such as a retained exon or intron.

In our transcriptome assembly, we have identified 888 contigs with  $2(k-1)$  bp sequences. We aligned these contigs to the Ensembl transcriptome reference as well as to the reference genome, hg18, using exonerate (Slater and Birney, 2005). We required that these alignments be ungapped, with 90% identity. We observed that:

- 287 (32%) of these contigs align only to the Ensembl transcriptome and represent known exon/exon junctions;



**Fig. 4.** (A) ABySS-Explorer and (B) UCSC viewer representations of the assembly of SLC15A4 and FP12591 genes, illustrating the expression of a skipped exon of SLC15A4 gene is higher compared to FP12591 gene.

- 118 (13%) align to both transcriptome and genome references and represent exon sequences;
- 27 (3%) align only to hg18 and represent reference sequence not in annotated exons; and
- 456 (51%) do not align, and represent putative novel events.

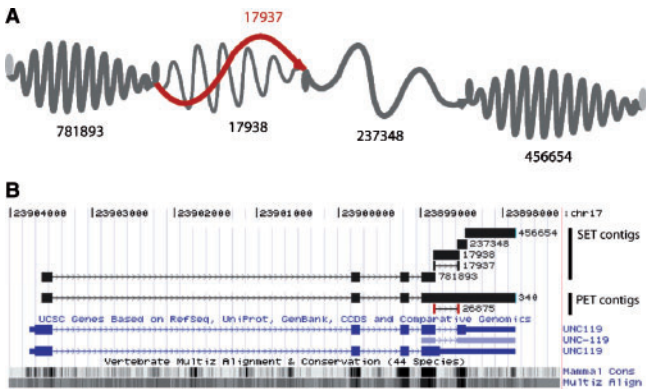
Contigs representing putative novel events were then split into two  $k-1$  bp halves, and aligned to hg18 using MAQ (Li *et al.*, 2008), allowing up to two mismatches. We found that 200 of these contigs aligned with 178 contigs having gap sizes of 50 bp or more, indicating either putative novel events or larger heterozygous indels. Only 22 contigs aligned with gaps of 6 bp or less, which we classified as putative heterozygous indels. Two of these aligned to intronic regions, and the remaining 20 aligned to 3'UTRs; hence none of them aligned to a coding region. Our list of novel events needs further investigation for biological significance and comparison with databases of known variations.

Below we present four examples, three of which come from the list of  $2(k-1)$  bp contigs and show a skipped exon, a retained intron, and an alternative splicing event. The fourth example shows a novel transcript we identified through gapped alignment of our PET contigs to the hg18 reference genome.

### 3.1 Skipped exon

Figure 4 shows an example of an exon skipping event. The path through SET contigs {253 799, 786 365, 413 875, 323 694, 163 543} in the ABySS-Explorer view (Fig. 4A) reconstructs the transcript of the gene SLC15A4, and is indicated by dark gray. Contig 786 366 (shown in red) is  $2(k-1)=54$  bp long, and defines an alternative path between contigs 253 799 and 413 875. As we have described above, this indicates the presence of multiple isoforms. Indeed, if we examine the UCSC browser view (Karolchik *et al.*, 2007) of the region described by the alignment of our contigs (Fig. 4B), we see that this corresponds to the skipping of the third exon of SLC15A4 gene. This alternative transcript is annotated as FP12591 gene.

The UCSC view (Fig. 4B) shows tracks for SET and PET contigs on top. The contig numbers in the SET contigs track correspond to those displayed in the ABySS-Explorer view. The PET contigs track shows the reconstructed SLC15A4 gene (PET contig 9789), and the exon skipping event (PET contig 757 439, illustrated in red). Below the tracks of genes are the conservation and multi-alignment tracks, indicating strong signals at the exon coordinates. ABySS-Explorer



**Fig. 5.** (A) ABySS-Explorer and (B) UCSC viewer representations of the assembly of UNC119 gene, illustrating a retained intron.

view suggests that the relative expression level of SLC15A4 gene is higher compared to FP12591 gene.

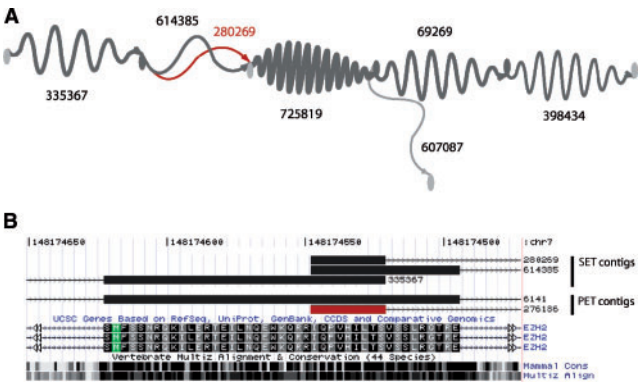
**3.2 Retained intron**

For retained introns, we expect a similar graph topology to that of the exon skipping events. Consider Figure 5A for example, which illustrates the ABySS-Explorer view of a subgraph with SET contigs aligning to UNC119 gene. The alternative paths through contigs 17937 (red) and 17938 (gray) correspond to observed alternative events as annotated in the UCSC browser ideograms in Figure 5B. The first and second UNC119 annotations show a gap that corresponds to the reconstruction indicated by SET contig 17937 (renumbered as 26875 during the PET stage and is illustrated in red in the PET contigs track). The third annotation of UNC119 indicates that this ‘intronic’ sequence corresponds to an untranslated region (UTR). Although PET contig 340 reconstructs the longer variant, the coverage information shown in the ABySS-Explorer view indicates that the expression level of the transcript with the retained intron is lower, and the SET path {781 893, 17 937, 237 348, 456 654} is the predominant variant.

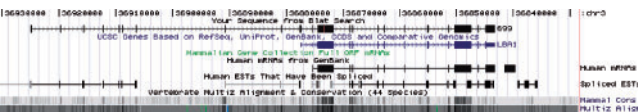
**3.3 Alternative 5’ splicing**

Figure 6 illustrates an example where we detect an alternative 5’ splicing event, represented in the ABySS-Explorer view by two SET contigs, 614 385 (gray) for the normal splicing event and 280 269 (red) for the alternative splicing event. The UCSC viewer representation indicates that this alternative splicing removes 27 bp or 9 amino acids from the 3’ end of an exon. When we investigate the raw data for this event, we see that eight reads align reasonably well to the normal transcript with up to five mismatches. Thus, if one were to analyze the data by alignment of short reads to the known transcriptome, allowing a certain number of mismatches and no indels, these are the reads which could identify this alternative junction, as other reads that span it would not be mappable. However, these reads and reads that tile across the alternative junction assemble easily to reconstruct this event, and the alignment of contigs with longer lengths enables its recovery.

The topology of this event is similar to those of the previous examples. Length of the contig with the shorter splicing (280 269 in SET, and 276 186 in PET assemblies) is  $2(k - 1) = 54$  bp, with  $k - 1$  bp of it anchored to SET contigs on either side.



**Fig. 6.** (A) ABySS-Explorer and (B) UCSC viewer representations of the assembly of EZH2 gene, illustrating a novel alternative splicing event.



**Fig. 7.** UCSC viewer representations of the assembly of LBA1 gene, illustrating a putative novel transcript.

Transcriptome data is typically contaminated by genomic data in small quantity, which nonetheless can be sufficient in some places to assemble short contigs. Reads derived from genomic sequence that cover an exon/intron boundary will assemble into short but low-coverage tips that branch from the portion of the graph that represents transcriptome sequence. Figure 6A shows one such tip (SET contig 607 087, illustrated in light gray) in the ABySS-Explorer view. The short length of the curve indicates the short length of the contig, and the narrow width of the curve indicates its low coverage.

**3.4 Novel transcript**

Graph topology of novel transcripts does not follow the previous cases of skipped exons, retained introns or other splice variant events. They are identified by aligning contigs to the reference genome, and the contig lengths are instrumental in their discovery, as longer contigs have improved alignment specificity. Consequently, the example we provide for this case does not require the graph representation of the SET stage, but the final contig of the PET stage.

PET contig 699 in Figure 7 represents a sequence over 10 kb in length that maps 100% to the reference genome with one mismatch in the first long exon, and mostly follows the annotated exon structure of LBA1 gene, but appears to have an alternative transcription start and additional exons. Interestingly, the extra nine upstream exons are all annotated as spliced ESTs, albeit in two disconnected groups, one of which points to an even further transcription start. Also of note here is the concordance between the estimated exons and the strong mammalian conservation signal depicted in the bottom ideogram as a gray-scale signal.

This event would be difficult, if not impossible to reconstruct through an analysis based on alignment of short reads. Although they would potentially yield hits to these estimated exons, they would not

only have lower coverage on the exon boundaries due to potential mismatches in reads that span junctions, but also would be difficult to discern from genomic sequence contamination.

## 4 DISCUSSION

To our knowledge, this is the first demonstration of *de novo* assembly of experimental human transcriptome data from a short read sequencing platform, and its results are presented in a form that allows high throughput analysis for comparative genomics. The preliminary analysis and the examples presented illustrate the utility of ABySS for *de novo* assembly of transcriptomes. Assembled contigs, their topological relationships and information about putative heterozygous nucleotide variations offer intriguing leads for the analysis of genomic events and their effects on biological functions. With the reported results in this manuscript we are only scratching the surface in analyzing the results of our transcriptome assembly. We are currently developing high throughput analysis methods based on the properties of the anecdotal evidence we present here, and will report a more thorough account of our results in due time.

ABySS-Explorer is an important component of this tools set that provides an interactive interface for visualizing ABySS assembly graphs. As illustrated by the sample events presented here, this graph view emphasizes alternative paths through an assembled region, making alternative events easy to identify. It captures coverage information as edge thickness and facilitates judgments of relative expression levels. Finally, this tool can readily display additional data not used in the assembly, such as annotations or additional paired end data, which are useful in inferring the nature of the putative isoforms presented in the graph.

Thus the set of tools we introduce in this article, coupled with alignment methods, will help researchers interrogate high throughput data from whole transcriptome shotgun sequencing experiments using short read sequencing platforms. As an enabling technology, these tools will make it possible to formulate new hypotheses for testing.

## ACKNOWLEDGEMENTS

The authors would like to thank Kim Wong, Matthew Field, Andrew Mungall and Karen Mungall for helpful discussions, and Pawan Pandoh, Helen McDonald and Jennifer Asano for their contributions in WTSS library construction. SJMJ is a senior scholar of the Michael Smith Foundation for Health Research.

**Funding:** Genome Canada, Genome British Columbia and the British Columbia Cancer Foundation; National Cancer Institute of

Canada Terry Fox Program Project award (# 016003) (to D.E.H., J.M.C., R.D.G. and M.A.M.).

**Conflict of Interest:** none declared.

## REFERENCES

- Bentley,D.R. *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53–59.
- Butler,J. *et al.* (2008) ALLPATHS: *de novo* assembly of whole-genome shotgun micreads. *Genome Res.*, **18**, 810–820.
- Chaisson,M.J. and Pevzner,P.A. (2008) Short read fragment assembly of bacterial genomes. *Genome Res.*, **18**, 324–330.
- de Bruijn,N.G. (1946). A combinatorial problem. *Koninklijke Nederlandse Akademie v. Wetenschappen*, **49**, 758–764.
- Dohm,J.C. *et al.* (2007) SHARCGS, a fast and highly accurate short-read assembly algorithm for *de novo* genomic sequencing. *Genome Res.*, **11**, 1697–1706.
- Farrer,R.A. *et al.* (2009) *De novo* assembly of the *Pseudomonas syringae* pv. *syringae* B728a genome using Illumina/Solexa short sequence reads. *FEMS Microbiol. Lett.*, **1**, 103–111.
- Fullwood,M.J. *et al.* (2009) Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. *Genome Res.*, **4**, 521–532.
- Hernandez,D. *et al.* (2008) *De novo* bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome Res.*, **5**, 802–809.
- Hsu,F. *et al.* (2006) The UCSC known genes. *Bioinformatics*, **9**, 1036–1046.
- Jackson,B.G. *et al.* (2009) Parallel short sequence assembly of transcriptomes. *BMC Bioinform.*, **10**, S1–S14.
- Karolchik,D. *et al.* (2007) The UCSC Genome Browser. *Curr. Protoc. Bioinform.*, Chapter 1, Unit 1.4.
- Kent,W.J. (2002) BLAT – the BLAST-like alignment tool. *Genome Res.*, **4**, 656–664.
- Kozarewa,I. *et al.* (2009) Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat. Methods*, **4**, 291–295.
- Li,H. *et al.* (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, **11**, 1851–8.
- Morin,R. *et al.* (2008) Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *Biotechniques*, **45**, 81–94.
- Nielsen,C.B. *et al.* (2009) ABySS-Explorer: visualizing genome sequence assemblies. *IEEE Trans. Vis. Comp. Graphics* (in revision).
- Ossowski,S. *et al.* (2008) Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Res.*, **12**, 2024–2033.
- Pevzner,P.A. *et al.* (2001) An Eulerian path approach to DNA fragment assembly. *Proc. Natl Acad. Sci. USA*, **98**, 9748–9753.
- Salzberg,S.L. *et al.* (2008) Gene-boosted assembly of a novel bacterial genome from very short reads. *PLoS Comput Biol.*, **9**, e1000186.
- Simpson,J.T. *et al.* (2009) ABySS: a parallel assembler for short read sequence data. *Genome Res.*, **19**, 1117–1123.
- Slater,G.S. and Birney,E. (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinform.*, **6**, 31
- Wang,E.T. *et al.* (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–476.
- Warren,R.L. *et al.* (2009) Profiling model T-cell metagenomes with short reads. *Bioinformatics*, **4**, 458–464.
- Yassour,M. *et al.* (2009) Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. *Proc. Natl Acad. Sci. USA*, **9**, 3264–3269.
- Zerbino,D.R. and Birney,E. (2008) Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.*, **18**, 821–829.