BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 17 2011, pages 2325–2329
doi:10.1093/bioinformatics/btr355

Genome analysis

Advance Access publication June 21, 2011

Identification of novel transcripts in annotated genomes using RNA-Seq

Adam Roberts[1], Harold Pimentel[1], Cole Trapnell[2]* and Lior Pachter[1,3,4,*]

[1]Department of Computer Science, UC Berkeley, Berkeley, CA, [2]Department of Stem Cell and Regenerative Biology, Harvard University and The Broad Institute of MIT and Harvard, [3]Department of Mathematics and [4]Department of and Molecular and Cell Biology, UC Berkeley, Berkeley, CA, USA

Associate Editor: Alex Bateman

ABSTRACT

Summary: We describe a new 'reference annotation based transcript assembly' problem for RNA-Seq data that involves assembling novel transcripts in the context of an existing annotation. This problem arises in the analysis of expression in model organisms, where it is desirable to leverage existing annotations for discovering novel transcripts. We present an algorithm for reference annotation-based transcript assembly and show how it can be used to rapidly investigate novel transcripts revealed by RNA-Seq in comparison with a reference annotation.

Availability: The methods described in this article are implemented in the Cufflinks suite of software for RNA-Seq, freely available from http://bio.math.berkeley.edu/cufflinks. The software is released under the BOOST license.

Contact: cole@broadinstitute.org; lpachter@math.berkeley.edu

Supplementary Information: Supplementary data are available at Bioinformatics online.

Received on March 2, 2011; revised on May 5, 2011; accepted on June 8, 2011

1 INTRODUCTION

Whole transcriptome sequencing, known as RNA-Seq (Cloonan *et al.*, 2009; Mortazavi *et al.*, 2008; Nagalakshmi *et al.*, 2008), leverages high-throughput sequencing technology to investigate the RNA content from a sample via the sequencing of cDNA. This technology has been the focus of numerous recent studies that demonstrate high resolution and accuracy in transcript abundance estimation, and it is being heralded as a possible replacement for microarray-based gene expression technology. This exciting development has partially eclipsed another important application of RNA-Seq: the improvement of existing genome annotations (Denoued *et al.*, 2008) and even the possibility of complete *de novo* genome annotation based on multiple experiments, as is being demonstrated by the RGASP consortium (http://www.gencodegenes.org/rgasp/).

In the context of genome annotation, RNA-Seq reads can be viewed as next-generation expressed sequence tags (ESTs) (Adams *et al.*, 1991). RNA-Seq offers the promise of rapid, comprehensive discovery of novel genes and transcripts in considerably less time and at lower cost than ESTs from conventional Sanger sequencing.

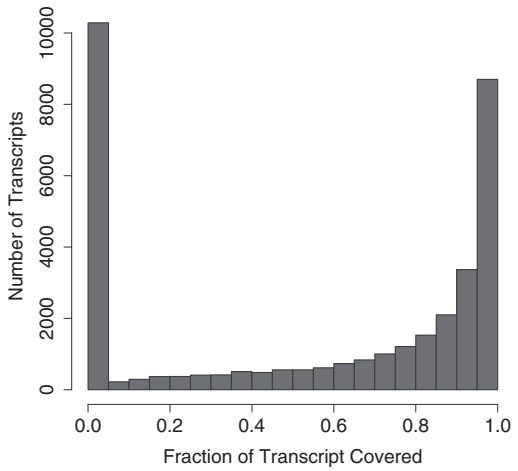

Fig. 1. Histogram showing the percent of bases covered in annotated reference transcripts. The low-coverage transcripts would not be properly assembled using sequencing reads alone. In genomes with extensive annotations, there is clearly room for improvement by taking advantage of the information contained in the annotated transcripts when assembling and calculating expression values.

However, a problem with using RNA-Seq for annotation is that genes that are expressed at a low level will be represented by few reads and may be only partially covered. To illustrate this point, we show in Figure 1 the coverage of RefSeq annotated transcripts in a typical RNA-Seq experiment (see Section 3 for details of the experiment). The histogram shows that 29.1% of the transcripts are not covered at all, which will happen because they are not expressed or are at levels undetectable in the experiment. The remaining transcripts have varying levels of coverage (based on their expression level and length), and although many transcripts are completely covered, we found that of the transcripts with partial coverage, 64.4% are incompletely covered (<95% coverage). This means that naïve assembly methods will fail to reconstruct the majority of full-length transcripts. One approach to predicting complete transcripts from partial coverage is to incorporate RNA-Seq data in a gene finding system. Such approaches have been successful, but they rely on models of genes structures that may limit predictions based on prior biases. Other approaches, such as Scripture (Guttman *et al.*, 2010) may be able to close gaps in coverage, but fail to incorporate existing annotation information

*To whom correspondence should be addressed.

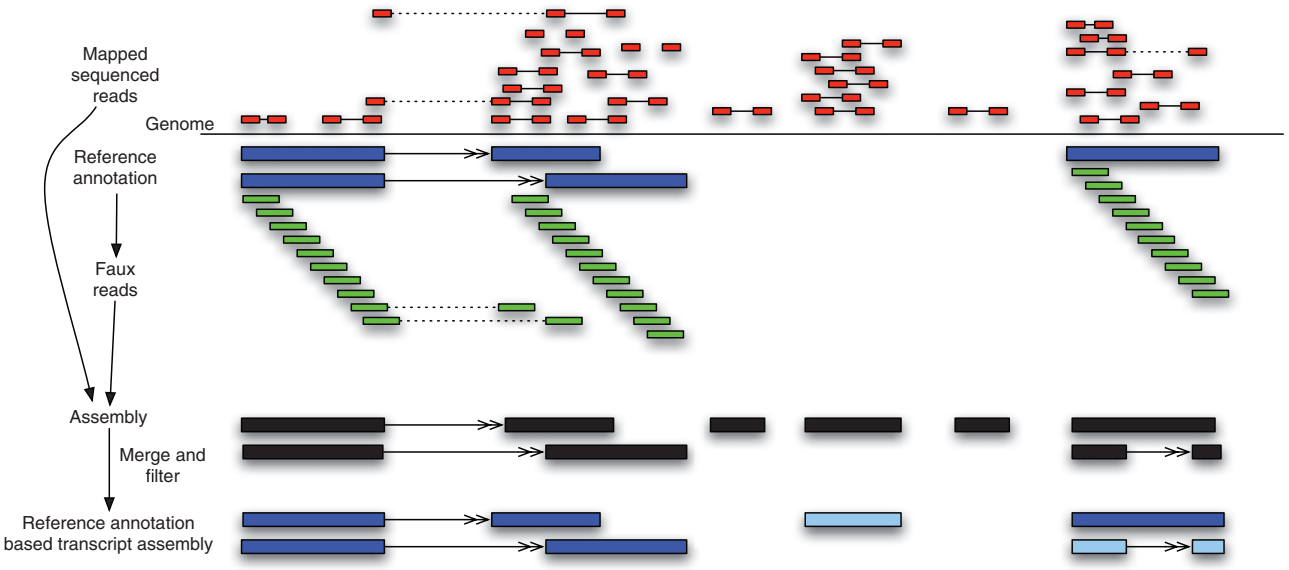

Fig. 2. An overview of our RABT assembly method. First paired-end reads (mates shown connected by solid lines) are mapped to the genome using a spliced read mapper that can map reads across junctions (shown in dotted lines). The reference annotation (blue) is used to generate faux-read alignments that tile the transcripts (green). The faux-read alignments are used together with the spliced read alignments to generate a reference genome-based assembly (black). This assembly is merged with the reference annotation, and 'noisy' read mappings are filtered resulting in all reference annotation transcripts in the output (blue) as well as novel transcripts (light blue).

during prediction. In the case of model organisms where gene annotations are based on years of effort, large consortia have combined state of the art computational predictions (Guigó *et al.*, 2006) with careful human curated annotations (Ashurst *et al.*, 2005). We have organized existing assembly strategies into three categories, and as far as we are aware, none of the programs in any of these categories can explicitly use existing annotations during assembly:

- *De novo* transcript assembly—the direct assembly of sequenced reads into transcripts without mapping to a reference genome. Examples include Simpson *et al.* (2009).
- Genome reference-based transcript assembly—assembly of transcripts by first mapping to a reference genome. These methods build on previous work using ESTs (Haas *et al.*, 2003; Heber *et al.*, 2002) or mapped pyrosequencing reads (Eriksson *et al.*, 2008). Examples include Guttman *et al.* (2010); Trapnell *et al.* (2010).
- RNA-Seq assisted protein coding gene annotation—the incorporation of read alignments as supporting evidence for *ab initio* gene finding algorithms. Examples include Allen and Salzberg (2005); Schweikert *et al.* (2009); Stanke and Waack (2003).

We address the need for a reference annotation-based assembler by developing a novel approach through modification of an existing assembler that we term *reference annotation based transcript assembly* (RABT assembly). We adopt the approach of (Trapnell *et al.*, 2010) which is to identify transcripts only based on read alignments (i.e. without regard to prior information about protein coding gene structure), and we employ the parsimony approach to find the fewest number of transcripts explaining the data (in this case, the aligned sequenced reads together with the reference annotation). A key feature of our approach is that it rapidly identifies novel transcripts (with respect to the reference annotation).

2 METHODS

Our RABT assembly method builds upon the Cufflinks assembler (Trapnell *et al.*, 2010) that determines the minimum number of transcripts needed to explain sets of reads aligned to a genome. The algorithm is based on finding a minimum path decomposition of a directed acyclic overlap graph constructed from the reads, and is efficient thanks to a reduction of the computational problem to graph matching. For details see the Supplementary Material in Trapnell *et al.* (2010).

We used the default parameters on the Cufflinks assembler, which include the removal of likely intronic reads (due to intron retention) as well as assembled transfrags with very low estimated abundance relative to other isoforms of the same gene. More details on these parameters can be found at the Cufflinks web site (http://bio.math.berkeley.edu/cufflinks). In order to incorporate a reference annotation into the assembly algorithm, we adopted the following approaches (see Fig. 2 for an overview):

(1) faux-reads were generated from reference transcripts in order to capture features in the reference that could be missing in the sequencing data due to low coverage. The faux-reads 'tiled' the reference transcripts so that every reference transcript position was covered by (multiple) reads;

(2) a parsimonious assembly was constructed by the original Cufflinks assembler (Trapnell *et al.*, 2010) using both the sequenced and faux-reads. This assembly contained the fewest number of transfrags that explained both the reference transcripts and the sequenced reads; and

(3) the reference transcripts were merged with the assembled transfrags and the resulting set was filtered to remove repeats.

Table 1. Results for two different versions of assembly on the MAQC Human Brain Reference

Human output set	No. of genes	No. of transfrags	Avg transfrag length	Isoforms per gene
Reference annotation	20 960	34 033	3127	1.62
Cufflinks assembly	43 757	58 030	1671	1.19
Cufflinks assembly (novel only)	15 483	32 738	1918	–
RABT assembly	36 494	70 241	2766	1.92
RABT assembly (novel only)	15 504	36 208	2422	–

The Cufflinks assembly used the original Cufflinks assembler, while the RABT assembly was generated using the method described in this article. Note that we judged an assembled transfrag to be novel if it was not removed by the filtering step described in Section 2. To compare just the novel portions of the assemblies, we also ran the Cufflinks assembly through the filtering step. Since it did not use faux-reads, multiple transfrags were often discarded from the Cufflinks assembly for a single reference transcript (2.03 on average). A gene is novel if it contains only novel transfrags. The average length of assembled transfrags from our new method are much more similar to those in the reference annotation than those produced by the original assembler. Isoforms per gene is undefined for the 'novel only' rows.

In the first step, the assembler generated faux- (aligned) reads of length 405 bp at 15 bp intervals along the reference transcripts, except within 405 bp of either end of each transcript, in which case the length of the generated faux-reads was the distance to the end. We chose these lengths and intervals in order to connect all potential gaps in coverage, while minimizing the number of unnecessary reads (and their effects on running time) and reducing the creation of assembled transfrags produced purely by mixing parts of different reference transcripts. In general, these parameters should be adjusted based on the properties of the transcriptome and the read lengths (see Section 4). In the second step, these reads were merged with the (aligned) sequenced reads for assembly. We note that our idea of merging faux-reads with sequenced reads was motivated by, and closely related to, the approach used to inform the Celera whole genome shotgun assembly with the human genome project assembly in Venter *et al.* (2001).

The set of transfrags generated in the second step was then compared with the reference transcripts to remove transfrags that were approximately equivalent to the whole or a portion of a reference transcript. Transfrags were discarded if a reference transcript was found such that all of the following criteria were met:

(1) Its 5′ endpoint was contained in the reference transcript.
(2) Its 3′ end point extended no more than 600 bp outside of the reference transcript, and this region contained no introns.
(3) It contained no introns that were not also in the reference transcript.
(4) It contained all introns in the reference transcript that fully lay within its boundaries.
(5) Its endpoints extended no more than the mean fragment length into the intron of the reference transcript.

If only the first criterion failed, and there were no additional introns in the region extending beyond the 5′ end of the reference transcript, the reference transcript was extended to match the 5′ end of the transfrag, and the transfrag was discarded. Again, the parameters in this step were chosen based on properties of the reads as well as the organism. We allowed 3′ overhang to account for errors in assembly due to imperfect transcription termination. Overlap with introns was allowed due to the common appearance of single fragment ends inside of annotated introns, perhaps due to mismapping or imperfect poly-A selection.

3 RESULTS

We ran our RABT assembler on a human RNA-Seq dataset sequenced from the Ambion Human Brain Reference by Illumina (accession no. SRA012427), and compared it to a reference genome assembly produced with Cufflinks (Trapnell *et al.*, 2010) on the same data. The reference annotation used was NCBI36/hg18 RefGene, downloaded from the UCSC genome browser database (http://genome.ucsc.edu). The reads were mapped using TopHat 1.2.0 with the splice junctions in the annotation provided. Both assembly algorithms required <1 h on an 8 core 2.26 GHz computer.

We focus our analysis on novel transcripts within annotated genes. All other transcripts are purely a product of the original Cufflinks assembler (Trapnell *et al.*, 2010) and their properties/validation is described in that article. Furthermore, we do not validate splice junctions as these are found by TopHat and validated in (Trapnell *et al.*, 2009). It is important to note that although we employed those tools, the RABT approach could be used with other mappers/assemblers.

Table 1 shows the comparison of the Cufflinks assembler with the RABT assembler. An average of 2.03 transfrags output by the Cufflinks assembler were found to be partial reference transcripts according to the filtering method described above. Furthermore, the transfrags assembled by RABT are 751 bp longer than those output by the Cufflinks assembler. This difference is due not only to the discarding of transfrags made from portions of known transcripts but also to the use of faux-reads generated from the reference annotation to predict complete novel isoforms of existing genes. For an example, see Figure 3.

We investigated the extent of novel isoform discovery in known genes and found an average of 0.95 new isoforms per gene. Such new annotated isoforms include novel junctions found by TopHat or novel combinations of junctions discovered by sequencing reads. To validate these novel transcripts, we calculated the phastCons44way conservation track in the UCSC Genome Browser. We calculated that the novel transcripts actually had higher average conservation probability than the reference transcripts (0.13 versus 0.12).

Finally, we note that the use of faux-reads led us to slightly increase the number of novel assembled transcripts output by the RABT assembler than we see after filtering the original Cufflinks method for partial reference transcripts (see Table 1). This is likely due to rare cases where a single novel feature led to multiple transfrags being assembled in combination with pieces of reference transcripts. To test the extent of such 'false positives', we ran the RABT assembler excluding the actual sequenced reads, i.e. using only faux-reads. Using this method, 1032 transcripts were assembled and falsely labeled 'novel'. Thus, we estimate that <3% of the transcripts defined as novel by the RABT assembler in the experiment above are false positives.

To show that our method is applicable to different organisms, we also repeated these experiments on *Drosophila melanogaster* using RNA-Seq data from the first embryo time point of the

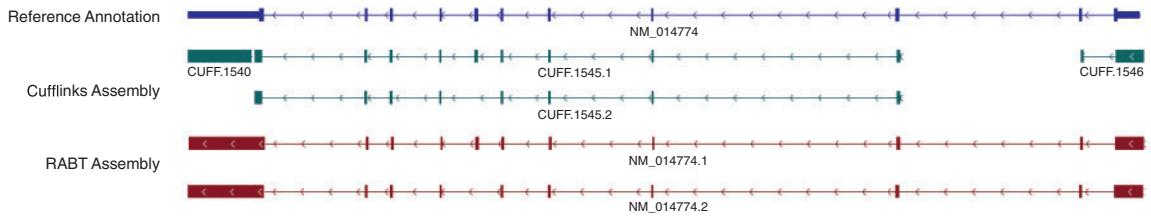

Fig. 3. Comparison of assembler output for an example gene. Lack of sequencing coverage in the UTR and across one splice junction caused the Cufflinks assembler (teal) to output three transfrags that match the reference (blue) and a fourth that contains a novel splice junction. The RABT assembler output (red) includes both the reference transcript (NM_014774.1) and a novel isoform (NM_014774.2) that is assembled from a combination of sequencing reads, which reveal the novel junction, and faux-reads, which connect the three sections to form a single transcript. Note that even with the addition of the reference transcript, the total number of transfrags output by the assembler has been reduced for this locus, and the transfrag lengths have increased.

Table 2. Results for two different versions of assembly on the first *D.melanogaster* embryo time-point from (Graveley *et al.*, 2010)

Drosophila melanogaster output set	No. of genes	No. of transfrags	Avg transfrag length	Isoforms per gene
Reference annotation	13 302	20 715	1629	1.56
Cufflinks assembly	7167	8701	2334	1.21
Cufflinks assembly (novel only)	350	3205	2741	–
RABT assembly	13 634	23 913	1815	1.75
RABT assembly (novel only)	332	3018	2719	–

The categories can be interpreted in the same manner as Table 1. These results show that the method also produces improved assemblies in fly.

modENCODE dataset (Graveley *et al.*, 2010). Our reference was the r5.22 annotation from FlyBase (http://flybase.org/), which did not make use of this data. Again, we found that the novel transfrags had average conservation probabilities similar to known transcripts (0.49 versus 0.47). Further results for this experiment are found in Table 2.

4 DISCUSSION

RNA-Seq is being increasingly adopted as the technology of choice for gene expression studies (Blow, 2009), and with large numbers of experiments producing partial transcripts of genes, it is expected that there will be rapid progress in the coming years in annotating genomes. As more complete genome annotations are produced, it is increasingly desirable to include them in analyses rather than assembling transcripts 'from scratch' with every new experiment. The reference annotation-based transcript assembly approach we have introduced addresses this problem, and allows for the incremental improvement of annotations with RNA-Seq experiments. It is also convenient in that novel genes and transcripts (with respect to an existing annotation) are easily extracted from the output of our assembler.

It is important to note that accurate genome annotation is crucial for accurate gene expression estimation. In previous work, it has been shown that incomplete annotations can bias gene expression estimates (Jiang and Wong, 2009; Trapnell *et al.*, 2010). The current practice of using RNA-Seq to estimate expression using known annotations when they are available is therefore liable to yield inaccurate results, especially in cases where genes have multiple isoforms, some of which may not yet be annotated. We, therefore, believe that RABT assembly is essential until annotations are improved and completed.

A key feature of RABT is that it is a 'pure' assembler. This means that it does not utilize information about the structure and content of coding genes or other external input (e.g. ESTs) during the assembly. We believe this is a feature (rather than a weakness of the method) because it means that RABT can assemble non-coding RNA transcripts. It is an interesting problem to extend RABT to allow for other external input, and we believe that similar approaches (based on faux-reads) may be fruitful. Along with this method, we present several open problems in RABT assembly. For example, deciding the optimal length and spacing of tiling reads that will connect transfrags while minimizing reorganizations of annotated transcripts is non-trivial and will vary depending on the properties of the experiment as well as the organism under investigation. The same is true for the parameters used in matching assembled transcripts with those in the annotation for filtering. While we believe we have chosen good parameters for this experiment, we do not believe there to be a 'one-size fits all' approach and it should be interesting to develop a statistically sound approach to automatically determine the best tiling method. We also note that using more complete annotations as references with RABT can lead to a larger number of false positives as portions of annotated transcripts can be re-assembled in new combinations. For example, we estimate through faux-read only assembly that using the more complete UCSC annotation as opposed to the RefSeq annotation would lead to ∼5 times the number of false positives. Therefore, we recommend this method for use on organisms where deep annotations do not already exist.

Continuing improvements in RNA-Seq technology will eventually result in the ability to sequence complete transcripts using long reads and fragments. Furthermore, large-scale surveys of transcripts in multiple developmental stages and tissues should be able to yield complete annotations of genomes solely based on RNA-Seq. However, until that time, it is imperative that genome

annotations be incrementally improved by building on, rather than discarding, previous work.

Funding: AR was funded in part by an NSF graduate fellowship.

Conflict of Interest: none declared.

REFERENCES

Adams,M. *et al.* (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, **252**, 1651–1656.

Allen,J. and Salzberg,S. (2005) JIGSAW: integration of multiple sources of evidence for gene prediction. *Bioinformatics*, **21**, 3596–3603.

Ashurst,J. *et al.* (2005) The Vertebrate Genome Annotation (VEGA) database. *Nucleic Acids Res.*, **33**, D459–D465.

Blow,N. (2009) Transcriptomics: the digital generation. *Nature*, **458**, 239–242.

Cloonan,N. *et al.* (2009) Rna-mate: a recursive mapping strategy for high-throughput rna-sequencing data. *Bioinformatics*.

Denoued,F. *et al.* (2008) Annotating genomes with massive-scale RNA-sequencing. *Genome Biol.*, **9**, R175.

Eriksson,N. *et al.* (2008) Viral population estimation using pyrosequencing. *PLoS Comput. Biol.*, **4**, e1000074.

Graveley,B. *et al.* (2010) The developmental transcriptome of Drosophila melanogaster. *Nature*, **461**, 473–479.

Guigó,R. *et al.* (2006) EGASP: the human ENCODE Genome Annotation Assessment Project. *Genome Biol.*, **7**, S2.

Guttman,M. *et al.* (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincrnas. *Nat. Biotechnol.*, **28**, 503–510.

Haas,B. *et al.* (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.*, **31**, 5654–5666.

Heber,S. *et al.* (2002) Splicing graphs and EST assembly problem. *Bioinformatics*, **18**, S181–S188.

Jiang,H. and Wong,W. (2009) Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics*, **25**, 1026–1032.

Mortazavi,A. *et al.* (2008) Mapping and quantifying mammalian transcriptomes by rna-seq. *Nat. Methods*, **5**.

Nagalakshmi,U. *et al.* (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, **320**, 1344–1349.

Schweikert,G. *et al.* (2009) mGene: accurate SVM-based gene finding with an application to nematode genomes. *Genome Res.*, **19**, 2133–2143.

Simpson,J. *et al.* (2009) Abyss: a parallel assembler for short read sequence data. *Genome Res.*, **19**.

Stanke,M. and Waack,S. (2003) Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*, **19**.

Trapnell,C. *et al.* (2009) Tophat: discovering splice junctions with rna-seq. *Bioinformatics*, **25**, 1105–11.

Trapnell,C. *et al.* (2010) Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–515.

Venter,J. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.