BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 14 2012, pages 1933–1934
doi:10.1093/bioinformatics/bts294

Sequence analysis

OSA: a fast and accurate alignment tool for RNA-Seq

Jun Hu[1,2,*], Huanying Ge[3], Matt Newman[1] and Kejun Liu[1,*]

[1]Division of Bioinformatics, Omicsoft Inc., 164 Quade Drive, Cary, NC 27513, USA, [2]Bioinformatics Research Center, North Carolina State University, Ricks Hall, 1 Lampe Dr., Raleigh, NC 27607, USA and [3]Genome Analysis Unit, Amgen Inc., One Amgen Center Drive, Thousand Oaks, CA 91320, USA

Associate Editor: Ivo Hofacker

ABSTRACT

Summary: Accurately mapping RNA-Seq reads to the reference genome is a critical step for performing downstream analysis such as transcript assembly, isoform detection and quantification. Many tools have been developed; however, given the huge size of the next generation sequencing datasets and the complexity of the transcriptome, RNA-Seq read mapping remains a challenge with the ever-increasing amount of data. We develop Omicsoft sequence aligner (OSA), a fast and accurate alignment tool for RNA-Seq data. Benchmarked with existing methods, OSA improves mapping speed 4–10-fold with better sensitivity and less false positives.

Availability: OSA can be downloaded from http://omicsoft.com/osa. It is free to academic users. OSA has been tested extensively on Linux, Mac OS X and Windows platforms.

Contact: john.hu@omicsoft.com; jhu7@ncsu.edu; jack.liu@omicsoft.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on March 2, 2012; revised on May 1, 2012; accepted on May 9, 2012

1 INTRODUCTION

RNA-Seq, a high-throughput sequencing technology used to sequence complementary DNA, has been widely thought of as 'a revolutionary tool for transcriptomics study' (Wang *et al*., 2009). Compared with microarray technology, conventional expressed sequence tag (EST) or other tag-based sequencing, it has less sampling bias, higher resolution and much broader expression range coverage (Marioni *et al*., 2008; Ozsolak and Milos, 2011; Wang, *et al*., 2008).

For human, mouse and other mammalian organisms with a relatively well-annotated genome, mapping short reads from next generation sequencing (NGS) to reference transcriptome/genome is usually the first and essential step for RNA-Seq data analysis. This 'reference-based' approach tends to be more sensitive and computationally efficient than a *de novo* assembly approach (Martin and Wang, 2011). Many tools have been developed in recent years (Au *et al*., 2010; Chen *et al*., 2011; Grant *et al*., 2011; Huang *et al*., 2011; Trapnell *et al*., 2009; Wang *et al*., 2010; Wu and Nacu, 2010). TopHat (Trapnell *et al*., 2009) maps the reads to reference genome using bowtie (Langmead *et al*., 2009). The consensus islands are first clustered; the splice junctions are then generated by enumerating the canonical donor and acceptor sites between the neighboring islands for putative introns. RNA-Seq unified mapper (RUM) (Grant *et al*., 2011) and RNASeqR (Chen *et al*., 2011) align the reads against both transcriptome and genome using bowtie. The results are merged and the un-mapped reads are further aligned to genome with blat (Kent, 2002). SOAPsplice (Huang *et al*., 2011) first maps reads to reference genome indexed with Burrows Wheeler transformation (BWT), and then unmapped reads are aligned again using two segment strategies. MapSplice (Wang *et al*., 2010) and SpliceMap (Au *et al*., 2010) also use similar break and assemble approach for longer reads.

We introduce Omicsoft sequence aligner (OSA), an ultra-fast RNA-Seq aligner/mapper, different from the majority of existing methods relying on Bowtie or BWA, which use a BWT-based index. OSA uses a k-mer indexed genome hash table and avoids breaking down longer reads into sub-reads and merging steps, which becomes increasingly computationally expensive with the longer reads from the newer sequencing machines. OSA is 4–10 times faster than popular RNA-Seq aligners developed and used so far. It works well for both short and long reads and its performance does not deteriorate for high-sequencing error data. OSA also significantly improves the ability to detect both short and long indels as well as exon junctions.

2 IMPLEMENTATION

OSA indexes the reference genomes into a 64-bit 14-mer library by default using an approach similar to GSNAP (Wu and Nacu, 2010) (see Supplementary Information for OSA index system detail). During RNA-Seq alignment, where a well-annotated gene model is available, OSA will first build the transcriptome reference library from the provided gtf(gff) gene model file. The alignment pipeline can be divided into four sub-steps. First, the RNA-Seq reads will be aligned to the transcriptome and translated to corresponding genome coordinates considering introns. From our benchmark results and previous publications, aligning the RNA-Seq reads to an annotated transcriptome can improve the speed and accuracy. Similar strategies have been successfully used by RUM, RNASEQR and latest TopHat 1.4 beta (Chen *et al*., 2011; Grant *et al*., 2011). Second, the unmapped reads will be mapped against the reference genome for un-annotated exons and novel exon junctions. The exon junction detection is based on an extended middle indel detection algorithm (Wu and Nacu, 2010). Only deletions were considered (as introns are deletions from the reference), and only canonical donor–acceptor patterns (GT-AG, GC-AG and AT-AC) are considered, as they constitute the majority of splice sites (Burset *et al*., 2001). Third, the exon junction detection algorithm will not be able to detect novel exon junctions with a distal short fragment. Thus, a step to rescue these reads was added. The remaining unmapped reads that fail to align to both the standard transcriptome and genome will be matched against the newly identified exon junctions in Step 2. Finally, all mapped reads from the previous steps will be combined for output.

*To whom correspondence should be addressed.

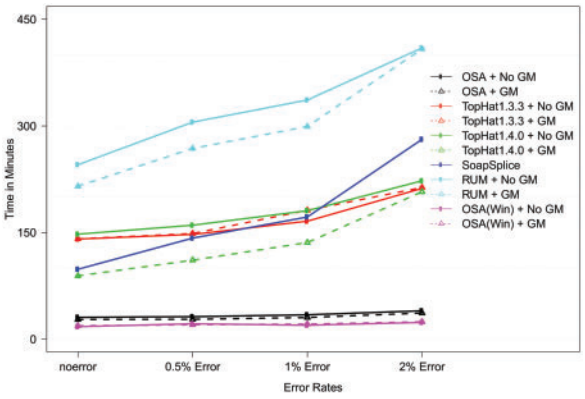

Fig. 1. Speed of various RNA-Seq alignment tools at different error levels. 10 M paired-end RNA-Seq reads are simulated at four different error rates. These reads are mapped to the mouse genome using different tools either with gene model (GM) or without. The time taken to align the reads is recorded. Results are averaged from three independent simulated datasets. The exact numeric values are summarized in Supplementary Table S1. (Win: Windows)

OSA implements various optimization strategies at different stages of mapping pipeline to improve alignment speed and accuracy. All the reference and read sequence matching are achieved through super-fast bitwise operations. For each read, where quality values are available, OSA will first scan the reads from 3′-end until the first acceptable quality nucleotide is found. This trimming process removes the ambiguous reads at the end and significantly improves the alignments speed (Falgueras *et al*., 2010). During alignment stage, a 'seed and extend' strategy is used to test each potential matching position. A similar strategy has been used in GSNAP (Wu and Nacu, 2010) and blat (Kent, 2002). In addition, OSA takes advantage of the properties of paired-end reads to filter the alignment results and limit alignment space.

3 RESULTS

We simulated 12 sets of 10 million pairs of paired-end RNA-Seq reads with three independent simulations at each of four error rates (0, 0.5, 1 and 2%) using the BEER pipeline (Grant *et al*., 2011). We compare OSA with TopHat (v1.3.3 and v1.4.0), RUM (v1.10) and SOAPsplice (v1.8) on a four-core 3 GHz AMD Phenom(tm) II X4 945 Processor desktop with 16 GB RAM running OpenSUSE Linux (See Supplementary Methods for details). The speed, the alignment accuracy and the percentage of indels and exon junctions correctly identified are evaluated for different tools either with a gene model or without gene model.

OSA has superior speed; it is about 4–10 times faster compared with other tools on Linux platform (Fig. 1). For all the tools, the alignment speed decreases with error rates; this is expected as the aligner has to try more scenarios for non-perfect matching. OSA's speed improvement is more dramatic at high error rates (Fig. 1). The benchmark results also show that using a gene model slightly improves the alignment speed (except Tophat1.3.3) (Fig. 1) and accuracy (Supplementary Fig. S1). When a gene model is provided, OSA matches most reads to the correct location of the transcriptome/genome; TopHat 1.3.3 is the close second. Both algorithms perform well even without a gene model. RUM and TopHat1.4.0 work much better with a gene model at low error rates and their performances deteriorate when error rate increases to 2% (Supplementary Fig. S1).

All tools detect >85% of exon junctions. OSA, SOAPsplice and RUM have higher sensitivity than TopHat. However, OSA has much fewer false positives than RUM and SOAPsplice (Supplementary Table S2A). RUM has the best sensitivity detecting indels. However, it also increases the percentage of the false positives and decreases the positive prediction value (PPV). TopHat detects a much smaller number of indels, this might be improved after integration with the latest Bowtie 2.0, which supports indel detection at the alignment stage. OSA achieves good sensitivity at detecting indels while controlling the false positives (Supplementary Table S3A). Due to sequencing errors (or random error introduced during simulation), one read can match to a different location on the genome with a better or equal matching score. A large portion of these false positives could be attributed to these sequencing/random errors. We tested the junction and indel detecting again after adjusting the metrics for these reads and we can see that the PPV of all tools improves, especially SOAPsplice and RUM. However, the overall detection pattern is still the same (Supplementary Tables S2B and S3B).

We also tested OSA against TopHat 1.3.3, which performs better than other tools we benchmarked against, with eight randomly selected NGS datasets from NCBI SRA. These datasets include both single- and paired-end reads, reads with different length, and from both human and mouse species with various sequencing depths. Similar to the simulation outcomes, we find that OSA is several times faster than TopHat and maps 3–4% more reads. Moreover, it also identifies much more known and novel exon junctions (Supplementary Table S4).

Conflict of Interest: J.H., M.N., and K.L. have financial interest in OmicSoft Corporation. The research reported in this paper is in the field in which OmicSoft is developing commercial software.

REFERENCES

Au,K.F. *et al.* (2010) Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic Acids Res.*, **38**, 4570–4578.

Burset,M. *et al.* (2001) SpliceDB: database of canonical and non-canonical mammalian splice sites. *Nucleic Acids Res.*, **29**, 255–259.

Chen,L.Y. *et al.* (2012) RNASEQR—a streamlined and accurate RNA-seq sequence analysis program. *Nucleic Acids Res*, **40**, e42.

Falgueras,J. *et al.* (2010) SeqTrim: a high-throughput pipeline for pre-processing any type of sequence read. *BMC Bioinformatics*, **11**, 38.

Grant,G.R. *et al.* (2011) Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). *Bioinformatics*, **27**, 2518–2528.

Huang,S. *et al.* (2011) SOAPsplice: genome-wide ab initio detection of splice junctions from RNA-Seq data. *Front. Genet.*, **2**, 46.

Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.

Langmead,B. *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

Marioni,J.C. *et al.* (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, **18**, 1509–1517.

Martin,J.A. and Wang,Z. (2011) Next-generation transcriptome assembly. *Nat. Rev. Genet.*, **12**, 671–682.

Ozsolak,F. and Milos,P.M. (2011) RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet.*, **12**, 87–98.

Trapnell,C. *et al.* (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.

Wang,E.T. *et al.* (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–476.

Wang,K. *et al.* (2010) MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.*, **38**, e178.

Wang,Z. *et al.* (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**, 57–63.

Wu,T.D. and Nacu,S. (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, **26**, 873–881.