Copyright Warning & Restrictions

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be "used for any purpose other than private study, scholarship, or research." If a, user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of "fair use" that user may be liable for copyright infringement,

This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve violation of copyright law.

Please Note: The author retains the copyright while the New Jersey Institute of Technology reserves the right to distribute this thesis or dissertation

Printing note: If you do not wish to print this page, then select "Pages from: first page # to: last page #" on the print dialog screen

ABSTRACT

PERFORMANCE COMPARISON OF FIVE RNA-Seq ALIGNMENT TOOLS

by
Yuanpeng Lu

Aligning millions of short reads to a reference genome is a critical task in high throughput sequencing. In recent years, a large number of mapping algorithms have been developed, all of which have in common that they align a vast number of reads to genomic or transcriptomic sequences. RNA-Seq data is discrete in nature, therefore with reasonable gene models and comparative metrics RNA-Seq data can be simulated to sufficient accuracy to enable meaningful benchmarking of alignment algorithms. To provide guidance in the choice of alignment algorithms, five different alignment tools for RNA-Seq data are evaluated. In order to compare the accuracy and sensitivity of the Bowtie, Bowtie2, GMAP, Tophat and GNUMAP tools, their alignment accuracy for approximately 1 million simulated reads of chromosome one was evaluated using these five alignment tools. Bowtie has the highest accuracy, which is 92.42%, while GMAP has the lowest, which is 49.63%. Tophat has the highest sensitivity, which is 71.35%, while GMAP has the lowest, which is 51.69%.

PERFORMANCE COMPARISON OF FIVE RNA-Seq ALIGNMENT TOOLS

by
Yuanpeng Lu

A Thesis
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Master of Science in Bioinformatics

Department of Computer Science

May 2013

.

APPROVAL PAGE

PERFORMANCE COMPARISON OF FIVE RNA-Seq ALIGNMENT TOOLS

Yuanpeng Lu

Dr. Zhi Wei, Thesis Advisor Assistant Professor of Computer Science, NJIT	Date
Dr. Usman W. Roshan, Committee Member Associate Professor of Computer Science, NJIT	Date
Dr. Egbert Ammicht, Committee Member Adjunct Professor of Mathematics, NJIT	Date

BIOGRAPHICAL SKETCH

Author: Yuanpeng Lu

Degree: Master of Science

Date: May 2013

Undergraduate and Graduate Education:

- Master of Science in Bioinformatics,
  New Jersey Institute of Technology, NJ, US, 2013

- Bachelor of Science in Bioinformatics,
  Harbin Institute of Technology, Heilongjiang, P. R. China, 2011

Major: Bioinformatics

Presentations and Publications:

Gai, L. Lu Y. et al, Separation Identification and Metabolic Pathway of Alkane Intermediate Degraded by Microbes, Geology and Development of Daqing Petroleum (Chinese), 2010, 5: 140-142

ACKNOWLEDGMENT

We thank Zhi Wei, Usman W. Roshan, Egbert Ammicht, Wei Wang and Xiao Ling for discussions and help.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure **Page**

CHAPTER 1

INTRODUCTION

RNA-Seq is a revolutionary technique to perform transcriptome studies based on next generation sequencing technologies. This technique is largely dependent on bioinformatics tools developed to support the different steps of the process.

In contrast to traditional Sanger sequencing, the next generation sequencing relies on multiple-fold coverage of each sequenced base by many short sequencing reads. Due to falling sequencing costs, next generation sequencing technologies have been extended to many more applications apart from genome sequencing, in particular transcriptome sequencing and quantification.

Previously used hybridization-based methods for quantification and characterization of transcripts require careful design of the array platform and knowledge about the transcriptome under investigation. Furthermore, they suffer from cross-hybridization effects and have a limited dynamic range. Earlier sequencing-based approaches to transcript quantification such as Serial Analysis of Gene Expression (SAGE) or Cap Analysis of Gene Expression (CAGE) had the advantage of providing count-based measures of transcript abundance, however due to high per-base sequencing costs, high throughput could only be achieved at the expense of small, often ambiguously mapping tag sizes. Furthermore, since transcripts were merely identified by their 3' or 5' terminal tags, these methods were oblivious to variations within the transcript. High-throughput RNA sequencing (RNA-seq) overcomes these limitations and provides a single methodology to assess transcript sequence, structure and abundance.

The mapping of sequencing reads to their sequence origin is the first step upon which any subsequent analyses are based. A specific challenge for the sequencing of eukaryotic transcriptomes is the mapping of reads from spliced transcripts to the genome. As read lengths increase, the number of reads spanning exon junctions increases such that alignment to an unspliced reference becomes impractical.

CHAPTER 2

MATERIALS AND METHODS

2.1 Alignment Algorithms

Bowtie is an ultrafast, memory-efficient alignment program for aligning short DNA sequence reads to large genomes. For the human genome, Burrows-Wheeler indexing allows Bowtie to align more than 25 million reads per CPU hour with a memory footprint of approximately 1.3 gigabytes. Bowtie extends previous Burrows-Wheeler techniques with a novel quality-aware backtracking algorithm that permits mismatches. Multiple processor cores can be used simultaneously to achieve even greater alignment speeds[1].

As the rate of sequencing increases, greater throughput is demanded from read aligners. The full-text minute index is often used to make alignment very fast and memory-efficient, but the approach is ill-suited to finding longer, gapped alignments. Bowtie 2 combines the strengths of the full-text minute index with the flexibility and speed of hardware accelerated dynamic programming algorithms to achieve a combination of high speed, sensitivity and accuracy[2].

GNUMAP performs alignment using a probabilistic Needleman-Wunsch algorithm. This tool is able to handle alignment in repetitive regions of a genome without losing information. The output of the program was developed to make possible easy visualization using available software[3].

The advent of next-generation sequencing technologies has increased the accuracy and quantity of sequence data, opening the door to greater opportunities in genomic research.

GNUMAP (Genomic Next- generation Universal MAPper), a program capable of overcoming two major obstacles in the mapping of reads from next-generation sequencing runs. First, an algorithm that probabilistically maps reads to repeat regions in the genome on a quantitative basis have been created. Second, a probabilistic Needleman–Wunsch algorithm which utilizes _prb.txt and _int.txt files produced in the Solexa/Illumina pipeline have been developed to improve the mapping accuracy for lower quality reads and increase the amount of usable data produced in a given experiment[4].

Gmap, a standalone program for mapping and aligning cDNA sequences to a genome. The program maps and aligns a single sequence with minimal startup time and memory requirements, and provides fast batch processing of large sequence sets. The program generates accurate gene structures, even in the presence of substantial polymorphisms and sequence errors, without using probabilistic splice site models. Methodology underlying the program includes a minimal sampling strategy for genomic mapping, oligomer chaining for approximate alignment, sandwich DP for splice site detection, and microexon identification with statistical significance testing[5].

On a set of human messenger RNAs with random mutations at a 1 and 3% rate, gmap identified all splice sites accurately in over 99.3% of the sequences, which was one-tenth the error rate of existing programs. On a large set of human expressed sequence tags, gmap provided higher-quality alignments more often than blat did. On a set of Arabidopsis cDNAs, gmap performed comparably with GeneSeqer. In these experiments, gmap demonstrated a several-fold increase in speed over existing programs[6].

TopHat is prepared to find de novo junctions. TopHat aligns reads in two steps. Firstly, unspliced reads are aligned with Bowtie. After, the aligned reads are assembled with Maq resulting islands of sequences. Secondly, the splice junctions are determined based on the initially unmapped reads and the possible canonical donor and acceptor sites within the island sequences[7].

A new protocol for sequencing the messenger RNA in a cell, known as RNA-Seq, generates millions of short sequence fragments in a single run. These fragments, or 'reads', can be used to measure levels of gene expression and to identify novel splice variants of genes. However, current software for aligning RNA-Seq data to a genome relies on known splice junctions and cannot identify novel ones. TopHat is an efficient read-mapping algorithm designed to align reads from an RNA-Seq experiment to a reference genome without relying on known splice sites[8].

The RNA-Seq reads from a recent mammalian RNA-Seq experiment was mapped and recovered more than 72% of the splice junctions reported by the annotation-based software from that study, along with nearly 20 000 previously unreported junctions. The TopHat pipeline is much faster than previous systems, mapping nearly 2.2 million reads per CPU hour, which is sufficient to process an entire RNA-Seq experiment in less than a day on a standard desktop computer. Several challenges unique to ab initio splice site discovery from RNA-Seq reads that will require further algorithm development have benn developed.

2.2 Simulation of RNA Sequencing Reads

FluxCapacitor is a computer program to predict splice form abundancies from reads of an RNA-seq experiment. FluxSimulator can generate simulated data for testing RNA-seq pipelines.The FluxSimulator is the part of the FLUX project that aims at providing an in silico reproduction of the experimental pipelines for RNA-Seq, adopting a minimal set of parameters. Corresponding models were established after analyzing RNA-Seq experiments from different cell types, sample preparation protocols and sequencing platforms.

CHAPTER 3

RESULTS

3.1 The number of reads and comparison to other methods

To evaluate the performance of RNA-Seq aligners, simulated data which was generated by FluxCapacitor was used. The data was generated from chromosome one of human RNA-Seq reads. The reads in the data set described above were aligned using Bowtie, Bowtie2, GMAP, Tophat and GNUMAP. As shown in Table 3.1 and Table 3.2, Bowtie2 and GMAP have relatively large number of alignment reads while Bowtie2 and Tophat have relatively large number of true alignment reads.

Table 3.1 Numbers of true alignment reads and total alignment reads from five alignment tools(75bp).

Alignment tools	True alignment reads	Total alignment reads
Bowtie	622,006	673,047
Bowtie2	653,871	1000,796
GMAP	517,316	1000,428
Tophat	714,109	780,733
GNUMAP	586,173	694,427

Table 3.2 Numbers of true alignment reads and total alignment reads from five alignment tools(100bp).

Alignment tools	True alignment reads	Total alignment reads
Bowtie	583,460	618,832
Bowtie2	622,460	999,704
GMAP	468,126	999,004
Tophat	676,476	731,499
GNUMAP	586,012	693,128

3.2 Accuracy and sensitivity of five alignment tools

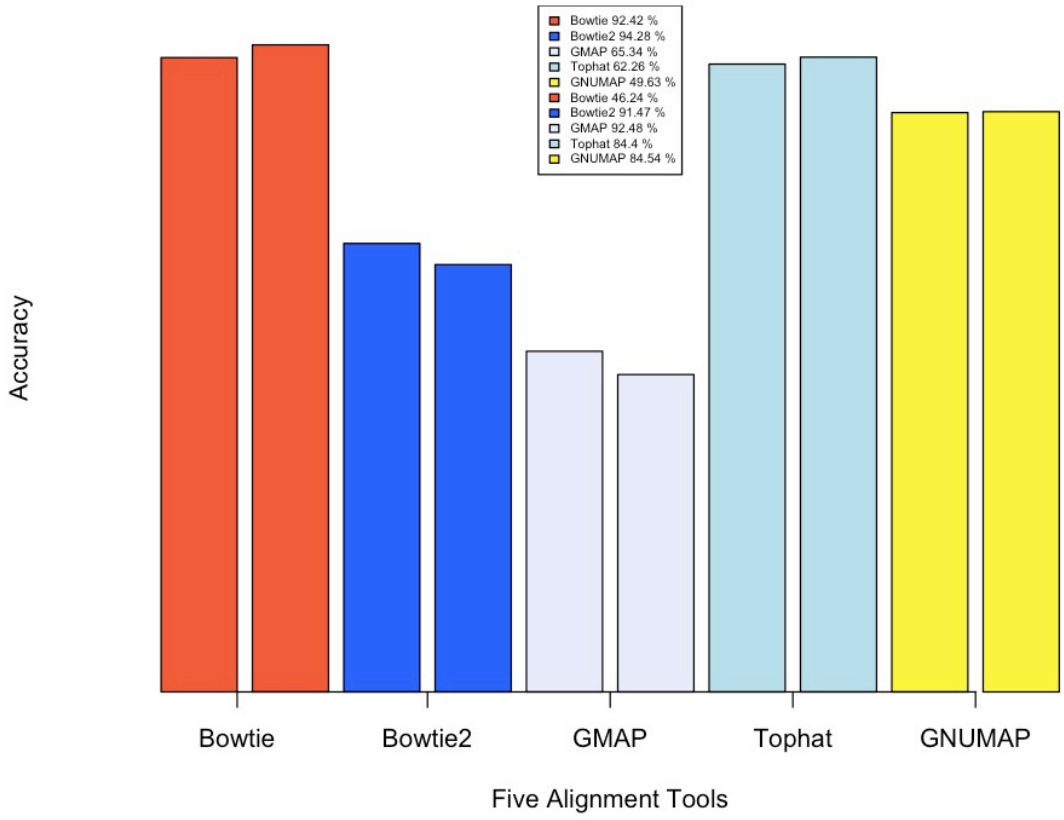

Figure 3.1 Accuracy statistics for analyses of simulated data sets. Bowtie has the highest accuracy which is 92.42, while GMAP has the lowest accuracy which is 49.63%. The accuracy of five RNA-Seq alignment tools have significant difference.

As shown in Figure 3.1, using the number of true alignment reads divided by the number of total alignment reads, Bowtie has relatively the highest accuracy among these five alignment tools while GMAP has the lowest accuracy.

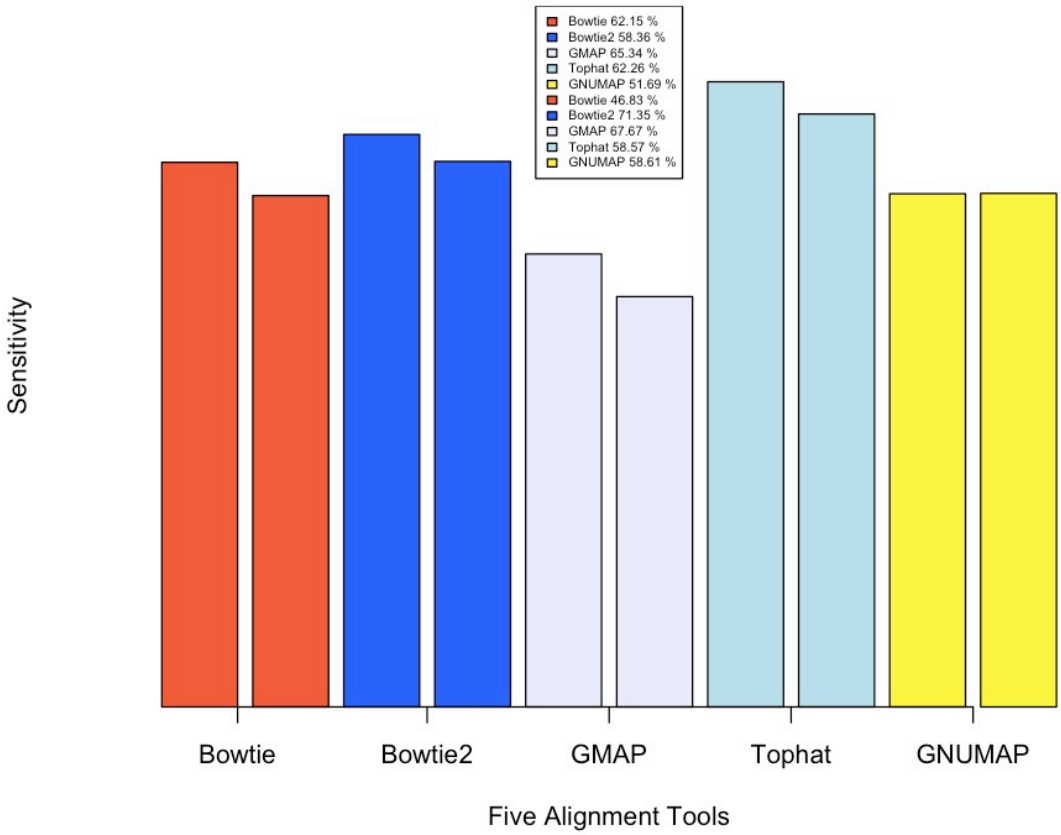

Figure 3.2 Sensitivity statistics for analyses of simulated data sets. Tophat has the highest accuracy which is 71.35,while GMAP has the lowest accuracy which is 51.69%. There was no significant difference among these five tools' sensitivity,all algorithms have sensitivity around 60%.

As shown in Figure 3.2, using the number of true alignment reads divided by the number of total reads from chromosome one, Tophat has relatively the highest sensitivity among these five alignment tools while GMAP has the lowest sensitivity.

CHAPTER 4

DISCUSSION

In this study, a comparison of read alignment algorithms with a particular focus on RNA-seq applications was performed. This is not an evaluation of mapping methods, but aims at evaluating the alignments algorithms and better reflect the characteristics of RNA-seq experiments.

A major difficulty in comparing the performance of different aligners is to set up a fair comparison in terms of the task and parameters chosen as evaluation metrics. To avoid a bias in parameter selection, optional parameters were trained on a smaller training set and these optional parameters were then used for evaluation.

This analysis also illustrates the importance of evaluating alignment quality compared to the actual number of mismatched and indels in a read and not the average error rate. The result showed that the reduction of recall due to mismatched and indels is not a gradual process.

CHAPTER 5

CONCLUSIONS

In summary, this study is relevant for the analysis of RNA-Seq data in several respects. First, it provides a comprehensive evaluation of the performance of algorithms from different algorithmic classes used for read alignments. Second, it provides guidance with regard to the choice of algorithm and parameters. Third, as alignment algorithms are generally used in a genetic fashion, exchanging the underlying alignment procedure for a better performing one provides one straightforward way to improve the overall performance of a mapping strategy. Here this study supports to some degree of using Tophat as alignment algorithm within RNA-seq mapping pipelines.

REFERENCES

1. Grant, G.R., et al., Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). Bioinformatics, 2011. **27**(18): p. 2518-28.

2. Lindner, R. and C.C. Friedel, A comprehensive evaluation of alignment algorithms in the context of RNA-seq. PLoS One, 2012. 7(12): p. e52403.

3. Wu, T.D. and S. Nacu, Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics, 2010. 26(7): p. 873-81.

4. Wu, T.D. and C.K. Watanabe, GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics, 2005. 21(9): p. 1859-75.

5. Clement, N.L., et al., The GNUMAP algorithm: unbiased probabilistic mapping of oligonucleotides from next-generation sequencing. Bioinformatics, 2010. 26(1): p. 38-45.

6. Kim, D. and S.L. Salzberg, TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. Genome Biol, 2011. 12(8): p. R72.

7. Langmead, B., et al., Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol, 2009. 10(3): p. R25.

8. Trapnell, C., L. Pachter, and S.L. Salzberg, TopHat: discovering splice junctions with RNA-Seq. Bioinformatics, 2009. 25(9): p. 1105-11.