Study Notes

Some Classic Papers in Bioinformatics

Jianyu Zhou

November 23, 2014

Contents

1 Relative Terms

Genome

Transcriptome The transcriptome is the set of all RNA molecules, including mRNA, rRNA, tRNA, and other non-coding RNA transcribed in one cell or a population of cells. It differs from the exome in that it includes only those RNA molecules found in a specified cell population, and usually includes the amount or concentration of each RNA molecule in addition to the molecular identities.

2 NGS Reads Mapping

2.1 Ultrafast and memory-efficient alignment of short DNA sequences to the human genome

Authors

Ben Langmead, Cole Trapnell, Mihai Pop and Steven L Salzberg

Abstract

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. Bowtie is open source http://bowtie.cbcb.umd.edu.

Link

http://www.biomedcentral.com/content/pdf/gb-2009-10-3-r25.pd

Notes

todo..

2.2 TopHat: discovering splice junctions with RNA-Seq

Authors

Cole Trapnell1, Lior Pachter and Steven L Salzberg

Abstract

Motivation: 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.
Results: We mapped the RNA-Seq reads from a recent mammalian RNA-Seq experiment 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. We describe several challenges unique to *ab initio* splice site discovery from RNA-Seq reads that will require further algorithm development.

Link

http://bioinformatics.oxfordjournals.org/content/25/9/1105.short

Notes

todo..

3 Transcript abundance estimation and reconstruction

3.1 Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation

Authors

Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold and Lior Pachter

Abstract

High-throughput mRNA sequencing (RNA-Seq) promises simultaneous transcript discovery and abundance estimation. However, this would require algorithms that are not restricted by prior gene annotations and that account for alternative transcription and splicing. Here we introduce such algorithms in an open-source software program called Cufflinks. To test Cufflinks, we sequenced and analyzed ¿430 million paired 75-bp RNA-Seq reads from a mouse myoblast cell line over a differentiation time series. We detected 13,692 known transcripts and 3,724 previously unannotated ones, 62% of which are supported by independent expression data or by homologous genes in other species. Over the time series, 330 genes showed complete switches in the dominant transcription start site (TSS) or splice isoform, and we observed more subtle shifts in 1,304 other genes. These results suggest that Cufflinks can illuminate the substantial

regulatory flexibility and complexity in even this well-studied model of muscle development and that it can improve transcriptome-based genome annotation.

Link
http://www.nature.com/nbt/journal/v28/n5/pdf/nbt.1621.pdf

Notes

todo..

4 Alignment-free Method

4.1 RNA-Skim: a rapid method for RNA-Seq quantification at transcript level

Authors
Zhaojun Zhang and Wei Wang

Abstract
Motivation: RNA-Seq technique has been demonstrated as a revolutionary means for exploring transcriptome because it provides deep coverage and base pair-level resolution. RNA-Seq quantification is proven to be an efficient alternative to Microarray technique in gene expression study, and it is a critical component in RNA-Seq differential expression analysis. Most existing RNA-Seq quantification tools require the alignments of fragments to either a genome or a transcriptome, entailing a time-consuming and intricate alignment step. To improve the performance of RNA-Seq quantification, an alignment-free method, Sailfish, has been recently proposed to quantify transcript abundances using all k-mers in the transcriptome, demonstrating the feasibility of designing an efficient alignment-free method for transcriptome quantification. Even though Sailfish is substantially faster than alternative alignment-dependent methods such as Cufflinks, using all k-mers in the transcriptome quantification impedes the scalability of the method.
Results: We propose a novel RNA-Seq quantification method, RNA-Skim, which partitions the transcriptome into disjoint transcript clusters based on sequence similarity, and introduces the notion of sig-mers, which are a special type of k-mers uniquely associated with each cluster. We demonstrate that the sig-mer counts within a cluster are sufficient for estimating transcript abundances with accuracy comparable with any state-of-the-art method. This enables RNA-Skim to perform transcript quantification on each cluster independently, reducing a complex optimization problem into smaller optimization tasks that can be run in parallel. As a result, RNA-Skim uses $< 4\%$ of the k-mers and $< 10\%$ of the CPU time required by Sailfish. It is able to finish transcriptome quantification in < 10 min per sample by using just a single thread on a

commodity computer, which represents > 100 speedup over the state-of-the-art alignment-based methods, while delivering comparable or higher accuracy.

Link

http://bioinformatics.oxfordjournals.org/content/30/12/i283.full

Notes

todo..