Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Applications Note

Subject Section

TieBrush: an efficient method for aggregating and summarizing mapped reads across large datasets

Ales Varabyou [1,2,*+], Geo Pertea [4+], Christopher Pockrandt [1,3] and Mihaela Pertea [1,2,3,*]

[1]Center for Computational Biology, Johns Hopkins University, Baltimore, MD 21211, USA
[2]Department of Computer Science, Johns Hopkins University, Baltimore, MD 21211, USA
[3]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21205, USA
[4]Lieber Institute for Brain Development, Johns Hopkins Medical Campus, Baltimore, MD 21205, USA.

*To whom correspondence should be addressed.
+These authors contributed equally.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: Although the ability to programmatically summarize and visually inspect sequencing data is an integral part of genome analysis, currently available methods are not capable of handling large numbers of samples. In particular, making a visual comparison of transcriptional landscapes between two sets of thousands of RNA-seq samples is limited by available computational resources, which can be overwhelmed due to the sheer size of the data. In this work we present TieBrush, a software package designed to process very large sequencing datasets (RNA, whole-genome, exome, etc) into a form that enables quick visual and computational inspection. TieBrush can also be used as a method for aggregating data for downstream computational analysis, and is compatible with most software tools that take aligned reads as input.
Availability: TieBrush is provided as a C++ package under the MIT License. Pre-compiled binaries, source code and example data are available on GitHub (https://github.com/alevar/tiebrush).
Contact: ales.varabyou@jhu.edu, mpertea@jhu.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

As sequencing costs have decreased, studies of gene expression have generated ever-larger numbers of samples per condition; e.g., the GTEx project (Lonsdale *et al*., 2013) generated hundreds of RNA-seq data sets for each tissue in the study. However, while computational methods have been developed to extract base or junction coverage from collections of samples (Li *et al*., 2009; Quinlan *et al*., 2010; Wilks *et al*., 2018; Pedersen *et al*., 2018), manual validation through visual inspection of regions of interest can be difficult or impossible due to the size of the data. Visual inspection remains a critical step in identifying unaccounted-for variables and gaining a better understanding of the data; for example, in RNA-seq experiments, the ability to visually compare the transcriptional profiles of a gene between two experimental conditions may reveal alternative splicing, transcriptional noise, and other important features. Yet with existing tools, inspecting large datasets and collections of samples may be extremely cumbersome, due to computational and memory constraints.

To enable the rapid manipulation of extremely large sequencing datasets, we designed TieBrush, an efficient method for merging redundant information from multiple alignment files. The method is designed to optimize investigations of sequencing experiments (eg., RNA, whole-genome, exome). TieBrush preserves much of the original information in a greatly condensed representation as a BAM file (Li *et al*., 2009), which allows manipulation and extraction of dataset and subset-specific statistics using tools within the package as well as other common utilities.

2 Overview and Features

2.1 Merging Datasets With TieBrush

Working with large collections of samples is constrained by available memory and computational costs. Visualizing datasets which may contain thousands of samples, can overwhelm the capabilities of currently available visualization tools such as IGV (Thorvaldsdóttir *et al*., 2013). TieBrush is designed as an efficient method for processing very large datasets

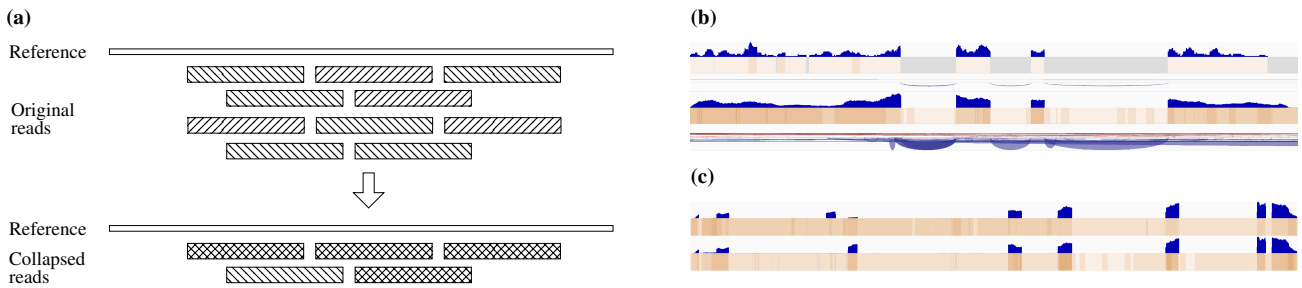

Fig. 1. Overview of the workflow of TieBrush and illustration of the summarized data produced by TieCov. (a) Redundant reads from 2 samples (left and right diagonals) are identified by TieBrush to produce a collapsed representation of 10 reads in 5 records. (b) Comparison of transcription of the NEFL gene in heart (top) and brain (bottom) tissues from GTEx. Each tissue is represented by three tracks produced by TieCov: read coverage (top), percent of samples containing the reads (middle), and splice junctions (bottom). The plot illustrates the higher prevalence and expression of the gene in brain tissue. (c) Comparison of transcription of the SLC25A3 gene in heart (top) and brain (bottom) tissues from GTEx. While the gene is expressed in both tissues, coverage data clearly indicate an exon switch where the 3rd and 4th exons are expressed at dramatically different levels in the two tissues.

of aligned RNA or DNA sequencing reads, exploiting the inherent redundancy in deeply covered regions. By identifying alignments within and between samples that are equivalent, i.e. sequences that are mapped to the same set of coordinates and can be considered duplicates of each other, based on user-defined criteria (eg., starting at the same position, having the same CIGAR string, or the same variants), TieBrush can collapse the alignments into a single entry in the output file. Information regarding the number of samples in the alignment is then stored efficiently within auxiliary tags. Furthermore, the method is designed to recognize "tiebrushed" files, allowing users to add data to pre-processed files efficiently. This property also allows for massively parallel execution of TieBrush on large datasets via splitting inputs into batches and processing them concurrently.

To illustrate the efficiency of TieBrush we processed 1,409 samples from brain tissue within the GTEx collection of RNA-seq experiments (Lonsdale *et al*., 2013). A total of over 166 billion individual alignments across all samples was compressed by TieBrush into fewer than 3 billion records, achieving a 98% decrease in the number of records and a 99.5% decrease in storage space (Supplementary Materials).

2.2 Summary and Visualization with TieCov

One of the most common methods for researchers to visualize genome sequence data is to load read alignments into a genome browser. To assist in this task, we implemented TieCov as part of TieBrush. TieCov facilitates efficient extraction of read coverage, splice junctions, and other features across a large number of samples. The format of TieCov's summary output is suitable for easy command-line parsing, and is tailored to present information in an intuitive way using IGV.

Currently TieCov supports three types of output for visualization in genome browsers (Fig 1 b, c): 1) the coverage track describes the cumulative sequencing depth at each sequenced base across a dataset; 2) the splice junction track lists all donor-acceptor pairs across the dataset and their multiplicity; 3) the sample count track is a heatmap that uses color intensity to show prevalence of regions among sequenced samples. Each type of output is tailored to highlight differences between regions and/or experiments. Additionally, collapsed alignments can also be loaded into the genome browser for inspection of non-redundant data.

To illustrate the use of TieBrush and TieCov, we randomly selected, collapsed, and computed statistics using 10 samples from brain and 10 more from heart tissue (Supplementary Tables 1, 2) within the GTEx dataset. Figure 1 shows IGV snapshots of processed data from two loci known to have different transcriptional profiles in heart and brain tissues, SLC25A3 and NEFL. In Figure 1b, expression data from the NEFL gene (which produces the neurofilament protein) is clearly much higher in brain, as expected for a gene that is highly expressed in neurons. Figure 1c shows expression data for SLC25A3, which has alternative dominant isoforms in the two tissues, as has been previously shown (Wang *et al*., 2008).

3 Conclusion

TieBrush is designed to facilitate visual inspection of large sets of experiments, particularly RNA-seq experiments, by first removing redundancies and then extracting key statistics from the data. By being able to process files sequentially, TieBrush can store and update tiebrushed files in an incremental way, as more data becomes available. Tiebrush can run both locally on desktop computers and remotely on compute nodes, which may be particularly important when working with very large datasets such as GTEx that cannot be easily transferred between devices. The summaries produced by TieCov and TieBrush can be visualized or reviewed using familiar utilities such as IGV, the UCSC genome browser, and samtools. The condensed representation of the data allows all of these tools to operate much more efficiently and helps users gain insights into large data sets.

Acknowledgements

We would like to thank Steven Salzberg for comments on a draft of the manuscript, and Martin Steinegger for testing beta versions of the software.

Funding

This work was supported in part by grant DBI-1759518 from the NSF, and grant R01-HG006677 from NIH.

References

Li,H. et al. (2009) The sequence alignment/map format and SAMtools, *Bioinformatics*, **25(16)**, 2078-2079.

Lonsdale,J. et al. (2013) The genotype-tissue expression (GTEx) project. *Nature genetics*, **45(6)**, 580-585.

Quinlan,A.R. et al. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26(6)**, 841-842.

Rosenbloom,K.R. et al. (2015) The UCSC genome browser database: 2015 update, *Nucleic acids research*, **43(D1)**, D670-D681.

Pedersen,B.S. et al. (2018). Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics*, **34(5)**, 867-868.

Thorvaldsdóttir,H. et al. (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration, *Briefings in bioinformatics*, **14(2)**, 178-192.

Wang,E.T. et al. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456(7221)**, 470-476.

Wilks,C. et al. (2018) Snaptron: querying splicing patterns across tens of thousands of RNA-seq samples. *Bioinformatics*, **34(1)**, 114-116.