BIOINFORMATICS APPLICATIONS NOTE

Vol. 26 no. 6 2010, pages 841–842
doi:10.1093/bioinformatics/btq033

Genome analysis

Advance Access publication January 28, 2010

BEDTools: a flexible suite of utilities for comparing genomic features

Aaron R. Quinlan[1,2,*] and Ira M. Hall[1,2,*]

[1]Department of Biochemistry and Molecular Genetics, University of Virginia School of Medicine and [2]Center for Public Health Genomics, University of Virginia, Charlottesville, VA 22908, USA

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Testing for correlations between different sets of genomic features is a fundamental task in genomics research. However, searching for overlaps between features with existing web-based methods is complicated by the massive datasets that are routinely produced with current sequencing technologies. Fast and flexible tools are therefore required to ask complex questions of these data in an efficient manner.

Results: This article introduces a new software suite for the comparison, manipulation and annotation of genomic features in Browser Extensible Data (BED) and General Feature Format (GFF) format. BEDTools also supports the comparison of sequence alignments in BAM format to both BED and GFF features. The tools are extremely efficient and allow the user to compare large datasets (e.g. next-generation sequencing data) with both public and custom genome annotation tracks. BEDTools can be combined with one another as well as with standard UNIX commands, thus facilitating routine genomics tasks as well as pipelines that can quickly answer intricate questions of large genomic datasets.

Availability and implementation: BEDTools was written in C++. Source code and a comprehensive user manual are freely available at http://code.google.com/p/bedtools

Contact: aaronquinlan@gmail.com; imh4y@virginia.edu

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

Received on November 24, 2009; revised on January 11, 2010; accepted on January 21, 2010

1 INTRODUCTION

Determining whether distinct sets of genomic features (e.g. aligned sequence reads, gene annotations, ESTs, genetic polymorphisms, mobile elements, etc.) overlap or are associated with one another is a fundamental task in genomics research. Such comparisons serve to characterize experimental results, infer causality or coincidence (or lack thereof) and assess the biological impact of genomic discoveries. Genomic features are commonly represented by the Browser Extensible Data (BED) or General Feature Format (GFF) formats and are typically compared using either the UCSC Genome Browser's (Kent *et al*., 2002) 'Table Browser' or using the Galaxy (Giardine *et al*., 2005) interface. While these tools offer a convenient and reliable method for such analyses, they are not amenable to large and/or *ad hoc* datasets owing to the inherent need to interact with a remote or local web site installation. Moreover, complicated analyses often require iterative testing and refinement. In this sense, faster and more flexible tools allow one to conduct a greater number and more diverse set of experiments. This necessity is made more acute by the data volume produced by current DNA sequencing technologies. In an effort to address these needs, we have developed BEDTools, a fast and flexible suite of utilities for common operations on genomic features.

*To whom correspondence should be addressed.

2 FEATURES AND METHODS

2.1 Common scenarios

Genomic analyses often seek to compare features that are discovered in an experiment to known annotations for the same species. When genomic features from two distinct sets share at least one base pair in common, they are defined as 'intersecting' or 'overlapping'. For example, a typical question might be '*Which of my novel genetic variants overlap with exons*?' One straightforward approach to identify overlapping features is to iterate through each feature in set A and repeatedly ask if it overlaps with any of the features in set B. While effective, this approach is unreasonably slow when screening for overlaps between, for example, millions of DNA sequence alignments and the RepeatMasker (Smit *et al*., 1996–2004) track for the human genome. This inefficiency is compounded when asking more complicated questions involving many disparate sets of genomic features. BEDTools was developed to efficiently address such questions without requiring an installation of the UCSC or Galaxy browsers. The BEDTools suite is designed for use in a UNIX environment and works seamlessly with existing UNIX utilities (e.g. *grep*, *awk*, *sort*, etc.), thereby allowing complex experiments to be conducted with a single UNIX pipeline.

2.2 Language and algorithmic approach

BEDTools incorporates the genome-binning algorithm used by the UCSC Genome Browser (Kent *et al*., 2002). This clever approach uses a hierarchical indexing scheme to assign genomic features to discrete 'bins' (e.g. 16 kb segments) along the length of a chromosome. This expedites searches for overlapping features, since one must only compare features between two sets that share the same (or nearby) bins. As illustrated in Supplementary Figure 1, calculating feature overlaps for large datasets (e.g. millions of sequence alignments) is substantially faster than using the tools available on the public Galaxy web site. The software is written in C++ and supports alignments in BAM format (Li *et al*., 2009) through use of the BAMTools libraries (Barnett *et al.*, http://sourceforge.net/projects/bamtools/).

2.3 Supported operations

Table 1 illustrates the wide range of operations that BEDTools support. Many of the tools have extensive parameters that allow user-defined overlap criteria and fine control over how results are reported. Importantly, we have also defined a concise format (BEDPE) to facilitate comparisons of discontinuous features (e.g. paired-end sequence reads) to each other (*pairToPair*), and to genomic features in traditional BED format (*pairToBed*). This functionality is crucial for interpreting genomic rearrangements detected by paired-end mapping, and for identifying fusion genes or alternative splicing patterns by RNA-seq. To facilitate comparisons with data produced by current DNA sequencing technologies, *intersectBed* and *pairToBed* compute overlaps between sequence alignments in BAM format (Li *et al*., 2009), and a general purpose tool is provided to convert BAM alignments to BED format, thus facilitating the use of BAM alignments with all other BEDTools (Table 1). The following examples illustrate the use of *intersectBed* to isolate single nucleotide polymorphisms (SNPs) that overlap with genes, *pairToBed* to create a BAM file containing only those alignments that overlap with exons and *intersectBed* coupled with *samtools* to create a SAM file of alignments that do not intersect (-v) with repeats.

```
$ intersectBed -a snps.bed -b genes.bed > out.bed

$ pairToBed -abam reads.bam -b exons.bed > out.bam

$ intersectBed -abam reads.bam -b repeats.bed -v |
samtools view - > reads.noRepeats.sam
```

Other notable tools include *coverageBed*, which calculates the depth and breadth of genomic coverage of one feature set (e.g. mapped sequence reads) relative to another; *shuffleBed*, which permutes the genomic positions of BED features to allow calculations of statistical enrichment; *mergeBed*, which combines overlapping features; and utilities that search for nearby yet non-overlapping features (*closestBed* and *windowBed*). BEDTools also includes utilities for extracting and masking FASTA sequences (Pearson and Lipman, 1988) based upon BED intervals. Tools with similar functionality to those provided by Galaxy were directly compared for correctness using the 'knownGene' and 'RepeatMasker' tracks from the hg19 build of the human genome. The results from all analogous tools were found to be identical (Table 1).

2.4 Other advantages

Except for the novel paired-end functionality and support for alignments in BAM format, many of the genomic comparisons supported by BEDTools can be performed in one way or another with available web-based tools. However, BEDTools offers several important advantages. First, it can read data from standard input and write to standard output, which allows complex set operations to be performed by combining BEDTools operations with each other or with existing UNIX utilities. Second, most of the tools can distinguish DNA strands when searching for overlaps, which allows orientation to be considered when interpreting paired-end mapping or RNA-seq data. Third, the use of BEDTools mitigates the need to interact with local or public instances of the UCSC Genome Browser or Galaxy, which can be a major bottleneck when working with large genomics datasets. Finally, the speed and extensive functionality of BEDTools allow greater flexibility in defining and refining genomic comparisons. These features allow for diverse and complex comparisons to be made between ever-larger genomic datasets.

Table 1. Summary of supported operations available in the BEDTools suite

Utility	Description
intersectBed*	Returns overlaps between two BED files.
pairToBed	Returns overlaps between a BEDPE file and a BED file.
bamToBed	Converts BAM alignments to BED or BEDPE format.
pairToPair	Returns overlaps between two BEDPE files.
windowBed	Returns overlaps between two BED files within a user-defined window.
closestBed	Returns the closest feature to each entry in a BED file.
subtractBed*	Removes the portion of an interval that is overlapped by another feature.
mergeBed*	Merges overlapping features into a single feature.
coverageBed*	Summarizes the depth and breadth of coverage of features in one BED file relative to another.
genomeCoverageBed	Histogram or a 'per base' report of genome coverage.
fastaFromBed	Creates FASTA sequences from BED intervals.
maskFastaFromBed	Masks a FASTA file based upon BED coordinates.
shuffleBed	Permutes the locations of features within a genome.
slopBed	Adjusts features by a requested number of base pairs.
sortBed	Sorts BED files in useful ways.
linksBed	Creates HTML links from a BED file.
complementBed*	Returns intervals not spanned by features in a BED file.

Utilities in bold support sequence alignments in BAM. Utilities with an asterisk were compared with Galaxy and found to yield identical results.

ACKNOWLEDGEMENTS

We thank Royden Clark for helpful algorithmic advice.

Funding: Ruth L. Kirschstein National Research Service Award from the National Institutes of Health [1F32HG005197-01 to A.R.Q.]; a Burroughs Wellcome Fund Career Award to I.M.H.; National Institutes of Health Director's New Innovator Award [DP2OD006493-01 to I.M.H.].

Conflict of Interest: none declared.

REFERENCES

Giardine,B. *et al*. (2005) Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.*, **15**, 1451–1455.

Kent,W.J. *et al*. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.

Li,H. *et al*. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.

Smit,A. *et al*. (1996–2004) RepeatMasker. Open-3.0. Available at http://www.repeatmasker.org/