Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Application Note

Application Note

annotatr: Associating genomic regions with genomic annotations

Raymond G. Cavalcante[1,*] and Maureen A. Sartor[1,2]

[1]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109; [2]Department of Biostatistics, University of Michigan, Ann Arbor, MI

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Analysis of next-generation sequencing data often results in a list of genomic regions. These may include differentially methylated CpGs/regions, transcription factor binding sites, interacting chromatin regions, or GWAS-associated SNPs, among others. A common analysis step is to annotate such genomic regions to genomic annotations (promoters, exons, enhancers, etc.). Existing tools are limited by requiring an artificial one-to-one region-to-annotation mapping, by a lack of visualization options to clearly and easily summarize annotations, by the time it takes to annotate regions, or by some combination thereof.
Results: We have developed the annotatr R package to easily and quickly summarize, and visualize the association of genomic regions with genomic annotations. The annotatr package reports all intersections of regions and annotations, giving a better understanding of the genomic context of the regions. A variety of visualization functions are implemented in annotatr to easily plot numerical or categorical data associated with the regions across the annotations, providing insight into how characteristics of the regions differ across the annotations. We also demonstrate that annotatr is up to 11x faster than the comparable R package, ChIPpeakAnno. Overall, annotatr facilitates easy and fast genomic annotation of genomic regions, enabling a richer biological interpretation of experiments.
Availability: http://www.github.com/rcavalcante/annotatr/
Contact: rcavalca@umich.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Genomic regions resulting from next-generation sequencing experiments and bioinformatics pipelines are more meaningful when annotated to genomic features. Hyper-methylation at promoters containing a CpG island may indicate different regulatory regimes in one condition compared to another. It may be of interest to find that a particular transcription factor overwhelmingly binds in promoters, while another binds mostly in 3'UTRs. A SNP occurring in an exon, or an enhancer, is likely of greater priority than one occurring in an inter-genic region.

While there exist tools to intersect genomic regions of interest with genomic annotations, we found the annotations, methods of intersection, and visualization options to be lacking. ChIPpeakAnno (Zhu *et al.*, 2010) is an R package that has been used in many studies across a variety of organisms. It returns only one genomic annotation per feature in a table. While providing the user with useful visualizations, these are limited by their inability to incorporate data that may be associated with the regions of interest. AnnotateGenomicRegions (Zammataro *et al.*, 2014) is a web-based annotator that is easy to use and appears to be quite fast. It returns all annotations intersecting regions of interest (a one-to-many mapping), but its only output is the table of intersections. The simplicity of a web-based tool also comes at the cost of not being able to incorporate the it into a pipeline. BEDtools (Quinlan & Hall, 2010), implemented in C++, intersects and aggregates genomic regions with annotations, and is very fast. However, its more general purpose means the user must provide annotations for intersection, and write their own code for visualization.

We have developed annotatr, an R package that reports all intersections of regions of interest with custom annotations or pre-computed genomic annotations from hg19, hg38, mm9, or mm10. annotatr enables users to associate an arbitrary number of numerical or categorical data columns with regions, enabling better understanding of the underlying experiments via summarization and visualization functions. annotatr is fast, easily incorporated into bioinformatics pipelines, and available on GitHub.

2 Implementation and Features

We downloaded CpG island (CGI), knownGene, and gaps tables for hg19, hg38, mm9, and mm10 from the UCSC Table Browser (Karolchik *et al.*, 2004), and processed them into the built-in annotations. See Supplementary Methods 1 and Figure S1 for more detail. For other genomes, users can provide any number of custom annotations.

annotatr is implemented in four modules that read, intersect, summarize, and visualize. The read module reads a BED6+ file, defined as BED6 plus any number of numerical or categorical data columns. The readr::read_tsv() function (Wickham & Francois, 2015b) enables large files to be read in very quickly. All region data are stored as GenomicRanges objects (Lawrence *et al.*, 2013). The intersect module uses the GenomicRanges::findOverlaps() function to quickly report the overlap of a region with all intersecting genomic annotations selected by the user. Overlaps are annotated and sorted with the dplyr package (Wickham & Francois, 2015a) to return a table of annotations. The summarize module uses the dplyr package to summarize any number of numerical data columns over the annotations or any grouping of categorical data columns.

The visualize module provides a simple interface for making a variety of plots with the ggplot2 package (Wickham, 2009). As an example, consider a file reporting the results of differential methylation (DM) tests between two groups for regions across the genome as a BED6+ file containing the following columns: chromosome, start, end, DM status, p-value, strand, methylation difference of the groups, group 1 methylation rate, and group 0 methylation rate. The annotatr package implements functions to show the number of regions in each annotation (Figure S2), the number of regions occurring in pairs of annotations (Figure S3), the distribution of numerical data across the annotations or any categorical variable (Figure 1A), the joint distribution of two data columns across the annotations or any categorical variable (Figure S4), the distribution of numerical data for regions in either of two annotations and the intersection (Figure S5), and the distribution of a categorical variable across the annotations or any categorical variable (Figure 1B).

We used the microbenchmark R package (Mersmann, 2015) to compare runtimes between ChIPpeakAnno (v3.4.1) and annotatr (v0.5.1) on four data sets varying in size from 27 000 to 25 000 000 lines. See Supplementary Methods 2 for details. On average, annotatr performs between 1.6 and 11.3 times faster than ChIPpeakAnno, with increasingly better performance as file size increases (Table S1).

3 Discussion

Associating regions of interest to genomic annotations is a standard part of many bioinformatics pipelines. annotatr improves upon existing annotation tools by returning *all* the genomic annotations associated with a region instead of artificially prioritizing one annotation type over another, enabling a clearer picture of the biological complexities at play. In addition to tabular output of the annotations, annotatr's built-in visualization functions provide an easy and flexible way to summarize the annotations and view how data associated with the regions changes in different genomic contexts. The annotatr package thus enables fast exploration and more complete interpretation of experiments.

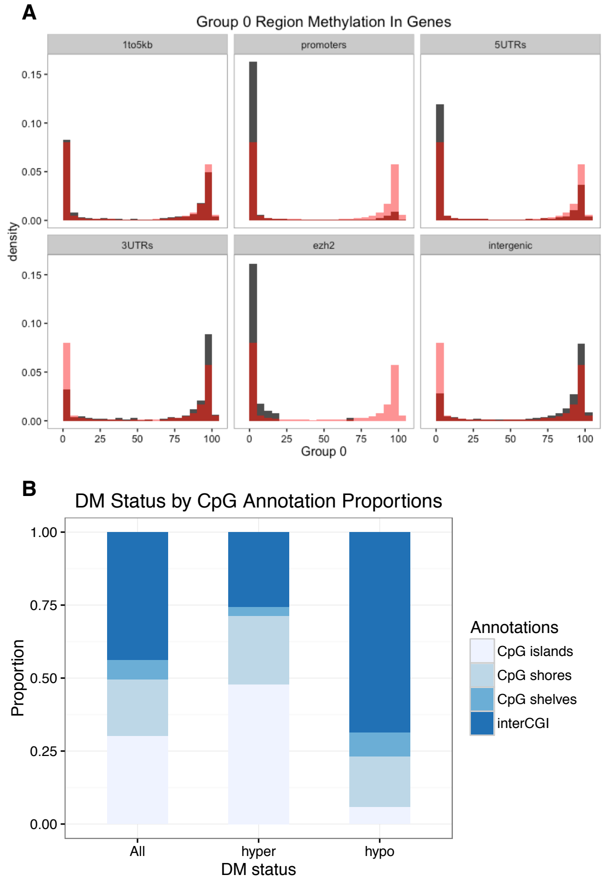

Figure 1: (A) The distribution of an experimental group's methylation rate across a number of annotations (darker histogram) with the background distribution overlaid (lighter histogram). (B) The proportion of hyper- and hypo-methylated regions associated with CpG annotations, with the background distribution (All) for comparison.

Acknowledgements

We thank Hani Habra and Jian Kang for their input on the conception and implementation of the package. We also thank Shweta Ramdas and Chee Lee for feedback on the manuscript and package vignette.

Funding

RC is funded on the Bioinformatics Training Grant (T32 GM070499).

Conflict of Interest: none declared.

References

Lawrence, M., *et al.* (2013). Software for computing and annotating genomic ranges. *PLoS Computational Biology*, *9*(8), e1003118.

Mersmann, O. (2015). microbenchmark: Accurate Timing Functions. R package version 1.4-2.1. http://CRAN.R-project.org/package=microbenchmark

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842.

Wickham, H. (2009) ggplot2: elegant graphics for data analysis. Springer NY

Wickham, H. and Francois, R. (2015a). dplyr: A Grammar of Data Manipulation. R package version 0.4.3. http://CRAN.R-project.org/package=dplyr

Wickham, H. and Francois, R. (2015b). readr: Read Tabular Data. R package version 0.2.2. http://CRAN.R-project.org/package=readr

Zammataro, L., *et al.* (2014). AnnotateGenomicRegions: a web application. *BMC Bioinformatics*, *15*(Suppl 1), S8.

Zhu, L. J., *et al.* (2010). ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics*, *11*, 237.