BIOINFORMATICS APPLICATIONS NOTE

Vol. 24 no. 15 2008, pages 1729–1730
doi:10.1093/bioinformatics/btn305

Genome analysis

FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology

Anthony P. Fejes[1,*], Gordon Robertson[1], Mikhail Bilenky[1], Richard Varhol[1], Matthew Bainbridge[2] and Steven J. M. Jones[1,*]

[1]Genome Sciences Centre, BC Cancer Agency, Suite 100 570 West 7th Avenue, Vancouver, British Columbia, Canada V5Z 4S6 and [2]College of Medicine, Houston, Texas, One Baylor Plaza, MS: BCM215, Houston, TX 77030, USA

Received on March 19, 2008; revised on May 8, 2008; accepted on June 9, 2008

Advance Access publication July 3, 2008

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: Next-generation sequencing can provide insight into protein–DNA association events on a genome-wide scale, and is being applied in an increasing number of applications in genomics and meta-genomics research. However, few software applications are available for interpreting these experiments. We present here an efficient application for use with chromatin-immunoprecipitation (ChIP-Seq) experimental data that includes novel functionality for identifying areas of gene enrichment and transcription factor binding site locations, as well as for estimating DNA fragment size distributions in enriched areas. The FindPeaks application can generate UCSC compatible custom 'WIG' track files from aligned-read files for short-read sequencing technology. The software application can be executed on any platform capable of running a Java Runtime Environment. Memory requirements are proportional to the number of sequencing reads analyzed; typically 4 GB permits processing of up to 40 million reads.

Availability: The FindPeaks 3.1 package and manual, containing algorithm descriptions, usage instructions and examples, are available at http://www.bcgsc.ca/platform/bioinfo/software/findpeaks Source files for FindPeaks 3.1 are available for academic use.

Contact: afejes@bcgsc.ca

1 INTRODUCTION

Next-generation sequencing technologies have begun to yield results that reflect the paradigm-shifting analytic power of cost-effective, massively parallel short-read enumeration of DNA or cDNA fragments. Publications indicate the potential of methods for detecting the genomic locations of DNA-binding proteins through ChIP-Seq experiments (Barski *et al.*, 2007; Johnson *et al.*, 2007; Robertson *et al.*, 2007), and novel applications such as the identification of miRNA sequences (Morin *et al.*, 2008), and the analysis of other enrichment strategies (Schones *et al.*, 2008; Taylor *et al.*, 2007). Additional new applications of short-read sequencing technologies will likely be developed.

For all applications outside of *de novo* genome sequence assembly, the first step in analysis is to align each sequence read to a reference genome in order to identify the location from which the sequence originated. Several programs are available for this (e.g. http://maq.sourceforge.net, http://www.ebi.ac.uk/∼guy/exonerate).

Many types of experiments also seek to identify locations in which sequence reads cluster into 'regions of enrichment', in which overlapping reads appear as peaks. For ChIP-Seq experiments, such areas represent *in vivo* locations where proteins of interest (e.g. modified histones or transcription factors) were associated with DNA; for transcriptome experiments, they correspond to the locations of transcribed exons.

Previously, enriched regions in ChIP-Seq experiments have been identified by enumerating genome-wide profiles of extended virtual fragments based upon the aligned position of the short sequencing reads. Several unpublished software packages are available for this goal ('Minimal ChipSeq Peak Finder', Wold lab, Caltech, 'Chip-Seq Data Analysis', Illumina Inc.). To locate and analyze such regions for ChIP-Seq experiments (Barski *et al.*, 2007; Johnson *et al.*, 2007; Robertson *et al.*, 2007), as well as for other enrichment-based protocols, we present a cross-platform software application that offers and extends this functionality and provides new features that can be used to develop new enrichment-based experiments.

2 METHODS

Architecture: One of the key features of FindPeaks 3.1 is a simplified, coherent software architecture that allows for rapid development of new features and excellent performance. The modular architecture allows for simple addition or modification of functions that address, for example, estimating false discovery rates or fragment length distributions, refining peaks and reading different aligned-read file formats.

Input formats: FindPeaks 3.1 is able to work directly with output from several aligners. All sequences read are internally mapped to a common software object that represents aligned sequence reads and is agnostic to the file format supplied. Currently, Exonerate (Slater and Birney, 2005) and Eland (Illumina Inc.) file formats can be translated and read into the common software object.

Directional reads: When enabled, this module analyzes each enriched region to identify DNA fragments that likely contribute to the formation of a peak. Thus, fragments observed on the forward strand occurring after a peak, or before a peak on the reverse strand are considered to be non-specific 'noise' and not included in the peak height.

Sub-peak identification: FindPeaks 3.1 is able to identify and separate multiple peaks within a complex region of overlapping DNA sequence

[*] To whom correspondence should be addressed.

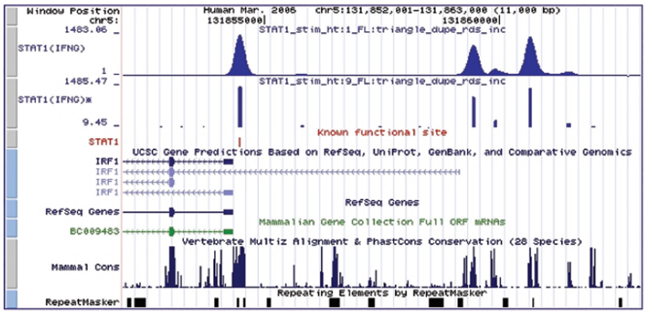
Window Position
chr5:
Human Mar. 2006 chr5:131,852,001-131,863,000 (11,000 bp)
131855000|
131860000|
1483.06 _
STAT1_stim_ht:1_FL:triangle_dupe_rds_inc
STAT1(IFNG)
1 _
1485.47 _
STAT1_stim_ht:9_FL:triangle_dupe_rds_inc
STAT1(IFNG)*
9.45 _
Known functional site
STAT1
UCSC Gene Predictions Based on RefSeq, UniProt, GenBank, and Comparative Genomics
IRF1
IRF1
IRF1
IRF1
RefSeq Genes
RefSeq Genes
Mammalian Gene Collection Full ORF mRNAs
BC009483
Vertebrate Multiz Alignment & PhastCons Conservation (28 Species)
Mammal Cons
Repeating Elements by RepeatMasker
RepeatMasker

Fig. 1. FindPeaks 3.1 has several functions for refining peak locations, which can be employed to obtain narrow boundaries of the peaks, likely to contain the sites of interest. The top line (STAT1(IFNG)) shows the unprocessed peaks obtained from a ChIP-Seq experiment (Robertson *et al*., 2007). The line below (STAT1(IFNG)*) shows the refined peaks, using several features of the FindPeaks 3.1 application, yielding narrow peaks which corresponds well to the known binding site (red).

alignments, given a user-provided value for the minimum relative valley depth required between adjacent peaks.

Peak trimming: FindPeaks provides a method for trimming peaks to retain only the core region of interest, allowing the user to refine peaks and create more visually intuitive wig files. For ChIP-Seq experiments with well-defined peaks (e.g. transcription factors), the combination of trimming and directional peak identification is well suited to identifying likely binding sites.

DNA fragment length distributions: Several distribution modes are available for FindPeaks, to expand the number of possible experiments that can be performed.

- *Fixed length:* This mode allows the user to preset the mean fragment size. This can be useful if an estimate of the fragment size is available.
- *Triangle:* This mode assumes that sequence reads were obtained from a gel size selection step that resulted in DNA fragments having minimum and maximum sizes of ∼100 bp and ∼300 bp, respectively, and a user selectable average fragment size within that range, adjustable at runtime. It is used for calculating adaptive distributions and is the standard mode for FindPeaks use.
- *Adaptive:* This mode can estimate the distribution of fragment lengths. While it is impossible to determine the original fragment lengths with single-end read data (as opposed to paired-end data), fragment lengths will be reflected in the observed distribution. To obtain this distribution, a single chromosome is scanned to identify the maximum peak heights of each area of enrichment, and the locations of the fragment ends relative to these maxima are accumulated. To filter out background noise, the width of the observed regions is extended to 500 bp on either side of the peak height. A cumulative distribution is then generated from this data, from which the region from 400–500 bp is then assumed not to be part of the peak, and the slope of that region can be used to determine the contribution of non-specific binding. This is subtracted and the remaining distribution is used.
- *Native:* This mode uses the length of the sequenced fragment to determine coverage. Only sequenced bases are included, ideal for visualization of reads, where alignments may not all have the same length.

Graphical representation: The FindPeaks program can be used to create visual representations of all forms of genome-aligned sequence data in the form of a UCSC web browser compatible wiggle (.wig) file. For genomes that have not been included in the UCSC browser, FindPeaks can also be utilized to generate an R script that creates postscript illustrations of each chromosome.

Monte Carlo-based false discovery rate modules: This module performs a Monte Carlo simulation using the effective genome size and the number of tags used in the analysis to determine the likelihood of observing a peak of a given height. The user can set the number of iterations performed. For most ChIP-Seq datasets, convergence of the simulations is observed rapidly, often on the first iteration.

AlignSlice: For more detailed investigation of a small region of chromosome, the stand-alone AlignSlice application exists as part of the FindPeaks package, which searches one or more source files for sequencing reads that overlap a specified region, producing either a text file or UCSC browser compatible wig file. Additionally, if read mappability files are available, a separate track will be shown.

Bed file generation: Tools have also been included for creating BED files for visualization of raw alignment results through the UCSC Genome Browser.

ACKNOWLEDGEMENTS

S.J.M.J. is a scholar of the Michael Smith Foundation for Health Research.

Funding: A.P.F. is supported by the Michael Smith Foundation for Health Reasearch. Funding was provided in part by Genome Canada and Genome British Columbia.

Conflict of Interest: none declared.

REFERENCES

Barski,A. *et al.* (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.

Johnson,D.S. *et al.* (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science*, **316**, 1497–1502.

Lander,E.S. and Waterman,M.S. (1988) Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics*, **2**, 231–239.

Morin,R.D. *et al.* (2008) Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Res.*, **18**, 610–621.

Robertson,G. *et al*. (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods*, **4**, 651–67l.

Schones,D.E. *et al.* (2008) Dynamic regulation of nucleosome positioning in the human genome. *Cell*, **132**, 887–898.

Slater,G.S. and Birney,E. (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, **6**, 31.

Taylor,K.H. *et al.* (2007) Ultradeep bisulfite sequencing analysis of DNA methylation patterns in multiple gene promoters by 454 sequencing. *Cancer Res.*, **67**, 8511–8518.