BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 19 2012, pages 2517–2519
doi:10.1093/bioinformatics/bts463

Genome analysis

Advance Access publication July 24, 2012

Nebula—a web-server for advanced ChIP-seq data analysis

Valentina Boeva[1,2,3,*], Alban Lermine[1,2,3], Camille Barette[1,2,3], Christel Guillouf[1,4] and Emmanuel Barillot[1,2,3]

[1]Institut Curie, [2]INSERM, U900, Bioinformatics and Computational Systems Biology of Cancer, Paris 75248, [3]Mines ParisTech, Fontainebleau 77300, and [4]INSERM, U830, Genetics and Biology of Cancers, Paris 75248, France

Associate Editor: Michael Brudno

ABSTRACT

Motivation: ChIP-seq consists of chromatin immunoprecipitation and deep sequencing of the extracted DNA fragments. It is the technique of choice for accurate characterization of the binding sites of transcription factors and other DNA-associated proteins. We present a web service, Nebula, which allows inexperienced users to perform a complete bioinformatics analysis of ChIP-seq data.

Results: Nebula was designed for both bioinformaticians and biologists. It is based on the Galaxy open source framework. Galaxy already includes a large number of functionalities for mapping reads and peak calling. We added the following to Galaxy: (i) peak calling with FindPeaks and a module for immunoprecipitation quality control, (ii) *de novo* motif discovery with ChIPMunk, (iii) calculation of the density and the cumulative distribution of peak locations relative to gene transcription start sites, (iv) annotation of peaks with genomic features and (v) annotation of genes with peak information. Nebula generates the graphs and the enrichment statistics at each step of the process. During Steps 3–5, Nebula optionally repeats the analysis on a control dataset and compares these results with those from the main dataset. Nebula can also incorporate gene expression (or gene modulation) data during these steps. In summary, Nebula is an innovative web service that provides an advanced ChIP-seq analysis pipeline providing ready-to-publish results.

Availability: Nebula is available at http://nebula.curie.fr/

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

Received on April 20, 2012; revised on June 27, 2012; accepted on July 16, 2012

1 INTRODUCTION

Chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) is a technique that allows identification of binding sites of DNA-associated proteins.

Most of the existing tools for ChIP-seq data analysis are ifficult to use by non-bioinformaticians. These tools map sequenced reads to the reference genome (Homer *et al.,* 2009; Langmead *et al.*, 2009) or predict binding site locations (ChIP-seq peaks) (Fejes *et al.*, 2008; Wilbanks and Facciotti, 2010, for a review; Zhang *et al.*, 2008). Several tools exist for peak filtering (Boeva *et al.*, 2010), motif discovery (Kulakovskiy *et al.*, 2010) and genome feature association (Shin *et al.*, 2009). Such tools are often command-line applications, R packages or their use on the web is restricted to a limited number of sequences/data quantity. ChIP-seq data analysis solutions offering a graphical interface such as the main Galaxy server (Goecks *et al.*, 2010), Cistrome (Liu *et al.*, 2011), CisGenome (Ji *et al.*, 2008) and the web version of CEAS (Ji *et al.*, 2006) can be very useful (in particular for biologists) when performing the first steps of ChIP-seq data analysis (Galaxy main) or further analysis of predicted binding sites (Cistrome, CisGenome and CEAS). However, none of these solutions provides the full set of tools needed for a complete ChIP-seq data analysis (Supplementary Table S1).

Our goal was to develop a framework in which biologists could analyze their ChIP-seq data with minimal help of bioinformaticians, from read mapping to the analysis of binding site properties. However, bioinformaticians can also benefit from using such a framework.

2 METHODS

Our web service, Nebula, is based on the Galaxy open source framework (Goecks *et al.*, 2010), which is highly used by the research community (Liu *et al.*, 2011). The Nebula pipeline contains about 20 tools. It allows read mapping, peak calling, peak/gene annotation and *de novo* motif discovery.

We used Perl and R to develop the tools that (i) calculate and visualize peak height distribution, (ii) calculate and visualize density and cumulative distribution of peak locations relative to gene transcription start sites (TSSs) provided by RefSeq (Pruitt *et al.*, 2012), (iii) annotate peaks with genomic features, (iv) annotate genes with peak information and (v) find genomic categories (promoter, enhancer, etc.) enriched in binding events. All these tools accept as a second input a file with control peak locations, so that the user can compare distributions calculated for ChIP and control experiments (see Supplementary Methods for more detail). In addition, tools (2–5) can use information about gene expression or gene modulation. This allows separate characterization of binding in expressed/silenced genes or in genes activated/repressed/non-modulated by a transcription factor (TF) of interest. The pipeline also includes several published tools (samtools, MACS, FindPeaks and ChIPMunk).

The pipeline can be used via Internet or can be downloaded and installed on a personal computer or server. A detail tutorial and a toy example are provided on the Nebula main page. The Nebula pipeline can be run from beginning to end in one stroke or each step can be run independently. Users can also build workflows based on their personal needs.

*To whom correspondence should be addressed.

© The Author 2012. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions.com

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

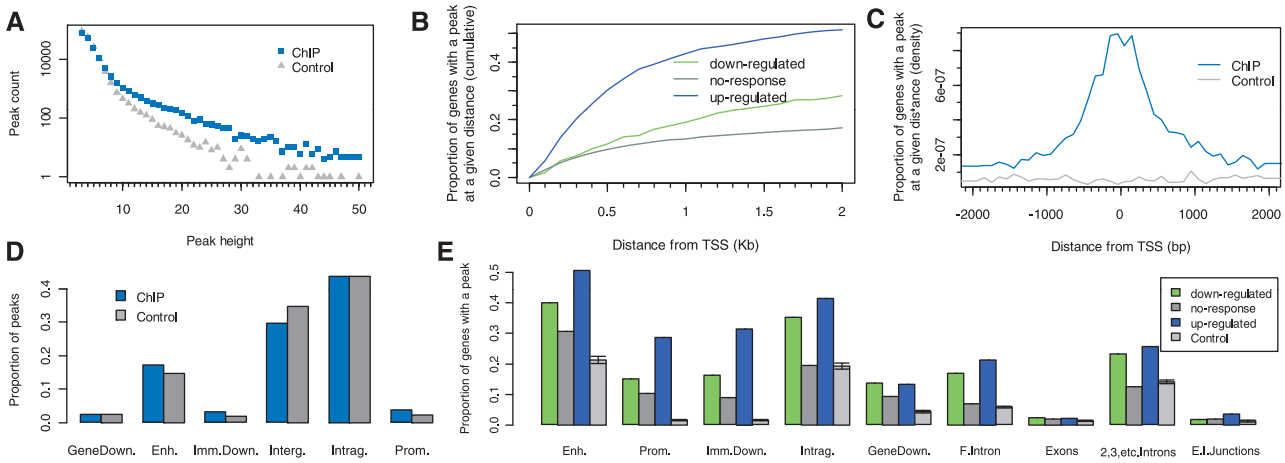

Fig. 1. Six of the 16 graphs produced by Nebula. (**A**) Peak height distribution; (**B**) cumulative proportion of genes with a peak within a given distance from TSS; (**C**) density of peak distribution relative to TSS; (**D**) peak frequency in each genomic category and (**E**) enrichment of each genomic category in peaks

3 RESULTS

We developed a pipeline for advanced ChIP-seq data analysis that (i) is web-based, (ii) modular, (iii) includes several tools for data quality control and (iv) produces graphical output and necessary statistics.

3.1 Nebula functionalities

3.1.1 Read mapping The primary step of ChIP-seq data analysis is read mapping. We provide capacities to do read mapping with Bowtie (Langmead *et al.*, 2009). Bowtie is an ultrafast gapless aligner, which is optimal for mapping ChIP-seq reads. The standard format of raw reads accepted by Bowtie is 'fastq'. Thus, we added a tool, which converts SOLiD 'csfasta' and 'qual' files to 'fastq' files (Homer *et al.*, 2009). When reads are mapped, the user can get information about the number and quality of reads using tools 'flagstat' (Li *et al.*, 2009) and 'FASTQC' (www.bioinformatics.babraham.ac.uk/projects/fastqc/).

3.1.2 Peak calling For prediction of binding sites (peak calling), we implemented FindPeaks (Fejes *et al.*, 2008) and maintained MACS (Zhang *et al.*, 2008), already existing in Galaxy. Findpeaks and MACS represent two families of peak calling tools: the former is based on tag extension and the latter on tag shift. We also developed a strategy to assess antibody quality and evaluate the false discovery rate (FDR) using FindPeaks. Briefly, we create a subset of control (input or nonspecific IgG) reads which contains the same number of reads as the ChIP dataset; we apply FindPeaks on both datasets and calculate peak height distribution (Fig. 1A). Under the hypothesis that the peak height distribution for the control subset reflects the level of noise in the ChIP dataset, we can evaluate FDR of a subset of ChIP peaks higher than a given threshold T as the ratio of number of peaks higher than T in the control over the number of peaks higher than T in the ChIP dataset. Duplicate reads are excluded from the analysis. The user can also perform local filtering by removing ChIP peaks overlapping with peaks in the control.

3.1.3 Peak and gene annotation Nebula calculates peak location distribution relative to gene TSSs (Fig. 1B and C). Using gene expression or gene modulation data, Nebula can separate curves for expressed/silenced genes or genes activated/inhibited/non-modulated by a given TF. When a user specifies boundaries of genomic categories (promoter, enhancer, gene downstream region, etc.), Nebula calculates the proportion of peaks falling in each category (Fig. 1D). For each gene, Nebula identifies peaks falling in each genomic category (Fig. 1E) and calculates enrichment relative to the control and/or to the average peak distribution (data not shown). The user can choose to apply bootstrapping to attain enrichment *P*-values.

3.1.4 De novo motif discovery With Nebula, the user can run *de novo* motif finding in the whole set of peak sequences or in the areas centered on peak summits. The ChIPMunk tool (Kulakovskiy *et al.*, 2010) provided for this purpose allows two modes for finding multiple motifs. Mode 'mask' hides already identified motifs before each subsequent round of motif discovery and mode 'filter' eliminates complete sequences containing already identified motifs. In addition, the user can select for motif discovery peaks falling in a given genomic category, e.g. peaks in promoters of TF-activated genes or enhancers of TF-repressed genes.

3.2 Application to Spi-1 ChIP-seq data

We applied Nebula to ChIP-seq data for Spi-1 in mouse erythroleukemic cells (Ridinger-Saison *et al.*, unpublished data). We predicted 17 781 Spi-1 binding sites. Out of 21 predictions tested using ChIP-quantitative PCR, 20 were validated. We obtained data about genomic regions preferentially bound by Spi-1 and Spi-1 binding motifs. We showed that the position of Spi-1 binding influences transcriptional activation of corresponding genes.

Funding: The 'Projet Incitatif et Collaboratif Bioinformatique et Biostatistiques' of the Institut Curie, and the Ligue Nationale Contre le Cancer.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

Conflict of Interest: none declared.

REFERENCES

Boeva,V. *et al.* (2010) *De novo* motif identification improves the accuracy of predicting transcription factor binding sites in ChIP-Seq data analysis. *Nucleic Acids Res.*, **38**, e126.

Fejes,A. *et al.* (2008) FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics*, **24**, 729–1730.

Goecks,J. *et al.* (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.*, **11**, R86.

Homer,N. *et al.* (2009) BFAST: an alignment tool for large scale genome resequencing. *PLoS One*, **4**, e7767.

Ji,H. *et al.* (2008) An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat. Biotechnol.*, **26**, 1293–1300.

Ji,X. *et al.* (2006) CEAS: cis-regulatory element annotation system. *Nucleic Acids Res.*, **34 (Web Server issue)**, W551–W554.

Kulakovskiy,I. *et al.* (2010) Deep and wide digging for binding motifs in ChIP-Seq data. *Bioinformatics*, **26**, 2622–2623.

Langmead,B. *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

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

Liu,T. *et al.* (2011) Cistrome: an integrative platform for transcriptional regulation studies. *Genome Biol.*, **12**, R83.

Pruitt,K.D. *et al.* (2012) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.*, **40 (Database issue)**, D130–D135.

Ridinger-Saison,M. *et al.* (2012) Spi-1/PU.1 activates transcription through clustered DNA occupancy in erythroleukemia. *Nucleic Acids Res.*, [Epub ahead of print] PubMed PMID: 22790984.

Shin,H. *et al.* (2009) CEAS: cis-regulatory element annotation system. *Bioinformatics*, **25**, 2605–2606.

Zhang,Y. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.

Wilbanks,E.G. and Facciotti,M.T. (2010) Evaluation of algorithm performance in ChIP-seq peak detection. *PLoS One*, **5**, e11471.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016