Background & 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. 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.

We present a web service, Nebula, which allows inexperienced users to perform a complete bioinformatics analysis of ChIP-seq data.

Methods & Results

Nebula 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:

- peak calling with FindPeaks and a module for immunoprecipitation quality control,
- de novo motif discovery with ChIPMunk,
- calculation of the density and the cumulative distribution of peak locations around gene TSSs,
- annotation of peaks with genomic features,
- annotation of genes with peak information.

Read mapping. We provide capacities to do read mapping with Bowtie (Langmead et al., 2009). 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/).

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.

Peak and gene annotation. Nebula calculates peak location distribution relative to gene TSSs. 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. For each gene, Nebula identifies peaks falling in each genomic category and calculates enrichment relative to the control and/or to the average peak distribution. The user can choose to apply bootstrapping to attain enrichment p-values.

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. Also, 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.

Conclusions

We applied Nebula to ChIP-seq data for transcription factor Spi-1 in mouse erythroleukemic cells (Ridinger-Saison et al., in press). 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.

In summary, Nebula is an innovative web service that provides an advanced ChIP-seq analysis pipeline providing ready-to-publish results.