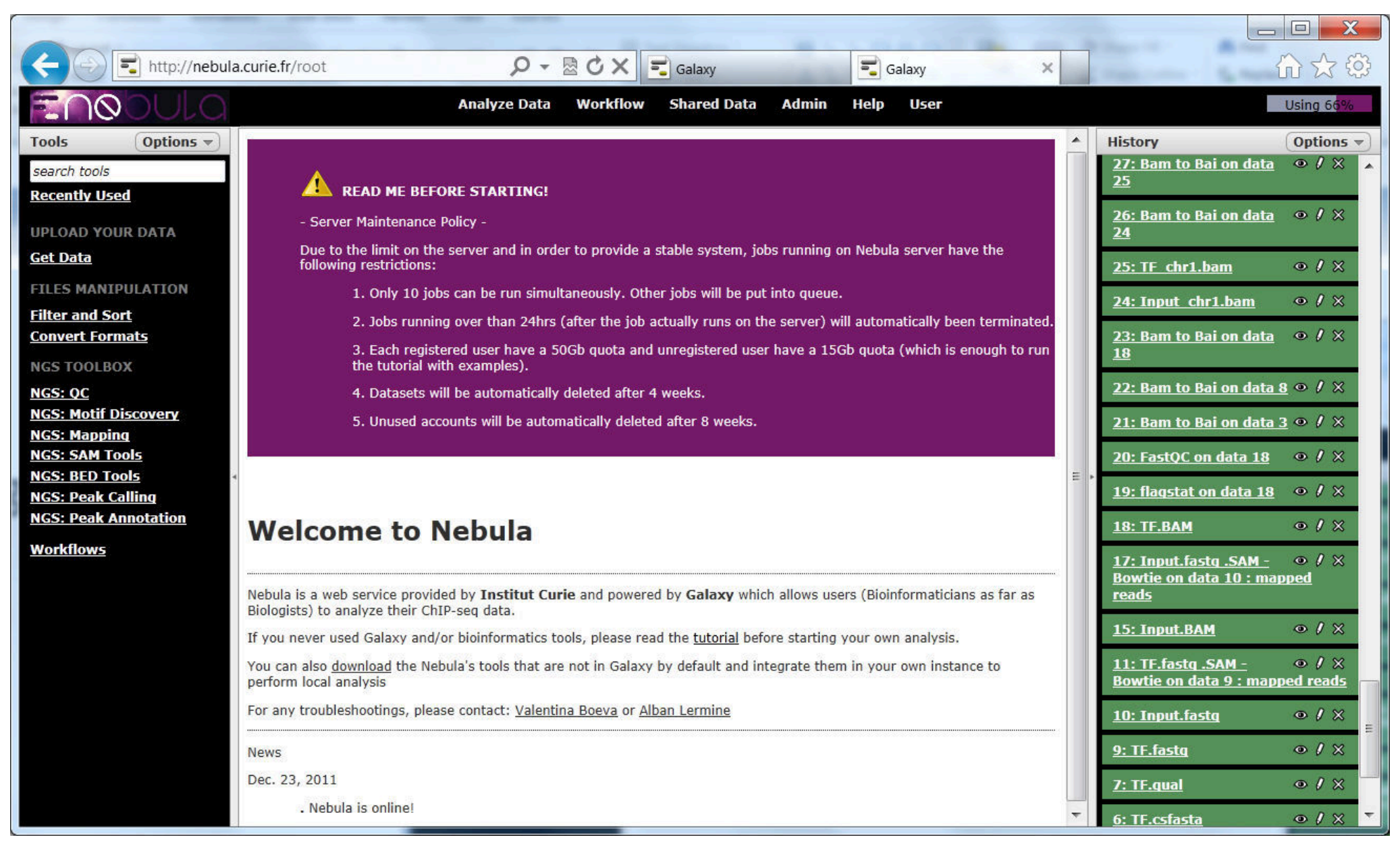
Nebula – a Web-Server for Advanced ChIP-Seq Data Analysis

Authors: Valentina Boeva, Alban Lermine and Emmanuel Barillot
Affiliation: Inserm U900, Mines ParisTech, Institut Curie, Paris, France
E-mails: valentina.boeva@curie.fr, alban.lermine@curie.fr

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.



READ ME BEFORE STARTING!

- Server Maintenance Policy -

Due to the limit on the server and in order to provide a stable system, jobs running on Nebula server have the following restrictions:

1. Only 10 jobs can be run simultaneously. Other jobs will be put into queue.
2. Jobs running over than 24hrs (after the job actually runs on the server) will automatically been terminated.
3. Each registered user have a 50Gb quota and unregistered user have a 15Gb quota (which is enough to run the tutorial with examples).
4. Datasets will be automatically deleted after 4 weeks.
5. Unused accounts will be automatically deleted after 8 weeks.

Welcome to Nebula

Nebula is a web service provided by **Institut Curie** and powered by **Galaxy** which allows users (Bioinformaticians as far as Biologists) to analyze their ChIP-seq data.

If you never used Galaxy and/or bioinformatics tools, please read the tutorial before starting your own analysis.

You can also download the Nebula's tools that are not in Galaxy by default and integrate them in your own instance to perform local analysis

For any troubleshootings, please contact: Valentina Boeva or Alban Lermine

News

Dec. 23, 2011

. Nebula is online!

27: Bam to Bai on data 25
26: Bam to Bai on data 24
25: TF_chr1.bam
24: Input_chr1.bam
23: Bam to Bai on data 18
22: Bam to Bai on data 8
21: Bam to Bai on data 3
20: FastQC on data 18
19: flagstat on data 18
18: TF.BAM
17: Input.fastq .SAM - Bowtie on data 10 : mapped reads
15: Input.BAM
11: TF.fastq .SAM - Bowtie on data 9 : mapped reads
10: Input.fastq
9: TF.fastq
7: TF.qual
6: TF.csfasta

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:

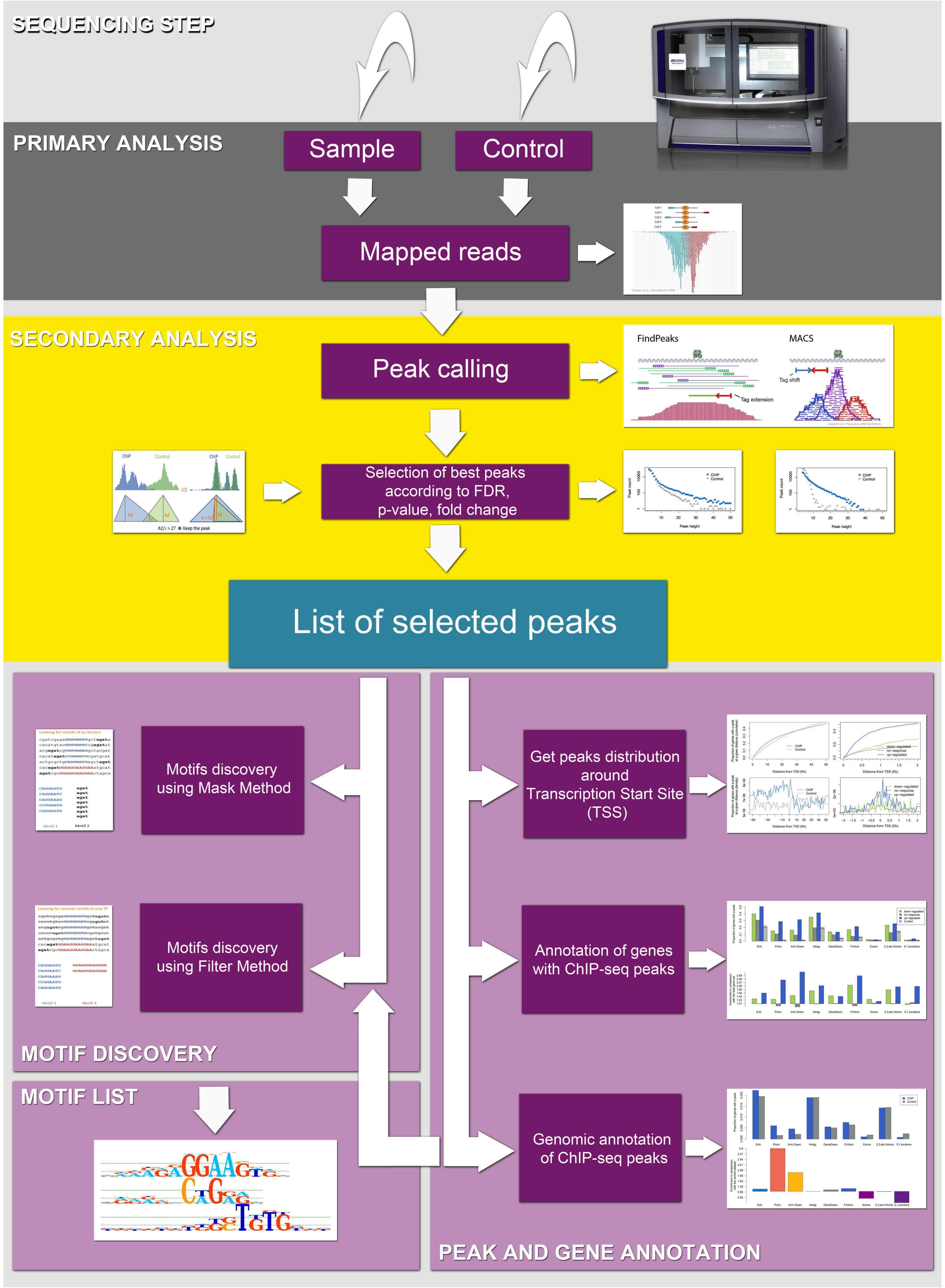
- peak calling with FindPeaks and a module for immunoprecipitation quality control,
- *de novo* motif discovery with ChIPMunk,
- calculation of the density and the cumulative distribution of peak locations around gene TSSs,
- annotation of peaks with genomic features,
- annotation of genes with peak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


SEQUENCING STEP

PRIMARY ANALYSIS

Sample

Control

Mapped reads

SECONDARY ANALYSIS

Peak calling

Selection of best peaks according to FDR, p-value, fold change

List of selected peaks

Motifs discovery using Mask Method

Motifs discovery using Filter Method

MOTIF DISCOVERY

MOTIF LIST

Get peaks distribution around Transcription Start Site (TSS)

Annotation of genes with ChIP-seq peaks

Genomic annotation of ChIP-seq peaks

PEAK AND GENE ANNOTATION

Conclusions

We applied Nebula to ChIP-seq data for transcription fact 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.