A web-based service for ChIP-Seq data analysis

D’Onorio De Meo P[1], D’Antonio M[1], Castrignanò T[1], Zambelli F[2], Pavesi G[2], Pesole G[3,4]

Motivation

The ChIP-seq technology (Chromatin ImmunoPrecipitation methods coupled with massive parallel sequencing) has amazingly fostered the genome-wide studies of DNA-protein interactions in vivo. However, the handling and the analysis of the huge amount of raw data and/or the short reads produced (over 10 million reads per single sequencing run, up to 75 bp long) requires non trivial skills, huge computer power and storage capacity which are generally not available in most research labs. A high-performance computing Chip-seq pipeline is thus needed to manage the size and complexity of sequence data.

Methods

The programs included in the pipeline are implemented in the CASPUR parallel environment. The steps of the pipeline can be summarized as the following: 1) Image Analysis; 2) Base calling; 3) Mapping sequence reads across the genome; 4) Peak detection using experimental or simulated backgrounds; 5) Peak functional classification based on their localization with respect the genome annotation (e.g. core promoters, intragenic, etc.). In particular, this pipeline has been fine tuned for the analysis of Illumina data, and the first three steps take advantage of the Illumina “Genome analyzer pipeline software”, parallelizing the analysis run to multiple simultaneous processes and allowing the automatic load-sharing to multiple CPUs. The procedure followed for peak detection and significance assessment merges and compares the results of different methods, according to the number of replicates, presence/absence of a control experiment or “input” DNA sequencing, and so on. A database management system has been used to manage peak data and optimized PHP scripts implemented for result retrieval.

Results

Conspicuous CASPUR hardware resources have been dedicated to this analysis service both in term of computational power and storage capacity. Our pipeline has been fully integrated with our cluster resources management middleware, both by optimizing the performance on the computing nodes and by maximizing job throughput to obtain faster results and lesser queue waiting. The system integrates

[1] Consorzio per le Applicazioni di Supercalcolo per Università e Ricerca, Rome, Italy [2] Dipartimento di Scienze Biomolecolari e Biotecnologie, University of Milan, Milan, Italy [3] Istituto Biomembrane e Bioenergetica, Consigli Nazionale delle Ricerche, Bari, Italy [4] Dipartimento di Biochimica e Biologia Molecolare, University of Bari, Bari, Italy

access to a large database of human genomics data with basic analytical visualization tools. All results are available through a web interface which provides a summary of input data, textual and graphical information of the mapping results, and the list of significant peaks using dynamic significance thresholds as well as a genome browser showing peaks in the sample and control under investigation in their genomic context. Finally, the retrieval system allows the selection of specific peak collections based on their statistical significance with respect to the background, on their genomic location (e.g. gene name, core-promoter, intragenic, intergenic, etc.).

Contact e-mail

donorio@caspur.it