Computational pipeline for automatic chromatin state identification in multiple tissues

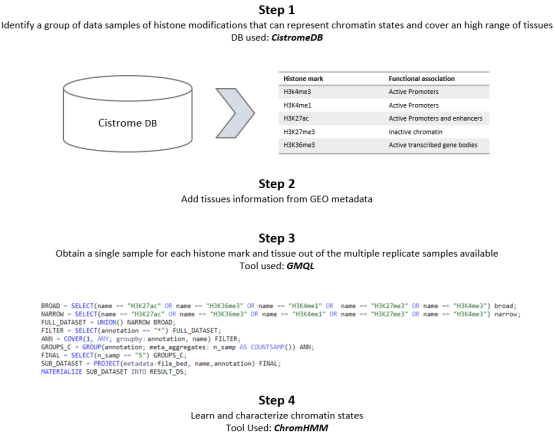
Leone M(1)† , Galeota E(2), Pelizzola M(2), Ceri S(1), Masseroli M(1)

(1) Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB), Politecnico di Milano, Milano
(2) Center for Genomic Science of IIT, Fondazione Istituto Italiano di Tecnologia, Milano

† Email: michele.leone@polimi.it

Motivation
Inside the cell, DNA is almost always associated with proteins. In fact, to form nucleosomes, the basic building blocks of chromatin, DNA sequences of about 150 base pairs are wrapped around octets of histone proteins [1]. In the last decades, researchers have started cataloguing chromatin proteins and their modifications. Chromatin, once considered as a simple scaffold to package DNA into each cell, has been recognized to have a dynamic role in genome organization and a multiplicity of functions in genome regulation. Lots of studies have been carried out with the aim of simplifying chromatin complexity by dividing it into a certain number of chromatin-states. This has led to the identification of a number of chromatin modifications or "marks" and the discovery of many regulatory elements throughout the genome [2]. Probably, in the next future, chromatin-states mapping will reveal a multitude of key aspects of genome functions and will pave the path to understand the mechanisms that regulate these functions. The aim of this work, based on the previous considerations, is to provide a computational pipeline to identify chromatin states in different tissues, in order to better understand conservation and differentiation of chromatin states in tissues.
Methods
Chromatin state annotation using combinations of chromatin modifications has emerged as a powerful approach for genome annotation and detection of regulatory activity [3]. A preliminary step of the pipeline developed in this work is to identify a set of available data samples of histone modifications that can accurately characterize chromatin states and cover a wide range of tissues. Another crucial point is to find information about the tissues in which the considered histone modifications are expressed. Based on the obtained information, available samples are grouped by histone modification and tissue annotation terms, leading to multiple samples for each group. The next step, thus, involves obtaining a single sample for each histone mark and tissue out of the multiple replicate samples available. Finally, the last step is to learn chromatin states and characterize their biological functions and correlations, with the possibility to visualize the resulting genome-wide maps of chromatin-state annotations.
Results
Following the described pipeline, we performed a chromatin states analysis in different tissues. A large number of ChIP-seq and chromatin accessibility data are

available in the GEO and ENA repositories. However, due to inconsistencies in the annotation of metadata, as well as the lack of uniform processing procedures, these resources, though precious, have been underutilized. To overcome this problem, we took advantage of Cistrome DB (http://cistrome.org/db), a comprehensive annotated resource of ChIP-seq and chromatin accessibility data publicly available in GEO. As a set of histone modifications that could accurately characterize chromatin states, we used the first five histone marks most represented in Cistrome DB. After enriching samples with tissues 'information from GEO metadata, to obtain one single sample for each histone mark and tissue out of the multiple replicate samples available, data were processed using pyGMQL, the python package of the GenoMetric Query Language [4], a new holistic approach to manage and query a large number of datasets, samples, DNA regions and metadata in order to discover interesting DNA regions and their relationships. Lastly, to learn and characterize chromatin states, the ChromHMM [5] tool was used. As a result of this pipeline, we obtained genome-wide information about 5 chromatin states for 80 different tissue types.



References

1. Kornberg RD. Chromatin structure: a repeating unit of histones and DNA. Science, 1974; 184: 868–871.

2. Heintzman, HD, et al. DNA methylation signatures link prenatal famine exposure to growth and metabolism Nature, 2009; 459: 108–112.

3. Filion GJ, et al. Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. Cell, 2010; 143: 212–224.

4. Masseroli M., et al. GenoMetric Query Language: A novel approach to large-scale genomic data management, Bioinformatics, 2015; 28: 691-693.

5. Ernst J, Kellis M. Chromatin-state discovery and genome annotation with ChromHMM. Nature Protocols,

2017; 12: 2478-2492