PIBIC Project

Identification of transcription regulatory elements using FAIRE-Seq data without reference genome of the same species

Responsible PI: **Dr. rer. nat. Diego Mauricio Riano-Pachon**
CNPEM unit: **Laboratorio Nacional de Ciencia e Tecnologia do Bioetanol (CTBE)**

Background

The identification of active transcriptional regulatory elements is an important step towards the understanding of transcriptional regulation [1-3]. One of the first step required to transcriptional activation is the release of the DNA form the chromatin, i.e., the local depletion of nucleosomes. It has been previously shown that cis-regulatory elements co-located with nucleosome depleted regions [2, 4]. One approach to identify those regions is the use of formaldehyde-assisted isolation of regulatory elements (FAIRE). In FAIRE, chromosomal regions are cross-linked to histones by formaldehyde treatment [5-7]. Thus, regulatory DNA regions would be open and barely cross-linked to proteins. Such genome segments can now be analyzed by quantitative PCR, hybridization to genome tiling arrays or next generation sequencing (NGS). These assays can be used to identify differences in chromatin structure across cell types or conditions. Differentially nucleosome-depleted regions identified, under certain conditions, could lead to the discovery of novel regulatory elements [8-9].

So far, FAIRE assays have been applied to organisms whose genome sequence is known. However, obtaining the genome sequence for economically important organisms is still a challenge, e.g, Sugarcane; due to several factors, including, high levels of ploidy, highly repetive nature, among others. There is a need to develop new methods using NGS technologies to uncover different characteristics on those genomes, e.g., transcriptional regulatory elements. Here we will explore the use of FAIRESeq to achieve this. We count with FAIRESeq data from the green algae *Chlamydomonas reinhardtii* and from the angiosperm model plant A*rabidopsis thaliana*. The genome sequence of those two species is fully known and well annotated. This will allow us to evaluate the power of FAIRESeq data to uncover transcriptional regulatory elements without the use of a reference genome. We will also evaluate the use of the genome sequence of a related species as a reference.

Objectives

General

Evaluate the power of FAIRESeq data in a case where no reference genome for the same species is available

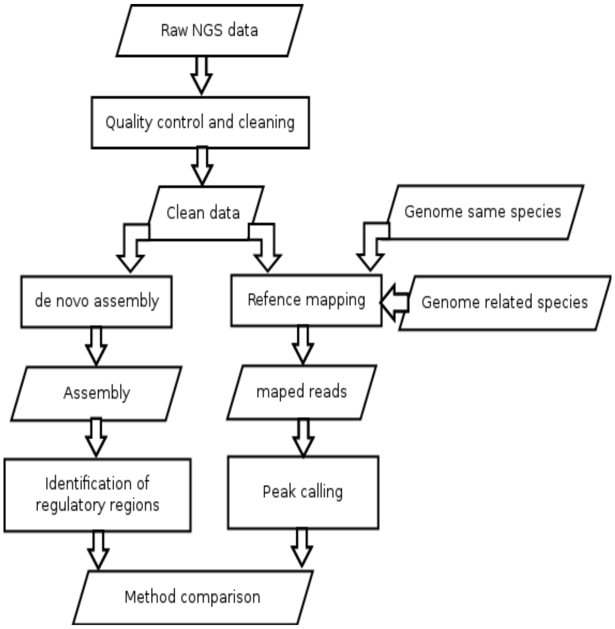
Specific

1. To familiarize with the pre-processing and processing of NGS data for the identification of transcription regulatory regions.
2. To *de novo* assemble FAIRESeq data in different species (*Chlamydomonas reinhardtii* and *Arabidopsis thaliana*) aiming at the identification of regulatory regions.
3. To run an existing pipeline for the traditional, i.e., using a reference genome for the same species, identification of FAIRESeq data.
4. To implement a workflow for the identification of FAIRESeq peaks using as reference the genome of a related species.
5. To compare the results of the different approaches (objectives 2, 3 and 4)

Methods

This projects has the following steps, for each we provide the name of the software packages that can be used.

1. Pre-processing of raw NGS data
   - FASTQC
   - FASTX Toolkit
   - NGS Backbone
2. *de novo* assembly of clean NGS data
   - Velvet
   - SOPAdenovo
3. Mapping of NGS data to reference genomes
   - Bowtie
   - BWA
   - SHRiMP
4. Peak Calling
   - MOSAICS
   - CSAR
   - MACS
5. Comparison of methods
   - Venn diagrams
   - BED tools
   - Matthews correlation



References

[1] Thurman RE, Rynes E, Humbert R, et al.: The accessible chromatin landscape of the human genome. Nature 2012, 489(7414):75–82, [[http://dx.doi.org/10.1038/nature11232]].

[2] Boyle AP, Song L, Lee BK, London D, Keefe D, Birney E, Iyer VR, Crawford GE, Furey TS: High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. Genome Res 2011, 21(3):456–464, [[http://dx.doi.org/10.1101/gr.112656.110]].

[3] Narlikar L, Ovcharenko I: Identifying regulatory elements in eukaryotic genomes. Brief Funct Genomic Proteomic 2009, 8(4):215–230, [[http://dx.doi.org/10.1093/bfgp/elp014]].

[4] Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van CalcarS, Qu C, Ching KA, Wang W, Weng Z, Green RD, Crawford GE, Ren B: Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet 2007, 39(3):311–318, [[http://dx.doi.org/10.1038/ng1966]].

[5] Nagy PL, Cleary ML, Brown PO, Lieb JD: Genomewide demarcation of RNA polymerase II transcription units revealed by physical fractionation of chromatin. Proc Natl Acad Sci U S A 2003, 100(11):6364–6369, [[http://dx.doi.org/10.1073/pnas.1131966100]].

[6] Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD: FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome Res 2007, 17(6):877–885, [[http://dx.doi.org/10.1101/gr.5533506]]. [10] Giresi PG, Lieb JD: Isolation of active regulatory elements from eukaryotic chromatin using FAIRE (Formaldehyde Assisted Isolation of Regulatory Elements). Methods 2009, 48(3):233–239, [[http://dx.doi.org/10.1016/j.ymeth.2009.03.003]].

[7] Simon JM, Giresi PG, Davis IJ, Lieb JD: Using formaldehyde-assisted isolation of regulatory elements (FAIRE) to isolate active regulatory DNA. Nat Protoc 2012, 7(2):256–267, [[http://dx.doi.org/10.1038/nprot.2011.444]].

[8] Nammo T, Rodríguez-Seguí SA, Ferrer J: Mapping open chromatin with formaldehydeassisted isolation of regulatory elements. Methods Mol Biol 2011, 791:287–296, [[http://dx.doi.org/10.1007/978-1-61779-316-5\s\do5(2)1]].

[9] Song L, Zhang Z, Grasfeder LL, Boyle AP, et al.: Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. Genome Res 2011, 21(10):1757–1767, [[http://dx.doi.org/10.1101/gr.121541.111]].