ChIP-Seq

Author: Mohamad Mahdi Ziaee
Student ID: 109 56 1770

Abstract - This document contains general information about ChIP sequencing. It explains the steps taken in order to achieve the task in details. However, there are limitations in today's algorithms used by the software built for this purposes. A new solution is to this obstacle is use DREME algorithm which is based on regular expression matches to find the motifs. Lastly, the commonly used tools are explained.

I. INTRODUCTION

Chromatin Immunoprecipitation sequencing is a one of the methods which analyses the relationship between DNA and proteins. In this method, the binding sites of DNA proteins are identified in two major steps: ChIP and Sequencing. An older version of this technology was called ChIP-on-chip. In this paper, the algorithm used in ChIP sequencing is explained. [1][2]

II. CHIP-CHIP VS CHIP-SEQ

ChIP-chip is an older technology compare to ChIP-seq. ChIP-chip is array-specific while ChIP-seq is based on single nucleotide. [1] In general, ChIP-seq produces result with better signal-to-noise ratio and it is much easier to spot narrow peaks. However, in order to produce an accurate result using ChIP-seq, high quality input DNA data is essential. [8]

III. CHiP SEQUENCING ANALYSIS

ChiP-seq analysis is under Gene regulation which is a part of DNA sequencing. There are different applications like protein-DNA interaction, nucleosome positioning and chromatin states. To achieve this analysis, 3 major steps are needed: Sequencing, Bioinformatics analysis and Downstream analysis. The first step is prepared in the lab which at the end results into raw reads in a file with fastq format. Later, the mappability is done by aligning the reads to the genome. For this step, small reads are used unless the duplicate regions are interesting for the analysis. To name some of the famous software used for mappability, BLAST and Bowtie could be used. later, peak calling and enriched areas will be done. The final result of the process is shown in different formats like genome browser or it is used to finding motifs. [6]

IIV. LIMITATIONS

Today many algorithms use Ab initio discovery algorithm to find the DNA-binding motifs and there are quite number of limitations in using this algorithm. For instance, due to the limitations of the algorithms, only a small part of the sequences can be used in the motif finding which affects the sensitivity of the overall result. Another limitation is single ChIP-seq dataset search; which means it does not allow two datasets to be used in finding the

motifs.

DREME is another algorithm which is designed to to find short motifs on very large data set. It is not only faster than other similar algorithms, but it is able to find more cofactor motifs (cofactor is a chemical compound that is needed for protein's biological activity [7]). This is achieved by looking up for motifs using regular expressions. It also limits its search to short motifs (up to 8 bp wise). [3]

V. TOOLS

There are different set of tools which can be used for ChIP sequencing. Illumina sequencer is the most commonly used tool for this purpose which maps a small region from both ends of each fragment. Base space labs uses HMAC2 to locate the enriched regions and HOVER to find the motifs. [4] Cistrome is another tool which its data set is based on *Galaxy* (an open source framework) and it has both ChIP-chip- and ChIP-seq-specific tools. [5]

REFERENCES

[1] Whole-genome Chromatin IP sequencing [online], Available:
http://www.illumina.com/Documents/products/datasheets/datasheet_chip_sequence.pdf

[2] Wikipedia [online], Available:
https://en.wikipedia.org/wiki/ChIP-sequencing

[3] Timothy L. Bailey, "DREME: motif discovery in transcription factor ChIP-seq data" *Oxford JournalsScience & Mathematics Bioinformatics* Volume 27, Issue 12Pp. 1653-1659
http://bioinformatics.oxfordjournals.org.hmlproxy.lib.csufresno.edu/content/27/12/1653.full
DOI: 10.1093/bioinformatics/btr261

[4] Dongjun Chung; Pei Fen Kuan; Bo Li; Rajendran Sanalkumar; Kun Liang; Emery H. Bresnick; Colin Dewey; Sunduz Keles, "Discovering Transcription Factor Binding Sites in Highly Repetitive Regions of Genomes with Multi-Read Analysis of ChIP-Seq" July 14, 2011
DOI: http://dx.doi.org/10.1371/journal.pcbi.1002111

[5] *Cistrome Project* [online], Available:
http://cistrome.org/Cistrome/Cistrome_Project.html

[6] *ChIP-seq analysis* [online], Available:
https://www.ebi.ac.uk/training/online/course/ebi-next-generation-sequencing-practical-course/gene-regulation/chip-seq-analysis

[7] *Cofactor (biochemistry)* [online], Available:
https://en.wikipedia.org/wiki/Cofactor_(biochemistry)

[8] Joshua WK Ho; Eric Bishop; Peter V Karchenko; Nicolas Nègre; Kevin P White; Peter J Park, "ChIP-chip versus ChIP-seq: Lessons for experimental design and data analysis", *BMC Genomics 2011*.
DOI: 10.1186/1471-2164-12-134

SNP

Author: Mohamad Mahdi Ziaee
Student ID: 109 56 1770

Abstract – In this document, single nucleotide polymorphisms are explained in simple words. Also the common methods of general approaches to find these SNPs are described. Lastly, the tools used to achieve the goal is mentioned.

I. INTRODUCTION

Single nucleotide polymorphisms (SNP), pronounced "snips", is a change in a single in DNA nucleotide which might change a gene's functionality. These variations are mostly found in DNA between different genes. SNP is currently used as a method of prediction for certain health issues and a patient response to certain drug. [1] These mutations in DNA must occur in more than 1 percentage of a large population to be considered as an SNP. This genetic variation can explain differences between two individuals' blood group, eye color, skin color etc. [5]

II. METHODS OF FINDING SNPS IN HUMAN GENOME

Finding SNPs in human genome can be categorized into two ways:

1. Genomic approaches: This approach looks at the bigger picture by processing a big set of data and finding the differences between genomes in different people. The final result is available to public over internet.
2. Functional approaches: In in this approach, unlike genomic approach, the scientists are targeting a specific case and study the result, for example the result of a certain drug usage in different people. [2]

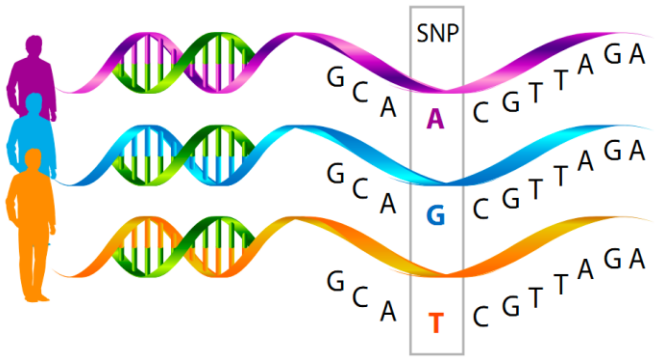

Figure 1. [4]

III. IMPLMENETATION STRATEGIES

3 different ways to find SNPs are described as part of an article called "*A tool for mapping Single Nucleotide Polymorphisms using Graphics Processing Units*":

- Two sequence to represent an SNP: In this approach, two sequences are mapped using a mapping tool aligned to the sequence database. Later, the result is analyzed for finding the SNP.
- Single sequence to present an SNP: Unlike the previous approach, the two sequences are combined together (N masked) and aligned to the sequence database. The result of this process needs to be corrected in order to find the exact position of SNP.
- G-SNPM mapping strategy: This is the

most complex strategy of all, in which there are 3 pipelines to update location of SNP. In this approach, SOAP3 is used in the first stage to map SNP against a reference sequence. In second phase, the unmapped or ambiguously mapped SNP will be mapped again using SHRiMP2. Lastly, the result is analyzed to identify the position of the SNP. [6]

IIV. HOW TO FIND SNP IN A GENOME ONLINE?

There are different ways to find SNP in a genome, however a fairly easy way is to look for SNP in DNA sequences is as follow:

1. Go to the BLAST home page and click "nucleotide blast" under Basic BLAST.
2. Paste the sequence in the query box.
3. Enter the name of the organism of interest in the "Organism" box. Click the BLAST button.
4. Click on the desired sequence from the results.
5. In the Links menu in the upper right, click on "GeneView in dbSNP". [1]

V. TOOLS

Short Oligonucleotide Analysis Package (SOAP 3) is a GPU-based software which can find all alignments with different number of mismatches. This tool is relatively faster than its older versions. However, the software cannot be used on all operating systems. It even requires a very specific hardware to be able to function. [3]

REFERENCES

[1] *U.S national library of medicine* [online], Available: https://ghr.nlm.nih.gov/primer/genomicresearch/snp

[2] *Making SNPs Make Sense* [online], Available: http://learn.genetics.utah.edu/content/pharma/snips/

[3] *Short Oligonucleotide Analysis Package* [online], Available: http://soap.genomics.org.cn/

[4] Image: https://biogeniq.ca/wp-content/uploads/2014/11/snp.png

[5] *Genetic variation* [online], Available: http://cisncancer.org/research/what_we_know/biology/genetic_variation.html

[6] Andrea Manconi; Alessandro Orro; Emanuele Manca; Giuliano Armano; Luciano Milanesi, "A tool for mapping Single Nucleotide Polymorphisms using Graphics Processing Units", *BMC Bioinformatics. 2014; 15(Suppl 1): S10.*
DOI: 10.1186/1471-2105-15-S1-S10