Research Proposal

Measurement and Analysis of the Read Coverage Distribution of ChIP-seq Experiments

Dinesh Adhithya Haridoss

Indian Institute of Science Education and Research Bhopal

Introduction:

ChIP-sequencing, is a method used to analyze protein interactions with DNA (Aparicio, Geisberg, et. al 2005). ChIP is a powerful method to selectively enrich DNA sequences bound by a particular protein in living cells. Such sequences are sequenced by fragmenting sequences into shorter sequences called reads. Aligning these reads to a reference genome leads to overlaps among various reads (Fig. 1).

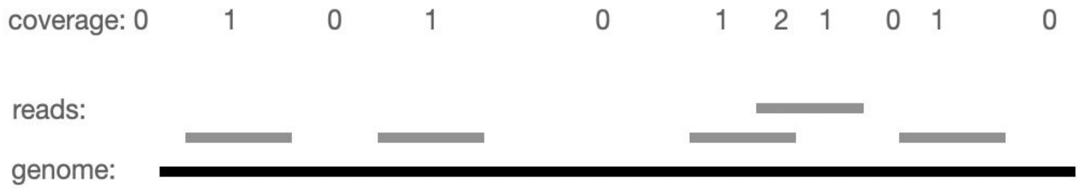

Fig. 1: Diagram describing coverage, reads and genome. Coverage would be no. of sequences overlapping a particular region.

We want to analyse the distribution of lengths of overlaps among various reads. Observed differences to a random model will allow us to get interesting insights on genome organization (see Massip and Arndt 2013 for a similar approach).

Literature Review:

A simple evolutionarily neutral model, which involves only point mutations and segmental duplications, and produces the same statistical features as observed for genomic data (Massip and Arndt, 2013).The model predicts analytically an exponent of power law of –3, which does not depend on the parameters nor the microscopic details of the model.

Research Design and Methods:

We want to develop a null model for the Read Coverage Distribution (RCD). This model would assume that all sequenced reads are uniformly distributed along the genome. My preliminary simulation of such a null model for small genome sizes revealed the following distributions:

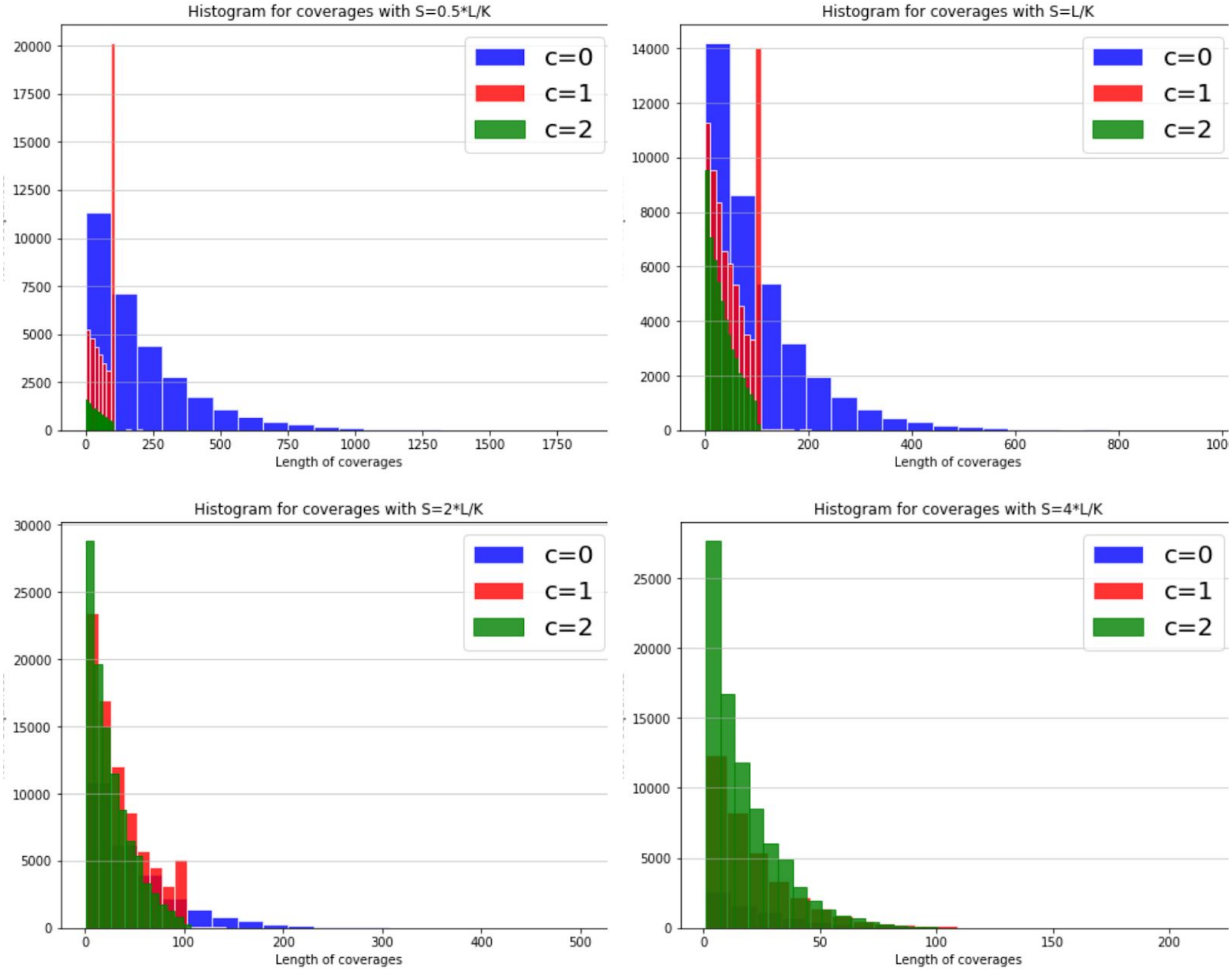

Fig. 2: Plot showing the coverage distribution for various numbers, denoted by S, of sequenced reads in the null model. The genome size L = 10^7 and the length of individual read K = 100 were fixed.

The simulation was performed using random generation of various read's starting positions and using dynamic programming techniques the lengths for overlaps values 0, 1 and 2 were found and plotted (see above Fig. 2).

During my internship I want to work on 2 research aims:

(1) What are the analytical functions describing the RCD for c = 0,1,2,…

Having analytical formulas is very important to judge whether our simulations are done correctly.

(2) I want to analyse observed ChIP-seq data and see whether the RCDs are different and what we can learn from the observed differences. ChIP-seq datasets from various types of experiments will be investigated.

All code will be written in Julia (Bezanson, Karpinski, Shah, Edelman et. al 2017), a new very efficient language to handle large dataset. To ensure reproducibility Jupyter Notebooks will be used to develop my analysis tools.

Implications and contribution to knowledge:

In our research project we want to measure the distribution for read coverage from ChIP-seq experiments. We have good reason to believe that this distribution from ChIP-seq experiments will not be the one expected for the above described random process. We want to analyze the distribution to be found and learn about genome organization and evolution.

Citations:

[1] Neutral evolution of duplicated DNA: an evolutionary stick-breaking process causes scale-invariant behavior.
Florian Massip and Peter F. Arndt, Physical Review Letter 2013

[2]Julia: A Fast Dynamic Language for Technical Computing.

Jeff Bezanson, Stefan Karpinski, Viral B. Shah, Alan Edelman 2017

[3] Chromatin Immunoprecipitation for Determining the Association of Proteins with Specific Genomic Sequences In Vivo.

Aparicio, O; Geisberg, J.V.; Sekinger, E.; Yang, A.; Moqtaderi, Z.; Struhl, K 2005