UNIVERSITY OF TARTU
Institute of Computer Science
Computer Science Curriculum

Yurii Toma

Predicting the impact of non-coding genetic variants on transcription factor binding with machine learning

Master's Thesis (30 ECTS)

Supervisor: Kaur Alasoo, PhD
Supervisor: Dmytro Fishman, MSc

Tartu 2018

Acknowledgments

First of all, I would like to thank both of my thesis supervisors Kaur Alasoo and Dmytro Fishman not only for their great understanding of the bioinformatics and machine learning and sharing that knowledge with me, but also for showing to me value and motivation of these fields.

Also, I would like to thank the Estonian Foreign Ministry's Development Cooperation and Humanitarian Aid funds, which supported my studies at the University of Tartu and made my experience of studying abroad incredible.

Another portion of thanks, I would like to express to Professor Jaak Vilo, dean's office workers of Institute of Computer Science, University of Tartu international student office workers and to all workers of the University of Tartu for the incredible experience that I have had during my studies, for their enthusiasm in trying to make study programmes very practical, interesting and useful, and for the opportunities to do internships at top tech companies during the studies.

Finally, I want to say many thanks my parents, friends and all engineers and interns who I have met during my internship at Google Waterloo for their ability to be always supportive and giving me good arguments and motivation to complete my masters thesis.

Predicting the impact of non-coding genetic variants on transcription factor binding with machine learning

Abstract:

Understanding how the human organism works is one of the most important problems in the science. A lot of research effort went into analysis of deoxyribonucleic acid (DNA) since the first human genome was sequenced. Despite these efforts, there are still a large number of poorly understood processes happening in the human organism. One of them is understanding the functional consequences of non-coding genetic variants in the DNA sequence of a human. These variants, if functional, are likely to influence the binding of transcription factors - regulatory proteins that control the expression of other genes by binding to regulatory elements across the genome. A diverse set of methods have been developed to predict the effect of genetic variants on transcription factor binding. However, all of these methods have been limited by the lack of high quality testing data to evaluate their accuracy. Here I combine and re-analyse three large genetic studies to identify a high quality set of likely causal genetic variants that regulate the binding of CTCF and PU.1 transcription factors. I then use these variants to evaluate the accuracy of three state-of-the art prediction algorithms. My results indicate that while the impact of some genetic variants with large effect can be readily predicted, most variants with smaller effects are missed by current prediction algorithms. My approach is generalisable to other transcription factors and can be used to benchmark the accuracy of novel prediction algorithms developed in the future.

Keywords:

Bioinformatics, DNA, transcription factor bindings, machine learning.

CERCS: P170 Computer Science, Numerical Analysis, Systems, Control

Masinõppe abil mittekodeerivate geneetiliste variantide mõju hindamine transkriptsioonifaktorite seondumisele

Lühikokkuvõte:

Inimorganismi toimimispõhimõtetest arusaamine on üks tänapäeva teaduse suurimaid väljakutseid. Esimese inimgenoomi sekveneerimise järgselt on palju resursse kulutatud DNA sekventsi ja selle varieeruvuse uurimiseks. Nendest jõupingutustest hoolimata ei saa me paljudest inimorganismis toimuvatest olulistes protsessidest endiselt väga hästi aru. Üks selliseid protsesse on mittekodeerivate geneetiliste variantide mõju hindamine inimese genoomis. Sellised geneetilised variandid, kui neil üldse peaks mingi mõju olema, mõjutavad suure tõenäosusega transkriptsioonifaktorite seondumist. Suur hulk erinevaid meetodeid on välja töötatud ennustamaks geneetiliste variantide mõju transkriptsioonifaktorite seondumisele. Täpsete testandmestike puudumise tõttu on aga nende meetodite täpsuse hindamine olnud raskendatud ja seetõttu on enamasti lähtutud kaudsetest mõõdikutest. Oma töös panen ma kokku kolm suurt geneetilist andmestikku, et kindlaks teha suur hulk geneetilisi variante, mis suure tõnäosusega mõjutavad põhjuslikult kahe transkriptsioonifaktori (CTCF ja PU.1) seondumist DNAle. Järgnevalt kasutan ma neid geneetilisi variante hindamaks kolme kaasaegse ennustusalgoritmi täpsust. Minu tulemused näitavad, et kuigi mõne suure mõjuga geneetilise variandi efekti hindamine on võimalik, jääb enamiku väiksema mõjuga variantide mõju kindlaks tegemata. See lähenemine on üldistatav teistele transkriptsioonifaktoritele ja seda saab kasutada uudsete ennustusalgoritmide täpsuse paremaks hindamiseks tulevikus.

Võtmesõnad:

bioinformaatika, DNA, transkriptsioonifaktorid, masinõpe.

CERCS: P170 Arvutiteadus, arvutusmeetodid, süsteemid, juhtimine

Contents

1 Introduction

Even though molecular biology is relatively young branch of science, it is already responsible for the huge jump in better understanding of the algorithms the human organism works (*1*). DNA (deoxyribonucleic acid) is one of the main parts of the molecular biology (*2*). It is a sequence of nucleotides which is present in every human cell.

DNA is particularly interesting for researching, because it has a potential of describing nature of different diseases: what is the reason these diseases happen, what do they change in genome, how to treat them or avoid them. The whole DNA sequence is classified into two types of regions: coding and non coding. Non coding regions of DNA are those regions of DNA that are not directly involved in protein producing in a given cell, while coding regions are involved. Different living organisms have different proportion of the non-coding DNA regions. Bacteria has only 2% of non-coding DNA, while the human DNA contain around 98% of non-coding nucleotides (*3*). Is it necessary to have all that non-coding nucleotides and why their proportion is so high? More than 90% of the genetic variants associated with human complex traits and diseases are in the non-coding regions of the genome(*4*), suggesting that non-coding regions are important for determining individual's genetic risk for those disease Another huge question is how and what regulates the producing of different proteins from the same DNA. All cells within one organism have same DNA sequence. However, the actual appearance, properties and functions of a cell can vary a lot, so the question about how and why it happens and is it possible to have control over it is one of the fundamental problems. Large body of evidence now supports the important role of non-coding regions in regulating protein abundance and thus cell function.

Mutations (*5*) in DNA sequence can happen for multiple reasons: lifestyle, environment, simple error while replicating DNA in the process of cell division. It may be only 1 nucleotide that will change from the long sequence of three billion

nucleotides for the human, but the actual consequences might be very good or bad for the individual, so it is also good to know what are the effects of these single nucleotide mutations on the protein production.

The work is structured as follows:

- in Chapter 2 general overview of the problem and basics of molecular biology are introduced,
- Chapter 3 - describes the process of getting fine tuned data for making the future evaluations much easier and reliable,
- Chapter 4 - the experiments with the data obtained in Chapter 3 and the existing approaches and evaluations of them,
- Chapter 5 - discussion of results and possible future work.

2 Background

In this chapter the general background of the problem is given. Shortly described how DNA is built, what is its role in the human organism and why is it important to investigate about how it works. Also in this chapter I shortly describe how the DNA is sequenced today.

2.1 DNA overview

DNA (*6*) is a sequence of nucleotides present in all living cells. Interesting part is that within one organism DNA is same in all cells, but cells itself have different properties.

DNA chains are formed with the four base nucleotides which are Adenine (A), Cytosine (C), Guanine (G) and Thymine (T). The nucleotides are chemical structures with nitrogenous base which is responsible for one of four different representations, they also contain sugar and phosphate group. Two adjacent nucleotides are then connected to each other through chemical sugar-phosphate connection, the sugar in each nucleotide consist of five carbon atoms, which form spatial structure and in chemistry it is common to standardize the way this sugar is described. It turns out that phosphate group is located in 5′ position of a nucleotide and OH group which also takes part in generating sequence of DNA is located in 3′ position (Figure 1). Thus when talking about DNA it is common to talk about 3′ end and 5′ end of DNA. When the DNA is created it is growing from 5′ end to the 3′ end, because it requires much less energy to grow that way comparing to the opposite direction.

When the sequence of nucleotides are connected a DNA strand is formed. Inside the DNA strand each of the nucleotides has free nitrogenous base, so what happens then - complementary strand of nucleotides is created (Figure 2). It is called complementary, because where in the original strand were located for example A

then in complementary strand it is known that T will be in that place. Same mappings are then applied to all nucleotides in both directions of mapping: A <-> T, C <-> G. The two complementary nucleotides in strands in the same position are then connected one to another via double hydrogen bonds.

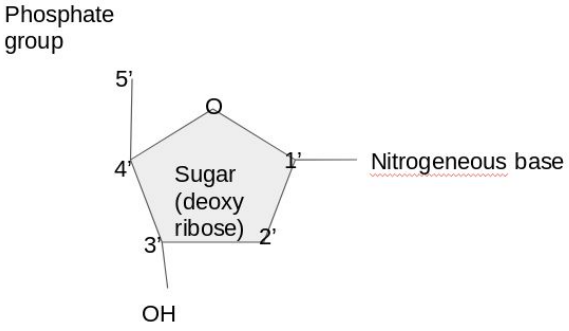

Figure 1: nucleotide structure, 3′ and 5′ ends, and the parts of nucleotide: 5 Carbon sugar base, Nitrogenous base connected at 1′, OH group at 3′, Phosphate group at 5′.

Thus, it makes the whole DNA to have a structure of double stranded helix (Figure 2) in space. Two strands of the DNA are directionally located opposite one to another: if one of the ends of the first strand is 5' then it automatically makes 3' end of the complementary strand.

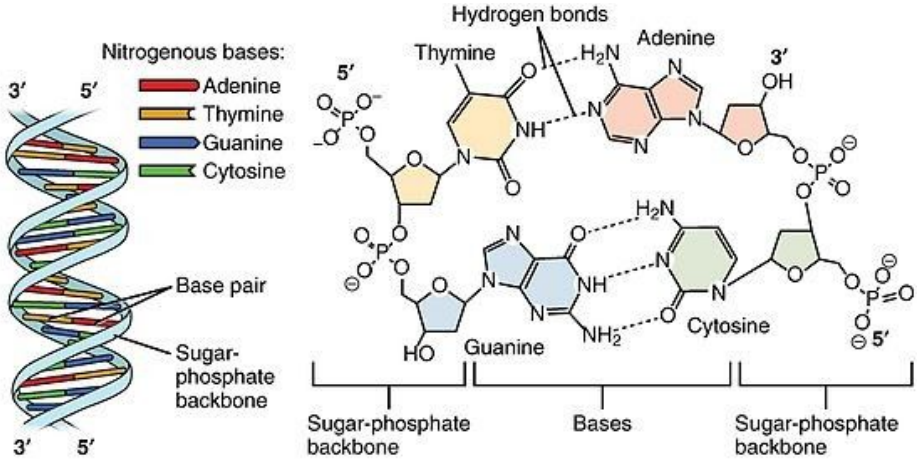

Figure 2: Left - spatial structure of DNA (double stranded helix) and nucleotides connected via hydrogen bond into two strands; Right - complementary bases hydrogen connections, forming strands from 5' to 3' via sugar-phosphate backbones (*7*).

DNA in human organism is very long. It has length of approximately three billion base pairs, so the storing algorithm of this huge amount of data in each cell is very complicated problem. In human cell DNA is split over 23 pairs of chromosomes (22 identical pairs and one defining sex: XX similar pair for females and partly similar XY for males). DNA in each chromosome is then tightly packed on histones forming lots of coils, coils are then combined into groups of supercoils which are then forming each of the chromosomes, making most of the DNA information hidden into this package.

2.2 Central dogma of molecular biology

The spatial structure of DNA described in the previous part was discovered by James Watson, Francis Crick and Rosalind Franklin in 1954, this discovery made possible to answer a lot of other questions. One of such was answered in 1958: Francis Crick proposed the way the proteins are created in any living organism. He proposed that once protein is created, it can not go back to nucleic acid, transfer from nucleic acid to nucleic acid or from nucleic acid to protein is possible, but transfer from protein to nucleic acid is not. This is known now as a central dogma of molecular biology in a bit different formulation (*8*). So, small portion of one strand of DNA is transcripted into messenger ribonucleic acid (mRNA), having the other strand of DNA and knowing the complementation rules for the nucleotides it is possible for the organism to replicate the other strand of DNA and have again two stranded DNA. mRNA is one strand that is created with four different nucleotides three of them are same as for DNA: A, C, G and Thymine is replaced with Uracil (U). mRNA is later translated via now known mapping of three consecutive base pairs (called codons) into one part of the protein, called amino acid. There codons also code the beginning of the translation and ending. Each of 64 possible codons is translated into exactly

one of 20 amino acid which are then connected into polypeptide chain to get the protein. There are lots of different types of proteins, and their role in the living organism is huge. This is one of the main reasons why the central dogma of molecular biology is called central. Central dogma can be summarised in a diagram in Figure 3:

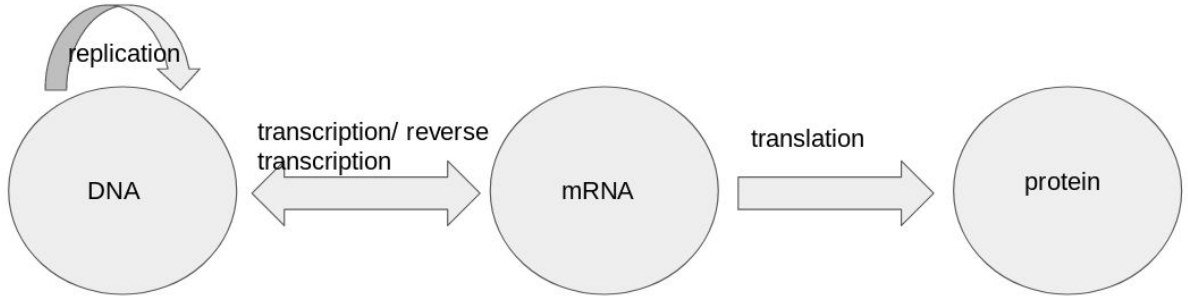

Figure 3: diagram to summarize the central dogma of molecular biology.

2.3 What is interesting about DNA?

Before the structure of the DNA was discovered, the exploring of DNA was not the main focus of research, because it seemed very boring: its spatial structure is known, it is created only of four well known nucleotides, even though it is very long ~3 billion base pairs for human, most of it is not translated into proteins (non-coding). The amount of known base blocks of proteins is 20 and that was making proteins much more interesting topic for researchers.

Everything changed after the discovery of the DNA structure and proposal of central dogmathe . Which allowed to connect different proteins with the DNA. It make the DNA main focus of the research for now. Because the resulting proteins are created from some subsequences of DNA, these regions of DNA should be freely accessible and open from their packagings. There are exist special proteins called transcription factors (Figure 4) to open the region of interest so then the mRNA can be transcribed and translated to the protein. There are also different transcription

factors in each cell that define which regions of the DNA are open in that cell type and which genes are going to be transcribed to mRNA. Transcription factors recognize specific subsequences of the DNA (called motifs), and when they find they target sequence, they bind to the DNA, open up the chromatin by displacing the nucleosomes and recruit other factors (such as RNA polymerase) to initiate the transcription of the mRNA (*9*). Thus, by controlling which genes are expressed in any given cell type, transcription factors define the identity of those cell types.

Any living organism is non ideal as well as the conditions lifestyles are. This influences different mutations in the DNA sequence. This thesis mostly focuses on the so called Single Nucleotide Polymorphisms (SNPs), that are basically change in one of the nucleotides in the DNA chain. It might seem not relevant to talk about one change in 3 billion base pairs, but actually the effect of this change can be very bad.

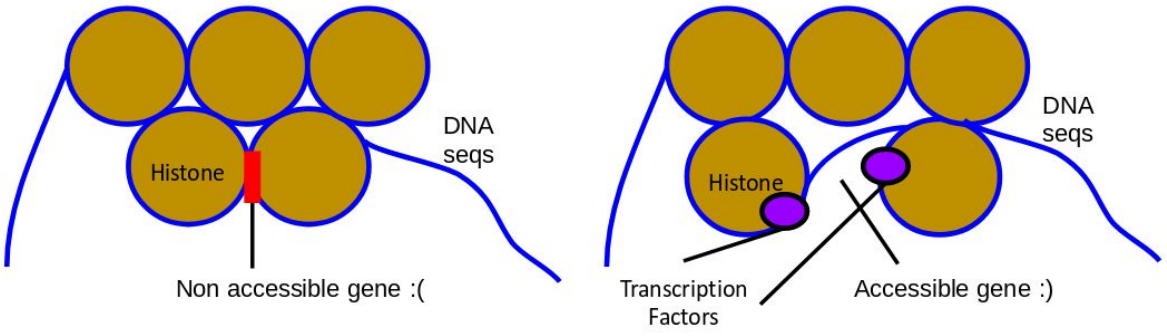

Figure 4: Left - DNA packed onto histones, red region - non accessible gene, because it is not accessible protein is not produces; Right - transcription factors opened the gene of interest and now protein can be produces.

One of the known issues is that such changes can activate cancer or increase the risk of many complex diseases. One mechanism by which non-coding SNPs could have these devastating effects is that they could disrupt normal transcription factor binding (*10*). So, studying the effect of SNPs is very important and can give a lot of information about how evolutionally DNAs were developing and how to get the best out of it.

2.4 Measuring transcription factor binding

One of the possible ways to understand what is the function of non-coding regions of the DNA is to analyze the data of locations and sequences of the DNA to which transcription factors are likely to bind. One of the methods to obtain such data is chromatin immunoprecipitation followed by sequencing (ChiP-seq) (*11*). For example, researcher may be interested in parts of the DNA sequence that are likely to be involved in the regulation of transcription to mRNA which is then translated to some specific protein. The method works in the following way: protein of interest is cross linked to the DNA, that DNA sequence is fragmented into huge number of small pieces. Antibody that recognizes the protein of interest is then taken and it will detect the subsequences of DNA that are involved in the process of creating the protein of interest. The results are called DNA reads and are usually stored in the FASTA (*12*) format. These results latter can be used in the analysis on the computer. For the analyses in this thesis I used published ChiP-seq datasets to get reads associated with CTCF and PU.1 transcription factors. Which are then processed into convenient formats and analyzed with the existing tools and approaches.

ChiP-seq is an experimental way for measuring the transcription factor binding. What if we have a sequence of nucleotides and are interested in *predicting* transcription factor binding at some particular position. In this case we can use position weight matrix (*9*, *13*) (PWM). PWM is a fixed length matrix with 4 rows, each row is corresponding to one of the 4 nucleotides it represents the motif (nucleotide pattern) for the transcription factor. Length is different for different transcription factors. So if a length of matrix is l, then the shape is 4xl and all elements in it are float numbers. It can then be aligned to some position in DNA sequence and then for the overlapped region the transcription factor binding score for the given motif can be computed by combining the actual nucleotides of that region and the numbers in

corresponding row and column of PWM by multiplication. Graphical representations of the PWM is called sequence logo. We can see sequence logos for CTCF and PU.1 transcription factors in Figure 16 and in Figure 17.

2.5 Measuring chromatin accessibility

Chromatin accessibility is the way of measuring how much protein can be produced from different parts of genome, thus it gives the information about how opened that region is. The main function of transcription factors is to control how many of the protein is produced. It is done by transcription factor binding to some sequence (binding site) in DNA and then opening the region of gene specific to the transcription factor, so that the protein can be produced. The more transcription factors have binded - the more protein can be produced. Then we can measure the effect of mutations on transcription factor bindings via measuring chromatin accessibility at different regions of DNA. The higher the chromatin accessibility - the higher the likeliness of transcription factor binding at that places and vice versa.

Assay for transposase-accessible chromatin using sequencing (ATAC-seq) (*11*) is a technique used nowadays to measure the chromatin accessibility. The general idea of ATAC-seq is that it inserts special enzymes into the accessible regions of the DNA and later inserted parts are sequenced. ATAC-seq does not differentiate between different transcription factors, it just gives the chromatin accessibility score and where it is located in the DNA. So the ATAC-seq data can be used as a different way to get transcription factors binding scores for sanity checks.

3 Data Preparation

In this chapter I will describe the data processing pipeline from the raw ChiP-seq signal to the final set of binding sites (peak sequences) associated with CTCF and PU.1 transcription factors. I will also describe how I identified the set of sequences that are not associated with these transcription factors. I will describe what and why was done and how the fine-grained data for CTCF and PU.1 transcription factors was obtained.

3.1 Data description and initial preprocessing

Data that was used for the two transcription factors (CTCF and PU.1) came from two previously published ChIP-seq experiments (*14*) (*15*). The raw CTCF data stored as 81 paired-end FASTQ files was downloaded from the ArrayExpress database (accession number E-ERAD-141). The raw PU.1 dataset consisting of 47 paired-end FASTQ files was also downloaded from ArrayExpress (accession number E-MTAB-3657). Sequences for some individuals for CTCF data were splitted into couple paired-end fasta, so the real number of individuals for CTCF is less than 81, it is 49. For the PU.1 there were same as number of files as the number of individuals. The actual meaning of each record is that there was binding of CTCF, PU.1 in this individual to that subsequence of nucleotides. But, because the number of sequences can be very high and each sequence can be very long, instead of reading out the full DNA sequences where the transcription factor was bound, the sequencing machine only reads the prefix and suffix of the original DNA sequences. The read lengths for the CTCF and PU.1 datasets were 50 bp and 38 bp, respectively for beginning and end of the read sequence.

Thus, the first step of data preprocessing was to obtain actual nucleotide sequences for each record in paired-end FASTA files. For this the “Burrows-Wheeler

Alignment Tool" (bwa) (*16*) was used. It is very fast way of getting the actual sequences from the fasta reads. First, the reference genome sequence is indexed to make bwa tool work on it. After this alignment procedure can begin, during the data processing maximum exact match algorithm of bwa was used. It tries to align FASTA record sequence to each position in reference genome and select then the positions which had the maximum number of matches and also when the sequences are paired-end the tool tries to estimate the lengths distributions of the sequences in the input FASTA files and get the final sequence with respect to the maximum exact match and length distribution, which is good, because we expect that the sequence lengths should not have very large differences. We ignored the secondary alignments provided by bwa mem, using only the most probable alignment for each read. The output of bwa mem is stream in SAM format which is then fed as input to samtools (*17*), the tool that can manipulate SAM(Sequence Alignment Map)/BAM files, which are essentially storing each sequence as a tuple of 11 elements that describe this sequence in terms of the coordinates of the reference genome.

One final step was to generate BED files out of BAM files. The BED file stores each nucleotide sequence as 3 main integer parameters: chromosome, start, end positions. Which is good for storing information about lots of sequences, without having to store them as a strings, so this saves a lot of memory and it is not very hard to retrieve the actual sequence back from some hashed/indexed genome. Very short and long aligned DNA fragments (from the start of the first read to the end of the second read) were removed during preprocessing of the data (<50 nucleotides or >5000). The short sequences were omitted due to fact that prefix and suffix read together are longer than 50 base pairs. Long sequences were omitted, because we are not expecting the binding site for the transcription factors to be very long.

3.2 Merging multiple CTCF FASTA files of same individuals into single FASTA file per individual

For the PU.1 dataset, each pair of the FASTQ files corresponded to a single individual. Furthermore, the authors clearly indicated which which files corresponded to which individuals, thus making it straightforward to link ChIP-seq signal with genetic variation data from the same individuals.

The situation was more complicated with the CTCF data, because, although, the study included 49 individuals, the published dataset on ArrayExpress had 81 pairs of FASTQ files. This suggested that data from some of the individuals were split between two pairs of FASTQ files. Unfortunately there was no easy way to determine this, because the authors of the study did not publish the mapping between the individuals and the the data files in the repository. Thus, I had to resort to genetic information present in the FASTQ files to link them back to the individuals from whom they originated from. This would allow me to later merge the files to have better data for analysis.

In order to do this the mbv mode of QTLtools (*18*) package was used. For each individual we had information about their genetic variants stored in a variant call format (VCF) file. Then the procedure done by mbv is to calculate the number of heterozygous (different copies of the allele on the two chromosomes) and homozygous (same alleles on the two chromosomes) genotypes in each individual and then in each file with reads. Then the genotypes that are not frequent enough are filtered out. It is assumed that if we have lots of the reads for the individual, they should cover sufficient number genetic variants to uniquely identify an individual. Next, for each pair of individual *versus* aligned read file (BAM) it calculates ratios between the number of heterozygous genotypes found in reads and in individual, same ratio is also computed for homozygous genotypes. Then results can be plotted as read file vs all individuals with previous proportions as points in 2D space.

It is expected that for the correct individual - file relation this proportions will be close to one, and for all other individuals they should be smaller and on the plot they should form two groups: correct individual as an outlier and then separate group of close points - all other individuals. Although this was not always the case for our data, so we excluded some sequence files from downstream analysis, because they were

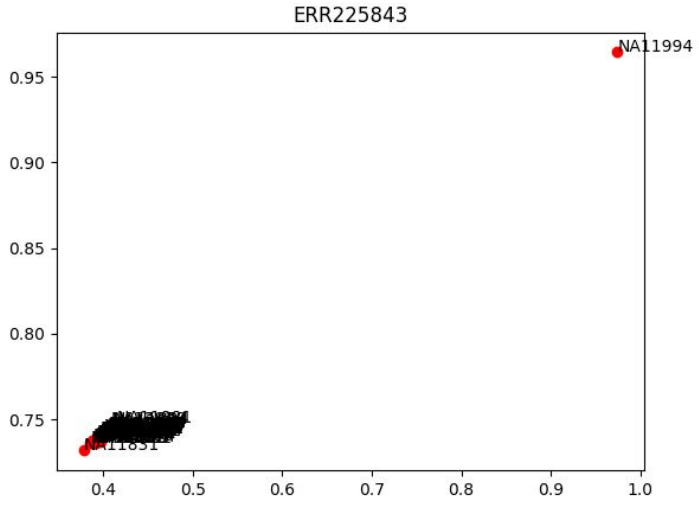

Figure 5: The sequence file ERR225843 shows high correspondence to the individual NA11994.

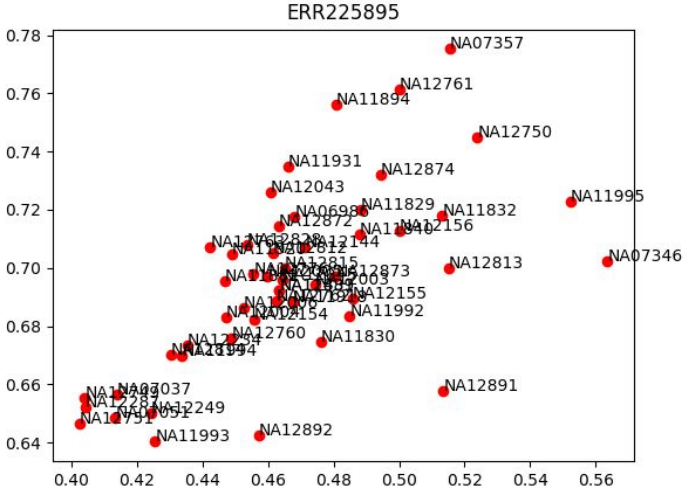

Figure 5: This sequence file does not have any good correspondence, such read files were omitted.

highly correlated with more than one individual, or there was no any visible good choices. Some of the examples can be seen in Figure 5 and Figure 6.

Next step was to create mapping between each sequence file and individual and merge sequence files grouped by individuals. So, one option was to manually and visually investigate all 81 plots and then create mapping but this is not efficient. Instead I developed the following algorithm:

1. For one fixed FASTQ read file and for each possible individual we compute the area of rectangles given by origin and point with proportions coordinates.
2. Then all the areas per fixed FASTQ read file are sorted and the individual with the largest area is treated as the one to which potentially given sequence file corresponds.
3. We then do steps 1, 2 for all FASTQ files and save maximum area values as well as also difference between largest and second largest area which we treat as distance between 2 individuals for the given FASTQ file. Ideally we want this distance to be as close to 1 as possible. So after this step there are 81 values for the areas of best matching individual per each FASTQ file, and 81 values of distances from this "best" individuals to the second "best" ones.
4. Based on the area values and distance values from previous step, we can now try to detect outliers. We do this by rejecting the FASTQ files with 10% smallest percent on both of the datasets: areas or distances. Reason - if area is small - then both proportions are low, if the distance is low, then it is not clear which individual correspond to the FASTQ file. 10% was used after looking at all plots (Figure 7) and seeing that there are not more than 8 outliers.
5. All of the FASTQ sequence files that are left are then mapped to individuals with the maximum area.

After this procedure all sequence files that are mapped to the same individual are merged together.

Examples of good and bad mappings can be seen in the Figure 5, Figure 6. Mappings for the all files can be seen in the Figure 7.

Only 5 files of read sequences were removed, as can be seen from the plots in the Figure 7 most of the files had good correspondence to exactly one individual.

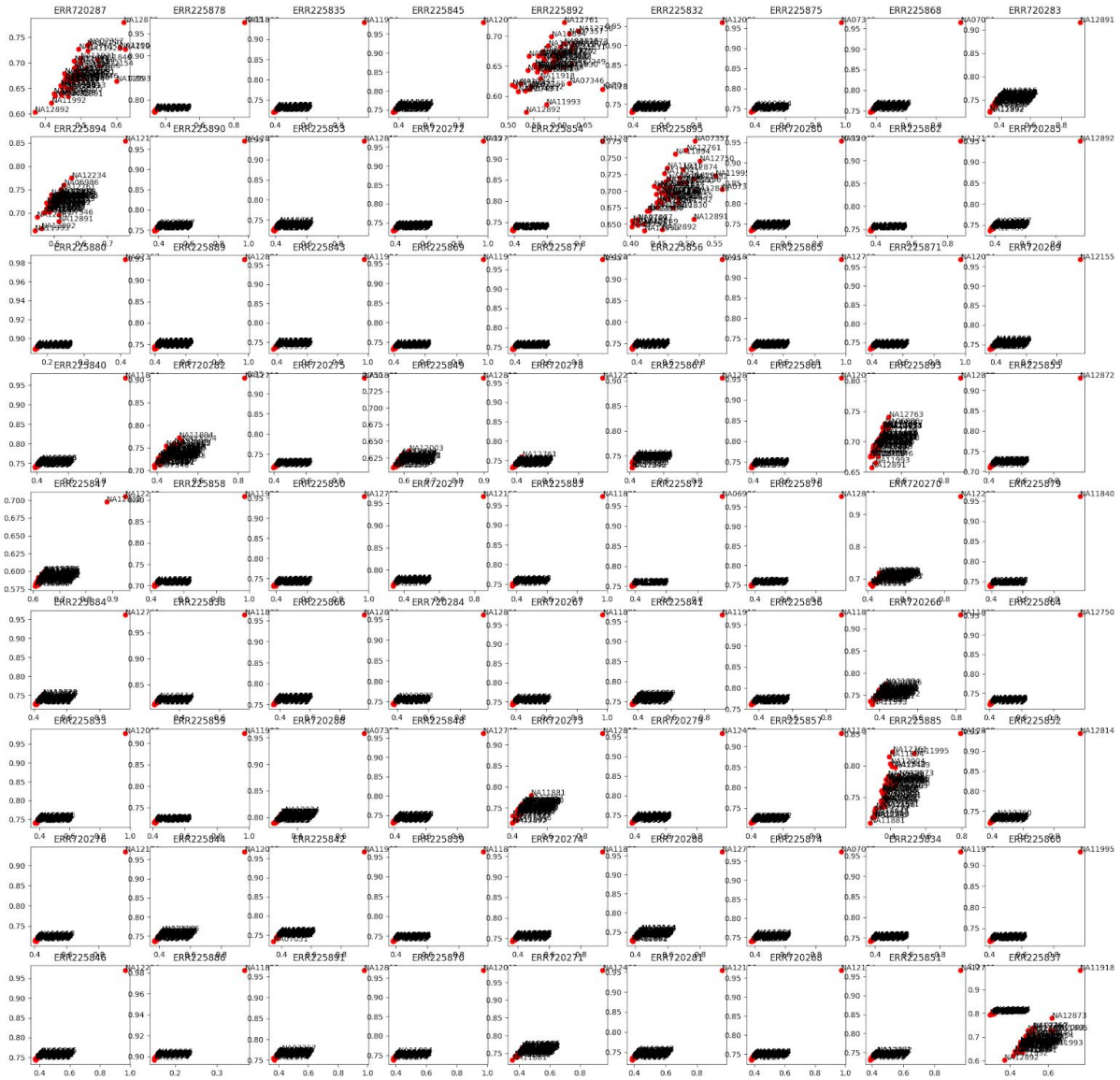

Figure 7: sequence file vs individual plots for all CTCF data files.

3.3 Identifying binding sites of transcription factors

In this part the bam files generated in previous steps are taken and manipulations with them are performed in order to get the CTCF and PU.1 peaks as well as their locations. Peaks are positions in the genome to which in given data the CTCF, PU.1 transcription factors are very likely to bind. Likeliness is mostly defined by frequency of occurrence of that binding site. For the purpose of peak calling the tool called MACS2 (*18*, *19*) was used. It was applied on per individual basis and the main idea is that it maps each of the reads to the reference genome and then calculates frequency of each nucleotide, after this it tries to predict which is a good position within the regions that is a position of a peak. The tool tries to model a shift of ChIP-seq for each factor given the input sequences with a Poisson distribution, and if it is equal to d, then it does sliding window over the genome and frequencies with size 2*d and takes point in each window with highest frequencies as potential peaks.

While modelling the peak distributions, MACS also tries to account for biases via making parameters dynamic. P-values for the potential peak are calculated to remove the potential false positives. Also False Discovery Rates (FDR) estimated for potential peak positions and then cutoffs either based on original p values or on adjusted by FDR q values it uses Benjamini-Hochberg procedure.

So, the result of applying the MACS tool should give us good estimations binding sites of CTCF and PU.1 in our data.

Then all the individuals are taken and all the peaks are merged to get the peaks data for the binding factor overall instead of for each individual separately. To do this R package GenomicRanges (*20*) used. All overlapping peak segments were merged into single peak range, to avoid duplicates and to treat the whole union of regions as a binding site.

3.4 Estimating the number of reads assigned to peaks

To assess the quality of the data and to detect outlier samples with worse than average quality, It is a good idea to see what proportion of the original reads correspond to some peaks. This is a type of signal-to-noise ratio, because reads originating from outside of peaks are likely to caused by background noise. Of course it is not expected to see lots of failures in this step. To do this part of analysis the tool called featureCounts (*21*) to count reads overlapping peaks was used and then MultiQC (*22*) was used to summarize that data with visualisations.

The results for the CTCF and for the PU.1 can be seen in the Figure 8, and Figure 9 respectively. Each line represents one sequence input file.

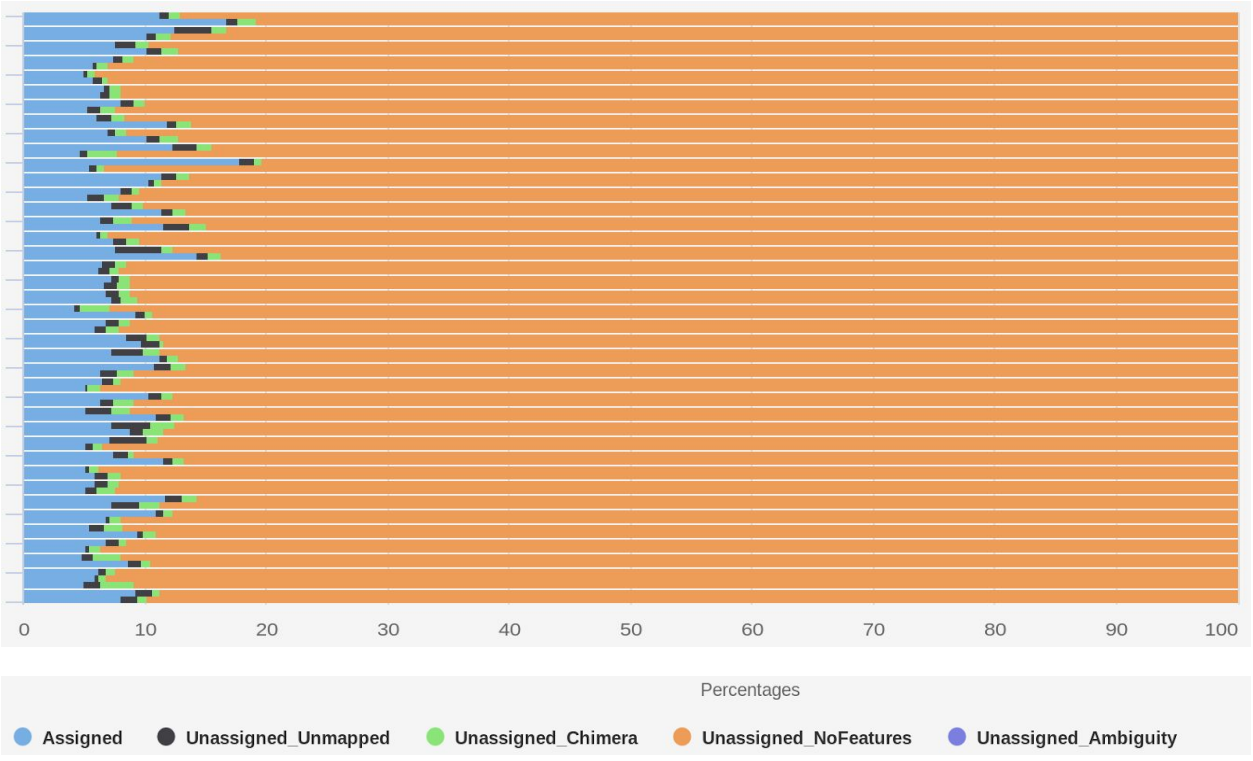

Figure 8: CTCF results of mapping reads to peaks, on average 10% of reads are assigned to peaks.

For CTCF the average number of assigned reads per individual was ~10%. And there are not huge amount of non regular unassigned reads. For PU.1 the situation is very similar: average number of assigned reads per individual was ~10%. And all other reads are just unassigned to any of the peaks. Furthermore, neither of the dataset contained obvious outlier samples that were very different from others.

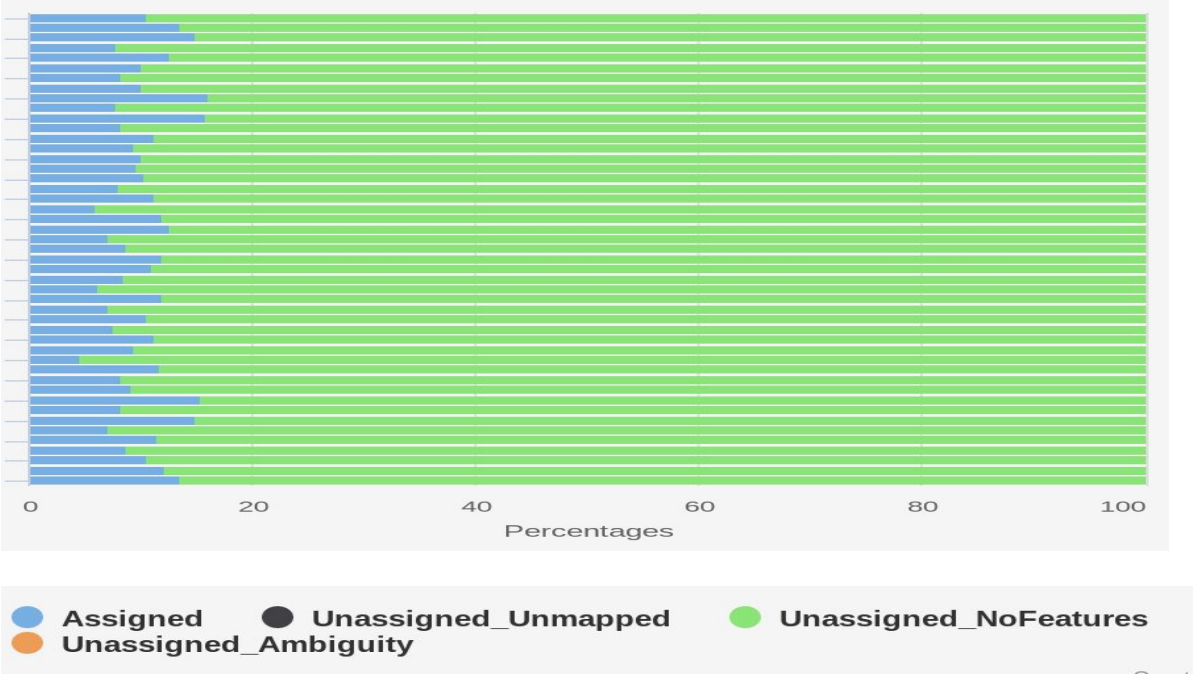

Figure 9: the mapping reads to peaks results for PU.1 per sequence file; ~10% on average reads were mapped to peaks.

3.5 Generating the QTL variants for CTCF and PU.1

At this point the nucleotide sequences for peak regions was obtained for both CTCF and PU.1 transcription factors. For each individual for the both datasets there was also available files with information about genotype of individual. Since almost all of these individuals were part of 1000 Genomes project, the genotype data was downloaded from the 1000 Genomes website (*23*). Given that data it is possible to

obtain the list of genetic variants (single nucleotide polymorphisms or SNPs) that are associated with the binding of the transcription factors under investigation.

For this purpose the cis tool from the QTLtools (*24*) package was used. In general QTL is a quantitative trait locus - is a statistical approach to understand the connection between the phenotype (observed properties) and genotype (the underlying sequence of nucleotides). For phenotype to genotype QTL mapping SNPs are usually used as to show places in genome that are responsible for the observed in phenotype traits. So, it should help in identifying variants in peak sequences that are responsible for changing the binding of the transcription factors. The tool helps to identify the SNPs and their effect with respect to the given data.

There are two types of QTLs: cis and trans, the difference is that if the given SNP affects the quantitative trait in on the same chromosome, then it is called cis QTL, otherwise it is trans QTL. Because in the experiments we are interested only in SNPs that have influence on transcription factor binding within the same peak we use the cis QTL to discover SNPs.

The tool also requires a matrix of covariates to be provided. The reason we are using the covariates is that we want to remove all variations obtained due to technical reasons, and be able to detect true genetic associations better (*25*). One can get it via doing principal component analysis (PCA) within the same QTL tools on the peak counts matrix. And then using its rows of the PCA matrix as a covariates.

For CTCF I received ~11 millions of variants, among which ~740k were significant (p-value smaller then 5%), and for PU.1 ~19 million, among which ~1.4 million where significant. In both these sets sequences that had at least one significant association with SNP where detected for further experiments, in CTCF there were 36k such peaks, and in PU.1 - 45k. The selected peaks were the ones that have overlap at least with one ATAC peak, and the FDR adjusted p-values for selected peaks was significant less than 0.1.

3.6 Normalisation of counts

Before using the PCA of peak counts as a covariates, it is a good idea to normalise the counts, so that the actual number of individuals and files will not have influence on further analysis. Especially taking into account that the data from another research for the peak scores was used as part of pipeline to get scores for the influence of SNPs for CTCF and PU.1.

For the normalisation R package cqn (*26*) based on conditional quantile normalisation (cqn) was used. It is one of the common ways of normalising count data from sequencing experiments and make it suitable for linear models data. All it requires is a fraction of GC nucleotides for each peak and length which can be easily evaluated, because we have peak sequences. And the result - is normalised same size as input matrix.

3.7 Chromatin accessibility data (ATAC-seq)

A recent study demonstrated that a clever modelling of ATAC-seq data from a large number of individuals together with their genotype data can be used to identify likely causal genetic variants responsible for differences in chromatin accessibility between individuals (*27*). This is the data from different research, the way this data was obtained is different from the way scores were obtained in our case. This data contains the information about peaks for the lots of different transcription factors, it also contains the different SNPs the nucleotide on allele0 and alternative nucleotide for all of the peaks. For each variant, their analysis also provides a posterior probability that the variant can cause the change in chromatin accessibility.. Thus, likely causal variants responsible for the change in chromatin accessibility can be defined. In the research they were defined as ones that have posterior probability of

changing the transcription factor binding higher than 50%. The total number of variants in ATAC-seq data was 170k, after defining the causal ones (posterior probability > 50%) the amount of variants was reduced to 3008. All of the 3008 variants are located within the accessible region.

One thing that the coordinates of the peak regions in this data as well as positions of the SNPs were defined with respect to the hg37 human genome assembly, so the mapping of all the positions was performed with the help of tool called CrossMap (*28*) to hg38 coordinates as well as mapping of the CTCF, PU.1 data was to the hg37 coordinates. This was done both ways mainly for the convenience. If this step was skipped, then the overlappings of different regions between ATAC data and CTCF, PU.1 peaks would have no sense sometimes, although sometimes the change in the coordinates was not very large.

3.8 Correlation between QTL scores and ATAC scores for CTCF, PU.1 peaks variants

A key limitation of the ATAC-seq analysis by (*27*) is that although they were able to identify likely causal variants responsible for changes in chromatin accessibility, they were not able to identify the transcription factors whose binding was affected. To overcome this limitation, I decided to focus on the 3008 ATAC-seq peaks for which they had identified a likely causal variant and overlapped this set of ATAC-seq peaks with PU.1. and CTCF peaks identified in this chapter. Next, I checked if the effects of these likely causal variants on chromatin accessibility CTCF and PU.1 binding were correlated with each other.

Among 3008 ATAC causal variants for chromatin accessibility peaks, first the CTCF peak variants were obtained and their scores compared to the scores obtained during the data generation described in this chapter. Among 3008 ATAC variants 279 were overlapping with some of the CTCF peaks obtained in this chapter

earlier, so these variants are likely to causally regulate CTCF binding. The correlation between the scores obtained in this chapter for CTCF peaks (genetic effect of the predicted causal variant on CTCF binding) and ATAC data scores (genetic effect of the predicted causal variant) was ~0.75, and the plot can be seen in the Figure 10. This confirms that genetic variants that regulate chromatin accessibility also regulate transcription factor binding at the same sites in the same direction, even though these two datasets have been generated using different approaches, in different labs using only partially overlapping cell lines.

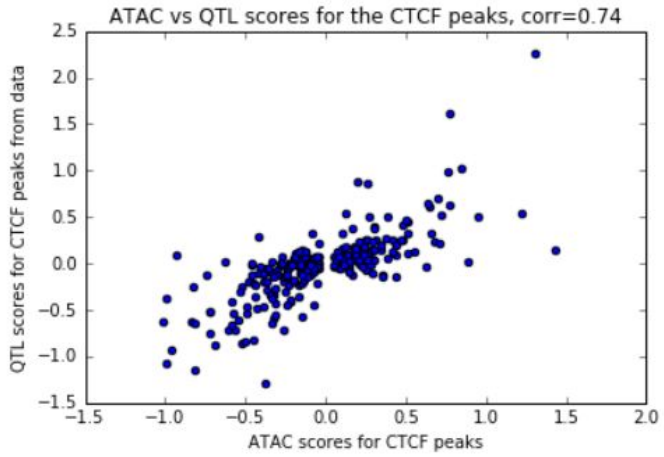

Figure 10: correlation between scores from ATAC data and scores obtained from data in the chapter 3 for CTCF peaks variants.

Then the histograms of the p-values from ATAC data for this 279 peak variants for CTCF can be seen in the Figure 11, they look as expected - higher number of peaks with lower p-values.

Similar processing was also then applied to the PU.1 data. Among 3008 ATAC causal variants for peaks of different transcription factors, the PU.1 peak variants were obtained and their scores compared to the scores obtained during the data generation described in this chapter. Among 3008 ATAC variants 1139 were overlapping with some of the PU.1 peaks obtained in this chapter earlier, so these variants are likely to causally regulate binding. The correlation between the scores obtained in this chapter for CTCF peaks and ATAC data scores was ~0.81, and the

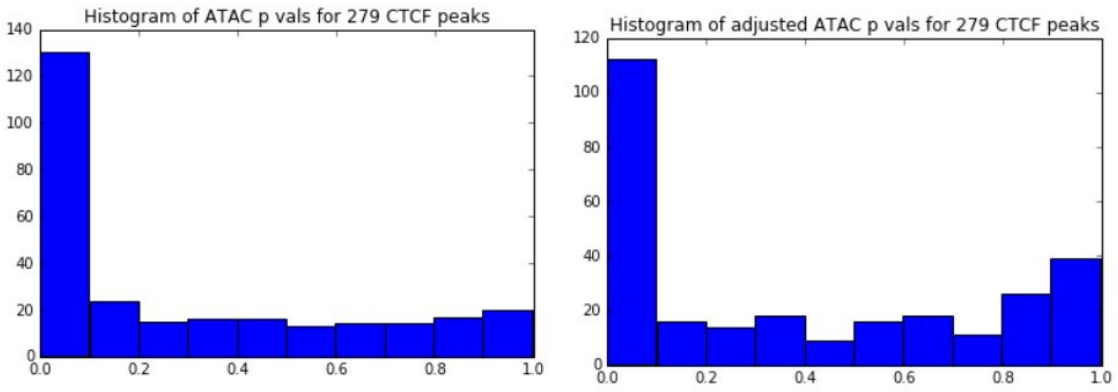

Figure 11: left - histogram of original p-vals for the 279 CTCF peak variants from ATAC data, right adjusted by FDR pvals for the same 279 p-vals.

plot can be seen in the Figure 12. Also the histograms of the p-values from ATAC data for this 1139 peak variants for PU.1 can be seen in the Figure 13, they look as expected - higher number of peaks with lower p-values.

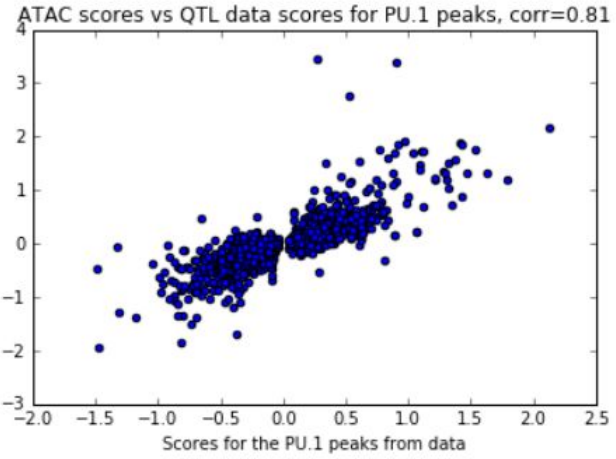

Figure 12: correlation between scores from ATAC data and scores obtained from data in the chapter 3 for PU.1 peaks variants.

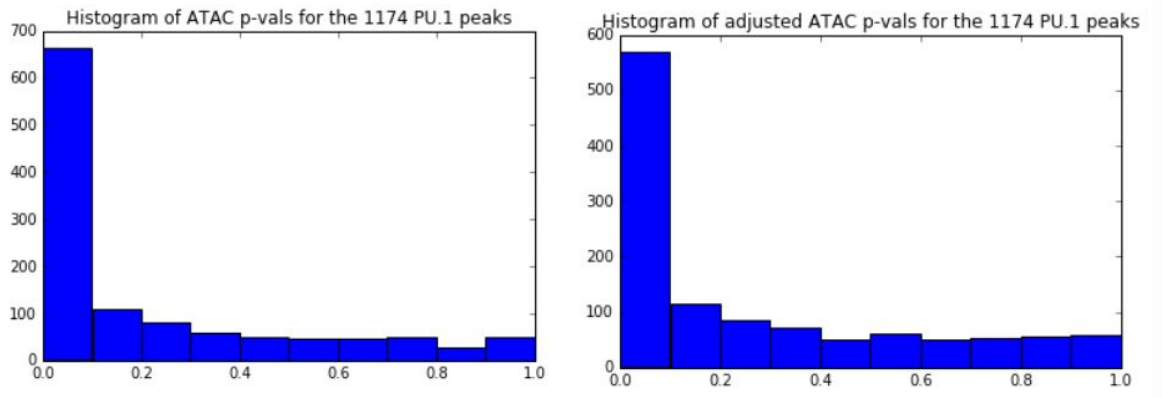

Figure 13: left - histogram of original p-vals for the 1139 PU.1 peak variants from ATAC data, right adjusted by FDR pvals for the same 1139 variants.

It can be observed that the scores that denote the change in transcription factor bindings for CTCF, PU.1 peak variants obtained from data in this chapter are very similar to the scores from totally different dataset ATAC, so the data generated should be of good quality.

4 Comparative study of existing approaches to estimate effect of SNPs on CTCF and PU.1 binding

In Chapter 3, I identified a subset of 3008 likely causal genetic variants that were responsible for changing chromatin accessibility. I also demonstrated that in regions overlapping CTCF and PU.1 binding sites, these variants also had highly correlated effects on transcription factor binding (Figure 10 and Figure 12). Thus, if a genetic variant is predicted to casually regulate chromatin accessibility and it also overlaps a transcription factor binding site from a ChiP-seq experiment where it shows concordant direction of effect on transcription factor binding, then this provides strong evidence that the same genetic variant also causally regulates the binding of the overlapping transcription factor.

In this chapter I explore how the effect of these likely causal genetic variants can be predicted from the DNA sequence context of the variant alone. I used the following three existing prediction methods: motifbreakR (*29*) - one of the most classical approaches based on position weight matrices (PWMs), gkmsvm (*30*) - more powerful, based on the classical machine learning support vector machine (SVM) classifier and DeFine (*31*) - potentially the most powerful among all three, because its based on deep convolutional neural networks (CNNs), which often outperform other methods in modern classification tasks.

4.1 motifbreakR

MotifbreakR (*29*) is an R package which can be used for the estimation of the consequences of SNPs on probability of transcription factor binding to DNA.

Especially interesting to see how well it will predict the change in likeliness of binding when the mutation occurs.

The way the motifbreakR works is mostly classical approach based on the PWMs. As an input it takes the file in the format similar to the BED format used before. This file has information about all the sequences of interest: position of start and position of end, chromosome location of the sequence, and the most important part - the description of a SNP - position within a sequence of it as well as original and alternative nucleotides.

Next the transcription factor of interest is defined. MotifbreakR uses motifDB (*32*) to retrieve the PWM for the transcription factor of interest in our case - CTCF or PU.1.For each row in the input BED file which represents one variation in one peak for the given transcription factor motifbreakR moves with a sliding window PWM over all positions that overlap the position of a SNP and calculates the score difference for the original sequence and the sequence with the SNP. It then produces the score difference which has the highest amplitude.

The motifbreakR was applied to the 279 CTCF peaks from the data obtained in Chapter 3 which has overlaps with the ATAC peaks data. It was also applied to the rest ~2.8k peaks to then see how good it can differentiate between CTCF and non CTCF peaks. The scores correlation between the predicted from QTL 279 CTCF peaks via motifbreakR and ATAC scores is 0.33 and the maximum possible that we can obtain from QTL data for correlation with ATAC from part 3.8 is ~0.75. This was very good result for very simple approach. Combined with the fact that correlation of scores for non CTCF peaks was only 0.009, it can be observed in Figure 14.

The precision vs recall curve (PRC) was then built based on the scores of motifbreakR for the CTCF and non CTCF peaks. Scores show the change in probability of transcription factor binding if the mutation happens in the DNA. For building PRC the ordered array of scores is taken, and assumption is if the classifier is good then this ordered array should have threshold to do easy classification between two classes. We assume that the classifier should produce higher absolute

valued scores for the CTCF peaks comparing to non CTCF peaks. Thus, all the scores here and in all other PRCs were taken by absolute value. The precision recall curve is used, because the dataset is not balanced and thus other metrics can not guarantee good intuition of the classifier that could potentially be built on top of motifbreakR. The sequences which overlap CTCF peaks are treated as positive sequences and the non overlapping - as negative class. The resulting PRC can be observed in Figure 15. The higher the area under the curve (AUC) - the more likely the classifier to put to random positive example score higher than to the random negative sample. It can be observed that the quality of such classifier is not particularly high, with AUC ~0.28.

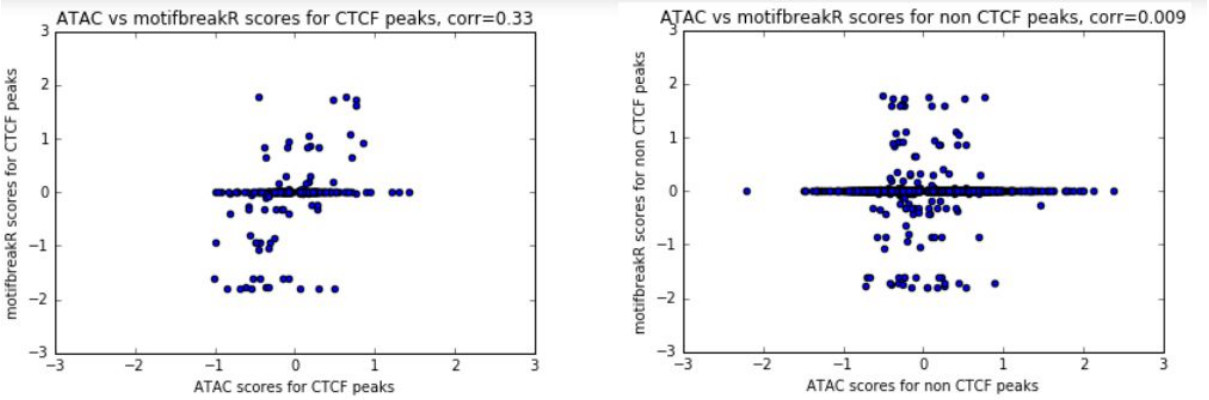

Figure 14: left ATAC scores correlation vs motifbreakR scores for CTCF peaks, right - for non CTCF peaks.

The motifbreakR was also applied to 1139 PU.1 peaks from the data obtained in Chapter 3 which has overlaps with the ATAC peaks data. It was also applied to the rest ~1.8k peaks to then see how good it can differentiate between PU.1 and non PU.1 peaks. The scores correlation between the predicted from QTL 1139 PU.1 peaks via motifbreakR and ATAC scores is 0.18 and the maximum possible that we can obtain from QTL data for correlation with ATAC from part 3.8 is ~0.81. This result

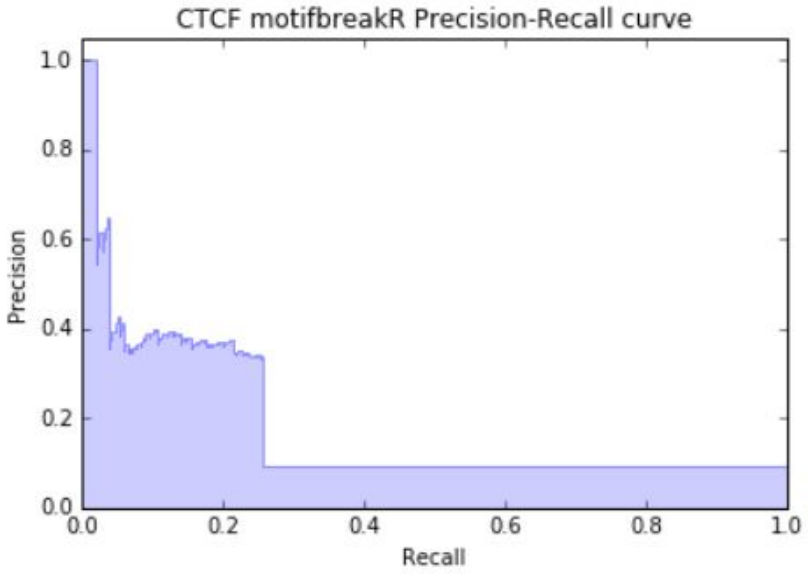

Figure 15: motifbreakR PRC for CTCF vs non CTCF peak classification, AUC is 0.28.

is worse then the respecting for the CTCF, the explanation for this fact is that the CTCF has longer and easier to catch nucleotide sequence that attracts it to bind, comparing to the same in PU.1, the actual sequence logos for CTCF and PU.1 are in Figure 16 and Figure 17. Combined with the fact that correlation of scores for non PU.1 peaks was 0.08, it can be observed in Figure 18.

Figure 16: sequence logo for CTCF transcription factor (*33*).

Figure 17: sequence logo for PU.1 transcription factor (*34*), shorter comparing to CTCF one, and most of the information is concentrated in the centre.

The PRC for the classification PU.1 peaks vs non PU.1 peaks is also non satisfiable - Figure 19, the AUC is 0.48.

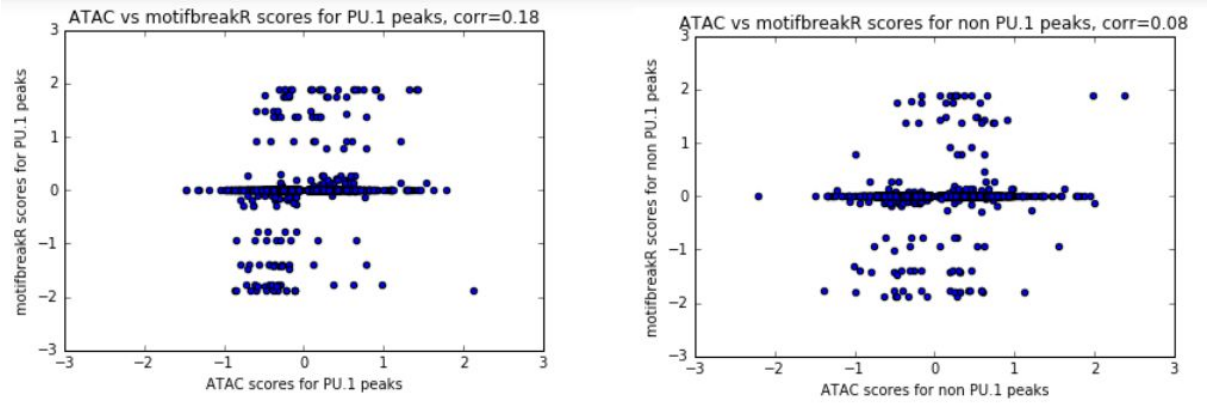

Figure 18: left ATAC scores correlation vs motifbreakR scores for PU.1 peaks, right - for non PU.1 peaks.

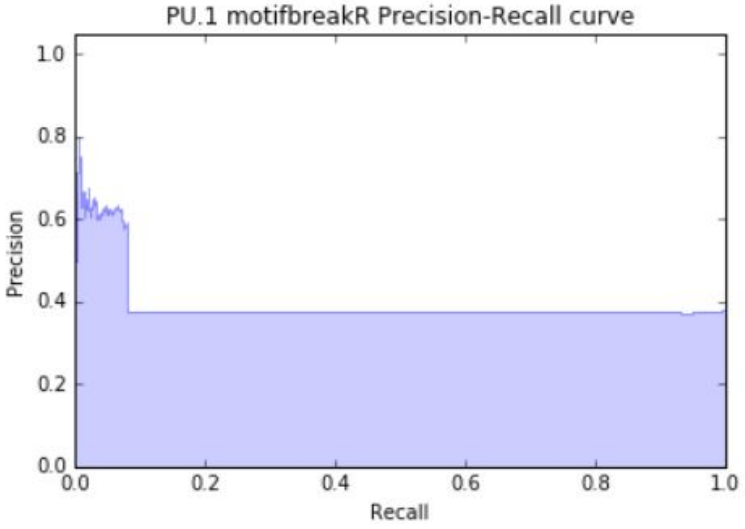

Figure 19: motifbreakR PRC for PU.1 vs non PU.1 peak classification, AUC is 0.48.

Overall the results for both PU.1 and CTCF transcription factors obtained by applying motifbreakR are good with respect to the fact that it is one of the simplest approaches. It shows that there is some potential for the development of more powerful classifiers that will be able to differentiate between genetic variants that

influence CTCF or PU.1 binding from those that have no effect on those factors but still regulate chromatin accessibility.

4.2 gkmsvm

Very often when using DNA sequences and machine learning the k-mers (nucleotide sequences of length k, k - variable) are used as a features to train a model to classify between binding and non-binding sites for example for some specific transcription factor. With the grows of k the k-mers become less frequent, and more noisy, which makes it then harder to train a good model based on such feature set.

Gkmsvm (*30*) is one of the approaches to overcome that problem and train a good classifier. Instead of k-mer features it uses gapped k-mer (gkm) features and as the authors claim they were able to achieve much better results comparing to the classifiers trained on k-mer features.

The classifier that is used - support vector machine (SVM) is a classical and powerful machine learning approach for binary classification. The SVM classifier aims to build a hyperplane in the space between positive and negative class examples such that the distance of the closest points to it is maximised. After the gkmsvm classifier is trained on the set of positive and negative sequences, given the new example it gives the distance from the example to the svm hyperplane. If the goal is to estimate the effect of the SNP on the binding of the transcription factor, then the 2 scores for the original sequence and the sequence with SNP can be calculated and the difference between them is then can be treated as an effect of SNP on the binding of transcription factor (*30*).

The gkmsvm classifier for CTCF vs non CTCF scores prediction was trained using approximately 20k sequences that are CTCF peaks sequences and 20k of non CTCF peaks sequences. They were collected by overlapping the CTCF peaks from

data obtained in chapter 3 with all peaks from ATAC data excluding the causal peaks.

Out of 280k ATAC sequences 37k were overlapped with the CTCF peaks, and used as a positive set, all others ~240k went to the negative dataset. The causal 3008 peaks of ATAC data were excluded from the training. Then out of the positive and negative sets the random subsample of size 20k for each class was selected. The gkmsvm is training for few days and then results were produced. The correlation for the causal CTCF peaks variants predicted scores by the model and ATAC data is ~0.36 and for the non CTCF peaks it was -0.18, Figure 20 has plots of the model scores vs ATAC data scores. The PRC was also built for this classifier, the results are not particularly good as can be observed in Figure 21, the AUC is 0.18. Comparing to the motifbreakR the correlation results are slightly better, but comparing to time spent on data preparation for the input to gkmsvm and time spent on training it is probably better to just use motifbreakR. Figure 22 has some insights on correlation in scores between gkmsvm predictions and motifbreakR the scores of the two classifiers are correlated with 0.65. Thus, the classifiers are learning similar things.

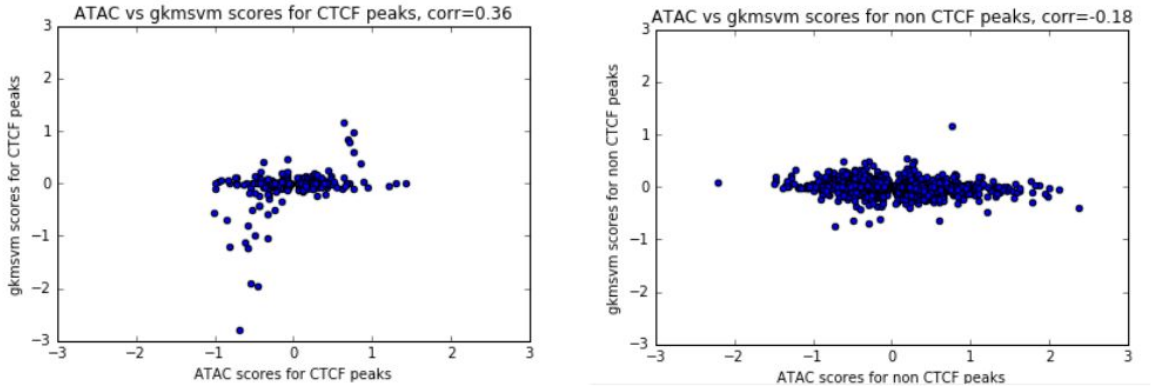

Figure 20: left ATAC scores correlation vs gkmsvm scores for CTCF peaks, right - for non CTCF peaks.

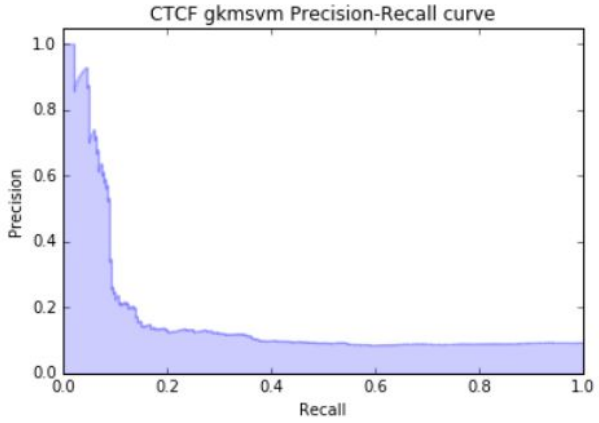

Figure 21: gkmsvm PRC for CTCF vs non CTCF peak classification, AUC is 0.18.

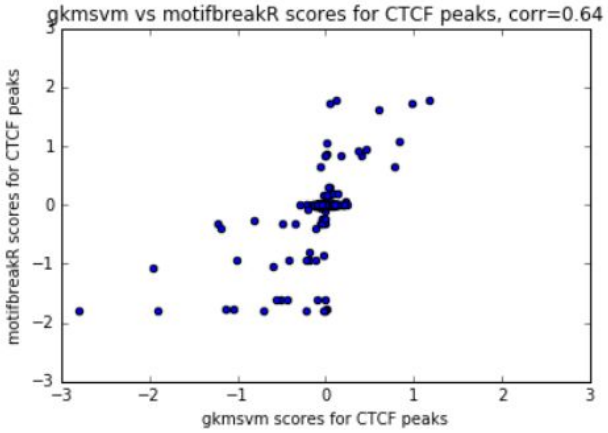

Figure 22: gkmsvm score vs motifbreakR scores, shows that the two classifiers seem to learn similar things.

Then similar processing for the PU.1 was performed. The positive dataset for PU.1 consisted of 45k sequences and ~230k sequences for the so called negative class the subset of 3008 causal ATAC peak variants was again excluded from the training set. Again the 20k subsets were subsampled from these to train the gkmsvm model. The reason for subsampling is amount of time spent for the model generation as well as amount of resources used, also regular gkmsvm does not work with huge number of sequences, so the modification for large scale gkmsvm was used (*35*). The gkmsvm results are again comparable to the PU.1 motifbreakr results, but this time gkmsvm is

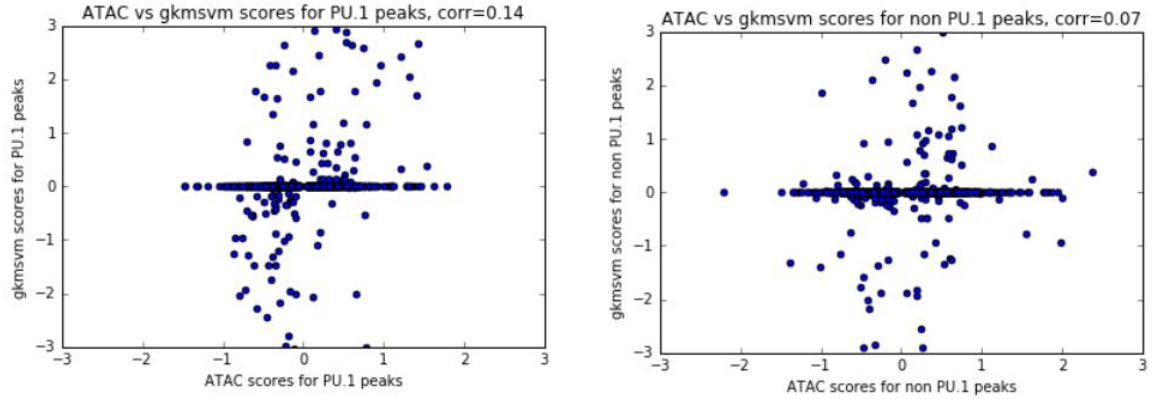

Figure 23: left ATAC scores correlation vs gkmsvm scores for PU.1 peaks, right - for non PU.1 peaks.

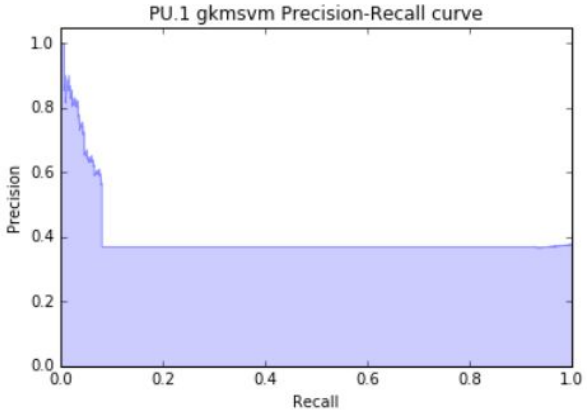

Figure 24: gkmsvm PRC for PU.1 vs non PU.1 peak classification, AUC is 0.45.

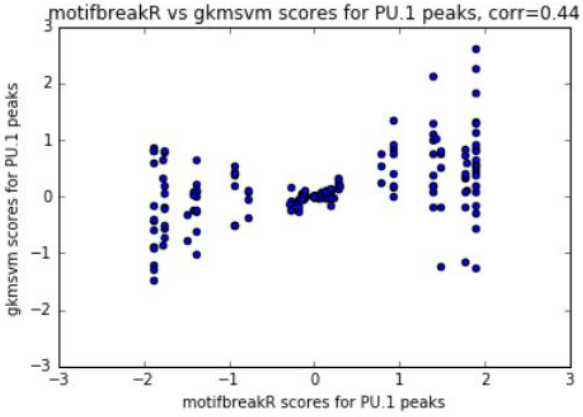

Figure 25: gkmsvm score vs motifbreakR scores for PU.1 classification, shows that 2 classifiers are likely to learn similar things.

slightly worse comparing to the motifbreakR. The results can be observed in Figure 23, Figure 24 and Figure 25.

4.3 DeFine

DeFine (*31*) is a modern and powerful approach to estimating the effect of mutations in DNA sequence on binding of huge amount of different transcription factors, including also CTCF and PU.1 that are factors of interest for the given thesis. It has a lot of pretrained models including the ones for CTCF and PU.1, these models were used. It is essentially deep convolutional neural network that nowadays is one popular and very powerful ways to do classification and regression problems. It takes raw sequence and the sequence where the mutation occurred as an input and then predicts the effect of that change on multiple different transcription factors. It has online interface available for everyone where one can provide sequence of interest begin and end positions as well as SNP parameters and then in couple minutes the result will be produced.

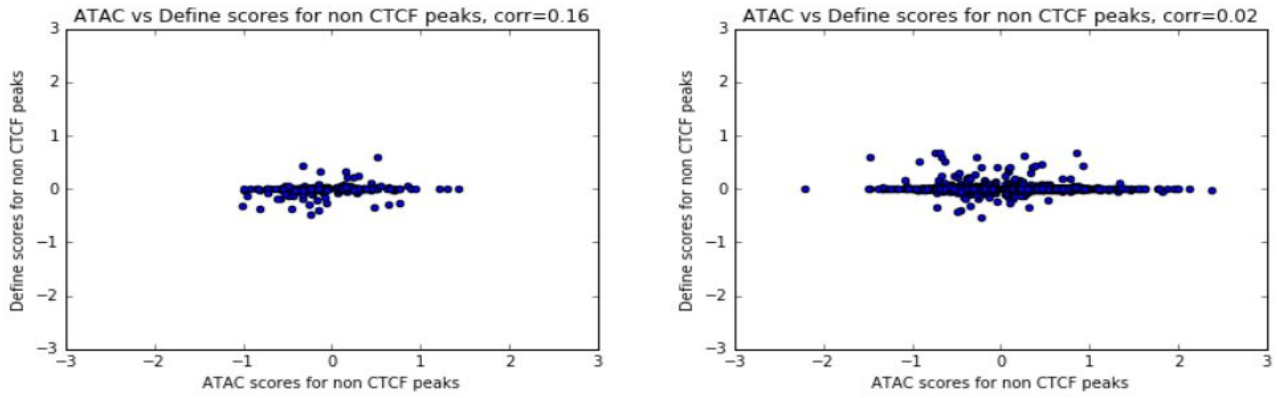

Figure 26: left ATAC scores correlation vs DeFine scores for CTCF peaks, right - for non CTCF peaks.

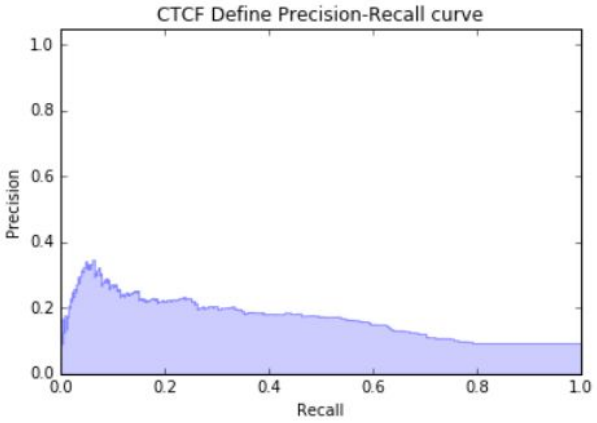

Figure 27: DeFine PRC for CTCF vs non CTCF peak classification, AUC 0.16.

For the CTCF 279 peaks were compared to 2.8k non CTCF peaks with the help of DeFine. Unexpectedly the results were worse compared to motifbreakR and to gkmsvm. Probably the reason is that DeFine is very generic prediction model and we are interested in one specific factor - CTCF and model that would be able to predict for CTCF. Figure 26 shows the correlations of DeFine scores with ATAC scores and Figure 27 - PRC for the DeFine, the AUC is 0.16.

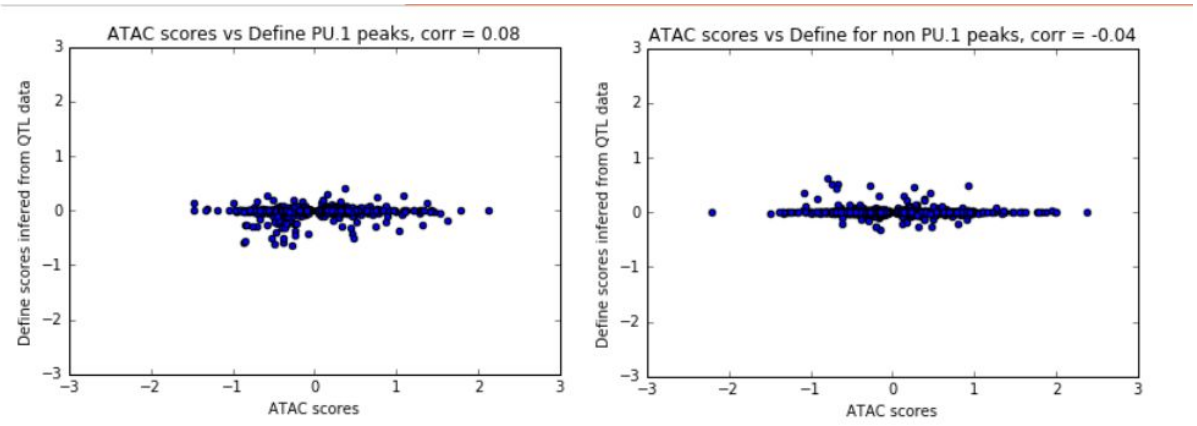

Figure 28: left ATAC scores correlation vs DeFine scores for PU.1 peaks, right - for non PU.1 peaks.

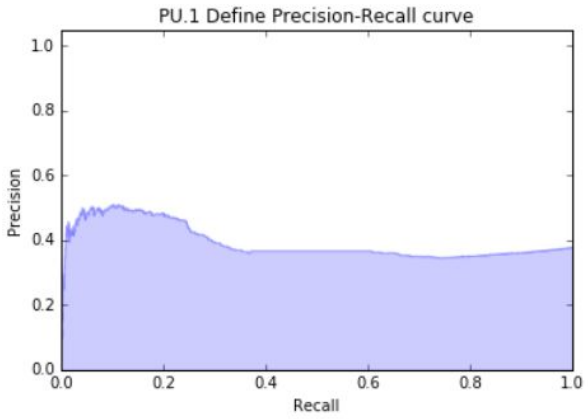

Figure 29: DeFine PRC for PU.1 vs non PU.1 peak classification, AUC is 0.38.

For the PU.1 1139 peaks compared to 1.8k non PU.1 peaks with the help of DeFine. Again same happened: the results were worse comparing to motifbreakR and to gkmsvm. Figure 28 shows the correlations of DeFine scores with ATAC scores and Figure 29 - PRC for the DeFine, the AUC is 0.38.

Interesting part is that correlations for peaks in CTCF and PU.1 were in both cases approximately 2 times worse comparing to gkmsvm and motifbreakR.

5 Discussions

The main amount of work is done in the intersection of different fields: computer science, bioinformatics, molecular biology and machine learning. While working on the thesis, lots of the bioinformatics tools were used. It looks like the learning curve for the using bioinformatics tools to combine computer science and molecular biology is unreasonably high, there is no (at least I haven't found) intuitive explanations to lot of definitions, there are no toy datasets (at least I haven't found) that the beginner of the field could play with develop better understanding of the topic. Most of the information is available only in the articles which are not always the best way for easy explanations and quick learning of the topic. Also all tools vary a lot and not necessarily have good documentation with explanations. So the direction of standardizing tools and developing playgrounds (e.g. OpenAI (*36*) gym for convenient familiarizing yourself with the reinforcement learning field) for the beginners seems like direction for the improvement in bioinformatics. This will potentially allow more people into this research field.

The first outcome of this work is that the high quality test data for estimating the effect of mutations in DNA on CTCF and PU.1 transcription factors is created and it can be reused in further studies. Most of the datasets that are present nowadays are for the effect on the lots of different transcription factors simultaneously, or if they are for some specific transcription factor, the development set of individuals was much lower than in this case. The approach is generalisable to other transcription factors if similar data becomes available.

The other outcome is that initial analysis with the obtained data was done with the help of the existing solutions and their results are compared one vs another.

The results for all 3 of the compared approaches were not satisfiable. The interesting thing is that these result showed one more time that sometimes the simple solution is the best one (motifbreakR), both: timewise and result wise. Although the more complicated one gkmsvm - showed comparable to motifbreakR

results (sometimes even better) it took much more time to perform the analysis with gkmsvm so it makes it less favourable. DeFine performed surprisingly poorly for specific task of estimation of effect of mutations on CTCF or PU.1 individually.

Even though, the results of these analysis are not specifically good, the conclusion is that existing solutions are either not very powerful to capture what is interesting for capturing or they are very generic. So, it seems that there is a big potential for further work that can involve developing, training and fine tuning of the custom deep learning models separately for predicting effect of SNPs in DNA on CTCF/PU.1 bindings. Furthermore, transcription factors often cooperate with each other in selecting their binding sites, so that a genetic variant that directly disrupts the binding of one transcription factor might also indirectly influence its partners (*10*). Thus, multi-task deep learning models that predict the binding of multiple transcription factors at the same time might be useful in those cases.

References

1. M. F. Perutz, Fundamental research in molecular biology: relevance to medicine*. *Nature*. **262**, 449–453 (1976).

2. P. B. Gahan, Molecular biology of the cell (4th edn) B. Alberts, A. Johnson, J. Lewis, K. Roberts and P. Walter (eds), Garland Science, 1463 pp., ISBN 0-8153-4072-9 (paperback) (2002). *Cell Biochem. Funct.* **23**, 150–150 (2005).

3. E. S. Lander *et al.*, Initial sequencing and analysis of the human genome. *Nature*. **409**, 860–921 (2001).

4. Discovery of DNA Structure and Function: Watson and Crick | Learn Science at Scitable, (available at https://www.nature.com/scitable/nated/article?action=showContentInPopup&contentPK=397).

5. Genetic Mutation | Learn Science at Scitable, (available at https://www.nature.com/scitable/nated/article?action=showContentInPopup&contentPK=1127).

6. Discovery of DNA Structure and Function: Watson and Crick | Learn Science at Scitable, (available at https://www.nature.com/scitable/nated/article?action=showContentInPopup&contentPK=397).

7. Nucleotide - Wikipedia, (available at https://en.wikipedia.org/wiki/Nucleotide).

8. F. Crick, Central Dogma of Molecular Biology. *Nature*. **227**, 561 (1970).

9. W. W. Wasserman, A. Sandelin, Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.* **5**, 276 (2004).

10. E. al Deplancke B, The Genetics of Transcription Factor DNA Binding Variation. - PubMed - NCBI, (available at https://www.ncbi.nlm.nih.gov/pubmed/27471964).

11. C. A. Meyer, X. Shirley Liu, Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nat. Rev. Genet.* **15**, 709 (2014).

12. BLAST TOPICS, (available at https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=BlastHelp).

13. S. Sinha, On counting position weight matrix matches in a sequence, with application to discriminative motif finding. *Bioinformatics*. **22**, e454–e463 (2006).

14. Z. Ding *et al.*, Quantitative genetics of CTCF binding reveal local sequence effects and different modes of X-chromosome association. *PLoS Genet.* **10**, e1004798–e1004798 (2014).

15. E. al Waszak SM, Population Variation and Genetic Control of Modular Chromatin Architecture in Humans. - PubMed - NCBI, (available at https://www.ncbi.nlm.nih.gov/pubmed/26300124).

16. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. **25**, 1754–1760 (2009).

17. H. Li *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. **25**, 2078–2079 (2009).

18. E. al Fort A, MBV: a method to solve sample mislabeling and detect technical bias in large combined genotype and sequencing assay datasets. - PubMed - NCBI, (available at https://www.ncbi.nlm.nih.gov/pubmed/28186259).

19. Y. Zhang *et al.*, Model-based Analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).

20. M. Lawrence *et al.*, Software for Computing and Annotating Genomic Ranges. *PLoS Comput. Biol.* **9**, e1003118 (2013).

21. E. al Liao Y, featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. - PubMed - NCBI, (available at https://www.ncbi.nlm.nih.gov/pubmed/24227677).

22. P. Ewels, M. Magnusson, S. Lundin, M. Käller, MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*. **32**, 3047–3048 (2016).

23. 1000 Genomes | A Deep Catalog of Human Genetic Variation, (available at http://www.internationalgenome.org/).

24. O. Delaneau *et al.*, A complete tool set for molecular QTL discovery and analysis. *Nat. Commun.* **8**, 15452 (2017).

25. O. Stegle, L. Parts, M. Piipari, J. Winn, R. Durbin, Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* **7**, 500 (2012).

26. K. D. Hansen, R. A. Irizarry, Z. Wu, Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics*. **13**, 204–216 (2012).

27. N. Kumasaka, A. Knights, D. Gaffney, High resolution genetic mapping of causal regulatory interactions in the human genome. *bioRxiv* (2017), p. 227389.

28. H. Zhao *et al.*, CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics*. **30**, 1006–1007 (2014).

29. E. al Coetzee SG, motifbreakR: an R/Bioconductor package for predicting variant effects at transcription factor binding sites. - PubMed - NCBI, (available at https://www.ncbi.nlm.nih.gov/pubmed/26272984).

30. D. Lee *et al.*, A method to predict the impact of regulatory variants from DNA sequence.

Nat. Genet. **47**, 955 (2015).

31. M. Wang, C. Tai, W. E, L. Wei, DeFine: deep convolutional neural networks accurately quantify intensities of transcription factor-DNA binding and facilitate evaluation of functional non-coding variants. *Nucleic Acids Res.* (2018), doi:10.1093/nar/gky215.

32. MotifDb. *Bioconductor*, (available at http://bioconductor.org/packages/MotifDb/).

33. [No title], (available at http://hocomoco11.autosome.ru/final_bundle/hocomoco11/full/HUMAN/mono/logo_large/CTCF_HUMAN.H11MO.0.A_direct.png).

34. [No title], (available at http://hocomoco11.autosome.ru/final_bundle/hocomoco11/full/HUMAN/mono/logo_large/SPI1_HUMAN.H11MO.0.A_direct.png).

35. D. Lee, LS-GKM: a new gkm-SVM for large-scale datasets. - PubMed - NCBI, (available at https://www.ncbi.nlm.nih.gov/pubmed/27153584).

36. OpenAI, Gym: A toolkit for developing and comparing reinforcement learning algorithms, (available at https://gym.openai.com).

Non-exclusive licence to reproduce thesis and make thesis public

I, Yurii Toma,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to:

   1.1. reproduce, for the purpose of preservation and making available to the public, including for addition to the DSpace digital archives until expiry of the term of validity of the copyright, and

   1.2. make available to the public via the web environment of the University of Tartu, including via the DSpace digital archives until expiry of the term of validity of the copyright,

   of my thesis
   Predicting the impact of non-coding genetic variants with machine learning
   Supervised by Kaur Alasoo and Dmytro Fishman

2. I am aware of the fact that the author retains these rights.
3. I certify that granting the non-exclusive licence does not infringe the intellectual property rights or rights arising from the Personal Data Protection Act.

Tartu/Waterloo, 21.05.2018