Tampereen yliopisto

Matias Partanen

COMPARING MOTIF ENRICHMENT IN BENIGN AND CANCEROUS PROSTATE TISSUES

Faculty of Medicine and Health Technology
Bachelor’s thesis
April 2021

ABSTRACT

Matias Partanen: Comparing motif enrichment in benign and cancerous prostate tissues
Bachelor's thesis
Tampere University
Bachelor's degree program in biotechnology
April 2021

Prostate cancer is one of the most common cancers in men. To better be able to treat cancers, they first need to be understood better. One avenue towards this understanding is to understand what differentiates cancer tissue from normal tissue and what differences there are between cancer types. One way to do this is to analyze different tissues at the epigenetic and chromatin level. This thesis used data that has been previously generated from different prostate tissues, including two prostate cancer types and benign prostatic hyperplasia, by using the technique assay for transposase-accessible chromatin using sequencing (ATAC-seq). The aim of this thesis was to try and find differences in motif enrichment between the tissue types.

The data generated by using ATAC-seq indicates stronger signals in areas of the genome where the chromatin is more accessible and open. This openness can indicate binding sites for transcription factors. By utilizing the genomic position of these signal peaks, it is possible to see if any position and its corresponding sequence are more common in certain samples than in others. This can be achieved in different ways, but for this study, a tool called Hypergeometric Optimization of Motif EnRichment (HOMER) was used.

The analysis results indicated several known and few possibly new motifs that were consistent across sample groups as well as motifs that seem to be more common in only one or two sample groups. Examples of consistent motifs found are Sp1, FOXA1, and CTCF. However, further analyses and validation of found sequences are needed to determine further significance.

Keywords: prostate cancer, motif enrichment, ATAC-seq, HOMER, transcription factor

TIIVISTELMÄ

Matias Partanen: Rikastuneiden transkriptiotekijöiden sitoutumissekvenssien vertailu
hyvänlaatuisessa eturauhasen liikakasvussa ja eturauhassyövässä
Kandidaatin tutkielma
Tampereen yliopisto
Bioteknologian tutkinto-ohjelma
Huhtikuu 2021

Eturauhassyöpä on yksi miesten yleisimmistä syövistä. Jotta syöpiä voidaan tulevaisuudessa hoitaa paremmin, pitää niitä ja niiden toimintaa ymmärtää paremmin. Eräs tapa edistää tätä ymmärrystä on käsittää, mitkä tekijät erottavat syövän terveestä kudoksesta ja mitkä tekijät erottavat syöpätyyppejä toisistaan. Eräs tapa tutkia tätä on keskittyä kudostyyppien välisiin eroihin epigeneettisellä ja kromatiinitasolla. Tämän tutkielman tarkoituksena on hyödyntää eri eturauhaskudoksista assay for transposase-accessible chromatin using sequencing (ATAC-seq) -tekniikalla aiemmin tuotettua dataa ja katsoa, löytyykö eturauhassyöpätyyppien ja hyvänlaatuisen eturauhasen liikakasvun väliltä eroja transkriptiotekijöiden sitoutumissekvenssien määrissä.

ATAC-seq:illä tuotettu data antaa vahvempia signaalipiikkejä kromatiinin alueilla, jotka ovat avoimempia. Avoimempi alue voi olla sitoutumisalue transkriptiotekijöille, jolloin tämä kromatiinin osa ja sekvenssi voivat olla siihen sitoutuvalle transkriptiotekijälle ominaisia. Näiden alueiden sijainteja ja niitä vastaavia signaalipiikkejä analysoimalla voidaan yrittää selvittää, esiintyykö jokin tietty sekvenssi, ja siihen sitoutuva transkriptiotekijä, tavanomaista yleisemmin tietyssä näytteessä. On monia eri tapoja analysoida näitä sijainteja ja signaalipiikkejä. Tässä tutkielmassa käytettiin työkalua nimeltä Hypergeometric Optimization of Motif EnRichment (HOMER).

Analyysin perusteella saatiin useita mahdollisia tunnettuja ja tuntemattomia sitoutumissekvenssejä, jotka voivat olla kudostyypistä riippumattomia tai riippuvaisia. Esimerkkeinä löytyneistä tunnetuista sitoutumissekvenssejä vastaavista transkriptiotekijöistä ovat FOXA1, Sp1 ja CTCF. Löytyneitä sekvenssejä pitää silti tutkia tarkemmin, jotta niiden merkitsevyys selviää paremmin.

Avainsanat: eturauhassyöpä, ATAC-seq, HOMER, sitoutumissekvenssi, transkriptiotekijä

PREFACE

This thesis was done as a project for Tampere university's Computational Biology research group in the Faculty of Medicine and Health Technology. I would like to thank my thesis advisor Matti Nykter for the opportunity to do this project and for his help in its successful completion. I would also like to thank my friends and family for their support during this process.

Tampere, 27.4.2021

Matias Partanen

TABLE OF CONTENTS

1. INTRODUCTION

Prostate cancer was the second most diagnosed cancer and the fifth most common cause of cancer death in men globally in 2020 (Sung et al., 2021). Cancers in general are difficult to treat because of their capabilities to adapt and mutate in response to the body's protective responses (Alberts et al., 2015). Prostate cancer and healthy prostate tissue can be categorized differently based on their malignancy and progression but for this text, the relevant ones are benign prostatic hyperplasia (BPH), castration-resistant prostate cancer (CRPC), and primary prostate cancer (PC).

To better treat these diseases in the future, it is imperative to know more about them and more about what differentiates them from each other. When more is found out about the changes that happen in healthy tissue and what changes might induce progress from healthy to cancerous tissue, it might pave way for future treatments and drugs.

One possibility in trying to identify differentiating traits in different prostate cancer types and benign hyperplastic tissue is focusing on their differences at the epigenetic level. One of the techniques that can be used to achieve this is a technique called assay for transposase-accessible chromatin using sequencing (ATAC-seq) (Buenrostro et al., 2013).

The technique relies on a hyperactive Tn5 transposase that has been loaded with adaptors. Tn5 can integrate these adaptors into open areas of chromatin during its integration. In essence, the more open the chromatin is in a region the higher the probability of the ATAC-seq signal being stronger in that region. The resulting chromatin landscape can then be used as an indication of what areas are open. These areas can then indicate areas of importance in the chromatin. This technique has been successfully used to map several cancer types' chromatin accessibility landscapes (Corces et al., 2018)

In their extensive study of prostate cancer progression, Uusi-Mäkelä et al. discovered that transcription factor binding syntaxes are similar between tumor sample types regardless of chromatin heterogeneity (Uusi-Mäkelä et al., 2020) using a predictive BPNET model (Avsec et al., 2019). The purpose of this thesis is to explore the ATAC-seq data generated by Uusi-Mäkelä et al. in another way to see if individual differences in transcription factor motifs could be found between sample groups.

2. MATERIALS AND METHODS

2.1 Data and programs

As previously mentioned, the data used in this thesis was generated by Uusi-Mäkelä et al. and further details on its generation can be found in the original publication (Uusi-Mäkelä et al., 2020). The data had already been processed so any additional normalization or cleaning of the data was unnecessary.

The data consists of about 180,000 rows which represent 500 base-pair widths in the genome. Each row contains information about a gap's precise location, the chromosome in which it is located, and additional information. Of particular importance are the last 38 columns which contain ATAC-seq signal strengths for each tissue sample. The data contains 11 benign prostatic hyperplasia (BPH) samples, 11 castration-resistant prostate cancer (CRPC) samples, and 16 primary prostate cancer (PC) samples. The ATAC-seq signals varied between [0,14938] with most signals being between [0,100] depending on the sample.

The code and functions that facilitated the analysis and manipulation of data were made with Python (Van Rossum, Drake, 2009) and utilizing mainly the modules NumPy (Harris et al., 2020) and Pandas (McKinney, 2010). Additionally, Microsoft Excel was used in the analysis of the relative appearance rates of the more notable motifs.

2.2 HOMER

Hypergeometric Optimization of Motif EnRichment (HOMER) (Heinz et al., 2010) is a host of tools that can be utilized in sequence analysis and motif enrichment. In this thesis HOMER's findMotifsGenome.pl was used. At its core, HOMER aims to find motifs and sequences that appear more often than could be expected when compared to a corresponding background.

HOMER accomplishes this by taking the genomic positions given to it and trying to provide a suitable background for the analysis. This includes trying to match GC%-content and normalizing other base weights. Afterward, HOMER tries to find *de novo* motifs by essentially calculating how many times certain bases and their combinations appear in the given genomic positions and comparing this to how often the same bases and base combinations appear in the background.

After this, the potential motifs are scored, and the best motifs are reported with a possible name based on similarities between known motifs and motifs found in JASPAR. At the same time, HOMER tries to match well-fitting known motifs to the given genomic positions as well, reporting them in the process. More details on the process can be found on HOMER's website (http://homer.ucsd.edu/homer/index.html; 26.4.2021).

2.3 Generating HOMER results

HOMER takes its genomic positions in the form of a file that has been formatted in a specific way. Namely the file needs to contain the chromosome, end-, and start-positions, a unique ID, and the strand direction as columns. Because ATAC-seq doesn't contain information about strand orientation, the strand column was made a uniform 0 denoting a positive-strand in all sample files.

An input file corresponding to a sample was created by omitting the rows in the source data based on the ATAC-seq signal strength in the corresponding sample's column. The signal threshold that signified a strong enough indication that a particular part of the chromatin is open was decided to be 5. To generate a corresponding input file meant to omit all genomic positions (i.e. rows) that had a signal strength less than 5 in that sample.

This was repeated for each of the 11 BPH, 11 CRPC, and 16 PC samples with additional signal thresholds of 10, 15, and 20. Higher signal thresholds omit an increasing number of genomic positions. To ensure an adequate number of genomic positions in the input files, the threshold wasn't raised higher than 20. In total 153 input files corresponding to 44 BPH, 44 CRPC, and 64 PC were analyzed, and an additional file, that was generated from the base data with no omissions in case it was found useful later.

2.4 Aggregating HOMER outputs

HOMER analyzes both already known motifs and *de novo* motifs and tries to match them to given genomic positions. The outputs from a single input file are given as several different motif files. These files contain a lot of information for a particular motif, but the most pertinent ones for this thesis were the sequences, the guessed or known motif name, and the associated p-value for that sequence.

By going through all the output files and the information contained in them, it is possible to arrange each sample with its corresponding motifs. The p-values can then be used to gauge the importance of the motifs in that specific sample. For *de novo* motifs HOMER also offers similar or possible reverse motifs for several found motifs as alternatives. These were ignored for simplicity resulting in only the primary suggested *de novo* motifs being included in the aggregation.

According to a motif's associated p-value in a specific sample, it was given a value of whether it was strong, present, or equivalent to a non-existent signal. The values were assigned as follows: p-value < 0.0001 equaled a strong signal, 0.0001 ≤ p-value < 0.001 equaled a present signal, and p-value ≥ 0.001 equaled a non-existent signal, but the motif was still present in the sample.

Based on these values the motifs were given numerical codes. Strong, present, non-existent but present, and not found at all, were given numerical values of 2, 1, 0, and -1 respectively. The sum of these values would help in getting an idea of a motif's presence across multiple samples forming a rough scoring system by which to sort the sequences by.

All the samples and their associated motifs were collected into 2 matrices consisting of unique *de novo* motifs and unique known motifs. By splitting the final matrices into 3 parts each corresponding to different tissue types it would be easier to compare motifs across sample types. By counting column sums and counts of numerical values in motif columns it is easy to get an indication of how many samples and how strongly a motif appears across the samples.

3. RESULTS

3.1 Ranking the motifs

The original data consists of 38 samples – 11 BPH, 11 CRPC, and 16 PC – and their respective ATAC-seq signals. By analyzing each sample 4 times at different signal thresholds the total number of analyses was 152. Thus, it was decided that for a motif to be considered significant, it must appear with a strong signal more than four times as this would denote that a particular motif appears in more than just a single sample.

When searching for both motifs in common and different between the sample types, 2 factors were considered to form a rule of thumb. The first factor is the number of strong signals that appear in the sample type. This was calculated by dividing the sum of strong signals by the number of analyses done on the sample group and multiplying the result by 100%. The second factor is essentially the same as the first one, but the divisible sum also includes the number of present signals.

The strong signal appearance rate had to be over 9.09 % for the BPH and CRPC groups and over 6.25 % for the PC group for a motif to be considered significant. These percentages correspond to the threshold of a motif appearing 4 times.

The combined strong and present signal appearance rate was used when considering the differences in motif appearance rates between sample groups and when filtering motifs of interest. For a motif to be considered of less meaning in a sample group, the appearance rate of the combined strong and present signal could at most be 9.09 % for the BPH and CRPC groups and at most 6.25 % for the PC group. This way, motifs that could have been found because of one sample would be excluded.

For a motif to be of initial interest either the appearance rate of the strong signal or the appearance rate of the combined strong and present signal had to be over 9.09 % for the BPH and CRPC groups and over 6.25 % for the PC group. It was enough that a motif had either of these values in a single sample group after which the values were inspected in the other groups as well.

3.2 Known motifs

In total, 431 unique known motifs were found across all samples. This is noteworthy because by default HOMER tries to match 440 known motifs to each input file. This leads to the conclusion several motif sequences are identical to each other in the known motifs. By looking at the duplicates in the analysis results it became apparent that HOMER's known motifs can have the same motif sequence but a different motif name depending on what database the motif name is from or depending on what sequencing experiment is associated with it.

In short, this number of known motifs means two things. At least for these samples, there is a very limited pool of known motifs to match. Additionally, all the known motifs can be found in each sample. Their p-values do vary which could be used as a differentiating factor.

Out of the 431 unique motifs, 35 motifs of initial interest were found and inspected more closely. 9 of these motifs had an appearance rate higher than could be explained by only one sample across all sample groups. The motifs and their appearance rates (strong signal rates in parentheses) in different sample groups can be found in table 1. The motifs are roughly in the order that they appear in the PC group sorted by the scoring system. The motif names are from HOMER's motif library and follow its naming conventions.

Table 1 *The notable known motifs that were consistent across sample groups.*

Motif name	Appearance in PC samples % (%)	Appearance in CRPC samples % (%)	Appearance in BPH samples % (%)
Sp1	59.38(43.75)	70.45(54.55)	64.63(38.64)
FOXA1	34.38(28.13)	27.27(25.00)	25.00(18.18)
CTCF	31.25(29.69)	31.82(29.55)	34.09(29.55)
FOXM1	32.81(25.00)	25.00(25.00)	20.45(18.18)
CDX4	28.13(18.75)	29.55(20.45)	25.00(22.73)
FoxD3	26.56(20.31)	20.45(13.64)	15.91(15.91)
Foxf1	28.13(18.75)	18.18(15.91)	13.46(13.46)
Foxa2	25.00(20.31)	13.64(13.64)	11.36(11.36)
Boris	21.88(18.75)	22.73(18.18)	25.00(25.00)

4 motifs had appearance rates that differed between groups. Either the motif was significant in only one sample group or it was insignificant in only one sample group. These motifs can be found in table 2. The format is the same as for table 1. As the motif names are from HOMER's motif library,

it is unsure if the motif name “Unknown” reported by HOMER is the literal name or if the name for that sequence is truly unknown.

***Table 2** The notable known motifs that differed between sample groups.*

Motif name	Appearance in PC samples % (%)	Appearance in CRPC samples % (%)	Appearance in BPH samples % (%)
Unknown	18.75(10.94)	9.09(4.55)	18.18(15.91)
FOXK2	15.63(10.64)	6.82(4.55)	2.27(2.27)
Foxo3	14.06(10.94)	15.91(11.36)	9.09(2.27)
Hoxd10	12.50(12.50)	0(0)	4.45(2.27)

The remaining 22 motifs were not included in the tables because they either did not have a strong enough signal or the signal appearance rates were ambiguous.

3.3 *De novo* motifs

5595 unique *de novo* motif sequences were found across all samples. As opposed to the known motif results, slightly different *de novo* motifs could have an identical motif name. This relates to the fact that HOMER guesses the de novo findings based on its library of known motifs and motifs in JASPAR. This is illustrated by the fact that 38 different *de novo* sequences had a guessed motif name of Sp1 with varying degrees of confidence. Other motif names also had several candidates for them.

Out of all *de novo* motifs, there were 15 notable ones found. Out of these 15, only 2 were consistent across all samples. They and their appearance rates can be found in table 3. Table 3 follows the same format as the previous tables.

***Table 3** The notable de novo motifs that were consistent across sample groups.*

Guessed motif name	Appearance in PC samples % (%)	Appearance in CRPC samples % (%)	Appearance in BPH samples % (%)
ZNF711	15.63(7.81)	27.27(13.64)	34.09(20.45)
PB0207.1_Zic3_2	14.06(14.06)	13.64(13.64)	22.73(22.73)

A total of 12 *de novo* motifs had differences between sample groups. These motifs can be found in table 4. Table 4 follows the same format as the previous tables. The 1 remaining notable *de novo* motif was not included in the results because of its signal appearance rates’ ambiguity. The motifs marked with an asterisk (*) have a counterpart in the known motifs with the same name. The sequences are not identical, but the sequences may be referring to the same possible transcription factor.

***Table 4** The notable de novo motifs that differed between sample groups.*

Guessed motif name	Appearance in PC samples % (%)	Appearance in CRPC samples % (%)	Appearance in BPH samples % (%)
ETV4	14.06(14.06)	0(0)	2.27(2.27)
NRF1*	14.06(14.06)	9.09(9.09)	18.18(18.18)
Tbx5	14.06(9.38)	2.27(2.27)	2.27(2.27)
MSANTD3	9.38(7.81)	2.27(0)	2.27(2.27)
ZNF460	10.94(9.38)	2.27(2.27)	13.46(11.36)
CTCFL	10.94(7.81)	0(0)	6.82(6.82)
PB0146.1_Mafk_2	9.38(9.38)	2.27(2.27)	2.27(2.27)
Sp1*	9.38(9.38)	0(0)	9.09(9.09)
ELF4	7.81(7.81)	2.27(2.27)	0(0)
PB0181.1_Spdef_2	7.81(7.81)	0(0)	0(0)
ZNF75D	4.69(4.69)	15.91(11.36)	4.55(4.55)
Sp1*	3.13(3.13)	2.27(2.27)	18.18(18.18)

4. DISCUSSION

The known motifs that were most significant across the sample groups (Table 1) are partly in line with the findings from the BPNET analysis that has been done with the same data. The motifs corresponding to FOXA1, Sp1, and CTCF were consistent across sample groups and these were also found in the original analysis (Uusi-Mäkelä et al., 2020). The AR and GRHL2 motifs were also found in the original analysis, but their presence in the HOMER analyses was very low compared to the other motifs shown in tables 1-4.

This can be a result of many factors. The parameters that were used during the HOMER analysis might have had an effect. HOMER's motif library might have differing or absent sequences and motif names corresponding to the motifs, though this is unlikely. The limits of what p-value constitutes a strong signal and what kind of an appearance rate a motif needs to be significant could need adjustment to modify the HOMER analysis's sensitivity.

Another factor to consider is the ignored alternative and reverse motifs that HOMER suggests for *de novo* motifs. Including these could have given better results. But this would have required the use of an additional tool for motif prediction because HOMER doesn't include guesses for these motifs. An option for this prediction would be e.g. Tomtom (Gupta et al., 2007). Tomtom could also be used to validate HOMER's guesses for the *de novo* motifs which could provide an interesting comparison for its accuracy.

Optimistically, the motifs that were consistent across sample groups – found in tables 1 and 3 – could correspond to sequences that could be used to mark or identify potential malignant tissues. And the motifs that had differences between sample groups – found in tables 2 and 4 – could be used as possible differentiating factors between both prostate cancer types and between malignant and nonmalignant prostate tissues.

Realistically, the incidences of many motifs are low compared to motifs that have been previously found. Additionally, just because a motif appears in an analysis that has been done on data gained from tumor samples doesn't mean that that motif has any malignant significance. The motifs' significance, validity, and purpose will have to be better determined with additional analyses, comparisons, and statistical tests.

In summary, by utilizing HOMER in analyzing ATAC-seq data it is possible to find both consistent and differing motif sequences between prostate cancer types and benign prostate hyperplastic tissue. These motifs can then be targets for additional study. By adjusting the workflow, significance limits, and HOMER parameters, it should be possible to improve the analysis further in the future.

5. REFERENCES

Alberts, B., Johnson, A., Lewis, J., et al. *Molecular biology of the cell,* Sixth Edition, Garland Science, Taylor & Francis Group.

Avsec, Z., Weilert, M., Shrikumar, A., et al. *Deep learning at base-resolution reveals motif syntax of the cis-regulatory code*.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., et al. "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position", *Nature Methods,* vol. 10, no. 12, pp. 1213-1218.

Corces, M.R., Granja, J.M., Shams, S., et al. "The chromatin accessibility landscape of primary human cancers", *Science,* vol. 362, no. 6413, pp. eaav1898.

Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L., et al. "Quantifying similarity between motifs", *Genome Biology,* vol. 8, pp. R24.

Harris, C.R., Millman, K.J., Van Der Walt, Stéfan J., et al. *Array programming with NumPy*, Springer Science and Business Media LLC.

Heinz, S., Benner, C., Spann, N., et al. "Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities", *Molecular Cell,* vol. 38, no. 4, pp. 576-589.

McKinney, W. 2010, "Data Structures for Statistical Computing in Python", *Proceedings of the 9th Python in Science Conference*, pp. 56.

Sung, H., Ferlay, J., Siegel, R.L., et al. "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries", *CA: a cancer journal for clinicians.*

Uusi-Mäkelä, J., Afyounian, E., Tabaro, et al. "Chromatin accessibility analysis uncovers regulatory element landscape in prostate cancer progression", *bioRxiv*, pp. 2020.09.08.287268.

Van Rossum, G. & Drake, F.L. 2009, *Python 3 Reference Manual*.