Motif analysis on OTX2 ChIP-seq peaks

LCG BEII personal work – 2019

Jesús Vélez Santiago: jvelez@lcg.unam.mx

2019-03-13

Contents

Introduction

Transcription factors are proteins that modulate the expression of target genes through the binding on DNA cis-regulatory elements. So the development of techniques that aim to identify DNA binding sites associated with proteins has had a breakthrough. One of those that has fulfilled this purpose on a large scale combines the immunoprecipitation of chromatin (ChIP) with DNA sequencing, giving rise to the ChIP-seq technique. A ChIP-seq experiment usually has hundreds of thousands of predicted sites (regularly called *ChIP-seq peaks*). These peaks are mapped to a reference genome to find possible binding patterns (also known as *motifs*) for the protein of interest. The present research seeks to find if there are other motifs in the human genome different from those found in databases such as Jaspar and Hocomoco of the transcriptional factor OTX2 (orthodenticle homeobox 2) that plays an important role in brain, craniofacial, and sensory organs development. To achieve this, the factor peaks from the ReMap database were used in combination with the tools provided by RSAT sofware.

Methods

Getting peaks from ReMap

A connection to ReMap was established. The OTX2 factor was searched and the peaks for MACS were downloaded in BED format (See Supplementary material, Table 3). Subsequently, the GSE ID was saved and the structure of the peak file was verified based on the fields specified in the ReMap entry to avoid inconsistencies.

Getting motifs from reference databases

Taking advantage of the fact that RSAT contains a database in which it stores matrices in transfac format of different *TF binding motifs databases* (e.g. Jaspar, Hocomoco) it was decided to create a script that established a connection with RSAT in order to obtain different matrices given a file with the identifiers and databases to search for the factors. (See Supplementary material, Table 4).

Discovering motifs with RSAT peak-motifs

To obtain the sequences of the peaks we used the *fetch-sequences* tool of RSAT. As a negative control, random genome fragments of the same length as the mapped sequences were generated (See Supplementary material, Table 3). Later these two datasets were used to find the motifs of the peak sequences using the RSAT *peak-motifs* web tool (See Supplementary material, Parameters RSAT web - peak-motifs).

Motif enrichment analysis

In order to know if the peaks were enriched in the reference motifs, the RSAT *matrix-quality* tool was used. The Markov order was changed to avoid the low representativeness that order 0 could mean. (See Supplementary material, Parameters RSAT web - matrix-quality).

Results and discussion

The results of the peak-motifs and quality-matrix tools are shown on a web page, so they were downloaded. The information to consult them is available in the supplementary material.
In the present study there are 15952 peaks of 15966 registered in the ReMap database. The reason for the difference is still unknown but since it is small it was decided to continue working with this data set.

Peak-motifs

Sequence composition

In this study it is observed that most of the sequences have an average length of 247bp, 159bp minimum, 1000bp maximum, and a total length of 3951kb (**Figure 1**). The nucleotide composition profile shows that there is a higher percentage of AT (60%) compared with GC (40%) (**Figure 2**), in addition, in the positional distribution for each dinucleotide shows a greater occurrence of AA, TT, AT and a lower occurrence of GC (**Figure 3**).

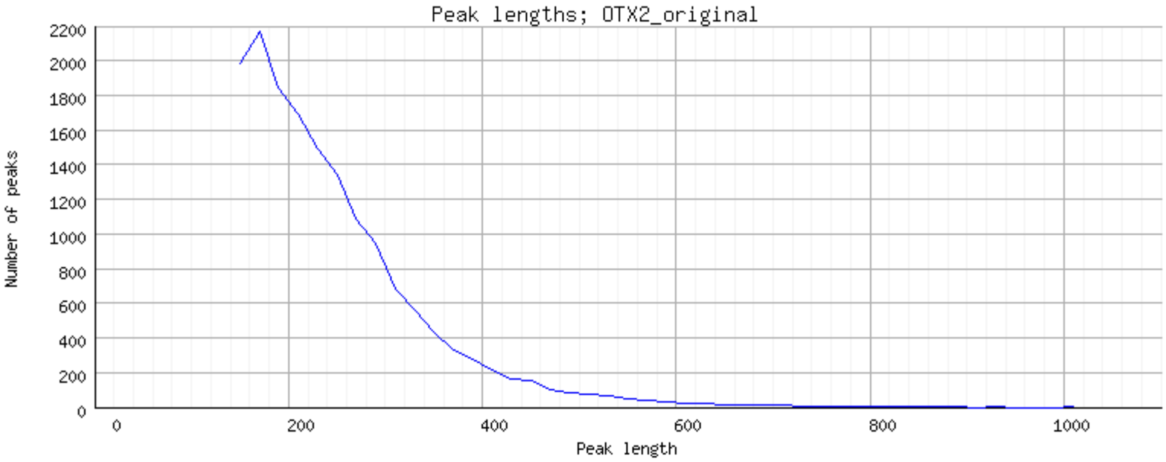

Figure 1: Length distribution of the sequences of OTX2. avg: 247bp, min:159bp, max: 1000, Total sequence size: 3951kb

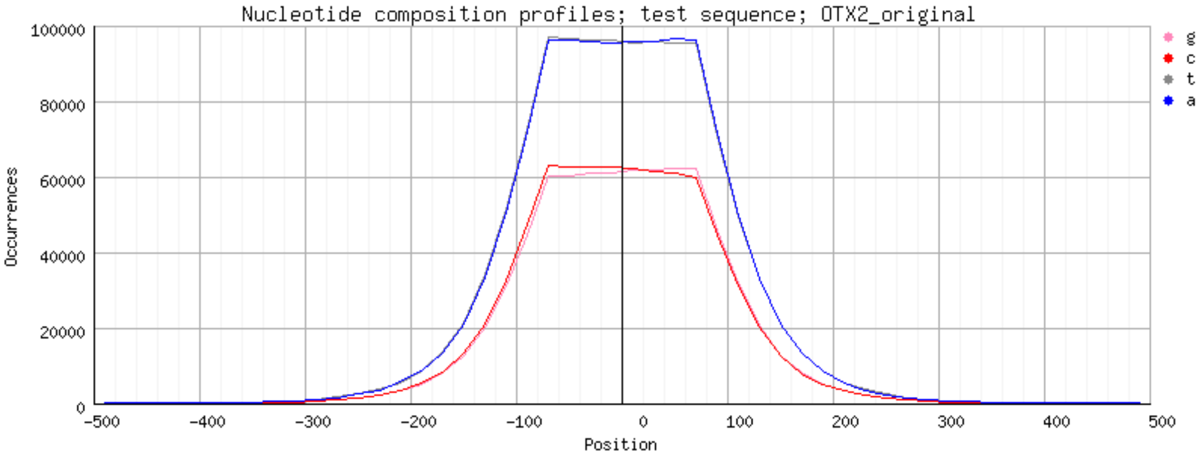

Figure 2: Individual nucleotide composition profile.

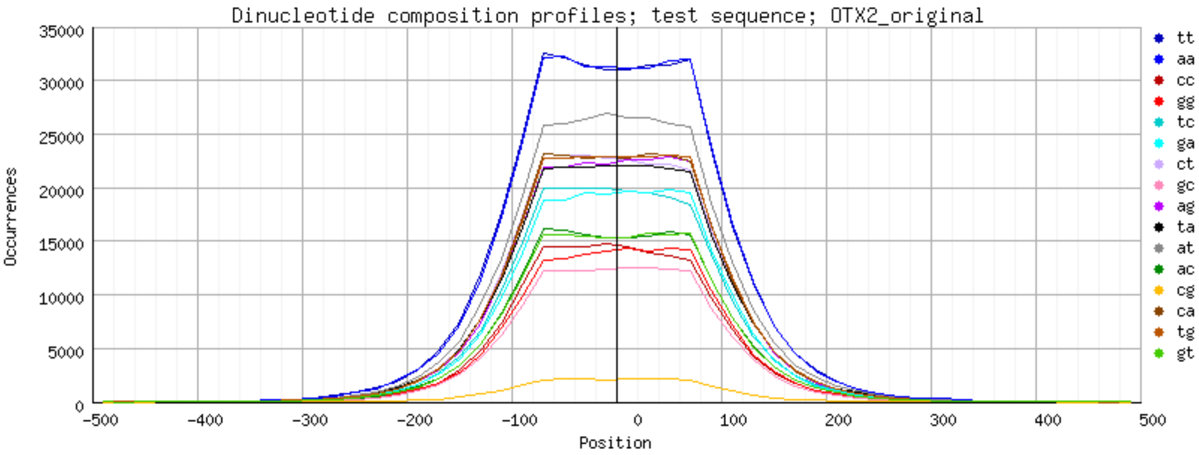

Figure 3: Dinucleotide composition profiles.

Reference Motifs

The reference motifs were downloaded from the database repositories of motifs contained in RSAT (transfac format) (See Supplementary material, Table 4), in this case, the selected databases were Jaspar and Hocomoco. Both reference motifs have similar information, resulting in almost identical logos but in complementary reverse (**Figure 4**). The main difference could be the length of the logos as well as the amount of information provided by each of these.

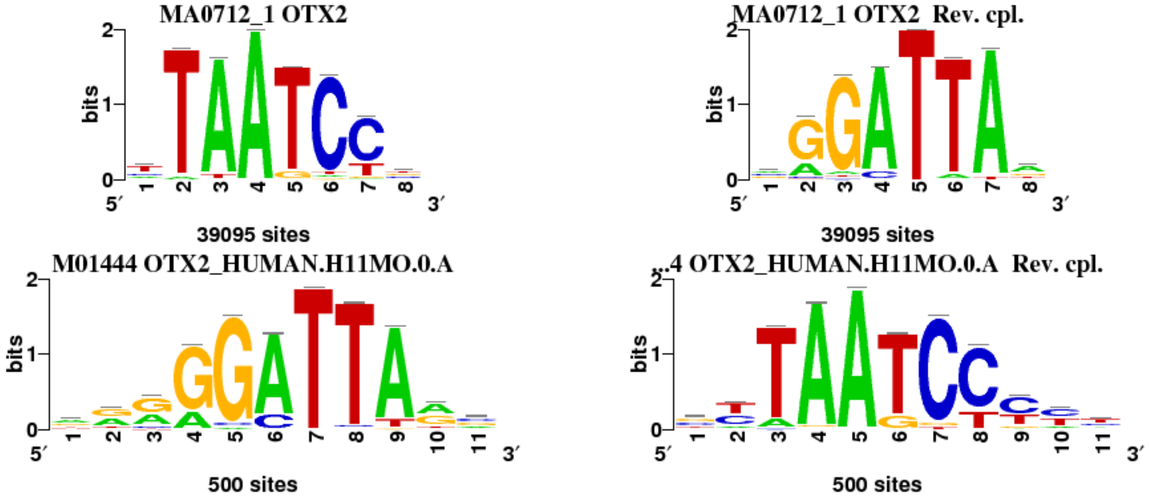

Figure 4: Jaspar, Hocomoco reference motifs of OTX2 per row, respectively.

Discovered motifs (by algorithm)

Supplementary material Parameters RSAT web - peak-motifs shows where to find the full list of discovered motifs by algorithm (*oligo-analysis*, *position-analysis*) for each selected k-mer size (6,7). It is also important to note that the algorithm converts the e-value (expected number of false positives) to a value of significance dictated by the formula: $-log10(e - value)$. The scores for the most significant motifs are shown in bold red. In this research, the $GGATTA$ motif is overrepresented with a significance of 152. The same motif in complementary reversal is found by *position-analysis* but with a significance of 300. It is noteworthy that these motifs correspond to the reference motives (not used in this analysis) which tells us that these are the most relevant in terms of overrepresentation, but there is another motif that is almost equally significant and is *TGA[CA]TCA* ({oligo, position}_6nt(?:_mkv4)_m2), with a significance of 143 and 300 (**Figure 5**). Comparing the results with the control (random peak sequences), the motifs discovered had significance scores in average of 12.

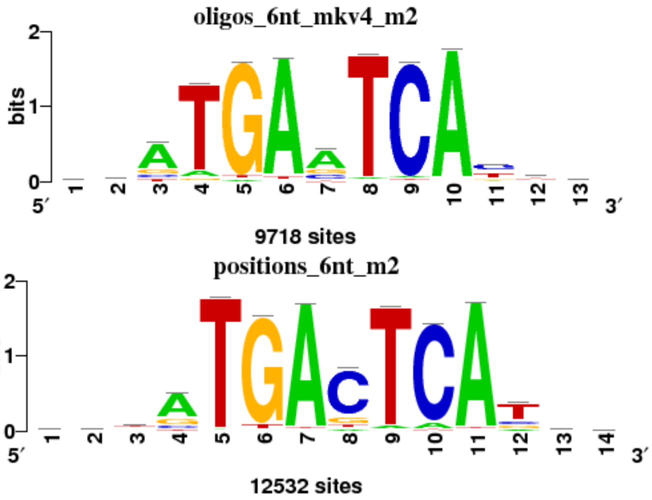

Figure 5: Discovered motifs found with oligo-analysis and position-analysis. Significance of 143, 300, respectively.

Discovered motifs (with motif comparison)

The motifs found with motif comparison with a high significance (bold red) fulfill the property of being extensions of the reference motifs (in total 9 motifs). Among the non-significant motifs is the one found in the **Figure 5** that despite not making a match with the reference motifs, it does so with different transcriptional factors. In future work it will be necessary to review in depth the characteristics of these factors to determine their exact relationship.

Matrix-quality

An analysis enrichment motifs was performed to highlight the binding sites in the datasets (i.e. Original peak sequences and random peak sequences), to assess enrichment the *matrix-quality* tool combines information from scores of theoretical and empirical distributions. As seen in the graph of decreasing cumulative distribution (**Figure 6** first column) the line representing the distribution of the original data set (red line) is above the theoretical (dark blue) and random (light blue) distributions, showing an enrichment - although small - for both motifs. Whereas, when looking at the Receiver Operating Characteristic (ROC) curve (**Figure 6** second column), the distributions can not be clearly differentiated. Although, just to mention, the original distribution seems to remain slightly above the rest. In spite of this, the ROC curve was considered as a bad curve. Concluding that the matrices do not have an enrichment in the original data set.

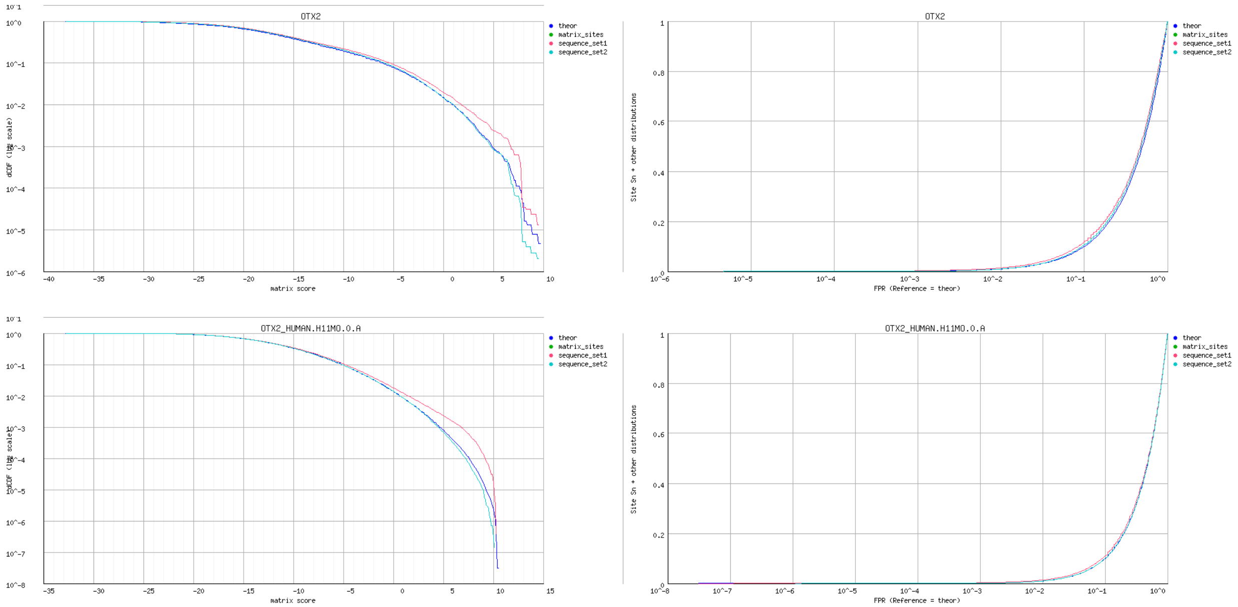

Figure 6: Decreasing cumulative distribution (CDF, logarithmic Y axis) and ROC curve (logarithmic X axis). Position as figure 4.

Conclusions and perspectives

In the research a pipeline is developed to characterize TF binding sites in sequences from ChIP-seq experiments. The method combines the use of transfac reference matrices together with databases of TF-binding sites with the aim of identifying motifs through algorithms such as *oligo-analysis* and *positional-analysis* (de novo discovery), and then comparing it against those that they reside in the databases and the reference motifs.
The result of the workflow for the ChIP-seq peaks for the OTX2 transcription factor found nine extended versions of the reference motifs, and two motifs that could be candidates as binding sites for the factor under study. However, the properties belonging to the factors that matched them have not yet been examined and will need to be taken into account in future research.
In future pipelines for the discovery of motifs it could be interesting to test clustering algorithms of the discovered motifs to eliminate redundancy if they exist, in addition to add some other negative control such as the permutation of the reference matrices and ensure the possible effect of the Markov model in the discovery algorithms. On the other hand the use of neural networks for the prediction of genomic characteristics has had a great boom in the current era, so it would not be wrong to take it as a promising new approach and see what happens.

Supplementary material

All the data and scripts used for this research are available on the github platform. So you can *clone* the repository in your own directory.

Copy and paste this line in your terminal:

- git clone https://github.com/jvelez-lcg/chip-seq.git

Note: all the paths specified in this section are relative to the cloned directory.

Questions and clarifications: jvelez@lcg.unam.mx

Bioinformatics resources used for this work

Table 1: Bioinformatics Sources.

Acronym	Description	URL
RSAT Metazoa	Regulatory Sequence Analysis Tools Metazoa	http://metazoa.rsat.eu/
ReMap	A database of ChIP-seq peaks	http://remap.cisreg.eu/
Jaspar	A database of eukaryote TF binding motifs, mainly built from CHIP-seq peaks	http://jaspar.genereg.net/
Hocomoco	A database of Human TF binding motifs, mainly built from CHIP-seq peaks	http://hocomoco11.autosome.ru/

Data sources

Table 2: Data Sources.

Data	Source/Value
Factor	OTX2
Tissue/Cell type	Retinal
Remap entry	OTX2
Peaks	GSE60024
Jaspar matrix	MA0712.1
Hocomoco matrix	M01444

Complete list of commands and parameters

Parameters RSAT Command Line

Parameters needed to download the genomic coordinates of the OTX2 factor, recover its sequences and generate negative control.

Table 3: RSAT command line parameters.

Parameter	Value
factor	OTX2
bed_url	url
genome	hg38
format	UCSC
organism	Homo_sapiens_GRCh38
bed	factor.bed
bed_fasta	factor.fasta
bed_rand_fasta	factor.random_genome_fragments.fasta

- Download bed file of transcription factor.

```
wget -O $bed.gz $bed_url && try gunzip $bed.gz
```

- RSAT - fetch-sequence.

```
fetch-sequences  -v 1 -genome $genome -header_format $format -i $bed -o $bed_fasta
```

- RSAT - random genome fragments
  `random-genome-fragments -i $bed_fasta -org $organism  -return seq  -o $bed_rand_fasta`

See documentation of `./chip_seq/get_requirements_for_the_RSAT_page.sh` file.

Parameters to download transfact matrixs

Table 4: Custom script parameters.

Parameter	Value
pythonf	python get_transfac_matrix_from_RSAT.py
tranfact_queries	transfac_queries.csv
outtype	all
outpath	./
sep	,

See documentation of `./chip_seq/get_requirements_for_the_RSAT_page.sh` file.
See documentation of `./chip_seq/get_transfac_matrix_from_RSAT.py` file.

- Download transfac matrix from queries file.
  `python get_transfac_matrix_from_RSAT.py -i transfac_queries.csv -o motifs`

Parameters RSAT web - peak-motifs

Files used:

- `OTX2.fasta`
- `OTX2.random_genome_fragments_hg38.fasta`
- `OTX2_reference_motifs.tf`

Compare motifs with databases:

- `JASPAR core nonredundant vertebrates (2018)`
- `Hocomoco (HUMAN TFs) (2017-10)`

Results available in:

- Original data: `./chip_seq/rsat_web/peak-motifs_OTX2_original.html`.
- Negative control: `./chip_seq/rsat_web/peak-motifs_OTX2_random.html`.

Parameters RSAT web - matrix-quality

Files used:

- `OTX2_reference_motifs.tf`
- `OTX2.fasta`
- `OTX2.random_genome_fragments_hg38.fasta`

Background:

- `Markov order = 3`
- `Organism = Homo sapiens GRCh38`

Result available in:

- Result: `./chip_seq/rsat_web/matrix-quality_result.html`

Motif comparisons graphs

Results availabe in:

- Result_ `./chip_seq/Cytoscape/`