BIOINFORMATICS

Vol. 00 no. 00 2017
Pages 1–3

Motif discovery and its analysis for binding sites of WhiH (White H) transcription factor in Streptomyces

Sergii Gladchuk*, Klas Flärdh and Björn Canbäck
Department of Biology, Box 118, 221 00, Lund University, Sweden

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: The aim of the developed procedure was to discover and analyze sequence motifs based on results of ChIP-seq experiments with MEME suit that may constitute binding sites for the WhiH protein. The initial procedure described in MEME manual required to much manual conversions and input data modifications in order to get appropriate input for MEME-ChIP program. Also further motif correlation with expression data has to be confirmed.

Results: Set of steps and scripts where developed to facilitate whole motif discovery procedure and ChIP-seq/motifs correlation with expression data. The final statistical analysis confirmed success of ChIP-seq experiment and produced list of motif-relevant significantly over-/under-expressed genes, which can illuminate the mechanism of transcriptional regulation.

Availability: Detailed description of full analysis together with all the results can be found on public repository (under results folder, notebook.html): https://github.com/sergiigladchuk/WhiH_motiff

Contact: se1522gl-s@student.lu.se

1 INTRODUCTION

WhiH is a transcriptional regulator of the GntR family that controls late stages of sporulation and cell division in Streptomyces. Chromatin Immuno Precipitation followed by next generation sequencing (ChIP-seq) experiment has been conducted to identify regions of DNA that are bound by WhiH during sporulation of the model organism Streptomyces venezuelae. Identifying a main motif in a large fraction of the peaks by motif analysis can confirm successful experiment and also identify the DNA-binding motifs of other proteins that bind in complex or in conjunction with the ChIPed protein, illuminating the mechanisms of transcriptional regulation (Timothy *et al.*, 2009).

In addition, microarray-based transcriptomic analyses have also been performed to monitor patterns of gene expression in wild type and whiH mutant strains during growth and sporulation.

Initial analysis of the data showed that WhiH has very complex regulon structure and motif discovery together with expression data correlation can better identify genes, which are under direct WhiH control.

2 METHODS

Chromatin immunoprecipitation, library construction, sequencing, and ChIP-seq data analysis were performed by The Genome Analysis Centre (TGAC), Norwich Research Park Norwich, United Kingdom, as described in Bush *et al.*, 2013.

Motif discovery based on ChIP-seq peaks data provided. Two lists (all significant peaks and top 36 peaks), which correspond to genomic coordinates in *S. venezuelae* genome (Gen-Bank accession number NC_018750, Pullan *et al.*, 2011), were separately used to extract 500-nucleotide-long ranges with python script *seq_extractor*. These ranges where fed to MEME-ChIP program (Timothy *et al.*, 2009). Numerous settings (different level Markov Models for background, palindromic only, discriminative mode) in different combinations were applied to make motiff more specific and show better back-check results based on FIMO program from MEME suit, which match given motif to the genome. Only 4 best motifs were selected for further statistical analysis:

- ’Best TOP non-palindromic discriminative motif discovered with MEME-ChIP (settings: discriminative mode against 500 random ranges with 0-model background, non-palindromic, based on TOP ChIP-seq regions)
- ’TOP palindromic motif’ - discovered with MEME (settings: background Markov model order 0 with palindrome only, based on TOP ChIP-seq regions)
- ’Best ALL-Peaks back-check discriminative motif’ discovered with MEME-ChIP (settings: 10000 random ranges with 0-model background, non-palindromic, based on ALL ChIP-seq regions)
- ’Best e-value from ALL peaks’ - discovered with MEME (settings: background Markov model order 0, non-palindromic, based on ALL ChIP-seq regions)

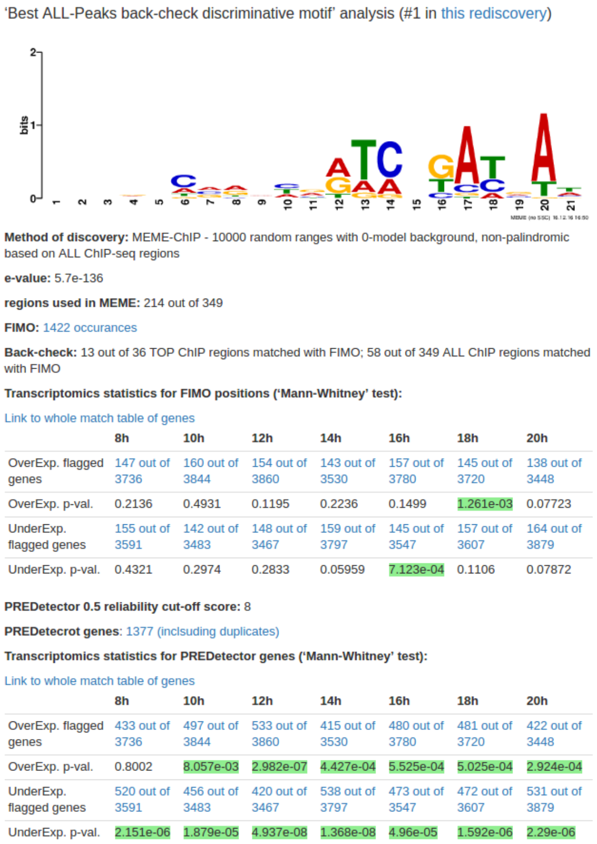
Statistical analysis for expression dependency of genes in close proximity to discovered motifs was done in R. For each selected motif list of positions from FIMO program was used to identify nearby genes on both DNA strands with developed python program *affy_log_creator*. This script takes 3 inputs: transcriptomics data file, annotation file of all genes for *S. venezuelae* and list of positions. It converts transcriptomics into AffyLog levels for each gene, and marks genes winch have motif position in region -300 away from the start codon and +50 after start codon.

Same *affy_log_creator* program was used to produce list of all genes with expression level and marked related genes from PREDetector program (Hiard *et al.*, 2007) and initial ChIP-seq peak positions.

Produced gene lists with AffyLogs and marked genes were used in actual statistical analysis in R. Procedure of the analysis:

1. whole list was processed 7 times (from 8h to 20h separately)

*to whom correspondence should be addressed


'Best ALL-Peaks back-check discriminative motif' analysis (#1 in this rediscovery)

Method of discovery: MEME-ChIP - 10000 random ranges with 0-model background, non-palindromic based on ALL ChIP-seq regions

e-value: 5.7e-136

regions used in MEME: 214 out of 349

FIMO: 1422 occurances

Back-check: 13 out of 36 TOP ChIP regions matched with FIMO; 58 out of 349 ALL ChIP regions matched with FIMO

Transcriptomics statistics for FIMO positions ('Mann-Whitney' test):

Link to whole match table of genes

	8h	10h	12h	14h	16h	18h	20h
OverExp. flagged genes	147 out of 3736	160 out of 3844	154 out of 3860	143 out of 3530	157 out of 3780	145 out of 3720	138 out of 3448
OverExp. p-val.	0.2136	0.4931	0.1195	0.2236	0.1499	1.261e-03	0.07723
UnderExp. flagged genes	155 out of 3591	142 out of 3483	148 out of 3467	159 out of 3797	145 out of 3547	157 out of 3607	164 out of 3879
UnderExp. p-val.	0.4321	0.2974	0.2833	0.05959	7.123e-04	0.1106	0.07872

PREDetector 0.5 reliability cut-off score: 8

PREDetecrot genes: 1377 (inclsuding duplicates)

Transcriptomics statistics for PREDetector genes ('Mann-Whitney' test):

Link to whole match table of genes

	8h	10h	12h	14h	16h	18h	20h
OverExp. flagged genes	433 out of 3736	497 out of 3844	533 out of 3860	415 out of 3530	480 out of 3780	481 out of 3720	422 out of 3448
OverExp. p-val.	0.8002	8.057e-03	2.982e-07	4.427e-04	5.525e-04	5.025e-04	2.924e-04
UnderExp. flagged genes	520 out of 3591	456 out of 3483	420 out of 3467	538 out of 3797	473 out of 3547	472 out of 3607	531 out of 3879
UnderExp. p-val.	2.151e-06	1.879e-05	4.937e-08	1.368e-08	4.96e-05	1.592e-06	2.29e-06

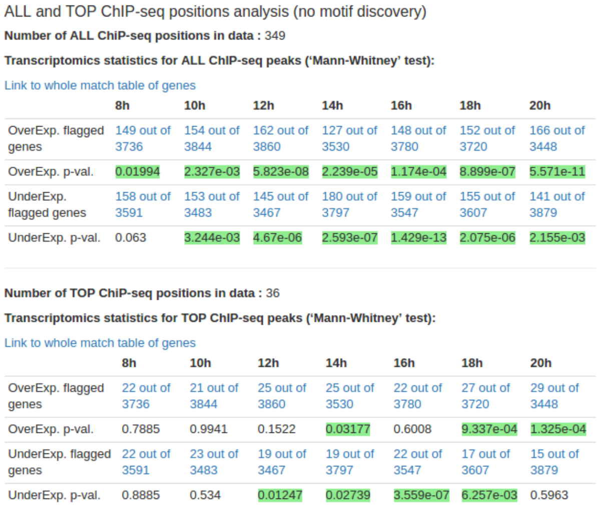
Fig. 1. Statistical analysis output for 'Best ALL-Peaks back-check discriminative motif'. Significant difference in expression between motif/peak related genes and other genes are colored in green

2. based on each time AffyLogs values gene list was split into two lists - over- and under- expressed genes (AffyLog - (-) indicates a decrease in expression of the gene in a whiH mutant compared to the wild-type; (+) indicates an increase in expression of the gene in a whiH mutant compared to the wild-type)
3. these sub-lists with two categories of genes (special genes marked due to closeness to Motif or ChIP-seq peak and all other non-related genes) were fed to one-sided MannWhitney non-parametric independent samples test to find significant difference in variation of expression data for two categories. This non-parametric test was chosen because data is not normally distributed.

For each set of positions there are 7 (times) * 2(over/under expr.) = 14 sub-lists and p-values that signify if there is true difference in transcriptomic expression between special genes and all the others.

R script produces table with p-values of Mann-Whitney test, and colors only significant (p-value less than 0.05) cells (Figure 1 and 2). Also there is a direct link to each list of genes so further analysis can be conducted.

Based on all significant lists of genes, rank tables of gene appearance were constructed separately for over- and under-expression (Only top 5 are present in Table 1). Hopefully, these tables can identify important genes, which are regulated by WhiH protein.


ALL and TOP ChIP-seq positions analysis (no motif discovery)

Number of ALL ChiP-seq positions in data : 349

Transcriptomics statistics for ALL ChIP-seq peaks ('Mann-Whitney' test):

Link to whole match table of genes

	8h	10h	12h	14h	16h	18h	20h
OverExp. flagged genes	149 out of 3736	154 out of 3844	162 out of 3860	127 out of 3530	148 out of 3780	152 out of 3720	166 out of 3448
OverExp. p-val.	0.01994	2.327e-03	5.823e-08	2.239e-05	1.174e-04	8.899e-07	5.571e-11
UnderExp. flagged genes	158 out of 3591	153 out of 3483	145 out of 3467	180 out of 3797	159 out of 3547	155 out of 3607	141 out of 3879
UnderExp. p-val.	0.063	3.244e-03	4.67e-06	2.593e-07	1.429e-13	2.075e-06	2.155e-03

Number of TOP ChiP-seq positions in data : 36

Transcriptomics statistics for TOP ChIP-seq peaks ('Mann-Whitney' test):

Link to whole match table of genes

	8h	10h	12h	14h	16h	18h	20h
OverExp. flagged genes	22 out of 3736	21 out of 3844	25 out of 3860	25 out of 3530	22 out of 3780	27 out of 3720	29 out of 3448
OverExp. p-val.	0.7885	0.9941	0.1522	0.03177	0.6008	9.337e-04	1.325e-04
UnderExp. flagged genes	22 out of 3591	23 out of 3483	19 out of 3467	19 out of 3797	22 out of 3547	17 out of 3607	15 out of 3879
UnderExp. p-val.	0.8885	0.534	0.01247	0.02739	3.559e-07	6.257e-03	0.5963

Fig. 2. Statistical analysis output for ALL and TOP ChIP-seq positions analysis (no motif discovery). Significant difference in expression between motif/peak related genes and other genes are colored in green

Table 1. Top 5 Over/Under-expressed gene apperance counts based on all significant lists

Gene	Product	Count
	Over-Expressed	
SVEN_1372	hypothetical protein	24
SVEN_1324	hypothetical protein	24
SVEN_1278	Gluconokinase	23
SVEN_1625	ATP-dependent RNA helicase	23
SVEN_4750	putative membrane protein	23
	Under-Expressed	
SVEN_5498	Transcriptional regulator GntR family	27
SVEN_4634	hypothetical protein	25
SVEN_4457	putative UDP-glucose or GDP-mannose dehydrogenase	24
SVEN_2614	hypothetical protein	24
SVEN_0269	hypothetical protein	23

These tables are also available in repository

3 DISCUSSION

Analyses of four motifs showed that PREDetector gene lists have better significance in statistical test than FIMO positions found from motif (also seen on Figure 1). That can explained by very simple approach used in FIMO program to identify matches of motif with no relation to genome structure.

Analysis of raw ChIP-seq peaks (Figure 2) showed that genes which are close to these peaks have significant difference of transcription comparing to other genes. So even without motif discovery, results produced based on ChIP-seq peaks only, are one of the best if number of significant times are compared.

4 CONCLUSION

The robust procedure and analyses described here confirmed the success of ChIP-seq experiment and produced the lists of ranked genes potentially controlled by WhiH, which will lead to better understanding of its overall regulon in future.

ACKNOWLEDGEMENT

We are grateful to people working in Bioinformatics department, Biology building, for constant support with overall analysis facilitation and small Linux/R/Rmarkdown tweaks and tricks, which were of great help to get this project done.

REFERENCES

Timothy,Bailey, Pawel,Krajewski, Istvan,Ladunga, Celine,Lefebvre, Qunhua,Li, Tao,Liu, Pedro,Madrigal, Cenny,Taslim, and Jie,Zhang (2013) Practical Guidelines for the Comprehensive Analysis of ChIP-seq Data. *PLoS Comput Biol*. 2013 Nov; 9(11): e1003326.

Bush,M.J., Bibb,M.J., Chandra,G., Findlay,K.C., Buttner,M.J. (2013) Genes required for aerial growth, cell division, and chromosome segregation are targets of WhiA before sporulation in *Streptomyces venezuelae*. *mBio*, **4(5)**: e00684-13. http://dx.doi.org/10.1128/mBio.00684-13.

Timothy,L.,Bailey, Mikael,Bodén, Fabian,A.,Buske, Martin,Frith, Charles,E.,Grant, Luca,Clementi, Jingyuan,Ren, Wilfred,W.,Li, William,S.,Noble (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Research*, 37:W202-W208.

Pullan,S.T., Chandra,G., Bibb,M.J. and Merrick,M. (2011) Genome-wide analysis of the role of GlnR in Streptomyces venezuelae provides new insights into global nitrogen regulation in actinomycetes. *BMC Genomics* **12**, 175

Hiard,S, Maree,R, Colson,S, Hoskisson,PA, Titgemeyer,F, van Wezel,GP, Joris,B, Wehenkel,L, Rigali,S. (2007) PREDetector: A new tool to identify regulatory elements in bacterial genomes. *Biochem Biophys Res Commun.* 357(4):861-4.