Bioinformatics, 32(3), 2016, 321–324
doi: 10.1093/bioinformatics/btv555
Advance Access Publication Date: 26 September 2015
Discovery Note

Genome analysis

Identification of coupling DNA motif pairs on long-range chromatin interactions in human K562 cells

Ka-Chun Wong[1],*, Yue Li[2] and Chengbin Peng[3]

[1]Department of Computer Science, City University of Hong Kong, Kowloon Tong, Hong Kong, [2]CSAIL, Massachusetts Institute of Technology, Cambridge, MA 02139-4307, USA and [3]CEMSE Division, King Abdullah University of Science and Technology, Thuwal, Jeddah, Kingdom of Saudi Arabia

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on June 25, 2015; revised on September 12, 2015; accepted on September 15, 2015

Abstract

Motivation: The protein–DNA interactions between transcription factors (TFs) and transcription factor binding sites (TFBSs, also known as DNA motifs) are critical activities in gene transcription. The identification of the DNA motifs is a vital task for downstream analysis. Unfortunately, the long-range coupling information between different DNA motifs is still lacking. To fill the void, as the first-of-its-kind study, we have identified the coupling DNA motif pairs on long-range chromatin interactions in human.

Results: The coupling DNA motif pairs exhibit substantially higher DNase accessibility than the background sequences. Half of the DNA motifs involved are matched to the existing motif databases, although nearly all of them are enriched with at least one gene ontology term. Their motif instances are also found statistically enriched on the promoter and enhancer regions. Especially, we introduce a novel measurement called motif pairing multiplicity which is defined as the number of motifs that are paired with a given motif on chromatin interactions. Interestingly, we observe that motif pairing multiplicity is linked to several characteristics such as regulatory region type, motif sequence degeneracy, DNase accessibility and pairing genomic distance. Taken into account together, we believe the coupling DNA motif pairs identified in this study can shed lights on the gene transcription mechanism under long-range chromatin interactions.

Availability and implementation: The identified motif pair data is compressed and available in the supplementary materials associated with this manuscript.

Contact: kc.w@cityu.edu.hk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

As one of the major gene regulation mechanisms, transcription factors (TF) can bind onto regulatory DNA elements in human and other eukaryotes. Different binding combinations of TFs may result in a gene being expressed in different tissues or at different developmental stages. To fully understand gene regulation, it is essential to identify the transcription factor binding sites (TFBSs) (Wong *et al.*, 2013). TFBS are relatively short (5–15 bp) and highly degenerate sequence motifs, which makes their effective identification a computationally challenging task. A number of high-throughput experimental and computational technologies were developed to determine TFBSs. Databases have been developed to store the TFBS (also known as DNA motif) data. Additional introduction can be found in supplementary materials.

In recent years, thanks to its relatively low costs, next generation sequencing technology is applied to thousands of genomes (Abecasis *et al.*, 2012). It creates unprecedented opportunities for understanding DNA motifs; for instance, the ChIP-seq data from the ENCODE consortium has enabled a systematic discovery and characterization of DNA motifs in human cell lines (Kheradpour and Kellis, 2014). Combining ChIP-seq and SELEX data, Jolma et al. have also characterized the DNA-binding specificity landscape of human TFs (Jolma *et al.*, 2013). On the other hand, the Hi-C technology has been developed and applied to reveal the three dimensional shapes of different cell lines by the chromosome conformation capture method (Belton *et al.*, 2012). In particular, there is increasing evidence that long-range chromatin interactions are related to gene co-expression (Babaei *et al.*, 2015) as well as protein–DNA interactions (Mifsud *et al.*, 2015). Therefore, it is essential to comprehensively identify the coupling DNA motif pairs on those long-range chromatin interactions for understanding gene transcription.

2 Methods

2.1 Collecting long-range chromatin interactions

We have collected the Hi-C chromatin interaction pairs in the human K562 cell line from Fit-Hi-C (Ay *et al.*, 2014). To accommodate the Hi-C data resolution, each interaction region mid-point is extended to 50 000 bp in both directions (Wong *et al.*, 2015). We set the *q*-value threshold to 0.1 to ensure that only highly confident chromatin interaction pairs are adopted.

2.2 Collecting long-range regulatory region pairs

We have collected the human genome (hg19) segmentation data for the human K562 cell line from ChromHMM and Segway (Hoffman *et al.*, 2012). To focus on gene regulation, we have limited our study to the regulatory regions designated by both ChromHMM and Segway (i.e. WE (Weak Enhancer), E (Enhancer), TSS (Promoter) and PF (Promoter-Flanking Region)). On the other hand, taking into account both regulatory region information and Hi-C chromatin interaction pairing information, we check their overlapping regions to create datasets of regulatory region pairs which are in close proximity to each other by long-range chromatin interactions. The regulatory region pairs with less than 30 nt on either side are discarded because the maximal DNA motif width is set to 25 nt (Kheradpour and Kellis, 2014). The resultant 74 552 long-range regulatory region pairs are listed in supplementary data which are summarized in Supplementary Figure S1. On Supplementary Figure S1, it can be observed that the number of E–WE (Enhancer–Weak Enhancer) pairs is the highest, followed by E–E (Enhancer–Enhancer) pairs and TSS–E (Promoter–Enhancer) pairs. The latter two pair types are expected to be of high frequency except the top one (E–WE), implying that E–WE pairs may play a previously uncharacterized but important role in long-range gene regulation.

2.3 De novo motif discovery

Based on the long-range regulatory region pairs collected, we have retrieved their corresponding sequences on which DNA motifs can be discovered. We have run the motif discovery program (MEME (Bailey and Elkan, 1994)) on the sequences of each side of each type of long-range regulatory pairs, resulting in 19 491 DNA motif pairs which are composed of 4227 DNA motifs discovered by MEME (Bailey and Elkan, 1994). The DNA motif pair results are described and listed in supplementary data and summarized in Supplementary Figure S2.

2.4 Motif pair type enrichment

To quantify the region pair enrichment, assuming each motif is equally likely to be coupled with another motif on the same chromosome, we have applied the binomial test to calculate the *P*-values (P_{XYC}) for the enrichment of the occurrence counts of different motif pair types ($X - Y$) on each chromosome (C). The mathematical calculations can be found in supplementary materials. In summary, the resultant enrichment *P*-values are visualized in Supplementary Figure S3. It can be observed that the DNA motif pairs on the promoter and enhancer region pairs (TSS–TSS, TSS–E and E–E) are significantly enriched and identified than the other DNA motif pairs. On the other hand, the DNA motif pairs on the weak enhancer region pairs are also slightly enriched if the other partner region is either promoter or enhancer (TSS–WE and E–WE). In contrast, the depletion has also been visualized in Supplementary Figure S4. It can be observed that the motif pairs involving promoter-flanking (PF) are highly depleted. On the other hand, it is also intriguing for us to look at the motif pairs under different genomic distance conditions. Therefore, we have further divided each motif pair type to different genomic distance subclasses based on their genomic distances. Specifically, for each chromosome, we observed that the empirical genomic distance distribution for all the motif pairs follows the normal distribution. Therefore, we define the following descriptive genomic distance terms 'near', 'mean' and 'far' for the motif pairs with genomic distances under the following intervals $(-\infty, \mu - \sigma), [\mu - \sigma, \mu + \sigma], (\mu + \sigma, +\infty)$. The results are visualized in Supplementary Figure S5. The exact binomial test calculation procedure can be found in supplementary materials. In summary, it can be observed that different chromosomes have different enrichment of motif pair types with different genomic distances.

3 Results

3.1 Matching to annotated DNA motifs using TomTom

After we have found the DNA motifs on those long-range chromatin interaction regions, we are especially interested in how those motifs are overlapped with the existing DNA motif data. Therefore, we have run TomTom (Gupta *et al.*, 2007) to perform database search with the DNA motifs (under the default parameter setting of TomTom). The results are visualized in Supplementary Figure S5. It can be observed that nearly half of the identified DNA motifs can be matched to existing DNA motif data.

3.2 Gene ontology enrichment using GOMO

On the other hand, the gene ontology enrichment of those identified DNA motifs are also important for us to understand their functional roles. Therefore, we have run GOMO (with its default setting) to calculate the gene ontology enrichment for each of those DNA motifs (Buske *et al.*, 2010). Briefly, GOMO scans all the human promoters using each DNA motif and determine if any of them is significantly associated with genes linked to one or more Gene Ontology (GO) terms which are valuable to our understanding on the functions of each DNA motif. The results are visualized in Supplementary Figure S7. In general, it can be observed that over 90% of the identified DNA motifs are enriched with at least one GO term (i.e. GO-enriched on the horizontal axis) by GOMO (details can be found in supplementary materials). In particular, we observe the DNA-motif-related GO terms among the top frequent terms (listed in supplementary materials) such as (GO:0010467 gene expression), (GO:0006139 nucleobase-containing compound metabolic process), (GO:003700 sequence-specific DNA binding transcription factor activity) and (GO:0043565 sequence-specific DNA binding). In addition, it may be interesting for us to study the overlapping functional roles between the two motifs coupled within each motif pair. Therefore, we have computed the overlap coefficient (or, Szymkiewicz-Simpson coefficient) between the enriched GO terms of the first motif and those of the second motif within each motif pair. The overlap coefficient results are visualized in Supplementary Figure S8. We can observe that their overlap coefficients are substantially higher than the expected one. In particular, the enriched GO terms of those DNA motifs on the promoter-flanking (PF) regions are very similar to those on the promoter-related (TSS and PF) regions coupled, comparing to the other motif pair type. It reflects that some functional consistency is expected if a DNA motif pair belongs to the two promoter types.

3.3 Motif pairing multiplicity

Since coupling DNA motif pairs are identified in this study, another interesting aspect is on the pairing multiplicity of those DNA motifs (i.e. the number of motifs coupled with a given motif on long-range regulatory pairs). Therefore, we have plotted the motif pairing multiplicity against different region types as depicted in Figure 1. Interestingly, we found that the motif pairing multiplicity distribution for the DNA motifs on the promoter (TSS) and enhancer (E) regions are quite different from those on the other (PF and WE) regions, reflecting the long-range coupling motif pairing complexity involved in the gene transcription initiation process. In addition, we observe that the multiplicity distribution of the DNA motifs identified on the weak enhancer (WE) regions are heavily tailed toward zero. It may indicate that those DNA motifs are rather singular than the others which are also reflected from the functional consistency within pairs in Supplementary Figure S8.

On motif pair coupling level, we may also wonder if there is any relationship between the first motif multiplicity and the second motif multiplicity. In other words, if a motif is coupled with many motifs, how likely its coupled motifs are also coupled with many other motifs. Interestingly, it can be observed from Supplementary Figure S9 that, given a motif pair, if the two motifs belong to the same type (i.e. E–E, TSS–TSS and WE–WE), their multiplicities are positively correlated to each other except PF-PF which does not have sufficient samples to be concluded with its confidence interval. We also observe that, given a motif pair, if the two motifs belong to one of those three classes (i.e. TSS–E, TSS–WE and E–WE), the positive correlation phenomenon still holds, although it appears not to be as strong as the previous cases.

3.4 Sequence pattern degeneracy

It is well known that DNA motif sequence degeneracy varies across different DNA motifs from different species (Jolma *et al.*, 2013) and different biotechnologies (Orenstein and Shamir, 2014). Therefore, it is intuitive for us to study the sequence degeneracy of the DNA motifs identified in this study. In particular, we have adopted the *de facto* measure, Shannon information entropy (Schug *et al.*, 2005), to measure the sequence degeneracy for each DNA motif matrix column. Since motif sequence degeneracy is linked to the protein–DNA binding dynamics (Stormo, 2000), it is essential to decipher its relationship with the others. In this study, we found an interesting relationship between motif sequence degeneracy and motif multiplicity as depicted in Figure 2. We observe that the motif sequence degeneracy gets heavily tailed toward the conserved direction as motif multiplicity is increased from left to right. One of the possible explanations is that a DNA motif has to be solid in its sequence pattern to play its functional roles, increasing its motif multiplicity.

3.5 DNase hypersensitivity

In addition to sequence pattern degeneracy, another important consideration is the spatial accessibility of those DNA motifs. Therefore, we have collected the DNase ChIP-seq peak-calling data in the K562 cell line (wgEncodeUWDukeDnaseK562.fdr01peaks.hg19.bb) from the ENCODE consortium (Consortium, 2012). For each DNA motif, we count how many of its motif instances overlap with those DNase hypersensitive sites (peaks) on the reference human genome (hg19), resulting in a measurement called 'DNase Peak Fraction' in this study. The overlapping results are visualized in Supplementary Figure S10. To quantify the statistical significance on the overlapping fractions, for each DNA motif instance, we randomly sample 100 positions of the same motif width on the regulatory regions and the entire regions on the same chromosome to compute for the regulatory region background DNase peak fractions and overall background DNase peak fractions for each chromosome respectively (denoted as regulatory region background (R) and background (BG) on Supplementary Fig. S10). We have also computed *t*-tests and Mann–Whitney tests to quantify the statistical significance of the difference between the discovered motifs and those background regions as listed in supplementary materials. We observe that all of the *P*-values are less than 0.01. It supports our observation that the DNA motifs identified are substantially overlapping with the DNase hypersensitive sites (peaks), comparing to the overall background (BG), although it is less pronounced for the regulatory background (R).

Another interesting phenomenon we observed is that the DNase peak fraction is positively correlated to the number of enriched GO terms (i.e. GOMO hits). It is depicted in Supplementary Figure S11. A possible explanation is that a DNA motif has to be spatially accessible to DNA-binding proteins to carry out its gene regulation

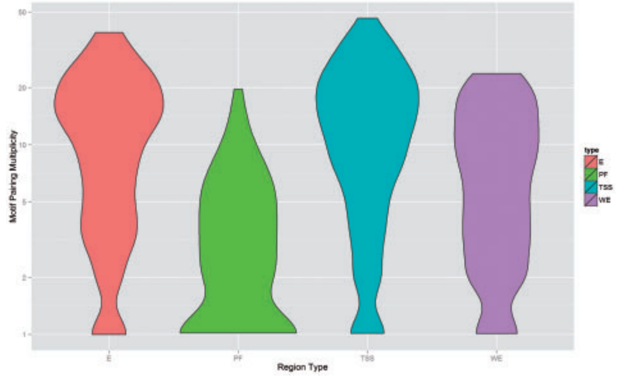

Fig. 1. Violin Plot on the motif pairing multiplicity of the DNA motifs sorted by region type (horizontal axis)

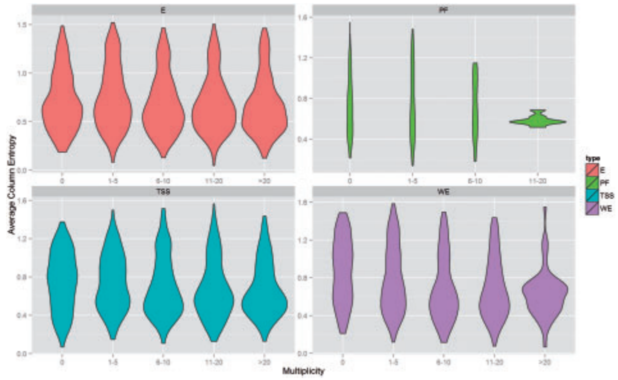

Fig. 2. Violin plots on the average column entropy of the DNA motifs found on different region types. The sub-figures are sorted by region type

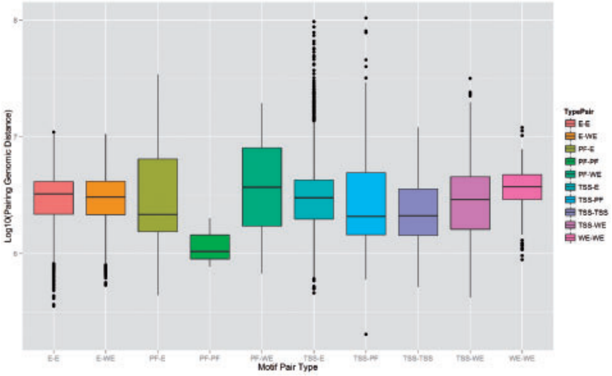

Fig. 3. Boxplot on the average genomic distances between the motif instances of the first motif and those of the second motif within each motif pair sorted by type (horizontal axis)

functions (i.e. open chromatin). Similar observation can be made from the motif pairing multiplicity with DNase peak fraction in Supplementary Figure S12, in which motif pairing multiplicity is negatively correlated to DNase peak fraction. It suggests that, if a DNA motif is highly coupled with other DNA motifs, its spatial accessibility may be suppressed in cells (i.e. closed chromatin).

3.6 Motif pairing distance

Since those identified motif pairs are found associative to each other on chromatin interaction regions, the actual genomic distances between their motif instances are worth for investigation further. In particular, it is generally believed that some enhancers can act in a very long range to regulate promoters, and thus gene transcription (Carter *et al.*, 2002). Therefore, we have grouped the motif pairs by types and plotted their genomic distances on Figure 3. Interestingly, we found that the motif pairs involving enhancer (E) or weak enhancer (WE) regions are more segregated than the others. It is consistent with the general belief that the regulatory components on enhancer regions are generally far from their acting partners by large genomic distances.

Next, we investigate how the motif pairing distances of the coupling DNA motif pairs may influence themselves. Interestingly, we observe that the more distant a coupled motif pair is separated in genomic distance, the lower its two DNA motifs' multiplicities are, as shown in Supplementary Figure S13. Such phenomenon is especially visible with the motif pairs found on promoter–enhancer (TSS–E) regions which are the core components for gene transcription.

4 Discussion

The DNA motifs on gene regulatory regions are believed to play central roles in gene transcription. In particular, their spatial organization can determine their functional roles in gene transcription.

Therefore, we have identified thousands of coupling DNA motif pairs on the human gene regulatory regions under long-range chromatin interactions. Those motif pairs have been characterized at different levels: region type enrichment, existing motif annotation matching, gene ontology enrichment and consistency, motif pairing multiplicity and transitivity, sequence pattern degeneracy, DNase accessibility and pairing genomic distance. Different relationships have been found between those attributes. In particular, we observe that motif pairing multiplicity is an important attribute which is correlated to some of the aforementioned attributes. In summary, this study provides a valuable resource for coupling DNA motif pairing information in human.

In the future, similar coupling DNA motif pairs in other cell lines from different species can be identified for comparative studies if the genome segmentation and Hi-C data availability issues can be alleviated. The data integration with ChIA-PET data is another promising direction (Li *et al.*, 2010), although each ChIA-PET run is limited to the DNA motifs of a single DNA-binding protein. The DNA motif pairing information in this study could be adopted for integrative analysis with the existing ChIP-seq studies, improving our understandings on protein–DNA binding. Finding the coupling sets of 3 or more DNA motifs is another interesting future direction.

Acknowledgements

The authors thank the three anonymous reviewers for their constructive reviews. The authors also thank the ENCODE consortium and Noble Research Lab for their public data release.

Funding

The work described in this article was substantially supported by a grant from City University of Hong Kong (Project No 7200444/CS).

Conflict of Interest: none declared.

References

Abecasis,G.R. *et al.* (2012) An integrated map of genetic variation from 1 092 human genomes. *Nature*, **491**, 56–65.

Ay,F. *et al.* (2014) Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome Res.*, **24**, 999–1011.

Babaei,S. *et al.* (2015) Hi-C chromatin interaction networks predict co-expression in the mouse cortex. *PLoS Comput. Biol.*, **11**, e1004221.

Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In: *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, pp. 28–36.*

Belton,J.M. *et al.* (2012) Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods*, **58**, 268–276.

Buske,F.A. *et al.* (2010) Assigning roles to DNA regulatory motifs using comparative genomics. *Bioinformatics*, **26**, 860–866.

Carter,D. *et al.* (2002) Long-range chromatin regulatory interactions in vivo. *Nat. Genet.*, **32**, 623–626.

Consortium,E. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.

Gupta,S. *et al.* (2007) Quantifying similarity between motifs. *Genome Biol.*, **8**, R24.

Hoffman,M.M. *et al.* (2012) Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods*, **9**, 473–476.

Jolma,A. *et al.* (2013) DNA-binding specificities of human transcription factors. *Cell*, **152**, 327–339.

Kheradpour,P. and Kellis,M. (2014) Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res.*, **42**, 2976–2987.

Li,G. *et al.* (2010) ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. *Genome Biol.*, **11**, R22.

Mifsud,B. *et al.* (2015) Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.*, **47**, 598–606.

Orenstein,Y. and Shamir,R. (2014) A comparative analysis of transcription factor binding models learned from PBM, HT-SELEX and ChIP data. *Nucleic Acids Res.*, **42**, e63.

Schug,J. *et al.* (2005) Promoter features related to tissue specificity as measured by Shannon entropy. *Genome Biol.*, **6**, R33.

Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.

Wong,K.C. *et al.* (2013) DNA motif elucidation using belief propagation. *Nucleic Acids Res.*, **41**, e153.

Wong,K.C. *et al.* (2015) SignalSpider: probabilistic pattern discovery on multiple normalized ChIP-Seq signal profiles. *Bioinformatics*, **31**, 17–24.