Maps of context-dependent putative regulatory regions and genomic signal interactions

Klev Diamanti1,†, Husen M. Umer1,†, Marcin Kruczyk1, Michał J. Dąbrowski2, Marco Cavalli3, Claes Wadelius3 and Jan Komorowski1, 2,*

1 Department of Cell and Molecular Biology, Uppsala University, Uppsala SE-751-24, Sweden
2 Institute of Computer Science, Polish Academy of Sciences, Warsaw, 012-48, Poland
3 Science for Life Laboratory, Department of Immunology, Genetics and Pathology, Uppsala University, Uppsala SE-751-08, Sweden

* To whom correspondence should be addressed. Tel: +46 (0) 18 471 66 92; Email: jan.komorowski@icm.uu.se
† The authors wish it to be known that, in their opinion, the first 2 authors should be regarded as joint First Authors

Supplementary Material
Supplementary Notes

Supplementary Note S1: Regulatory region detection with tfNet

The algorithm

Our region detection method is based on the observation that genomic loci accumulating various TF signals are more likely to play a role in the genomic regulatory mechanism. Hence, genomic groups of signals (peaks) can be compiled into clusters considering their pairwise point-sources (summits) distance. For the construction of putative regulatory regions a threshold distance and a collection of genomic-signal in UCSC BED format should be provided.

All the operations of the algorithm are performed in parallel for each chromosome (one thread per chromosome). If chromosome\(_A\) is the chromosome with the largest number of TFBS peaks, then the algorithm takes:

- Worst case:
  \[
  |chromosome|^2 + |peaks_{chromosome_1}|^2 + |peaks_{chromosome_2}|^2 + \ldots + |peaks_{chromosome_N}|^2 = O\left(|peaks_{chromosome_A}|^2\right)
  \]

- Average case:
  \[
  |chromosome| \times \log|chromosome| + |peaks_{chromosome_1}| \times \log|peaks_{chromosome_1}| + |peaks_{chromosome_2}| \times \log|peaks_{chromosome_2}| + \ldots + |peaks_{chromosome_N}| \times \log|peaks_{chromosome_N}|
  \]
  \[
  = O\left(|peaks_{chromosome_A}| \times \log|peaks_{chromosome_A}|\right)
  \]
  time to sort the input list of peaks. As for the region detection process it needs \(3 \times |peaks_{chromosome_A}| = O\left(|peaks_{chromosome_A}|\right)\) time.

As shown before the overall complexity of the algorithm depends on the complexity of the sorting algorithm (QuickSort). Hence the time complexity of the algorithm will be

```
1  L ← sort(peaks)
2  d_m
3  region, currPeak ← L[first]
4  while (nextPeak ← L[next])
5      region ← currPeak
6      if ((nextSummit − currSummit) ≤ d_m)
7          currPeak ← nextPeak
8      else
9          close the region
10         start a new region
11         currPeak ← nextPeak
12     end if
13 end while
```

Pseudocode of the proposed algorithm. The input is a list of peaks L and the maximum distance between consecutive peak summits \(d_m\). There are 3 main variables in the algorithm: region – the region that is currently constructed and points to a list of TF peaks; currPeak – the currently examined peak signal; nextPeak – the next peak in the list.
\(O\left(\left| \text{peaks}_{\text{chromosome}} \right|^2 \right)\) and \(O\left(\left| \text{peaks}_{\text{chromosome}} \right| \times \log \left| \text{peaks}_{\text{chromosome}} \right| \right)\), in the worst case and in the average case respectively.

### Implementation and functionalities

In order to validate the reliability of the algorithm we implemented a cross-platform fast parallel computational tool, tfNet executables and user manual: [http://figshare.com/articles/tfNet_manual/1408532](http://figshare.com/articles/tfNet_manual/1408532) (1) (Supplementary Figure S1). The key functionalities of the tool include strand-specific region detection, on-the-fly region filtering, statistically significant interactions of the genomic-signal pairs, output in the UCSC BED format and in the information-rich XML format. The tool is provided as a single executable file, the results can be visualized in genome browsers and the genomic signals interactions are modeled with weighted networks.

### Construction of a null model of regions

We created artificial datasets by applying the shuffleBed functionality of BEDTools v2.21(2) to a synthetic TF dataset constructed using the mean peak length (335bp) and the mean peak number (14633 peaks) of the full analysis dataset. We constructed collections of TFBS datasets varying from 1 to 1000 and we recursively ran our tfNet pipeline while observing the properties of the generated regulatory regions (quantity, genome coverage and time efficiency of the tool) (Supplementary Figure S4 and S19). For each run we computed the running time of tfNet on a 12 core Intel(R) Xeon(R) CPU X5650 @ 2.67GHz CentOS7 machine.

### Supplementary Note S2: Data processing

#### Mus musculus data processing

We downloaded transcription factor binding site data for Mus musculus from ENCODE (Supplementary Table S1) and renamed the downloaded files according to TF data they contained. We divided the TFBS files into 5 groups based on their cell line (C2C12, CH12.LX, ES-E14, MEL, myocyte). Finally, we merged TFBSs files originating from different replicates of the same TF and run the tfNet tool on the data (Methods - Identification of the regulatory landscape). The input data statistics and the number of putative regulatory regions calculated are presented in (Supplementary Table S7; Supplementary Figure S7A). The TF interactions networks for each of the five cell lines are shown in (Supplementary Figure S8).

#### Drosophila melanogaster data processing

We downloaded all the available TFBSs for drosophila melanogaster from modENCODE, Berkeley Drosophila Transcription Network Project (BDTNP) and Literature (5-8). We also downloaded all the available cis-regulatory elements (CRMs) from the redFly database (9) and the DNaseI hypersensitive sites from modENCODE (10). All data sources, the corresponding URLs and the database download criteria are listed in (Supplementary Table S1). First we converted any gff3 files to BED format. Secondly, we merged TFBSs files originating from different replicates and different databases of the same TF, and run the tfNet tool on the data (Methods - Identification of the regulatory landscape). We named this analysis as cumulative. The input data statistics and the number of putative regulatory regions calculated are presented in (Supplementary Table S7; Supplementary Figure S7B). The TF interactions networks for the cumulative dataset are shown in (Supplementary Figure S10).

Next, we divided the data into developmental stages and selected those containing a sufficient number of TFs (Embryos 0-8, Embryos 0-12, Embryos 8-16 and Embryos 16-24). Then we merged TFBSs files originating from different replicates and different databases of the same TF, and run the tfNet tool on the data (Methods - Identification of the regulatory landscape). The input data statistics and the number of putative regulatory regions calculated are presented in
Colorectal cancer data processing

We downloaded all the high throughput (HT) ChIP-seq data generated by (11) (GSE49402). We used only the TF ChIP-seq peaks that had passed the quality control pipeline developed by (11). In total we were left with 117 TFs. On average there were 13628 peaks per TF. We used the ENCODE liftOver tool to convert the genomic coordinates from hg18 to hg19 and then we run tfNet (Methods - Identification of the regulatory landscape).

Prostate cancer data processing

We downloaded ChIP-seq raw data for multiple prostate cancer cell lines from (12) (Supplementary Table S6). We used only TF ChIP-seq experiments that were performed under normal conditions. We aligned the reads against hg19 with bwa (13) and we ran the ENCODE ChIP-seq peak calling pipeline based on Kundaje’s Lab implementation (ENCODE3 pipeline v1 - https://github.com/kundajelab/TF_chipseq_pipeline) with MACS2 (14). We used the overlapped peaks for each replicate since the Irreproducibility Discovery Rate (IDR) framework (15) could not be applied to the majority of the experiments due to lack of controls or lack of exactly 2 replicates per TF. We pooled TF replicates originating from different prostate cancer cell lines. We merged the overlapping TFBSs originating from different replicates or studies of the same TF into single peaks in order to avoid artefacts and misleading TF interactions (multiple peaks of the same TF originating from different replicates binding the same genomic loci) using the function mergeBed of bedtools (2). Next we run tfNet (Methods - Identification of the regulatory landscape).

Supplementary Note S3: Experimental Validations

Selection and validation of regulatory regions

The following steps were taken in order to select putative regulatory regions for experimental validation: (a) we created a list of keywords associated with various liver diseases (Supplementary Table S9). (b) we listed all SNPs from genome-wide association study catalogue (GWAS)(3) to find SNPs in linkage disequilibrium \((r^2=0.8)\) with the GWAS SNPs. (d) we selected putative regulatory regions from the pools of filtered regions that contained TFBSs with GWAS SNPs or SNPs in LD with GWAS SNPs. We found 17 putative regulatory regions that contained SNPs related to hepatocellular diseases. To obtain preliminary results of biological validation we selected 8 putative regulatory regions starting with the shortest ones. We adopted a hypothesis that the presence of putative regulatory region should significantly increase the signal of luciferase assay (protocol described in detail below).

Construction of cloning plasmids and luciferase report assays

The luciferase expression constructs were built based on pGL4.23 (Promega). The ccdB expression cassette was inserted into KpnI and EcoRV sites of pGL4.23 to construct pGL4.23-ccdB, which was used as a basal vector to diminish false positive signal during the cloning process. The designed primers (Sigma-Aldrich) amplified the putative regulatory region using Phusion Hot Start Flex DNA polymerase (NEB) and HepG2 genomic DNA as a template. The PCR product was separated on an agarose gel and followed by isolation of a specific gel band (QIAquick Gel Extraction Kit from QIAGEN). The purified amplification product was inserted upstream of the minimal promoter sequence of pGL4.23 by SLiCE cloning methods. In transformation step we used Library Efficiency™ DH5α™ Competent Cells (life technologies) and followed the manufacturer’s transformation procedure with minor changes. To annotate a SNP variant with a sequence of a putative regulatory region, multiple individual clones were picked.
up and subjected to Sanger sequencing. HepG2 cells were transfected using X-tremeGENE HP DNA Transfection Reagent (Roche), one day after plating with approximately 90% confluence in 96-well plate. Each well was transfected with 100 ng of firefly luciferase reporter vector harboring respective putative regulatory region together with 1 ng of renilla luciferase reporter vector pGL4.74, which was used to normalize the transfection and lysis efficiency. Twenty-four hours after transfection, the cells were harvested and lysed in 1X passive lysis buffer (Promega) on a rocking platform for 45 min at room temperature. Luminescence of firefly and luciferase activity were measured by Dual-Luciferase® Reporter (DLR™) Assay System (Promega) on an Infinite® M200 pro reader (TECAN) following the manufacturer’s protocol. Simultaneously, we fulfilled the same procedure for an empty vector control. The luminescence measurements came from six replicate wells. The luciferase normalized ratios of experimental samples (Supplementary Table S5) were compared to empty vector using non-parametric Mann-Whitney test.

**Supplementary Note S4: Gene expression level comparison**

We compared the gene expression levels between promoters of genes annotated as heterochromatic, insulators and promoters from ChromHMM for 3 selected cell lines (GM12878, HepG2 and K562). The gene set together with the average reads per kilobase per million mapped reads (RPKM) was obtained from the Gene expression matrix over ENCODE2 in GENCODEv19 (Supplementary Table S1) and it constituted of 48807 genes. To obtain the gene coordinates we employed the GENCODEv23 gene annotations mapped to hg19 coordinates (Methods - Regulatory region annotation). Then we intersected the promoters ±1.5kb with putative regulatory regions of heterochromatic, insulator or promoter annotation for all three cell lines while maintaining the RPKM1 value information. In order to calculate the statistical difference between gene expression levels (p-value) for the annotations we used Wilcoxon rank-sum test.
Equations

\[ A \quad P(X = k) = \frac{\binom{K}{k} \times \binom{N - K}{n-k}}{\binom{N}{n}} \]

\[ B \quad P(X = k) = \binom{n}{k} \times p^k \times (1 - p)^{n-k} \]

Equation S1: Hypergeometric and binomial distribution applied for the p-Value calculation of the TF interaction networks. (a) \( N \) is the total number of TF pairs, \( K \) is the total number of occurring pairs, \( n \) is the sum of the TFs participating in a specific pair and \( k \) is the number of successful pairs. (b) \( n \) is the number of all possible TF pairs, \( k \) is the number of successful pairs and \( p \) is the probability of a specific pair to occur by chance.
Supplementary Figures

Figure S1: Graphical representation for the detection of the putative regulatory regions by tfNet according to Supplementary Note S1.
Figure S2: Distance between consecutive peak summits. The peaks are sorted based on their chromosome and their summit coordinate. All the pairwise distances between peaks belonging to the same chromosome are considered. The label above the percentage represents the distance between consecutive summits while the percentage represents the number of peaks having that distance.
Figure S3: The number of putative regulatory regions including DNaseI (Methods – Identification of the regulatory landscape) for 5 different human cell lines (green). Next we interested the putative regulatory region datasets with DNaseI (orange) and ChromHMM (blue). For the cell lines where ChromHMM was not available the corresponding bar is missing (HeLaS3).
Figure S4: Results of region construction on synthetic datasets (Supplementary Note S1). The dots in the plot represent the real data obtained from the experimental ENCODE datasets. The color represents the different cell lines. The curve represents A) the number of constructed regions, and B) The total genome coverage, in a range of 0 to 1000 TF datasets.
Figure S5: Heatmap networks modelling the significant TF-TF interactions in putative regulatory regions of LoVo cell line for colorectal cancer. The colour intensity in each cell represents the TF-TF interaction significance for each network type.
Figure S6: Heatmap networks modelling the significant TF-TF interactions in putative regulatory regions generated from curated ChIP-seq datasets in multiple prostate cancer cell lines. The colour intensity in each cell represents the TF-TF interaction significance for each network type.
Figure S7: The number of putative regulatory regions generated by tfNet for M. musculus (A) and D. melanogaster (B) (Methods - Identification of the regulatory landscape). (A) Each bar represents the number of regulatory regions obtained for 5 different M. musculus cell lines (Supplementary Note S2). (B) Each bar represents the number of regulatory regions obtained for four different developmental stages and for the cumulative dataset for D. melanogaster. (C) The stacked bar represents the number of known cis regulatory modules (CRMs) obtained from the redFly database (Supplementary Table S1) for D. melanogaster that overlap with cumulative putative regulatory regions detected by tfNet (red) and the ones that do not overlap (blue).
Figure S8: TF interaction networks for M. musculus generated by tfNet (Methods - Identification of the regulatory landscape). We used TFBS data for M. musculus downloaded from ENCODE and we divided them according to their cell lines (headers of the rows). The data pre-processing procedure is described in detail in the section “Mus musculus data processing” above. The columns of the figure represent the type of TF interaction network (Methods – TF complexes detection). The co-occurring network for C2C12 is not available since there was no statistically significant TF-interaction detected.
Figure S9: TF interaction networks for four different developmental stages of D. melanogaster generated by tfNet (Methods - Identification of the regulatory landscape). We used TFBS data for D. melanogaster downloaded from modENCODE, BDTNP and literature sources (Supplementary Note S2). The columns of the figure represent the type of TF interaction network (Methods – TF complexes detection) and the rows the developmental stage.
cooccurring

neighboring

overlapping
Figure S10: TF interaction networks for the cumulative dataset of *D. melanogaster* generated by tfNet (Methods - Identification of the regulatory landscape). We used TFBS data for *D. melanogaster* downloaded from modENCODE, BDTNP and literature sources (Supplementary Note S2). The columns of the figure represent the type of TF interaction network (Methods – TF complexes detection).
Figure S11: Heatmap network modeling the interactions of TFs co-occurring in heterochromatin (Methods - TF complexes detection). Each cell contains 4 signals originating from each of the 4 cell lines as shown in the legend on the bottom (Cell lines). All TFs in the network are present in at least 2 cell lines. The red color intensity is calculated as $-\log_{10} p$ where $p$ is the Bonferroni corrected $p$-value of the specific TF interaction.
Figure S12: Heatmap networks modelling the TF interactions with statistically significant interactions in at least 2 cell lines in putative regulatory regions of heterochromatin annotation. Each cell contains 4 signals (rows) originating from each of the 4 cell lines as shown in the legend bottom-left. The red colour intensity is calculated as where p is the calculated p-value of the specific TF interaction (cf. Methods). A) Interactions defined as neighbouring (TF summits in a distance of 20-60bp). B) Interactions defined as overlapping (TF summits in a distance of ≤20bp).
Figure S13: Heatmap network modeling the interactions of TFs co-occurring in heterochromatin (Methods - TF complexes detection). Each cell contains 4 signals originating from each of the 4 cell lines as shown in the legend on the bottom (Cell lines). All TFs in the network are present in at least 1 cell line. The red color intensity is calculated as $-\log_{10} p$ where $p$ is the Bonferroni corrected p-value of the specific TF interaction.
Figure S14: Heatmap network modeling the interactions of TFs neighboring in heterochromatin (Methods - TF complexes detection). Each cell contains 4 signals originating from each of the 4 cell lines as shown in the legend on the bottom (Cell lines). All TFs in the network are present in at least 1 cell line. The red color intensity is calculated as $-\log_{10} p$ where $p$ is the Bonferroni corrected p-value of the specific TF interaction.
Figure S15: Heatmap network modeling the interactions of TFs overlapping in heterochromatin (Methods - TF complexes detection). Each cell contains 4 signals originating from each of the 4 cell lines as shown in the legend on the bottom (Cell lines). All TFs in the network are present in at least 1 cell line. The red color intensity is calculated as $-\log_{10} p$ where $p$ is the Bonferroni corrected p-value of the specific TF interaction.
Figure S16: Number of pairs of interacting domains from the study (17) that interact with regulatory regions as predicted in our study. In pink are the interacting domains where both the upstream and the downstream domains intersect with putative regulatory regions. In green are the pairs harboring putative regulatory regions in either the upstream or the downstream domains. In blue there are the pairs of regulatory domains that did not intersect with any of our predicted regulatory regions.
Figure S17: Participation of the putative regulatory regions (including DNaseI), as calculated by tfNet (Methods - Identification of the regulatory landscape), in genomic loops as defined in (17) (Putative regulatory regions marked in yellow are those intersecting with the interacting domains while those marked as “out” (red) are those not intersecting with any of the proposed loops or the interacting domains. A) ChromHMM annotated cell lines (GM12878 and K562). B) ChromHMM unannotated cell line (HeLaS3).
Figure S18: Participation of the putative regulatory regions (including DNaseI) in interacting domains of HeLaS3 as defined by (17, 18). Region annotations are shown outside the circles. The percentages represent the whole range of putative regulatory regions (some putative regulatory regions have been counted multiple times since we have assumed a complete graph for interacting domains harboring more than one pair of putative regulatory regions there). The numbers between the inner and the outer circle represent putative regulatory regions of a specific annotation interacting with other regions (duplicates are included). The number colored in pink stands for the real number of putative regulatory regions as detected by tfNet (duplicates excluded). The color code for the putative regulatory regions annotation remains the same as in Figure 1A in the main text. The thickness of the ribbon shows the number of interacting regions of each annotation. The arks of the thin innermost circle denote the origin of the ribbon.
Figure S19: Time needed for tfNet to run on synthetic datasets (Supplementary Note S1).
References
1. Diamanti K, Umer HM, Kruczyk M, Dabrowski MJ, Cavalli M, Wadelius C, et al. tfNet. figshare: figshare; 2015.
2. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010 Mar 15;26(6):841-2. PubMed PMID: 20110278. PubMed Central PMCID: 2832824.
3. Beck T, Hastings RK, Gollapudi S, Free RC, Brookes AJ. GWAS Central: a comprehensive resource for the comparison and interrogation of genome-wide association studies. European journal of human genetics : EJHG. 2014 Jul;22(7):949-52. PubMed PMID: 24301061. PubMed Central PMCID: 4060122.
4. Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, de Bakker PI. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics. 2008 Dec 15;24(24):2938-9. PubMed PMID: 18974171. PubMed Central PMCID: 2720775.
5. Holmqvist PH, Boija A, Philip P, Crona F, Stenberg P, Mannervik M. Preferential genome targeting of the CBP co-activator by Rel and Smad proteins in early Drosophila melanogaster embryos. PLoS genetics. 2012;8(6):e1002769. PubMed Central PMCID: 3380834.
6. Perez-Lluch S, Blanco E, Carbonell A, Raha D, Snyder M, Serras F, et al. Genome-wide chromatin occupancy analysis reveals a role for ASH2 in transcriptional pausing. Nucleic acids research. 2011 Jun;39(11):4628-39. PubMed PMID: 21310711. PubMed Central PMCID: 3113561.
7. Harrison MM, Li XY, Kaplan T, Botchan MR, Eisen MB. Zelda binding in the early Drosophila melanogaster embryo marks regions subsequently activated at the maternal-to-zygotic transition. PLoS genetics. 2011 Oct;7(10):e1002266. PubMed PMID: 22028662. PubMed Central PMCID: 3197655.
8. Herzig HM, Mohan M, Garrett AS, Miller C, Casto D, Zhang Y, et al. Polycomb repressive complex 2-dependent and -independent functions of Jarid2 in transcriptional regulation in Drosophila. Molecular and cellular biology. 2012 May;32(9):1683-93. PubMed PMID: 22354997. PubMed Central PMCID: 3347239.
9. Gallo SM, Gerrard DT, Miner D, Simich M, Des Soye B, Bergman CM, et al. REDfly v3.0: toward a comprehensive database of transcriptional regulatory elements in Drosophila. Nucleic acids research. 2011 Jan;39(Database issue):D118-23. PubMed PMID: 20965965. PubMed Central PMCID: 3013816.
10. Krzywinski M, Schein J, Birol I, et al. Circos: an information aesthetic for comparative genomics. Genome research. 2009 Jul 15;20(14):1754-60. PubMed PMID: 19451168. PubMed Central PMCID: 2705234.
11. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014 Dec 18;159(7):1665-80. PubMed PMID: 25497547.
12. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. Genome research. 2009 Sep;19(9):1639-45. PubMed PMID: 19541911. PubMed Central PMCID: 2752132.