BMC Research Notes

DATA NOTE

Open Access

Genome-wide map of human and mouse transcription factor binding sites aggregated from ChIP-Seq data

Ilya E. Vorontsov[1], Alla D. Fedorova[1], Ivan S. Yevshin[2], Ruslan N. Sharipov[2,3,4], Fedor A. Kolpakov[2,3], Vsevolod J. Makeev[1,5,6] and Ivan V. Kulakovskiy[1,5,7*]

Abstract

Objectives: Mammalian genomics studies, especially those focusing on transcriptional regulation, require information on genomic locations of regulatory regions, particularly, transcription factor (TF) binding sites. There are plenty of published ChIP-Seq data on in vivo binding of transcription factors in different cell types and conditions. However, handling of thousands of separate data sets is often impractical and it is desirable to have a single global map of genomic regions potentially bound by a particular TF in any of studied cell types and conditions.

Data description: Here we report human and mouse cistromes, the maps of genomic regions that are routinely identified as TF binding sites, organized by TF. We provide cistromes for 349 mouse and 599 human TFs. Given a TF, its cistrome regions are supported by evidence from several ChIP-Seq experiments or several computational tools, and, as an optional filter, contain occurrences of sequence motifs recognized by the TF. Using the cistrome, we provide an annotation of TF binding sites in the vicinity of human and mouse transcription start sites. This information is useful for selecting potential gene targets of transcription factors and detecting co-regulated genes in differential gene expression data.

Keywords: Transcription factor binding sites, ChIP-Seq, Cistrome, Regulatory regions, Target genes, Human and mouse

Objective

Locations of genomic regions responsible for transcriptional regulation are valuable for many genomics and genetics studies, from analysis of gene regulatory networks to prediction of the functional impact of non-coding genetic variants. There are thousands of experimental data sets related to in vivo binding of human and mouse transcription factors (TFs), with chromatin immunoprecipitation followed by deep sequencing (ChIP-Seq) as the gold standard method. Many existing databases such as GTRD [1] or ReMAP [2] focus on systematic reprocessing and user-friendly access to published ChIP-Seq data.

By design, ChIP-Seq data provide information on cell-type specific binding. Binding profiles of the same TF can be quite different in different cell types or experimental conditions, and, for a particular transcription factor, it is not always feasible to separately analyze hundreds of individual data sets. Instead, a list of reproducible TF binding regions routinely identified as TF binding sites could be valuable for preliminary selection of putative target genes or for the discovery of key regulators for differentially expressed genes. For example, meta-clusters provided in GTRD are the genome segments bound by the studied TF across several data sets. For practical applications it is useful to have such constituent binding regions being separated into different reproducibility categories, and annotated with occurrences of transcription factor binding motifs, to highlight likely genuine binding sites of each particular TF. In particular, transcription

*Correspondence: ivan.kulakovskiy@gmail.com
[1] Vavilov Institute of General Genetics, Russian Academy of Sciences, GSP-1, Gubkina 3, Moscow, Russia 119991
Full list of author information is available at the end of the article

factor binding sites in the vicinity of the transcription start site are of special interest allowing to identify putative TF target genes. Finally, it would be convenient to have available genomic coordinates of a TF binding map for each of several commonly used genome releases.

Data description

Here we present the human and mouse cistromes [3], the genome-wide maps of regions bound by TFs, obtained through systematic analysis of ChIP-Seq data. The cistromes include data for 349 mouse and 599 human TFs. Cistromes provide an important information layer for detecting putative target genes of the corresponding TFs, for detecting regulators bound to known promoters and enhancers, and for intersection and enrichment analysis of various genomic features including regulatory sequence variants.

For each TF, the cistrome consists of sets of non-overlapping regions with assigned reliability categories. For convenience, we provide genome-wide (Table 1, Data set 1–4) and gene-centric (Table 1, Data set 5–8) maps for two major human (hg19, hg38) and mouse (mm9, mm10) genome assemblies [4, 5]. The genome-wide map contains global genomic coordinates of TF binding regions. The gene-centric map contains the relative locations of the nearest cistrome segments for each transcription start site (TSS).

Cistrome aggregation and motif annotation

The initial set of TF binding regions from ChIP-Seq (the ChIP-Seq peaks) was extracted from GTRD (release 17 April 2017). GTRD provided ChIP-Seq peak calls from four different peak calling software (see Data file 1 for details) executed with default parameters. Using the approach described in [6], some data sets were excluded as unreliable. Then we applied BEDTools 2.26.0 [7] to merge the overlapping intervals from different experiments and ChIP-Seq peak callers. The resulting regions were classified into four reliability categories in the following manner:

A (the highest reliability, experimental and technical reproducibility): this group contains cistrome regions consisting of overlapping peaks detected in at least two experimental data sets and by at least two peak calling tools, i.e. supported by at least two experiments and at least two peak callers;

B (high reliability, experimental reproducibility): regions supported by at least two experiments;

C (medium reliability, technical reproducibility): regions supported by at least two peak callers.

For segments of A, B, and C sub-cistromes, it is required that each segment overlaps at least one ChIP-Seq peak from a data set that was accompanied by the experimental control data. All other reproducible segments fall into D category (limited reliability). The technical details of the cistrome construction and overall statistics of the cistrome are provided in the Data file 1 (see Table 1).

All the cistrome categories were annotated by predictions of TF binding sites with HOCOMOCO v11 [6] sequence motifs to obtain a subset of regions with genuine binding sites recognized by a particular TF.

Table 1 Overview of data files/data sets

Label	Name of data file/data set	File types (file extension)	Data repository and identifier (DOI or accession number)
Data set 1	cistrome_hg19.zip	Archive file (.zip) containing genomic regions files (.bed)	Figshare (https://doi.org/10.6084/m9.figshare.7087697)
Data set 2	cistrome_hg38.zip	Archive file (.zip) containing genomic regions files (.bed)	Figshare (https://doi.org/10.6084/m9.figshare.7087697)
Data set 3	cistrome_mm9.zip	Archive file (.zip) containing genomic regions files (.bed)	Figshare (https://doi.org/10.6084/m9.figshare.7087697)
Data set 4	cistrome_mm10.zip	Archive file (.zip) containing genomic regions files (.bed)	Figshare (https://doi.org/10.6084/m9.figshare.7087697)
Data set 5	cistrome2genes_hg19.zip	Archive file (.zip) containing tab-separated files (.tsv)	Figshare (https://doi.org/10.6084/m9.figshare.7087697)
Data set 6	cistrome2genes_hg38.zip	Archive file (.zip) containing tab-separated files (.tsv)	Figshare (https://doi.org/10.6084/m9.figshare.7087697)
Data set 7	cistrome2genes_mm9.zip	Archive file (.zip) containing tab-separated files (.tsv)	Figshare (https://doi.org/10.6084/m9.figshare.7087697)
Data set 8	cistrome2genes_mm10.zip	Archive file (.zip) containing tab-separated files (.tsv)	Figshare (https://doi.org/10.6084/m9.figshare.7087697)
Data file 1	cistrome_overview.xlsx	MS Excel file (.xslx)	Figshare (https://doi.org/10.6084/m9.figshare.7087697)

Data for human hg38 and mouse mm10 genome assemblies was produced directly from GTRD peak calls, data for human hg19 and mouse mm9 assemblies was produced with liftOver (v353) [8].

A (the best constitutively bound sites) and joint ABC (the compromise) cistromes are the most informative. We used those along with the motif annotation to construct a gene-centric map of TF binding (Table 1, Data set 5–8) using the GENCODE [9] annotation (GTF, main annotation files) and PyRanges 0.0.13 [10]. For each TSS, the gene-centric map contains the absolute distance from a TSS to the nearest cistrome segment corresponding to binding of a particular TF.

Limitations

- The cistrome lacks metadata regarding cell types, antibodies or experimental conditions. For studies of particular genes or particular binding sites, the user is advised to address a detailed database, such as GTRD.
- The cistrome coverage and reliability heavily depends on a volume of experimental data available for a particular TF. In the presented map, many TFs have very sparse maps with only a few bound regions or only low-reliability cistrome categories, or with no cistrome regions at all.
- For many TFs it was not possible to perform the motif annotation due to absence of reliable information on binding sequence preferences, the corresponding entries are explicitly marked in the gene-centric map.

Abbreviations

TF: transcription factor; ChIP-Seq: chromatin immunoprecipitation followed by deep sequencing; TSS: transcription start site.

Authors' contributions

IVK, IEV and VJM designed the cistrome construction pipeline. ADF and IVK technically implemented the pipeline. ISY, RNS, and FAK performed GTRD data and metadata extraction. IVK, IEV and VJM wrote the manuscript. All authors read and approved the final manuscript.

Author details

[1] Vavilov Institute of General Genetics, Russian Academy of Sciences, GSP-1, Gubkina 3, Moscow, Russia 119991. [2] BIOSOFT.RU Ltd, Russkaya 41/1, Novosibirsk, Russia 630058. [3] Institute of Computational Technologies, Siberian Branch of the Russian Academy of Sciences, Akad. Rzhanova 6, Novosibirsk, Russia 630090. [4] Novosibirsk State University, Pirogova 2, Novosibirsk, Russia 630090. [5] Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, GSP-1, Vavilova 32, Moscow, Russia 119991. [6] Moscow Institute of Physics and Technology (State University), 9 Institutskiy per, Dolgoprudny, Russia 141700. [7] Institute of Mathematical Problems of Biology RAS-the Branch of Keldysh Institute of Applied Mathematics of Russian Academy of Sciences, Vitkevicha 1, Pushchino, Russia 142290.

Acknowledgements

We thank prof. Robert Kelsh for valuable comments.

Competing interests

The authors declare that they have no competing interests.

Availability of data materials

The human and mouse cistrome assembly and annotation datasets generated during the current study are available in the figshare repository: https://doi.org/10.6084/m9.figshare.7087697.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Funding

The human cistrome construction was supported by the Russian Science Foundation grant to I.V.K. [17-74-10188]. The mouse cistrome construction was supported by the Program of fundamental research for state academies for 2013–2020 years (No. 01201363825) and by Presidium of the Russian Academy of Sciences program in molecular and cellular biology. Development of motif annotation tools was supported by the Russian Science Foundation Grant 14-50-00060. I.E.V. was personally supported by the Skoltech Systems Biology Fellowship.

Publisher's Note

Received: 21 September 2018 Accepted: 16 October 2018
Published online: 23 October 2018

References

1. Yevshin I, Sharipov R, Valeev T, Kel A, Kolpakov F. GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments. Nucleic Acids Res. 2017;45:D61–7. https://doi.org/10.1093/nar/gkw951.
2. Chèneby J, Gheorghe M, Artufel M, Mathelier A, Ballester B. ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. Nucleic Acids Res. 2018;46:D267–75. https://doi.org/10.1093/nar/gkx1092.
3. Vorontsov IE, Fedorova AD, Yevshin IS, Sharipov RN, Kolpakov FA, Makeev VJ, Kulakovskiy IV. Human and mouse cistromes: genomic maps of putative cis-regulatory regions bound by transcription factors. figshare. 2018. https://doi.org/10.6084/m9.figshare.7087697.
4. The Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860–921.
5. Church DM, Schneider VA, Graves T, Auger K, Cunningham F, Bouk N, et al. Modernizing reference genome assemblies. PLoS Biol. 2011;9(7):e1001091. https://doi.org/10.1371/journal.pbio.1001091.
6. Kulakovskiy IV, Vorontsov IE, Yevshin IS, Sharipov RN, Fedorova AD, Rumynskiy EI, et al. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. Nucleic Acids Res. 2018;46:D252–9. https://doi.org/10.1093/nar/gkx1106.
7. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.
8. Hinrichs AS, Karolchik D, Baertsch R, Barber GP, Bejerano G, Clawson H, et al. The UCSC Genome Browser Database: update 2006. Nucleic Acids Res. 2006;34(suppl_1):D590–8.
9. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: the reference human genome annotation for the ENCODE project. Genome Res. 2012;22:1760–74.
10. Performant Pythonic GenomicRanges. https://github.com/endrebak/pyranges. Accessed 11 Sep 2018.