Cell Stem Cell
Letter

BindDB: An Integrated Database and Webtool Platform for "Reverse-ChIP" Epigenomic Analysis

Ilana Livyatan,[1,3] Yair Aaronson,[1,3] David Gokhman,[1] Ran Ashkenazi,[1] and Eran Meshorer[1,2,*]
[1]Department of Genetics, The Institute of Life Sciences
[2]The Edmond and Lily Center for Brain Sciences
The Hebrew University of Jerusalem, Edmond J. Safra Campus, Jerusalem 91904, Israel
[3]Co-first author
*Correspondence: meshorer@huji.ac.il
http://dx.doi.org/10.1016/j.stem.2015.11.015

The high-throughput revolution has brought an unprecedented amount of epigenomic data for embryonic stem cells (ESCs), including genome-wide profiles of chromatin-bound proteins and histone modifications generated by chromatin immunoprecipitation assays (ChIP-seq). As dataset after dataset of ChIP-seq data is added to the pool, the time is ripe to reverse the viewpoint from being factor-oriented to the perspective of genomic locations in order to offer a comprehensive view of chromatin characteristics and regulatory elements that govern coherent gene groups or chromosomal regions. Previously, we collected over 50 such genome-wide datasets in mouse (mESCs), and, using a bioinformatic pipeline which we developed, were able to identify novel regulators of histone genes (Gokhman et al., 2013), demonstrating the power of this approach. We have now built a dynamic database and webtool platform called BindDB that enables in silico "reverse-ChIP" analyses for widespread use within the stem cell community. BindDB includes a significantly expanded epigenomic database, which comprises a collection of over 450 genome-wide datasets in over 40 ESC and iPSC lines from mouse and human, integrated with a webtool that emulates our analysis pipeline.

Our database foundation includes the ENCODE (ENCODE Project Consortium, 2012) and Roadmap (Bernstein et al., 2010) databases as well as hundreds of datasets from the GEO repository. It is kept current and up to date to incorporate new data as it becomes available. In addition, the interactive webtool enables any scientist, computational and non-computational alike, to query any number of genes or regions of interest against the database and receive a comprehensive epigenomic profile in ESCs. Anything from a single gene promoter to a set of thousands of unannotated chromosomal regions can be queried. For larger queries, qualitative enrichment scores and statistical assessment of significance are performed to focus the user on the factors and histone modifications specifically enriched within the query group. Results are also hierarchically clustered in an effort to facilitate not only the correlations between the factors themselves but also the subdivisions of the genes/regions within the query group vis-a-vis epigenomic regulation.

The BindDB webtool can be found at http://bind-db.huji.ac.il/bindDB/default_new.php or by link from http://www.meshorerlab.huji.ac.il. The BindDB webtool receives a query of genomic regions or genes from the user as input via an interactive webform and then queries the database in order to determine which factors in the database have evidence of binding to the queried genes or regions. The single query section allows the user to query one gene and the portion of the gene to explore (promoter [proximal/distal] and/or the gene body) or one unannotated location in the genome in the form of "chrN:start-end." The multi-query section allows a user to upload a file containing either a list of gene symbols (Entrez, Refseq, or UCSC annotations) or a list of any genome coordinates in BED format. Once the "Get Epigenomic Profile" button is pressed, the query initiates to the database and provides several outputs: (1) a comma-separated-vector (.csv, Excel compatible) formatted raw results file of the epigenomic "barcodes" of the query regions, in the form of a binary matrix, where cell (i,j) contains the value of "1" if factor "j" binds queried region "i" and 0 if not; and (2) a dynamic (searchable, sortable) table of those factors, which bind one or more of the queried regions with extra columns of information for each factor such as the binding enrichment scores and p values of their statistical significance. The enrichment scores are also displayed in bar graph format for visual enhancement of these values. Furthermore, a downstream clustering analysis is performed on the resulting epigenomic barcodes of the query regions and displayed as a heat map. The factors can then be filtered according to enrichment score levels, statistical significance, and prevalence to generate new clustered profiles and heat maps, enabling the user to focus on the more impactful epigenomic aspects of the queried regions.

We have also incorporated an "Enhancer Finder" functionality for a single-gene query whereby the BindDB tool scans up to 50 kb upstream of the gene TSS and calculates a running "enhancer score" based on the number of binding factors (Chen et al., 2008); footprints of super-enhancers in mESCs such as presence of mediator complex components, Cohesin complex components, p300, and Nibpl (Chen et al., 2012; Hnisz et al., 2013); and histone modifications supporting enhancer identification such as H3K4me1 (Heintzman et al., 2007), H3K27ac (Creyghton et al., 2010), and a newly identified enhancer mark, H3K56ac. BindDB now makes it easier than ever to locate the potential upstream enhancer elements of a gene.

To test the functional relevance of our analysis, we also collected 290 different knockdown/knockout/overexpression (KD/KO/OE) experiments, which were followed by expression analyses in mESCs. Identified factors, for which KD/KO/OE data are available, can be thus further tested for potential functional implications for their enrichment.

The usage and utility of the BindDB webtool can be demonstrated in several

case studies (Oct4, http://dx.doi.org/10.17632/z9vv7tvpfp.1; pseudogenes, http://dx.doi.org/10.17632/7d2fb4m7gn.1; bivalent genes, http://dx.doi.org/10.17632/537hdf9zwz.1; and lincRNA, http://dx.doi.org/10.17632/j6z69cyyyy.1, which can be found at https://data.mendeley.com/) in which we point out different types of queries, different usages of the tool, and the specific insights they offer. We find that epigenetic regulation of bivalent genes is largely confined to histone modifications only and that pseudogenes do not, by and large, have evidence of factor binding in ESCs (negative control). We even define an expanded epigenetic profile for lincRNA genes beyond the basal "K4–K36" and detect subclasses of lincRNAs that may be indicative of their potential function.

To summarize, BindDB offers the non-computational biologist the opportunity to take advantage of the amounting epigenomic data in the field of ESC biology without having to write one line of code. The tool emulates our approach to epigenomic analysis by combining the advantage of an integrative view with an underlying flexibility of querying any genomic location (not only annotated genes). This combination positions BindDB as an ideal tool for analysis of yet unformulated biological questions as they arise.

ACKNOWLEDGMENTS

This work was supported by grants from the Israel Science Foundation (ISF 1252/12) and the European Research Council (ERC-281781).

REFERENCES

Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M.A., Beaudet, A.L., Ecker, J.R., et al. (2010). Nat. Biotechnol. *28*, 1045–1048.

Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V.B., Wong, E., Orlov, Y.L., Zhang, W., Jiang, J., et al. (2008). Cell *133*, 1106–1117.

Chen, C.Y., Morris, Q., and Mitchell, J.A. (2012). BMC Genomics *13*, 152.

Creyghton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T., Carey, B.W., Steine, E.J., Hanna, J., Lodato, M.A., Frampton, G.M., Sharp, P.A., et al. (2010). Proc. Natl. Acad. Sci. USA *107*, 21931–21936.

ENCODE Project Consortium (2012). Nature *489*, 57–74.

Gokhman, D., Livyatan, I., Sailaja, B.S., Melcer, S., and Meshorer, E. (2013). Nat. Struct. Mol. Biol. *20*, 119–126.

Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K.A., et al. (2007). Nat. Genet. *39*, 311–318.

Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-André, V., Sigova, A.A., Hoke, H.A., and Young, R.A. (2013). Cell *155*, 934–947.