D658–D662 *Nucleic Acids Research, 2017, Vol. 45, Database issue* *Published online 26 October 2016*
doi: 10.1093/nar/gkw983

Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse

Shenglin Mei[1,2,†], Qian Qin[1,2,†], Qiu Wu[1,2,†], Hanfei Sun[2], Rongbin Zheng[2], Chongzhi Zang[3,4], Muyuan Zhu[2], Jiaxin Wu[6], Xiaohui Shi[2], Len Taing[3], Tao Liu[7], Myles Brown[4,5], Clifford A. Meyer[3,4,*] and X. Shirley Liu[1,2,3,4,*]

[1]Clinical Translational Research Center, Shanghai Pulmonary Hospital, Tongji University, Shanghai 200433, China, [2]Department of Bioinformatics, School of Life Sciences and Technology, Tongji University, Shanghai 200092, China, [3]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Harvard T.H. Chan School of Public Health, Boston, MA 02215, USA, [4]Center for Functional Cancer Epigenetics, Dana-Farber Cancer Institute, Boston, MA 02215, USA, [5]Department of Medical Oncology, Dana-Farber Cancer Institute and Harvard Medical School, Boston, MA 02215, USA, [6]MOE Key Laboratory of Bioinformatics, Bioinformatics Division and Center for Synthetic & Systems Biology, TNLIST; Department of Automation, Tsinghua University, Beijing 100084, China and [7]Department of Biochemistry, University at Buffalo, Buffalo, NY 14214, USA

Received August 13, 2016; Revised September 23, 2016; Editorial Decision October 10, 2016; Accepted October 15, 2016

ABSTRACT

Chromatin immunoprecipitation, DNase I hypersensitivity and transposase-accessibility assays combined with high-throughput sequencing enable the genome-wide study of chromatin dynamics, transcription factor binding and gene regulation. Although rapidly accumulating publicly available ChIP-seq, DNase-seq and ATAC-seq data are a valuable resource for the systematic investigation of gene regulation processes, a lack of standardized curation, quality control and analysis procedures have hindered extensive reuse of these data. To overcome this challenge, we built the Cistrome database, a collection of ChIP-seq and chromatin accessibility data (DNase-seq and ATAC-seq) published before January 1, 2016, including 13 366 human and 9953 mouse samples. All the data have been carefully curated and processed with a streamlined analysis pipeline and evaluated with comprehensive quality control metrics. We have also created a user-friendly web server for data query, exploration and visualization. The resulting Cistrome DB (Cistrome Data Browser), available online at http://cistrome.org/db, is expected to become a valuable resource for transcriptional and epigenetic regulation studies.

INTRODUCTION

Genome-wide identification of transcription factor (TF) and chromatin regulator binding sites, histone modifications and chromatin accessibility is important for understanding transcriptional control governing biological processes such as differentiation, oncogenesis and cellular response to environmental perturbations (1–3). Massively parallel DNA sequencing combined with chromatin immunoprecipitation (ChIP-seq), DNase I hypersensitivity (DNase-seq) and the transposase-accessible chromatin assay (ATAC-seq) enable the genome-wide study of transcriptional regulation, histone modification and cis-regulatory elements (4–7).

Over the past decade, ChIP-seq, DNase-seq and ATAC-seq data have rapidly accumulated in large repositories such as gene expression omnibus (GEO) (8,9) and European nucleotide archive (ENA) (10). Many experimental biologists may not have the bioinformatics expertise to effectively use these valuable resources. The Encyclopedia of DNA Elements (ENCODE) Consortium has provided high quality processed genome-wide histone modification, chromatin regulator and transcription factor binding data in a selected set of human cell lines (3,11) and the NIH Roadmap Epigenomics Project has built a similar resource for human stem cells and tissues (12). These projects, however, do not support data generated outside the consortia. Other ChIP-seq databases, such as Cistrome CR (13), BloodChIP (14), CTCFBSDS (15), only contain a limited number of samples, each focusing on a narrow selection of factor or tissue types. Standardized quality control and streamlined analy-

*To whom correspondence should be addressed. Tel: +1 617 632 3012/3498; Fax: +1 617 632 2444; Email: xsliu@jimmy.harvard.edu
Correspondence may also be addressed to Clifford A. Meyer. Tel: +1 617 632 3012/3498; Fax: +1 617 632 2444; Email: cliff@jimmy.harvard.edu
†These authors contributed equally to this work as the first authors.

sis of ChIP-seq and chromatin accessibility data have also been lacking.

Most of the publicly available ChIP-seq and chromatin accessibility data are available in the GEO (8,9) and ENA (10) repositories. However, due to inconsistencies of metadata annotation, as well as lack of unified data processing procedures, quality control measures and interfaces for visualization and exploration, these valuable resources have been underutilized. We collected about 23 000 (pre January 1, 2016) samples including more than 800 TFs and 80 histone marks and processed these samples with a uniform pipeline, producing quality control metrics and analysis results. To make these resources more accessible to the research community, we developed Cistrome DB at http://cistrome.org/db, an integrated data analysis and visualization portal for ChIP-seq and chromatin accessibility data in human and mouse.

DATA BROWSER CONTENTS

All data sources and web-interface features are summarized in Figure 1. Cistrome DB consists of three parts: a curated metadata collection, processed data and a web interface.

Data sources and metadata annotation

Cistrome DB is a comprehensive annotated resource of publicly available ChIP-seq and chromatin accessibility data in human and mouse. Samples collected include those from the NCBI GEO database (8,9), ENCODE database (16) and Roadmap Epigenomics project (12). We have systematically annotated the following metadata for each sample: species, biological sources (cell line/population, cell type, tissue origin, strain, disease state), factor name, PubMed ID and citation. Metadata were automatically parsed from GEO entries, followed by manual curation of factor names and biological sources to ensure annotation consistency. The number of ChIP-seq and chromatin accessibility experiments has increased dramatically since 2007, with 2015 alone contributing 8941 new samples (Figure 2A). In total, our database contains 23 319 ChIP-seq and chromatin accessibility samples, 9953 for mouse and 13 366 for human. Among them, there are 10 276 ChIP-seq samples for transcription factors and chromatin regulators, 10 680 ChIP-seq samples for histone modifications and variants, 1370 chromatin accessibility samples and the remaining 993 are classified as other (Figure 2B). These samples include 713 different TFs and 76 histone modifications/variants in human, 480 TFs and 71 histone modifications/variants in mouse (Figure 2C).

Data processing and quality control metrics

In order to keep data consistent, all ChIP-seq, DNase-seq and ATAC-seq samples were processed by ChiLin (17,18), a streamlined pipeline for chromatin profiling data analysis and quality control. Analysis included three steps: read mapping, peak calling and peak annotation. We first downloaded raw ChIP-seq and chromatin accessibility data from GEO or ENA, and mapped them to the human (hg38) or the mouse (mm10) genome using BWA (19). Peak calling was done using MACS2 (20). Finally, we performed annotation analysis, including average conservation profiles across peak regions, motif analysis (21) and putative gene target identification (22). Comprehensive information about the tools and parameters used for data analysis can be found on the 'About' page of the Cistrome DB website.

A set of quality control (QC) metrics is an important feature of Cistrome DB. We provide seven different QC criteria across three layers (Figure 2D). In the reads layer, the median quality score is used to evaluate the raw sequencing quality; uniquely mapped reads is used to reflect mapping quality; PCR bottleneck coefficient (PBC) (23) is used to identify potential over-amplification by PCR. In the ChIP layer, the FRiP score (23) (fraction of non-mitochondrial reads in peak regions) and the number of high quality peaks with 10- or 20-fold enrichment over background were calculated to show data quality at the ChIP level. QC measures in the annotation layer include the proportions of peaks in promoters, introns and intergenic regions, along with the proportion of peaks overlapping with a union of DNase-seq peaks across diverse cell types. Using our collection of 23 319 samples, we established the thresholds of quality control characteristics based on the overall distributions. The Cistrome DB result page displays whether or not the quality control characteristics of each sample meet these quality control thresholds. See the Supplementary Material for details on the definition and calculation of QC statistics. Distributions of quality control statistics and thresholds are shown in Supplementary Figure S1 and on the Cistrome DB website.

Peak calling algorithms are specialized in identifying narrow or broad enrichment although there is no precise threshold that distinguishes one category from the other (24). Cistrome DB QC metrics are mostly developed for TFs and histone marks with sharp enrichment; for factors or marks with broad enrichment patterns the current QC measures might not be as reliable. The accuracy of ChIP-seq experiments is highly reliant on antibody specificity and quality and it is common for antibodies to recognize several proteins apart from the stated target. Current Cistrome DB QC metrics do not provide information on this issue. It is suggested that Cistrome DB users that are unfamiliar with potential pitfalls in ChIP-seq, DNase-seq or ATAC-seq understand the nature of biases in these data types before using Cistrome DB results in their research (18).

Data visualization and extensions

Cistrome DB also provides visualization functions that allow users to view peaks and signal intensity in either the UCSC (25) or WashU (26) genome browsers. Visualization of both single or batch samples is supported. For example, a 'super-enhancer' region of the genome contains multiple SOX2 and NANOG binding sites and is enriched in mediator and H3K27ac signals (Figure 2E). Using Cistrome DB, users can select the relevant ChIP-seq samples and visualize the co-binding pattern between master transcription factors on the WashU genome browser using the sample batch view function (Figure 2E). In addition, Cistrome DB can export data to our previously created Cistrome analysis pipeline (Cistrome AP) (21) for downstream analysis.

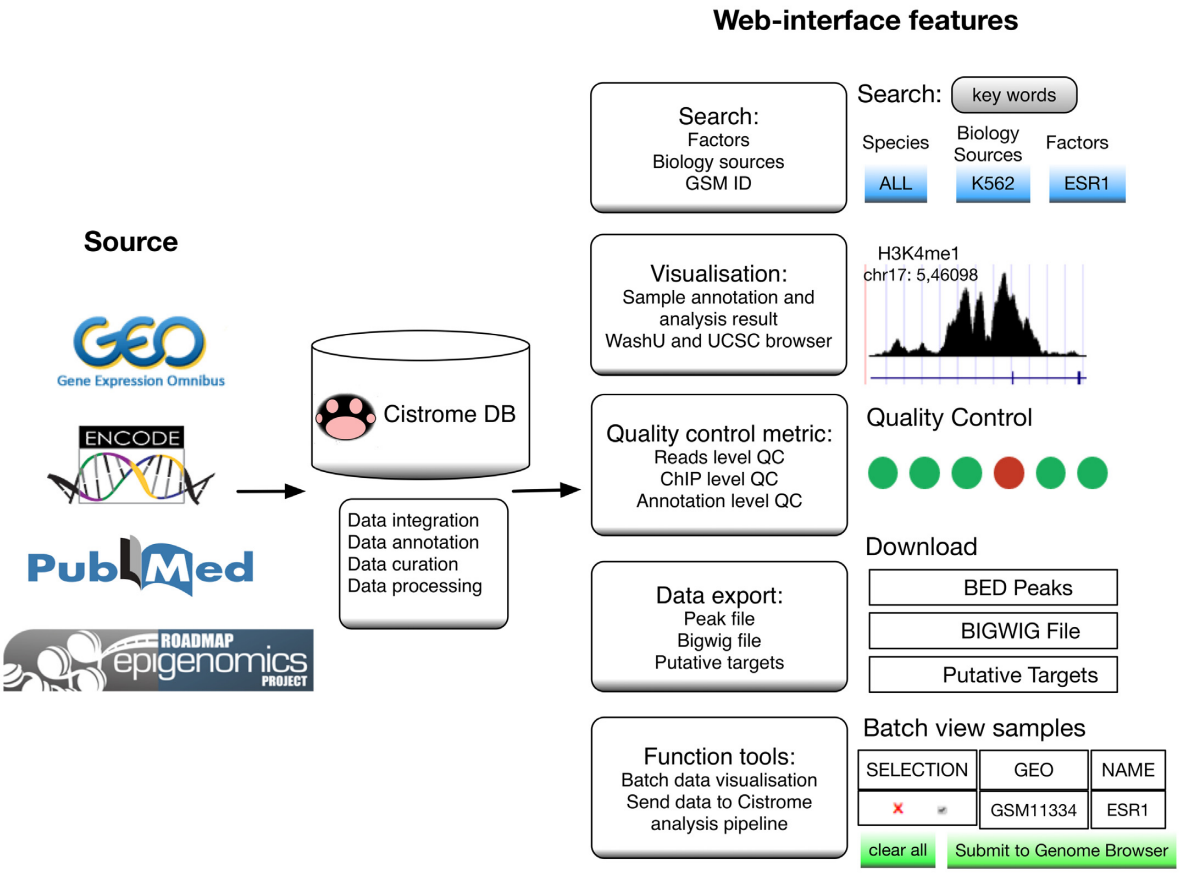

Figure 1. Schematic of Cistrome DB data sources and web-interface features. Cistrome DB collects publically available ChIP-seq, DNase-seq and ATAC-seq data from gene expression omnibus (GEO), Encyclopedia of DNA Elements (ENCODE) and Roadmap Epigenomics. Metadata is manually curated and annotated with PubMed information. All data are processed by a streamlined analysis pipeline and stored in a MySQL relationship database. Cistrome DB provides methods to query and visualize data. Users can search by key words or select by term. Detailed metadata annotations, analysis results and quality control (QC) metrics are presented for each sample. Data can be explored in more detail using the Cistrome analysis pipeline (Cistrome AP) and visualized using the UCSC and WashU genome browsers.

DATA RETRIEVAL

Samples

Cistrome DB provides automatically parsed and subsequent manual curation of metadata annotations for each sample, including species, factor name, biological source, publication and process status, which are stored in a local MySQL relational database. Each ChIP-seq or chromatin accessibility sample has a unique sample identifier. A web interface has been designed to provide user-friendly access and visualization. The result page displays detailed annotations, analysis results and quality control metrics for each sample. To track the source of the original sample, links to citations and the data repository are provided. Users can also download analysis results, send data to WashU or UCSC genome browsers for visualization, or Cistrome AP for subsequent analysis. In addition, Cistrome DB provides a list of putative target genes of the factor for each sample. Users can view the complete ranked list of putative target genes or search for a gene of interest by the gene symbol.

Queries

Cistrome DB contains two options for searching. One is through a selection list and the other is based on an advanced search menu. Users can select a sample of interest from a list of factors and biological sources, or search by factor name, cell type, GSM accession number or other keywords. Each search produces a table of matched samples. Users can then view detailed data annotations and analyze results by clicking on the table entry. Any sample of interest can be added to a batch view list and visualized in the UCSC and WashU genome browser.

Exploring data

Users can start their data explorations either by keyword search or by selecting a species and looking at lists of factors or biology sources. Searching produces a table of matched samples. Cistrome DB makes it easy to query one factor in multiple cell types or multiple factor types within a single cell type. Cistrome DB provides four layers of content for each sample. First, it provides a manually curated metadata annotation, including the species, factor name, biological source, citation and data accession number. Sec-

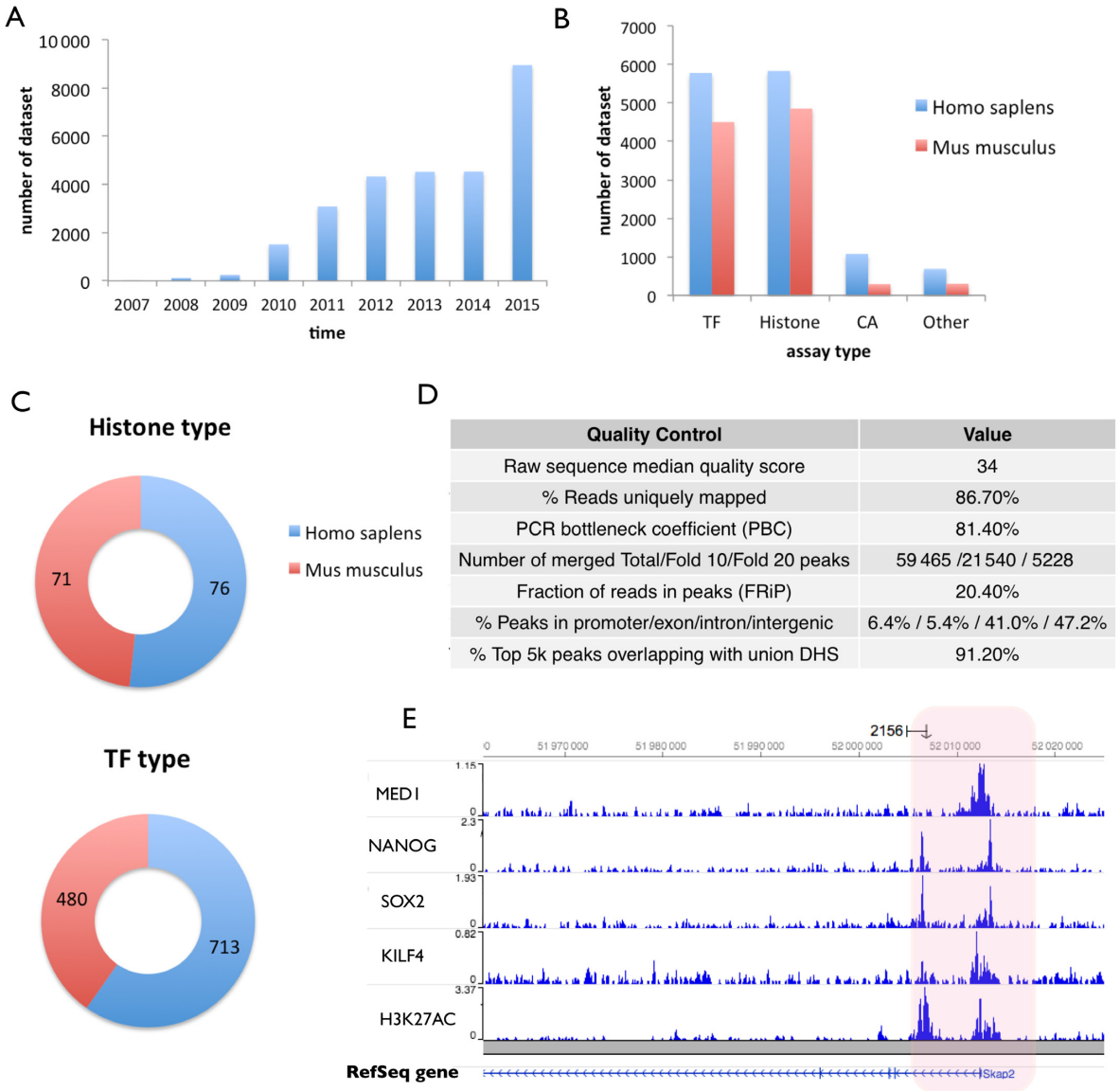

Figure 2. Database content. (**A**) Growth statistics of ChIP-seq and chromatin accessibility data. (**B**) Statistics of processed ChIP-seq and chromatin accessibility (CA) data in Cistrome DB. (**C**) Statistics of transcription factor and histone modification type. (**D**) Example of Quality control metric in Cistrome DB. (**E**) Batch sample visualization through WashU browser showing the co-binding pattern between master transcription factors in embryonic stem cells.

ond, it presents analysis results, including a peak file, a read density file, motif scan results, putative target genes and summaries of the distribution of peaks across different genomic locus categories. Third, it provides comprehensive QC metrics at the read, peak and annotation levels. Finally, it provides functions to analyze and visualize these samples; users can directly send data to the Cistrome analysis pipeline (Cistrome AP) or load data to the UCSC and WashU genome browsers for visualization. Both single-sample and batch visualization are supported.

DISCUSSION AND FUTURE DIRECTIONS

We present Cistrome DB, the most comprehensive knowledgebase and data portal for ChIP-seq and chromatin accessibility data in human and mouse. Cistrome DB is a valuable resource for transcriptional and epigenetic regulation studies. Transcriptional regulation is a complex process, which is controlled by hundreds of TFs, cofactors and chromatin regulators (27). With the Cistrome DB data, users can systematically investigate patterns of transcription factor or chromatin regulator binding and histone modifications related to their research questions. Chromatin profiles in diverse cell types and under various experimental conditions can be used in tissue specific or cell developmental studies (28). Cistrome DB provides quality control guidelines for ChIP-seq, DNase-seq and ATAC-seq experiments that allow users to disregard low quality results. Experimentalists can use our historical quality control data to evaluate the quality of their own ChIP-seq or chromatin accessibility experiments.

We update Cistrome DB on a regular basis to incorporate newly published ChIP-seq and chromatin accessibility samples. In the future, the utility of Cistrome DB will be improved in several ways. Cell information, QC metrics, TF type and histone modifications can be further classified. In addition we will also integrate Cistrome DB with other data. Cistrome DB is more than a data repository, it also allows users to visualize and explore the data. In comparison with other resources, Cistrome DB is by far the most comprehensive database for curated and analyzed ChIP-seq and chromatin accessibility data.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to acknowledge Sujun Chen, Qi Liao and Sheng'en Hu for their helpful discussions about the project.

FUNDING

National Institutes of Health [U01 CA180980]; National Natural Science Foundation of China [31329003]; Campaign Technology Fund of Dana-Farber Cancer Institute. Funding for open access charge: National Natural Science Foundation of China [31329003].
Conflict of interest statement. None declared.

REFERENCES

1. Park,P.J. (2009) ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.*, **10**, 669–680.
2. Furey,T.S. (2012) ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. *Nat. Rev. Genet.*, **13**, 840–852.
3. Thurman,R.E., Rynes,E., Humbert,R., Vierstra,J., Maurano,M.T., Haugen,E., Sheffield,N.C., Stergachis,A.B., Wang,H., Vernot,B. *et al.* (2012) The accessible chromatin landscape of the human genome. *Nature*, **489**, 75–82.
4. Schmidt,D., Wilson,M.D., Spyrou,C., Brown,G.D., Hadfield,J. and Odom,D.T. (2009) ChIP-seq: using high-throughput sequencing to discover protein-DNA interactions. *Methods*, **48**, 240–248.
5. Song,L. and Crawford,G.E. (2010) DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb. Protoc.*, **2010**, doi:10.1101/pdb.prot5384.
6. Barski,A., Cuddapah,S., Cui,K., Roh,T.Y., Schones,D.E., Wang,Z., Wei,G., Chepelev,I. and Zhao,K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.
7. Buenrostro,J.D., Giresi,P.G., Zaba,L.C., Chang,H.Y. and Greenleaf,W.J. (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods*, **10**, 1213–1218.
8. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M. *et al.* (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.*, **41**, D991–D995.
9. Edgar,R., Domrachev,M. and Lash,A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.
10. Gibson,R., Alako,B., Amid,C., Cerdeno-Tarraga,A., Cleland,I., Goodgame,N., Ten Hoopen,P., Jayathilaka,S., Kay,S., Leinonen,R. *et al.* (2016) Biocuration of functional annotation at the European nucleotide archive. *Nucleic Acids Res.*, **44**, D58–D66.
11. Djebali,S., Davis,C.A., Merkel,A., Dobin,A., Lassmann,T., Mortazavi,A., Tanzer,A., Lagarde,J., Lin,W., Schlesinger,F. *et al.* (2012) Landscape of transcription in human cells. *Nature*, **489**, 101–108.
12. 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) The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.*, **28**, 1045–1048.
13. Wang,Q., Huang,J., Sun,H., Liu,J., Wang,J., Wang,Q., Qin,Q., Mei,S., Zhao,C., Yang,X. *et al.* (2014) CR Cistrome: a ChIP-Seq database for chromatin regulators and histone modification linkages in human and mouse. *Nucleic Acids Res.*, **42**, D450–D458.
14. Chacon,D., Beck,D., Perera,D., Wong,J.W. and Pimanda,J.E. (2014) BloodChIP: a database of comparative genome-wide transcription factor binding profiles in human blood cells. *Nucleic Acids Res.*, **42**, D172–D177.
15. Ziebarth,J.D., Bhattacharya,A. and Cui,Y. (2013) CTCFBSDB 2.0: a database for CTCF-binding sites and genome organization. *Nucleic Acids Res.*, **41**, D188–D194.
16. Gerstein,M.B., Kundaje,A., Hariharan,M., Landt,S.G., Yan,K.K., Cheng,C., Mu,X.J., Khurana,E., Rozowsky,J., Alexander,R. *et al.* (2012) Architecture of the human regulatory network derived from ENCODE data. *Nature*, **489**, 91–100.
17. Qin,Q., Mei,S., Wu,Q., Sun,H., Li,L., Taing,L., Chen,S., Li,F., Liu,T., Zang,C. *et al.* (2016) ChiLin: a comprehensive ChIP-seq and DNase-seq quality control and analysis pipeline.*BMC Bioinformatics*, **17** , 404.
18. Meyer,C.A. and Liu,X.S. (2014) Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nat. Rev. Genet.*, **15**, 709–721.
19. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
20. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W. *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
21. Liu,T., Ortiz,J.A., Taing,L., Meyer,C.A., Lee,B., Zhang,Y., Shin,H., Wong,S.S., Ma,J., Lei,Y. *et al.* (2011) Cistrome: an integrative platform for transcriptional regulation studies. *Genome Biol.*, **12**, R83.
22. Wang,S., Sun,H., Ma,J., Zang,C., Wang,C., Wang,J., Tang,Q., Meyer,C.A., Zhang,Y. and Liu,X.S. (2013) Target analysis by integration of transcriptome and ChIP-seq data with BETA. *Nat. Protoc.*, **8**, 2502–2515.
23. Landt,S.G., Marinov,G.K., Kundaje,A., Kheradpour,P., Pauli,F., Batzoglou,S., Bernstein,B.E., Bickel,P., Brown,J.B., Cayting,P. *et al.* (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.*, **22**, 1813–1831.
24. Mendoza-Parra,M.A., Van Gool,W., Mohamed Saleem,M.A., Ceschin,D.G. and Gronemeyer,H. (2013) A quality control system for profiles obtained by ChIP sequencing. *Nucleic Acids Res.*, **41**, e196.
25. Speir,M.L., Zweig,A.S., Rosenbloom,K.R., Raney,B.J., Paten,B., Nejad,P., Lee,B.T., Learned,K., Karolchik,D., Hinrichs,A.S. *et al.* (2016) The UCSC Genome Browser database: 2016 update. *Nucleic Acids Res.*, **44**, D717–D725.
26. Zhou,X., Lowdon,R.F., Li,D., Lawson,H.A., Madden,P.A., Costello,J.F. and Wang,T. (2013) Exploring long-range genome interactions using the WashU Epigenome Browser. *Nat. Methods*, **10**, 375–376.
27. Lee,T.I. and Young,R.A. (2013) Transcriptional regulation and its misregulation in disease. *Cell*, **152**, 1237–1251.
28. Heintzman,N.D., Hon,G.C., Hawkins,R.D., Kheradpour,P., Stark,A., Harp,L.F., Ye,Z., Lee,L.K., Stuart,R.K., Ching,C.W. *et al.* (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, **459**, 108–112.