



epiCOLOC: Integrating Large-Scale and Context-Dependent Epigenomics Features for Comprehensive Colocalization Analysis

Yao Zhou¹, Yongzheng Sun¹, Dandan Huang¹ and Mulin Jun Li^{1,2*}

¹ Department of Pharmacology, Tianjin Key Laboratory of Inflammation Biology, School of Basic Medical Sciences, National Clinical Research Center for Cancer, Tianjin Medical University Cancer Institute and Hospital, Tianjin Medical University, Tianjin, China, ² Collaborative Innovation Center of Tianjin for Medical Epigenetics, Tianjin Key Laboratory of Medical Epigenetics, Tianjin Medical University, Tianjin, China

OPEN ACCESS

Edited by:

Geir Kjetil Sandve,
University of Oslo,
Oslo, Norway

Reviewed by:

Mikhail Dozmorov,
Virginia Commonwealth University,
Richmond, United States
Enrique Medina-Acosta,
Universidade Estadual do Norte
Fluminense Darcy Ribeiro,
Brazil

*Correspondence:

Mulin Jun Li
mulinli@connect.hku.hk

Specialty section:

This article was submitted to
Bioinformatics and
Computational Biology,
a section of the journal
Frontiers in Genetics

Received: 08 September 2019

Accepted: 17 January 2020

Published: 12 February 2020

Citation:

Zhou Y, Sun Y, Huang D and Li MJ
(2020) epiCOLOC: Integrating Large-
Scale and Context-Dependent
Epigenomics Features
for Comprehensive
Colocalization Analysis.
Front. Genet. 11:53.
doi: 10.3389/fgene.2020.00053

High-throughput genome-wide epigenomic assays, such as ChIP-seq, DNase-seq and ATAC-seq, have profiled a huge number of functional elements across numerous human tissues/cell types, which provide an unprecedented opportunity to interpret human genome and disease in context-dependent manner. Colocalization analysis determines whether genomic features are functionally related to a given search and will facilitate identifying the underlying biological functions characterizing intricate relationships with queries for genomic regions. Existing colocalization methods leveraged diverse assumptions and background models to assess the significance of enrichment, however, they only provided limited and predefined sets of epigenomic features. Here, we comprehensively collected and integrated over 44,385 bulk or single-cell epigenomic assays across 53 human tissues/cell types, such as transcription factor binding, histone modification, open chromatin and transcriptional event. By classifying these profiles into hierarchy of tissue/cell type, we developed a web portal, epiCOLOC (<http://mulinlab.org/epicoloc> or <http://mulinlab.tmu.edu.cn/epicoloc>), for users to perform context-dependent colocalization analysis in a convenient way.

Keywords: colocalization, epigenomics and epigenetics, functional annotation analysis, genetic variants, cell type specific, web server

INTRODUCTION

The epigenome, beyond genome sequence, has been increasingly recognized as key component in the gene regulation to drive certain biological processes and associate with many human diseases (Lawrence et al., 2016; Dor and Cedar, 2018; Feinberg, 2018). In the past decades, high-throughput epigenomic sequencing assays have profiled large numbers of functional elements across numerous human tissues/cell types, such as histone modification, DNA methylation, open chromatin, transcription factor binding site (TFBS), etc. The International Human Epigenome Consortium (IHEC) project (Bujold et al., 2016) have been initialized, across different countries and consortiums, to coordinate the production of reference maps of human epigenomes for key cellular states relevant to health and diseases. These unprecedented growths of epigenetic profiles and following comprehensive analysis of tissue/cell type-specific epigenomes will ultimately lead

to a better understanding of how human population and genome function are shaped in response to the environment (Egtex, 2017).

To facilitate convenient and accurate utilization of increasing volume of epigenomic data, several commonly-used resources have uniformly processed raw profiles and made them easily accessible, including ENCODE (Consortium, 2012), Roadmap Epigenomics (Roadmap Epigenomics et al., 2015), Blueprint Epigenome (Stunnenberg et al., 2016) and CistromeDB (Mei et al., 2017; Zheng et al., 2019). Furthermore, comprehensive epigenomics accumulation has motivated novel computational methods of modelling functional elements across many tissues/cell types, such as ChromHMM (Roadmap Epigenomics et al., 2015) and Segway (Libbrecht et al., 2019). Therefore, integrating such large-scale and context-dependent epigenomics features for novel biological findings is in urgent demand (Dozmorov, 2017; Cazaly et al., 2019). To this end, colocalization analysis was frequently used to study the interplay of various functional elements in different biological processes and conditions, where potential enrichment of a given genomic/epigenomic profile in pre-defined dataset could be drawn from the global perspective (Kanduri et al., 2019). Integrated with large-scale tissue/cell type-specific epigenomics data, colocalization analysis provides a powerful avenue to investigate biological relations and cell type specificities, such as identifying co-occurrence of transcription regulators (Yan et al., 2013) and inferring causal tissues/cell types from disease-associated variants identified by genome-wide association study (GWAS) (Farh et al., 2015).

Many colocalization tools have been developed by holding diverse assumptions and background models to assess the significance of enrichment. For instances, GSuite HyperBrowser is a web-based tool that performs colocalization analysis using either analytical approaches or Monte Carlo simulations (Simovski et al., 2017). LOLA utilizes Fisher's exact test based on universe regions to inspect enrichment and provides a web-based portal LOLAweb (Sheffield and Bock, 2016; Nagraj et al., 2018). GoShifter (Trynka et al., 2015) and GARFIELD (Iotchkova et al., 2019), which were implemented into standalone tools, specifically quantify enrichment of overlaps between GWAS variants and genomic annotations by considering linkage disequilibrium (LD). To overcome the discordant enrichment among existing methods, Coloc-stats integrates multiple colocalization analysis tools in a single web interface (Simovski et al., 2018). This integrated system serves as a one-stop shop for performing comprehensive colocalization analysis and assesses the consistency of the conclusions across seven different methods. However, some critical issues remain unaddressed. First, existing tools only provide limited pre-defined sets for genomic features in different biological domains. Current web-based tools, such as GSuite HyperBrowser, GenomeRunner (Dozmorov et al., 2016) and LOLAweb, only incorporate a small number of epigenomic profiles from ENCODE, Cistrome and other specific annotation datasets, which restrict the broader applications of online colocalization analysis. Second, the descriptions of tissue and cell type information are disordered and only based on free

text, making current tools unable to properly classify or group tissues/cell types to inspect the specificity of enrichment. Therefore, a uniform human tissue/cell-type definition is needed. Furthermore, the growing volume of epigenomic profiles on extensive tissues/cell types, collection and integration of these genomic features require a great effort to download. Most colocalization web tools are time-consuming for features intersection and background generation when dealing with such accumulating data scale. To ease the comprehensive colocalization analysis for biologists and geneticists, a faster and versatile online platform would be welcome.

For this study we comprehensively collected and integrated over 44,385 bulk or single cell epigenomic profiles across 53 human tissues/cell types. By classifying and mapping these profiles into hierarchy of tissue/cell type, we developed a web portal, epiCOLOC, for users to perform context-dependent colocalization analysis in a convenient way. We leveraged a recent ultrafast genomics search engine, GIGGLE, to identify and prioritize the enrichment of genomic loci shared between query features and our pre-defined epigenomic interval files (Layer et al., 2018). epiCOLOC equips many visualization functions and is freely available at <http://mulinlab.org/epicoloc> or <http://mulinlab.tmu.edu.cn/epicoloc>.

EPIGENOMIC PROFILES INTEGRATION AND PROCESSING

Data Collection

We collected human genomic and epigenomic data from various public resources including ENCODE (Consortium, 2012), Roadmap Epigenomics (Roadmap Epigenomics et al., 2015), Cistrome (Mei et al., 2017), ReMap (Cheneby et al., 2018), ChIP-Atlas (Oki et al., 2018), DeepBlue (Albrecht et al., 2017), BOCA (Fullard et al., 2018), TCGA (Corces et al., 2018) and HACER (Wang et al., 2019) (**Supplementary Table 3**). According to data sources and corresponding attributes, we classified collected features into following categories: 1) Transcriptional regulator, which incorporates ChIP-seq profiles of large number of transcriptional factors and chromatin remodelers; 2) Histone modification, which incorporates ChIP-seq profiles of different histone modifications; 3) Chromatin accessibility, which contains DNase-seq, ATAC-seq and FAIRE-seq profiles of open chromatin; We also curated several single cell ATAC-seq assays in this category; 4) Transcriptional event, which contains CAGE-seq, GRO-seq and PRO-seq profiles of nascent transcription signals; 5) Chromatin segmentation, which introduces tissue/cell type-specific chromatin states predicted by ChromHMM and Segway (**Figure 1A** and **Supplementary Table 1**). In order to improve accuracy and robustness of epiCOLOC backend database, we removed low-quality profiles according to the quality control scheme provided in the original resource. For example, we removed ChIP-seq data not passing two Cistrome quality metrics, including fraction of reads in peaks, and sufficient number of peaks with good enrichment. We also

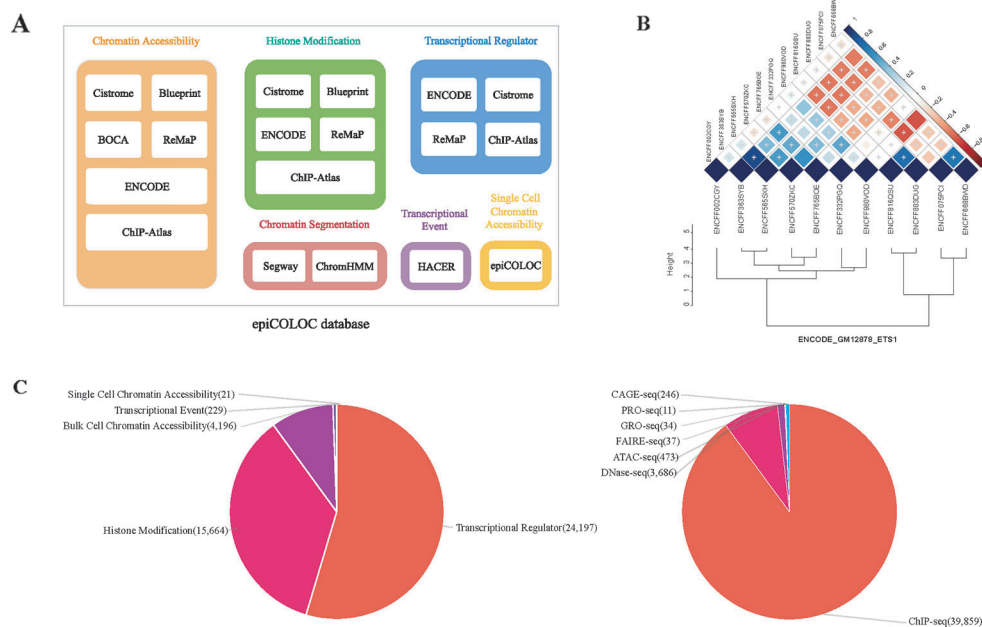


FIGURE 1 | The overview of epiCOLOC design and datasets. **(A)** The source schema of epiCOLOC data collection. **(B)** An example to illustrate outlier profiles removing. **(C)** The summary of data types in the current version of epiCOLOC.

excluded ENCODE profiles with error audit flags, such as extremely low read length, not tagged antibody, etc. Current epiCOLOC database covers 1,631 chromatin markers, which comprises 88 histone modifications, 1,538 transcriptional regulators, open chromatin and transcriptional event.

Data Processing

Tissue Organization and Mapping

We mapped cell lines to tissues by accounting for some auxiliary information from original epigenomic studies and several standards from GTEx (Consortium et al., 2017), Expression Atlas (Papatheodorou et al., 2018), Cellosaurus (Bairoch, 2018), ATCC (www.atcc.org), and BRENDA Tissue Ontologies (www.ebi.ac.uk/ols/ontologies/bto), yielding 53 main human tissues in total. For some main tissues that contain multiple well characterized components or some cell lines that cannot simply map to specific main tissues, we set independent terms in tissue set and finally generated 137 sub-tissues (**Supplementary Table 2**). We then manually mapped tissue/cell type name of each profile to our uniformly defined tissue set.

Cell Type Mapping

To reduce the complexity of cell type description in our collected epigenomic profiles, we performed cell type mapping using Cellosaurus that collected almost all cell line synonyms in a reference database (Bairoch, 2018). We acquired the Cellosaurus accession numbers and corresponding synonyms for all recorded cell lines, and assigned uniform synonyms identifiers to epigenomic profiles, which greatly reduces the heterogeneity of cell type descriptions. For cancer cell types mapping, we

borrowed DepMap which provides standard terms for over thousands of cancer cell lines and organoid models (Van Der Meer et al., 2019). Since DepMap provides Cellosaurus accession numbers, we were able to easily map cancer cell lines to consistent reference.

Profile Grouping

Since the epigenomic data were generated by different laboratories or produced using different protocols, replicates and analysis methods among collected sources, we sought to identify profiles describing similar biological processes in each source. We grouped all collected profiles according to source + assay type + tissue/cell type + biological target, and assigned unique group identifiers to them.

Outlier Profiles Removal

To further ensure informative profiles in each group, we designed a strategy to eliminate potential outlier profiles that may deviate from underlying biological process of the group (**Supplementary Methods**). For each group with at least three profiles, we first constructed a pair-wise similarity matrix for all profiles based on GIGGLE combo score (Layer et al., 2018). Then, hierarchical clustering was used to cluster these profiles based on Euclidean distance and the optimal number of clusters was automatically determined by inconsistency coefficient method (Zahn, 1971). Furthermore, we only retained profiles within the largest cluster as representatives in this group. For example, we identified that four outlier profiles among 11 ETS1 ChIP-seq peak profiles in GM12878 cell line, and excluded them in the colocalization analysis (**Figure 1B**).

epiCOLOC Web Tool Implementation

The current version of epiCOLOC incorporates 44,385 tissue/cell type-specific functional profiles from 44,364 bulk-cell studies and 21 single-cell studies after quality control (**Supplementary Table 4**). Most of these profiles (89.8%) are derived from ChIP-seq for transcription regulators and histone modifications, while, 9.5% profiles came from DNase-seq and ATAC-seq for chromatin accessibility (**Figure 1C**).

Colocalization Method

To achieve a fast and efficient colocalization based on high volume epigenomic features, we embedded a genomic feature search engine, GIGGLE, into epiCOLOC web server (Layer et al., 2018). GIGGLE uses Fisher's exact test and odds ratio of "observed" versus "expected" to measure enrichment between query features and pre-indexed genomic intervals. It also creates a combination score called GIGGLE combo score, which is the product of $-\log_{10}$ (Fisher's exact test P -value) and $\log_2$ (odds ratio). Given thousands of epigenomic profiles in epiCOLOC database, GIGGLE can significantly reduce the running time from hours to minutes. For example, epiCOLOC takes about 6 minutes to finish colocalization analysis on transcriptional regulator profiles of all blood cells for a set of 10k intervals (randomly generated genomic intervals with varying length). For each profile group, we calculated median score to represent group-level enrichment. With the aid of efficient colocalization strategy, epiCOLOC tries to provide powerful context-specific epigenomic evidences, leading to novel biological problems identification, such as "Are two transcription factors (TFs) colocalized and forming cooperation" or "Are the query variants/intervals enriched in chromatin open regions of specific tissues?" or "Are the query variants/intervals overlap with transcribed enhancers regions more than would be expected by chance?" More biological examples can be found in our website <http://mulinlab.org/epicoloc/Introduction/#Biological-examples>.

Web Interface and Usage

epiCOLOC was implemented in a web-based tool with built-in large-scale and context-dependent epigenomic annotations. The epigenomic profiles were indexed using GIGGLE. The web server was developed by Python, jQuery, igv.js, amcharts.js and related JavaScript modules.

Queries

epiCOLOC accepts two types of genomic format: BED-like format and VCF-like format. Both plain text and uploaded file of regions of interest (ROIs) or variant positions are well supported. Uploaded file can be BED or VCF text file or compressed gzip file (<20Mb).

Options

epiCOLOC provides several options for users to customize colocalization analysis, including 1) select tissues (53 tissues/137 sub-tissues); 2) select profile categories (Transcriptional regulator, Histone modification, Chromatin accessibility, Transcriptional event, Chromatin segmentation); 3) change

human genome assembly (GRCh37 and GRCh38); 4) define background genome size (3,095,677,412 for GRCh37 and 3,088,269,832 for GRCh38 in default); 5) set maximal interval length (500bp in default, and ROIs which exceed maximum length will be removed); 6) set extended length on both sides (no extension by default); 7) set central window size (cut the central area of genomic intervals, no central window by default).

Job Submission

Once submitted, the job will be sent to the backend of the web server for colocalization analysis. epiCOLOC displays a progress bar to track the execution status. It allows job retrieval by searching for the job ID in the home page, or by using a fixed URL (<http://mulinlab.org/epicoloc/<jobid>>) to check results directly, or through email notification.

Results Visualization

We used GIGGLE combo scores to prioritize colocalization results. Higher combo score indicates better enrichment on a specific profile, while negative combo scores suggest depleted enrichment (**Supplementary Figure 1**). Users can inspect and visualize the results in four different manners: 1) Prioritization table, which shows statistics metrics of colocalization including combo score, Fisher's exact P -value, odds ratio, the number of overlaps and extra information of enriched profiles (**Figure 2A**); 2) Tissue-wise pie charts for enrichment and depletion, which depict the per tissue proportion in all enriched (positive combo score) or depleted (negative combo score) profiles (**Figure 2B**). Users can click the slice of each tissue in the pie chart to see detailed sub-tissue results; 3) Tissue-wise bar plots, which display the representative enriched or depleted profiles in each tissue (**Figure 2C**). The user can search, scroll, zoom and hover over the bar plot to get detailed information of enrichment (only assay IDs for the best profiles in each group are displayed in hover tooltip). Once the label under the tissue-wise bar plots is clicked, cell type-wise bars which depict enrichment patterns for the top 20 enriched cell types appear in a pop-up window. 4) The IGV dashboard displays relative genomic location for queries genomic intervals and top five enriched profiles in colocalization analysis.

Download

epiCOLOC allows users to download colocalization results in csv format and result figures in png, jpg or pdf formats.

CASE STUDIES AND EVALUATIONS

By integrating large-scale tissue/cell type-specific epigenomic profiles, epiCOLOC could be used to investigate many biological questions. Here, we used several examples to demonstrate the performances and potential usages of epiCOLOC.

To identify potential disease-relevant genomic features and tissues using GWAS variants, we first performed colocalization analysis on disease-associated variants for inflammatory bowel



FIGURE 2 | Results page of epiCOLOC. Colocalization result for IBD GWAS variants in open chromatin regions, **(A)** Prioritization table. **(B)** Pie chart that depicts the number of significant enriched or depleted profiles in each tissue. **(C)** Bar plots that display ordered combo score, *P*-value, odds ratio in tissue-wise manner.

disease (IBD) (Liu et al., 2015) to test the tissue-specific enrichment. Using chromatin accessibility features, we found that IBD GWAS variants (*P*-value < 5E-8) were significantly enriched in blood tissue, where open chromatin profiles on monocyte, lymphocyte and granulocyte macrophage progenitor received highest enrichment scores. (Figure 2, and also see

colocalization result from: <http://mulinlab.org/epicoloc/results/bc2fa49a-6dfa-40f1-bb61-1349c9118168>). This result was consistent with GARFIELD results using functional annotations from ENCODE and Roadmap Epigenomics (Iotchkova et al., 2019). We then used coronary artery disease (CAD) GWAS variants (P -value $< 5E-8$) to perform colocalization in open chromatin regions (Van Der Harst and Verweij, 2018). Consistent with GARFIELD reports, we observed that most of tissues showed similar enrichment patterns, without distinct tissue specificity at open chromatin (<http://mulinlab.org/epicoloc/results/63b0cd1b-f22f-43dd-9452-fdea114f6c3d>). However, when using fine-mapped CAD variants, we observed several highly enriched signals in tissues like liver and artery blood vessel (<http://mulinlab.org/epicoloc/results/04bf79a8-f7cd-4960-913e-5c5c84c05753>), implying that the importance of selecting informative ROIs before colocalization analysis.

Next we sought to demonstrate that whether epiCOLOC could be used to identify potential cooperative factors for given TF. Transcription factor 7-like 2 (TCF7L2), a TF in the Wnt-signaling pathway, has been proven to play a central role in coordinating the expression of proinsulin and forming mature insulin (Zhou et al., 2014). TCF7L2 binding sites had been reported to colocalize with HNF4alpha and FOXA2 in HepG2 cell (Fietze et al., 2012). We hence used TCF7L2 ChIP-seq in HepG2 to perform colocalization analysis using epiCOLOC. In our colocalization results, TCF7L2 ChIP-seq peaks were significantly enriched in EP300, CREM, SP1, FOXA2 and HNF4alpha ChIP-seq profiles in various tissues/cell types (<http://mulinlab.org/epicoloc/results/d736578a-59a4-4160-a6fe-1a9c420c4adf>). Furthermore, we used two motif finding tools, PscanChIP (Zambelli et al., 2013) and HOMER (Heinz et al., 2010), with the same query input to investigated enriched TF motifs. We found that TF motifs including HNF4alpha, FOXA2, TCF7, GATA4, FOXF1, FOXA1, FOXK2 and FOXO3 can be simultaneously identified among two motif finding tools and our epiCOLOC, which also validates the efficacy of our tool.

DISCUSSION

In this study, we have integrated a comprehensive and tissue/cell type-specific epigenomics profiles database. With strict pre-processing, quality control and tissue mapping, we established a user-friendly web portal, epiCOLOC, which to perform fast and context-dependent colocalization analysis; and provide a series of visualization functions to interpret results; and significantly distinguish between existing web-based tools (**Supplementary Table 5**). In the applied examples, we demonstrated the accuracy and practicality of epiCOLOC in identifying causal tissues/cell types from GWAS disease-associated variants and inferring co-occurrence of transcription regulators.

There are some limitations in this work which deserve optimization in our future works. First, the statistical assumption of GIGGLE is simple and could be sub-optimal in several cases. We strongly recommend users to prioritize results by combo score and set stringent thresholds. As observed from the

combo scores distribution when $P \leq 0.05$ using query intervals that randomly generated in genome (**Supplementary Figure 2**), we propose to use an empirical combo score cutoff, 5 for enrichment and -2 for depletion, as advisable criteria to further filter enrichment or depletion results. Although GIGGLE can greatly speed up colocalization analysis, as compared with GenomeRunner (Dozmorov et al., 2016) and LOLAweb (Nagraj et al., 2018), it limits the usage of user-specific background of genomic regions and the analysis of multiple genomic intervals. Second, although epiCOLOC is applicable to perform colocalization analysis using genetic variants, but it cannot account for LD and allele frequency. Third, there are uneven epigenomic profiles for different tissues/cell types. It may potentially affect the robustness of colocalization when applying epiCOLOC to the tissues/cell types having fewer data available, and it also cannot determine the missing enrichment for tissues/cell types lacking sufficient data. In addition, single-cell technologies, such as single-cell ATAC-seq and single-cell ChIP-seq (Grosselin et al., 2019), have been developed to analyze genome-wide epigenomic features. Such approaches pave the way to study the role of epigenetic heterogeneity in many biological conditions and will be largely incorporated into epiCOLOC in the next stage. Recently, a novel algorithm named Augmented Interval List (AIList) (Feng et al., 2019), which introduces a new data structure and provides a significantly improved fundamental operation for highly scalable genomic data analysis. This method together with upcoming large-scale genomic features will be added in the epiCOLOC future updates.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found from ENCODE, Roadmap Epigenomics, etc and also related sources has been listed here: <http://mulinlab.org/epicoloc/Introduction/>.

AUTHOR CONTRIBUTIONS

ML designed and guided the study, YZ, YS and DH developed the tool, YZ and ML wrote the manuscript.

FUNDING

This work was supported by grants from the National Natural Science Foundation of China 31871327, 31701143 (ML), Natural Science Foundation of Tianjin 18JCZDJC34700, 19JCJQJC63600 (ML). We also appreciate all tool and resource providers.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2020.00053/full#supplementary-material>

REFERENCES

- Albrecht, F., List, M., Bock, C., and Lengauer, T. (2017). DeepBlueR: large-scale epigenomic analysis in R. *Bioinformatics* 33, 2063–2064. doi: 10.1093/bioinformatics/btx099
- Bairoch, A. (2018). The cellosaurus, a cell-line knowledge resource. *J. Biomol. Tech.* 29, 25–38. doi: 10.7171/jbt.18-2902-002
- Bujold, D., Morais, D., Gauthier, C., Cote, C., Caron, M., Kwan, T., et al. (2016). The international human epigenome consortium data portal. *Cell Syst.* 3, 496–499. e492. doi: 10.1016/j.cels.2016.10.019
- Cazaly, E., Saad, J., Wang, W., Heckman, C., Ollikainen, M., and Tang, J. (2019). Making sense of the epigenome using data integration approaches. *Front. Pharmacol.* 10, 126. doi: 10.3389/fphar.2019.00126
- Cheney, J., Gheorghe, M., Artufel, M., Mathelier, A., and Ballester, B. (2018). ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. *Nucleic Acids Res.* 46, D267–D275. doi: 10.1093/nar/gkx1092
- Consortium, G. T., Laboratory, D. A. Coordinating Center -Analysis Working, G. Statistical Methods Groups-Analysis Working, G., Enhancing, G. G., and Fund, N. I. H. C., et al. (2017). Genetic effects on gene expression across human tissues. *Nature* 550, 204–213. doi: 10.1038/nature24277
- Consortium, E. P. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74. doi: 10.1038/nature11247
- Corces, M. R., Granja, J. M., Shams, S., Louie, B. H., Seoane, J. A., Zhou, W., et al. (2018). The chromatin accessibility landscape of primary human cancers. *Science* 362, eaav1898. doi: 10.1126/science.aav1898
- Dor, Y., and Cedar, H. (2018). Principles of DNA methylation and their implications for biology and medicine. *Lancet* 392, 777–786. doi: 10.1016/S0140-6736(18)31268-6
- Dozmorov, M. G., Cara, L. R., Giles, C. B., and Wren, J. D. (2016). GenomeRunner web server: regulatory similarity and differences define the functional impact of SNP sets. *Bioinformatics* 32, 2256–2263. doi: 10.1093/bioinformatics/btw169
- Dozmorov, M. G. (2017). Epigenomic annotation-based interpretation of genomic data: from enrichment analysis to machine learning. *Bioinformatics* 33, 3323–3330. doi: 10.1093/bioinformatics/btx414
- Egtex, G. P. (2017). Enhancing GTEx by bridging the gaps between genotype, gene expression, and disease. *Nat. Genet.* 49, 1664–1670. doi: 10.1038/ng.3969
- Farh, K. K., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W. J., Beik, S., et al. (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* 518, 337–343. doi: 10.1038/nature13835
- Feinberg, A. P. (2018). The key role of epigenetics in human disease prevention and mitigation. *N. Engl. J. Med.* 378, 1323–1334. doi: 10.1056/NEJMr1402513
- Feng, J., Ratan, A., and Sheffield, N. C. (2019). Augmented interval list: a novel data structure for efficient genomic interval search. *Bioinformatics* 35, 4907–4911. doi: 10.1093/bioinformatics/btz407
- Frietze, S., Wang, R., Yao, L. J., Tak, Y. G., Ye, Z. Q., Gaddis, M., et al. (2012). Cell type-specific binding patterns reveal that TCF7L2 can be tethered to the genome by association with GATA3. *Genome Biol.* 13, R52. doi: 10.1186/gb-2012-13-9-r52
- Fullard, J. F., Hauberg, M. E., Bendl, J., Egervari, G., Cirnar, M. D., Reach, S. M., et al. (2018). An atlas of chromatin accessibility in the adult human brain. *Genome Res.* 28, 1243–1252. doi: 10.1101/gr.232488.117
- Grosselin, K., Durand, A., Marsolier, J., Poitou, A., Marangoni, E., Nemati, F., et al. (2019). High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer. *Nat. Genet.* 51, 1060–1066. doi: 10.1038/s41588-019-0424-9
- Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., et al. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* 38, 576–589. doi: 10.1016/j.molcel.2010.05.004
- Iotchkova, V., Ritchie, G. R. S., Geihs, M., Morganello, S., Min, J. L., Walter, K., et al. (2019). GARFIELD classifies disease-relevant genomic features through integration of functional annotations with association signals. *Nat. Genet.* 51, 343–344. doi: 10.1038/s41588-018-0322-6
- Kanduri, C., Bock, C., Gundersen, S., Hovig, E., and Sandve, G. K. (2019). Colocalization analyses of genomic elements: approaches, recommendations and challenges. *Bioinformatics* 35, 1615–1624. doi: 10.1093/bioinformatics/bty835
- Lawrence, M., Daujat, S., and Schneider, R. (2016). Lateral thinking: how histone modifications regulate gene expression. *Trends Genet.* 32, 42–56. doi: 10.1016/j.tig.2015.10.007
- Layer, R. M., Pedersen, B. S., Disera, T., Marth, G. T., Gertz, J., and Quinlan, A. R. (2018). GIGGLE: a search engine for large-scale integrated genome analysis. *Nat. Methods* 15, 123–126. doi: 10.1038/nmeth.4556
- Libbrecht, M. W., Rodriguez, O. L., Weng, Z., Bilmes, J. A., Hoffman, M. M., and Noble, W. S. (2019). A unified encyclopedia of human functional DNA elements through fully automated annotation of 164 human cell types. *Genome Biol.* 20, 180. doi: 10.1186/s13059-019-1784-2
- Liu, J. Z., Van Sommeren, S., Huang, H., Ng, S. C., Alberts, R., Takahashi, A., et al. (2015). Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* 47, 979–986. doi: 10.1038/ng.3359
- Mei, S., Qin, Q., Wu, Q., Sun, H., Zheng, R., Zang, C., et al. (2017). Cistrome data browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. *Nucleic Acids Res.* 45, D658–D662. doi: 10.1093/nar/gkw983
- Nagraj, V. P., Magee, N. E., and Sheffield, N. C. (2018). LOLAweb: a containerized web server for interactive genomic locus overlap enrichment analysis. *Nucleic Acids Res.* 46, W194–W199. doi: 10.1093/nar/gky464
- Oki, S., Ohta, T., Shioi, G., Hatanaka, H., Ogasawara, O., Okuda, Y., et al. (2018). ChIP-Atlas: a data-mining suite powered by full integration of public ChIP-seq data. *EMBO Rep.* 19, e46255. doi: 10.15252/embr.201846255
- Papathodorou, I., Fonseca, N. A., Keays, M., Tang, Y. A., Barrera, E., Bazant, W., et al. (2018). Expression Atlas: gene and protein expression across multiple studies and organisms. *Nucleic Acids Res.* 46, D246–D251. doi: 10.1093/nar/gkx1158
- Roadmap Epigenomics, C., Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., et al. (2015). Integrative analysis of 111 reference human epigenomes. *Nature* 518, 317–330. doi: 10.1038/nature14248
- Sheffield, N. C., and Bock, C. (2016). LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. *Bioinformatics* 32, 587–589. doi: 10.1093/bioinformatics/btv612
- Simovski, B., Vodak, D., Gundersen, S., Domanska, D., Azab, A., Holden, L., et al. (2017). GSuite HyperBrowser: integrative analysis of dataset collections across the genome and epigenome. *Gigascience* 6, 1–12. doi: 10.1093/gigascience/gix032
- Simovski, B., Kanduri, C., Gundersen, S., Titov, D., Domanska, D., Bock, C., et al. (2018). Coloc-stats: a unified web interface to perform colocalization analysis of genomic features. *Nucleic Acids Res.* 46, W186–W193. doi: 10.1093/nar/gky474
- Stunnenberg, H. G. International Human Epigenome, C., , Hirst, M. (2016). The international human epigenome consortium: a blueprint for scientific collaboration and discovery. *Cell* 167, 1145–1149. doi: 10.1016/j.cell.2016.11.007
- Trynka, G., Westra, H. J., Slowikowski, K., Hu, X. L., Xu, H., Stranger, B. E., et al. (2015). Disentangling the effects of colocalizing genomic annotations to functionally prioritize non-coding variants within complex-trait loci. *Am. J. Hum. Genet.* 97, 139–152. doi: 10.1016/j.ajhg.2015.05.016
- Van Der Harst, P., and Verweij, N. (2018). Identification of 64 novel genetic loci provides an expanded view on the genetic architecture of coronary artery disease. *Circ. Res.* 122, 433–443. doi: 10.1161/CIRCRESAHA.117.312086
- Van Der Meer, D., Barthorpe, S., Yang, W., Lightfoot, H., Hall, C., Gilbert, J., et al. (2019). Cell model passports-a hub for clinical, genetic and functional datasets of preclinical cancer models. *Nucleic Acids Res.* 47, D923–D929. doi: 10.1093/nar/gky872
- Wang, J., Dai, X., Berry, L. D., Cogan, J. D., Liu, Q., and Shyr, Y. (2019). HACER: an atlas of human active enhancers to interpret regulatory variants. *Nucleic Acids Res.* 47, D106–D112. doi: 10.1093/nar/gky864
- Yan, J., Enge, M., Whittington, T., Dave, K., Liu, J., Sur, I., et al. (2013). Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites. *Cell* 154, 801–813. doi: 10.1016/j.cell.2013.07.034
- Zahn, C. T. (1971). Graph-theoretical methods for detecting and describing gestalt clusters. *IEEE Trans. Comput.* 20, 68–86. doi: 10.1109/T-C.1971.223083

- Zambelli, F., Pesole, G., and Pavesi, G. (2013). PscanChIP: finding over-represented transcription factor-binding site motifs and their correlations in sequences from ChIP-Seq experiments. *Nucleic Acids Res.* 41, W535–W543. doi: 10.1093/nar/gkt448
- Zheng, R., Wan, C., Mei, S., Qin, Q., Wu, Q., Sun, H., et al. (2019). Cistrome data browser: expanded datasets and new tools for gene regulatory analysis. *Nucleic Acids Res.* 47, D729–D735. doi: 10.1093/nar/gky1094
- Zhou, Y. D., Park, S. Y., Su, J., Bailey, K., Ottosson-Laakso, E., Shcherbina, L., et al. (2014). TCF7L2 is a master regulator of insulin production and processing. *Hum. Mol. Genet.* 23, 6419–6431. doi: 10.1093/hmg/ddu359

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2020 Zhou, Sun, Huang and Li. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.