scientific reports

OPEN

Developing OCHROdb, a comprehensive quality checked database of open chromatin regions from sequencing data

Parisa Shooshtari[1,2,3,4,5✉], Samantha Feng[6,7], Viswateja Nelakuditi[8], Reza Asakereh[1,3], Nader Hosseini Naghavi[2,3], Justin Foong[9], Michael Brudno[9,10✉] & Chris Cotsapas[11,12✉]

International consortia, including ENCODE, Roadmap Epigenomics, Genomics of Gene Regulation and Blueprint Epigenome have made large-scale datasets of open chromatin regions publicly available. While these datasets are extremely useful for studying mechanisms of gene regulation in disease and cell development, they only identify open chromatin regions in individual samples. A uniform comparison of accessibility of the same regulatory sites across multiple samples is necessary to correlate open chromatin accessibility and expression of target genes across matched cell types. Additionally, although replicate samples are available for majority of cell types, a comprehensive replication-based quality checking of individual regulatory sites is still lacking. We have integrated 828 DNase-I hypersensitive sequencing samples, which we have uniformly processed and then clustered their regulatory regions across all samples. We checked the quality of open-chromatin regions using our replication test. This has resulted in a comprehensive, quality-checked database of Open CHROmatin (OCHROdb) regions for 194 unique human cell types and cell lines which can serve as a reference for gene regulatory studies involving open chromatin. We have made this resource publicly available: users can download the whole database, or query it for their genomic regions of interest and visualize the results in an interactive genome browser.

In recent years, multiple international consortia-based projects, such as ENCODE[1], NIH Roadmap Epigenomics Mapping Consortium (REMC) Project[2], NHGRI Genomics of Gene Regulation (GGR) Project[3] and Blueprint Epigenome[4] have made several large scale chromatin accessibility datasets publicly available. ENCODE, GGR and REMC datasets contain DNase-I sequencing for hundreds of primary cell types, stem cells, cell lines and tissues. Blueprint Epigenome project has also made open chromatin data available for hundreds of haematopoietic cell types. These datasets are extremely useful for studying mechanisms of gene regulation in different contexts, including disease and cell development. However, current regulatory region identification pipelines process samples within these datasets individually, making it difficult to annotate and analyze DNase-I Hypersensitive Sites (DHS) across multiple samples in a study, and across studies. Accessibility of regions of open chromatin varies across cell types, and may even vary within biological replicates of the same cell type. The absence of a well-defined, reference map for genomic locations of accessible chromatin that is also consistent across replicates frequently poses challenges to many downstream analyses, such as investigations into the association between accessibility state of specific regulatory elements and the expression of the genes which they control.

[1]Department of Pathology and Lab Medicine, University of Western Ontario, London, ON N6A 5C1, Canada. [2]Department of Computer Science, University of Western Ontario, London, ON N6A 5C1, Canada. [3]Children's Health Research Institute, Lawson Health Sciences Center, London, ON N6C 2R5, Canada. [4]Ontario Institute for Cancer Research, Toronto, ON M5G 0A3, Canada. [5]Present address: Children's Health Research Institute, A5-144, 800 Commissioners Road East, London, ON N6C 2V5, Canada. [6]Microsoft Canada, Toronto, ON M5K 1E7, Canada. [7]Biomedical Engineering, University of Waterloo, Waterloo, ON N2I 3G1, Canada. [8]Center for Computational Medicine, Hospital for Sick Children (SickKids), Toronto, ON M5G 0A4, Canada. [9]University Health Network, Toronto, ON M5G 1L7, Canada. [10]Department of Computer Science, University of Toronto, Toronto, ON M5T 3A1, Canada. [11]Department of Neurology, Yale University, New Haven, CT 06510, USA. [12]Medical and Population Genetics and Stanley Center for Psychiatric Research, Broad Institute, Cambridge, MA 02142, USA. ✉email: pshoosh@uwo.ca; brudno@cs.toronto.edu; cotsapas@broadinstitute.org

Although most epigenomics datasets such as ENCODE, GGR, REMC and Blueprint Epigenome data contain multiple biological and/or technical replicates of the same cell type, a comprehensive replication-based quality check of individual regulatory regions identified by peak calling algorithms is still lacking. Quality checking measurements that are used in the majority of other open chromatin databases check quality of individual samples and filter unreliable samples[5,6]; however, they do not assess the quality of individual regulatory sites across multiple samples. It is already well-recognized that several peaks identified in a sample by peak calling algorithms may not be supported in replicate samples[7,8]. Quality checking methods such as jMOSAiCS[7], MSPC[8] and irreproducibility discovery rate (IDR)[9] have been developed to compare peaks identified in pairs of replicates in order to identify replicable peaks in sequencing experiments with a main focus on ChIP-seq experiments. IDR particularly has been used in the ENCODE Project, and has been applied to various sequencing-based experiments such as ATAC-seq, DNase-seq and ChIP-seq data. However, QC methods are limited by the fact that they compare replicates from one cell type only, and subsequently can be quite stringent at filtering true peaks in the absence of orthogonal information.

In order to address the above problems we designed and implemented a pipeline that obtained DNase-I Hypersensitive (DHS) peaks of 828 data samples obtained from multiple resources (including ENCODE, REMC, Blueprint Epigenome and GGR), aligned them across samples, assessed replicability of each DHS site across several cell types using our recently developed replication test[10] and filters unreliable DHS sites[10]. This has resulted in a quality-checked database of regions of open chromatin comprising 1,455,046 reliable DHS for 194 unique cell types across the autosomal genome.

In a related study, Meuleman et al. has developed a high-resolution map of DNase-I hypersensitive sites using the 733 biosamples from ENCODE and REMC projects. Here, we have incorporated the samples from two additional consortium-based projects (NHGRI GGR and Blueprint Epigenome), increasing the number of immune-related cell types and states. We have also incorporated a replication test to remove false positive DHS sites and improve quality and reliability of the detected DHS sites. Additionally, in our analysis, DHS sites are aligned across different cell types and tissues, facilitating correlating DHS sites with matched gene expression data.

In summary, OCHROdb is a comprehensive DHS database obtained by (a) aligning the same regulatory sites across more than 800 samples from multiple public datasets and (b) checking quality of individual DHS sites, and it can serve as a reference for studies concerned with DHS activities and their role in gene regulation. This database can be accessed through our website (https://dhs.ccm.sickkids.ca/), where users can query and download the data either fully or at specific genomic regions of interest. We have also prepared an interactive genome browser for effective visualization of DHS sites across several cell types.

Methods

Data collection.
We collected DHS data generated by multiple international consortia-based projects including ENCODE, REMC, GGR and Blueprint Epigenome. The data generated by ENCODE, REMC and GGR are hosted by ENCODE and publicly available through (https://www.encodeproject.org). Blueprint DHS data samples from Haematopoietic Epigenomes are available at (http://www.blueprint-epigenome.eu). In August 2017 we collected from these resources the following types of the data:

a. Data samples in BED format, where regions of open chromatin were identified by Hotspot peak calling algorithm (narrow peaks; FDR of 0.05)[11]. This was available for 362 samples from ENCODE, 318 samples from REMC, 51 samples from GGR and 97 samples from Blueprint Epigenome project (see Table S1 for a full list of samples).
b. Metadata describing file accession ID, cell type, project, assembly and file download URL (see metadata in Table S1). We used ENCODEExplorer R Bioconductor package[12] to get access to and query metadata for ENCODE, REMC and GGR data files. The metadata of Blueprint Epigenome data was downloaded from ftp://ftp.ebi.ac.uk/pub/databases/blueprint/data_index/homo_sapiens/data.index.

Data processing workflow.
Step 1: Analyzing raw DNase-I sequencing data to identify DHS peaks. All DHS samples released from ENCODE project website (i.e. GGR, ENCODE and REMC data samples) had previously gone through the ENCODE uniform processing pipeline for DNase-I Hypersensitive experiments (https://github.com/ENCODE-DCC/dnase_pipeline/). Briefly, first a bwa index and a mappability file were produced by using a reference (e.g. GCRh38) fasta file and a DNase mappability file specific to a read size. Then BWA was used to generate bam files containing all reads mapped to the reference genome, and the bam files were merged and filtered to include only high quality mapped reads. At the last step, hotspot peak calling algorithm[11] was applied to the filtered bam files to identify enriched regions (i.e. peaks and hotspots). DHS data samples released through Blueprint Epigenome project were also analyzed following a similar procedure. First, the reads were mapped to human genome GCRh38 reference using BWA 0.7.7. Bam files were sorted and duplicates were marked using Picard Tools[13]. Then, the output bam files were filtered and reads with Mapping Quality of less than 15 were removed. Finally, the Hotspot peak calling algorithm was applied to identify peaks and hotspots.

Step 2: Checking quality of samples. We originally obtained a total of 847 samples containing DHS peaks in BED format from ENCODE and Blueprint site (Fig. 1A). We found that five Blueprint data samples do not have a cell type and/or donor ID assigned to them in the metadata and removed them. Additionally, by considering distribution of number of DHS peaks per sample, we identified and removed 14 outlying samples with too many (> 500,000 peaks) or too few (< 40,000 peaks) DHS peaks (see Fig. 3A for distribution of number of DHS peaks per samples). The total number of samples remaining after the above quality checking steps was 828 (Fig. 1A). Of these 828 samples, 45 were mapped to GRCh37 genome reference and the rest were mapped to GRCh38 genome
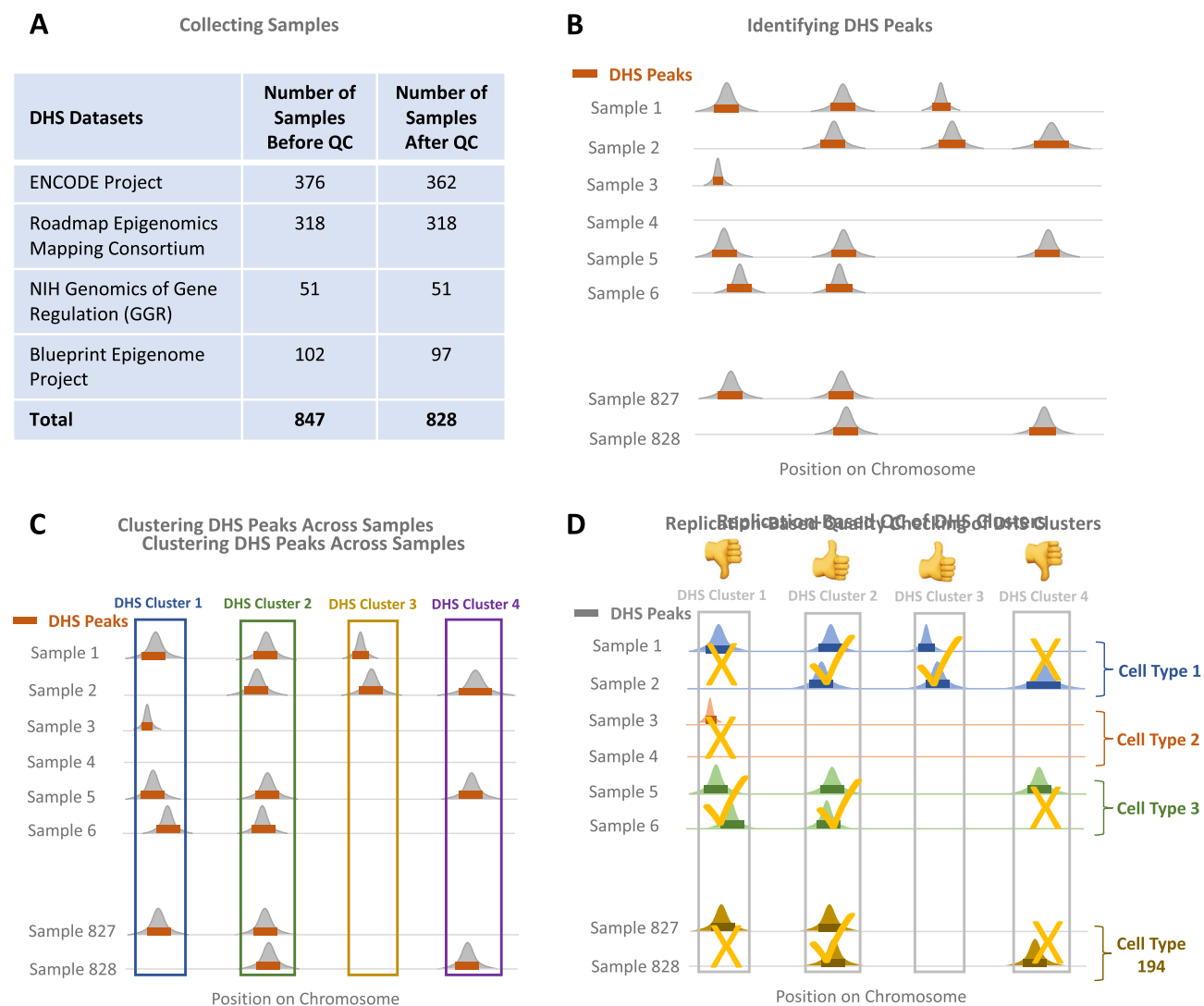
A Collecting Samples

DHS Datasets	Number of Samples Before QC	Number of Samples After QC
ENCODE Project	376	362
Roadmap Epigenomics Mapping Consortium	318	318
NIH Genomics of Gene Regulation (GGR)	51	51
Blueprint Epigenome Project	102	97
Total	847	828

Figure 1. Post processing of DNAse-I hypersensitive (DHS) data. (**A**) A total of 847 DHS data samples were collected from ENCODE, REMC, GGR and Blueprint Projects. 828/847 passed initial quality checking. (**B**) All the DNase-I sequencing data samples were already pre-processed by ENCODE and Blueprint consortia and narrow peaks were identified by Hotspot peak calling algorithm[11]. (**C**) Following the method we previously developed[10], we clustered DHS peaks across samples and (**D**) checked the quality of each DHS cluster using our replication test.

build. We converted GRCh37 samples to GRCh38 using the GATK API[14]. Figure S1 shows a histogram of the number of biological and technical replicates per cell type. As shown in this figure, these 828 samples come from 194 unique cell types. 161/194 cell types have at least two replicate samples.

Step 3: Clustering peaks. In the 828 samples, we found a total of 117,488,197 DHS peaks on the autosomes (i.e. excluding X and Y chromosomes). Peak calling algorithms such as Hotspot[11] and MACS[15] identify peaks in each individual sample separately (Fig. 1B). In order to identify the same DHS sites over multiple samples, we followed the method that we developed earlier[10], where we used Markov clustering (MCL)[16] with its default parameters to cluster the peaks across multiple samples (Fig. 1C). In our application, each DHS peak is treated as a graph node. Similarity between the two nodes is defined as the length of intersection between the two peaks divided by the length of union of the intervals containing the two peaks. Although any other clustering algorithm can be used at this stage, we chose MCL for two reasons. First, MCL does not require estimation of the number of clusters a priori. Second, the complexity of MCL is $O(Nk_2)$, where N is the number of nodes (i.e. peaks), and k is the number of resources allocated per node (a relatively small number for sparse graphs). The fact that MCL complexity is linear with the number of peaks makes it quite fast for our application. Using MCL we were able to group 117,488,197 DHS peaks across 828 samples into 4,020,940 DHS clusters across the autosomal chromosomes.

Step 4: Replication-based quality checking of DHS clusters.   Both peak calling and peak clustering ignore sample labels (i.e. cell types of origin), so we can check DHS cluster quality by looking for evidence that DHS peaks are replicated in this analysis. We expect that for a DHS cluster that represents a true regulatory site, the DHS accessibility is consistent across replicates. We therefore employed our previously designed replication test[10] to assess consistency across replicates (Fig. 1D). As required by our replication test, we selected cell types with at least two replicates (161/194 of cell types). Since for 84/161 cell types, we had more than two replicates per cell type (Fig. S1), we repeated our replication test ten times, each time we randomly selected two replicates per cell type. We then merged the test statistics obtained from each test. The output of each run of our replication test follows a Chi-squared distribution with one degree of freedom (2 (df = 1))[10] and thus, the sum of the test statistics resulting from running our replication test ten times follows a 2 (df = 10) distribution[17]. We selected DHS clusters that pass a nominal significance threshold of $p \leq 0.05$ for the combined replication test and called them replicable DHS. In order to examine how many runs of replication test is required in order to obtain a stable set of replicable DHS, we applied the replication test 20 times. Then for any N between 1 and 19, we found a list of replication tests obtained by combining N replication tests. Our results indicated that the Area Under the Curve (AUC) obtained for sensitivity/specificity analysis remained above 0.95 after nine iterations, confirming that running the replication test ten times is appropriate (Fig. S2). To verify the effects of changing P value threshold in detecting reproducible DHS peaks, we measured the proportion of DHS peaks that might be preserved if we had set the p value thresholds slightly above 0.05. We found that the numbers turned out to be quite small. More specifically, 0 (0%) and 76 DHS peaks (0.0019%) would have been preserved, if we had considered P value thresholds of 0.0501 and 0.0505, respectively.

Figure S3 shows a histogram of the p-values of all DHS clusters. 1,460,986 out of 4,020,940 DHS clusters (36.3%) passed the combined statistical test. We reasoned that if our replication test underpowered or DHS clusters were poorly detected, we would only capture about one third of the useful information across DHS peaks. To evaluate this, we asked if the 1,460,986 DHS clusters that pass our replication test explain a high proportion of disease heritability (h2g). We computed the enrichment of proportion of disease heritability[18] in each tissue for (1) all DHS peaks called across the sample replicates; and (2) the subset of the 1,460,986 replicable DHS active in that tissue. For these analyses, we used publicly available summary statistics data from genome-wide association studies for multiple sclerosis[19] as a test case. We found that autosomal genome covered by all DHS peaks and by the replicable DHS explained effectively the same disease heritability enrichment, indicating that our approach identifies most DHS relevant to disease pathogenesis (Fig. S4). This is consistent with our previous finding[10], where we showed active replicable DHSs capture the majority of proportion of disease heritability (h2g) explained by all DHS-detected peaks in a tissue, suggesting the non-replicating peaks are spurious. In addition, we overlapped our replicable DHS with the list of experimentally validated enhancers (from ENdb and VISTA databases) and promotors (from hsEPDnew database). We found that 80.2% and 93.7% of known enhancers from the ENdb and VISTA overlap replicable DHS peaks, respectively. In addition, 67.2% of known promotors from the hsEPDnew database overlap replicable DHS peaks. These results confirm that our replicable DHS peaks contain a good portion of experimentally tested enhancers and promotors from public databases; The replicable DHS peaks in our dataset that do not overlap the list of known promotors and enhancer (from ENdb, VISTA, and hsEPDnew) could be the promotors and enhancers which have not been experimentally validated yet in these databases. Also, these DHS peaks are likely to be specific to cell types not considered in the previous studies.

Step 5: Incorporating DHS accessibility significance.   Chromatin accessibility as a quantitative measure is the density of mapped DNase-I cleavage at different genomic locations. Chromatin accessibility varies across replicable DHS sites of the same sample. Also different samples have different accessibilities at the same replicable DHS site. Since DHS accessibility can be an indicator of how active a DHS site is in a particular sample, we incorporated DHS accessibility information to our data. To measure significance of accessibility of each replicable DHS in each sample, we first identified all DHS peaks that belong to a replicable DHS, and assigned $-\log10(p$ value) of the most significant DHS peak of a sample to that replicable DHS in the sample of interest.

There exist batch effects in DHS intensities (i.e. $-\log10(p$ value of accessibility)) due to the fact that DHS data were generated and processed in multiple centres (ENCODE, REMC, Blueprint and GGR) (Fig. 2A). As the result of this batch effect, in the tSNE plot, samples generated by the same project grouped together disregard of their cell type similarity (Fig. 2B). We applied the following steps to remove these batch effects. First, we normalized DHS intensities within each sample by linearly scaling them between 0 and 1. The number of accessible DHS varied from sample to sample. We therefore randomly selected 10,000 accessible DHS sites (i.e. with non-zero intensities) from each sample to estimate DHS intensity distributions for each sample, and adopted a method similar to ComBat[20,21] to remove batch effects. Briefly, we sorted intensities of randomly selected DHS sites within each sample, quantile normalized them to make distributions of intensities similar across samples. We then interpolated intensities of the DHS sites that were not initially chosen in the random selection and estimated their quantile normalized values. This process resulted in removing batch effects of DHS intensities, and made them comparable across samples generated by different centres, and therefore samples grouped together based on their cell type similarity (Fig. 2C).

Finally, we collapsed DHS intensities of samples from the same cell types by measuring median intensities over these samples. 1,455,046 out of 1,460,986 (99.5%) replicable DHS have median intensities higher than 0.25 in at least one cell type. We considered these 1,455,046 sites as the final replicable DHS sites to build our database of regions of open chromatin for 194 cell types.
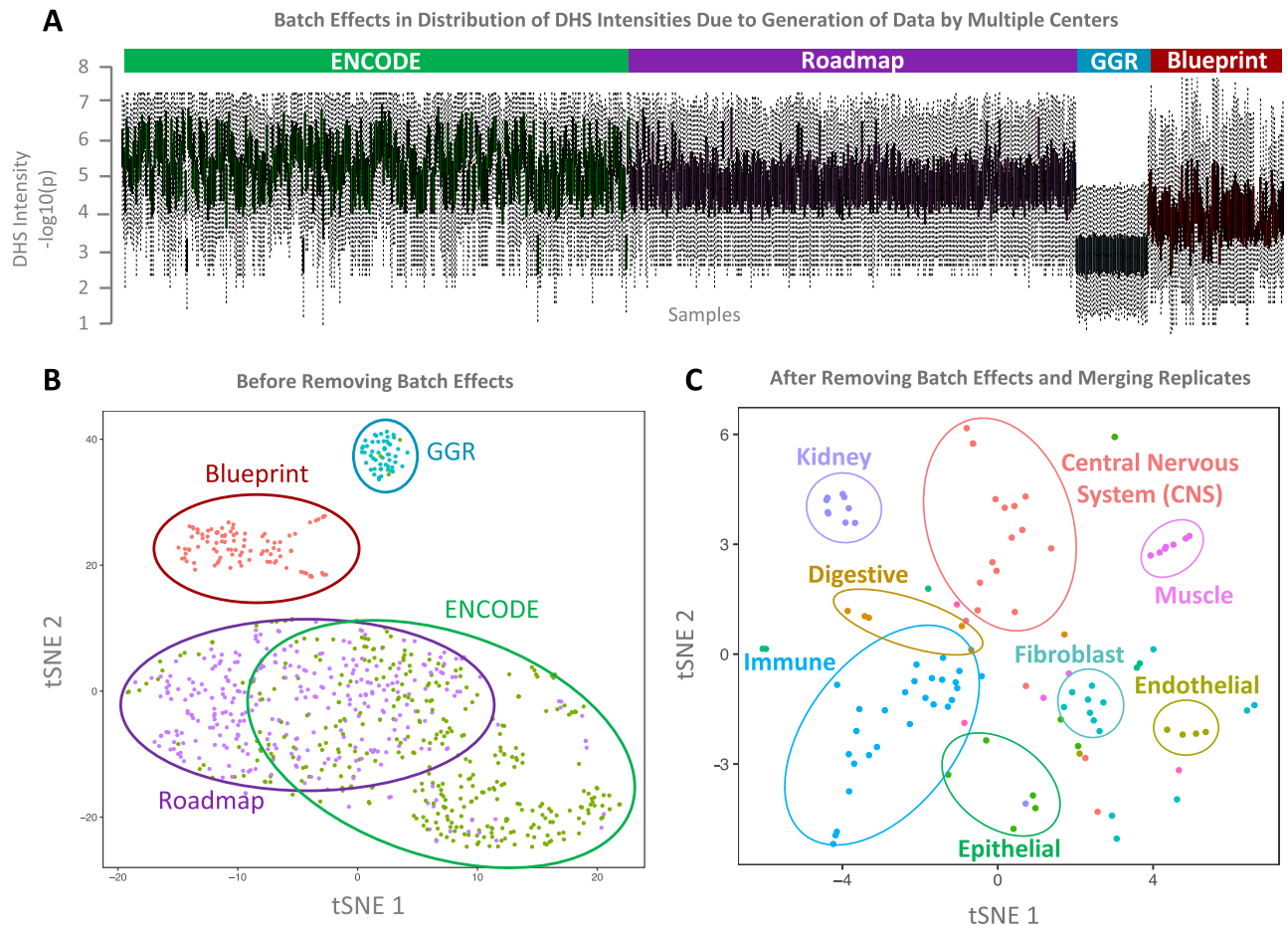
Figure 2. Removing batch effects caused by generation of data by multiple centers. (**A**) There exist batch effects in accessibility levels of replicable DHS clusters as a result of generation of data by multiple centers. (**B**) This resulted in grouping samples together based on the project that generated them disrespect of cell type similarity. (**C**) We removed batch effects and merge intensities of multiple replicates of the same cell types. After batch effect removal, in tSNE plot cell types grouped together based on their cell type similarities.

Results

Database content and statistics. We have prepared a comprehensive and quality-checked database of regions of open chromatin using DNase-I Hypersensitive datasets generated by REMC, ENCODE, GGR and Blueprint Epigenome Projects. OCHROdb contains DHS intensities for 1,455,046 DHS sites aligned over 828 samples. These 828 samples comprise of 194 distinct cell types, including 27 immune cell types, 18 CNS, 9 muscle, 57 cell lines, 10 stem cells, 6 digestive system cell types in addition to lung, cardiovascular, fibroblasts, kidney, extra embryo, male and female reproductive system and endothelial cell types (see Table S1 for the full list of cell types). tSNE plots of DHS activities of the 194 cell types shows that cell types from similar tissues group together confirming they have similar DHS activities (Fig. 2C).

As it is shown in Fig. 3C, the number of accessible replicable DHS varies across cell types. While number of DHS peaks varies across samples, majority of samples have < 200,000 DHS peaks (Fig. 3A). Additionally, the number of cell types where a replicable DHS is accessible can be any number between one cell type (i.e. cell-type specific DHS) to 194 cell types (i.e. constitutive DHS). The median number of active cell types per replicable DHS is eight (mean = 21.3; sd = 33.4) (Fig. 3B). Around 10.2% (148,142/1,455,046) of replicable DHS are cell-type specific, meaning that they are active in one cell type, and only 1,773 out of 1,455,046 replicable DHS are constitutive, meaning that they are active in all 194 cell types. Median length of replicable DHS is 310 base pairs (sd = 112 bp) (Fig. 3B).

Web interface and genome browser. We developed a web interface that permits users and researchers access to OCHROdb by primarily utilizing React, Tabix, Node.js, and Express.js. React[22], a component-based JavaScript library developed by Facebook, was selected for the front-end to establish a single page application with high performance. React helps to render HTML for web applications without refreshing the page or website templates and allows the client and server to communicate faster. Tabix is a software package that indexes tab-delimited files to efficiently perform queries and was developed specifically for biological data. Data indexing involves sorting data based on specific fields and allows a query to be completed without reading the entire data file, greatly reducing query speeds if implemented properly. Several actions were executed to prepare the
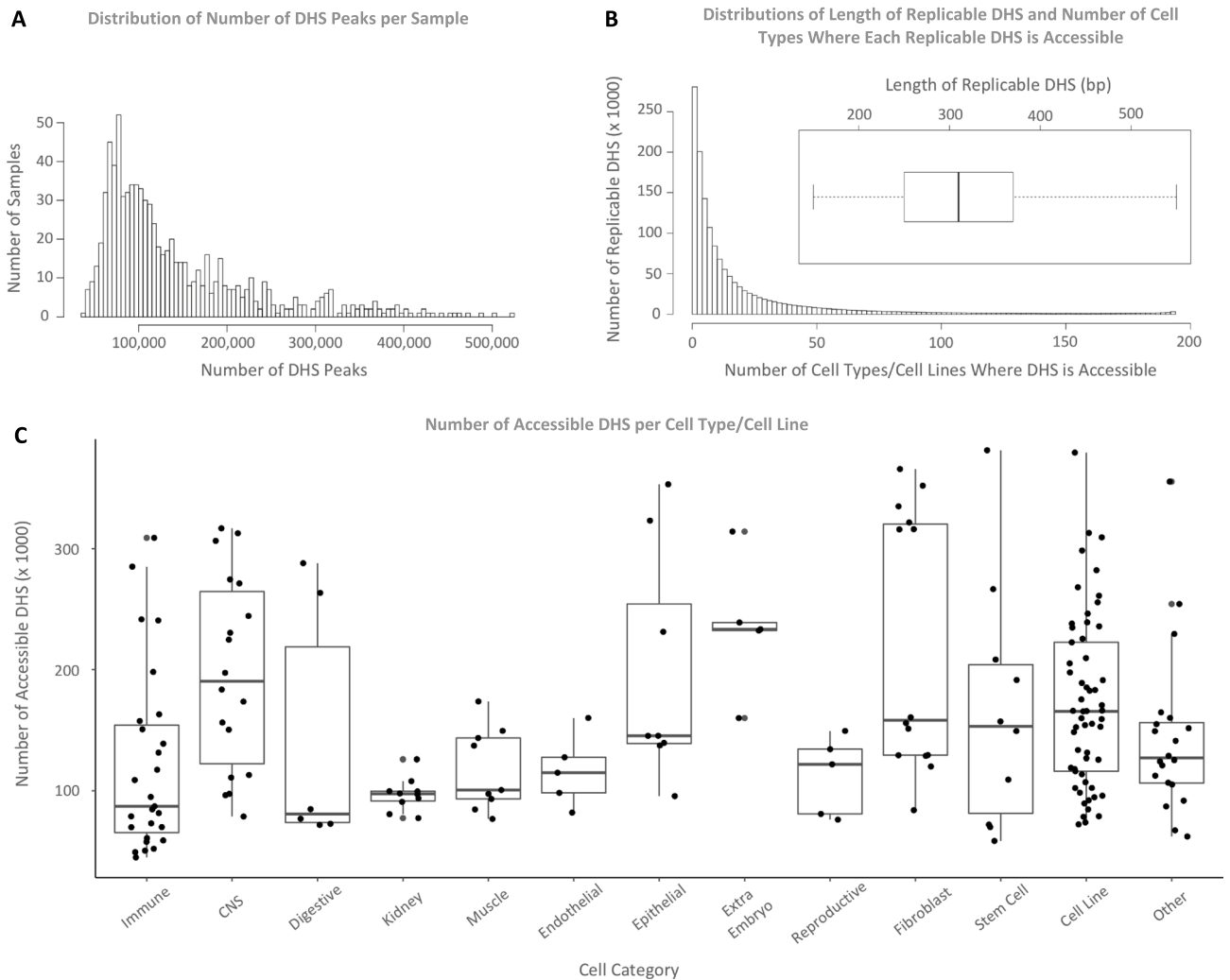
Figure 3. Statistics on DHS data. (**A**) Distribution of number of DHS peaks per sample. (**B**) Distributions of length of replicable DHS and number of cell types where each replicable DHS is accessible. Average length of replicable DHSs is 310 base pairs. Around 10.2% of replicable DHS are active in one cell type only and a very small portion of DHS (1,773/1,455,046) are active in all cell types. (**C**) Number of replicable DHS is between around 40,000 to 400,000, and it varies substantially across cell types. The percentage of genome covered by replicable DHS per cell type is between around 0.4% to 4% depending on the cell type.

DHS dataset for Tabix queries, such as sorting the files by chromosome and start position, compressing the file, and indexing the file with Tabix. The back-end of our web application is composed of the server-side JavaScript runtime environment Node.js[23], in conjunction with Express.js[24], a web application framework that is used for the web server. Node.js forms a connection between Tabix[25] and React front-end and allows the user to view and interact with the DHS data on the web application. Express.js is a flexible and highly documented web framework built on top of Node.js and greatly simplifies the complexity of back-end code written.

We have made the metadata, entire curated DHS dataset and data specific to each chromosome downloadable from the web interface. Through the web interface, the user can also query the database by specifying a region of interest (i.e. entering a specific chromosome number and start and end coordinates). After submitting coordinates of a region of interest, an exportable table of the results matching the user input is generated by a JavaScript library called DataTables.net[26] (Fig. S5A).

Additionally, the user can visualize the replicable DHS data through JBrowse[27], an embeddable genome browser, by specifying cell types of interest, and the coordinates of the genomic region or the name of the genes of interest (Fig. 4). To do this, a GFF (general feature format) file per cell type is generated by reading the original BED files using a Python script. Information specific to each replicable DHS (e.g. intensity of the DHS for any given cell-type and genomic coordinates of the DHS) can be found in the pop-up window that opens after clicking on a given DHS (Fig. S5B). Replicable DHS are only visible if they are accessible (i.e. having non-zero intensities) at the cell types of interest. The replicable DHS that are accessible in multiple cell types get the same color on different tracks (each corresponding to a cell type) of the genome browser, to make adjacent DHS more trackable visually across several cell types. In addition, we have a track called *RefSeq Genes,* where users can click on the corresponding checkbox to add a track for the gene names and locations.
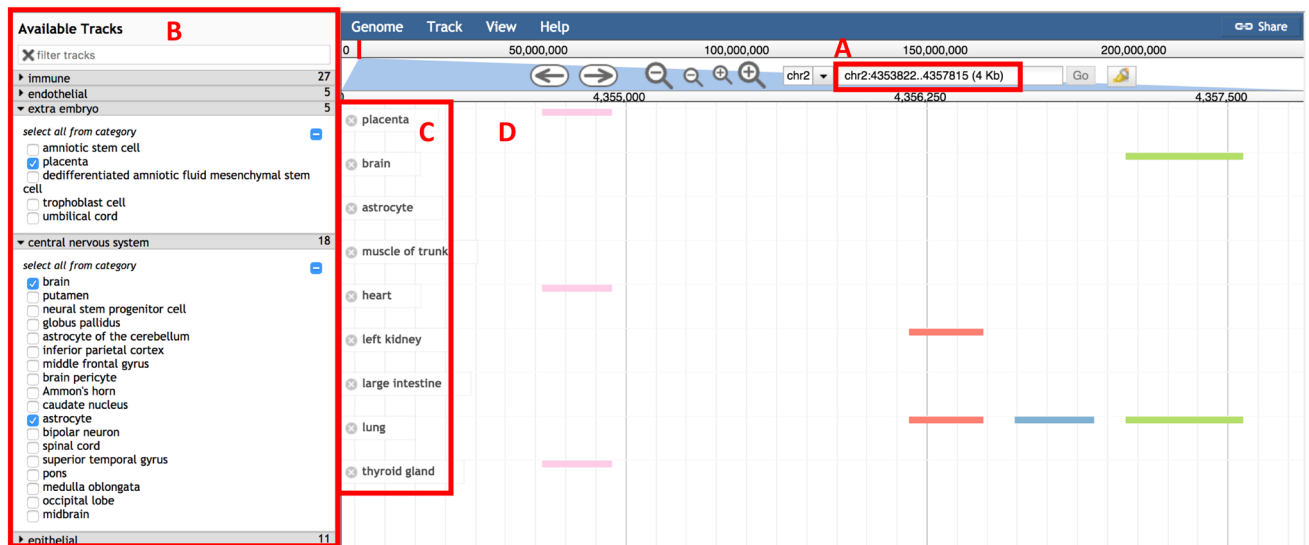
Figure 4. Interactive genome browser. We developed a customized genome browser to visualize DHS data. Here users can (**A**) specify their genomic region of interest, and (**B**) select from different cell categories and different cell types within each category. (**C**) The selected cell types appear on the left side of the browser. (**D**) Genomic location of replicable DHS appears in the middle of the browser, where each replicable DHS, aligned across multiple cell types, appears in a different colour to make it distinguishable from other replicable DHS adjacent to it.

Planned future development. Although we have explained our processing pipeline in the context of DHS data samples, it can also be applied to the narrow peaks obtained by other assays such as ATAC-seq. We therefore plan to process open chromatin datasets generated by other assays (e.g. ATAC-seq) and make these results available in our database. Additionally, we aim to process non-human (e.g. mouse) open chromatin datasets and release them separately in our open chromatin database.

Recently, the International Human Epigenome Consortium (IHEC) have collected and processed individual epigenome data samples from multiple large-scale international projects, such as ENCODE, REMC and Blueprint Epigenome. It is expected that more epigenome samples from other projects become available through IHEC portal. This includes the data from Deutsches Epigenome Program (DEEP), McGill Epigenomics Mapping Center, 4D Nucleome, Hong Kong Epigenomes Project (EpiHK), Multiple MS and SYSCID. As more open chromatin data samples are generated and released by IHEC and other projects[28], we plan to apply our processing pipeline to these samples and release the processed data as part of our open chromatin database. Making a comprehensive database of regions of open chromatin using the data released by IHEC project aligns with IHEC goals and will benefit the scientific community. We will maintain our database and release new updates every six months.

Discussion

Multiple large-scale consortia-based projects, including ENCODE, REMC, Blueprint and GGR have generated thousands of sequencing-based data samples that capture regions of open chromatin for the whole genome in hundreds of cell types. Building on this immensely informative DHS datasets, we have developed an analysis pipeline that gets hundreds of pre-processed DHS data samples (i.e. in narrow peaks format) as the input, aligns regions of open chromatin across samples, checks quality of each region using a replication-based test, and outputs a database of open chromatin accessibility across the whole genome.

Through applying our processing pipeline to 828 DHS data samples obtained from multiple public datasets, we have built a database of 1,455,046 regions across the whole genome for 828 samples consisting of 194 cell types. The main advantages of OCHROdb compared to previous datasets that are available publicly (e.g. ENCODE, REMC, Blueprint and GGR) are as follows.

(1) In the publicly available datasets, the DHS peaks are identified for each individual sample separately, while our method incorporates peaks from multiple samples together, aligns them across samples, and releases them as a coherent DHS database consisting of replicable peaks from multiple samples. Since in the previous datasets, each sample is analyzed separately, annotation of the same DHS sites across several samples is not available.

(2) Additionally, the batch effects that exist across samples due to the fact that multiple centers generate and analyze the data are not adjusted in the previous DHS datasets. In our database, we removed the batch effects that exist in the data samples collected from multiple centers.

(3) We have employed our replication test that considers multiple samples from a diverse range of cell types in order to check replicability of each DHS site. This results in identifying replicable DHS that present true regulatory sites (Fig. S4), while preserving regulatory sites filtered by more stringent tests such as IDR[9].

In summary, we believe that our open chromatin database, which contains a wide range of cell types can serve as a reference map for regions of open chromatin and will find many applications in studies concerned with uncovering gene regulatory mechanisms of disease and cell development.

Data availability

OCHROdb is a database of open chromatin regions from sequencing data. All data generated are available to download through https://dhs.ccm.sickkids.ca/. Additionally, our interactive open chromatin browser is accessible through the same link. A list of data samples that are analysed during this study are included in the supplementary information files (Table S1).

References

1. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
2. Roadmap Epigenomics Consortium. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
3. Song, L. & Crawford, G. E. DNase-seq: A high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb. Protoc.* **2010**, 5384 (2010).
4. Adams, D. *et al.* BLUEPRINT to decode the epigenetic signature written in blood. *Nat. Biotechnol.* **30**, 224–226 (2012).
5. Mei, S. *et al.* Cistrome data browser: A data portal for ChIP-Seq and chromatin accessibility data in human and mouse. *Nucleic Acids Res.* **45**, D658–D662 (2017).
6. Wang, F. *et al.* ATACdb: A comprehensive human chromatin accessibility database. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gkaa943 (2020).
7. Zeng, X. *et al.* jMOSAiCS: Joint analysis of multiple ChIP-seq datasets. *Genome Biol.* **14**, R38 (2013).
8. Jalili, V., Matteucci, M., Masseroli, M. & Morelli, M. J. Using combined evidence from replicates to evaluate ChIP-seq peaks. *Bioinformatics* **34**, 2338 (2018).
9. Li, Q., Brown, J. B., Huang, H. & Bickel, P. J. Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.* **5**, 1752–1779 (2011).
10. Shooshtari, P., Huang, H. & Cotsapas, C. Integrative genetic and epigenetic analysis uncovers regulatory mechanisms of autoimmune disease. *Am. J. Hum. Genet.* **101**, 75–86 (2017).
11. John, S. *et al.* Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat. Genet.* **43**, 264–268 (2011).
12. Beauparlant, C. J., Lemacon, A. & Droit, A. ENCODExplorer: A compilation of ENCODE metadata. *R Package Version*, Vol. 1 (2015).
13. *Picard Tools—By Broad Institute.* http://broadinstitute.github.io/picard/.
14. McKenna, A. *et al.* The Genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
15. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
16. Van Dongen, S. M. *Graph Clustering by Flow Simulation* (2000).
17. Hogg, R. V. & Tanis, E. A. *Probability and Statistical Inference* Global. (Pearson Education Limited, 2015).
18. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
19. Patsopoulos, N. A. *et al.* Genome-wide meta-analysis identifies novel multiple sclerosis susceptibility loci. *Ann. Neurol.* **70**, 897–912 (2011).
20. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
21. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882–883 (2012).
22. *React—A JavaScript Library for Building User Interfaces.* https://reactjs.org.
23. *Node. js Foundation. Node.js. Node.js.* https://nodejs.org/en/.
24. *Express—Node.js Web Application Framework.* https://expressjs.com.
25. Li, H. Tabix: Fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics* **27**, 718–719 (2011).
26. Jardine, A. Datatables (table plug-in for jquery). *Poslední čtení* **20**, 12 (2012).
27. Buels, R. *et al.* JBrowse: A dynamic web platform for genome visualization and analysis. *Genome Biol.* **17**, 66 (2016).
28. Fullard, J. F. *et al.* An atlas of chromatin accessibility in the adult human brain. *Genome Res.* **28**, 1243–1252 (2018).

Author contributions

P.S. and C.C. have contributed to the conception and design of the work. P.S. and N.H.N. has analyzed the data. P.S., S.F., R.A., V.N. and J.F. contributed to the creation of the website and the new database developed in the work. P.S., M.B. and C.C. contributed to the interpretation of data and have drafted the work.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-26791-x.

Correspondence and requests for materials should be addressed to P.S., M.B. or C.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.