HHS Public Access
Author manuscript
Curr Opin Genet Dev. Author manuscript; available in PMC 2022 April 01.

Published in final edited form as:
Curr Opin Genet Dev. 2021 April ; 67: 103–110. doi:10.1016/j.gde.2020.12.009.

Resources and challenges for integrative analysis of nuclear architecture data

Youngsook L Jung[1], **Koray Kirli**[1], **Burak H Alver**[1], **Peter J Park**[1]

[1]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

Abstract

A large amount of genomic data for profiling three-dimensional genome architecture have accumulated from large-scale consortium projects as well as from individual laboratories. In this review, we summarize recent landmark datasets and collections in the field. We describe the challenges in collection, annotation, and analysis of these data, particularly for integration of sequencing and microscopy data. We introduce efforts from consortia and independent groups to harmonize diverse datasets. As the resolution and throughput of sequencing and imaging technologies continue to increase, more efficient utilization and integration of collected data will be critical for a better understanding of nuclear architecture.

Keywords

epigenomics; Hi-C; data integration; Bioinformatics

Introduction

The rapid pace of technology development in genome and epigenome profiling has led to major advances in our understanding of genome architecture and function. The initial techniques for measuring three-dimensional interactions among genomic loci based on chromosome conformation capture [1–3] have matured in terms of protocol optimization and have led to the development of numerous related techniques, e.g., enriching for interactions with a protein of interest [4,5]. Aided by decreasing sequencing cost, researchers can now produce high-quality data that allow for more sensitive detection of long-range interactions.

In addition to published data from individual laboratories, the US National Institutes of Health (NIH) as well as other governments' agencies have launched consortium efforts to systematically profile epigenomes across many cell lines and tissue types, generating a large

Corresponding author: Peter J Park (peter_park@hms.harvard.edu).

Conflict of interest statement
Nothing declared.

amount of data including 3D interaction data. These data provide an opportunity for researchers to engage in integrative analysis that combines their DNA, RNA, and/or local epigenetic data with publicly available 3D interactions data.

In this review, we will first summarize the resources currently available for those interested in 3D data analysis. Then, we will describe several challenges in collection, curation, and integration of data, as well as steps that can be taken to maximize the value of the data resources for the scientific community. We will focus on nuclear architecture data, but the issues and approaches are also relevant for other data types.

Landmark nuclear architecture datasets

Here, we highlight several datasets that represent key advances in terms of data quality and resolution. For chromosome conformation capture assays, advances in experimental protocols improved the spatial resolution of long-range interactions. The first Hi-C maps with more than a billion reads, using *in situ* Hi-C, was in 2014, providing resolution reaching 1kb and identifying ~10k loops anchored by CTCF [6]. A subsequent dataset with a similar resolution was in mouse, resolving dynamic enhancer-promoter interactions genome-wide during development [7]. Another high-depth dataset was derived using induced human pluripotent stem cells to study differentiation and revealed the role of active HERV-H retrotransposons in demarcating topological domains [8]. Optimized protocols using micrococcal nuclease (MNase) for chromatin digestion have further increased the resolution of genome-wide 3C assays to nucleosome levels [9,10]. These five datasets likely constitute the deepest genome-wide 3C assays generated to date. Some of the data were generated as part of the NIH 4D Nucleome initiative (http://www.4dnucleome.org), and the uniformly curated and processed versions of all five datasets are available at the 4DN Data Portal (http://data.4dnucleome.org).

Whereas genomic assays are typically carried out on millions of cells and therefore capture the average measurements, single cell sequencing technologies provide opportunities to study cell-to-cell variability. Recent single cell Hi-C data include thousands of cells with an average of >120,000 contacts per cell to study chromosomal organization in cell cycle [11] and cells with >1 million contacts per cell for oocyte-to-zygote transition [12] and for modeling of diploid genomes [13].

Major advances in probing the genome architecture have been achieved recently by high-throughput and high-resolution imaging. Low-throughput Fluorescent in situ hybridization (FISH) has been used since the 1990s to map two or three individual genomic DNA elements relative to cellular features like nucleolus and nuclear membrane. Emerging microscopy techniques such as hiFISH [14], OligoDNA-PAINT [15], multiplexed-FISH [16], ChromEMT [17] and OligoFISSEQ [18] enable a detailed view of the chromosome organization in high-throughput. Moreover, limited by the wavelength of light, traditional light microscopy (200–500 nm) is not precise enough to resolve genomic loci within hundreds of kilobases. Super-resolution microscopy methods improve the resolution to 10–30 nm and enable tracing of chromosomes [15]. These imaging data provide a comprehensive picture at the single cell level, highlighting the variability in chromatin

structure between cells. In combination with sequencing-based methods, they can be used to characterize differences across cell types and stages and guide integrative modelling efforts.

Databases for nuclear architecture and epigenomics data

The largest coordinated initiative focusing on 3D genome architecture is the 4D Nucleome Network (the authors are associated with the Data Coordination and Integration Center of this project) [19]. This initiative aims to understand the principles underlying nuclear organization in space and time (hence the '4D'), the role of nuclear organization in gene expression and cellular function, and the impact of changing nuclear organization in various diseases. 4D Nucleome in Phase I (2015–2020) has generated ~1200 replicate experiment sets that span sequencing and imaging datasets. The second phase has just begun in September 2020 and is expected to produce even a greater amount of data. The 4DN Data Portal also hosts 336 replicate experiment sets from landmark publications produced outside the 4DN Network.

Multiple consortia have generated epigenomic data that are highly relevant to understanding nuclear architecture. The largest of these consortia is the Encyclopedia of DNA Elements (ENCODE) project [20]. Now in its Phase 4 (18th year), it has produced more than 10,000 experiments that map regions of transcription, regulatory elements, transcription factor binding sites, chromatin structure, and histone modification. In particular, it also includes 56 Hi-C profiles across 34 human cell lines and tissues. Several dozens of additional Hi-C and ChIA-PET profiles are planned for the current phase of the ENCODE Project. The ENCODE Data Portal also hosts data from related consortia, such as Roadmap Epigenomics [21], model organism ENCODE (modENCODE) [22], and Genomics of Gene Regulation (GGR, https://www.genome.gov/Funded-Programs-Projects/Genomics-of-Gene-Regulation). Roadmap Epigenomics is another notable consortium that generated valuable data, with 111 reference epigenomes (e.g., histone modification, DNA methylation, chromatin accessibility) in a variety of healthy human cells and tissues. These and other epigenomics projects are summarized in Table 1.

There have also been efforts from individual laboratories to collect and uniformly process public epigenomics and nuclear architecture data. One of the successful examples is Cistrome Data Browser (http://cistrome.org/db/) [26], which provides uniformly processed data such as peak calls and profile tracks for >56,000 transcription factor (TF) and histone modification ChIP-seq, DNase-seq, and ATAC-seq datasets in human and mouse, along with quality control metrics and the list of the tools and parameters used. Another related database is SEA (now in version 3, http://sea.edbc.org) [27], which provides super-enhancer calls in 266 cell/tissue types based on ChIP-seq data. Databases that focus on 3D architecture data include 3DIV (https://www.kobic.kr/3div/) [28], which has collected public Hi-C and promoter capture Hi-C data in 80 human cell/tissue types and provides normalized contact matrices and significant interactions, and 3D Genome Browser (http://promoter.bx.psu.edu/hi-c/index.html) [29], which profiled chromatin loops using a machine learning model [30] for all available ENCODE Hi-C data in 56 cell/tissue types.

Data visualization tools

Exploratory analysis of Hi-C or other 3D interaction data typically begins with visual inspection of the interaction matrix, which shows the estimated frequency of interactions between every pair of loci. These datasets are large in size: the minimum number of reads required for a Hi-C experiment in the 4D Nucleome consortium is 600 million (a standard RNA-seq may contain on the order of 10–40 million reads). Thus, a tool that allows visualization of the interaction maps quickly without having to download and process the raw data is beneficial. One recent advance in this area is HiGlass [31], which allows for interactive visualization embedded in a web browser. When the user finds a Hi-C dataset in the 4D Nucleome Data Portal, for instance, the interaction map is already in place, ready to be browsed. The challenge in visualization of the matrix is its sheer size at high resolution: at 10kb resolution, the matrix is 300k bins by 300k bins (3 billion bases in the genome divided by 10k), containing nearly 100 billion entries. To display such a matrix while zooming in and out, HiGlass utilizes a Google Map-like technology. With only the necessary data streaming from the cloud as the user navigates the contact map, this tool obviates the need for downloading the full data. Juicebox [32], is another popular tool that allows users to interactively navigate interaction maps.

While HiGlass and Juicebox primarily focus on Hi-C data, 3D genome browser [29] and WashU epigenome browser [33] can visualize other 3D chromatin interaction data such as ChIA-PET, capture Hi-C, and PLAC-seq. One of the challenges in 3D data visualization is how to efficiently integrate other epigenomic profiles (e.g., ChIP-seq and ATAC-seq). These browsers host a large number of epigenetic profiles from ENCODE and Roadmap Epigenomics so that functional elements can be easily linked with chromatin interactions.

Additional algorithm development is needed to produce more accurate interaction matrices. A standard procedure for generating an interaction matrix fails to account for any copy number variants and translocations in the genome—when there is a copy number gain, for instance, that region will show more interaction counts simply because there is more genetic material to interact. While the effects of copy number variations are partially removed by the canonical matrix balance methods [34], several approaches have been proposed to explicitly account for the copy number changes [35,36]. Alternatively, one could detect copy number variation and translocations directly from Hi-C data [37] and use this information to normalize contact frequencies.

Challenges and best practices in the analysis of chromatin interaction data

To ensure the validity of a study based on chromatin interaction data, evaluation of data quality and reproducibility is essential. In addition to the common statistics on read alignments, several additional measures specific to 3D data are often informative, such as the fraction of valid pairs, the ratio between intra- and inter-chromosomal contacts, and the fraction of short-range compared to long-range interactions [38]. Many Hi-C analysis pipelines, such as HiC-Pro [39], generate similar QC statistics. To access reproducibility between replicates, several 3D data-specific methods [40–42] have been developed, as summarized and evaluated recently [43]. These methods propose similarity measures for

contact maps that perform better than conventional correlation coefficients. Single cell Hi-C data are much noisier, and new methods will be needed [44].

Identification of chromatin structures such as topologically-associated domains (TADs) and chromatin loops is crucial to understanding how spatial organization affects gene regulation. Although TADs are evident in contact maps visually, appearing as blocks with high interaction frequencies within, their accurate delineation remains challenging due to their hierarchical structure and a lack of clear correlation with other features, e.g., TAD boundaries often, but not always, colocalize with insulation proteins. A recent comparison of thirteen algorithms for Hi-C data has found that there is wide variability in the chromatin interactions found among the algorithms, but the TAD detection results were more comparable [45]. For chromatin loops, detection methods typically build a background model of interaction frequencies between two loci and access significance of the observed interaction frequency [46,47]. One of the main issues in methods development for 3D feature detection has been the lack of gold standards; however, integration of data from genomics and new microscopy techniques in the coming years will enable more accurate evaluation and thus better tools.

Opportunities and challenges for data reuse

The key datasets highlighted above and the hundreds of other published datasets present many opportunities for deriving new insights without the need to perform expensive experiments. For instance, a cancer biologist may have found a recurrent non-coding mutation in colorectal cancers that, based on the histone mark H3K27ac or H3K4me1, appears to be in an enhancer region. To identify which genes may be regulated by the enhancer, she could generate her own data. Alternatively, she could first search for a chromatin interaction map in a relevant cell line from a public database.

Such analysis, however, requires several elements. First, a sufficiently large number of samples need to be profiled by the research community so that a researcher is likely to find applicable profiles. In the above example, 3D data from a colorectal cancer tissue would be ideal, but, since it is still difficult to do 3C-based experiments on clinical samples, data from cell lines derived from colorectal cancer could serve as a substitute. Most Hi-C data do not have sufficient resolution for linking an enhancer to a complete set of targets, so capture-based experiments would be ideal. Second, the researcher must be able to find relevant samples. A simple approach may be to find papers of interest and look for datasets associated with the publications; however, this process can be laborious, as tracking down the datasets associated with each paper is time-consuming. To search data repositories, the investigator must know which repository contains relevant data. Currently, there exists no centralized catalog of datasets across repositories—to find an RNA-seq profile of a cancer cell line, for instance, one must visit individual data portals one at a time. There is an effort to coordinate among the NIH Common Fund programs (projects of broad interest jointly funded by multiple NIH institutes), with the goal of providing a common interface that catalogs multiple portals (https://commonfund.nih.gov/dataecosystem). However, building such an interface is not trivial, as the scope and vocabulary used in annotations are heterogeneous. Even when a relevant dataset is found, the investigator *(i)* should be

confident that the data are suitable based on the metadata provided and are of high-quality based on the description of quality control steps and replication; *(ii)* should be able to find the data at the desired level of analysis, e.g., for Hi-C, raw data, contact matrix at multiple resolutions, or gene-level summaries; *(iii)* should be able to explore the data, unless a bioinformatics analyst is available to download, process, visualize, and interpret the data.

To combine datasets from multiple sources, data processing steps and parameters must be examined carefully. For instance, sequenced reads are often aligned to different genome versions. As this discrepancy could lead to significant differences in analysis results, re-processing of the raw data may be necessary. Besides genome versions, there are many other sources of variations, including different data quality standards, experimental protocols and reagents, and algorithms used for features identifications (e.g., gene expression levels, regions of chromatin accessibility, and chromatin compartments). It is not unusual to have dramatically different results depending on the analytical steps (and the parameters used); thus, attention to details is necessary when working with published data, especially when metadata explaining the analytical pipelines are inadequate.

Importance of metadata collection for reproducible science

To take full advantage of the existing data, proper metadata ('data about data') must be available at the repositories. Lack of proper metadata is one of the main factors that hinder reproducibility of published results. To increase scientific rigor and transparency, NIH has implemented policies that emphasize the "FAIR" principle: findability, accessibility, interoperability, and reusability [48]. The idea behind this principle is to encourage data producers and publishers to provide sufficiently rich metadata and persistent identifiers in such a way that the data can be found easily by a potential user and the results can be reproduced.

Although the FAIR principles are reasonable and straightforward, their implementation is complicated by a number of factors. A complete set of metadata necessary for replicating one's analysis is substantially more than what is currently required for data submission to common repositories. For instance, what information does an RNA-seq dataset require to ensure full reproducibility? In addition to clearly annotated raw sequence data (FASTQ or BAM files), several pieces of information are needed: how replicates were handled, what control samples were used, what normalization was used, which version of which aligner was used with which parameters; which transcriptome annotation was used; how gene expression levels were quantified; and how subsequent analyses were performed. The problem is compounded when laboratory experiments must be described. Once collected, the metadata must be stored in a standardized manner—in a 'data model'—so that the information can be organized and searched by other tools.

In many NIH consortia, the primary objective is to generate resources for the scientific community. As such, there is a great deal of emphasis on proper metadata curation: key components of domain-specific metadata are decided in relevant working groups and formed into a metadata model. Whenever possible, the metadata model incorporates existing ontologies; in other cases, at least standardized controlled vocabulary is used [49]. In the 4D

Nucleome project, for instance, a working group consensus for the required metadata information for Hi-C experiment included experiment types, biosample metadata, biosample amount, protocol, enzyme, library preparation methods, enzyme lot number, digestion time, crosslinking method, time and temperature, tagging method, ligation time and temperature, ligation volume, whether biotin is removed, average fragment size, fragment size range, fragmentation method, fragment size selection method, raw FASTQ files, and replicate information. In ENCODE and 4D Nucleome, metadata are stored in the JSON (JavaScript Object Notation) format, organized as objects linked to each other.

Although submission of full metadata is desirable, it presents practical challenges for individual investigators. For most data types, there is no consensus on the required metadata fields or data models, and common repositories such as Gene Expression Omnibus (GEO) and Short Read Archive (SRA) require only very limited metadata with no controlled vocabularies. Investigators are also not incentivized to spend the additional effort necessary to obtain and submit full metadata. As a result, a large proportion of data in public repositories (other than those in consortium data portals) lack sufficiently detailed and structured metadata to allow efficient searching or full reproducibility. Published datasets are a valuable resource for the community, and a greater emphasis on proper metadata curation in the scientific community will enable a more efficient use of the resource.

Collecting imaging data

Collection, curation, and re-analysis of microscopy data present additional challenges to the ones we have outlined above for genomic assays. Whereas sequencing experiments have common data formats (e.g., FASTQ) and common coordinate systems (genome builds), microscopy experiments are diverse in many aspects, including image resolutions, biological sample preparation methods, imaging modalities, and data formats. Imaging experiments are sometimes performed with extensive protocol variations even for the same technique. The signal in the microscopy data may be encoded in the cell line, added by a modification, or transiently activated by a treatment. Microscopy equipment is produced by various manufacturers in numerous models, each with its own hardware and software customizations. Microscopes often come with their own software that stores the data acquisition metadata and output imaging files in proprietary formats.

Capturing all metadata and making them interoperable across different technologies and with genomic experiments require an extensive metadata model and file format standardization. There are multiple community-driven efforts in this direction. The most prominent of these is Bio-Formats by Open Microscopy Environment consortium (OME) [50]. Bio-Formats enables conversion of different microscopy file formats and acquisition metadata to the common formats OME-TIFF and OME-XML. Any data portal or consortium dealing with imaging data from multiple sources would benefit from building on such tools.

In addition to the lack of metadata and data standards, sharing of microscopy experiments is also hindered by the large data volumes. For example, super-resolution microscopy experiments produce terabytes of data per day, making long term storage of raw imaging

files prohibitive. Reproducible methods to reliably summarize these raw data to concise localization formats will make future imaging data sharing more feasible. However, the field has not yet converged on such methods. As an interim solution, the 4DN Network's pilot for microscopy data sharing presents example subsets of the raw microscopy data and focuses on sharing the most interoperable processed results, such as the pixel positions of recorded signals, drift profiles, and constructed 3D models.

Conclusion

In recent years, we have seen major advances in our understanding of nuclear architecture, aided by the increase in the resolution and throughput with which we can probe chromatin organization. An important byproduct of these advances are the high-quality datasets that have been generated. We have highlighted some datasets that provide the highest resolutions of genomic interactions to date. We have described how data portals such as those by 4DN and ENCODE increase the utility of datasets with uniform curation, processing, quality control, and visualization. We provided a perspective on remaining challenges in extending such efforts to the whole field, especially for microscopy data. We hope that with rising awareness for FAIR principles across the whole scientific community and funding agencies, our community will tackle these challenges effectively.

Acknowledgements

This work was supported by the National Institutes of Health (U01CA200059).

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

·· of special interest

···· of outstanding interest

1. Dekker J, Rippe K, Dekker M, Kleckner N: Capturing chromosome conformation. Science (80-) 2002, 295:1306–1311.
2. Lieberman-aiden E, Van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al.: Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science (80-) 2009, 33292:289–294.
3. Nagano T, Lubling Y, Stevens TJ, Schoenfelder S, Yaffe E, Dean W, Laue ED, Tanay A, Fraser P: Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature 2013, 502:59–64. [PubMed: 24067610]
4. Fullwood MJ, Ruan Y: ChIP-based methods for the identification of long-range chromatin interactions. J Cell Biochem 2009, 107:30–39. [PubMed: 19247990]
5. Mumbach MR, Rubin AJ, Flynn RA, Dai C, Khavari PA, Greenleaf WJ, Chang HY: HiChIP: Efficient and sensitive analysis of protein-directed genome architecture. Nat Methods 2016, 13:919–922. [PubMed: 27643841]
6. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, et al.: A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 2014, 159:1665–1680. [PubMed: 25497547]
7. Bonev B, Mendelson Cohen N, Szabo Q, Fritsch L, Papadopoulos GL, Lubling Y, Xu X, Lv X, Hugnot JP, Tanay A, et al.: Multiscale 3D Genome Rewiring during Mouse Neural Development. Cell 2017, 171:557–572.e24. [PubMed: 29053968] * A study of chromosome contacts during

mouse neural differentiation, from ES cell through neural progenitors to cortical neurons. The study includes generation of ultrahigh Hi-C maps at each stage and analyzes dynamic regulatory interactions around polycomb regulated and activating regions.

8. Zhang Y, Li T, Preissl S, Amaral ML, Grinstein JD, Farah EN, Destici E, Qiu Y, Hu R, Lee AY, et al.: Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells. Nat Genet 2019, 51:1380–1388. [PubMed: 31427791]
9. Krietenstein N, Abraham S, Venev SV., Abdennur N, Gibcus J, Hsieh THS, Parsi KM, Yang L, Maehr R, Mirny LA, et al.: Ultrastructural Details of Mammalian Chromosome Architecture. Mol Cell 2020, 78:554–565.e7. [PubMed: 32213324]
10. Hsieh THS, Cattoglio C, Slobodyanyuk E, Hansen AS, Rando OJ, Tjian R, Darzacq X: Resolving the 3D Landscape of Transcription-Linked Mammalian Chromatin Folding. Mol Cell 2020, 78:539–553.e8. [PubMed: 32213323] * This study compares chromatin conformation between transcriptionally active and inactive cells using micro-C. The study resolves the functional interplay between transcription and fine-scale chromatin structure.
11. Nagano T, Lubling Y, Várnai C, Dudley C, Leung W, Baran Y, Mendelson Cohen N, Wingett S, Fraser P, Tanay A: Cell-cycle dynamics of chromosomal organization at single-cell resolution. Nature 2017, 547:61–67. [PubMed: 28682332]
12. Flyamer IM, Gassler J, Imakaev M, Brandão HB, Ulianov SV., Abdennur N, Razin SV., Mirny LA, Tachibana-Konwalski K: Single-nucleus Hi-C reveals unique chromatin reorganization at oocyte-to-zygote transition. Nature 2017, 544:110–114. [PubMed: 28355183]
13. Tan L, Xing D, Chang CH, Li H, Xie XS: Three-dimensional genome structures of single diploid human cells. Science (80-) 2018, 361:924–928.* This study uses transposon-based whole-genome amplification to capture a median of 1M contacts from 17 individual cells. This allows reconstruction of the diploid genome with 20kb resolution, identifying cell type specific structural features in individual cells.
14. Finn EH, Pegoraro G, Brandão HB, Valton AL, Oomen ME, Dekker J, Mirny L, Misteli T: Extensive Heterogeneity and Intrinsic Variation in Spatial Genome Organization. Cell 2019, 176:1502–1515.e10. [PubMed: 30799036]
15. Nir G, Farabella I, Pérez Estrada C, Ebeling CG, Beliveau BJ, Sasaki HM, Lee SH, Nguyen SC, McCole RB, Chattoraj S, et al.: Walking along chromosomes with super-resolution imaging, contact maps, and integrative modeling. PLoS Genet 2018, 14:1–35.
16. Wang S, Su J, Beliveau BJ, Bintu B, Moffitt JR, Wu C: Spatial Organization of Chromatin Domains and Compartments in Single Chromosomes. 2016, 353:598–602.
17. Ou HD, Phan S, Deerinck TJ, Thor A, Ellisman MH, O'Shea CC: ChromEMT: Visualizing 3D chromatin structure and compaction in interphase and mitotic cells. Science (80-) 2017, 357.
18. Nguyen HQ, Chattoraj S, Castillo D, Nguyen SC, Nir G, Lioutas A, Hershberg EA, Martins NMC, Reginato PL, Hannan M, et al.: 3D mapping and accelerated super-resolution imaging of the human genome using in situ sequencing. Nat Methods 2020, 17:822–832. [PubMed: 32719531]
19. Dekker J, Belmont AS, Guttman M, Leshyk VO, Lis JT, Lomvardas S, Mirny LA, O'Shea CC, Park PJ, Ren B, et al.: The 4D nucleome project. Nature 2017, 549:219–226. [PubMed: 28905911] ** This perspective describes the goals and strategies of the 4D Nucleome project as well as the genomic and imaging technologies used in the consortium.
20. Abascal F, Acosta R, Addleman NJ, Adrian J, Afzal V, Aken B, Akiyama JA, Jammal O Al, Amrhein H, Anderson SM, et al.: Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature 2020, 583:699–710. [PubMed: 32728249] ** This paper provides the overview of 6000 human and mouse datasets generated in the latest phase of ENCODE and contains analysis of RNA transcription, chromatin structure and modification, DNA methylation, chromatin looping, and occupancy by transcription factors and RNA-binding proteins.
21. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, et al.: Integrative analysis of 111 reference human epigenomes. Nature 2015, 518:317–329. [PubMed: 25693563]
22. Kharchenko PV, Alekseyenko AA, Schwartz YB, Minoda A, Riddle NC, Ernst J, Sabo PJ, Larschan E, Gorchakov AA, Gu T, et al.: Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. Nature 2011, 471.

23. Stunnenberg HG, Abrignani S, Adams D, de Almeida M, Altucci L, Amin V, Amit I, Antonarakis SE, Aparicio S, Arima T, et al.: The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. Cell 2016, 167:1145–1149. [PubMed: 27863232]
24. Chen L, Ge B, Casale FP, Vasquez L, Kwan T, Garrido-Martín D, Watt S, Yan Y, Kundu K, Ecker S, et al.: Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. Cell 2016, 167:1398–1414.e24. [PubMed: 27863251]
25. Wang T, Pehrsson EC, Purushotham D, Li D, Zhuo X, Zhang B, Lawson HA, Province MA, Krapp C, Lan Y, et al.: The NIEHS TaRGET II Consortium and environmental epigenomics. Nat Biotechnol 2018, 36:225–227. [PubMed: 29509741]
26. Zheng R, Wan C, Mei S, Qin Q, Wu Q, Sun H, Chen CH, Brown M, Zhang X, Meyer CA, et al.: Cistrome Data Browser: Expanded datasets and new tools for gene regulatory analysis. Nucleic Acids Res 2019, 47:D729–D735. [PubMed: 30462313]
27. Chen C, Zhou D, Gu Y, Wang C, Zhang M, Lin X, Xing J, Wang H, Zhang Y: SEA version 3.0: A comprehensive extension and update of the Super-Enhancer archive. Nucleic Acids Res 2020, 48:D198–D203. [PubMed: 31667506]
28. Yang D, Jang I, Choi J, Kim MS, Lee AJ, Kim H, Eom J, Kim D, Jung I, Lee B: 3DIV: A 3D-genome Interaction Viewer and database. Nucleic Acids Res 2018, 46:D52–D57. [PubMed: 29106613]
29. Wang Y, Song F, Zhang B, Zhang L, Xu J, Kuang D, Li D, Choudhary MNK, Li Y, Hu M, et al.: The 3D Genome Browser: A web-based browser for visualizing 3D genome organization and long-range chromatin interactions. Genome Biol 2018, 19:1–12. [PubMed: 29301551]
30. Salameh TJ, Wang X, Song F, Zhang B, Wright SM, Khunsriraksakul C, Ruan Y, Yue F: A supervised learning framework for chromatin loop detection in genome-wide contact maps. Nat Commun 2020, 11:1–12. [PubMed: 31911652]
31. Kerpedjiev P, Abdennur N, Lekschas F, McCallum C, Dinkla K, Strobelt H, Luber JM, Ouellette SB, Azhir A, Kumar N, et al.: HiGlass: Web-based visual exploration and analysis of genome interaction maps. Genome Biol 2018, 19:1–12. [PubMed: 29301551] ** HiGlass is an ultra-fast genome browser for 1D and 2D data with a modern client-server architecture. It features linked views for comparison of multiple Hi-C maps at dynamic resolutions.
32. Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, Aiden EL: Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. Cell Syst 2016, 3:99–101. [PubMed: 27467250]
33. Li D, Hsu S, Purushotham D, Sears RL, Wang T: WashU Epigenome Browser update 2019. Nucleic Acids Res 2019, 47:W158–W165. [PubMed: 31165883]
34. Yaffe E, Tanay A: Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nat Genet 2011, 43:1059–1065. [PubMed: 22001755]
35. Vidal E, le Dily F, Quilez J, Stadhouders R, Cuartero Y, Graf T, Marti-Renom MA, Beato M, Filion GJ: OneD: increasing reproducibility of Hi-C samples with abnormal karyotypes. Nucleic Acids Res 2018, 46:e49. [PubMed: 29394371]
36. Servant N, Varoquaux N, Heard E, Barillot E, Vert JP: Effective normalization for copy number variation in Hi-C data. BMC Bioinformatics 2018, 19:1–16. [PubMed: 29291722]
37. Wang S, Lee S, Chu C, Jain D, Kerpedjiev P, Nelson GM, Walsh JM, Alver BH, Park PJ: HiNT: A computational method for detecting copy number variations and translocations from Hi-C data. Genome Biol 2020, 21:1–15.
38. Lajoie BR, Dekker J, Kaplan N: The Hitchhiker's guide to Hi-C analysis: Practical guidelines. Methods 2015, 72:65–75. [PubMed: 25448293]
39. Servant N, Varoquaux N, Lajoie BR, Viara E, Chen CJ, Vert JP, Heard E, Dekker J, Barillot E: HiC-Pro: An optimized and flexible pipeline for Hi-C data processing. Genome Biol 2015, 16:1–11. [PubMed: 25583448]
40. Yang T, Zhang F, Yardımci GG, Song F, Hardison RC, Noble WS, Yue F, Li Q: HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. Genome Res 2017, 27:1939–1949. [PubMed: 28855260]

41. Yan KK, Yardlmcl GG, Yan C, Noble WS, Gerstein M: HiC-spector: A matrix library for spectral and reproducibility analysis of Hi-C contact maps. Bioinformatics 2017, 33:2199–2201. [PubMed: 28369339]
42. Ursu O, Boley N, Taranova M, Wang YXR, Yardimci GG, Noble WS, Kundaje A: GenomeDISCO: A concordance score for chromosome conformation capture experiments using random walks on contact map graphs. Bioinformatics 2018, 34:2701–2707. [PubMed: 29554289]
43. Yardımcı GG, Ozadam H, Sauria MEG, Ursu O, Yan KK, Yang T, Chakraborty A, Kaul A, Lajoie BR, Song F, et al.: Measuring the reproducibility and quality of Hi-C data. Genome Biol 2019, 20(1):1–19. [PubMed: 30606230]
44. Horton CA, Alver BH, Park PJ: GiniQC: A measure for quantifying noise in single-cell Hi-C data. Bioinformatics 2020, 36:2902–2904. [PubMed: 32003786]
45. Forcato M, Nicoletti C, Pal K, Livi CM, Ferrari F, Bicciato S: Comparison of computational methods for Hi-C data analysis. Nat Methods 2017, 14:679–685. [PubMed: 28604721]
46. Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, Aiden EL: Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. Cell Syst 2016, 3:95–98. [PubMed: 27467249]
47. Kaul A, Bhattacharyya S, Ay F: Identifying statistically significant chromatin contacts from Hi-C data with FitHiC2. Nat Protoc 2020, 15:991–1012. [PubMed: 31980751]
48. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, et al.: Comment: The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016, 3:1–9.** This paper describes the FAIR principle--findability, accessibility, interoperability, and reusability—that should serve to guide data producers and publishers to enhance reusability of published data.
49. Hong EL, Sloan CA, Chan ET, Davidson JM, Malladi VS, Strattan JS, Hitz BC, Gabdank I, Narayanan AK, Ho M, et al.: Principles of metadata organization at the ENCODE data coordination center. Database 2016, 2016:1–10.
50. Linkert M, Rueden CT, Allan C, Burel JM, Moore W, Patterson A, Loranger B, Moore J, Neves C, MacDonald D, et al.: Metadata matters: Access to image data in the real world. J Cell Biol 2010, 189:777–782. [PubMed: 20513764]

Table 1.

Consortium projects that have generated large-scale epigenomics data

Project	Brief description	Years	Data portal	Cell types/tissues	Data types	Ref
ENCODE (the Encyclopedia of DNA Elements)	NIH-funded project to map functional elements primarily for human and later extended to mouse and other model organisms	2003 - current	https://www.encodeproject.org/	In Phases 1–2, it profiled focusing on several human cell lines (including cancer cell lines). Now it contains datasets in more than 200 cell lines. Later phases also included data from healthy tissues. Others include primary cells, in vitro differentiated cells, whole organisms, single cells, organoids, cell-free samples.	>900 ChIP-seq (TF, histone modification), 769 DNase-seq, 200 ATAC-seq, >1000 RNA-seq (sc, total, polyA, sm etc), CAGE, Repli-seq, RRBS etc. ***59 Hi-C (mouse data included), 57 ChIA-PET, 12 5C**	[20]
Roadmap Epigenomics	NIH-funded project for 111 human reference epigenomes primarily from normal, healthy individuals	2008 – 2017	http://www.roadmapepigenomics.org/	It contains datasets in 111 human cell types, tissues from healthy individuals. It also provided several ENCODE cell line data uniformly processed.	>1000 experiment sets consisting of ChIP-seq (histone modification), DNase-seq, RNA-seq, Bisulfite-seq, MeDIP-seq, MRE-seq, RRBS, DGF	[21]
4D Nu cleome	NIH-funded project to map the genome structure and dynamics in space and time mainly for human and mouse	2015 - current	https://data.4dnucleome.org/	Most datasets in cell lines (>1000 datasets from immortalized, stem cell derived or primary cell lines)	***295 Hi-C datasets (in situ, dilution, dnase, capture, micro-c, methyl, sn, sci, single cell, single cell methyl, MC) with other data types including SPRITE, PLAC-seq, ChIA-PET, ChIA-Drop, TCC, and MARGI** ** **275 DNA FISH, 3 Electron Tomography** Although its primary data type is Hi-C, it also includes nonarchitecture data such as ChIP-seq, RNA-seq, 2-stage Repli-seq, ATAC-seq, DamID-seq, NAD-seq, CUT&RUN, TSA-seq, TRIP. In addition, it hosts 359 experiment sets from other projects.	[19]
CEEHRC (Canadian Epigenetics, Environment and Health Research)	Canadian project to generate human reference epigenomes focusing on diseases including cancer, inflammatory,	2012 - current	https://epigenomesportal.ca/ihec/about.html	30 tissues, cell types, cell lines (mostly tissues or cells) in human blood, brain, breast, thyroids, skin, sperm, and tonsils, both from	> 1000 experiment sets in ChIP-seq (histone modification), Bisulfite-seq and RNA-seq	[23]

Project	Brief description	Years	Data portal	Cell types/tissues	Data types	Ref
	cardio-metabolic, and neuropsychiatric diseases			healthy and diseased conditions		
Blueprint epigenome	European project to provide 100 reference epigenomes from healthy and disease individuals, focused on blood diseases	2012 – 2017	http://dcc.blueprint-epigenome.eu/#/home	> 90 tissues and cell types in human blood, bone marrow, thymus, tonsil and liver	2602 datasets in ChIP-seq (histone modification), Bisulfite-seq, RNA-seq, DNase-seq	[24]
IHEC (International Human Epigenome Consortium)	International efforts to generate 1000 human reference epigenomes	2010 - current	https://epigenomesportal.ca/ihec/about.html	human tissues and cells in diseased and normal conditions	> 5000 profiles collected from multiple projects including ENCODE, Roadmap, CEEHRC, Blueprint, AMED-CREST for ChIP-seq (histone modification), Bisulfite-seq, RNA-seq	[23]
TaRGET (Toxicant Exposures and Responses by Genomic and Epigenomic Regulators of Transcription)	NIEHS-funded project for the study of epigenetic changes in toxicant exposure, primarily in human liver and blood tissues	2013 - current	https://data.targetepigenomics.org/	human liver and blood tissues before and after toxican exposure	382 ATAC-seq, 387 RNA-seq	[25]

*3D genomic data,

**Imaging data