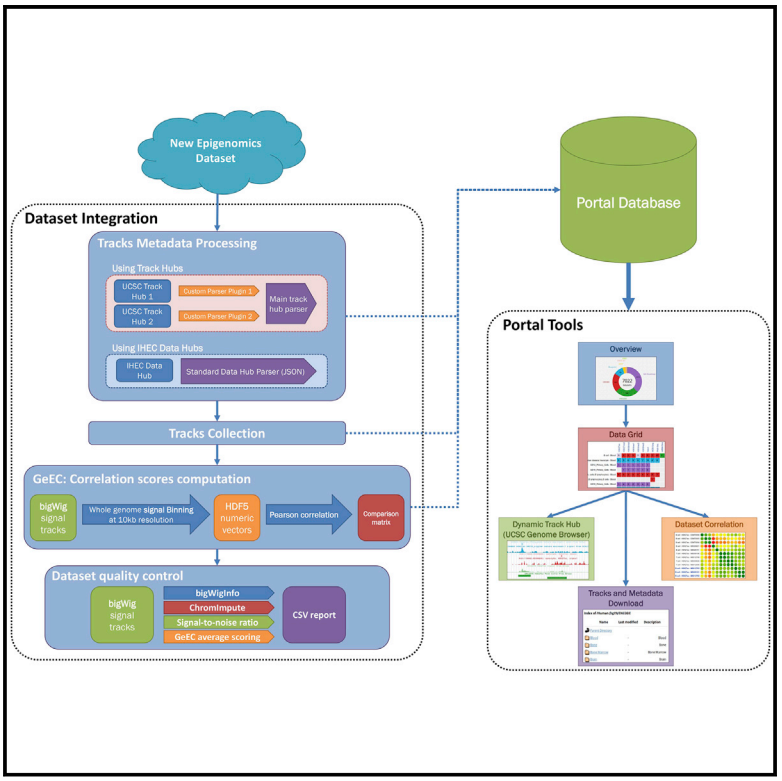
Tool

Cell Systems

The International Human Epigenome Consortium Data Portal

Graphical Abstract



Highlights

- The IHEC Data Portal is a web resource for epigenomic assays data
- Integrates datasets produced by the International Human Epigenome Consortium
- Offers discovery, visualization, and comparison tools
- Enables sharing of sessions among collaborators and for publication purposes

Authors

David Bujold,
David Anderson de Lima Morais,
Carol Gauthier, ..., Alain Veilleux,
Pierre-Étienne Jacques,
Guillaume Bourque

Correspondence

guil.bourque@mcgill.ca

In Brief

The International Human Epigenome Consortium (IHEC) Data Portal is an online resource that enables researchers to discover, integrate, visualize, compare, download, and share epigenomics datasets that have been generated by various international consortia. Explore the Cell Press IHEC website at http://www.cell.com/consortium/IHEC.

Bujold et al., 2016, Cell Systems *3*, 496–499
November 23, 2016
http://dx.doi.org/10.1016/j.cels.2016.10.019

CellPress

Cell Systems
Tool

The International Human Epigenome Consortium Data Portal

David Bujold,[1] David Anderson de Lima Morais,[2] Carol Gauthier,[2] Catherine Côté,[1] Maxime Caron,[1] Tony Kwan,[1] Kuang Chung Chen,[3] Jonathan Laperle,[4] Alexei Nordell Markovits,[5] Tomi Pastinen,[1,6] Bryan Caron,[3] Alain Veilleux,[2] Pierre-Étienne Jacques,[2,4,5,7] and Guillaume Bourque[1,6,8,*]

[1]McGill University and Génome Québec Innovation Center, Montréal, QC H3A 0G1, Canada
[2]Centre de Calcul Scientifique, Université de Sherbrooke, Sherbrooke, QC J1K 2R1, Canada
[3]McGill High Performance Computing Centre, McGill University, Montréal, QC H3C 1K3, Canada
[4]Département d'Informatique, Université de Sherbrooke, Sherbrooke, QC J1K 2R1, Canada
[5]Département de Biologie, Université de Sherbrooke, Sherbrooke, QC J1K 2R1, Canada
[6]Department of Human Genetics, McGill University, Montréal, QC H3A 0G1, Canada
[7]Centre de Recherche du Centre Hospitalier Universitaire de Sherbrooke, Sherbrooke, QC J1H 5N4, Canada
[8]Lead Contact
*Correspondence: guil.bourque@mcgill.ca
http://dx.doi.org/10.1016/j.cels.2016.10.019

SUMMARY

The International Human Epigenome Consortium (IHEC) coordinates the production of reference epigenome maps through the characterization of the regulome, methylome, and transcriptome from a wide range of tissues and cell types. To define conventions ensuring the compatibility of datasets and establish an infrastructure enabling data integration, analysis, and sharing, we developed the IHEC Data Portal (http://epigenomesportal.ca/ihec). The portal provides access to >7,000 reference epigenomic datasets, generated from >600 tissues, which have been contributed by seven international consortia: ENCODE, NIH Roadmap, CEEHRC, Blueprint, DEEP, AMED-CREST, and KNIH. The portal enhances the utility of these reference maps by facilitating the discovery, visualization, analysis, download, and sharing of epigenomics data. The IHEC Data Portal is the official source to navigate through IHEC datasets and represents a strategy for unifying the distributed data produced by international research consortia.

The International Human Epigenome Consortium (IHEC) Data Portal (http://epigenomesportal.ca/ihec) was developed to address the need for data integration across distributed research consortia. IHEC generates and releases reference epigenome maps of normal and disease tissues along with appropriate metadata using a distributed data model (STAR Methods; Table S1). Consortia contributing data through IHEC include ENCODE, NIH Roadmap, CEEHRC, Blueprint, DEEP, AMED-CREST, and KNIH (Table S2). The datasets available in the portal have grown rapidly since early beta versions (Figure S1), and the research community has used it for >15,000 sessions from >100 different countries (Figure S2). As a result, the portal has become the official platform to navigate the reference epigenomic datasets generated by the consortium.

IHEC members follow the principles described in the Fort Lauderdale Guidelines (Fort Lauderdale Guidelines, 2003) and in the Toronto Statement (Birney et al., 2009), which specify that publically funded studies should make their datasets available to the community as soon as they become available. Datasets generated by different groups are downloaded regularly from their respective public site and added into a relational database, central to all the features of the portal.

The IHEC Data Portal allows epigenomic data integration, discovery, visualization, analysis, download, and sharing. Previously, other solutions had been developed to enable some of these features. For instance, the ENCODE (ENCODE Project Consortium, 2004) and the NIH Roadmap (Bernstein et al., 2010) made commendable efforts to standardize protocols, enforce systematic data submission, and build resources to facilitate data discovery and visualization (Table S3). However, most of these solutions are not directly applicable to IHEC, where the datasets are generated by various groups funded by different organizations that are working asynchronously and whose data are archived at different locations.

Data Integration

The portal relies on IHEC Data Hub JSON documents for retrieval and distribution of consortium data. This format is well tailored for information exchange across IHEC members and uses a specification that is extensible to include all metadata available on published datasets (IHEC Ecosystem GitHub, 2016). Data Hub documents also contain links to public data tracks in bigBed and bigWig formats, which are binary indexed files of the sequencing analysis results (Kent et al., 2010). The portal initially relied on member consortia publishing their datasets using UCSC Genome Browser track hubs (Raney et al., 2014). While this format can still be used to populate the IHEC Data Portal database, it was primarily intended for display in the UCSC Genome Browser (Kent et al., 2002), which made it more difficult to organize metadata across IHEC members. Extraction and integration of new datasets in the portal are done biannually, and previous versions of datasets are always archived for future referral.

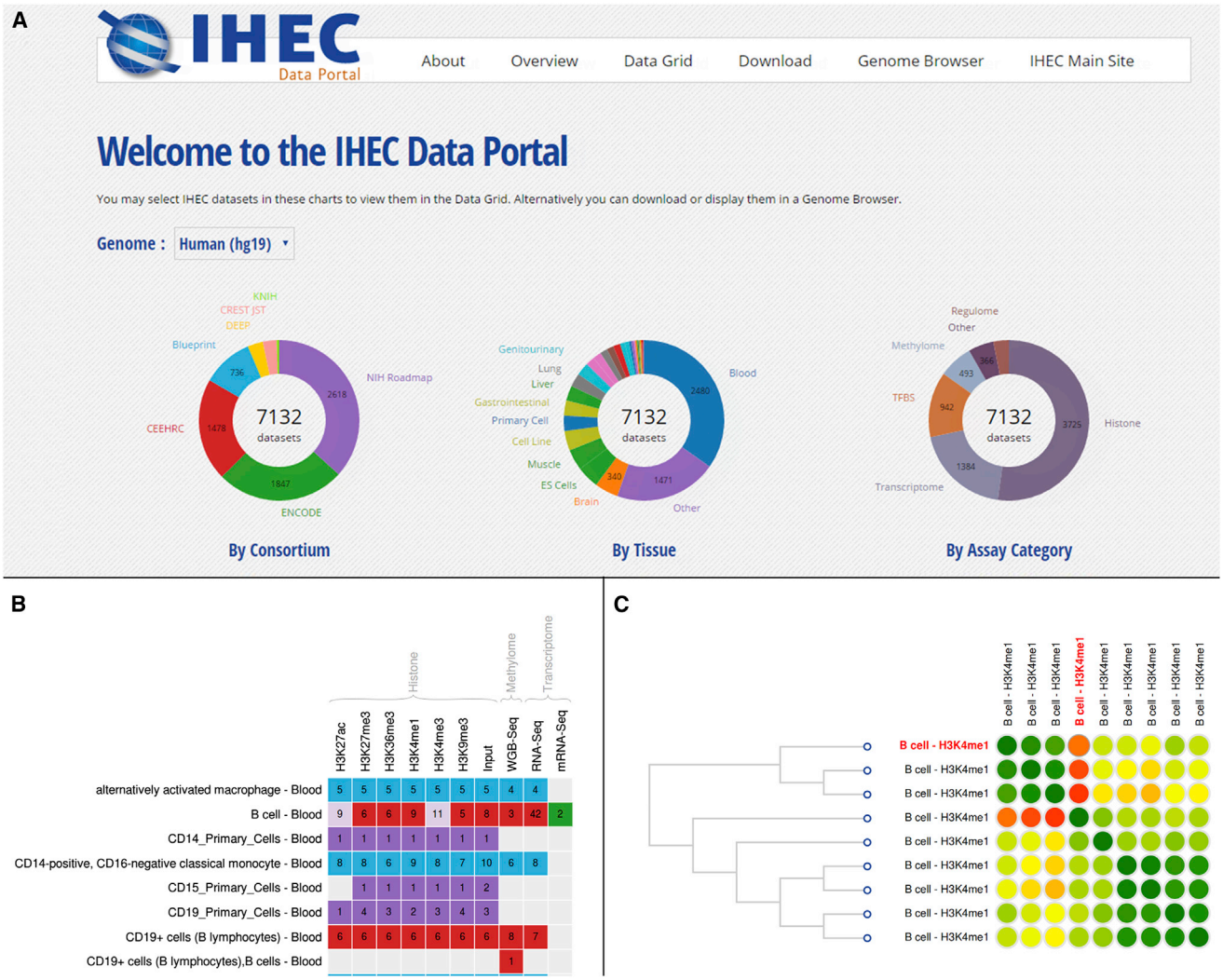


Figure 1. IHEC Data Portal

(A) The portal overview page displays dataset counts split by consortium, tissue, and assay types.

(B) A customizable grid shows available datasets produced by all IHEC consortium members, matching user-specified filtering criteria. These datasets are grouped and displayed by cell type in rows, assay in columns, and publishing consortium in colors; the grid cell number indicates how many datasets are available by combination. The picture shown displays datasets for IHEC core assays in the blood cell type category.

(C) The correlation tool facilitates identification and removal of outlier datasets from a user's tracks selection, allowing the comparison of samples in a correlation matrix using Pearson correlation coefficients. It also allows characterization of dataset group similarity based on attributes such as cell types and assays. The figure shows the correlation tool that clusters available datasets, generated by different groups, for the "B Cell" cell type.

Data Discovery

The portal implements a relational database to power its features through a dynamic web interface and an online application programming interface (API) (STAR Methods). It currently provides access to >7,000 epigenomic datasets from >600 tissues. The front page of the portal provides an overview of the datasets currently available using wheels (Figure 1A) and an interactive Data Grid (Figure 1B) for selecting and navigating through datasets by cell type, assay, and publishing consortium.

Data Visualization

User-selected datasets can be inspected without downloading them by creating a dynamic UCSC Genome Browser track hub that can be visualized on a local mirror of the UCSC Browser or on any other browser instance, such as the main UCSC server (Figure S3). Other visualization tools supporting the track hub format, such as Ensembl (Herrero et al., 2016), can also be used through the provided dynamic track hub link.

Data Analysis

The portal provides a correlation tool making it possible to compare selected datasets (Figure 1C). Once datasets are selected, the "Correlate datasets" button will display a clustered heatmap for assessing similarity based on correlation scores. This functionality facilitates the identification and removal of outliers from user-selected datasets using the interactive Data Grid. The scores are precomputed using the Genomic Efficient Correlator (GeEC) tool by averaging track signal in 10 kb bins

Table 1. Main Features of the IHEC Data Portal

Feature	Component	Description
Distributed data submission[a]	integration	IHEC Data Hub documents allow IHEC members to automate submission of new datasets and metadata to the Portal database.
Data overview[a]	discovery	Pie charts display available datasets available by producing consortia, tissues and assays.
Dynamic data grid	discovery	A dynamic grid displays available datasets for the chosen assembly with extensive filtering and selection tools.
Export track hubs	visualization	Track hubs are generated dynamically to enable visualization in the UCSC Genome Browser, Ensembl, or other browsers.
Datasets correlation[a]	analysis	A heatmap and dendrogram display clustering of selected datasets in the grid based on signal similarity to allow detection of outliers.
Download tracks	download	All public tracks published by IHEC members are downloadable directly from a browser or using a batch download tool such as *wget*.
Export metadata	download	JSON documents that follow the IHEC Data Hub specification can be generated with the metadata for the datasets selected.
Permanent sessions	sharing	Generates a permanent session accession ID that can be shared with collaborators or used in publications.
Session overview[a]	sharing	Session report with information on metadata and original location of public tracks and raw data at controlled access repositories.

[a]Features are unique to the IHEC Data Portal as compared to other epigenomic data portals.

spanning the whole genome and computing Pearson correlation coefficients between all datasets (STAR Methods; Breeze et al., 2016). This method is reapplied every time new datasets are added to the portal.

Data Download

Data retrieval is offered through a download tool providing a directory-like structure of tracks for selected datasets that allows individual file download and also serves a plain text list of track URLs to be used with a batch download tool, such as *wget* (Figure S4). Because the tracks displayed in this section are the result of downstream analysis tools, they do not contain identifiable information (STAR Methods). In contrast, the raw data files for most of the IHEC datasets are located in controlled access repositories, such as EGA (Leinonen et al., 2011) or dbGap (Tryka et al., 2014), and obtaining them requires a data access request to be sent to the appropriate consortium. Nevertheless, the portal provides links to these raw files to facilitate locating them. Finally, users can also obtain metadata on samples, experiments, and analyses by using a metadata download Web API or by clicking the "Get metadata" button below the data grid. Parameters such as the assembly and the publishing consortium can also be used when invoking the API directly. The metadata properties are exported using a JSON document that follows the IHEC Data Hub specification (IHEC Ecosystem GitHub, 2016).

Data Sharing

The portal provides users with a session ID and URL that allow dataset selection and filtering options to be reused in any web browser. The session report page also makes it possible to obtain the full listing of datasets used in a given analysis project with their sources, location of raw data at controlled access repositories, and the rest of the metadata. As tracks are permanently stored on the portal's server, these sessions are more reliable than using track hubs from each of the IHEC members. Once generated, portal sessions are unmodifiable and permanently accessible and can therefore be referred to in journal publications that used IHEC data.

Several features separate the portal from other epigenomic online resources (Table 1; Figure S5). For instance, unlike most portals, where data were processed through a single data coordination center, IHEC data are analyzed and released through several different consortia. These consortia use their own servers to distribute their data with varying levels of reliability and different analysis methods and data storage standards. This was one of the major challenges that the IHEC Data Portal had to address to organize the data of the consortium in a unified way.

The development of the portal is ongoing, and additional features are gradually being added, including more tools to assess the quality of datasets. Another planned improvement is to enrich the metadata querying and exportation features from the database using the Web API, allowing this metadata to be fetched manually from a web browser or programmatically. The querying system will enable obtaining metadata reports based on various filtering criteria rather than just on grid dataset selection.

The IHEC Data Portal is being built as a comprehensive discovery tool to enable the research community to share epigenomic data and collaborate more effectively. Establishing at an early stage the assay standards and metadata formats was paramount to the success of the project. More generally, the strategy that was implemented in the portal, to overcome the difficulties of integrating data coming from different sources, could potentially be adapted to unify and distribute data produced by other international research consortia.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING

SUPPLEMENTAL INFORMATION

Supplemental Information includes five figures, three tables, and supplemental text and can be found with this article online at http://dx.doi.org/10.1016/j.cels.2016.10.019.

AUTHOR CONTRIBUTIONS

D.B., D.A.L.M., C.G., and C.C. worked on the design and implementation of the portal and the UCSC Genome Browser mirror. D.B., D.A.L.M., and M.C. worked on the portal's data and metadata management. T.K. and T.P. provided input at various steps of portal development. J.L., A.N.M., and P.E.J. developed the datasets correlation approach. C.G., D.M., K.C.C., B.C., and A.V. provided network and hardware infrastructure support on Compute Canada HPC. D.B. and G.B. wrote the manuscript. G.B. provided overall direction for the project.

ACKNOWLEDGMENTS

The IHEC Data Portal is a service offered by the Canadian Center for Computational Genomics (C3G) and was developed by the McGill Epigenomics Data Coordination Centre (EDCC). It is funded under the Canadian Epigenetics, Environment, and Health Research Consortium (CEEHRC) by the Canadian Institutes of Health Research (CIHR-EP2-120609) and by Genome Québec, with additional support from Genome Canada. The correlation matrix functionalities have been developed at the Université de Sherbrooke and funded by the Natural Sciences and Engineering Research Council of Canada (NSERC-435710–2013). The computing and networking infrastructure, and part of the software development, are provided by Calcul Québec, Compute Canada, and CANARIE. Useful suggestions at various stages of development were made by Laura Clarke, David Richardson, and Paul Flicek from the Blueprint team at EBI. We would also like to thank the IHEC Scientific Steering Committee for supporting the project.

Received: January 19, 2016
Revised: July 21, 2016
Accepted: October 19, 2016
Published: November 15, 2016

REFERENCES

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

Birney, E., Hudson, T.J., Green, E.D., Gunter, C., Eddy, S., Rogers, J., Harris, J.R., Ehrlich, S.D., Apweiler, R., Austin, C.P., et al.; Toronto International Data Release Workshop Authors (2009). Prepublication data sharing. Nature *461*, 168–170.

Bostock, M., Ogievetsky, V., and Heer, J. (2011). D^3: data-driven documents. IEEE Trans. Vis. Comput. Graph. *17*, 2301–2309.

Breeze, C.E., Paul, D.S., van Dongen, J., Butcher, L.M., Ambrose, J.C., Barrett, J.E., Lowe, R., Rakyan, V.K., Iotchkova, V., Frontini, M., et al. (2016). eFORGE: a tool for identifying cell type-specific signal in epigenomic data. Cell Rep., in press.

ENCODE Project Consortium (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. Science *306*, 636–640.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Herrero, J., Muffato, M., Beal, K., Fitzgerald, S., Gordon, L., Pignatelli, M., Vilella, A.J., Searle, S.M., Amode, R., Brent, S., et al. (2016). Ensembl comparative genomics resources. Database (Oxford) *2016*, baw053.

IHEC Ecosystem GitHub (2016). IHEC Data Hub specification: GitHub source control repository. https://github.com/IHEC/ihec-ecosystems.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res. *12*, 996–1006.

Kent, W.J., Zweig, A.S., Barber, G., Hinrichs, A.S., and Karolchik, D. (2010). BigWig and BigBed: enabling browsing of large distributed datasets. Bioinformatics *26*, 2204–2207.

Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tárraga, A., Cheng, Y., Cleland, I., Faruque, N., Goodgame, N., Gibson, R., et al. (2011). The European Nucleotide Archive. Nucleic Acids Res. *39*, D28–D31.

National Information Standards Organization (2004). Understanding Metadata (NISO Press).

Raney, B.J., Dreszer, T.R., Barber, G.P., Clawson, H., Fujita, P.A., Wang, T., Nguyen, N., Paten, B., Zweig, A.S., Karolchik, D., and Kent, W.J. (2014). Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. Bioinformatics *30*, 1003–1005.

Fort Lauderdale Guidelines (2003). Sharing data from large-scale biological research projects: a system of tripartite responsibility. https://wellcome.ac.uk/sites/default/files/wtd003207.pdf.

Tryka, K.A., Hao, L., Sturcke, A., Jin, Y., Wang, Z.Y., Ziyabari, L., Lee, M., Popova, N., Sharopova, N., Kimura, M., and Feolo, M. (2014). NCBI's Database of Genotypes and Phenotypes: dbGaP. Nucleic Acids Res. *42*, D975–D979.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
IHEC publicly-accessible dataset tracks	Respective IHEC member producing consortia	N/A
human reference genome NCBI build 37, GRCh37	Genome Reference Consortium	http://www.ncbi.nlm.nih.gov/grc/human
human reference genome NCBI build 38, GRCh38	Genome Reference Consortium	https://www.ncbi.nlm.nih.gov/grc/human
mouse reference genome NCBI build 10, GRCm38	Genome Reference Consortium	http://www.ncbi.nlm.nih.gov/grc/mouse
Software and Algorithms		
UCSC Genome Browser	UCSC Genome Bioinformatics	http://genome.ucsc.edu/
GeEC	P.-E.J., unpublished data	N/A

CONTACT FOR REAGENT AND RESOURCE SHARING

As Lead Contact, Guillaume Bourque is responsible for all reagent and resource requests. Please contact Guillaume Bourque at guil.bourque@mcgill.ca with requests and inquiries.

METHOD DETAILS

Implementation

The IHEC Data Portal was implemented using Perl CGI and Python in the back-end, and Javascript with the D3.js and jQuery libraries for the front-end. Code examples from D3 online gallery were especially useful in early development of the Portal (Bostock et al., 2011). The source code to DIGViewer, the Javascript data grid module developed in-house, is freely available under the MIT license, at https://bitbucket.org/genap/digviewer. An introductory video to the IHEC Data Portal is available at the following address: https://www.youtube.com/watch?v=5_oW5_uVgt8.

Testing

Testing procedures for the features of the IHEC Data Portal are available in Document S1 "Testing procedures."

What Constitutes a Complete IHEC Dataset

The IHEC consortium aims at making available a set of full reference epigenomes, by defining a core group of assays that should be run on a given sample. A reference epigenome that meets IHEC requirements is composed, at minimum, datasets for whole-genome bisulfite sequencing, RNA-Seq and ChIP-Seq for 6 histone marks (H3K27me3, H3K36me3, H3K4me1, H3K4me3, H3K27ac, H3K9me3) with input. Archiving complete and accurate datasets for these experiments is the first step toward ensuring that the produced data is useful to the research community. Such datasets include sensitive and non-sensitive data, and well-defined accompanying metadata.

IHEC Metadata

Well-defined metadata is an essential but often neglected component of published datasets. The National Information Standards Organization (NISO) describes metadata as "data about data" or "information about information," highlighting the importance of well-defined metadata: "Metadata is key to ensuring that resources will survive and continue to be accessible into the future" (National Information Standards Organization, 2004). This is especially important in large scale projects such as IHEC, which aims at producing datasets that will be useful not only for the producing data center, but for the scientific community at large. There are multiple reasons for well-defined metadata, including: 1) resource discovery, allowing data to be searched and located using criteria that are meaningful to the user and 2) interoperability, to ensure that data will keep its usefulness not only in its originating system, but on other systems as well. A practical example of this is to use ontologies and controlled vocabulary to define the realm of possible values for a specific field. 3) Archiving and preservation ensures that as technology evolves, well-kept metadata describing a dataset content will ensure the survival of this dataset, giving information on its organization, formats, etc.

Controlled Access versus Publicly Accessible Components

Multiple components constitute a complete epigenomic dataset, and these can be split into two categories: controlled access data and publicly accessible data. Controlled access data includes data and metadata that is considered as personally identifiable. This includes raw data coming from sequencers and clinical information that is considered confidential. Such data is usually archived at controlled access repositories such as EGA and dbGaP. Publicly accessible data describes annotation tracks produced by downstream analysis, such as signal tracks and peak calls in the case of ChIP-Seq experiments, and non-confidential metadata such as library, experiment and analysis metadata, as well as portions of sample metadata. This freely downloadable data is made available through the IHEC Data Portal.

QUANTIFICATION AND STATISTICAL ANALYSIS

Correlation of Datasets

Using the Genomic Efficient Correlator (GeEC) tool (Breeze et al., 2016), correlation scores in the Portal are pre-computed using the local copy of signal tracks for transcriptome and histone modification experiments, and the CpG sites methylation ratio for methylome experiments. Briefly, the signal is averaged in bins of 10 kb covering the whole genome, excluding the ENCODE blacklisted regions (ENCODE Project Consortium, 2012), and Pearson correlation coefficients are calculated between all datasets. This method was validated by computing the correlation matrix for the whole IHEC dataset, which demonstrated that even when strictly using publicly released bigWig signal tracks, datasets correlated primarily on assays and tissue types, before correlating on artifacts such as data producing consortium (Breeze et al., 2016). Note that once released, the GeEC tool will also allow users to submit and compare their data with the IHEC datasets.

DATA AND SOFTWARE AVAILABILITY

The IHEC Data Portal set of tools is entirely available from the main portal page, at http://epigenomesportal.ca/ihec. The documentation to use the public API is available here: http://epigenomesportal.ca/ihec/doc/using_the_metadata_api.html.

ADDITIONAL RESOURCES

IHEC Main Website: http://ihec-epigenomes.org/