Published in final edited form as:
Cell Syst. 2016 November 23; 3(5): 491–495.e5. doi:10.1016/j.cels.2016.10.021.

The BLUEPRINT Data Analysis Portal

José María Fernández[1,2], **Victor de la Torre**[1,2], **David Richardson**[3], **Romina Royo**[2,4], **Montserrat Puiggròs**[4], **Valentí Moncunill**[4], **Stamatina Fragkogianni**[4], **Laura Clarke**[3], **Paul Flicek**[3], **Daniel Rico**[1,5], **David Torrents**[4,6], **Enrique Carrillo de Santa Pau**[1,*], and **Alfonso Valencia**[1,2,*] **on behalf The BLUEPRINT consortium**

[1]Structural Biology and BioComputing Programme, Spanish National Cancer Research Centre (CNIO), Madrid, Spain

[2]Spanish National Bioinformatics Institute (INB), Spain

[3]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK

[4]BSC - CRG - IRB. Barcelona Supercomputing Center (BSC). Joint BSC-CRG-IRB. Research Program in Computational Biology

[6]Institució Catalana de Recerca i Estudis Avançats (ICREA)

Summary

The impact of large and complex epigenomic datasets on biological insights or clinical applications is limited by the lack of accessibility by easy, intuitive, and fast tools. Here we describe epigenomics comparative cyber-infrastructure (EPICO), an open-access reference set of libraries to develop comparative epigenomic data portals. Using EPICO, large epigenome projects can make available their rich datasets to the community without requiring specific technical skills. As a first instance of EPICO, we implemented the BLUEPRINT Data Analysis Portal (BDAP). BDAP provides a desktop for the comparative analysis of epigenomes of hematopoietic cell types based on results, such as the position of epigenetic features, from basic analysis pipelines. The BDAP interface facilitates interactive exploration of genomic regions, genes, and pathways in the context of differentiation of hematopoietic lineages. This work represents initial steps toward broadly accessible integrative analysis of epigenomic data across international consortia. EPICO can be accessed at https://github.com/inab and BDAP at http://blueprint-data.bsc.es.

[*]Corresponding author: ecarrillo@cnio.es, valencia@cnio.es (Lead contact).
[5]present address: Institute of Cellular Medicine, Newcastle University, Newcastle upon Tyne NE2 4HH, United Kingdom

Author Contributions

J. M. Fernández designed and wrote the BP-Schema-tools, BP-DCC-model, EPICO-REST-API, EPICO-data-loading-scripts and most of BDAP web client, as well as the technical part of the manuscript.
V. de la Torre designed and wrote the first prototype of BDAP web client.
D. Richardson and L. Clarke provided insightful discussions and advice about the data model behind BDAP, as well as feedback on the different BDAP releases.
R. Royo, M. Puiggròs, V. Moncunill and S. Fragkogianni have been responsible for the different public releases of BDAP at BSC, synchronized with BLUEPRINT data releases, and they provided feedback on the data model.
E. Carrillo de Santa Pau and D. Rico prepared the practical cases, beta-tested BDAP web client and suggested improvements, as well as coordinating the preparation of the manuscript.
E. Carrillo de Santa Pau coordinated the communication with the journal and the response to the reviewers, as well as the different manuscript versions.
P. Flicek, D. Torrens and A. Valencia coordinated the project.

Europe PMC Funders Author Manuscripts

The International Human Epigenome Consortium (IHEC; IHEC, 2016) coordinates standards for the production, distribution, and accessibility of reference epigenomes generated by several large consortia, including BLUEPRINT (Adams et al., 2012), CEMT (CEMT, 2016), CREST (CREST, 2016), DEEP (DEEP, 2016), ENCODE (ENCODE, 2016), CEEHRC (CEEHRC, 2016) and NIH ROADMAP (ROADMAP, 2016). Each consortium keeps its original data on their own Data Coordination Centre (DCC) portal, which provides some additional analysis results (for example, chromatin state or intron retention) in different formats (text, BED, BigWig for the raw signal, and BigBed for regions highly enriched in raw signal) and metadata (Table S1). Moreover, some consortia provide genome browsers to visualize and compare the primary data (Table S1).

Nevertheless, additional bioinformatics skills are needed to identify, download, process, and analyze the large and complex data sets (Table S1). Indeed, the majority of potential users interested in epigenomic datasets, including most biologists and physicians, are not able to exploit the data satisfactorily. Therefore, novel efficient exploration tools to quickly test biological hypotheses are needed.

Here we describe an Epigenomics Comparative cyber-infrastructure (EPICO; https://github.com/inab) to facilitate the production of user-friendly interactive web portals, and a portal for BLUEPRINT Data (http://blueprint-data.bsc.es) implemented using EPICO. The BLUEPRINT Consortium is a flagship European project that aims to provide reference epigenomes from hematopoietic cell lineages (Adams et al., 2012). Portals created with EPICO enable the comparison of the epigenetic structure of different cell types and related diseases.

The EPICO platform is based on five components: (i) a data model (EPICO-data-model, 2016); (ii) data validation and loading programs (EPICO-data-loading-scripts, 2016), which must be adapted to the data and metadata acquisition of the particular project; (iii) an empty database which will store all the data and metadata produced by the data validation and loading programs; (iv) the EPICO REST API (EPICO-REST-API, 2016), which implements the queries to the database, providing a programmatic access; and (v) the data analysis portal itself (BP-Data_analysis, 2016), which queries the databases through the EPICO REST API (Figure 1).

The minimum infrastructure needed to generate epigenomic data portals with EPICO are the five components described above, storage space to create the database, a connection to fetch the primary data to be integrated into the database, and the modules to receive the queries over the stored data and send the results to be visualized. A detailed technical description of the EPICO components and the instructions to create custom data portals are provided at https://github.com/inab/epico-data-analysis-portal/wiki.

The epigenomic information displayed by BDAP is based on data obtained from ChIP-Seq experiments (H3K4me3, H3K4me1, H3K27ac, H3K36me3, H3K27me3, H3K9me3 or H2A.Zac); DNaseI-Seq; WGBS (whole-genome bisulfite sequencing of hypo and hyper-methylated regions); and RNA-Seq (at the gene or transcript level). As of the 2016-08 data release, the platform contains the analysis of 2,757 products from 2,558 experiments,

involving 487 donors, 11 pool donors and 7 cell lines, and summarizing a total of 62 different cell types that cover 17 diseases. The BDAP allows users to visualize and compare all the epigenomic and transcriptomic data for blood cell types of interest. This query is performed following the three-step process implemented in EPICO (STAR Methods). We illustrate its utility by analyzing two genes as examples, Formyl Peptide Receptor 1 (FPR1; Figures S1 and S2) and interferon regulatory factor 8 (IRF8; Figures S3 and S4).

FPR1 is one of the most extensively studied G protein-coupled receptors involved in neutrophil chemotaxis (reviewed by Ye et al., 2009), We used BDAP to explore in which phase of neutrophil differentiation FPR1 expression is regulated (STAR Methods). A gene expression box-plot shows a clear increase in FPR1 expression as neutrophil differentiation progresses from the neutrophilic myelocyte to the mature neutrophil, and more modest increase for FPR2 (Figures S2A and S2B). The cell types with the strongest expression were the segmented neutrophils of the bone marrow and the mature neutrophils. In addition, the ChIP-Seq data revealed active histone modifications in the start codon and transcribed regions of the principal isoforms for FPR1 and FPR2 in the cell types with the strongest expression (Figures S2C and S2D). These results suggest that the chemotactic properties associated with FPR1 and FPR2 are acquired during neutrophil differentiation, and they reach their peak in the segmented neutrophils of the bone marrow.

The interferon regulatory factor 8 (IRF8) has been identified as a key transcription factor that regulates myeloid cell production. IRF8 maintains the balance between monocytes and neutrophils, and a lack of this gene increases the number of neutrophils and diminishes the monocyte population (Kurotaki et al., 2014). We used BDAP to explore the differences between neutrophils and monocytes at the transcriptome and epigenome levels (STAR Methods). A gene expression box-plots clearly show that IRF8 is expressed in classical monocytes and macrophages, whereas neutrophils do not express this transcription factor (Figure S4B). In addition, differences in the transcripts expressed are observed (Figure S4A). Moreover, peaks of active histone modifications H3K27ac were observed along the gene body and at the start codon in the two cell types that express IRF8 (Figure S4D), while in neutrophilic myelocytes repressive histone modifications were situated in the region around the start codon (H3K27me3 and H3K9me3, Figures S4C and S4E). These results confirmed previous observations suggesting that IRF8 is a marker of the macrophage lineage (Kurotaki et al., 2014).

Data portals facilitate access to different data analyses, by reducing the need to deal with issues related to the different formats in which data are stored. Unfortunately, the current lack of standards for data and metadata in epigenomics limits the possibility of developing a single portal to compare data of the different IHEC consortia. Given this situation, we propose the use of EPICO to create project-specific data analysis portals. EPICO includes a common template and standards for the description of data and metadata.

The BLUEPRINT Data Analysis Portal complements the BLUEPRINT-DCC portal (DCC_portal, 2016) by providing facilities to analyze multiple epigenetic data types at once, for instance DNA methylation and histone marks, and to deal with multiple samples from different cell types, rather than dealing with individual samples and specific data types.

Europe PMC Funders Author Manuscripts

Moreover, BDAP answers queries about specific genomic regions, genes, or pathways, providing summary statistics and comparative analysis grouping samples by cell type or tissue of origin. BDAP is accessible to the many biologists and doctors without programmatic or technical skills. This is different from other solutions, such as the one recently proposed by DeepBlue (Albrecht et al., 2016; Table S1). BDAP is the first platform generated with EPICO. The main condition for generating a portal is that data and metadata have to be converted to EPICO standard format (STAR Methods). In the future, additional efforts will have to be made to homogenize across IHEC experimental procedures, quality controls, and primary data analysis workflows, a situation reminiscent of the current developments in other large scale consortia, like ICGC (ICGC, 2016) or ENCODE (ENCODE, 2016).

Future improvements to our platform will include the integration of our visualization tools (e.g., box-plots or scatter plots) and results (e.g., consensus peaks) with standard genomic browsers, such as EnsEMBL (Flicek et al., 2014) or UCSC Genome Browser (Kent et al., 2002). In summary, EPICO provides the infrastructure and a standard template to create powerful tools that brings complex epigenomic data to the hands of researchers that want to test biological hypotheses, as shown in the two use cases of the BDAP implementation of EPICO fed with BLUEPRINT primary data.

STAR Methods

Contact for Reagent and Resource Sharing

Alfonso Valencia contact: valencia@cnio.es

Experimental Model and Subject Details

The experiments, datasets, and primary analysis that support the BLUEPRINT Data Analysis Portal are available at http://dcc.blueprint-epigenome.eu/#/home and http://www.blueprint-epigenome.eu/

Method Details

EPICO cyber-infrastructure description

The storage and querying of epigenomic data poses significant challenges to analysis portals. The main problem is that data are semi-structured and not fully organized. The analysis pipelines uses as input the results obtained by the primary analysis done by the consortium from large epigenomic experiments, ChIP-Seq histone peaks, consolidated methylated regions and gene/transcript expression values. All this information is difficult to index with traditional database due to both the size and nature of the datasets.

EPICO requires a file index with the metadata, including donor, specimen (blood or other tissue types), sample identifiers, status (healthy or disease), cellular type (referred to specimen i.e. neutrophil, monocyte, etc) and the paths to access the files with the results from the analyses. The EGA/IHEC XML files by sample and experiment are required, with a description of the sample identifier, information about origin, a description about the experiment type (i.e. ChIP-Seq, WGBS or RNA-Seq) and the type of analysis performed.

EPICO data model includes the necessary sample tracking metadata (donors, specimens and samples), along with the details of the experiments performed (i.e. chromatin accessibility, WGBS, MeDIP-Seq, ChIP-Seq, mRNA-Seq and others). The results of the primary analysis pipelines have their analysis identifiers (IDs), their corresponding metrics (z-score, -$\log_{10}$ q-value, FPKM and methylation levels) and their genomic locations. These results may be a genomic region, the EnsEMBL gene ID or the EnsEMBL transcript ID with its associated metrics. Each result is mapped to its physical genomic region and linked to the corresponding metadata, such that EPICO web client can follow the path from the consolidated methylated regions, regulatory regions, expression, or histone peaks to the samples or donors through the analysis and experiments.

EPICO platform relies on a NoSQL database infrastructure to handle large volumes of semi-structured data to be stored. The EPICO data model validation is a key step in cases of unstructured data usually associated to the insertions in databases (i.e. requiring strict types, range values restriction, check valid values against a controlled vocabulary among others). We have developed the EPICO infrastructure, which take into account both the EPICO conceptual model and the physical database technology, applying concepts from object oriented programming and extended entity relationship (EER) model (Chen, 1976; Codd, 1979). The bridge data model describes the concepts, specifications and restrictions that must be validated before storing the results from the analysis pipelines. Both the results and metadata are stored in a NoSQL database instance according to the definitions and restrictions of the data model (for instance, controlled vocabularies and ontologies). EPICO software components are open access, and they are available at https://github.com/inab

The EPICO platform is comprised of the following modules:

- The data model (EPICO-data-model, 2016), which was initially inspired on earlier data models from ICGC DCC (ICGC-DCC-Docs, 2016), and the validation logic (BP-Schema-tools, 2015).
- The data validation and loading programs (EPICO-data-loading-scripts, 2016), which have specific parts for each project (e.g. BLUEPRINT). These specific parts store the public data results produced by each project for genes and transcripts. The generic programs also store the mappings of genes to complexes, reactions and metabolic pathways registered on REACTOME (Croft et al., 2014), as well as additional public data (for instance, principal isoform, start and stop codons or TSSs) from EnsEMBL (Flicek et al., 2014), GENCODE (Harrow et al., 2012) and APPRIS (Rodriguez et al, 2013).
- The underlying database, where all the entries are kept, as well as a copy of the data model and the controlled vocabularies (used by EPICO web client). The NoSQL database system used for BLUEPRINT deployments is Elasticsearch.
- The EPICO REST API, which implements the database queries, providing a programmatic access to the data independent from the database technology (EPICO-REST-API, 2016)

Europe PMC Funders Author Manuscripts

- The data analysis portal itself (BP-Data_analysis, 2016), which fetches the data relevant to the queries, consolidating them on the fly in mean data series for the different charts and views. EPICO web client has been built using web technologies and libraries, including HTML5, SVG, ES5; AngularJS, UI Bootstrap, Jallete, Ng-CSV, D3, Highcharts, Simple Statistics, Plotly and Angular Plotly.

BLUEPRINT Data Analysis Portal general usage description

In the first step (Figures S1A and S3A) the user introduces the query of interest, which can be a combination of genomic coordinate ranges, gene names, transcripts, exons, TSS, complexes and pathway identifiers or names from EnsEMBL (Flicek et al., 2014), GENCODE (Harrow et al., 2012) or REACTOME (Croft et al., 2014). In addition, the user can define a flanking window around the feature to explore the genomic context. In the case of complex features, like pathways or complexes, the genomic regions of all the involved genes are shown. Moreover, the search box understands a basic search language that includes coupling (e.g. "gene:BRCA1 + gene:BRCA2") and difference operators (e.g. "pathway:Cell Cycle, Mitotic – gene:PLK1").

In the second step BDAP shows a set of tabs with all the genomic regions obtained from the initial user query. In each one of these tabs the web platform shows the query, a textual description of the genome layout, (i.e. the genes and transcripts in that region), and it allows the metadata of the samples and the products of the analysis related to the query to be inspected and saved. The 62 primary cell types from the haematopoietic lineage tree involved into 2016-08 release are mapped to Cell Ontology terms (Smith et al., 2007), as implemented in EPICO. The cell lines are described in EFO (Experimental Factor Ontology; Malone et al., 2010) and Cell Line Ontology (Sarntivijai et al., 2014), and the terms have links to the corresponding ontology descriptions. These ontologies do not necessarily reflect all the aspects of differentiation, but as they are built connecting each term (cell type) by relationships of “is_a” and “develops_from” (Bard et al., 2005), the ontology structures in part recapitulate the haematological differentiation. For instance, Cell Ontology has been used to study cell identity along the haematological differentiation pathways (Meehan et al., 2013).

The user selects the cell types and cell lines of his/her interest from the simplified ontology trees (Figures S1B and S3B). The number of samples is shown in function of the cell type and name for each of the selected terms, along with the results of the query region. Procedures that facilitate easy ontological sub-tree selection and ancestor de-selection are implemented in the system.

Finally, in the third step the user can select the initial analytical charts corresponding to the cell types for which epigenomics information is available, including ChIP-Seq (H3K4me3, H3K4me1, H3K27ac, H3K36me3, H3K27me3, H3K9me3 or H2A.Zac), DNaseI-Seq, WGBS (hypo and hyper-methylated regions) and RNA-Seq (gene or transcript level; Figures S1C and S3C). Also, the user can choose to focus on a specific disease in the shown cell types, as long as there are samples from those cell types diagnosed by that disease.

Once the three-step process is completed BDAP displays the results in a graphic form and in tables (BP-analysis, 2016). Results are distributed in three different views, the “General View” uses the different cell types as the data series for the plots, and allows to filter by disease status. The “By tissue” view allows finer grain analysis of specific tissues of origin. The “Diseases by cellular type” view uses as data series the different diseases and normal states, and allows the comparison with a set of cellular types.

RNA-Seq data is represented in box-plots as an expression value for the whole gene or for each transcript (Fragments Per Kilobase Million; BP-FPKM, 2016) and by heatmaps used to show pairwise t-test comparisons (at p-value level) between pairs of cellular types on the same gene or transcript. The graphs of the ChIP-Seq, DNaseI-Seq and Methylation data are spline+ribbon-based scatter plots that represent the genomic coordinates on the x-axis and the averaged (solid line), minimum and maximum (shadow area) -log10(p-value), z-values or methylation levels for the ChIP-Seq, DNA-Seq and WGBS experiments on the y-axis, respectively.

These graphs also include a graphical representation of the genomic layout on the inspected genomic coordinates as special series. Genes in the genomic region are included in this graphical representation, showing in its condensed mode the transcript, exons, UTR, CDS and TSS (start and stop codons) corresponding to the principal isoform of each gene, as identified by APPRIS (Rodriguez et al, 2013).

The graphs are distributed in a fluid grid, which adapts to different screen sizes and resolutions for a better user experience when using BDAP in mobile devices. Moreover, these graphs have interactive features, like the capacity to zoom in and out on the data (all the graphs are updated at once), show and hide the legend, and switch the shown genomic layout between the condensed representation and a complete one, which includes all the transcripts. High quality renderings of each plot can be saved using their context menu for publication in png, jpeg, pdf and svg formats. The entire data series can also be downloaded in csv and xls formats for further analysis. The disease filtering menu, the list of available charts and the list of cell types with data in the query region are on the left of the grid. The user can show or hide the data series related to each cell type by clicking on it, as well as inspect the sample names and the number of samples with data in the query region.

The user can also inspect the first data entries used to compute the visible series on a specific chart. This supporting data contains the coordinates of the ChIP-Seq and the DNaseI-Seq peaks, the hypo- and hypermethylated regions and/or the RNA-Seq expression levels. From this view, all the supporting data can be downloaded in a tabular format.

BDAP’s Browser URL is rewritten on each query, cell type and chart selection allowing to bookmark them for later inspection.

Additional links to the BLUEPRINT DCC portal (DCC_portal, 2016) and the main web page of the project are also provided. Through these links the user can browse the raw and processed data produced by the consortium, as well as a description of the methods and results, the groups participating, and the publications associated to the BLUEPRINT data.

Step-by-step example of BDAP usage: FPR1

In the first step the user introduces the query of interest, FPR1, in the search box and in this case, we selected 500 bp in the flanking window size box to also explore the gene's local upstream and downstream regions (Figure S1A). In the second step the 62 primary cell types from the haematopoietic lineage tree involved into 2016-08 release are mapped to Cell Ontology terms (Smith et al., 2007). We selected the neutrophil terms based on the cell ontology hierarchy: neutrophilic myelocyte; neutrophilic metamyelocyte; band form neutrophil; segmented neutrophil of bone marrow; and mature neutrophil (Figure S1B). Finally, in the third step, from the epigenomic information available, we selected the gene expression and pairwise t-test comparisons charts (from RNA-Seq) and all the histone peaks (from ChIP-Seq experiments) for H3K27Ac, H3K36me3 and H3K4me3 filtering by normal cell types to explore the FPR1 data from the neutrophil differentiation lineage (Figure S1C).

BDAP displayed the data for FPR1, and for its paralogues FPR2 and FPR3, as these genes overlap completely or partially with the longest FPR1 transcript annotated in EnsEMBL (Figure S2).

Step-by-step example of BDAP usage: IRF8

We started the search introducing the gene symbol "IRF8" in the first step and with a flanking window size of 500 bps (Figure S3A). We then selected the neutrophilic myelocyte term, the classical monocyte and the macrophage terms in the second step (Figure S3B), and for the experimental results in the third step, we selected gene and transcript expression and the histone modification charts for H3K27ac, H3K27me3 and H3K9me3 filtering by normal cell types (Figure S3C).

Quantification and Statistical Analysis

BLUEPRINT Data Analysis Portal display information from the official pipelines to analyze the data produced by the project. This information is public available across BLUEPRINT-DCC or ftp site.

Detailed information about the parameters and statistics produced by BLUEPRINT and displayed by BDAP is available at:

For hypo and hyper-methylated regions: http://dcc.blueprint-epigenome.eu/#/md/bs_seq_grch38

For ChIP-seq experiments: http://dcc.blueprint-epigenome.eu/#/md/chip_seq_grch38

For DNAse-seq experiments: http://dcc.blueprint-epigenome.eu/#/md/dnase_seq_grch38

For RNA-seq experiments: http://dcc.blueprint-epigenome.eu/#/md/rna_seq_grch38

The statistics displayed in the boxplot charts (median, mean, quartiles, maximum, minimum and outliers) are computed and represented with Plotly library (https://plot.ly/)

Europe PMC Funders Author Manuscripts

The Welch's t-test calculated for the pairwise t-test comparison charts is done by Simple Statistics library (http://simplestatistics.org/). These charts present the p-value resulted from the test without cut-off selection. The t-test comparisons are made on the fly.

Data and Software Availability

The BLUEPRINT Data Analysis Portal is available at http://blueprint-data.bsc.es/#!/

The EPICO components can be downloaded at:

1. EPICO-data-model: https://github.com/inab/EPICO-data-model
2. EPICO-data-loading-scripts: https://github.com/inab/EPICO-data-loading-scripts
3. EPICO-REST-API: https://github.com/inab/EPICO-REST-API

The BLUEPRINT Data Analysis Portal components can be downloaded at:

1. BP-Schema-Tools: https://github.com/inab/BP-Schema-tools
2. Epico-data-analysis-portal: https://github.com/inab/epico-data-analysis-portal

Additional Resources

The EPICO guide installation and usage is available at https://github.com/inab/epico-data-analysis-portal/wiki

A first steps tutorial about BLUEPRINT Data Analysis Portal usage is available at http://blueprint-data.bsc.es/#!/first-steps

An example usage of BLUEPRINT Data Analysis Portal is available at: http://blueprint-data.bsc.es/#!/doing-a-search

The experiments, datasets and primary analysis that support the BLUEPRINT Data Analysis Portal are available at http://dcc.blueprint-epigenome.eu/#/home

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors thank David Pisano and Miriam Rubio from UBio-CNIO for their contributions at the early stages of the conceptual data model used by EPICO, as well as Avik Data (EMBL-EBI) for feedback on each BLUEPRINT data release. INB-CNIO unit is a member of ProteoRed, PRB2-ISCIII and is supported by grant PT13/0001, of the PE I+D+i 2013-2016, funded by ISCIII and FEDER. BSC-CRG-IRB thanks to the funding support of the Spanish Ministry of Health, ISCIII, in the project Instituto Nacional de Bioinformática - PRB2: PT13/0001/0028. The research leading to these results was funded from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 282510 (BLUEPRINT) and by the European Molecular Biology Laboratory and the Spanish National Bioinformatics Institute.

Europe PMC Funders Author Manuscripts

References

Adams D, Altucci L, Antonarakis SE, Ballesteros J, Beck S, Bird A, Bock C, Boehm B, Campo E, Caricasole A, et al. BLUEPRINT to decode the epigenetic signature written in blood. Nat Biotechnol. 2012; 30(3):224–226. [PubMed: 22398613]

Albrecht F, List M, Bock C, Lengauer T. DeepBlue epigenomic data server: programmatic data retrieval and analysis of epigenome region sets. Nucleic Acids Research. 2016; doi: 10.1093/nar/gkw211

Bard JL, Rhee SY, Ashburner M. An ontology for cell types. Genome Biol. 2005; 6(2):R21. [PubMed: 15693950]

BP-analysis. BLUEPRINT Analysis descriptions release 20160816. 2016. ftp://ftp.ebi.ac.uk/pub/databases/blueprint/releases/20160816/homo_sapiens/

BP-Data_analysis. BLUEPRINT Data Analysis Portal GitHub repository. 2016. https://github.com/inab/epico-data-analysis-portal

BP-FPKM. A description file about how FPKMs were calculated in release 20160816. 2016. ftp://ftp.ebi.ac.uk/pub/databases/blueprint/releases/20160816/homo_sapiens/README_rnaseq_analysis_crg_20160816

BP-Schema-tools. Bioinformatic Pantry Schema tools GitHub repository. 2015. https://github.com/inab/BP-Schema-tools

CEMT. Canadian Epigenetics, Environment and Health Research Consortium (CEEHRC) Network. 2016. http://www.epigenomes.ca/

CEEHRC. McGill EMC (CEEHRC). 2016. http://epigenomesportal.ca/edcc/

Chen PP-S. The Entity-relationship Model—Toward a Unified View of Data. ACM Trans Database Syst. 1976; 1:9–36.

Codd EF. Extending the Database Relational Model to Capture More Meaning. ACM Trans Database Syst. 1979; 4:397–434.

Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, Caudy M, Garapati P, Gillespie M, Kamdar MR, et al. The Reactome pathway knowledgebase. Nucleic Acids Res. 2014; 42(Database issue):D472–477. [PubMed: 24243840]

CREST. International Human Epigenome Consortium, IHEC, Team Japan. 2016. http://crest-ihec.jp/english/index.html

DCC_portal. The BLUEPRINT DCC Portal. 2016. http://dcc.blueprint-epigenome.eu

DEEP. The German epigenome programme ‘DEEP’. 2016. http://www.deutsches-epigenom-programm.de/

ENCODE. The ENCODE Project: ENCyclopedia Of DNA Elements. 2016. https://www.genome.gov/encode/

EPICO-data-model. EPICO data model GitHub repository, designed using BP-Schema-tools. 2016. https://github.com/inab/EPICO-data-model

EPICO-data-loading-scripts. EPICO database loading scripts GitHub repository. 2016. https://github.com/inab/EPICO-data-loading-scripts

EPICO-REST-API. EPICO REST API GitHub repository. 2016. https://github.com/inab/EPICO-REST-API

Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, et al. Ensembl 2014. Nucleic Acids Res. 2014; 42(Database issue):D749–55. [PubMed: 24316576]

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome research. 2012; 9:1760–1774.

ICGC. International Cancer Genome Consortium web page. 2016. http://icgc.org

ICGC-DCC-Docs. ICGC DCC documents. 2016. http://docs.icgc.org/

IHEC. International Human Epigenome Consortium web page. 2016. http://ihec-epigenomes.org/

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The Human Genome Browser at UCSC. Genome Research. 2002; 12:996–1006. [PubMed: 12045153]

Kurotaki D, Yamamoto M, Nishiyama A, Uno K, Ban T, Ichino M, Sasaki H, Matsunaga S, Yoshinari M, Ryo A, et al. IRF8 inhibits C/EBPα activity to restrain mononuclear phagocyte progenitors from differentiating into neutrophils. Nat Commun. 2014; 5:4978. [PubMed: 25236377]

Malone J, Holloway E, Adamusiak T, Kapushesky M, Zheng J, Kolesnikov N, Zhukova A, Brazma A, Parkinson H. Modeling sample variables with an Experimental Factor Ontology. Bioinformatics. 2010; 26:1112–1118. [PubMed: 20200009]

Meehan TF, Vasilevsky NA, Mungall CJ, Dougall DS, Haendel MA, Blake JA, Diehl AD. Ontology based molecular signatures for immune cell types via gene expression analysis. BMC Bioinformatics. 2013; 14:263. [PubMed: 24004649]

ROADMAP. The NIH Roadmap Epigenomics Mapping Consortium. 2016. http://www.roadmapepigenomics.org/

Rodriguez JM, Maietta P, Ezkurdia I, Pietrelli A, Wesselink J-J, Lopez G, Valencia A, Tress ML. APPRIS: annotation of principal and alternative splice isoforms. Nucl Acids Res. 2013; 41:D110–D117. [PubMed: 23161672]

Sarntivijai S, Lin Y, Xiang Z, Meehan TF, Diehl AD, Vempati UD, Schürer SC, Pang C, Malone J, Parkinson H, et al. CLO: The cell line ontology. Journal of Biomedical Semantics. 2014; 5:37. [PubMed: 25852852]

Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol. 2007; 25:1251. [PubMed: 17989687]

Ye RD, Boulay F, Wang JM, Dahlgren C, Gerard C, Parmentier M, Serhan CN, Murphy PM. International Union of Basic and Clinical Pharmacology. LXXIII. Nomenclature for the formyl peptide receptor (FPR) family. Pharmacol Rev. 2009; 61(2):119–161. [PubMed: 19498085]

Zhou X, Wang T. Using the Wash U Epigenome Browser to Examine Genome-Wide Sequencing Data. Current Protocols in Bioinformatics. 2012 40:10.10:10.10.1–10.10.14.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

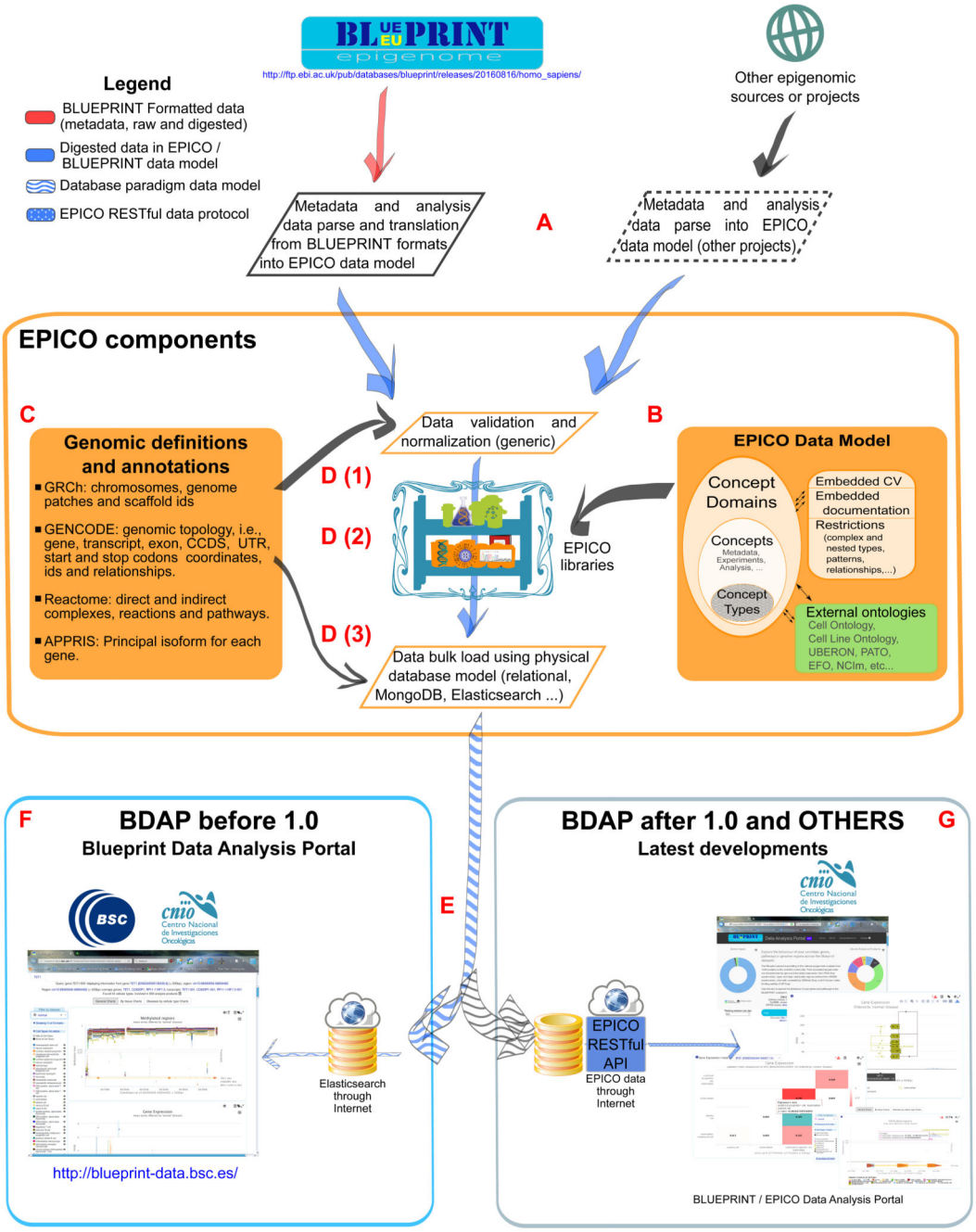

Figure 1. EPICO infrastructure flowchart

A. Each epigenomic data set usually has its own file formats and conventions, so this step is custom.

B. EPICO data model concepts, ontologies and restrictions are common. Only details like the versions of reference EnsEMBL, GENCODE, GRCh and other primary database resources or project name have to be tweaked.

C. As genomic definitions and annotations are published in common sites, and their data formats are stable from release to release, this step is done by EPICO.

D. The metadata and data insertion (which should be following the EPICO data model at this point) is composed by several steps, all of them generic: data validation and normalization (1) using EPICO libraries (2), which later translate it into the dependent database model (3) (currently supported relational, MongoDB and Elasticsearch). In the case of BLUEPRINT we have used Elasticsearch.

E. The data is massively inserted into the database, which already contains the database definitions mapped from the EPICO data model, as well as the ontologies, and the genomic coordinates of the known features, like genes, transcripts, direct complexes, reactions and pathways.

F. BLUEPRINT Data Analysis Portal prior to version 1.0 was issuing its queries to the read-only instance of Elasticsearch which contained all the BLUEPRINT metadata + primary analysis data.

G. BDAP 1.0 issue its queries to the EPICO REST API, which manages the different databases, and implements the queries to Elasticsearch. EPICO Data Analysis Portal is going to be a superset of BDAP, able to work with one or more project data sets at once. Data from different epigenomic projects usually cannot be mixed on comparisons, due different experimental, normalization and analysis protocols.

Europe PMC Funders Author Manuscripts