EGO: a biomedical ontology for integrative epigenome representation and analysis

Yongqun He[1], Jie Zheng[2], Zhaohui Qin[3]

[1] **University of Michigan, Ann Arbor, MI, USA** (Email: yongqunh@med.umich.edu) [2] **University of Pennsylvania, Philadelphia, PA, USA** (jiezheng@upenn.edu); [3] **Emory University, Atlanta, GA, USA** (zhaohui.qin@emory.edu)

***Abstract*—Epigenomics is crucial to understand biological mechanisms beyond genome DNA. To better represent epigenomic knowledge and support data integration, we developed a prototype Epigenome Ontology (EGO). EGO top level hierarchy and design pattern are provided with a use case illustration. EGO is proposed to be used for statistically analyzing enriched epigenomic features based on given sequence data input using statistical methods.**

Keywords—Epigenome; ontology; EGO; enrichment analysis; ENCODE; ChIP-seq

I. Introduction

The majority of eukaryotic genomes such as those of the human and mouse is noncoding. In eukaryote genomes, basic biological functions such as gene expression are affected by many regulatory elements located outside the coding region of the genome. Unlike the genome, which is largely static across tissues and cells within an individual, the epigenome is cell type specific and can be dynamically altered by environmental conditions. Better epigenomic understanding is critical for uncovering biological mechanisms and disease etiology.

Our understanding of the epigenome has dramatically improved in the past decade thanks to the efforts of several large international consortia, e.g., the Encyclopedia of DNA Elements (ENCODE; https://www.encodeproject.org/). As more and more cells and cell lines are being profiled, combined with more factors being studied, it becomes difficult to track all the experiments and resulting knowledge. In particular, many of the experiments are related due to the fact that either the cell types, or the experimental factors are related or both. Thus it is inadvisable to treat these data as independent.

For better interpretation, the subtle and complicated relationships among these experiments should be more fully considered to achieve a better interpretation. Specifically, a well-defined ontology system that handles complex relationships within a rigorous framework and offers annotation at various levels of granularity would be particularly effective. Towards such goals, we developed a new ontology named Epigenome Ontology (EGO).

II. Methods

A. *EGO ontology development*

EGO development follows the OBO Foundry principles (http://obofoundry.org/). Existing terms from other ontologies were imported to EGO using OntoFox (http://ontofox.hegroup.org). New terms were added and edited using the Protégé OWL editor.

An EGO ontology design pattern was generated with a use case in ChIP-seq, a method that combines chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing to identify DNA-protein binding sites [1].

EGO is open and freely available at GitHub: https://github.com/EGO-ontology/EGO.

B. *EGO applications*

Beyond the ChIP-seq use case illustration, different EGO applications were identified in this study.

III. Results

A. *EGO top level design*

Fig. 1 illustrates the top level EFO hierarchy.

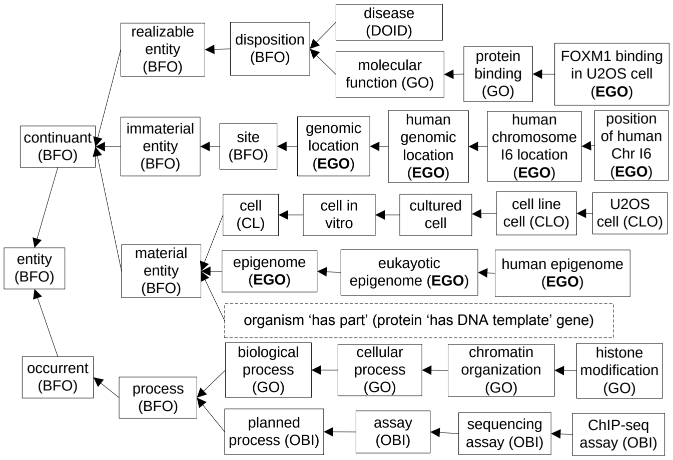


Fig. 1. EGO top level design. BFO: Basic Formal Ontology; OBI: Ontology for Biomedical Investigations (OBI); CL: Cell Type Ontology (CL); CLO: Cell Line Ontology (CLO); DOID: Disease Ontology; GO: Gene Ontology. All terms are aligned together under the BFO structure.

Current prototype EGO includes >600 terms. EGO is aligned with the BFO (http://ifomis.uni-saarland.de/bfo/) (Fig. 1). EGO also imports many terms and relations from existing ontologies (Fig. 1). EGO-specific terms focus on terms in the domain of epigenome. EGO defines ‘epigenome’ as a material entity that is made up of chemical compounds and proteins that can attach to DNA and direct various actions (e.g., turning genes on or off) (https://www.genome.gov/27532724). Since EGO targets genetic intervals at base-pair resolution, we will represent different positions of human and mouse chromosomes and the events that occur at these positions.

B. *EGO ontology design pattern with example*

EGO is aimed to catalog and organize a large amount of high quality, genome-wide profiling data and label every base with observed "events" denoted as "EGO terms" such as ChIP-seq read coverage for a transcription factor (TF) in a specific cell line. Fig. 2 illustrates an example design of how EGO represents ChiP-seq read results for human Forkhead TF FOXM1 binding in the cell line cell U2OS, which occurs at different chromosomal positions of the cell line genome DNA [1]. For example, FOXM1 binds at a promoter region of gene *PLK1* [1] (Fig. 2). FOXM1 is represented with the Protein Ontology (PR), and its corresponding gene is represented by the Ontology of Genes and Genomes (OGG). EGO marks the chromosome positions where FOXM1 binding is indicated by the landing of a ChIP-seq read in the cell line. EGO includes shortcut relations such as 'function in cell' and 'binding at position' to lay out semantic axioms and make queries (Fig. 2).

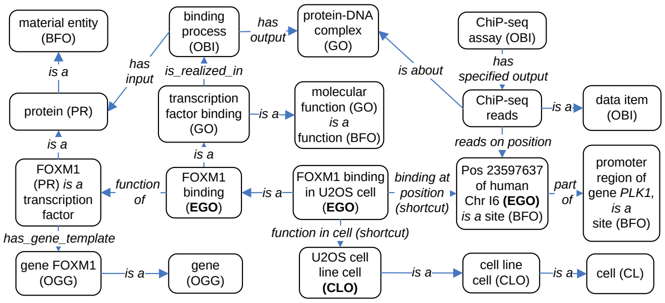
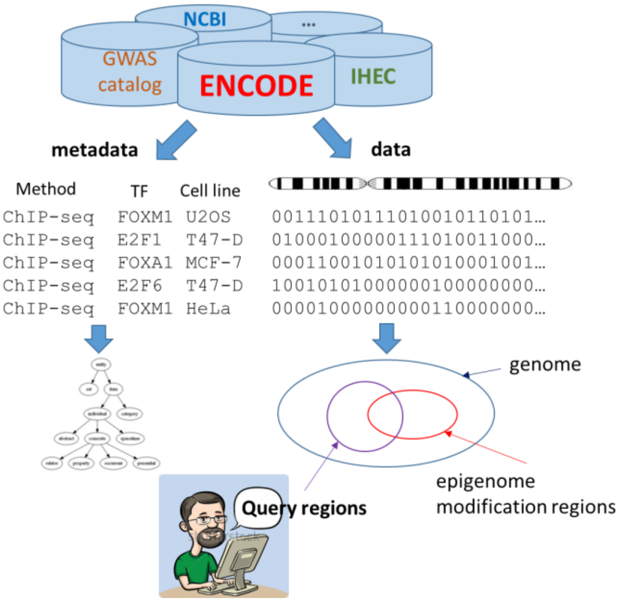


Fig. 2. Design pattern illustrated with example. The pattern links contents from different ontologies (e.g., BFO, CLO, GO, OBI, OGG and PRO) in EGO. More detail about this example is provided in the text.

C. *EGO applications*

With the EGO support, we will be able to leverage dense, genome-wide annotations of various types to facilitate enrichment analysis of "EGO terms" for any give genomic region(s), similar to the GO term enrichment analyses or gene set enrichment analysis (GSEA). As an illustration, suppose a researcher conducted ATAC-seq experiments [2] on primary tumor cells of prostate cancer as well as normal surrounding tissues. She is interested in finding out in the thousands of genomic loci that harbor a peak in the tumor cells but not in the normal cells, what transcription factor (TF) or histone marks showed elevated in vivo binding in relevant cell types (such as CLO_0000748: human prostate gland-derived cell lines http://purl.obolibrary.org/obo/CLO_0000748). EGO enrichment analysis will return a ranked list of all EGO terms satisfied with the cell type constraint and highlight the EGO terms that show significant enrichment. In another example, suppose a researcher has identified two lists of genes that show differential expression between psoriasis skin tissue and adjacent normal tissue (one up-regulated and one down-regulated). She can then collect the promoter region (10 kb upstream to 10 kb downstream of the transcription start site) for the two sets of genes, and ask the question which TF or TFs are bind preferentially to each set of these regions. Such information will inform whether different TFs are likely to be responsible for turning on and off these genes during the process of disease pathogenesis.



Method	TF	Cell line	
ChIP-seq	FOXM1	U2OS	00111010111010010110101…
ChIP-seq	E2F1	T47-D	01000100000111010011000…
ChIP-seq	FOXA1	MCF-7	00011001010101010001001…
ChIP-seq	E2F6	T47-D	10010101000000100000000…
ChIP-seq	FOXM1	HeLa	00001000000000110000000…

Fig. 3. Illustration of the EGO project pipeline and research goals. Different transcription factor (TF) binding sites were identified by ChiP-seq in different cell lines. The results are summarized in EGO and used for advanced queries and analyses given different sequence regions.

Such an enrichment analysis enables researchers to learn the key properties of a set of genomic regions, collectively, avoid the distraction of noise and diversity that are found common place in set of regions identified through high-throughput experiments. A key advantage of using an ontology system is that annotation is provided at different level of granularity, beyond individual dataset level. For example, instead of querying individual TFs in which many are highly related, EGO analysis enables query against TF families which provide concise and clear interpretation. Similarly, often times it is highly desirable to compare the binding patters of a factor across a handful of cell type classes over many closely-related individual cell types.

IV. Discussion

EGO aims to represent the terms and complex relations in the domain of epigenome. With EGO, we are hoping to shed light on the less understood, "dark matter" regions of the genomes while producing insights, enhancing interpretation, and generating new hypotheses for genomic regions of interest.

References

[1] X. Chen, G. A. Muller, M. Quaas, M. Fischer, N. Han, B. Stutchbury, *et al.*, "The forkhead transcription factor FOXM1 controls cell cycle-dependent gene expression through an atypical chromatin binding mechanism," *Mol Cell Biol*, vol. 33, pp. 227-36, Jan 2013.

[2] J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, and W. J. Greenleaf, "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position," *Nat Methods*, vol. 10, pp. 1213-8, Dec 2013.