Goal: to create an ontology system for annotating human and mouse genome about protein-DNA interaction, using ChIP-seq data.

The annotation will be in the form of EGO terms. Each EGO term is consisted of two parts: factor and cell type. An ontology will be built on each of them.

EGO term: factor term | cell type term

Cell type

Factor

Cell type

Factor

Cell type

Factor

Data:

We will use 359 SYDH ENCODE ChIP-seq data (from Snyder lab) and 692 HAIB ENCODE ChIP-seq data (from Hudson Alpha). Initially, we will use narrow peaks (genomic regions) and broad peaks (genomic regions) provided by ENCODE to annotate. To be specific, all bases within a peak region is labeled by the EGO term.

For each EGO term, we need to provide information that how many datasets it uses and genomic coverage (%).

Testing:

Test enrichment of EGO terms on several lists of genomic regions.

ChIP-seqs, DNase-seq, ATAC-seq peaks.

Promoter, enhancer or coding regions.

Any regions of interest?

Issues need to consider:

1. Antibody. Do we add antibody information to the EGO term?
2. How to address the inbalance of genomic coverage, and number of datasets associated with each EGO term.
3. Need more ideas in testing.