ENCODE DREAM Challenge

John Reid

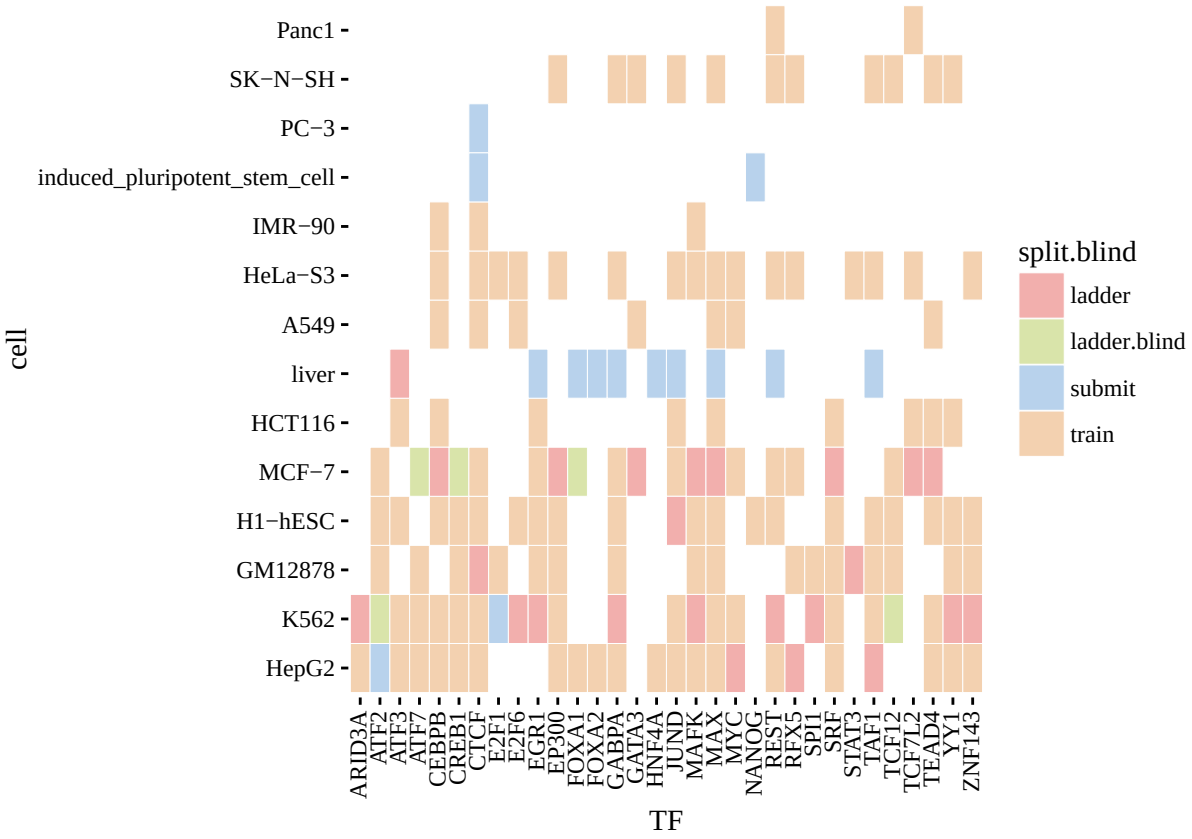
The ENCODE-DREAM challenge

The challenge is to predict cell-type specific binding of transcription factors (TFs) using four types of data:

- DNA sequence: a reference human genomic sequence
- DNA shape: the physical shape of the genome
- DNase-seq: experimental evidence for how open and accessible regions of the genome are
- RNA-seq: gene expression levels

TF/cell-type combinations

TFs have different binding profiles in different cell types. 32 TFs and 14 cell types are represented in the data although not all combinations are present.



The prediction task boils down to predicting binding of the TF in a cell type in which you have no binding data for that specific TF. In addition you are evaluated on your predictions only on chromosomes for which you have no data for any TF/cell type combination. Liver is by far the most prevalent cell type in those cell types used in the evaluation of the final submissions. There is no training data for liver for any TFs although it is a ladder cell type for ATF3.

Data

TF binding

TF binding is measured using the ChIP-seq protocol and converted to a binary score for sliding windows of 200bp. The windows slide by 50bp. For each 200bp location, binding is defined as bound (B), unbound (U) or ambiguous (A). In addition the challenge provides more detailed information from the ChIP-seq experiments including conservative and relaxed estimates of peaks and fold-control signals showing how enriched the ChIP-seq experiment was over a control background experiment. It is not clear how useful the extra information will be as it will obviously not be available on the held-out data.

DNA sequence

The human reference genome is over 3 billion base pairs long. Each base is represented as a character: adenine (A); cytosine (C); guanine (G) and thymine (T). The challenge is restricted to chromosomes 1-22 and chromosome X. All data in the challenge are defined with respect to release GRCh37/hg19. TFs tend to prefer to bind specific sequences. These preferences are summarised in binding motifs but these are not known for all TFs. The only external data that is allowed to be used in this challenge are libraries of TF-DNA binding motifs. Obviously this data is not cell-type specific.

DNase-seq

Information on chromatin accessibility on a per-cell-type basis will be available in four formats:

- conservative peaks
- relaxed peaks
- filtered BAM alignment files
- fold-enrichment signal coverage tracks

The first two will be much easier to use as they summarise the last two.

DNA shape

Participants are encouraged to use the DNAshapeR to calculate DNA shape features across the genome. This information is not cell-type specific but has been shown to be predictive of TF-DNA binding. Note that `DNAshapeR` requires a version of R > 3.3.

Gene expression

Gene expression is regulated by TF binding and so should be useful indirectly to infer binding. The major difficulties are that:

- It is not known which locations on the genome regulate which genes. Commonly TF binding sites regulate the closest gene but this is not always the case.
- Several TFs can combine in an unknown way to regulate a gene.
- There are other mechanisms of gene regulation that will confound the relationship.

Transcription factors

Motifs

- Hard to find motif for *TAF1*. Is *TBP* a good motif to use?

Cell types

There are 14 cell types, 3 of which have never had any TF-binding released under ENCODE

- PC-3
- induced pluripotent stem cells
- liver

Most of the final submission will be predictions of binding for the whole genome in these 3 cell types for a subset of TFs. In addition whole genome predictions will be required for ATF2 in HEPG2 cells and E2F1 in K562 cells. Some metadata on the cell types:

cell	lineage	tissue	karyotype	sex
HepG2	endoderm	liver	cancer	M
K562	mesoderm	blood	cancer	F
GM12878	mesoderm	blood	normal	F
H1-hESC	inner cell mass	embryonic stem cell	normal	M
MCF-7	ectoderm	breast	cancer	F
HCT116	NA	colon	cancer	M
liver	NA	liver	NA	NA
A549	endoderm	epithelium	cancer	M
HeLa-S3	ectoderm	cervix	cancer	F
IMR-90	endoderm	lung	normal	F
induced_pluripotent_stem_cell	NA	NA	NA	NA
PC-3	NA	prostate	cancer	M
SK-N-SH	ectoderm	brain	cancer	F
Panc1	NA	pancreas	cancer	M