ENCODE DREAM Challenge

John Reid

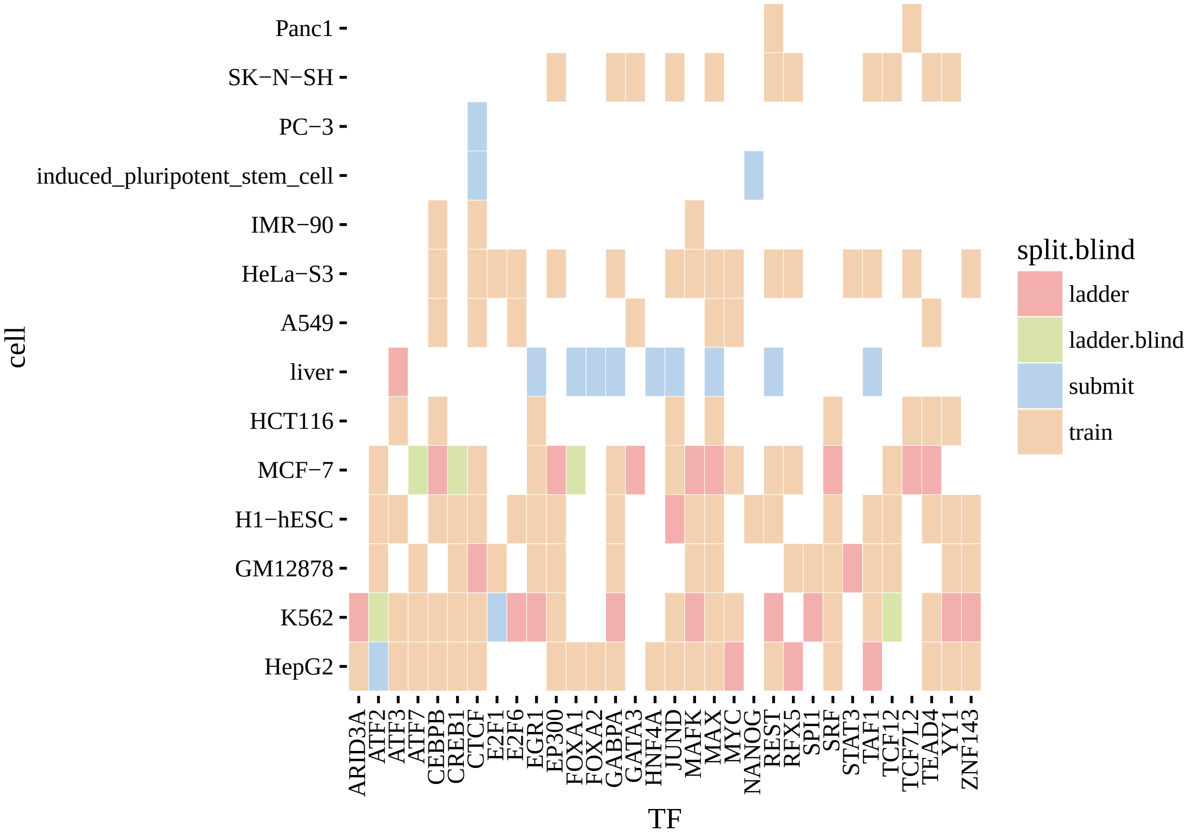
The ENCODE-DREAM challenge

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.



There are essentially three (two!?) prediction tasks:

- Held-out chromosomes: For each TF/cell-type combination, the ChIP-seq data will not be available for 3 chromosomes. Methods will be assessed by their predictive performance on these chromosomes.
- Across cell-type: Methods will be assessed by their performance on cell-types for which no training data has been made available (PC-3, induced pluripotent stem cells, liver).
- Within cell-type: Methods will be assessed on

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 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

```
cells.csv <- 'cell,lineage,tissue,karyotype,sex,description
"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",,colon,cancer,M,
"liver",,liver,,,
"A549",endoderm,epithelium,cancer,M,
"HeLa-S3",ectoderm,cervix,cancer,F,
"IMR-90",endoderm,lung,normal,F,
"induced_pluripotent_stem_cell",,,,,
"PC-3",,prostate,cancer,M,
"SK-N-SH",ectoderm,brain,cancer,F,
"Panc1",,pancreas,cancer,M,'
cells <-
    as.data.frame(readr::read_csv(cells.csv)) %>%
    mutate(
        cell = factor(cell, levels=levels(tfs$cell)),
        lineage = factor(lineage),
        tissue = factor(tissue),
        karyotype = factor(karyotype),
        sex = factor(sex))
cells
```

```
##                             cell         lineage              tissue
## 1                          HepG2        endoderm               liver
## 2                           K562        mesoderm               blood
## 3                        GM12878        mesoderm               blood
## 4                        H1-hESC inner cell mass embryonic stem cell
## 5                          MCF-7        ectoderm              breast
## 6                         HCT116            <NA>               colon
## 7                          liver            <NA>               liver
## 8                           A549        endoderm          epithelium
## 9                        HeLa-S3        ectoderm              cervix
## 10                        IMR-90        endoderm                lung
## 11 induced_pluripotent_stem_cell            <NA>                <NA>
## 12                          PC-3            <NA>            prostate
## 13                       SK-N-SH        ectoderm               brain
## 14                         Panc1            <NA>            pancreas
##    karyotype  sex description
## 1     cancer    M        <NA>
## 2     cancer    F        <NA>
## 3     normal    F        <NA>
## 4     normal    M        <NA>
## 5     cancer    F        <NA>
## 6     cancer    M        <NA>
## 7       <NA> <NA>        <NA>
## 8     cancer    M        <NA>
## 9     cancer    F        <NA>
## 10    normal    F        <NA>
## 11      <NA> <NA>        <NA>
## 12    cancer    M        <NA>
## 13    cancer    F        <NA>
## 14    cancer    M        <NA>
```

```
# devtools::use_data(cells, overwrite=TRUE)
```