ENCODE DREAM Challenge

John Reid

```
## Joining, by = c("TF", "cell")
```

```
##         TF                          cell  split true.blind  split.blind
## 1   ARID3A                         HepG2  train      FALSE        train
## 2   ARID3A                          K562 ladder      FALSE       ladder
## 3     ATF2                       GM12878  train      FALSE        train
## 4     ATF2                       H1-hESC  train      FALSE        train
## 5     ATF2                         MCF-7  train      FALSE        train
## 6     ATF2                          K562 ladder       TRUE ladder.blind
## 7     ATF2                         HepG2 submit       TRUE       submit
## 8     ATF3                        HCT116  train      FALSE        train
## 9     ATF3                       H1-hESC  train      FALSE        train
## 10    ATF3                         HepG2  train      FALSE        train
## 11    ATF3                          K562  train      FALSE        train
## 12    ATF3                         liver ladder      FALSE       ladder
## 13    ATF7                       GM12878  train      FALSE        train
## 14    ATF7                         HepG2  train      FALSE        train
## 15    ATF7                          K562  train      FALSE        train
## 16    ATF7                         MCF-7 ladder       TRUE ladder.blind
## 17   CEBPB                          A549  train      FALSE        train
## 18   CEBPB                       H1-hESC  train      FALSE        train
## 19   CEBPB                        HCT116  train      FALSE        train
## 20   CEBPB                       HeLa-S3  train      FALSE        train
## 21   CEBPB                         HepG2  train      FALSE        train
## 22   CEBPB                        IMR-90  train      FALSE        train
## 23   CEBPB                          K562  train      FALSE        train
## 24   CEBPB                         MCF-7 ladder      FALSE       ladder
## 25   CREB1                       GM12878  train      FALSE        train
## 26   CREB1                       H1-hESC  train      FALSE        train
## 27   CREB1                         HepG2  train      FALSE        train
## 28   CREB1                          K562  train      FALSE        train
## 29   CREB1                         MCF-7 ladder       TRUE ladder.blind
## 30    CTCF                          A549  train      FALSE        train
## 31    CTCF                       H1-hESC  train      FALSE        train
## 32    CTCF                       HeLa-S3  train      FALSE        train
## 33    CTCF                         HepG2  train      FALSE        train
## 34    CTCF                        IMR-90  train      FALSE        train
## 35    CTCF                          K562  train      FALSE        train
## 36    CTCF                         MCF-7  train      FALSE        train
## 37    CTCF                       GM12878 ladder      FALSE       ladder
## 38    CTCF                          PC-3 submit       TRUE       submit
## 39    CTCF induced_pluripotent_stem_cell submit       TRUE       submit
## 40    E2F1                       GM12878  train      FALSE        train
## 41    E2F1                       HeLa-S3  train      FALSE        train
## 42    E2F1                          K562 submit       TRUE       submit
## 43    E2F6                          A549  train      FALSE        train
## 44    E2F6                       H1-hESC  train      FALSE        train
## 45    E2F6                       HeLa-S3  train      FALSE        train
## 46    E2F6                          K562 ladder      FALSE       ladder
## 47    EGR1                       GM12878  train      FALSE        train
## 48    EGR1                       H1-hESC  train      FALSE        train
## 49    EGR1                        HCT116  train      FALSE        train
## 50    EGR1                         MCF-7  train      FALSE        train
## 51    EGR1                          K562 ladder      FALSE       ladder
## 52    EGR1                         liver submit       TRUE       submit
## 53   EP300                       GM12878  train      FALSE        train
## 54   EP300                       H1-hESC  train      FALSE        train
## 55   EP300                       HeLa-S3  train      FALSE        train
## 56   EP300                         HepG2  train      FALSE        train
## 57   EP300                          K562  train      FALSE        train
## 58   EP300                       SK-N-SH  train      FALSE        train
## 59   EP300                         MCF-7 ladder      FALSE       ladder
## 60   FOXA1                         HepG2  train      FALSE        train
## 61   FOXA1                         MCF-7 ladder       TRUE ladder.blind
## 62   FOXA1                         liver submit       TRUE       submit
## 63   FOXA2                         HepG2  train      FALSE        train
## 64   FOXA2                         liver submit       TRUE       submit
## 65   GABPA                       GM12878  train      FALSE        train
## 66   GABPA                       H1-hESC  train      FALSE        train
## 67   GABPA                       HeLa-S3  train      FALSE        train
## 68   GABPA                         HepG2  train      FALSE        train
## 69   GABPA                         MCF-7  train      FALSE        train
## 70   GABPA                       SK-N-SH  train      FALSE        train
## 71   GABPA                          K562 ladder      FALSE       ladder
## 72   GABPA                         liver submit       TRUE       submit
## 73   GATA3                          A549  train      FALSE        train
## 74   GATA3                       SK-N-SH  train      FALSE        train
## 75   GATA3                         MCF-7 ladder      FALSE       ladder
## 76   HNF4A                         HepG2  train      FALSE        train
## 77   HNF4A                         liver submit       TRUE       submit
## 78    JUND                        HCT116  train      FALSE        train
## 79    JUND                       HeLa-S3  train      FALSE        train
## 80    JUND                         HepG2  train      FALSE        train
## 81    JUND                          K562  train      FALSE        train
## 82    JUND                         MCF-7  train      FALSE        train
## 83    JUND                       SK-N-SH  train      FALSE        train
## 84    JUND                       H1-hESC ladder      FALSE       ladder
## 85    JUND                         liver submit       TRUE       submit
## 86    MAFK                       GM12878  train      FALSE        train
## 87    MAFK                       H1-hESC  train      FALSE        train
## 88    MAFK                       HeLa-S3  train      FALSE        train
## 89    MAFK                         HepG2  train      FALSE        train
## 90    MAFK                        IMR-90  train      FALSE        train
## 91    MAFK                          K562 ladder      FALSE       ladder
## 92    MAFK                         MCF-7 ladder      FALSE       ladder
## 93     MAX                          A549  train      FALSE        train
## 94     MAX                       GM12878  train      FALSE        train
## 95     MAX                       H1-hESC  train      FALSE        train
## 96     MAX                        HCT116  train      FALSE        train
## 97     MAX                       HeLa-S3  train      FALSE        train
## 98     MAX                         HepG2  train      FALSE        train
## 99     MAX                          K562  train      FALSE        train
## 100    MAX                       SK-N-SH  train      FALSE        train
## 101    MAX                         MCF-7 ladder      FALSE       ladder
## 102    MAX                         liver submit       TRUE       submit
## 103    MYC                          A549  train      FALSE        train
## 104    MYC                       HeLa-S3  train      FALSE        train
## 105    MYC                          K562  train      FALSE        train
## 106    MYC                         MCF-7  train      FALSE        train
## 107    MYC                         HepG2 ladder      FALSE       ladder
## 108  NANOG                       H1-hESC  train      FALSE        train
## 109  NANOG induced_pluripotent_stem_cell submit       TRUE       submit
## 110   REST                       H1-hESC  train      FALSE        train
## 111   REST                       HeLa-S3  train      FALSE        train
## 112   REST                         HepG2  train      FALSE        train
## 113   REST                         MCF-7  train      FALSE        train
## 114   REST                         Panc1  train      FALSE        train
## 115   REST                       SK-N-SH  train      FALSE        train
## 116   REST                          K562 ladder      FALSE       ladder
## 117   REST                         liver submit       TRUE       submit
## 118   RFX5                       GM12878  train      FALSE        train
## 119   RFX5                       HeLa-S3  train      FALSE        train
## 120   RFX5                         MCF-7  train      FALSE        train
## 121   RFX5                       SK-N-SH  train      FALSE        train
## 122   RFX5                         HepG2 ladder      FALSE       ladder
## 123   SPI1                       GM12878  train      FALSE        train
## 124   SPI1                          K562 ladder      FALSE       ladder
## 125    SRF                       GM12878  train      FALSE        train
## 126    SRF                       H1-hESC  train      FALSE        train
## 127    SRF                        HCT116  train      FALSE        train
## 128    SRF                         HepG2  train      FALSE        train
## 129    SRF                          K562  train      FALSE        train
## 130    SRF                         MCF-7 ladder      FALSE       ladder
## 131  STAT3                       HeLa-S3  train      FALSE        train
## 132  STAT3                       GM12878 ladder      FALSE       ladder
## 133   TAF1                       GM12878  train      FALSE        train
## 134   TAF1                       H1-hESC  train      FALSE        train
## 135   TAF1                       HeLa-S3  train      FALSE        train
## 136   TAF1                          K562  train      FALSE        train
## 137   TAF1                       SK-N-SH  train      FALSE        train
## 138   TAF1                         HepG2 ladder      FALSE       ladder
## 139   TAF1                         liver submit       TRUE       submit
## 140  TCF12                       GM12878  train      FALSE        train
## 141  TCF12                       H1-hESC  train      FALSE        train
## 142  TCF12                         MCF-7  train      FALSE        train
## 143  TCF12                       SK-N-SH  train      FALSE        train
## 144  TCF12                          K562 ladder       TRUE ladder.blind
## 145 TCF7L2                        HCT116  train      FALSE        train
## 146 TCF7L2                       HeLa-S3  train      FALSE        train
## 147 TCF7L2                         Panc1  train      FALSE        train
## 148 TCF7L2                         MCF-7 ladder      FALSE       ladder
## 149  TEAD4                          A549  train      FALSE        train
## 150  TEAD4                       H1-hESC  train      FALSE        train
## 151  TEAD4                        HCT116  train      FALSE        train
## 152  TEAD4                         HepG2  train      FALSE        train
## 153  TEAD4                          K562  train      FALSE        train
## 154  TEAD4                       SK-N-SH  train      FALSE        train
## 155  TEAD4                         MCF-7 ladder      FALSE       ladder
## 156    YY1                       GM12878  train      FALSE        train
## 157    YY1                       H1-hESC  train      FALSE        train
## 158    YY1                        HCT116  train      FALSE        train
## 159    YY1                         HepG2  train      FALSE        train
## 160    YY1                       SK-N-SH  train      FALSE        train
## 161    YY1                          K562 ladder      FALSE       ladder
## 162 ZNF143                       GM12878  train      FALSE        train
## 163 ZNF143                       H1-hESC  train      FALSE        train
## 164 ZNF143                       HeLa-S3  train      FALSE        train
## 165 ZNF143                         HepG2  train      FALSE        train
## 166 ZNF143                          K562 ladder      FALSE       ladder
```

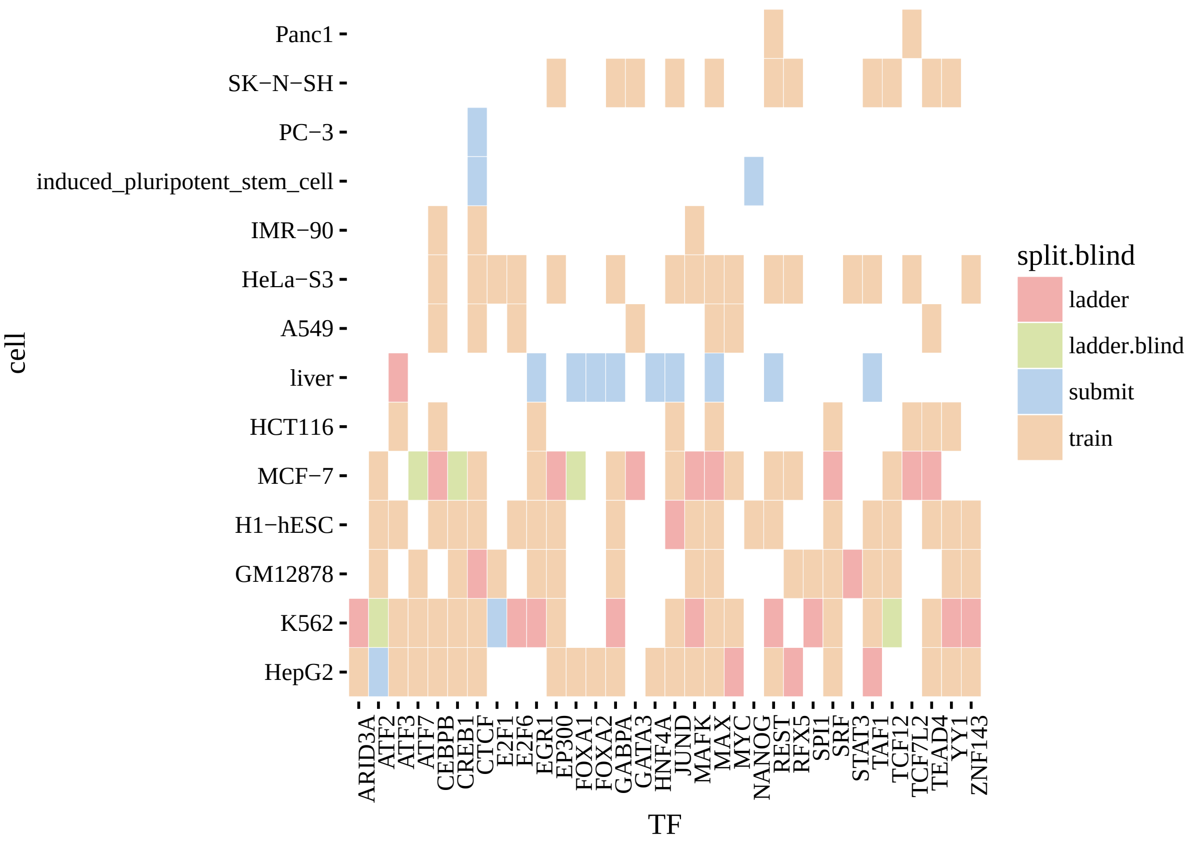
The ENCODE-DREAM challenge

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
- *In-vitro* DNA shape: the physical shape of the genome in an *in-vitro* system
- DNase-seq: how open and accessible regions of the genome are
- RNA-seq: levels of gene expression

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?