Build and run Saturn DREAM-ENCODE project

Clone the repository

```
git clone git@github.com:JohnReid/Saturn.git
```

Create and active a python virtual environment

```
cd Saturn/python
bin/create-virtualenv-py2
. ./activate-py2
pip install -r reqd-pkgs.txt
pip install -r reqd-pkgs-py2.txt
```

Create a data directory

```
cd ..
mkdir Data
```

Download the challenge data into the `'Data` directory

```
annotations/hg19.genome.fa.gz
annotations/ladder_regions.blacklistfiltered.bed.gz
annotations/ladder_regions.blacklistfiltered.merged.bed
annotations/ladder_regions.processed.bed
annotations/test_regions.blacklistfiltered.bed.gz
annotations/test_regions.blacklistfiltered.merged.bed
annotations/train_regions.blacklistfiltered.bed.gz
annotations/train_regions.blacklistfiltered.merged.bed

DNASE/peaks/conservative/DNASE.A549.conservative.narrowPeak.gz
DNASE/peaks/conservative/DNASE.GM12878.conservative.narrowPeak.gz
DNASE/peaks/conservative/DNASE.H1-hESC.conservative.bed
...

DNASE/peaks/relaxed/DNASE.A549.relaxed.narrowPeak.gz
DNASE/peaks/relaxed/DNASE.GM12878.relaxed.narrowPeak.gz
DNASE/peaks/relaxed/DNASE.H1-hESC.relaxed.bed
...

DNASE/bams/DNASE.H1-hESC.biorep1.techrep1.bam
```

...

```
ChIPseq/labels/ARID3A.train.labels.tsv.gz
ChIPseq/labels/ATF2.train.labels.tsv.gz
ChIPseq/labels/ATF3.train.labels.tsv.gz
```
...

DNase

Create DNase features from the p-values of the relaxed peaks for each cell type by

```
scripts/feature-DNase.R <cell type>
```

Merge the DNase sequencing reads for each cell type by

```
merge-dnase <cell type>
```

Run wellington on the relaxed peaks for each cell type by

```
run-wellington <cell type> relaxed
```

Known motifs

Download the necessary known motif databases from the MEME suite website into the following locations

```
/etc/STEME/MEME-dbs/EUKARYOTE/jolma2010.meme
/etc/STEME/MEME-dbs/EUKARYOTE/jolma2013.meme
/etc/STEME/MEME-dbs/EUKARYOTE/SwissRegulon_human_and_mouse.meme
/etc/STEME/MEME-dbs/EUKARYOTE/zhao2011.meme
/etc/STEME/MEME-dbs/JASPAR/JASPAR_CORE_2016_vertebrates.meme
/etc/STEME/MEME-dbs/CIS-BP/Homo_sapiens.meme
```

Filter those motifs we are interested in.

```
KNOWNMOTIFSDIR=$SATURNDIR/Data/Motifs/Known
get-motifs >$KNOWNMOTIFSDIR/tf-motifs.meme
```

Install STEME[1].

Scan the genome for the known motifs using STEME.

```
STEME_USE_GENOME_INDEX=1 steme-pwm-scan \
    --prediction-Z-threshold=.7 \
    --lambda=.001 \
    --cache-index \
    -o $KNOWNMOTIFSDIR \
```

[1] https://github.com/JohnReid/STEME

```
    $SATURNDIR/Data/Motifs/Known/tf-motifs.meme \
    $SATURNDIR/Data/annotations/hg19.chrs.fa
```

Create a set of features from the scan.

```
scripts/feature-scan.R $KNOWNMOTIFSDIR Known
```

Create a set of features from the scan that are filtered by the Wellington footprints for each cell type.

```
scripts/feature-scan.R \
    --cell=<cell type 1> \
    --cell=<cell type 2> \
    ... \
    --wellington \
    $KNOWNMOTIFSDIR KnownWell
```

De novo motifs

Install DREME[2].

Generate positive and negative sequences for each TF that DREME can discriminate between

```
fasta-get <TF>
```

Run DREME on the sequences for each TF (this can take up to 8 days on my machine)

```
tf-dreme <TF>
```

Scan the genome for each TF

```
$SATURNDIR/python/bin/scan-genome $SATURNDIR/Data/ChIPseq/seqs/DREME-<TF>
```

Create a set of features from the scans.

```
scripts/feature-scan.R --tf=<TF> $SATURNDIR/Data/ChIPseq/seqs/DREME-<TF> DREME
```

Create a set of features from the scan that are filtered by the Wellington footprints for each cell type relevant for each TF.

```
scripts/feature-scan.R \
    --tf=<TF> \
    --cell=<cell type 1> \
    --cell=<cell type 2> \
    ... \
    --wellington \
    $SATURNDIR/Data/ChIPseq/seqs/DREME-<TF>
    DREMEWell
```

[2] http://meme-suite.org/doc/download.html?man_type=web

Now all the features should be ready for prediction.

Predict

Run the prediction script on each TF/cell type combination of interest

```
scripts/predict.R \
    -f DNase \
    -f Known \
    -f KnownWell \
    -f DREME \
    -f DREMEWell \
    <TF> \
    <cell type>
```

Smooth the predictions using the TF-specific parameters given in `$SATURNDIR/R/smooth.R`. Note we do not smooth the predictions for each TF.

```
scripts/predictions-smooth.R \
    --logodds \
    --width=<width> \
    --length-scale=<length scale> \
    $SATURNDIR/Data/Predictions/predictions-....tsv \
    <output file>
```