Segway for ENCODE4 - code documentation

Marjan Farahbod

September 2023

This is the code documentation for the ENCODE Segway project. Each section is a component of the project, with a diagram showing the input and output files, the code files and the process.

- Code is in the Github repository:
https://github.com/marjanfarahbod/SegwayClustering

- List of the project compartments:

1. Segway train and run on Cedar

2. Segway interpretation, train and test

3. The transcriptomic segway-chrom comparison analyses

4. The Enhancer analyses for Segway and Chrom

5. Obtaining data from ENCODE API, Chrom and RNAseq

6. The GWAS wing. No code on my part, just the analyses documentation Other prep, meta

7. Miscellaneous

1 Segway train and run on Cedar

Figure 1
Folder: SegwayTrainAndRun

2 Segway interpretation; train, test, apply and plots

Figure 2
Folder: SegwayInterpretation

3 The transcriptomic plots and analyses
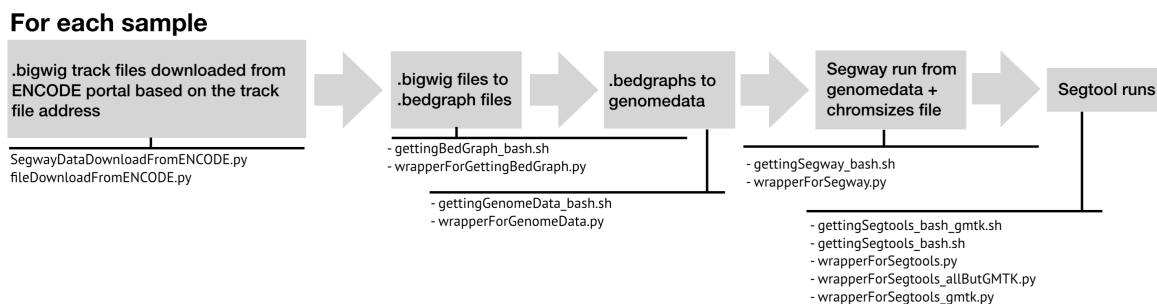
Figure 3
Folder: transcriptomicAnalysis

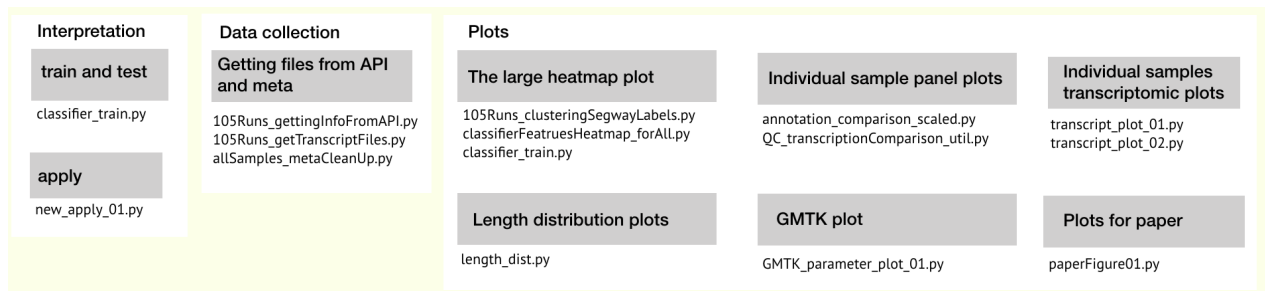Figure 1: Segway train and run on Cedar

Interpretation

train and test

classifier_train.py

apply

new_apply_01.py

Data collection

Getting files from API and meta

105Runs_gettingInfoFromAPI.py
105Runs_getTranscriptFiles.py
allSamples_metaCleanUp.py

Plots

The large heatmap plot

105Runs_clusteringSegwayLabels.py
classifierFeatruesHeatmap_forAll.py
classifier_train.py

Individual sample panel plots

annotation_comparison_scaled.py
QC_transcriptionComparison_util.py

Individual samples transcriptomic plots

transcript_plot_01.py
transcript_plot_02.py

Length distribution plots

length_dist.py

GMTK plot

GMTK_parameter_plot_01.py

Plots for paper

paperFigure01.py

Figure 2: Segway interpretation; train, test, apply and plots

The enrichment plots (the original heatmaps)

For each sample

Preporcessing the initial RNAseq files. Sort and pre-process the original .bed files.

QC_transcriptionComparison_03.py
QC_transcriptionComparison_util.py

Obtaining the enrichment matrix

QC_transcriptionComparison_03
Transcription_overlap.py

Plots

Transcript_plot02.py

The chromHMM and Segway transcription comparison

For each sample

Preporcessing the initial RNAseq files.

QC_transcriptionComparison_03.py
QC_transcriptionComparison_util.py

Obtaining the label coverage for each gene

ChromHMMSegway_comparison_transcription_01.py
Transcription_overlap.py

Apply the regression

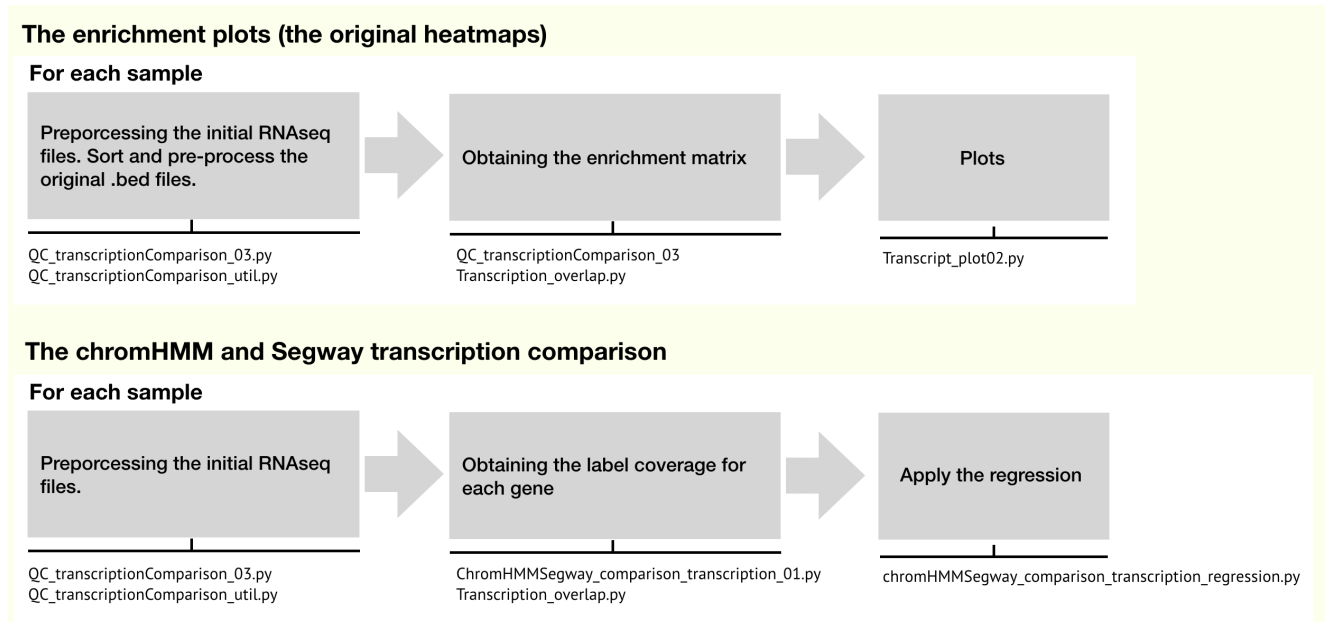chromHMMSegway_comparison_transcription_regression.py
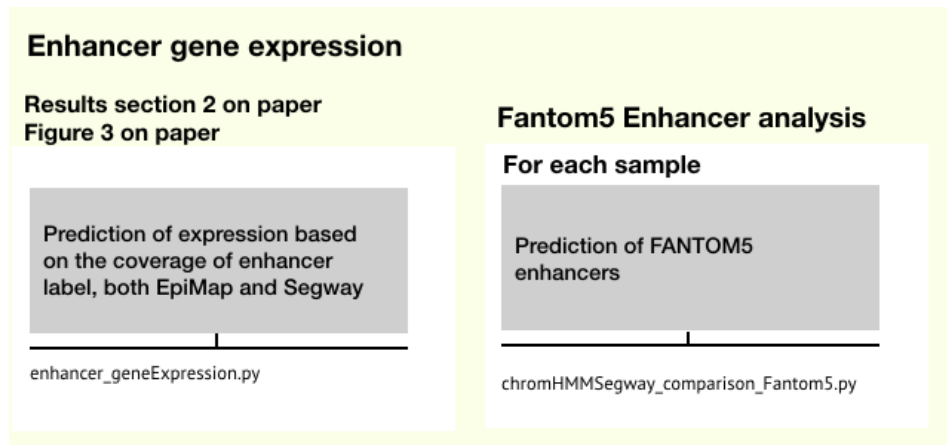
Figure 3: The transcriptomic plots and analyses

Figure 4: Enhancer analyses

4 The Enhancer analyses for Segway and Chrom

Figure 4
Folder: enhancerAnalysis

5 Obtaining data from ENCODE API, Chrom and RNAseq

Folder: Miscellaneuou

```
- get_data.py
- get_the_runID_accession_matching_from_portal.py
```

6 The GWAS wing

No code in this repository.

7 Specifically for the ENCODE paper

Folder: paper

```
- paperFigure01.py
- paperFigure02.py
- paper_supp_plots.py
```