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. Current and obsolete code files are listed as well.

- All 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

Please see the document RunningSegwayOnCedar.pdf for the environment settings used to run Segway and Segtools.

Code list – main:

- `gettingBedGraph_bash.sh`
- `gettingGenomeData_bash.sh`
- `gettingSegtools_bash_gmtk.sh`
- `gettingSegtools_bash.sh`
- `gettingSegtools_bash_oneSample.sh`
- `gettingGenomedata_bash_oneSample.sh`
- `gettingGenomedata_bash.sh`
- `gettingSegway_bash.sh`
- `gettingSegway_bash_oneSample.sh`
- `wrapperForGenomeData_oneSample.py`
- `wrapperForGenomeData.py`
- `wrapperForGettingBedGraph.py`
- `wrapperForSegtools.py`
- `wrapperForSegtools_allButGMTK.py`
- `wrapperForSegtools_gmtk.py`
- `wrapperForSegway.py`
- `wrapperForSegway_oneSample.py`

Code list – auxiliary:

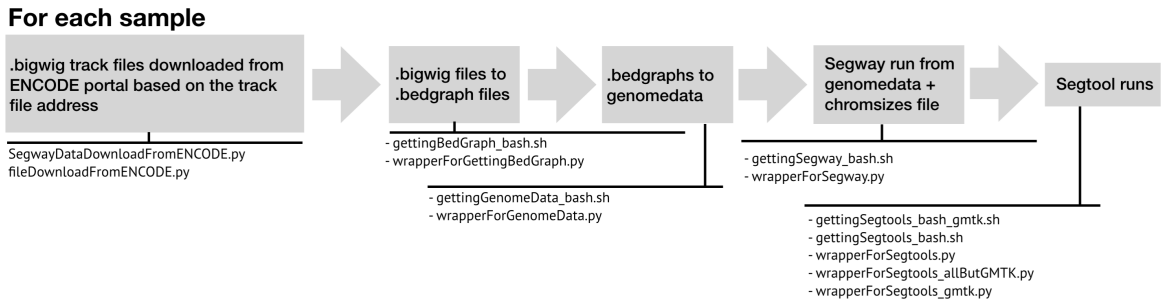

Figure 1: Segway train and run on Cedar

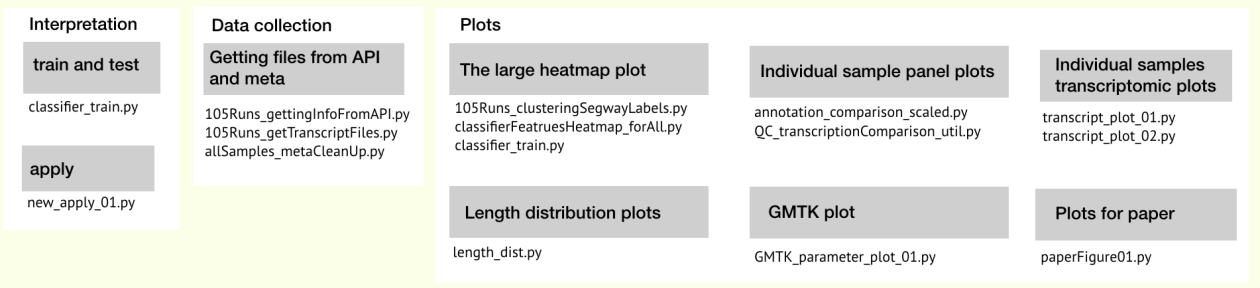

Figure 2: Segway interpretation; train, test, apply and plots

- copyAnnotationsFromCedar.py
- fileDeleteOnCedar.py
- fileDownloadFromENCODE.py
- getSegwayAccessionFromSheet.py
- ifWeHaveGenomeData.py
- sampleSelectionFromThe112Batch.py
- SegwayDataDownloadFromENCODE.py
- the112Batch_organismSelection.py
- unzipbeds_cedar_bash.sh
- unzipbeds_cedar.py
- writingAccessionListToText.py
- zipbeds_cedar_bash.sh

2 Segway interpretation; train, test, apply and plots

Code list – main:

- 105Runs_clusteringSegwayLabels.py
- annotation_comparison.py
- annotation_comparison_scaled.py
- classifier_train.py
- classifierFeatruesHeatmap_forAll.py
- GMTK_parameter_plot_01.py
- length_dist.py
- meta_interpretation_explore.py
- new_apply_01.py
- prob_filter.py
- transcript_plot_01.py
- transcript_plot_02.py
- transcription_overlap.py
- QC_transcriptionComparison_03.py
- QC_transcriptionComparison_util.py
- sampleQuality_interpretationBased.py

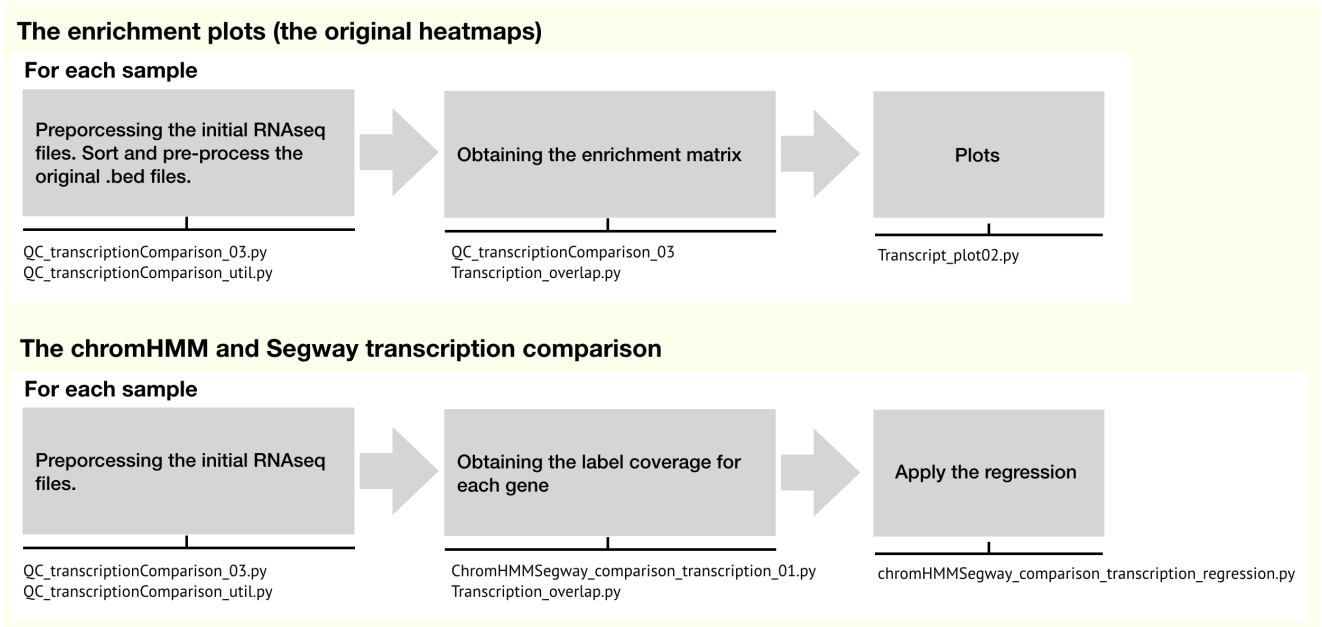

Figure 3: The transcriptomic plots and analyses

Code list – auxiliary:

- 105Runs_gettingInfoFromAPI.py
- 105Runs_getTranscriptFiles.py
- allSamples_metaCleanUp.py
- annotation_fileDownload.py
- get_classifier_data_from_cedar.py
- plot_pdf.py
- util.py

3 The transcriptomic plots and analyses

- chromHMMSegway_comparison_transcription.py
- chromHMMSegway_comparison_transcription_01.py
- chromHMMSegway_comparison_transcription_regression.py
- chromHMMSegway_comparison_transcription_regression_otherSamples.py
- transcription_overlap.py
- transcript_plot_01.py
- transcript_plot_02.py
- QC_transcriptionComparison_03.py
- QC_transcriptionComparison_main.py
- QC_transcriptionComparison_util.py

4 The Enhancer analyses for Segway and Chrom

- chromHMMSegway_comparison_enhancers.py
- chromHMMSegway_comparison_Fantom5.py
- Enhancer_common.py
- Enhancer_distribution.py
- Enhancer_geneExpression.py
- Enhancer_regions.py
- fantom5EnhancerProcessing.py

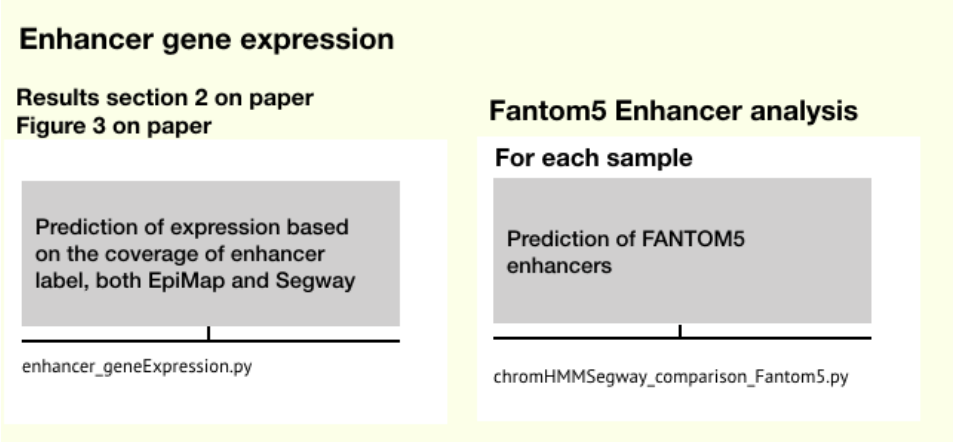

Figure 4: Enhancer analyses

5 Obtaining data from ENCODE API, Chrom and RNAseq

- get_data.py
- get_the_runID_accession_matching_from_portal.py

6 The GWAS wing

No code in this repository. Just the analysis documentation.

7 Miscellaneous

- browser_mod.py
- ccre_annotation_comparison.py
- ENCODE_submission_dataPrep.py
- file_storage_adjustments.py
- genomicRegionsOverGenome.py
- getCCRe.py
- getThatISMBplot.py
- poster_histPlot.py
- SNP_regionAnnotation.py

8 Specifically for the ENCODE paper

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

Code glossary

Code section and category are added. main: m, auxiliary: a

- 105Runs_clusteringSegwayLabels.py 2-m
- 105Runs_getTranscriptFiles.py 2-a
- 105Runs_gettingInfoFromAPI.py 2-a
- allSamples_metaCleanUp.py 2-a
- annotation_comparison.py 2-m
- annotation_comparison_scaled.py 2-m
- annotation_fileDownload.py 2-m
- browser_mod.py 7
- ccre_annotation_comparison.py 7

- chromHMMSegway_comparison_enhancers.py 4
- chromHMMSegway_comparison_Fantom5.py 4
- chromHMMSegway_comparison_transcription.py 3
- chromHMMSegway_comparison_transcription_01.py 3
- chromHMMSegway_comparison_transcription_regression.py 3
- chromHMMSegway_comparison_transcription_regression_otherSamples.py 3
- classifierFeatruesHeatmap_forAll.py 2-m
- copyAnnotationsFromCedar.py 1-a
- classifier_train.py 2-m
- Enhancer_common.py 4
- Enhancer_distribution.py 4
- Enhancer_geneExpression.py 4
- Enhancer_regions.py 4
- fileDeleteOnCedar.py 1-a
- fileDownloadFromENCODE.py 1-a
- genomicRegionsOverGenome.py 7
- get_data.py 5
- getCCRe.py 7
- get_classifier_data_from_cedar.py 2-a
- getSegwayAccessionFromSheet.py 1-a
- get_the_runID_accession_matching_from_portal.py 5
- gettingGenomedata_bash_oneSample.sh 1-m
- gettingGenomeData_bash.sh 1-m
- gettingSegtools_bash_gmtk.sh 1-m
- gettingSegtools_bash.sh 1-m
- gettingSegtools_bash_oneSample.sh 1-m
- gettingSegway_bash.sh 1-m
- gettingSegway_bash_oneSample.sh 1-m
- GMTK_parameter_plot_01.py 2-m
- ifWeHaveGenomeData.py 1-a
- length_dist.py 2-m
- meta_interpretation_explore.py 2-m
- new_apply_01.py 2-m
- paperFigure01.py 8
- paperFigure02.py 8
- paper_supp_plots.py 8
- plot_pdf.py 2-a
- poster_histPlot.py 7
- prob_filter.py 2-m
- QC_transcriptionComparison_03.py 2-m, 3
- QC_transcriptionComparison_main.py 3
- QC_transcriptionComparison_util.py 2-m, 3
- sampleQuality_interpretationBased.py 2
- sampleSelectionFromThe112Batch.py 1-a
- SegwayDataDownloadFromENCODE.py 1-a
- SNP_regionAnnotation.py 7
- transcription_overlap.py 2-m, 3
- transcript_plot_01.py 3, 2-m
- transcript_plot_02.py 3, 2-m
- the112Batch_organismSelection.py 1-a
- unzipbeds_cedar_bash.sh 1-a
- unzipbeds_cedar.py 1-a
- util.py 2-a
- wrapperForGettingBedGraph.py 1-m
- wrapperForGenomeData_oneSample.py 1-m
- wrapperForGenomeData.py 1-m
- wrapperForSegtools.py 1-m
- wrapperForSegtools_allButGMTK.py 1-m
- wrapperForSegtools_gmtk.py 1-m
- wrapperForSegway.py 1-m

- wrapperForSegway_oneSample.py 1-a
- writingAccessionListToText.py 1-a
- zipbeds_cedar_bash.sh 1-a