Directory Structure

Each analysis or experiment should have its input and output files contained within a single folder. Note that the chromosome names in the input files should be the same as those in the mappability files and chromosome lengths files. A separate output folder for each model can also be used instead, e.g. output_model1 and output_model2.

- hiHMM
  - MATLAB_scripts
    - driv_hiHMM.m *(script to run hiHMM)*
  - R_scripts
  - sample_analysis
    - Chromosome_Lengths
    - Mappability_files
    - Mapped_input_files *(output from hiHMM_Pre1.r in Step 2)*
    - Unmapped_input_files
    - Output *(output from driv_hiHMM.m in Step 3)*

The R scripts use a number of packages, which are automatically detected and will need to be installed if not already done so by the user. These include:

- hiHMM_Pre1.r: `IRanges`
- hiHMM_Post1.r: `gplots`, `RColorBrewer`
- hiHMM_Post2.r: `IRanges`

Step 1: Create hiHMM input files

There should be one file for each condition and each chromosome, and should be named as "condition_chromosome.txt". Each file should contain all the ChIP-seq tracks (e.g. histone modifications, transcription factors etc).

The sample files provided contain the signal values for 200 bp bins and in the following structure:
fly_chr2L.txt

	H3K4me3	H3K4me1	H3K27ac	H3K27me3	H3K9me3	H3K36me1	H3K36me3	H3K79me1	H3K79me2	H3K79me3
100	1.273	1.845	1.255	6.678	3.992	3.008	0.4813	2.13	0.2253	0.5487
300	1.263	2.094	1.314	6.704	4.17	3.26	0.2024	2.321	0.2718	0.597
500	-0.1197	-0.3921	-0.3642	1.862	0.8911	0.4991	-0.3099	0.2276	-0.6492	-0.5131
700	0.1116	-0.8932	-0.8005	1.477	-0.2313	0.2389	-0.4453	-0.4695	-0.9227	-0.6949
900	0.5436	-0.4834	0.02368	2.255	1.632	0.6649	-0.3674	0.1754	-0.5208	-0.415
1100	-0.05895	-0.4038	-0.1332	1.656	0.6821	0.2218	-0.4263	-0.4604	-0.6122	-0.4382

These raw files should be stored in **sample_analysis/Unmapped_input_files**.

Step 2: Remove unmappable regions using hiHMM_Pre1.r

Unmappable regions are removed from the hiHMM input files using the information contained in the **sample_analysis/Mappability_files** folder, which contains the mappable regions for each chromosome for each condition. These files are .bed files in the format "chromosome \t start position \t end position". For example, for fly:

mappable.dm3.lucy.bed

chr2L	4991	47665
chr2L	52412	64471
chr2L	64621	348304
chr2L	348486	348763
chr2L	355057	359711
chr2L	359723	359996

The R script then removes all regions that are not mappable, for example the input file for fly chr2L now begins at 5100 bp, since the region 1 bp – 4991 bp is unmappable according to the mappability file.

fly_chr2L.txt

	H3K4me3	H3K4me1	H3K27ac	H3K27me3	H3K9me3	H3K36me1	H3K36me3	H3K79me1	H3K79me2	H3K79me3
5100	-0.3123	0.2516	0.1117	-0.7232	-0.5201	1.086	0.1961	1.322	-0.3769	-0.5807
5300	-0.6504	-0.09875	0.02368	-0.9396	-0.4945	0.4427	0.4978	0.9172	-0.3569	-0.517
5500	-0.9367	-0.3014	0.1918	-0.8595	-0.4755	0.4098	-0.09351	1.05	-0.2972	-0.5385
5700	-1.032	-0.6962	-0.3121	-1.011	-0.4128	0.04305	-0.2534	-0.2515	-0.382	-0.8122
5900	-0.5397	-0.1498	0.3542	-1.143	-0.5445	0.0981	0.2207	0.2507	-0.3796	-0.2164
6100	-0.1392	-0.5817	-0.584	-1.352	-0.1226	0.4569	1.308	1.374	-0.2677	-0.04611

To run this script, users need to specify:

- The working directory to be the analysis folder
- `unmapped_file_dir`: the location of the hiHMM unmapped input files
- `mapped_file_dir`: the location where the mapped files should be written to
- `source_dir`: the location of the Mappability folder
- `Mappable`: the mappable files and a species/condition ID, for example “fly” and “worm”

The files with the unmappable regions removed are written to **sample_analysis/Mapped_input_files**.

Step 3: Run hiHMM using driv_hihmm.m

The analysis folder, input and output folders need to be specified. Parameters such as the conditions, mapped file names, chromosome labels, model number, bin size (same as the input files e.g. 200 bp), etc, also need to be specified.

In the example given, the results from hiHMM, including the chromatin annotations (bed files e.g. *hihmm.model2.K7.fly.bed for fly*) and emission/transition matrices, will be written to the **sample_analysis/output** folder.

Step 4: Annotate states in the emission matrix

The emission matrix (*sample_analysis/output/ train-hihmm-model2.emission.csv*) will contain the *K* chromatin states along the rows and the ChIP-seq tracks used along the columns. Note that for Model 1, the output file will contain all the emission matrices for all conditions. For Model 2, there will only be one emission matrix that has been jointly inferred across all conditions.

The states need to be functionally annotated (e.g. promoter, enhancer) in the following format, and separated by spaces:

`State_Number State_Name Species/Condition`

There are a number of state names and colourings that have been pre-defined in the `colourise` function in the hiHMM_Post1.r script, including promoter, enhancer, gene, transcription, repressed, heterochromatin, and low signal. Note that the last element is optional for Model 2, but should be included in Model 1 as a sample identifier. See hiHMM_Post1.r for more details on naming states. An example of the named emission matrices are given below:

train-hihmm-model1.emission_named.csv: Note that the first emission matrix is for Worm (W) and the second for Fly (F), as specified by the order in the `condition` parameter in driv_hihmm.m

	H3K4me3	H3K4me1	H3K27ac	H3K27me3	H3K9me3	H3K36me1	H3K36me3	H3K79me1	H3K79me2	H3K79me3
2 Enhancer 1 W	-0.0323	0.8691	0.4588	-0.6168	-0.3788	-0.0487	0.0318	-0.8239	1.4910	1.3210
1 Promoter W	5.7863	-0.3360	3.3966	-0.5563	-1.0302	-0.1242	0.9694	-1.2709	1.1744	1.7336
5 PC Repressed W	-0.1054	-1.2029	-0.7792	1.4668	0.3089	-0.8273	-0.8080	-0.6410	-1.0192	-1.1357
6 Heterochromatin W	-0.0705	-0.2804	-0.4956	2.3319	3.8049	-0.4730	-0.2953	0.3221	-0.5007	-0.3229
4 Gene W	-0.0026	-0.0837	-0.3803	-0.6069	-0.1091	-0.3945	1.7930	0.2083	0.3334	0.5038
3 Enhancer 2 W	0.0683	0.5548	0.2642	-0.1012	0.2018	1.0572	-0.3768	1.1546	-0.2198	-0.3697
7 Low Signal W	-0.4918	-0.2480	-0.1779	-0.3383	-0.3760	-0.2093	-0.0410	0.0988	-0.4792	-0.5034
9 Enhancer 1 F	-0.1331	1.0825	0.5356	-0.6573	-0.5845	0.1439	0.0737	-0.4927	1.7550	1.4558
8 Promoter F	5.1553	-0.0312	3.2958	-1.1117	-1.0726	0.4939	1.9017	-0.8159	1.7085	2.3959
12 PC Repressed F	-0.1897	-0.8121	-0.5553	1.3283	0.1452	-0.6437	-0.4787	-0.3432	-1.0586	-1.0005
13 Heterochromatin F	-0.0585	-0.8531	-0.6156	1.3020	2.4621	-0.4543	-0.3100	-0.1041	-0.6831	-0.3218
11 Gene F	-0.0478	0.2996	-0.3738	-0.9096	-0.3130	-0.4386	1.8722	0.1259	0.5530	0.5229
10 Enhancer 2 F	0.1194	0.6623	0.1069	-0.4766	0.0926	1.4390	-0.1899	1.3008	-0.0623	-0.0971
14 Low Signal F	-0.1755	-0.0979	-0.1767	-0.0897	-0.3549	-0.2582	-0.3205	-0.0934	-0.3807	-0.3298

train-hihmm-model2.emission_named.csv: A single set of states jointly inferred

	H3K4me3	H3K4me1	H3K27ac	H3K27me3	H3K9me3	H3K36me1	H3K36me3	H3K79me1	H3K79me2	H3K79me3
2 Enhancer 1	0.1706	1.0422	0.8509	-0.3931	-0.3950	0.2080	-0.6256	-0.5108	1.4658	1.4575
1 Promoter	5.5983	0.0317	3.3322	-0.8904	-0.9702	0.1381	1.1576	-0.7392	1.2823	1.9605
5 PC Repressed	-0.1231	-0.9440	-0.7557	1.4891	0.2976	-0.7011	-0.6193	-0.6147	-1.2136	-0.9854
6 Heterochromatin	0.0520	-0.7432	-0.6946	1.7472	2.8661	-0.6583	-0.4104	-0.0474	-0.6389	-0.2472
4 Gene	-0.0986	0.1945	-0.1259	-0.7067	-0.4945	-0.4578	1.3367	-0.5149	1.0945	0.9962
3 Enhancer 2	0.0434	0.4099	0.2496	-0.2682	0.1239	1.1627	-0.0734	1.2474	-0.3145	-0.2573
7 Low Signal	-0.3452	-0.2534	-0.1762	-0.1939	-0.2541	-0.2377	-0.1895	-0.0452	-0.4448	-0.4510

The emission matrix with annotated states needs to be saved with the "_named" suffix, i.e. **train-hihmm-model2.emission_named.csv**.

Step 5: Plot the emission and transition matrices and recolour the output bed files using hiHMM_Post1.r

The emission and transition matrices can be plotted as PDF files using this script. For Model 1, an emission matrix will be plotted for each condition, while only a single emission matrix will be plotted for Model 2. A transition matrix will be plotted for each condition regardless of which model is used. These will be stored in the same **sample_analysis/output** folder.

To run this script, users need to specify:

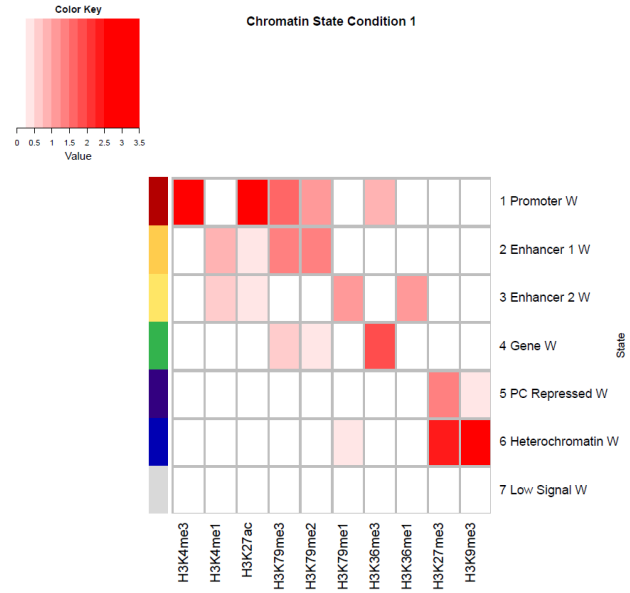
- The working directory to be the hiHMM output folder (subfolder in the analysis folder)
- `m`: the model number (1 or 2)

- `c`: the number of conditions or samples. In this example the number of samples is 2, for fly and worm
- `fworder`: ensure that the ChIP-seq tracks (e.g. histone modifications) used in the analysis are present in this list and in the desired order
- `colourise`: ensure all annotated states are accounted for in this function with the desired colours

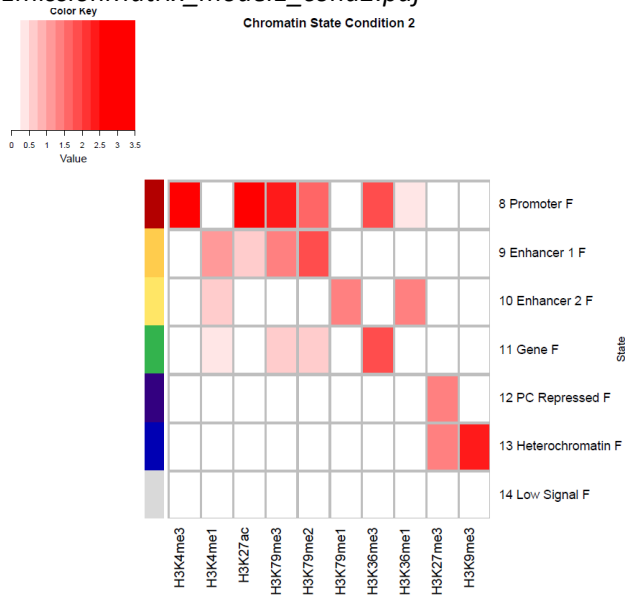
Examples of plotted emission and transition matrices for the two models are provided below:

Model 1

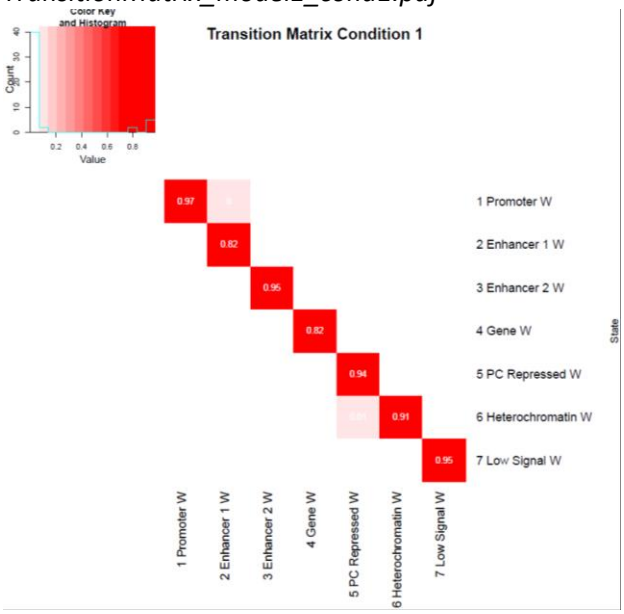
EmissionMatrix_model1_cond1.pdf



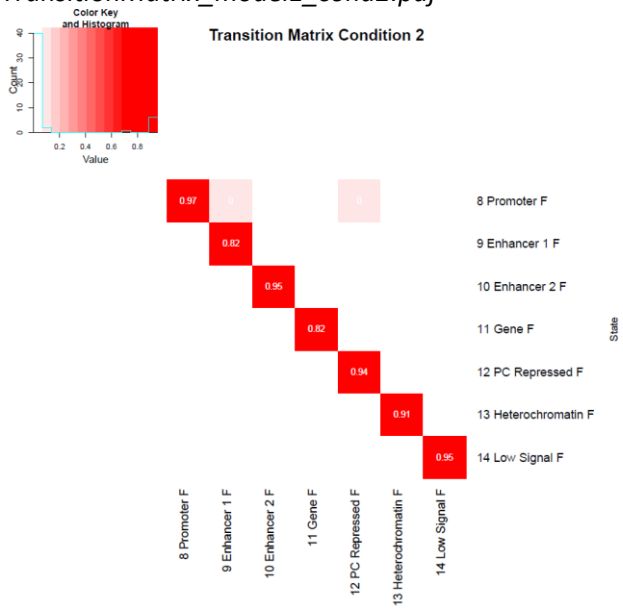
EmissionMatrix_model1_cond2.pdf



TransitionMatrix_model1_cond1.pdf

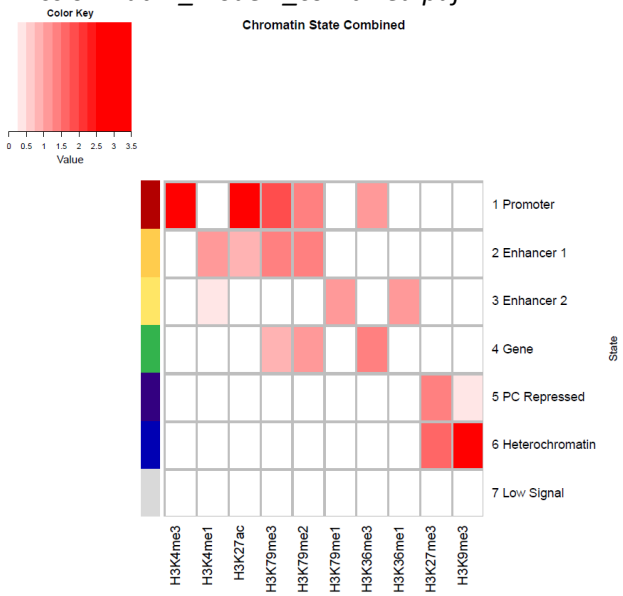


TransitionMatrix_model1_cond2.pdf

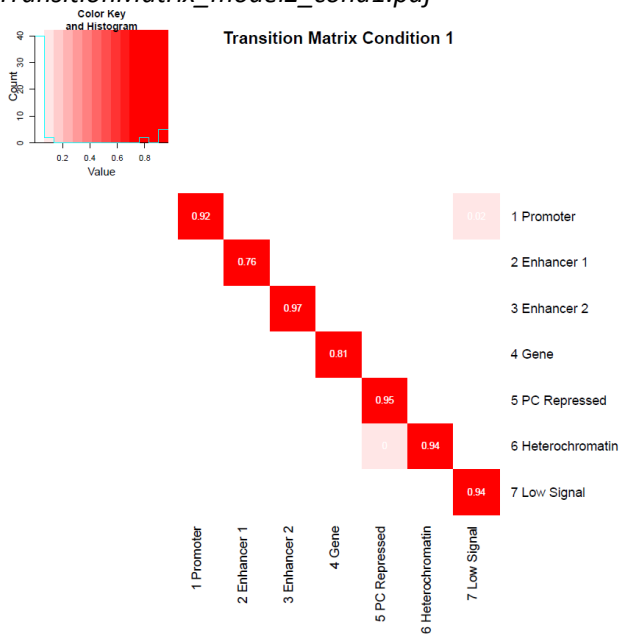


Model 2

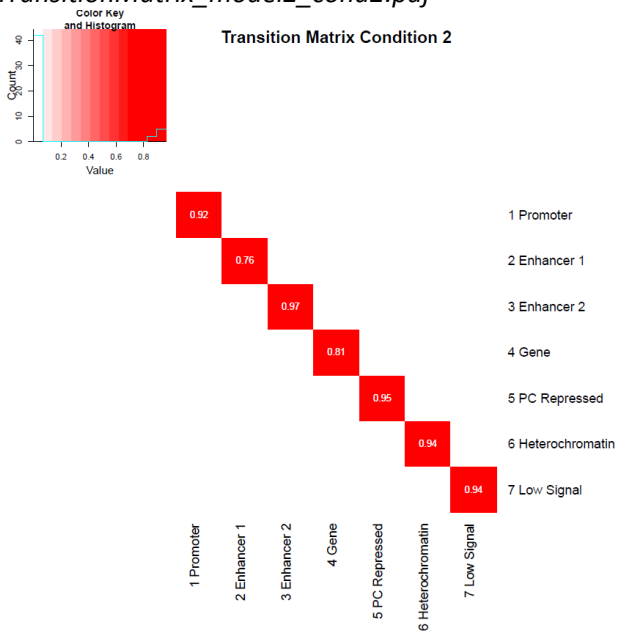
EmissionMatrix_model2_combined.pdf



TransitionMatrix_model2_cond1.pdf



TransitionMatrix_model2_cond2.pdf



The chromatin state segments in the .bed files will also be recoloured according to the same colouring pattern in the emission matrix. These files will have the "recoloured" suffix, for example "*hihmm.model2.K7.fly.bed*" will become "*hihmm.model2.K7.fly.recoloured.bed*" and stored in the same **sample_analysis/output** folder.

Step 6: Reintroduce unmappable regions as a new state using hiHMM_Post2.r

The unmappable regions that were removed in Step 2 prior to running hiHMM will now be added back to the .bed files as "State 0".

To run this script, users need to specify:

- the working directory to be the hiHMM output folder (subfolder in the analysis folder)
- `outdir`: the output directory where the remapped files will be written to
- `bin_size`: bin size (same as the hiHMM input files)
- `mappability_dir`: the location of the Mappability folder
- `Mappable`: the mappable files along with a species/condition ID, for example "fly" and "worm"
- `chr_lengths_dir`: the Chromosome Lengths folder
- `chr_lengths`: the chromosome lengths files along with the same species/condition ID i.e. "fly" and "worm"

The output files will have the "ReMapped" suffix, so for example, "*hihmm.model2.K7.fly.recoloured.bed*" will become *"hihmm.model2.K7.fly.recoloured.ReMapped.bed"* and written to the **sample_analysis/output/ReMapped** folder.

Results

The following image shows a screenshot of the IGV Genome Browser for fly showing the different types of .bed files that are produced from Steps 3, 5 and 6 respectively. Note that the colours for each state in the recoloured and remapped bed files correspond to those in the emission matrix, with the unmappable state 0 coloured as black.