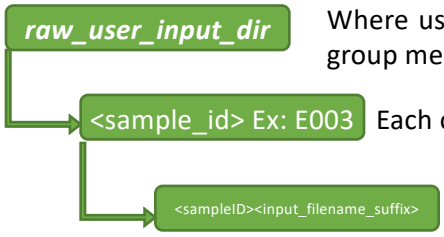
Cheat sheet for the users parameters, and output file structures of CSREP

This cheat sheet aims at helping you easily follow along the tutorial, and later easily modify the config/config.yml for your own usage of CSREP or base_count method.

Files/folders that contain CSREP's output (or supporting files produced by the pipeline)

Files/folders that users provide

- ***Bold, black and italic words*** represent variable names as appeared in config/config.yml
- **Bold and black words** show the example file path as presented in the testdata folder provided along with the tutorial



Where users store all input segmentation data files for all samples, regardless of their group memberships. **testdata/raw_data**

Each of these folder corresponds to one sample. **testdata/raw_data/E003/**

Input segmentation data for sample <**sampleID**>. <input_filename_suffix> is shared across all samples, and is a user input param. **testdata/raw_data/E003/E003_chr22_core_K27ac_segments.bed.gz** → **_chr22_core_K27ac_segments.bed.gz** correspond to ***input_filename_suffix*** in **config/config.yml**

state_annot_fn

This file specifies different characteristics of the states . Two mandatory columns: **state** and **mnemonic. testdata/state_annot.txt**

chrom_length_fn

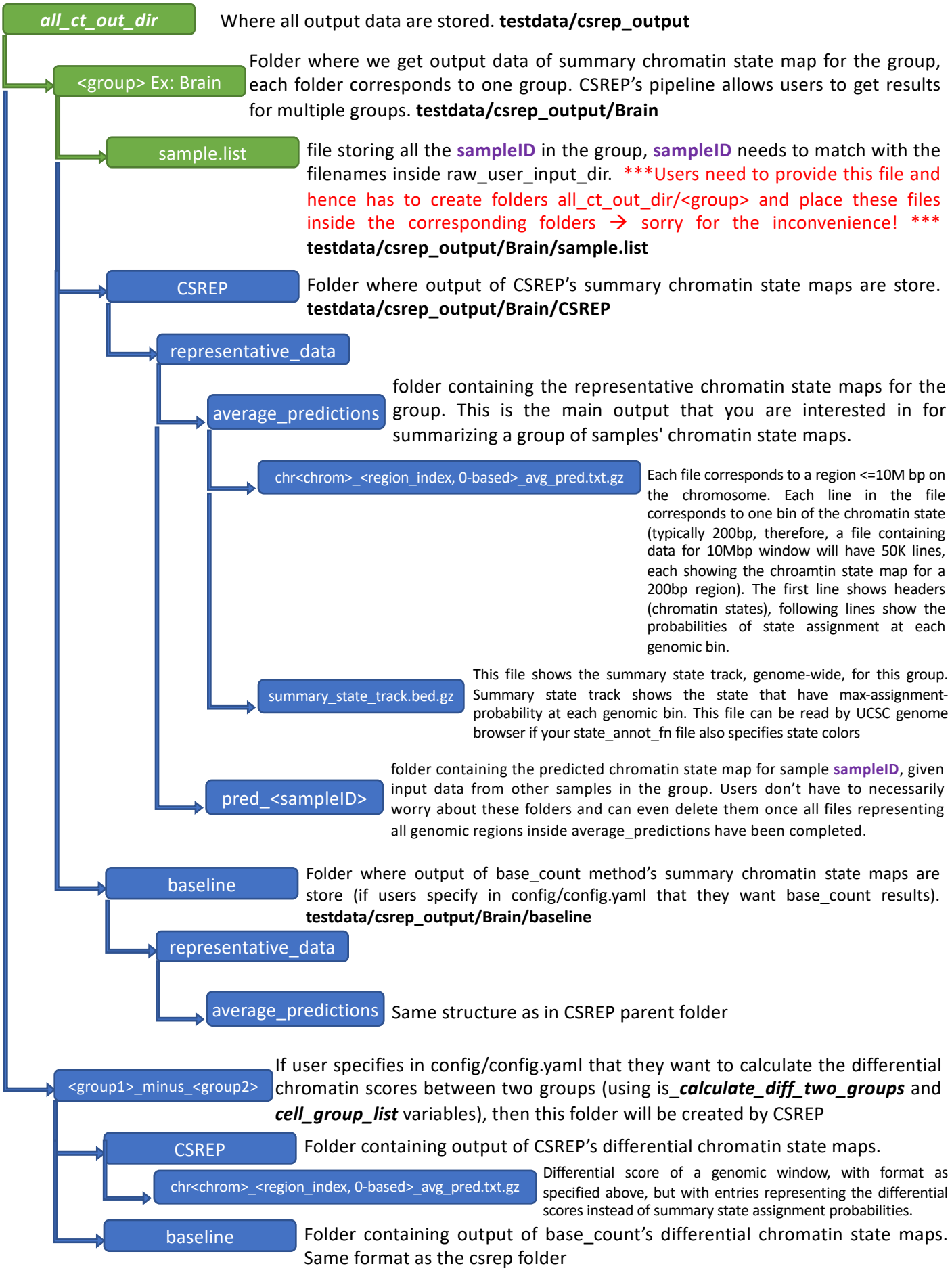
bed file showing the length of each chromosome, used when CSREP try to generate random sample regions across the genome for training. **testdata/roadmap_18state_chromlength.bed**

sample_genome_fn

where the data of regions that we sample for training data are stored (we recommend specifying this file path inside all_ct_out_dir). **testdata/sample_genome.bed.gz**

training_data_folder

the folder where we will store chromatin state segmentation data used for training, which corresponds to 10% of the genome, for each sample.


all_ct_out_dir
Where all output data are stored. testdata/csrep_output
<group> Ex: Brain
Folder where we get output data of summary chromatin state map for the group, each folder corresponds to one group. CSREP's pipeline allows users to get results for multiple groups. testdata/csrep_output/Brain
sample.list
file storing all the sampleID in the group, sampleID needs to match with the filenames inside raw_user_input_dir. ***Users need to provide this file and hence has to create folders all_ct_out_dir/<group> and place these files inside the corresponding folders → sorry for the inconvenience! *** testdata/csrep_output/Brain/sample.list
CSREP
Folder where output of CSREP's summary chromatin state maps are store. testdata/csrep_output/Brain/CSREP
representative_data
average_predictions
folder containing the representative chromatin state maps for the group. This is the main output that you are interested in for summarizing a group of samples' chromatin state maps.
chr<chrom>_<region_index, 0-based>_avg_pred.txt.gz
Each file corresponds to a region <=10M bp on the chromosome. Each line in the file corresponds to one bin of the chromatin state (typically 200bp, therefore, a file containing data for 10Mbp window will have 50K lines, each showing the chroamtin state map for a 200bp region). The first line shows headers (chromatin states), following lines show the probabilities of state assignment at each genomic bin.
summary_state_track.bed.gz
This file shows the summary state track, genome-wide, for this group. Summary state track shows the state that have max-assignment-probability at each genomic bin. This file can be read by UCSC genome browser if your state_annot_fn file also specifies state colors
pred_<sampleID>
folder containing the predicted chromatin state map for sample sampleID, given input data from other samples in the group. Users don't have to necessarily worry about these folders and can even delete them once all files representing all genomic regions inside average_predictions have been completed.
baseline
Folder where output of base_count method's summary chromatin state maps are store (if users specify in config/config.yaml that they want base_count results). testdata/csrep_output/Brain/baseline
representative_data
average_predictions
Same structure as in CSREP parent folder
<group1>_minus_<group2>
If user specifies in config/config.yaml that they want to calculate the differential chromatin scores between two groups (using is_calculate_diff_two_groups and cell_group_list variables), then this folder will be created by CSREP
CSREP
Folder containing output of CSREP's differential chromatin state maps.
chr<chrom>_<region_index, 0-based>_avg_pred.txt.gz
Differential score of a genomic window, with format as specified above, but with entries representing the differential scores instead of summary state assignment probabilities.
baseline
Folder containing output of base_count's differential chromatin state maps. Same format as the csrep folder