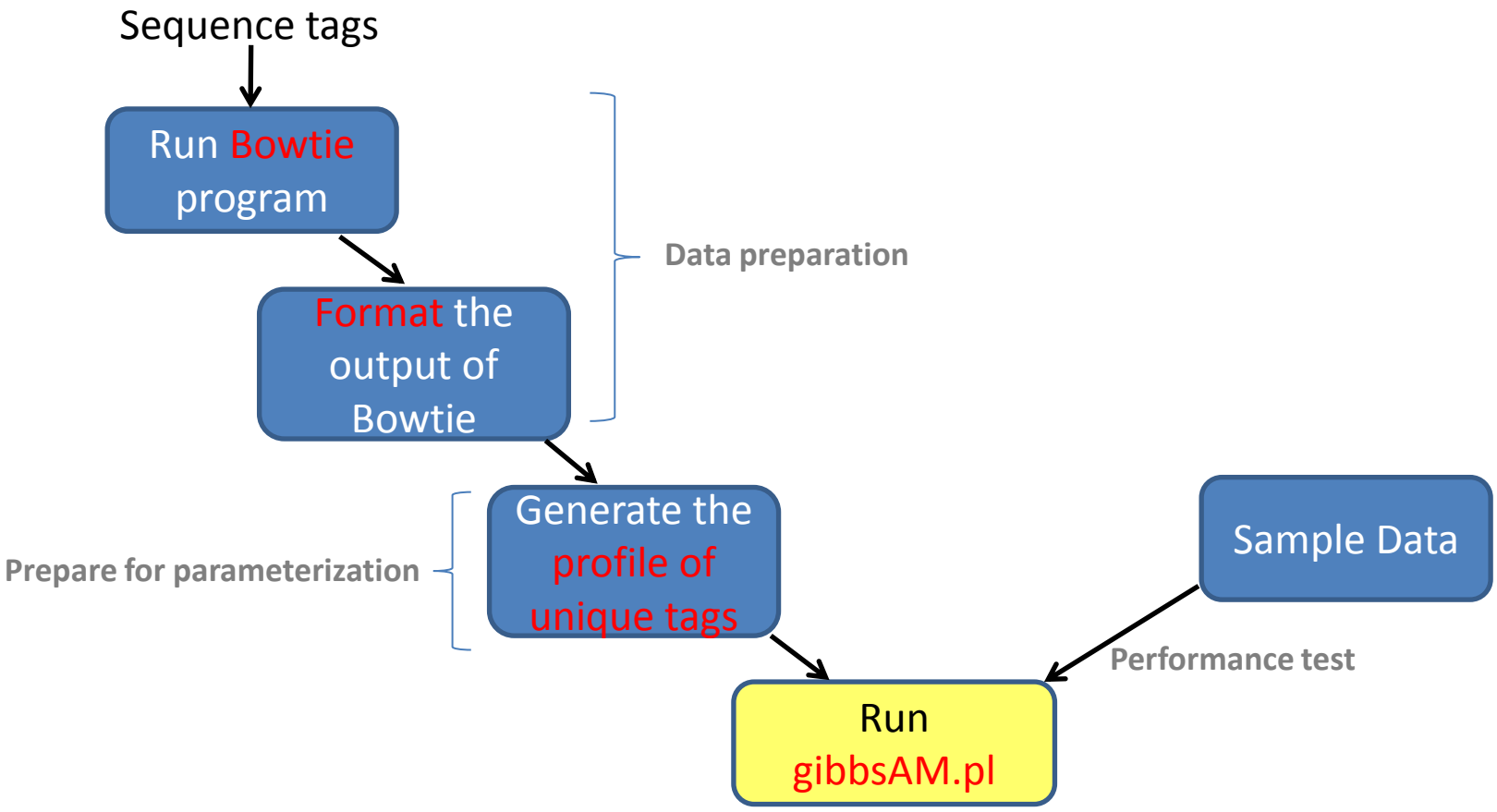
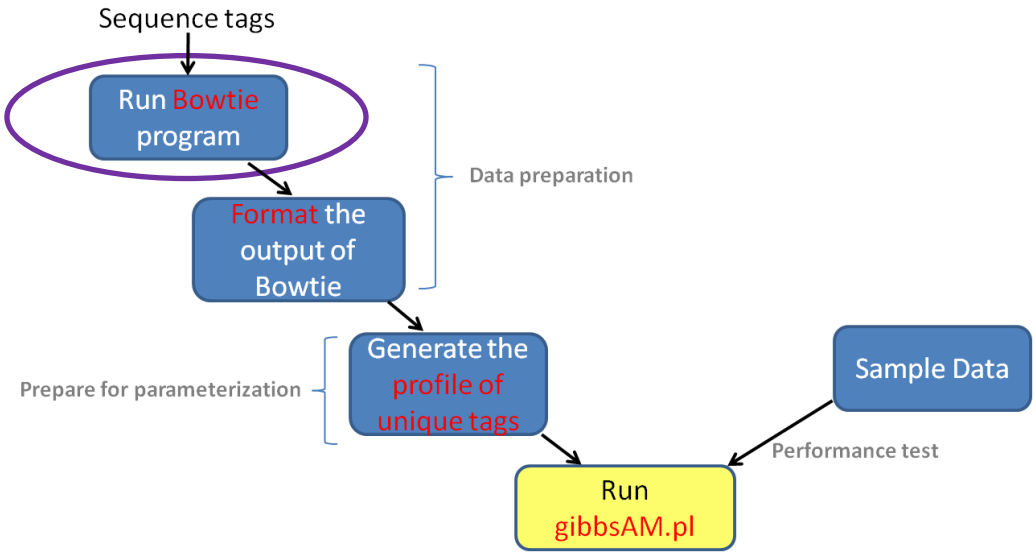
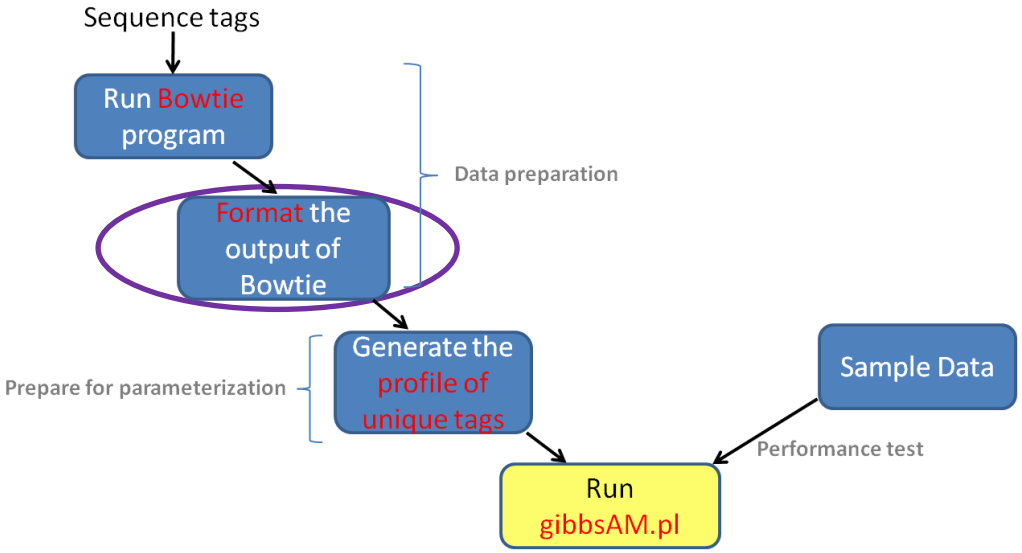
Documentation for “A Gibbs sampling strategy applied to the mapping of ambiguous short sequence tags”

In this file, we explain the procedure of using the Gibbs sampling algorithm to map ambiguous sequence tags step by step.





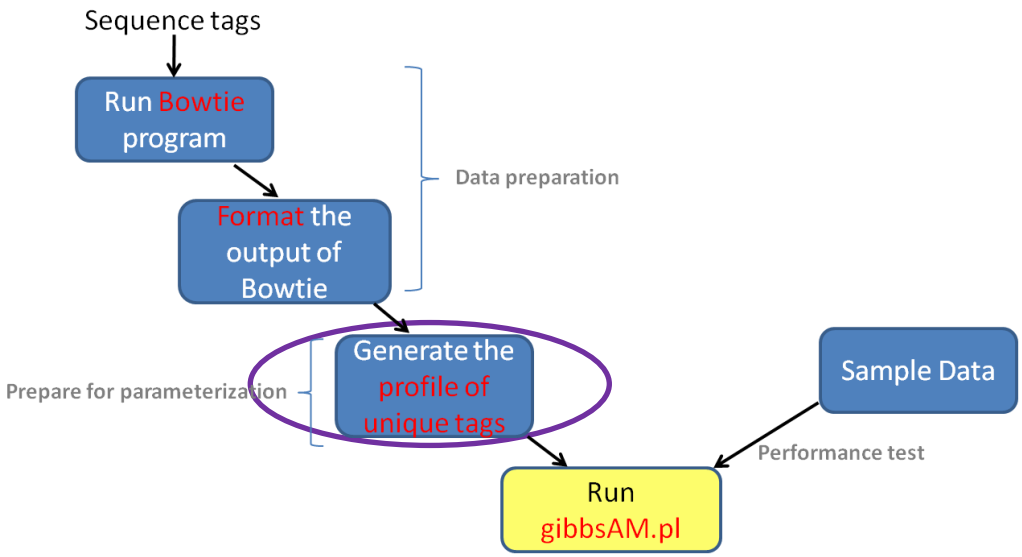
1. In order to map ambiguous tags, you need first get the initial mapping of all tags to the genome using the program Bowtie.
2. The Bowtie program could be obtained here: http://bowtie-bio.sourceforge.net/index.shtml
3. Reference:
   Langmead, B. et al, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, Genome Biology, 10, R25



1. The output file from Bowtie program needs to be formatted in order to run "gibbsAM.pl".
2. Run the script "format_bowtie_result.pl" to change the format.
3. Command line:
   perl format_bowtie_result.pl –p directory –i Bowtie output file -o name of output file
   -p: directory where the Bowtie output file (initial mapping of tags) is located;
   -i : the Bowtie output file (initial mapping of tags), including both unique and ambiguous tags;
   -o: the name of the output file. The default name is "mapping_result.bed"
4. The resulting output file is in this format:
   "tag_id chr>position>strand,chr>position>strand,…"

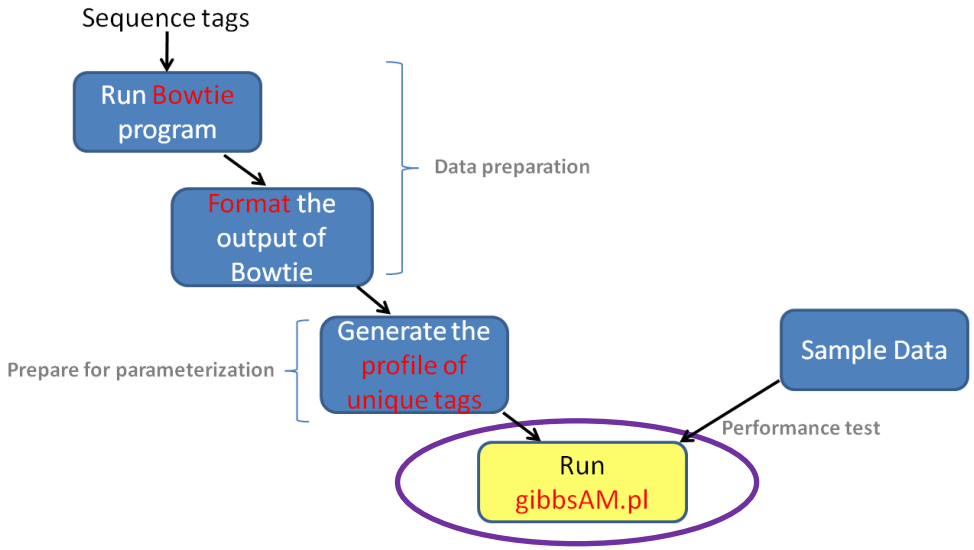
   e.g.

   HWI-EAS229_75_30DY0AAXX:4:1:0:1282/1 chr18>6452262>+,
   HWI-EAS229_75_30DY0AAXX:4:1:0:1282/2 chr18>6452351>+,chr4>66122359>-,



1. In order to set the parameters for the algorithm, the unique tags need to be organized.
2. Run "get_unique_file.pl" to organize the unique tags.
3. Command line:
   perl get_unique_file.pl -p directory -i the formatted mapping file -o name of output file -l length of adjacent region
   -p: directory where the formatted mapping file is located
   -i : the formatted mapping file generated by "format_bowtie_result.pl"
   -o: name of output file, the default name is "unique_screen_result.bed"
   -l : length of adjacent region for co-located tags, the default value is 147
4. The resulting output file has this format
   "chr position_start position_end unique_tag_count"
   e.g.

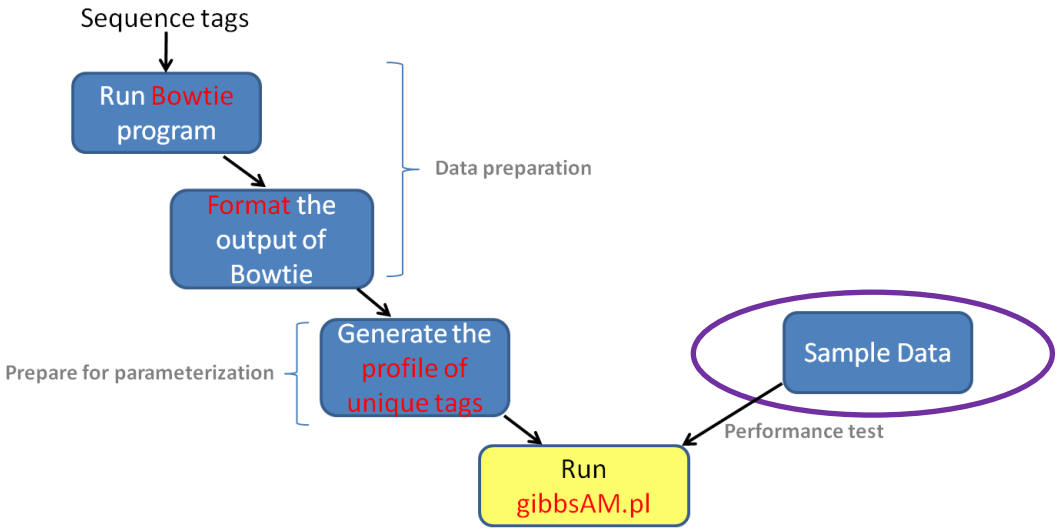
chr1	795260	795406	226
chr1	830067	830213	166



1. After the preparation through the steps above, run "gibbsAM.pl" to apply the Gibbs sampling method to assign each ambiguous tag to a specific genomic site.
2. Command line:

   perl gibbsAM.pl -p path -f mapping_result.file -u unique_mapping.file -o output_name -l region_length -r maximal_tag_number -a ambiguous_confidence -m iteration_number

   -p: the directory where the files (formatted mapping file & the file for unique tags) are located.
   -f : the formatted mapping file generated by "format_bowtie_result.pl".
   -u: the file for unique tags generated by "get_unique_file.pl".
   -o: name of the output file.
   -l : the length of the adjacent region for co-located tags. The default value is 147.
   -r : the maximal tag count used to construct the likelihood table. The default value is 50.
   -a: the relative confidence of ambiguous tags. The default value is 0.2.
   -m: the number of iterations. The default value is 5.



1. The sample data are the libraries we used to evaluate the algorithm performance.
2. "benchmark.zip": files with the real genomic sites in the benchmarks. It contains 2 files. "benchmark_8lib.bed" is the benchmark for the 8 smaller sequence libraries and "benchmark_biglib.bed" is for the bigger library.
3. "tag.zip": files with the short sequence tags. There are 9 files inside the folder. "tag_lib*.fastq" are tags for the 8 smaller libraries respectively. "tag_biglib.fastq" is the file with tags for the bigger library.
4. "initial_map.zip": contains
   1) "mapping_result_lib*.bed" & "mapping_result_biglib.bed": the formatted files (generated by "format_bowtie_result.pl")with initial mapping results from Bowtie program;
   2) "unique_screen_result_lib*.bed" & "unique_screen_result_biglib.bed": the files for unique tags (generated by "get_unique_file.pl").
5. In order to test the algorithm, files in "initial_map.zip" are enough. If you want to make sure the initial mappings are correct, you can run Bowtie on files in "tag.zip" and then run "format_bowtie_result.pl" and "get_unique_file.pl".
6. Files in the "benchmark.zip" could be used to compare the final map of tags to evaluated the performances.