ChIP-AP – Creating the genome files yourself

1. You need a master folder to contain all the genome files. Technically, you can name it anything you want and put it anywhere you want. However, ChIP-AP GUI have this genome folder path set to **/home/username/genomes** by default. While you can easily change the genome folder path there in two clicks and one second, for simplicity reason, it will be **/home/username/genomes** in this guide.

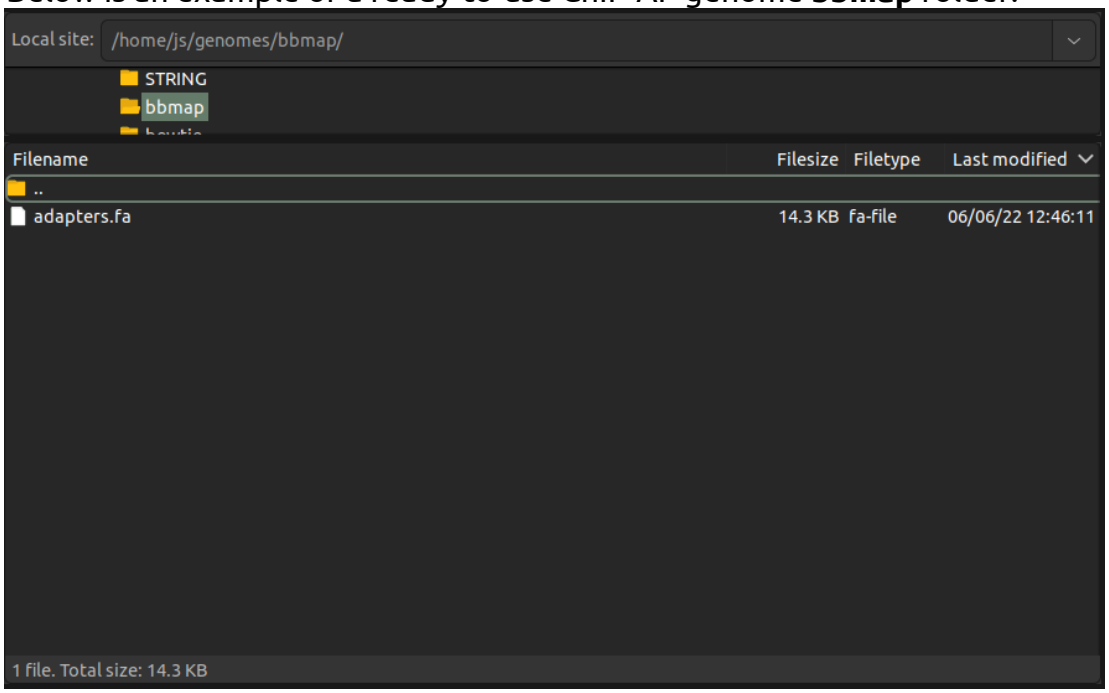
2. You need three folders inside the master folder: **bbmap**, **bwa**, and **GEM**. These ones you have to name them correctly.

3. For the **bbmap** folder, you just need one file: **adapters.fa**, which a the list of known sequencing adapters, provided by bbmap. You can find it here:
   https://github.com/BioInfoTools/BBMap/blob/master/resources/adapters.fa

   Or if you want an updated version of it, you can download the bbmap package:
   https://sourceforge.net/projects/bbmap/
   ... extract it, then look for **adapters.fa** in the resource subfolder.

   Either way, put the **adapters.fa** file inside the bbmap folder.

   Below is an example of a ready-to-use ChIP-AP genome **bbmap** folder:

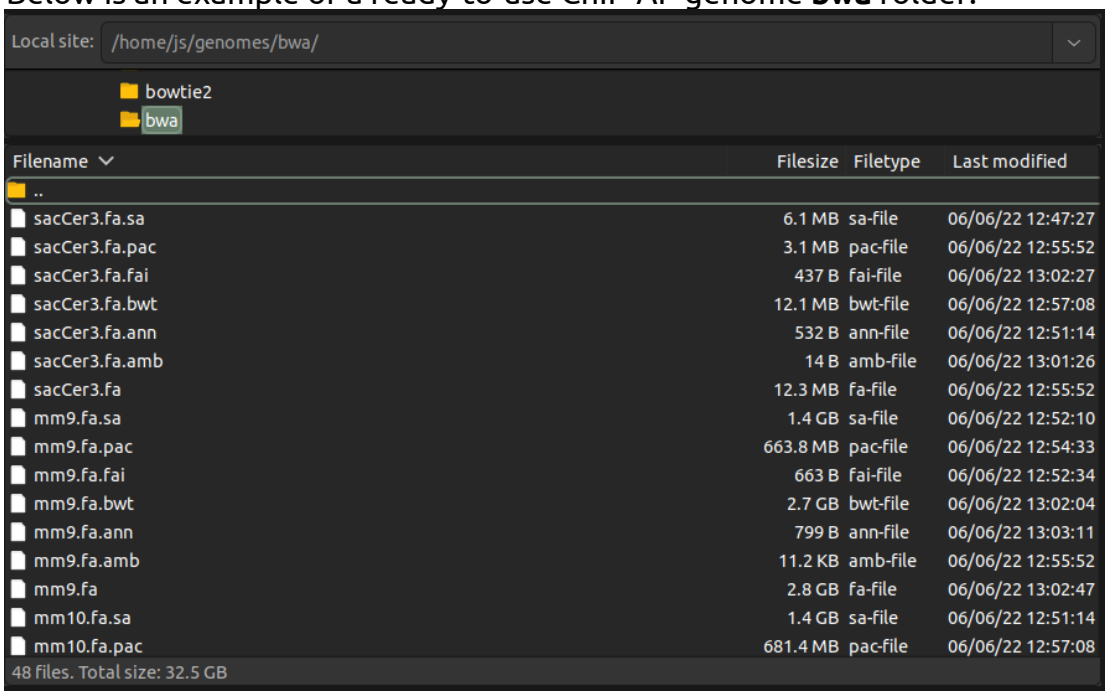


4. For the **bwa** folder, you will first need the genome FASTA file (**.fa**) for the sample organism of your ChIP-seq dataset. By default, ChIP-seq comes with six genome assemblies, which **.fa** compressed (**.gz**) files can be downloaded from these links:
   http://hgdownload.cse.ucsc.edu/goldenPath/hg38/bigZips/hg38.fa.gz
   http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/hg19.fa.gz
   http://hgdownload.cse.ucsc.edu/goldenPath/mm10/bigZips/mm10.fa.gz
   http://hgdownload.cse.ucsc.edu/goldenPath/mm9/bigZips/mm9.fa.gz
   http://hgdownload.cse.ucsc.edu/goldenPath/dm6/bigZips/dm6.fa.gz
   http://hgdownload.cse.ucsc.edu/goldenPath/sacCer3/bigZips/sacCer3.fa.gz
   ... then put the downloaded .fa.gz file inside the bwa folder.

For each of the genome assembly **.fa.gz** file, you need to:

1. Unzip the archive with this command: ***gunzip [path to .fa.gz file]***. Done correctly, this will decompress the **.fa.gz** extension into **.fa.**
2. Generate the required indices for bwa with this command: ***bwa index [path to .fa file]***. Done correctly, you will have five index files for one **.fa** file (**.sa**, **.pac**, **.bwt**, **.ann**, **.amb**).
3. Generate the required index for samtools with this command: ***samtools faidx [path to .fa file]***. Done correctly, you will have one index file for one **.fa** file (**.fai**)

Below is an example of a ready-to-use ChIP-AP genome **bwa** folder:



5. For the **GEM** folder, for each genome assembly, you will need a list chromosome sizes (**.chrom.sizes**), which can be downloaded from these links:
http://hgdownload.cse.ucsc.edu/goldenPath/hg38/bigZips/hg38.chrom.sizes
http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/hg19.chrom.sizes
http://hgdownload.cse.ucsc.edu/goldenPath/mm10/bigZips/mm10.chrom.sizes
http://hgdownload.cse.ucsc.edu/goldenPath/mm9/bigZips/mm9.chrom.sizes
http://hgdownload.cse.ucsc.edu/goldenPath/dm6/bigZips/dm6.chrom.sizes
http://hgdownload.cse.ucsc.edu/goldenPath/sacCer3/bigZips/sacCer3.chrom.sizes

   Then, for each genome assembly, you will need to create a folder. The folder name must be **[genome assembly]_Chr_FASTA**. For example: **hg38_Chr_FASTA** for the **hg38** genome assembly.

   Afterwards, use the '**chromfa_splitter.py**' to generate chromosome-wise **.fa** files inside the folder created above from the genome-wide **.fa** file in **bwa** folder. To do so, use the command: ***chromfa_splitter.py --fa [path to .fa file] --output [path to the chromosome-wise .fa folder]***. For example:

   chromfa_splitter --fa /home/username/genomes/bwa/hg38.fa --output home/username/genomes/GEM/hg38_Chr_FASTA/

Done correctly, you will have these chromosome-wise **.fa** files inside the folder:

Local site: /home/js/genomes/GEM/hg38_Chr_FASTA/

hg19_Chr_FASTA
hg38_Chr_FASTA

Filename	Filesize	Filetype	Last modified
..			
chrY.fa	58.4 MB	fa-file	06/06/22 13:04:43
chrX.fa	159.2 MB	fa-file	06/06/22 13:05:17
chr9.fa	141.2 MB	fa-file	06/06/22 13:04:47
chr8.fa	148.1 MB	fa-file	06/06/22 13:05:02
chr7.fa	162.6 MB	fa-file	06/06/22 13:04:50
chr6.fa	174.3 MB	fa-file	06/06/22 13:05:22
chr5.fa	185.2 MB	fa-file	06/06/22 13:04:39
chr4.fa	194.1 MB	fa-file	06/06/22 13:04:57
chr3.fa	202.3 MB	fa-file	06/06/22 13:05:15
chr22.fa	51.9 MB	fa-file	06/06/22 13:04:51
chr21.fa	47.7 MB	fa-file	06/06/22 13:04:33
chr20.fa	65.8 MB	fa-file	06/06/22 13:04:41
chr2.fa	247.1 MB	fa-file	06/06/22 13:05:12
chr19.fa	59.8 MB	fa-file	06/06/22 13:04:58
chr18.fa	82.0 MB	fa-file	06/06/22 13:04:45
chr17.fa	85.0 MB	fa-file	06/06/22 13:04:56

24 files. Total size: 3.2 GB

*Note: Your chromosome-wise **.fa** files may be different than what are shown, because ChIP-AP's genome-wide **.fa** file has been filtered. However, this should not break the pipeline or cause any difference in the final results.*

6. Lastly, you need the **read distribution .txt** files provided by **GEM**. Available here: https://groups.csail.mit.edu/cgs/gem/download/Read_Distribution_default.txt

   *Note: There are other read distributions provided by GEM, but for ChIP-AP you will only ever need the **default read distribution file**.*

   Below is an example of a ready-to-use ChIP-AP genome **GEM** folder:

Local site: /home/js/genomes/GEM/

genomes
GEM

Filename	Filesize	Filetype	Last modified
..			
saccer3.chrom.sizes	218 B	sizes-file	06/06/22 13:03:11
mm9.chrom.sizes	320 B	sizes-file	06/06/22 13:03:11
mm10.chrom.sizes	320 B	sizes-file	06/06/22 13:03:12
hg38.chrom.sizes	365 B	sizes-file	06/06/22 13:03:11
hg19.chrom.sizes	365 B	sizes-file	06/06/22 13:03:11
dm6.chrom.sizes	100 B	sizes-file	06/06/22 13:03:11
Read_Distribution_default.txt	12.8 KB	txt-file	06/06/22 13:03:11
Read_Distribution_ChIP-exo.txt	7.6 KB	txt-file	06/06/22 13:03:11
Read_Distribution_CLIP.txt	25.4 KB	txt-file	06/06/22 13:03:11
Read_Distribution_BP.txt	494 B	txt-file	06/06/22 13:03:11
sacCer3_Chr_FASTA		Directory	06/06/22 13:06:09
mm9_Chr_FASTA		Directory	06/06/22 13:06:45
mm10_Chr_FASTA		Directory	06/06/22 13:04:31
hg38_Chr_FASTA		Directory	06/06/22 13:05:17
hg19_Chr_FASTA		Directory	06/06/22 13:06:03
dm6_Chr_FASTA		Directory	06/06/22 13:06:08

10 files and 6 directories. Total size: 47.9 KB