

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, the ChIP-AP GUI has this genome folder path set to **/home/username/genomes** by default. While you can easily change the genome folder path there in two clicks, for simplicity, it will be kept as **/home/username/genomes** in this guide.

NOTE: This guide assumes that all required programs for ChIP-AP are installed and set up correctly.

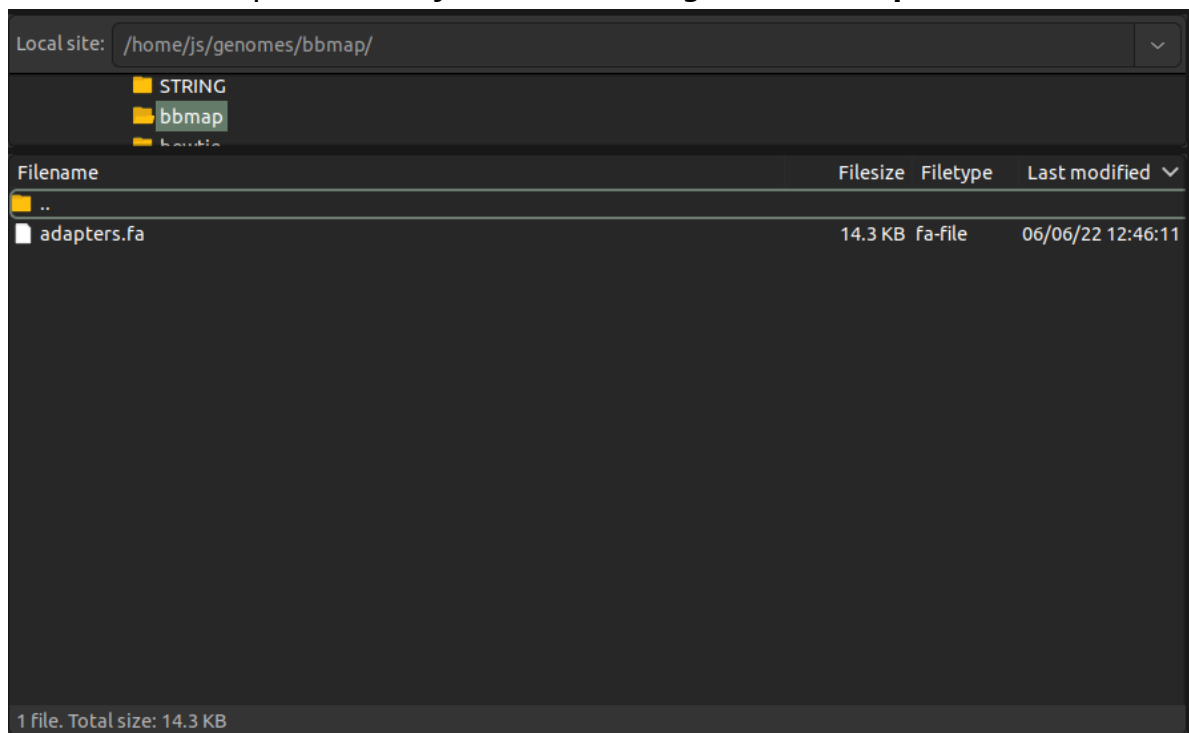
2. You need three folders inside the master folder: **bbmap**, **bwa**, and **GEM**. For these ones, you will have to name them correctly and are CASE SENSITIVE.
3. For the **bbmap** folder, you just need one file: **adapters.fa**, which is the list of known sequencing adapters, provided by bbmap. You 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.

NOTE: If you are using custom adaptors for any reason in your sequencing, make sure to insert those sequences into the adapters.fa file. (Good idea to back up the default file before modifying it though...)

Once you have located/updated the **adapters.fa** file, put it inside the **bbmap** folder.

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



- 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 were downloaded from these links:

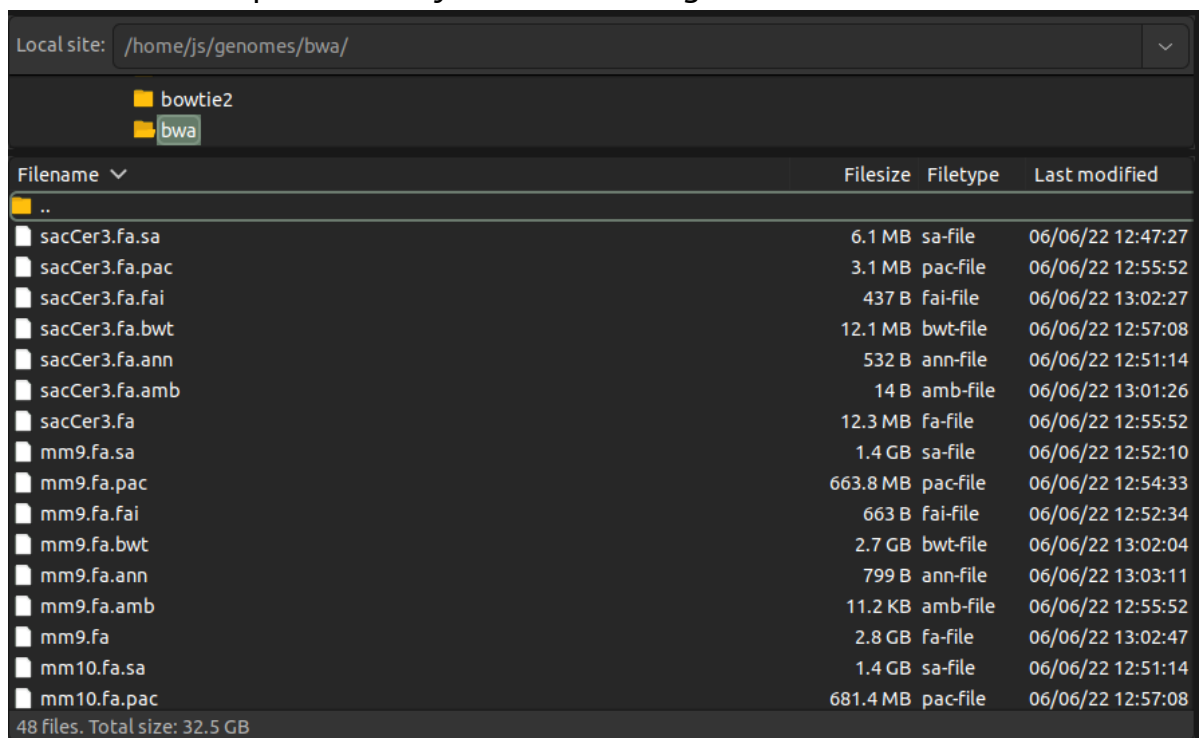
<http://hgdownload.cse.ucsc.edu/goldenPath/hg38/bigZips/hg38.fa.gz>
<http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/hg19.fa.gz>
<http://hgdownload.cse.ucsc.edu/goldenPath/mm10/bigZips/mm10.fa.gz>
<http://hgdownload.cse.ucsc.edu/goldenPath/mm9/bigZips/mm9.fa.gz>
<http://hgdownload.cse.ucsc.edu/goldenPath/dm6/bigZips/dm6.fa.gz>
<http://hgdownload.cse.ucsc.edu/goldenPath/sacCer3/bigZips/sacCer3.fa.gz>

Once downloaded, place the downloaded **.fa.gz** file inside the **bwa** folder.

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

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

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



Filename	Filesize	Filetype	Last modified
..			
sacCer3.fa.sa	6.1 MB	sa-file	06/06/22 12:47:27
sacCer3.fa.pac	3.1 MB	pac-file	06/06/22 12:55:52
sacCer3.fa.fai	437 B	fai-file	06/06/22 13:02:27
sacCer3.fa.bwt	12.1 MB	bwt-file	06/06/22 12:57:08
sacCer3.fa.ann	532 B	ann-file	06/06/22 12:51:14
sacCer3.fa.amb	14 B	amb-file	06/06/22 13:01:26
sacCer3.fa	12.3 MB	fa-file	06/06/22 12:55:52
mm9.fa.sa	1.4 GB	sa-file	06/06/22 12:52:10
mm9.fa.pac	663.8 MB	pac-file	06/06/22 12:54:33
mm9.fa.fai	663 B	fai-file	06/06/22 12:52:34
mm9.fa.bwt	2.7 GB	bwt-file	06/06/22 13:02:04
mm9.fa.ann	799 B	ann-file	06/06/22 13:03:11
mm9.fa.amb	11.2 KB	amb-file	06/06/22 12:55:52
mm9.fa	2.8 GB	fa-file	06/06/22 13:02:47
mm10.fa.sa	1.4 GB	sa-file	06/06/22 12:51:14
mm10.fa.pac	681.4 MB	pac-file	06/06/22 12:57:08

48 files. Total size: 32.5 GB

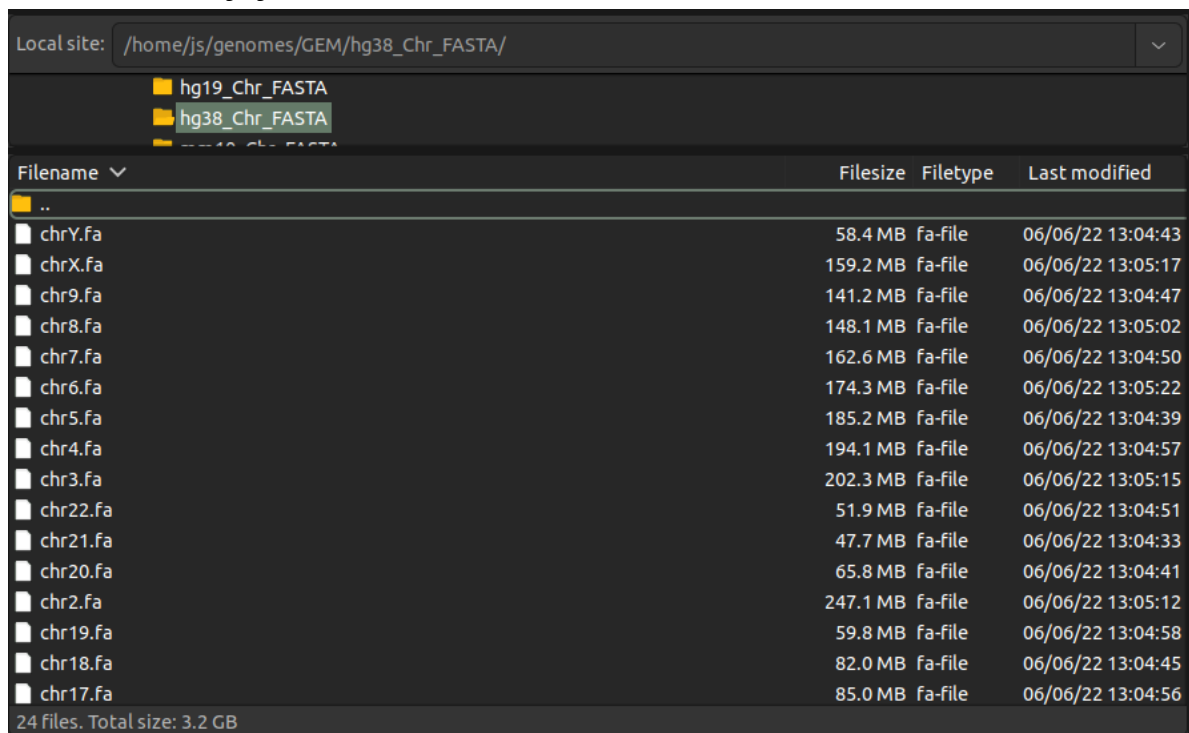
- For the **GEM** folder, for each genome assembly, you will need a list chromosome sizes (**.chrom.sizes**), which can be downloaded from these links:

<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**' ([link](#)) script to generate chromosome-wise **.fa** files inside the folder created above from the genome-wide **.fa** file in the **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/

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



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:

> 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_Chromosome_FASTA		Directory	06/06/22 13:06:09
mm9_Chromosome_FASTA		Directory	06/06/22 13:06:45
mm10_Chromosome_FASTA		Directory	06/06/22 13:04:31
hg38_Chromosome_FASTA		Directory	06/06/22 13:05:17
hg19_Chromosome_FASTA		Directory	06/06/22 13:06:03
dm6_Chromosome_FASTA		Directory	06/06/22 13:06:08

10 files and 6 directories. Total size: 47.9 KB