

1. Prepare the chromosome map (cmap) file:

```
#chr_ID(must be digit)<TAB>chr_name<TAB>FASTA reference(must exist)<TAB>
Double-Encoded Reference<NEWLINE>
1      1      /share/reference/genomes/huref/upper_case/chr1.fa
        /share/reference/genomes/double_encoded/huref/de_chr1.fa
2      2      /share/reference/genomes/huref/upper_case/chr2.fa
        /share/reference/genomes/double_encoded/huref/de_chr2.fa
...
23     X      /share/reference/genomes/huref/upper_case/chr23.fa
        /share/reference/genomes/double_encoded/huref/de_chr23.fa
24     Y      /share/reference/genomes/huref/upper_case/chr24.fa
        /share/reference/genomes/double_encoded/huref/de_chr24.fa
```

The 4th field is the doubleEncode sequence, required by SNP pipeline but not required by matching or mate-pairing pipeline, so you can ignore it here.

2. Run the matching wrapper: `matching_large_genomes_cmap_save_script.pl`

Usage:

```
matching_large_genomes_cmap_save_script.pl [options] --help -csfasta
-dir -cmap -t -e -p -a -z -reads -post -incremental -[no]compact
-tempdir -schema
```

Options:

`-help` Display the help message you are looking at.

[REQUIRED]

`-csfasta` csfasta file

`-dir` Output Directory

`-cmap` Chromosome map file

`-t` Tag Length

`-e` Number of Errors

[OPTIONAL]

`-p` Pattern: defaults all to 1's

`-a` Count Adjacent Errors as 1: 0 = no : 1 = valid adjacent errors : 2 = all adjacent errors : defaults to 0

`-z` Maximum Number of Hits: Default 1000

`-reads` Number of Reads per Subset: defaults to no subsets

`-post` Run Post Matching Only

`-incremental` Flag to perform incremental matching by removing reads which are already mapped.

`-compact` Flag to output intermediate matching results in compact form, default is on, use `-nocompact` to turn it off.

`-tempdir` Space used for scratch files, default /scratch.

`-schema` Schema file for the mapping, if not set, standard schema

will be used.

Here are more information:

- csfata: the reads file, either F3 or R3
- dir: output
- cmap: the cmap file explained at the first step
- t, -e: For human, we recommend 25\_2 if the tag length is 25bp.
- p: default set to all 1's at the length of the tag, but you can mask a position by changing the number to 0 at that position
- z: To reduce match file size, we recommend setting z=10 for human matching
- reads: used for splitting input reads file
- compact: the raw matching results in each chr dir saved without sequence lines and lines without hits are omitted as well.
- tempdir: directory used by the mapper to write scratch files, a drive that is local to each compute node is preferred. If no such drive is available, one can choose "." to use the network drives where each matching job is taking place.
- schema: used for custom schema.

The scripts are saved in the scripts directory under -dir, and there is a file named JOB\_LIST.txt in -dir which contains all the locations and names for all the scripts.

3. Submit the jobs described in JOB\_LIST.txt to compute farms as described in RELEASE\_NOTES.