

Short Sequencing Read Mapping

BWA for short read alignment

Read the manual/documentation for BWA

```
module load bwa
bwa
bwa index
bwa mem
```

There is an alternative faster implementation of bwa called **bwa-mem2**. See some of the documentation.

```
module load bwa-mem2
bwa-mem2
bwa-mem2 index
bwa-mem2 mem
```

Index genome

It is necessary to index the genome in preparation for alignment.

```
#!/bin/bash -l
#SBATCH -p short -N 1 -n 1 -c 1 --mem 2gb --time 1:00:00
#SBATCH -J bwaindex
#SBATCH -o logs/%x.%j.log
module load bwa
```

```
set -euo pipefail
GENOME=S_enterica_CT18.fasta
bwa index $GENOME
```

`bwa index` runs on a single CPU, so `-c 1` is enough. Remember to `mkdir -p logs` before you `sbatch` a script that writes its log to `logs/`.

Align reads

```
#!/bin/bash -l
#SBATCH -p short -N 1 -n 1 -c 8 --mem 16gb --time 2:00:00
#SBATCH -J bwamem
#SBATCH -o logs/%x.%j.log
```

```
module load bwa
module load samtools
```

```
set -euo pipefail
CPU=${SLURM_CPUS_PER_TASK:-1}
mkdir -p ~/bigdata/Short_read_aligning
cd ~/bigdata/Short_read_aligning
mkdir -p fastq
# -f: replace the links if they already exist (so the script can be rerun)
ln -sf /bigdata/gen220/shared/data/S_enterica/*.fastq.gz fastq
ln -sf /bigdata/gen220/shared/data/S_enterica/S_enterica_CT18.fasta .
ln -sf /bigdata/gen220/shared/data/S_enterica/acc.txt .
GENOME=S_enterica_CT18.fasta
if [ ! -f $GENOME.sa ]; then
    bwa index $GENOME
fi
```

```

for acc in $(cat acc.txt)
do
    FWDREAD=fastq/${acc}_1.fastq.gz
    REVREAD=fastq/${acc}_2.fastq.gz

    bwa mem -t $CPU $GENOME $FWDREAD $REVREAD > ${acc}.sam
    samtools fixmate --threads $CPU -O bam ${acc}.sam ${acc}_fixmate.bam
    samtools sort --threads $CPU -O BAM -o ${acc}.bam ${acc}_fixmate.bam
    samtools index ${acc}.bam
done

```

`samtools sort` uses about 768 MB of memory per thread by default, which is why this asks for more memory than `bwa` alone would need. See [UNIX IV](#) for how to choose `-c`, `--mem` and `--time`.

samtools

A multi-use tool for investigating SAM/BAM file format data.

```

module load samtools
samtools

```

See the menu of options and explore a bit.

Also see `sambamba`

```

module load sambamba
sambamba

```

Visualizing depth of coverage

Interactively - you can use `samtools`

```

module load samtools
samtools tview SRR10574912.bam
# to see the reference genome loaded as well add this option
samtools tview SRR10574912.bam --reference S_enterica_CT18.fasta

```

SNP calling

There are many standardized SNP calling pipelines. [GATK](#) provides a robust pipeline that can be used.

[Samtools/BCFTools](#) are also useful and straight forward.

[freebayes](#) is another very useful pipeline for non-model systems.

Samtools/BCFTools SNP and INDEL calling

[Workflows from the htlib](#)

```

#!/bin/bash -l
#SBATCH -p short -N 1 -n 1 -c 4 --mem 16gb --time 2:00:00
#SBATCH -J bcftools_call
#SBATCH -o logs/%x.%j.log
module load samtools
module load bcftools

set -euo pipefail
CPU=${SLURM_CPUS_PER_TASK:-1}
GENOME=S_enterica_CT18.fasta

```

```

# need to make a string which is all the bam files you want to process
# but if we do *.bam it will catch the intermediate bam files that are in the folder
m=""
for a in $(cat acc.txt)
do
    m="$a.bam $m"
done

VCF=Salmonella.vcf.gz
VCFFILTER=Salmonella.filtered.vcf.gz
bcftools mpileup -Ou -f $GENOME $m | bcftools call --threads $CPU -vm0 z -o $VCF
tabix -p vcf $VCF
bcftools stats -F $GENOME -s - $VCF > $VCF.stats
mkdir -p plots
# plot-vcfstats needs python3 with matplotlib; don't stop the job if it fails
plot-vcfstats -p plots/ $VCF.stats || echo "plot-vcfstats failed - skipping the plots"
bcftools filter -O z -o $VCFFILTER -s LOWQUAL -i '%QUAL>10' $VCF

```

Genome Browsers

We will do more on genome browsers later in the course. But if you want to see how to visualize genome you can see some of these tools.

IGV

[IGV](#) - High-performance visualization tool for interactive exploration of large, integrated genomic datasets. It supports a wide variety of data types, including array-based and next-generation sequence data, and genomic annotations.

JBrowse

[JBrowse2](#) provides google-maps style interface to genomes

Public genome browsers

Many browsers allow upload of aligned data (bam files) to integrate local data with public genome resources.

- [Ensembl](#), [Ensembl Genomes](#)
- [UCSC Genome Browser](#)
- [WormBase](#), [FlyBase](#)
- [TAIR](#) - Arabidopsis, [Phytozome](#)
- [EuPathDB](#), [JGI Genomes](#)
- [IMG/M](#) - [JGI](#)

Displaying data in EnSEMBL

Go to [Ensembl Site for Salmonella enterica subsp. enterica serovar Typhi str. CT18](#)

See [the EnSEMBL tutorial](#) on how to add a BAM file track (note this only works if you have aligned reads to the SAME ASSEMBLY that is in Ensembl).

Click on “Display your data in Ensembl Bacteria”

Make a link on the web for your data. Follow directions on [HPCC site](#)

```

mkdir -p ~/.html/share
cd ~/.html/share
ln -s ~/bigdata/Short_read_aligning . # or wherever you were doing

```

Now you can use the URL `http://cluster.hpcc.ucr.edu/~YOURLOGIN/share/Short_read_aligning` and the `.bam` files that are in there.