Short Sequencing Read Maping

BWA for short read alignment

Index genome

It is necessary to index the genome in preparation of alignemnt.

```
#SBATCH -p short -N 1 -n 2 --mem 2gb
module load bwa
GENOME=S_enterica_CT18.fasta
bwa index $GENOME
```

Align reads

```
#SBATCH -p short -N 1 -n 16 --mem 4gb

module load bwa
module load samtools
CPU=16
mkdir -p fastq
ln -s /bigdata/gen220/shared/data/S_enterica/*.fastq.gz fastq
ln -s /bigdata/gen220/shared/data/S_enterica/S_enterica_CT18.fasta
ln -s /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 -O bam ${acc}.sam ${acc}_fixmate.bam
    samtools sort --threads $CPU -O BAM -o ${acc}.bam ${acc}_fixmate.bam
    samtools index ${acc}.bam
done
```

Visualizing depth of coverage

Interactively - you can use `samtools`

```
module load samtools
samtools tview SRR10574912.bam
```

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 htslib

```
#SBATCH -p batch -N 1 -n 4 --mem 16gb
module unload perl
module load samtools
module load bcftools
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
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 -vmO z -o $VCF
tabix -p vcf $VCF
bcftools stats -F $GENOME -s - $VCF $VCF.stats
mkdir -p plots
plot-vcfstats -p plots/ $VCF.stats
bcftools filter -O z -o $FILTERED -s LOWQUAL -i'%QUAL>10' $VCF
```

Genome Browsers

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.

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