Read mapping and simple variants

Adam m France BINF 6203 UNCC
afrance3@uncc.edu

Download files SRR1763770 - 1763781 via fastq-dump :
$ fastq-dump SRR1763770 - 781

and chloroplast reference file(s) NC_007898:
https://www.dropbox.com/sh/mwmt6twf5dvcggf/AAD_OtrIhtdMTs4fhUSShgA5a?dl=0

Outline of Workflow:

- Quality control with Fastqc and Trimmomatic.
- Mapping reads with bowtie2.
- Converting a SAM into a BAM (binary, smaller file, same information) using samtools
- Indexing a BAM using samtools
- Loading a BAM and corresponding reference genome into a genome browser

1. Create the bowtie2 index with the reference file:
$ bowtie2-build **NC_007898.fasta** NC_007898
-This creates several files with the extension ".bt2"

2. Trimmomatic trimming:
This step was repeated for all 12 samples.
$ java -jar trimmomatic-0.39.jar SE **input_SRR_file output_file_name**
ILLUMINACLIP:ionseq.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:20 MINLEN:10

Note that ionseq.fa is with the following sequences:
> A1 5'
CCTCTCTATGGGCAGTCGGTGAT
> TRP1 3'
CCATATCATCCCTGCGTGTCTCCCACTCAG

3. Run mapping with bowtie:
Repeat this step for all 12 samples:
$ bowtie2 -x NC_007898 -U **SRR-trim.fastq** -S **SRR.sam**

4. Converting SAM to a BAM
Repeat for all 12 samples:
Convert SAM to BAM files with samtools, then sort bam file:
$ samtools view -uS **SRR.sam** | samtools sort - -o **SRR-sort.bam**

5. Bam Index - Repeat for all 12
$ samtools index **SRR-sort.bam**

6. Produce VCF files - Repeat for all 12
$ samtools mpileup -uf NC_007898.fasta **SRR-sort.bam** | bcftools call -c - > **SRR-sort.vcf**

7. Loading a BAM and corresponding reference genome into a genome browser:
A . Select “Genomes” then “Load Genome from file” and browse for NC_007898.fasta
B. Select “File” then “Load from file” and browse for NC_007898.gff and SRR-sort.BAM files
C. go to “Tools” menu, the select “IGVtools” and then “Index” and input the VCF files

Results:

Figure1. A table of all 12 alignments. About half were above 85%, the other half had ranging % below 85. This could be due to the QC step, where there was not a specific trimming parameter for ion torrent data.

SRR	Cultivar	% Alignment
1763770	Cherry	86.42
1763771	Purple Calabash	91.73
1763772	LA2377	97.30
1763773	Hardin’s Miniature Tomato	35.01
1763774	Baxter’s Early Bush Cherry	87.52
1763775	Black Cherry	86.75
1763776	Cherokee Purple	35.14
1763777	Delicious	52.54
1763778	Jersey Devil	78.22
1763779	Yellow Surfer	34.59
1763780	German Cherry	96.69
1763781	Mortgage Lifter	95.81

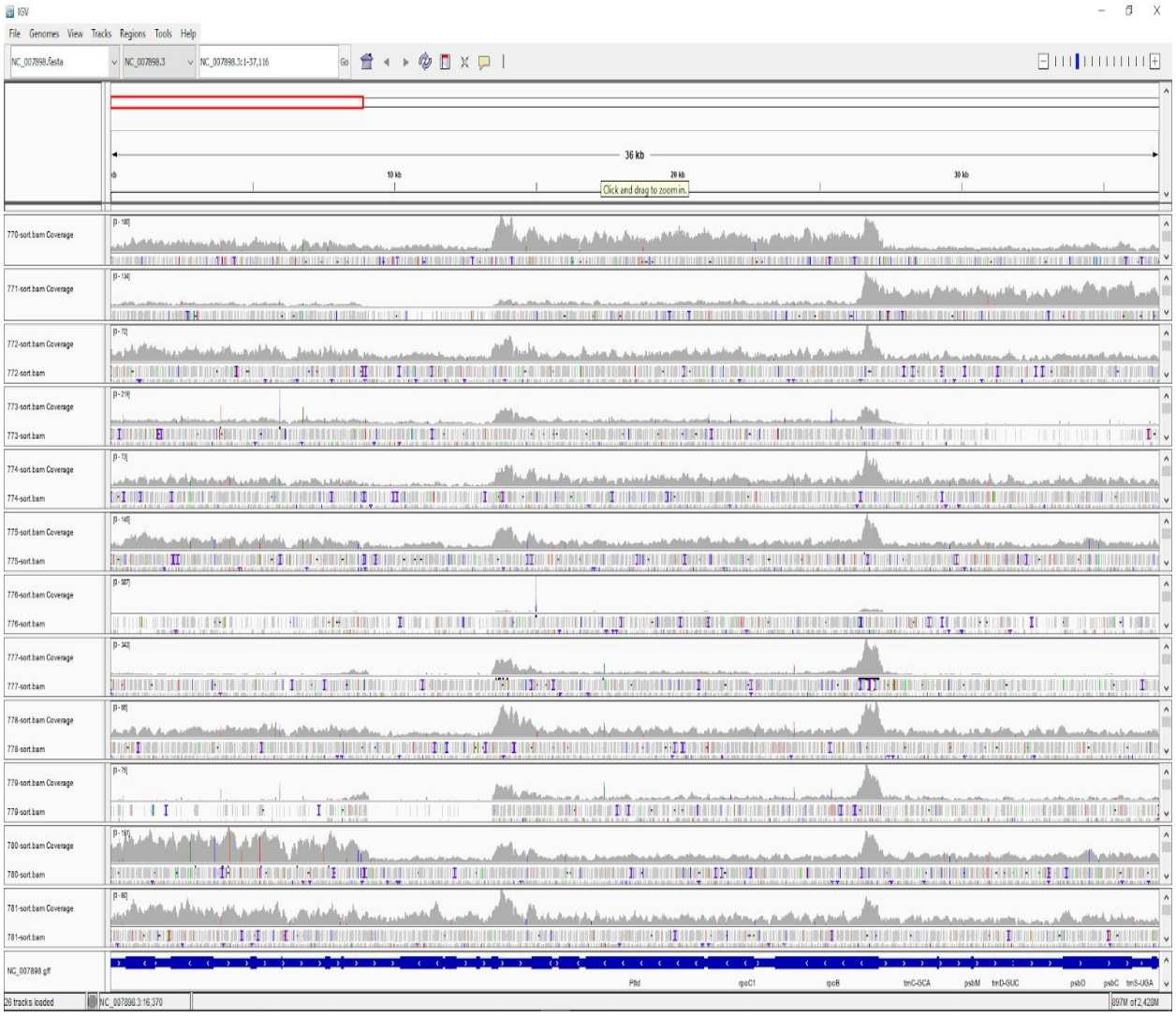

Figure 2. The IGV browser with all 12 tracks loaded. The peaks represent coverage, where higher peaks means a greater number of contig coverage for that part of the sequence. The purple bars represent insertions, of which there appears to be quite a few across each sample. Each sample seems to have good coverage, with the exception of 776, the reason for this is unknown.

Discussion: From figure 2 it is obvious that there are thousands of both insertions (purple) and SNPs (red,green,blue,orange). It is not clear what the biological significance is, if any. Further investigation would reveal any connections between SNP and insertions and biological impact?