Assembly Exercise

Turning reads into genomes

J. Craig Venter™ INSTITUTE

Where we are

- 13:30-14:00 – Primer Design to Amplify Microbial Genomes for Sequencing
- 14:00-14:15 – Primer Design Exercise
- 14:15-14:45 – Molecular Barcoding to Allow Multiplexed NGS
- 14:45-15:15 – Processing NGS Data – de novo and mapping assembly
- 15:15-15:30 – Break
- **15:30-15:45 – Assembly Exercise**
- 15:45-16:15 – Annotation
- 16:15-16:30 – Annotation Exercise
- 16:30-17:00 – Submitting Data to GenBank

Log onto ILRI cluster

- Log in to HPC using ILRI instructions
- NOTE: All the commands here are also in the file -

`assembly_hands_on_steps.txt`

- If you are like me, it may be easier to cut and paste Linux commands from this file instead of typing them in from the slides

Start an interactive session on larger servers

- The *interactive* command will start a session on a server better equipped to do genome assembly

```
$ interactive
```

- Switch to *csh* (I use some csh features)

```
$ csh
```

- Set up Newbler software that will be used

```
$ module load 454
```

A norovirus sample sequenced on both 454 and Illumina

- The vendors use different file formats

```
unknown_norovirus_454.GACT.sff
unknown_norovirus_illumina.fastq
```

- I have converted these files to additional formats for use with the assembly tools

```
unknown_norovirus_454_convert.fasta
unknown_norovirus_454_convert.fastq
unknown_norovirus_illumina_convert.fasta
```

J. Craig Venter™
INSTITUTE

Set up and run the Newbler de novo assembler

- Create a new de novo assembly project

```
$ newAssembly de_novo_assembly
```

- Add read data to the project

```
$ addRun de_novo_assembly
unknown_norovirus_454.GACT.sff
```

```
$ addRun de_novo_assembly
unknown_norovirus_illumina_convert.fasta
```

- Run the project

```
$ runProject de_novo_assembly
```

J. Craig Venter™ INSTITUTE

Look at the output sequence and use it to find a reference

```
$ more de_novo_assembly/assembly/454LargeContigs.fna
```

- Find the longest contig and use it in a BLAST search to find a promising reference genome at NCBI

```
In a browser, go to http://www.ncbi.nlm.nih.gov/ ->
BLAST -> nucleotide blast, paste in sequence from
de_novo_assembly/assembly/454Scaffolds.fna, and run
```

- Copy and paste the fasta sequence of a complete genome

```
$ vi reference_genome.fasta
```

Set up and run the Newbler mapping assembler

- Create a new mapping assembly project

```
$ newMapping mapping_assembly
```

- Set the reference sequence

```
$ setRef mapping_assembly reference_genome.fasta
```

- Add read data to the project

```
$ addRun mapping_assembly
unknown_norovirus_454.GACT.sff
$ addRun mapping_assembly
unknown_norovirus_illumina_convert.fasta
```

- Run the project

```
$ runProject mapping_assembly
```

J. Craig Venter™
INSTITUTE

Look at the output data from the mapping assembly

- Look at High Confidence Differences between the NGS data and the reference

```
$ more mapping_assembly/mapping/454HCDiffs.txt
```

- Look at our genome sequence

```
$ more mapping_assembly/mapping/454LargeContigs.fna
```

- We will use this genome sequence later for annotation exercise

J. Craig Venter™ INSTITUTE

Mapping Assembly with BWA and SAMTOOLS

- Set up tools and data

```
$ module load bwa/0.7.4
$ module load samtools/0.1.19
$ cp reference_genome.fasta bwa_reference_genome.fasta
$ mkdir bwa_mapping_assembly
$ cd bwa_mapping_assembly
$ newMapping mapping_assembly
```

- Set the reference sequence and read data

```
$ set best_refs_file = ../bwa_reference_genome.fasta
$ set final_sff_fastq = ../unknown_norovirus_454_convert.fastq
$ set final_illumina = ../unknown_norovirus_illumina.fastq
```

Mapping Assembly with BWA and SAMTOOLS

- Build BWA index on reference

```
$ bwa index -a is ${best_refs_file}
```

- Align long reads (454)

```
$  (bwa bwasw ${best_refs_file} ${final_sff_fastq} >
final_sff_mapping.sam) >& final_sff_bwa_bwasw.stderr
$ samtools view -bS -o final_sff_mapping.bam
final_sff_mapping.sam >& final_sff_samtools_view.stderr
$ samtools sort final_sff_mapping.bam final_sff_mapping.sorted
```

Mapping Assembly with BWA and SAMTOOLS

- Align short reads (Illumina)

```
$ (bwa aln ${best_refs_file} ${final_illumina} >
final_illumina_mapping.sai) >& final_illumina_bwa_aln.stderr
$ (bwa samse ${best_refs_file} final_illumina_mapping.sai
${final_illumina} > final_illumina_mapping.sam) >&
final_illumina_bwa_samse.stderr
$ samtools view -bS -o final_illumina_mapping.bam
final_illumina_mapping.sam >& final_illumina_samtools_view.stderr
$ samtools sort final_illumina_mapping.bam
final_illumina_mapping.sorted
```

Mapping Assembly with BWA and SAMTOOLS

- Merge the alignments and output consensus

```
$ samtools merge final_all.sorted.bam
final_sff_mapping.sorted.bam final_illumina_mapping.sorted.bam
$ echo ">sample_hybrid_refs_consensus" >
sample_hybrid_refs_consensus.fasta
$ (samtools mpileup -uf ${best_refs_file} final_all.sorted.bam |
bcftools view -cg - | gawk '{if($0 !~ /^#/){printf("%s",$4);}}' |
sed -e "s/.\{60\}/&\n/g" >> sample_hybrid_refs_consensus.fasta)
>& sample_hybrid_bcf.stderr
```

- Look at the consensus

```
$ more sample_hybrid_refs_consensus.fasta
```