Genome Assembly Continued

Genome Assembly Exercise

Download fastq files from server

Quality assessment using FastQC

Create hash table in Velvet

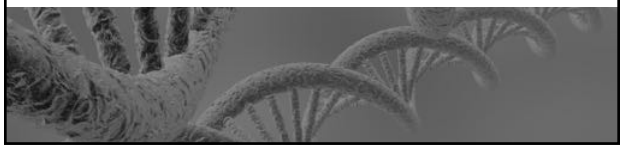
Assembly genome using Velvet

Visualize data in Tablet

Sequences and Annotations

Some common genomics data file formats

- FASTA: gene and genome sequence format (fa).
- FASTQ: high-throughput sequencing (HTS) data format (fastq).
- GFF: feature format typically general feature format (gff).
- SAM/BAM: HTS read alignment data (sam or bam).



FASTA Format

FASTA: DNA sequence alignment software. The software gave rise to the fasta format, now ubiquitous sequence file format.

```
>sequence1_description
AGCTAGCATCGACTAGCTACGATCGATCGATCACGAGCTACGACGTAGGCATGGGGGCTTACGATGCTA
CGGCGGAGCTACGGCGACTGCGATCTACGGCGATCGACGGACGGACGTCAGGCGACGATCTATCATCTA
TCGAGCGAGCTACTTACTCTTCTCTATCTACTTATCCCCTTCTTAGGGGTTGATTAGTCTAGCTGGTAC
GATCGAGCGATCTAGAGCGATCGACGAGCTGACGGACGTACTTACTATCGTAGCGACTACTTC

>sequence2_description
CTCTAGCATCGACTAGCTACGATCGATCGATCACGAGCTACGACGTAGGCATGGGGGCTTACGATGCTA
CCCCGGAGCTACGGCGACTGCGATCTACGGCGATCGACGGACGGACGTCAGGCGACGATCTATCATCTA
TCAAGCGAGCTACTTACTCTTCTCTATCTACTTATCCCCTTCTTAGGGGTTGATTAGTCTAGCTGGTAC
GATCTTTCTAGCGAGCGATCTAGAGCGATCGACGAGCTGACGGACGTACTTACTATCGTAGCGACTACT
TC
```

*DNA, RNA, or amino acid sequence

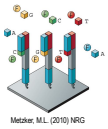
FASTQ Format

Index sequence

Read 1
```
1 @D64TDFP1:248:C50DMACXX:5:1101:1241:2095 1:N:0:ATCACG
2 CACCGCCCGTCGCTATCCGGGACTGGAATTCTCGGGTGCCAAGGAACTCCA
3 +
4 CCCFFFFFHHHHHJIJGHJJJIJJJJJGGGFFFFEADDHHHFHFF@@DD>
```

Read 2
```
1 @D64TDFP1:248:C50DMACXX:5:1101:1371:2154 1:N:0:ATCACG
2 TCAATATTTGCATAGGGTATCTGGAATTCTCGGGTGCCAAGGAACTCCAGT
3 +
4 CCCFFFFFHHHHHJJJJGFHIJJJJJJJJJJJJJFHHIIJJHGHJFGHJJI
```

Read 3
```
1 @D64TDFP1:248:C50DMACXX:5:1101:1461:2205 1:N:0:ATCACG
2 GAAAGACGTCTTCCTAGATTATGGAATTCTCGGGTGCCAAGGAACTCCAGT
3 +
4 CCCFFFFFHHHHHJJJJJJJJJJJIJJJJJJJJJHIJJJJJGIIJFGIJJJ
```



Metzker, M.L. (2010) NRG

Line 1: sequence ID, description, and index; begins with @

Line 2: sequence; contains only A, C, T, G, and N

Line 3: optional sequence ID; begins with +

Line 4: signal quality of each base, cryptic code, ASCII_base 33 or 64

Compressing and Decompressing Large Files

gzip...compress or decompress a file.
usage (compress file) $ gzip file.txt
usage (decompress a gz file) $ gzip -d file.gz

tar...combine files into a single archive – a tarball.
usage $ tar cf archive_name file1 file2 file3
usage (extract files from a tarball) $ tar xf archive_name.tar

gzip+tar... compress files and combine into a single archive
usage $ tar cvzf newname.tar file1 file2 fileN
usage (extract and decompress) $ tar xvzf file.tar.gz

See unix cheat sheet for additional options.

Datasets can be stored in a variety of locations: portable external hard drives, storage arrays, etc.