

Bioinformatics Important Terms and Concepts

Read this to learn the definitions of words we will use in this lab

Bioinformatics

- the science of collecting and analyzing complex biological data such as genetic codes.

Cluster

- A group of DNA molecules on an Illumina sequencing flow cell
- All DNA molecules in a cluster are identical
- There are billions of clusters per flow cell, each clonally amplified from a single DNA molecule using Bridge PCR

Sequence read

- A single read of the DNA sequence of a single cluster

Forward and reverse reads

- The forward read starts at one end of the DNA molecule
- The reverse read starts at the other end of the molecule, and is the sequence of the complementary strand (the reverse complement)

Read pair

- The forward and reverse sequence reads from the same cluster

Sequence Assembly

- aligning and merging sequence fragments from a longer DNA molecule in order to reconstruct the sequence of the original DNA molecule
- Sequence fragments are aligned by finding matching sequences in different fragments

Consensus sequence

- Within a cell
 - A **consensus sequence** is **determined** by aligning many nucleotide (or protein) **sequences** that share a common function, then **determining** the most commonly expressed nucleotide (or amino acid) at each position. Often conserved **sequences** reflect a common function or binding domain.
- In DNA sequence analysis like ours
 - After aligning multiple sequences, the consensus sequence is the most common base at each site

FASTA format

- A file format for DNA sequences

The first line starts with ">" and the name of the sequence.

The following lines contain the the DNA sequence.

The next line starting with ">" starts the next sequence

Example:

>first sequence name

AAATTGCTCC

>second sequence name

AGAAT

FASTQ format

- A file format for DNA sequences that includes quality information

Example:

The first line starting with @ is the sequence ID

The following lines are the sequence

The lines after the + are quality codes for each base

@SEQ_ID

GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT

+

!''*((((**+)) %%%++) (%%%) .1***-+*'') **55CCF>>>>>CCCCCCC65

Genbank

- An online database that stores DNA sequence data and information about where the organisms the sequences were obtained from

NCBI

- National Center for Biotechnology Information
- A part of National Institutes of Health (NIH)
- They maintain Genbank and many other biological databases

BLAST (Basic Local Alignment Search Tool):

- a program that compares your DNA sequences to those in GenBank - it searches for similar DNA sequences.
 - a. BLAST searches GenBank for sequences that are similar to your query

Query

- your sequence. The sequence you entered into a BLAST search

Subject

- a sequence from Genbank that is similar to your query

The following software were used to analyze your DNA sequence data:

SeekDeep

- Demultiplex (separate sequences by their indices)
- Trim (remove) adaptors, indices, low quality sequence data
- Group similar sequences together
- Determine a consensus sequence for sequences that are grouped

RDP Classifier

- Use statistical analysis to determine what species your consensus sequences are most likely to belong to, and whether we can reliably classify the sequence to species, or only to genus or family.

KRONA

- Make graphs to represent the results of RDP Classifier