ntro to umina Sequence Ana ysis

Illumina or Short-Read Sequencing

- Allows the rapid and inexpensive sequencing of billions of base pairs of DNA or RNA in a single reaction.
- Revolutionized many aspects of biology over the last decade.
- Analyzing Illumina data is a critical skill for any bioinformaticist.
- We will spend the next six labs working with an Illumina data set.

Three Videos for more info on Illumina sequencing

- https://www.youtube.com/watch?v=mI0Fo9kaWqo
- https://www.youtube.com/watch?v=WneZp3fSJIk&t=13s
- https://www.youtube.com/watch?v=fCd6B5HRaZ8

The Data Set

- Illumina technology can be used to sequence RNA or DNA.
- In this experiment we purified mRNA from:
  - 2 varieties of *Brassica rapa*
  - Multiple growth conditions:
    - Growth chamber: simulated sun and shade
    - Greenhouse: crowded and uncrowded plantings
    - Field
  - Multiple tissues: (see lab manual).
- *What can we learn from sequencing RNA?*

What can we learn from sequencing RNA?

- Transcript abundance (gene expression levels)
- Intron/exon junctions (gene structure)
- Transcript start and stop sites (gene structure)
- Genetic variants (SNP and in/del discovery and genotyping)

Goals

1. Learn about Illumina reads, how to map them, and quality control (Tuesday)
2. How to view reads in a genome browser and how to find single nucleotide polymorphisms (Thursday)
3. Find genes that are differentially expressed between genotypes or treatments (Next week)
4. Ask if differentially expressed genes have any common functionality (gene ontologies) or promoter motifs
5. Build a gene regulatory network to determine how genes connect to one another.

Illumina Data

FASTQ

```
@HWUSI-EAS100R:6:73:941:1973#0/1
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCC
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**
```

1.

wiki

FASTQ

```
@HWUSI-EAS100R:6:73:941:1973#0/1
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCC
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**
```

1. @SEQID
2. Sequence
3. Starts with “+” and then usually blank
4. Quality information

PHRED QUALITY

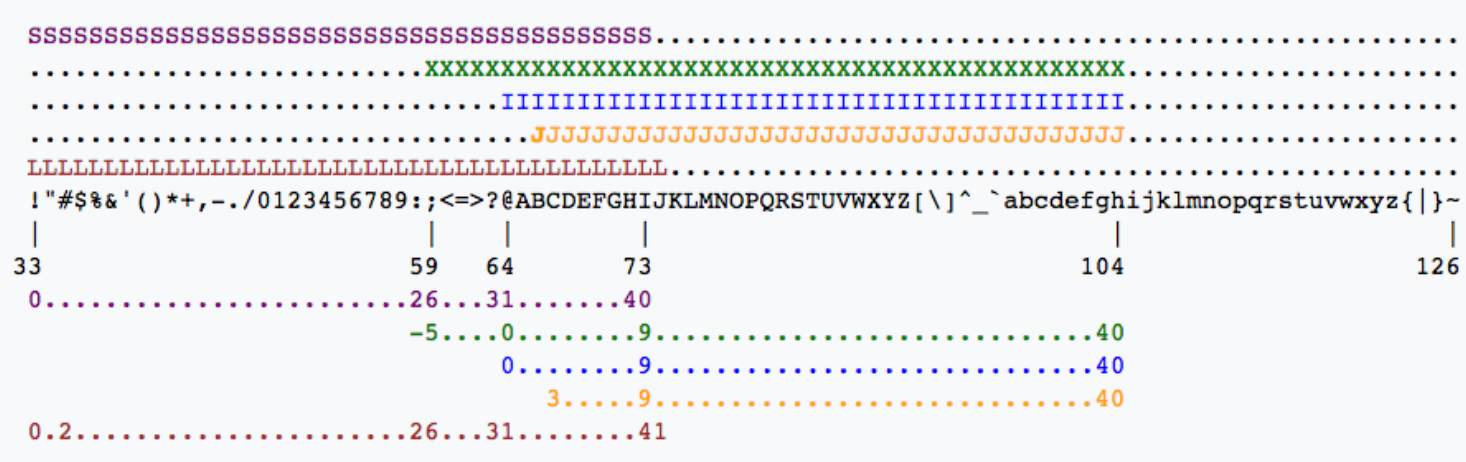
-
-
- **QUESTION: If there is a 1 in 100 chance that the base is wrong, what is the PHRED score?** (Try this in R)

PHRED Qualities, part 2

```
!''*(((***+))%%%++)(%%%%).1***-+*''))**
```

Why use characters instead of numbers?

PHRED Qualities, part 3



S - Sanger Phred+33, raw reads typically (0, 40)
X - Solexa Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64, raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64, raw reads typically (3, 40)
with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
(Note: See discussion above).
L - Illumina 1.8+ Phred+33, raw reads typically (0, 41)

Barcodes and sample indexing

- For RNAseq one typically needs 10 - 20 Million reads per sample.
- However the sequencer gives 350 Million reads per flow cell.
- “Barcodes” or “Indexes” are used to uniquely associate reads with samples.

Summary: Barcodes and sample indexing

- Allow multiple samples to be sequenced in a single lane.
- Tag each DNA fragment with a sequence that is unique for each sample
- “Indexes”
  - Tag or index is internal in the adapter and is sequenced in a separate reacion
  - Reads are automatically separated for the different samples
- “Barcodes”
  - Tag or barcode is at the end of the adapter
  - The barcode is sequenced in the same reaction used to sequence the insert DNA
  - The reads must be sorted and barcodes must be trimmed by the end user.

What to do with your sequences

- If the sequences come from an organism with an already sequenced genome, then you will want to **map them to the reference sequence** so that you know where they came from.
  - Look for polymorphisms and structural changes
  - If RNA, examine expression levels differences
- There are **many mapping programs**. Some popular ones:
  - **BWA**. Non-splicing. Use for mapping genomic reads to a genomic reference or mRNA reads to a cDNA reference
  - **Tophat / Bowtie**. Splicing. Use for mapping mRNA reads to a genomic reference.
  - **STAR**. Splicing. Use for mapping mRNA reads to a genomic reference.
  - **kallisto**. Non-splicing. Use for mapping mRNA reads to a cDNA reference.

What to do with your sequences

- If the sequences come from an organism without a reference, then you will need to perform a ***de novo* assembly.** (not covered in this class)

Workflow for tomorrow's lab

1. Check sequence quality with `fastqc`
2. Filter reads based on quality with `Trimmomatic`
3. Split into samples based on barcodes with `auto_barcode`
4. Map reads to find where the came from in the genome

File types

- .fastq – file of short read data
- .fa – fasta files for reference genome
- .sam – **sequence alignment/map file for mapped reads**
- .bam – the binary version of a sam file
- .bai – index for bam files
- .gff – genome annotation: information about where the genes are in the genome