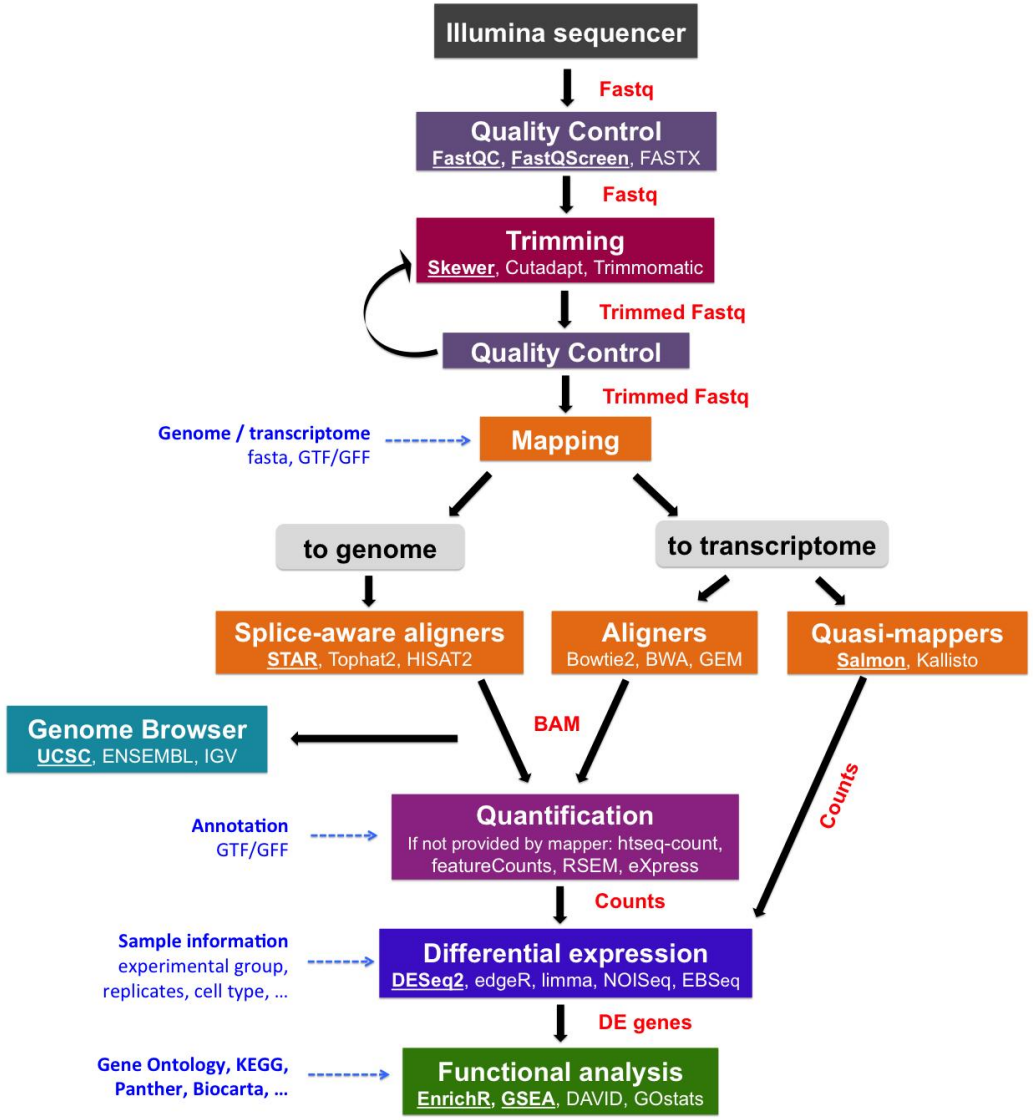
BIOL5488: Genomics Lab II

Table of Contents

- Review
- Assignment 2
- Q & A

Adapted from
Bio5488 Genomics 2023 TAs

Review


Illumina sequencer
Fastq
Quality Control
FastQC, FastQScreen, FASTX
Fastq
Trimming
Skewer, Cutadapt, Trimmomatic
Trimmed Fastq
Quality Control
Trimmed Fastq
Genome / transcriptome
fasta, GTF/GFF
Mapping
to genome
to transcriptome
Splice-aware aligners
STAR, Tophat2, HISAT2
Aligners
Bowtie2, BWA, GEM
Quasi-mappers
Salmon, Kallisto
Genome Browser
UCSC, ENSEMBL, IGV
BAM
Counts
Annotation
GTF/GFF
Quantification
If not provided by mapper: htseq-count, featureCounts, RSEM, eXpress
Counts
Sample information
experimental group, replicates, cell type, ...
Differential expression
DESeq2, edgeR, limma, NOISeq, EBSeq
DE genes
Gene Ontology, KEGG, Panther, Biocarta, ...
Functional analysis
EnrichR, GSEA, DAVID, GOstats

We are here

FASTA format

1. Represent nucleotide sequences or amino acid sequences
2. The filename ends with .fa/.fna/.fasta
3. The first line start with “>” as the title & description
   (sometimes “;” will use to start a comment line)
4. The following lines are data

```
>NG_008679.1:5001-38170 Homo sapiens paired box 6 (PAX6)
ACCCTCTTTTCTTATCATTGACATTTAAACTCTGGGGCAGGTCCTCGCGTAGAACGCGGCTGTCAGATCT
GCCACTTCCCCTGCCGAGCGGCGGTGAGAAGTGTGGGAACCGGCGCTGCCAGGCTCACCTGCCTCCCCGC
CCTCCGCTCCCAGGTAACCGCCCGGGCTCCGGCCCCGGCCCGGCTCGGGGCCCGCGGGGCCTCTCCGCTG
CCAGCGACTGCTGTCCCCAAATCAAAGCCCGCCCCAAGTGGCCCCGGGGCTTGATTTTTGCTTTTAAAAG
GAGGCATACAAAGATGGAAGCGAGTTACTGAGGGAGGGATAGGAAGGGGGGTGGAGGAGGGACTTGTCTT
TGCCGAGTGTGCTCTTCTGCAAAAGTAGCAAAATGTTCCACTCCTAAGAGTGGACTTCCAGTCCGGCCCT
GAGCTGGGAGTAGGGGGCGGGAGTCTGCTGCTGCTGTCTGCTAAAGCCACTCGCGACCGCGAAAAATGCA
GGAGGTGGGGACGCACTTTGCATCCAGACCTCCTCTGCATCGCAGTTCACGACATCCACGCTTGGGAAAG
TCCGTACCCGCGCCTGGAGCGCTTAAAGACACCCTGCCGCGGGTCGGGCGAGGTGCAGCAGAAGTTTCCC
GCGGTTGCAAAGTGCAGATGGCTGGACCGCAACAAAGTCTAGAGATGGGGTTCGTTTCTCAGAAAGACGC
```

FASTQ format

1. Storing both nucleotide sequence and corresponding quality scores
   -> usually handle reads from the sequencing machine
2. The filename ends with .fq/.fastq
3. Each sequence has four lines:

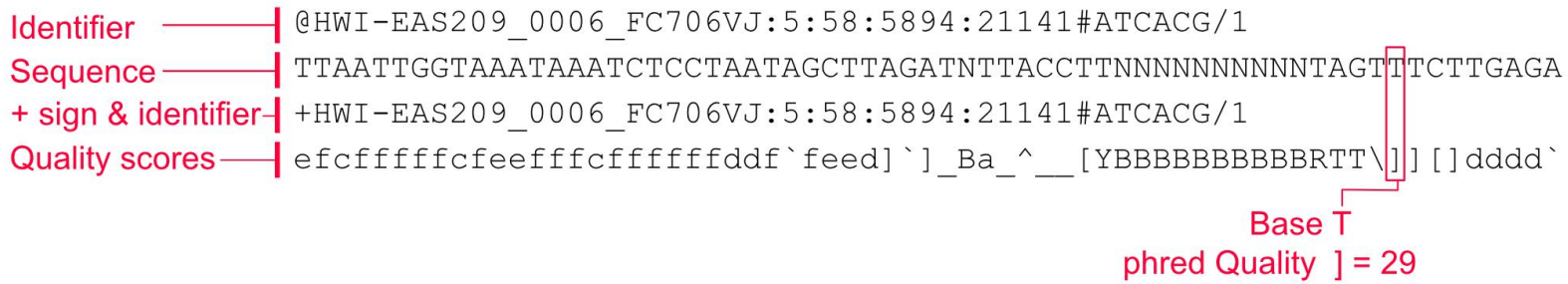
Line 1: begins with '@', the following context is about sequence identifier (sequence tech, flow cell IDs, info about read pairs etc)

Line 2: the sequence

Line 3: begins with '+', the following context could be the same as Line 1

Line 4: encodes the quality of the sequence in line 2 by ASCII. The letters in Line 4 should be the same as the letters in Line2.

FASTQ format: example



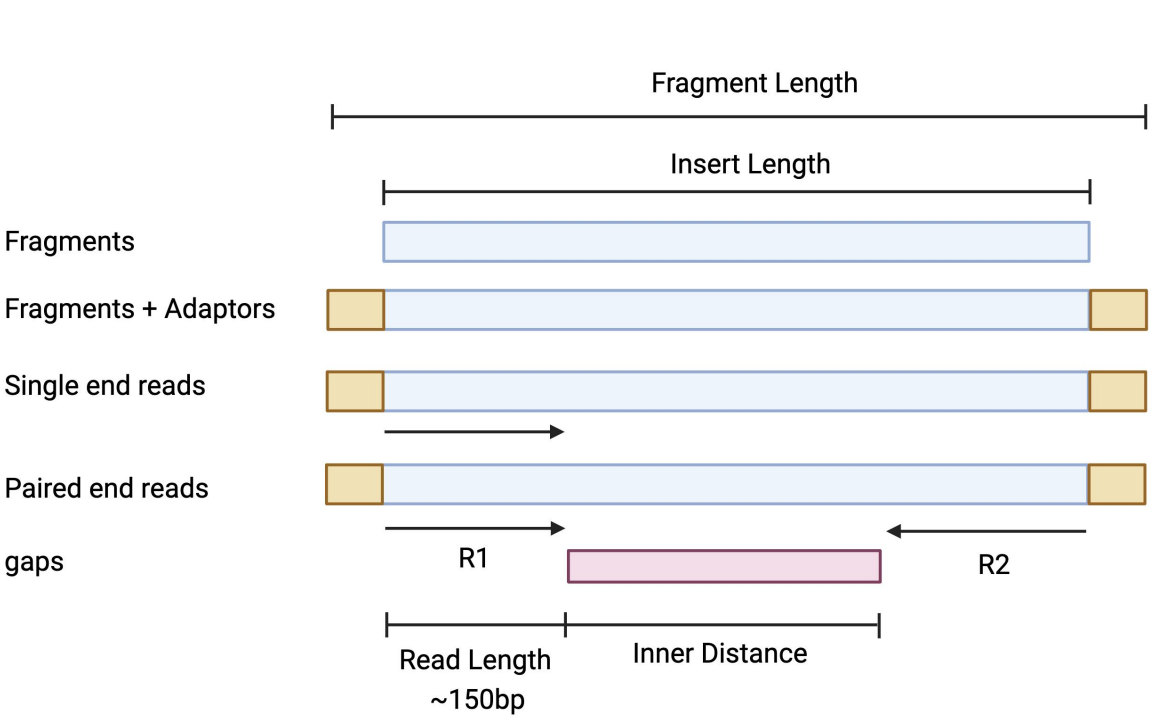
Each letter in line 4 represents a *Qphred value*

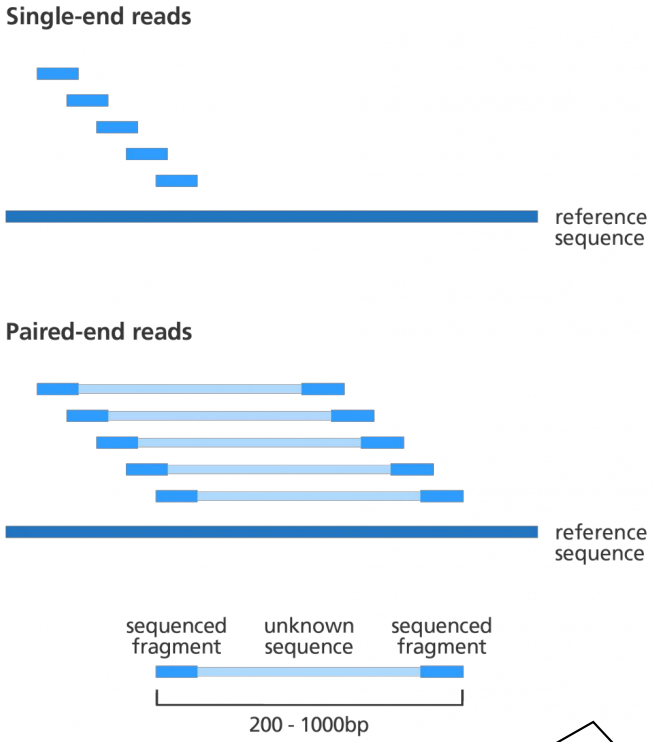
$$Q_{phred} = -10log_{10}p$$

Where *p* is the estimated probability of a base being wrong

Higher *Qphred,* better quality

Single-end vs. Paired-end





Single-end vs. Paired-end

Single-end: 1. Economic friendly
2. when the fragment is relatively short
e.g., small RNA sequencing, CHIP-seq

Paired-end: 1. Better mapping ability (especially in repetitive sequence)
2. Detect splice isoforms

Choose your aligners

1. What is the data source?
   -- Do the reads contain splice junctions (i.e., RNA-seq)
   -- Map the spliced reads against genome using STAR instead of Bowtie2
   -- Map the non-spliced DNA-seq reads using Bowtie2

2. What is the output?
   -- Typically SAM file. Make sure the out put file matches downstream analysis.

3. Trade-off between speed and completeness