

BIOINFORMATICS & HIGH- PERFORMANCE COMPUTING

WHAT WE WILL DO TODAY ...

- This is not a “Tool X” vs “Tool Y” workshop
- We will not go into the details of alignment algorithms

Who are you???



Bioinformatics Scientist

Identified that you need to align reads

Identified that you need to use Rivanna

What you will learn???

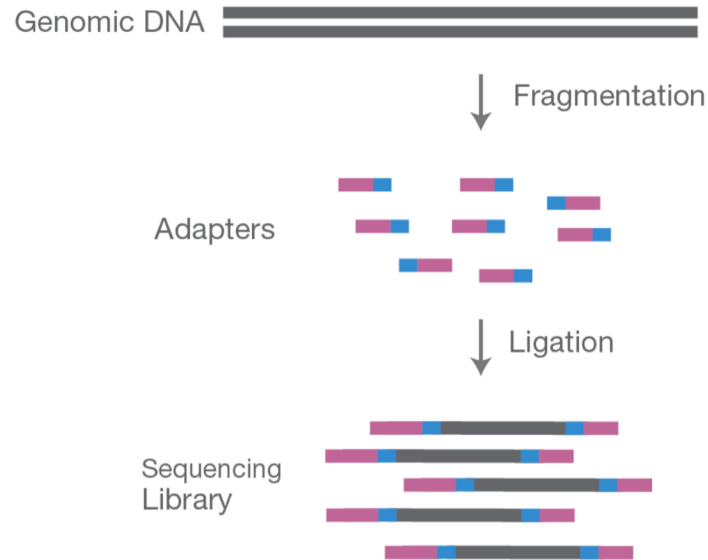
How to use bowtie2 on Rivanna

Understand SAM/BAM format

Perform downstream manipulations

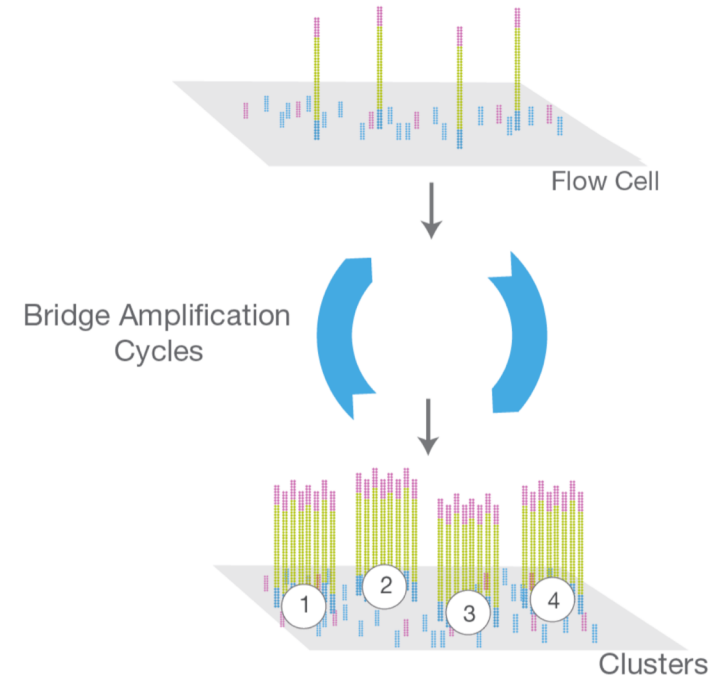
Illumina Sequencing

A. Library Preparation



NGS library is prepared by fragmenting a gDNA sample and ligating specialized adapters to both fragment ends.

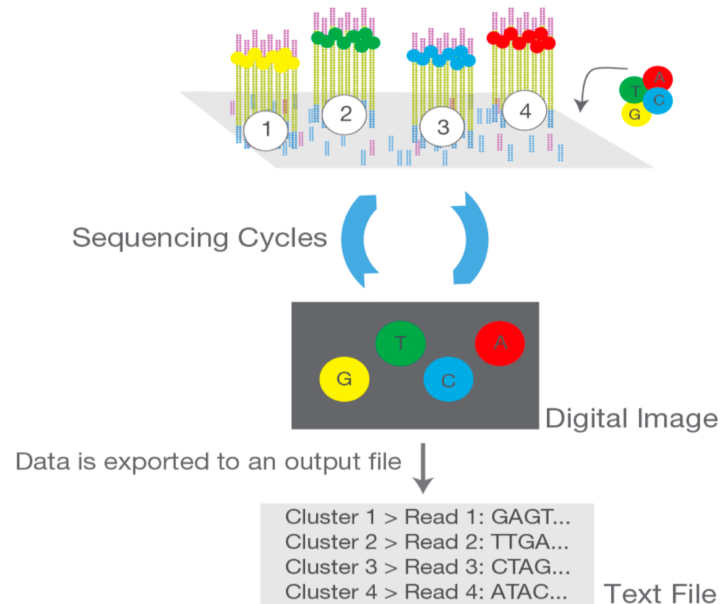
B. Cluster Amplification



Library is loaded into a flow cell and the fragments are hybridized to the flow cell surface. Each bound fragment is amplified into a clonal cluster through bridge amplification.

Illumina Sequencing

C. Sequencing



Sequencing reagents, including fluorescently labeled nucleotides, are added and the first base is incorporated. The flow cell is imaged and the emission from each cluster is recorded. The emission wavelength and intensity are used to identify the base. This cycle is repeated “n” times to create a read length of “n” bases.

D. Alignment and Data Analysis

Reads

```
ATGGCATTGCAATTTGACAT
TGGCATTGCAATTTG
AGATGGTATTG
GATGGCATTGCAA
GCATTGCAATTTGAC
ATGGCATTGCAATT
AGATGGCATTGCAATTTG
```

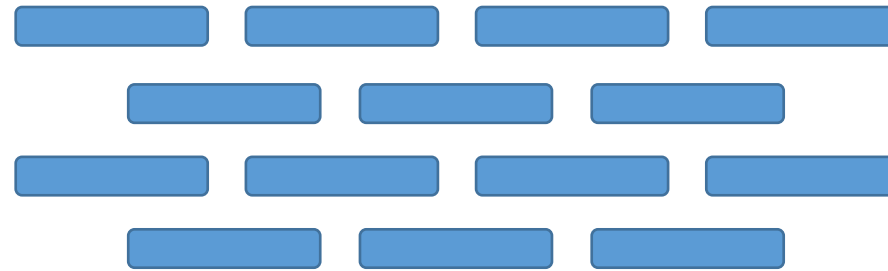
Reference Genome

```
AGATGGTATTGCAATTTGACAT
```

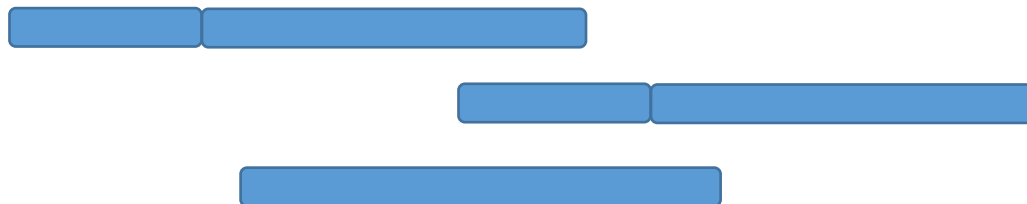
Reads are aligned to a reference sequence with bioinformatics software. After alignment, differences between the reference genome and the newly sequenced reads can be identified.

https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

Short Read Processing



Absence of Reference Genome
de novo Assembly



Align to Reference Genome



When to Align ?

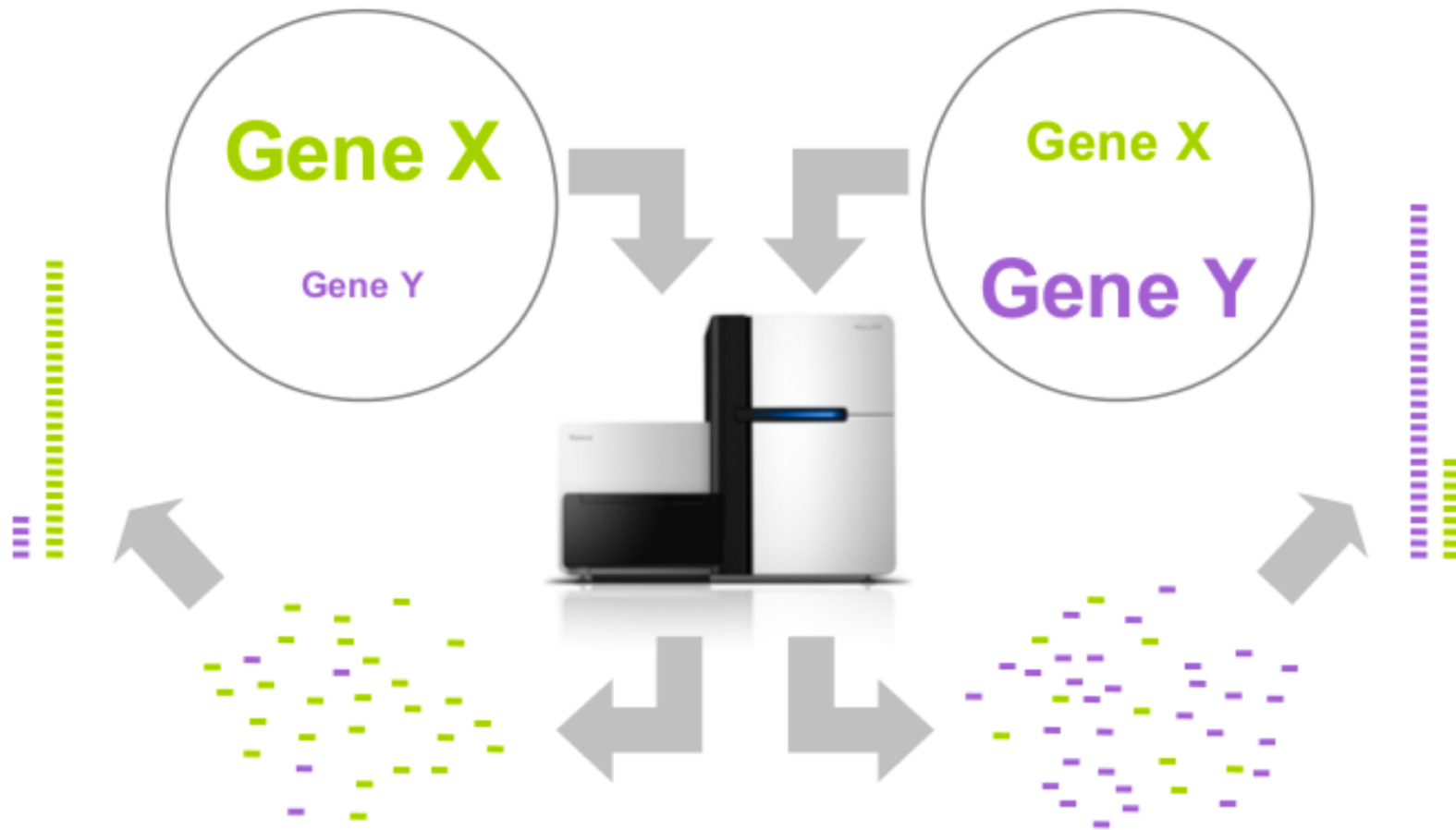
- Identify Variation in Individuals



ATGATAGCATCGTCGGGTGTCTGCTCAATAATAGTGCCGTATCATGCTGGTGTTATAATCGCCGCATGACATGATCAATGG
CAATAAAAGTGCCGTATCATGCTGGTGTTACAATCGCCGCA
CGTATCATGCTGGTGTTACAATCGCCGCATGACATGATCAATGG
TGTCTGCTCAATAAAAGTGCCGTATCATGCTGGTGTTACAATC
ATCGTCGGGTGTCTGCTCAATAAAAGTGCCGTATCATG--GGTGTTATAA
CTCAATAAGAGTGCCGTATCATG--GGTGTTATAATCGCCGCA
GTTATAATCGCCGCATGACATGATCAATGG

When to Align ?

- Quantify Abundance



Popular Aligners

- BWA / Bowtie2

They both use the Burrows – Wheeler Transform to index the reference genome

This massively reduces the memory footprint of long reference genomes, while allowing for rapid identification of potential origin of query sequence

- Choice of aligner

Documentation → can I figure out how it works?

Input features → what input can it handle?

Output → will the output be useful for downstream analysis?

Performance → do I have the computational resources to run?

- **DO NOT QUOTE ME:** In some tests, Bowtie2 was slightly faster at marginal expense of sensitivity.