Intro to NGS Bioinformatics using Tufts HPC

Rebecca Batorsky

Sr Bioinformatics Specialist

May 2020

Requirements

- HPC Cluster Account available to Tufts affiliates
- VPN if working off campus
- Basic knowledge of Linux and HPC:
  - Intro to Linux
  - HPC Quick Start guide or Intro to HPC

We'll test out access together during this session.
Depending on the number/type of questions, we may choose to follow up after the session.

Course Format

1-hour Zoom Introduction

~3 hours of self-guided material on github, suggested to be completed over the **next week:** https://rbatorsky.github.io/intro-to-ngs-bioinformatics/

(working with a partner is encouraged)

Piazza

- Please ask and answer questions liberally on Piazza
- Steps to enroll in class if you are not auto-enrolled:
  - https://piazza.com/tufts
  - 1: Intro to NGS Bioinformatics
  - Join as student
- If you can't access Piazza for some reason please let me know Rebecca.Batorsky@tufts.edu

Bioinformatics goals

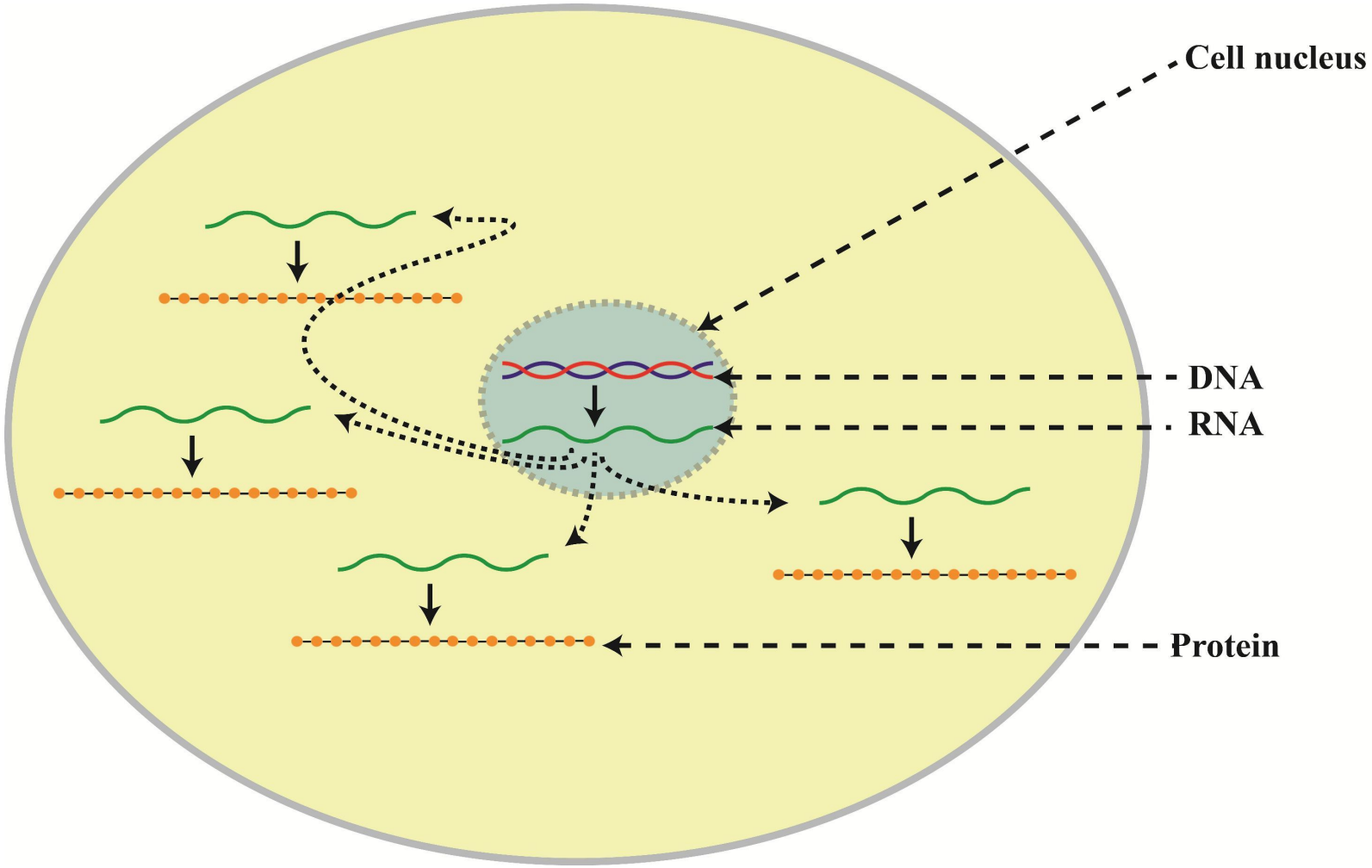
Variant Calling and Interpretation for a **human exome** sample

Writing and running bash scripts

Intro to several common bioinformatics tools: BWA, Samtools, Picard, GATK, IGV

Using modules on the HPC

DNA and RNA in a cell

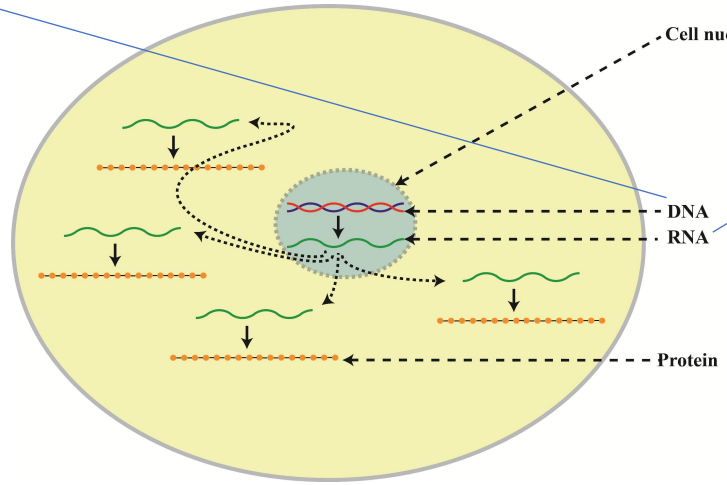


https://i0.wp.com/science-explained.com/wp-content/uploads/2013/08/Cell.jpg

Two common analysis goals

DNA Sequencing

- Fixed copy of a gene per cell
- Analysis goal:
  Variant calling and interpretation



RNA Sequencing

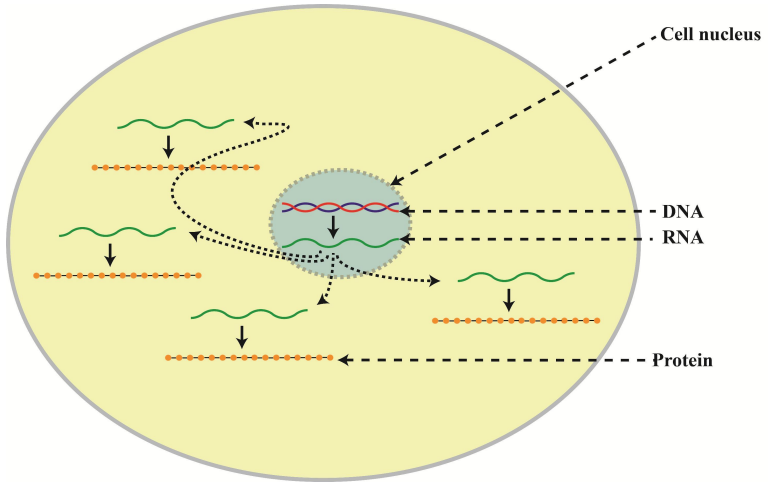
- Copy of a transcript per cell depends on gene expression
- Analysis goal: Differential expression and interpretation

https://i0.wp.com/science-explained.com/wp-content/uploads/2013/08/Cell.jpg

This workshop will cover DNA sequencing

DNA Sequencing

- Fixed copy of a gene per cell
- Analysis goal:
  Variant calling and interpretation



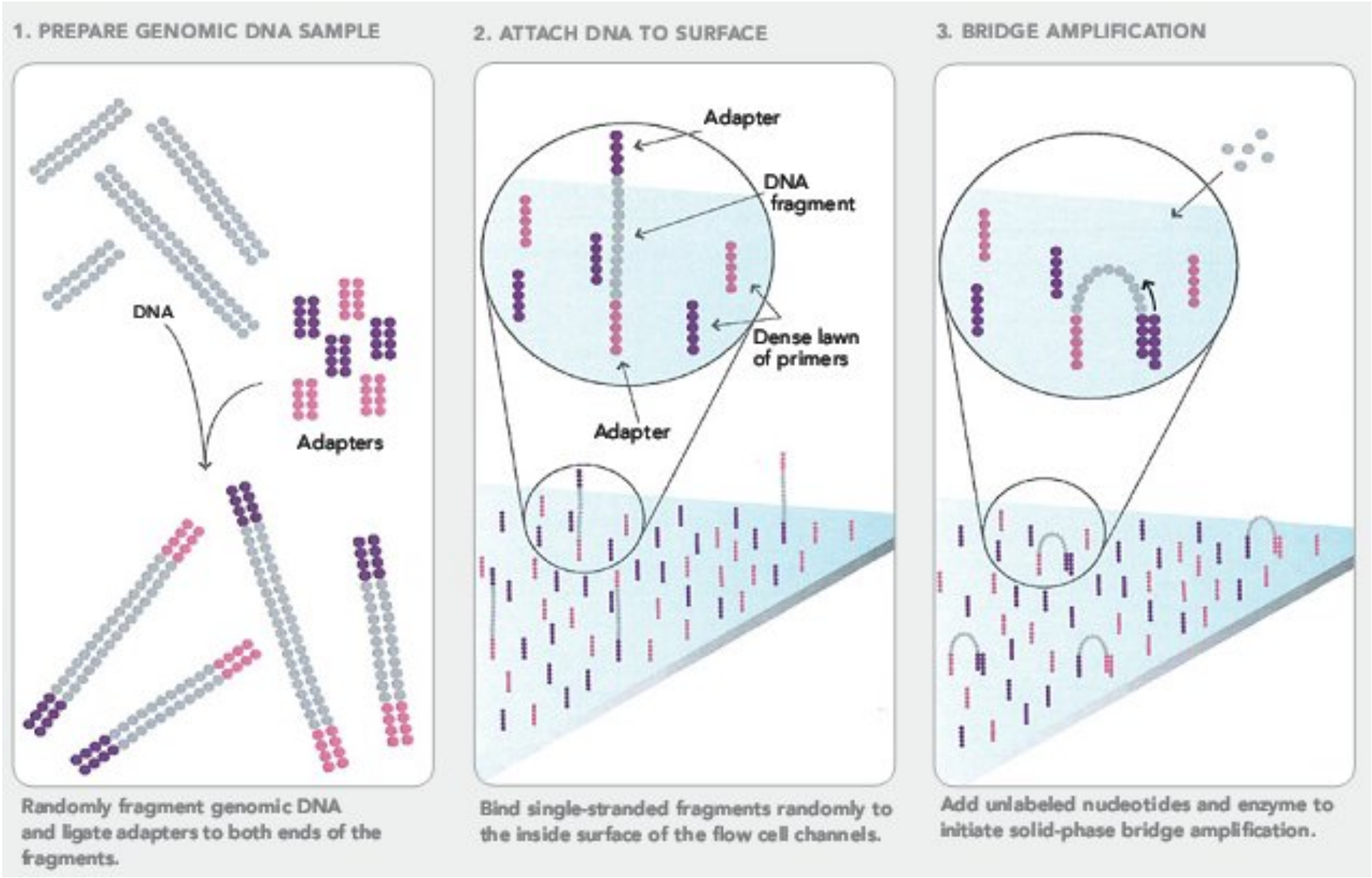
Not today!
Check out our 6/2/20 workshop:
https://tufts.libcal.com/event/6716203

RNA Sequen

- Copy of a g cell depends on gene ex
- Analy l: Diffe l expr n and inter tion

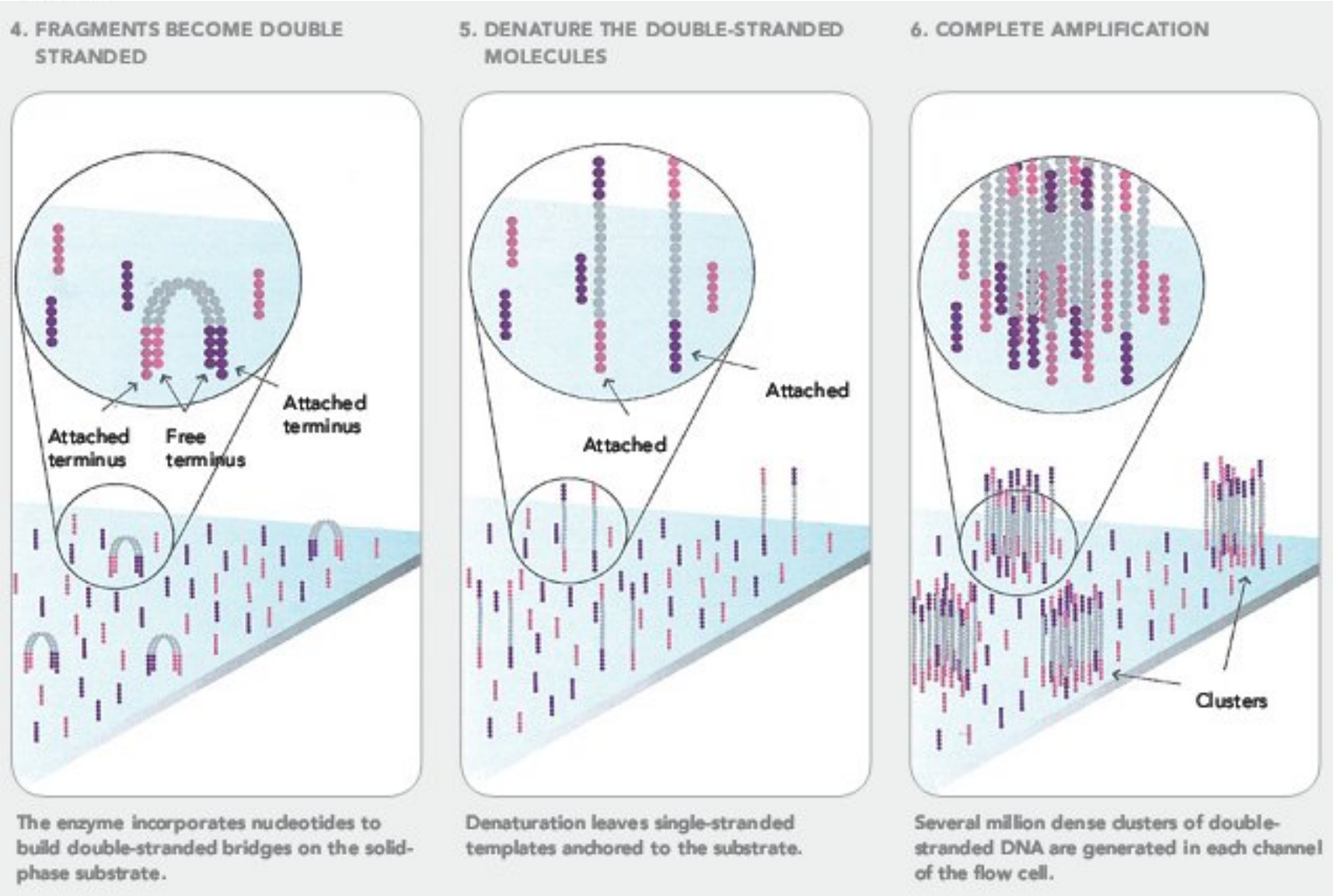
https://i0.wp.com/science-explained.com/wp-content/uploads/2013/08/Cell.jpg

Next Generation Sequencing (NGS)



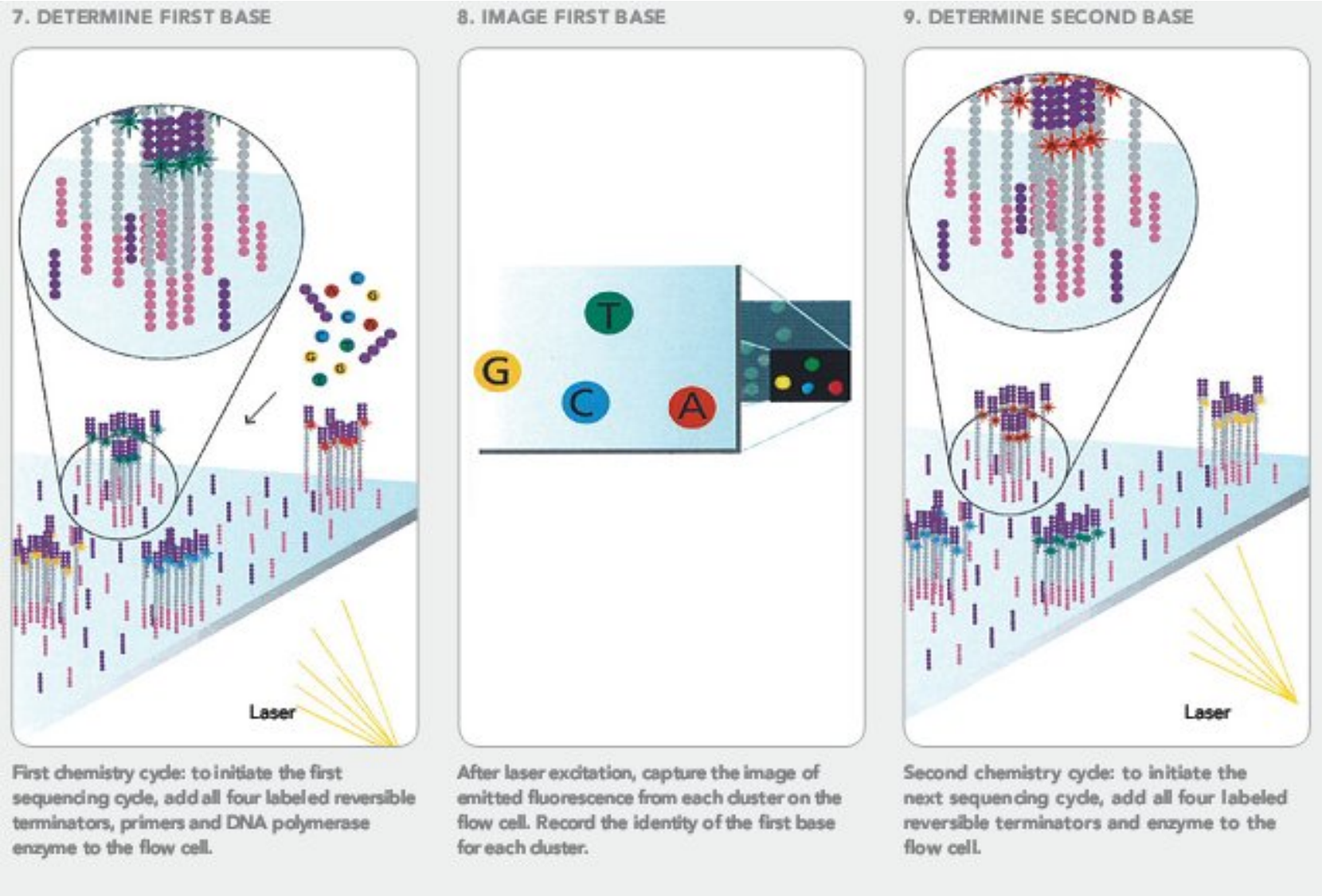
https://sites.google.com/site/himbcorelab/illumina_sequencing

Next Generation Sequencing (NGS)



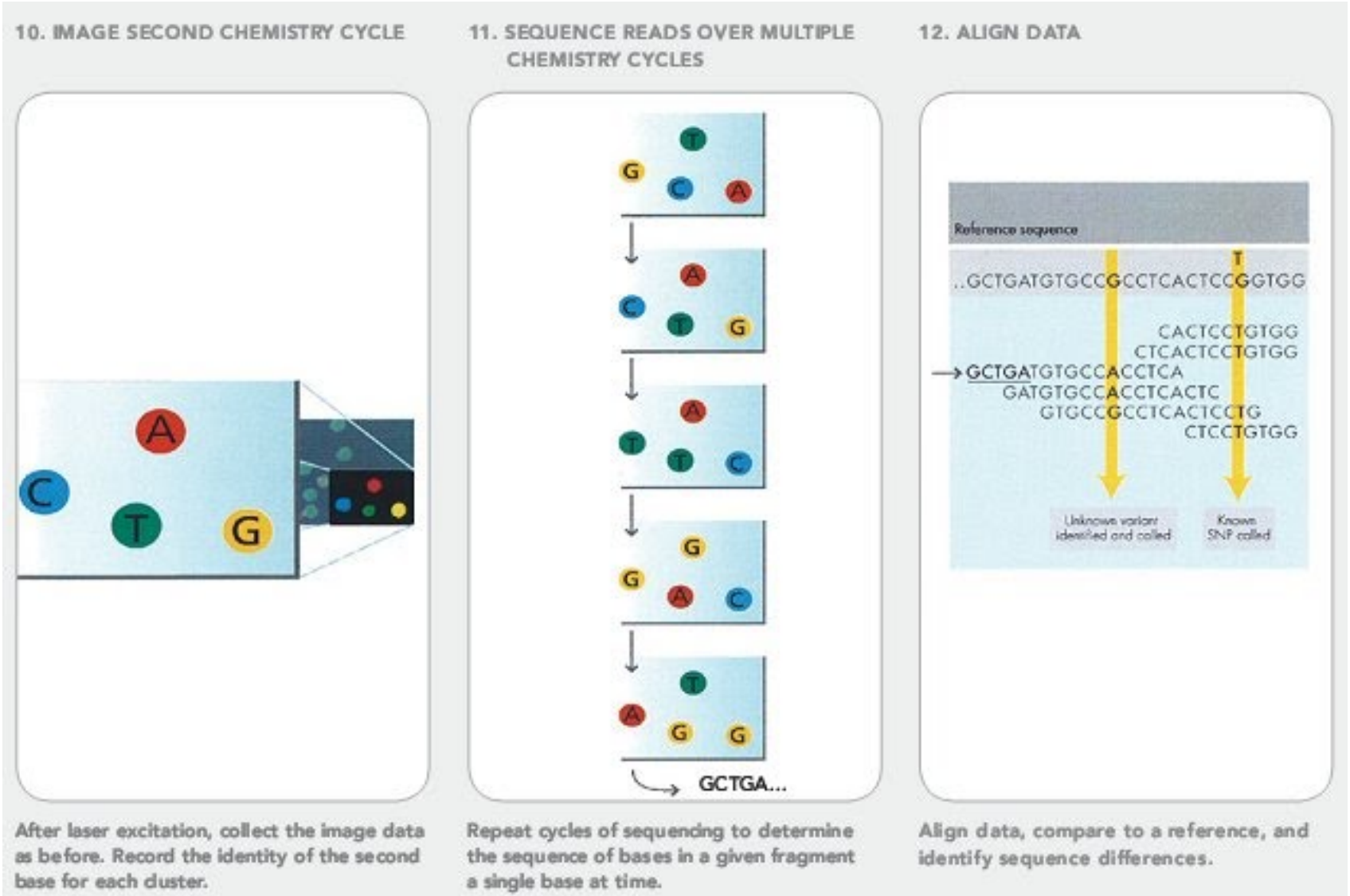
https://sites.google.com/site/himbcorelab/illumina_sequencing

Next Generation Sequencing (NGS)



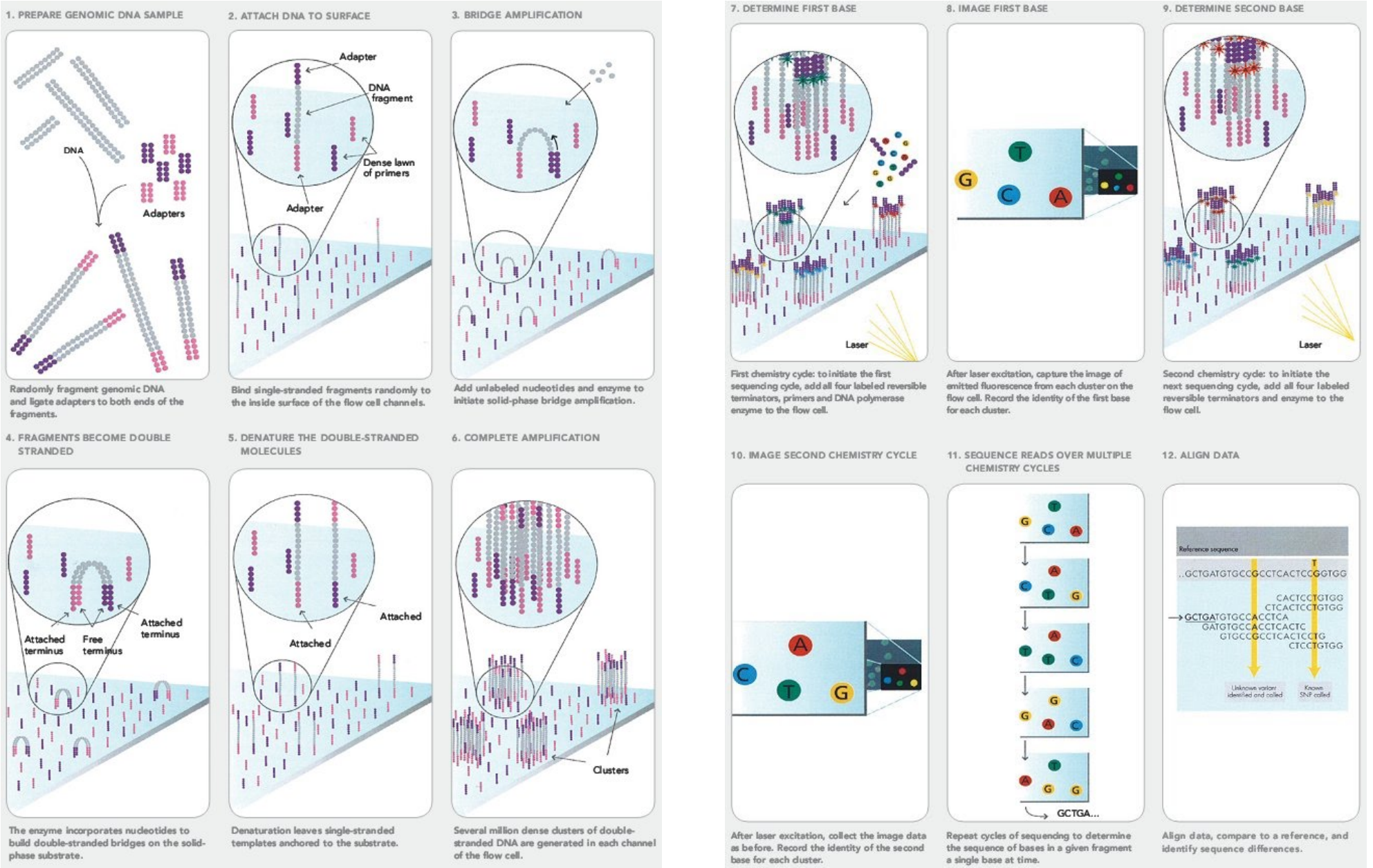
https://sites.google.com/site/himbcorelab/illumina_sequencing

Next Generation Sequencing (NGS)



https://sites.google.com/site/himbcorelab/illumina_sequencing

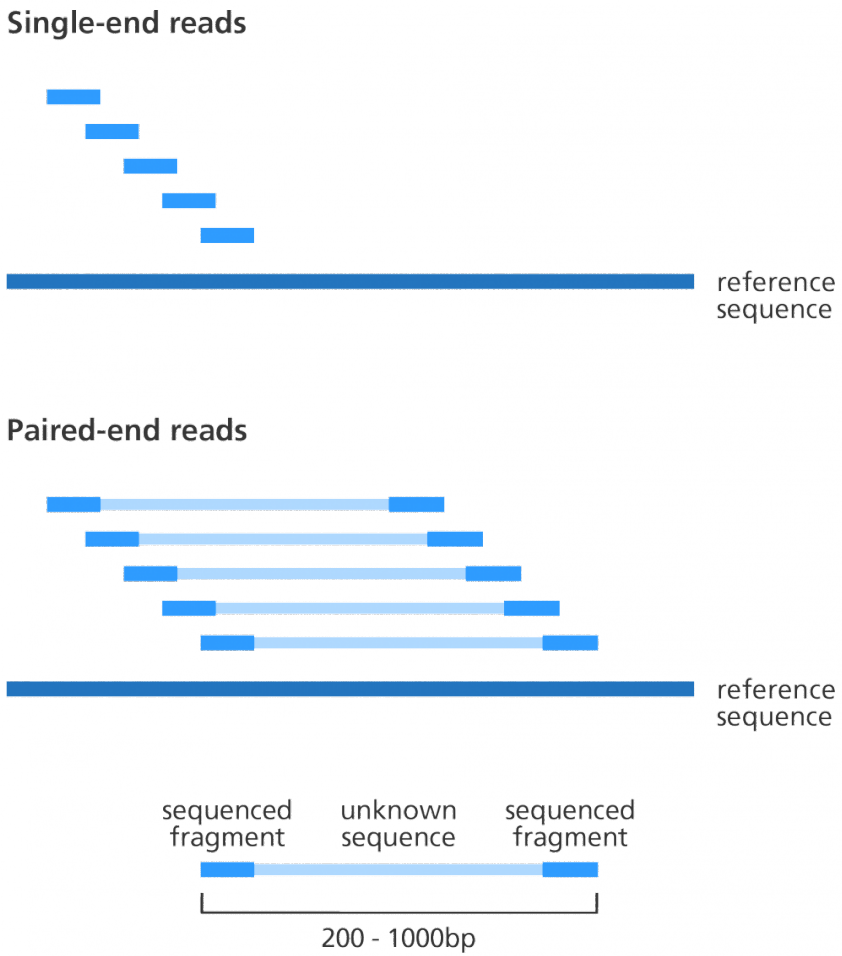
Next Generation Sequencing (NGS)



This Illumina Video is helpful for visualization!

Paired end vs Single end reads

- In single-end reads, only one end of the fragment is sequenced.
- In paired-end reads, both ends of the fragment are sequenced.

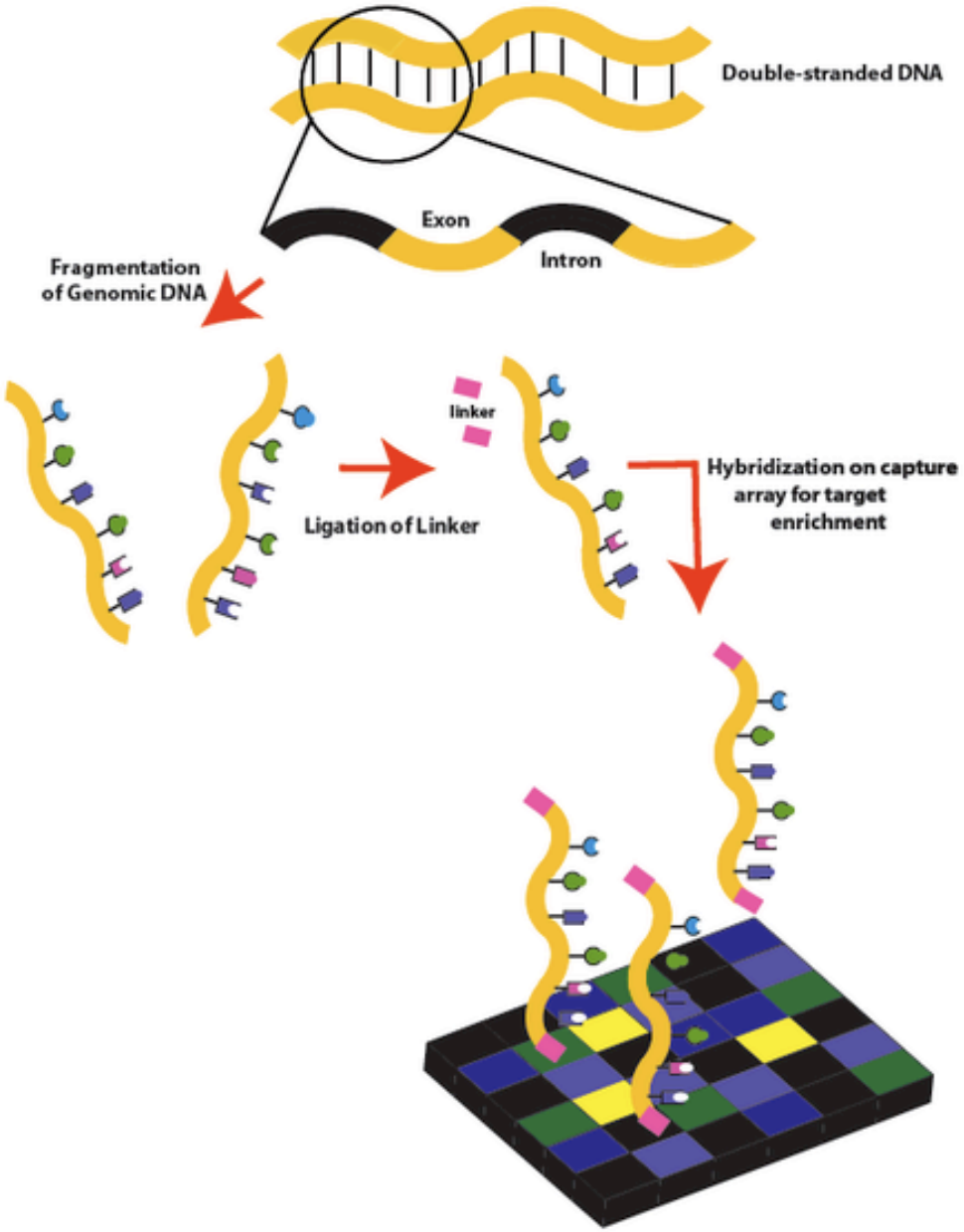


“Insert Size”

https://www.biostars.org/p/267167/

Exome Sequencing

- **Whole Exome Sequencing (WES)** aims to sequence all protein-coding regions of genes in a genome, called **exons**
- **Exons** comprise ~1% of the human genome and cause 80% of characterized inherited disordered
- **Array-based capture** is an extra step in library preparation that enriches for exons.
- Sequences that are complementary to the exons are used as probes to capture exonic DNA fragments, uncacptured fragments are washed away.



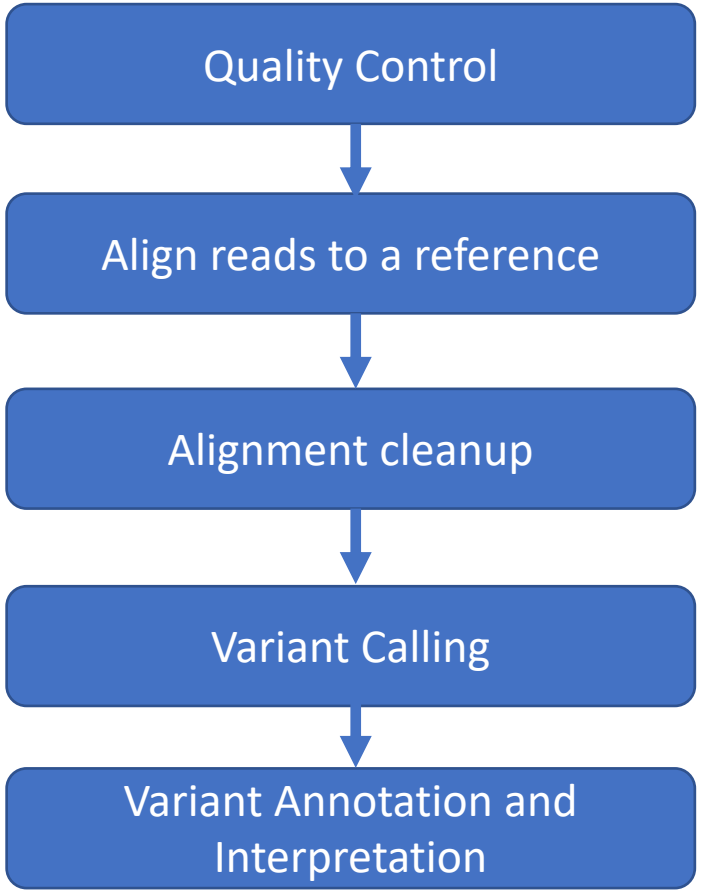
https://en.wikipedia.org/wiki/Exome_sequencing

The result: lots of short reads

How do we make sense of these?

Today: we'll **align** to a **reference sequence** and look for **variants**

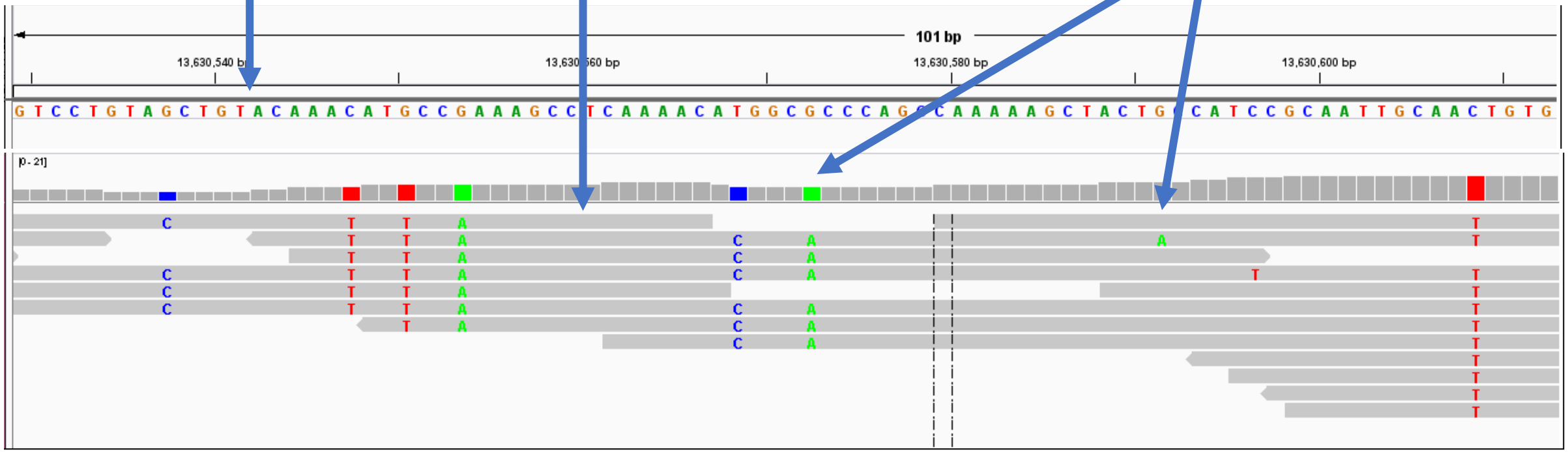
Variant Calling workflow



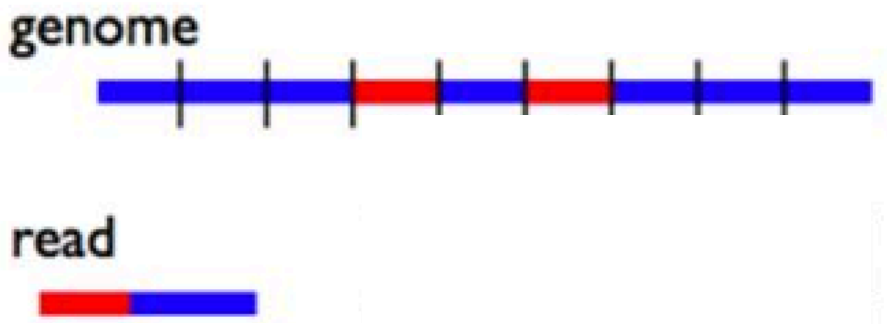
https://github.com/hbctraining/In-depth-NGS-Data-Analysis-Course

Overview

- A **reference sequence** is a previously determined sequence from your organism
- **Reads** are aligned to the reference based on sequence similarity
- **Variants** are positions where your sequences differ from the reference

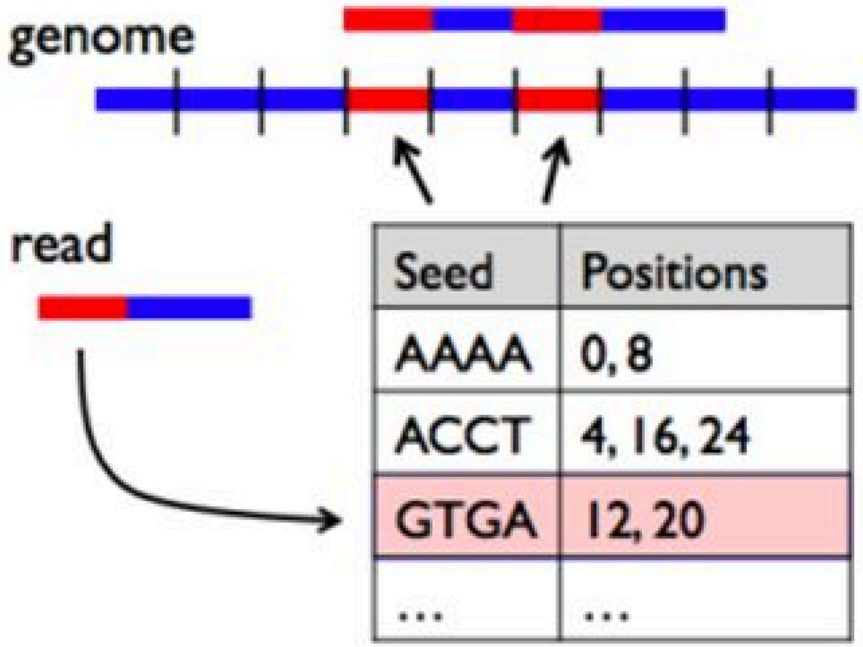


Alignment



- The goal of read alignment is to find the correct location in a reference genome from which the short read originated
- Insertions, deletions, and mismatches are allowed
- There may be >1 equally good choices
- Comparing millions of reads to billions of reference positions (human genome) is very time consuming
  - For a single read of length ***m*** and a genome of length ***n*** : O(mxn) comparisons

Alignment



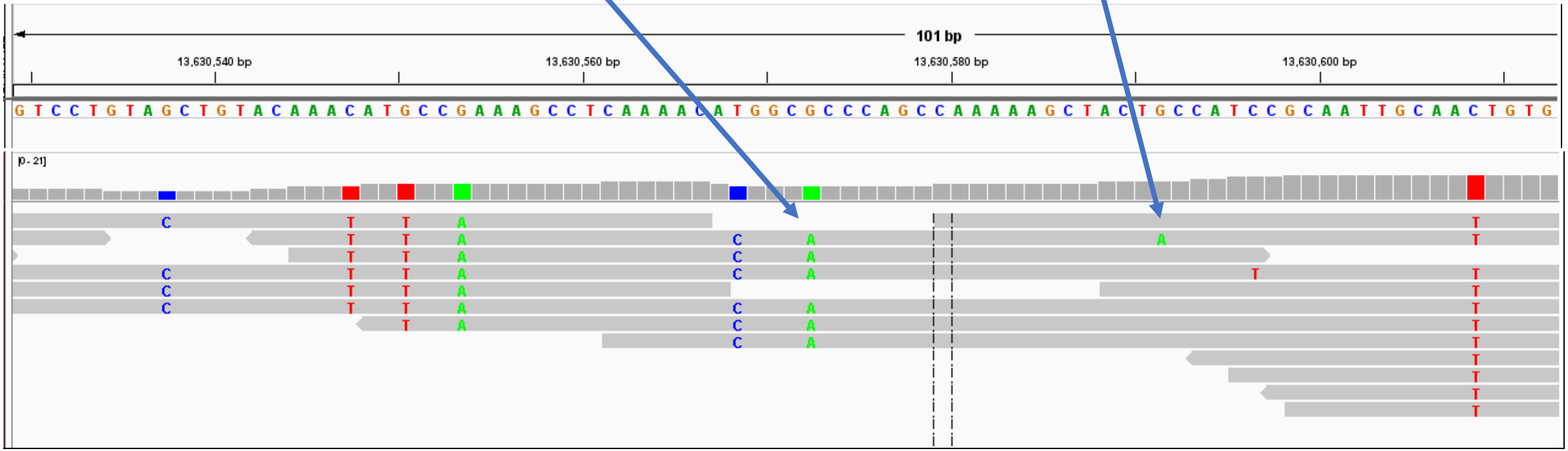
- Creating an **index** of our **reference sequence** speeds things up
- An index is a lookup table, where for each short sequence in the reference genome (**seed**), a list of all positions in the reference genome where that sequence is found.
- The index is created only once for a given genome
- For read alignment: look up the positions for the first 4 bases (seed) of my read in my index table
    - For a single read of length ***m*** and a genome of length ***n*** : O(mxlog2(n))

Variant Calling

- Our variant caller provides a list of positions where the sequenced base is different from the reference base
- Quality metrics are also provided to help us judge whether the variant is a technical artifact

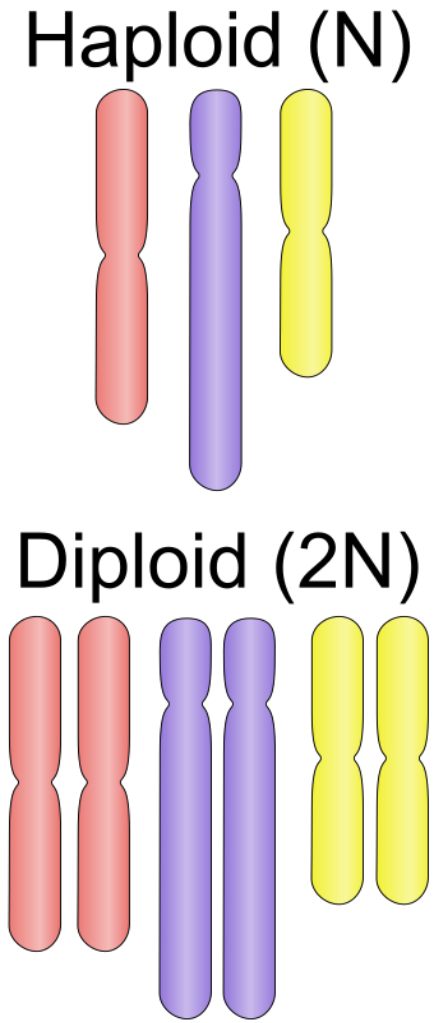
Reference position 13,635,567
G -> A
6/6 reads -> High confidence

Reference position 13,630,586
G -> A
1/8 reads -> Low confidence



Ploidy and Variant Calling

- Ploidy is the number of copies of each chromosomes
  - Humans cells are diploid for autosomal chromosome and haploid for sex chromosomes
  - Bacteria are haploid
  - Viruses and Yeast can by haploid or diploid



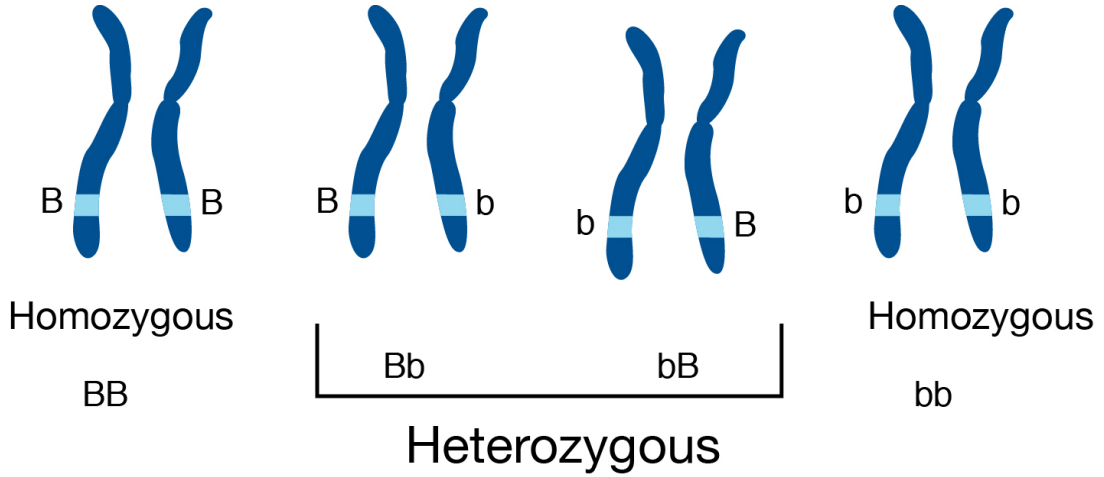
https://en.wikipedia.org/wiki/Ploidy

Ploidy and Variant Calling

Variant callers can use ploidy to improve specificity (avoid false positives) because there are expected variant frequencies, e.g. for diploid:

- Homozygous
  - both copies contain variant
  - fraction of the reads ~1

- Heterozygous –
  - one copy of variant
  - fraction of reads with variant ~0.5



https://en.wikipedia.org/wiki/Ploidy

Interpretation

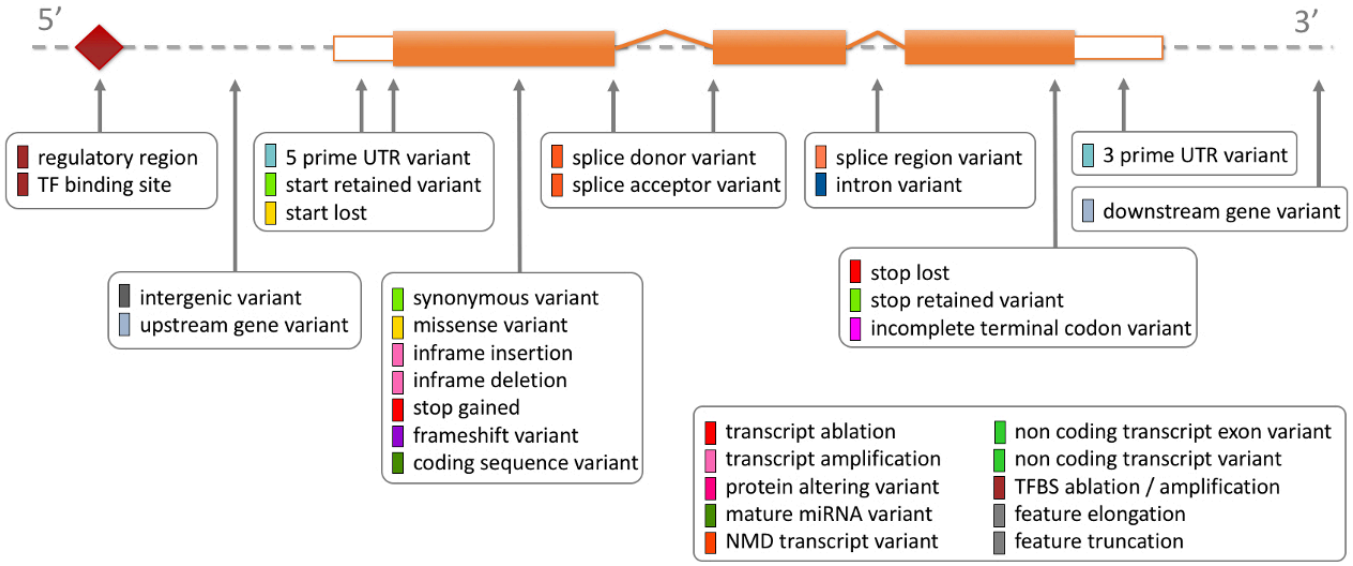
Position 13,635,567
G -> A
6/6 reads -> High confidence

ClinVar: Database of variants in relation to human health

NM_005902.3(SMAD3):c.364G>A (p.Val122Met) — Cite this record

Interpretation:	**Conflicting interpretations of pathogenicity** **Likely pathogenic(1);Uncertain significance(1)**
Review status:	★☆☆☆ criteria provided, conflicting interpretations
Submissions:	2 (Most recent: Jun 10, 2016)
Last evaluated:	Feb 24, 2016
Accession:	VCV000155836.1
Variation ID:	155836
Description:	single nucleotide variant

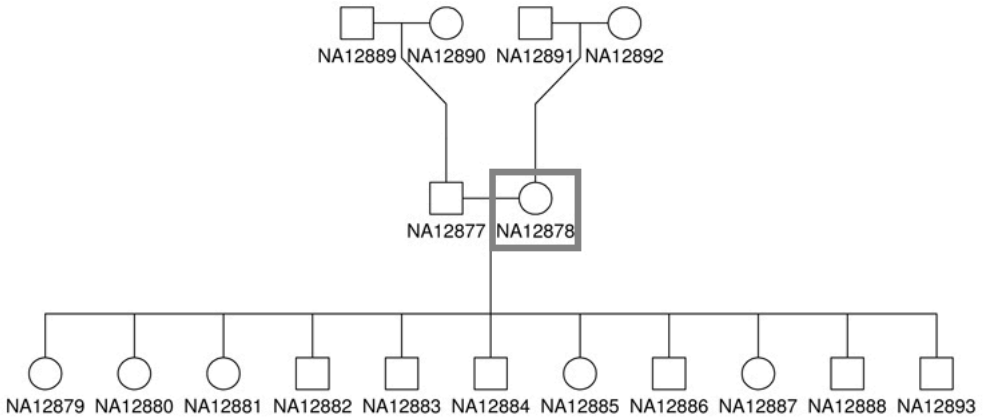
Variant Effect Predictor (VEP) : what is the predicted consequence of the variant in a gene transcript?



Data for this class

GIAB was initiated in 2011 by the National Institute of Standards and Technology "to develop the technical infrastructure (reference standards, reference methods, and reference data) to enable translation of whole human genome sequencing to clinical practice" [1]

The source DNA, known as NA12878, was taken from a single person: the daughter in a father-mother-child 'trio' (she is also mother to 11 children of her own) [4]. Father-mother-child 'trios' are often sequenced to utilize genetic links between family members.



https://github.com/hbctraining/In-depth-NGS-Data-Analysis-Course/blob/master/sessionVI/lessons/01_alignment.md

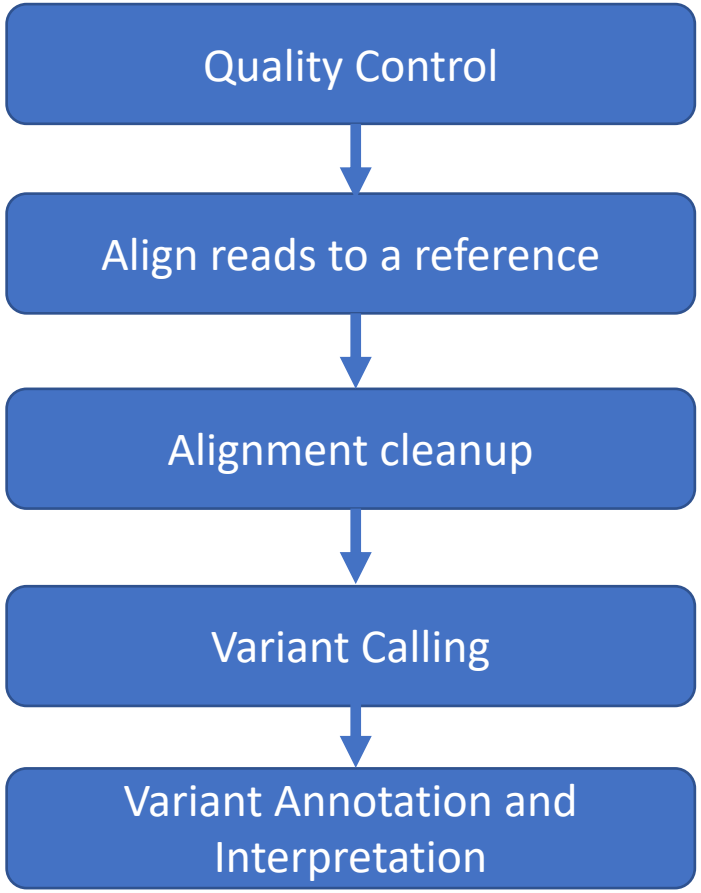
For this class, I've created a small dataset

Sample: NA12878

Gene: Cyp2c19 on chromosome 10

Sequencing: Illumina, **Paired End**, **Exome**

Variant Calling workflow



https://github.com/hbctraining/In-depth-NGS-Data-Analysis-Course

Thank you

Especially to:
Wenwen Huo, postdoctoral research scholar Isberg Lab, Tufts Medical School
Shawn Doughty, Research Computing Manager, TTS
Delilah Maloney, High Performance Computing Specialist, TTS
Susi Remondi, Senior Technical Training Specialist, TTS

For more tutorials like these on doing Bioinformatics on the Tufts HPC cluster:
https://sites.tufts.edu/biotools/tutorials/

For more great bioinformatics tutorials:
https://github.com/hbctraining/

For questions on Bioinformatics or the Tufts HPC, contact tts-research@tufts.edu