Intro to Bioinformatics using Tufts HPC

Rebecca Batorsky

Sr Bioinformatics Specialist

Dec 2019

Outline

You’ll need:

- Cluster Account – please let me know if you don’t have one
- Basic knowledge of Linux

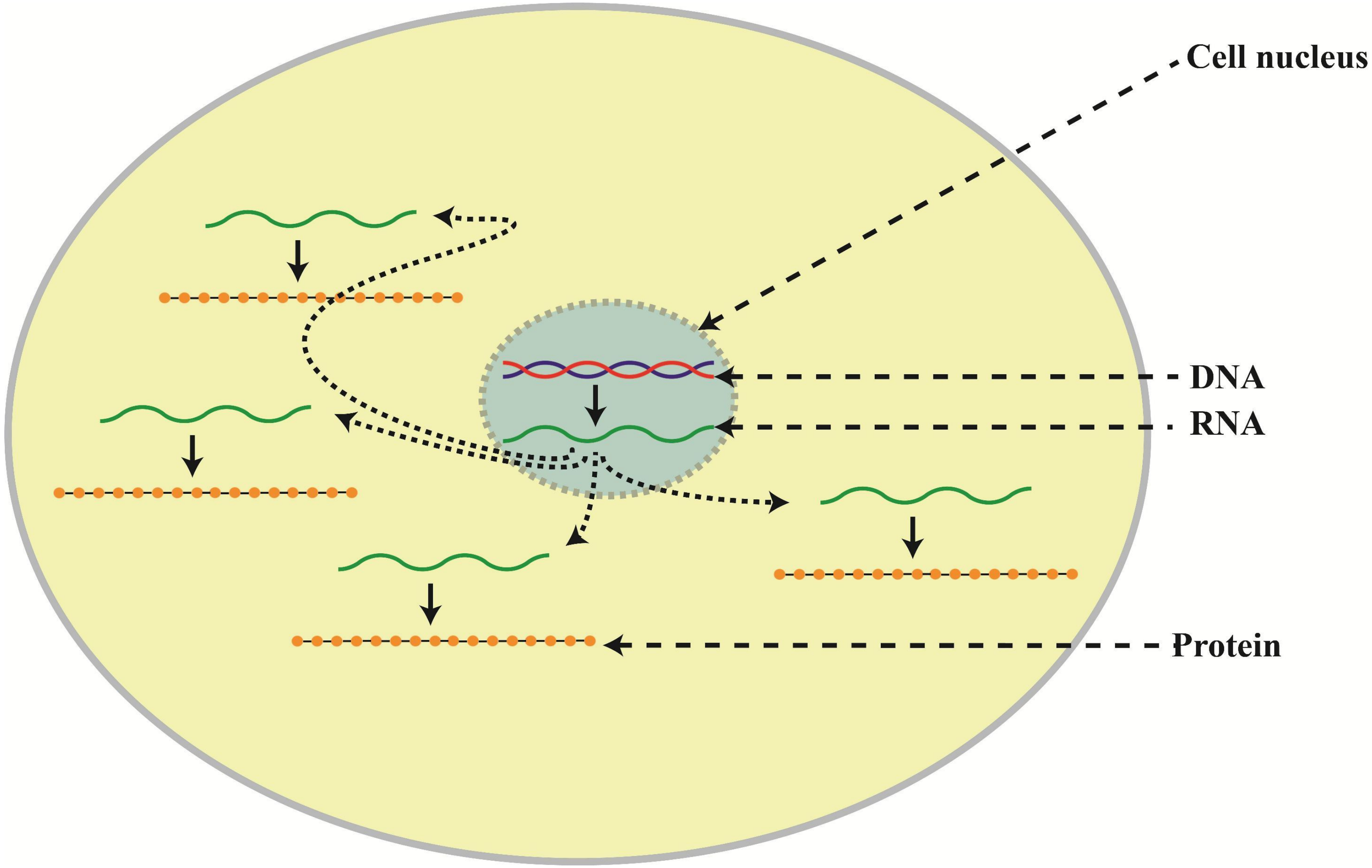
Our goals:

- Writing and running bash scripts
- Intro to several bioinformatics tools: BWA, Samtools, Picard, GATK
- Variant Calling and Interpretation

Course format:

- Short explanations followed by hands on exercises
- Working with a partner is encouraged
- Please ask questions!

DNA and RNA in a cell

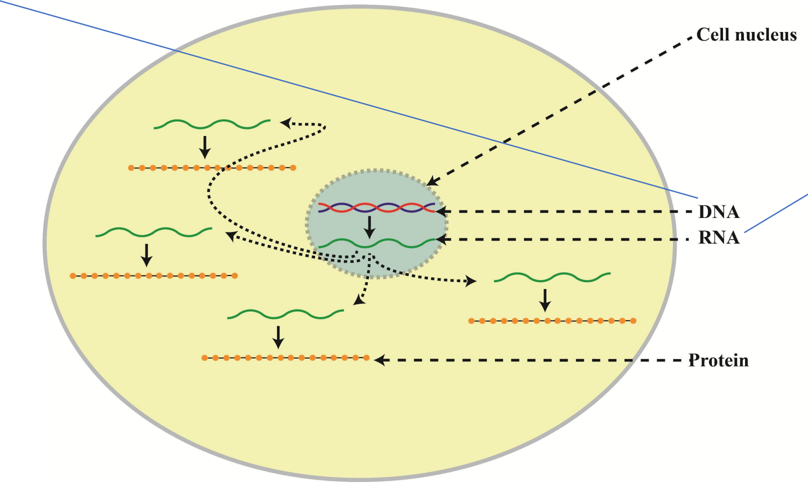


https://i0.wp.com/science-explained.com/wp-content/uploads/2013/08/Cell.jpg

Two common analysis goals

DNA Sequencing

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



RNA Sequencing

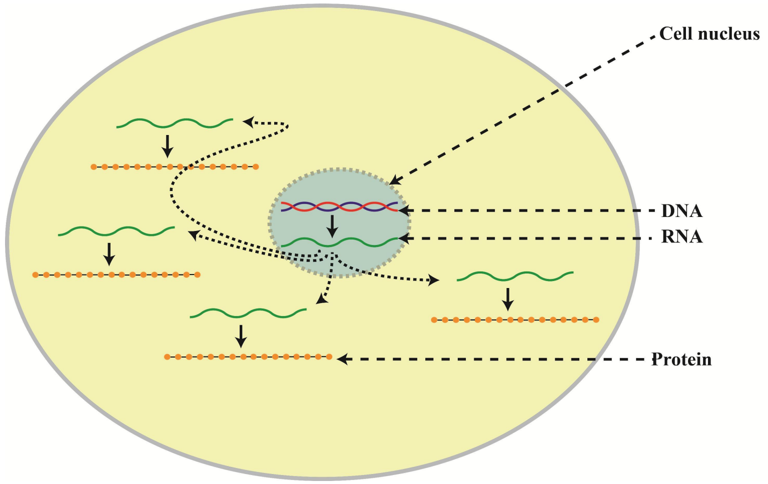
- Copy of a transcript per cell depends on gene expression
- Analysis goal: Differential expression and interpretation

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

Today we will cover DNA sequencing

DNA Sequencing

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



Not today!

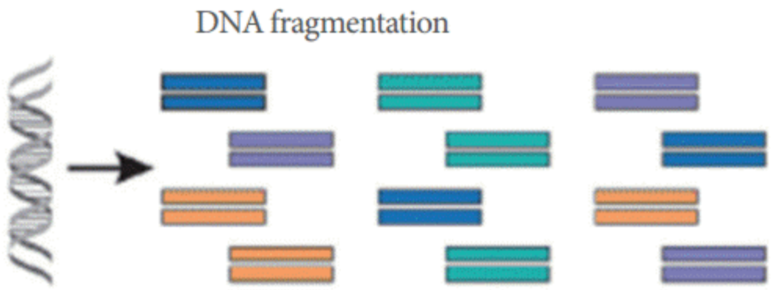
RNA Sequen

- Copy of a ge cell depends on gene ex
- Analy l: Diffe l expre n and interp tion

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

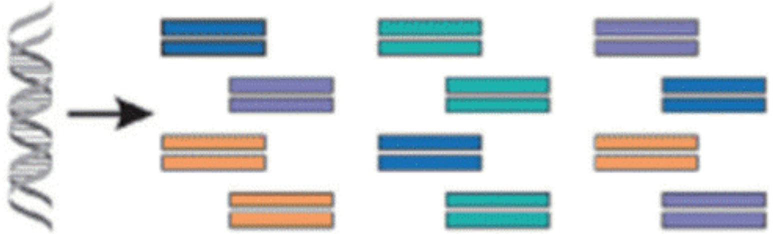
Next Generation Sequencing (NGS)

1) DNA is fragmented
2) **Adaptors** ligated to fragments
3) **Cluster** generation
4) Extension of fragments with fluorescently tagged nucleotides
5) Cyclic readout by imaging the array

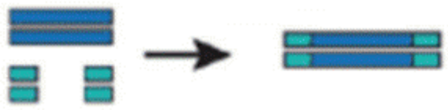


Next Generation Sequencing (NGS) -1

1) DNA is fragmented
2) **Adaptors** ligated to fragments
3) **Cluster** generation
4) Extension of fragments with fluorescently tagged nucleotides
5) Cyclic readout by imaging the array



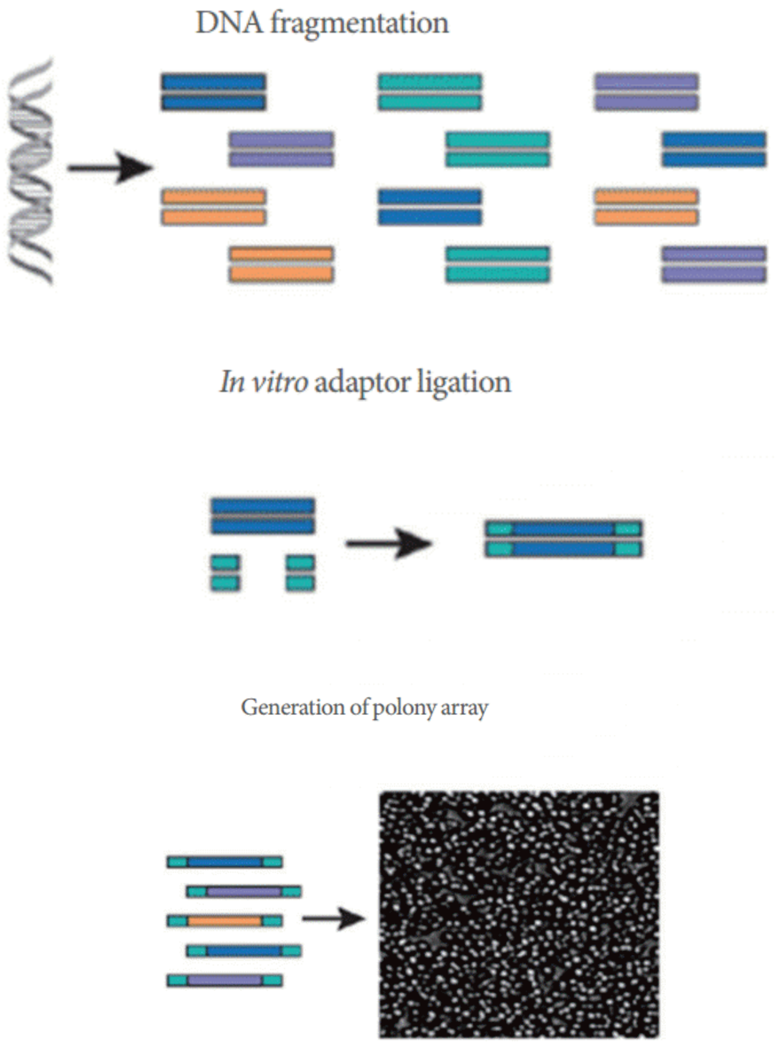
DNA fragmentation



In vitro adaptor ligation

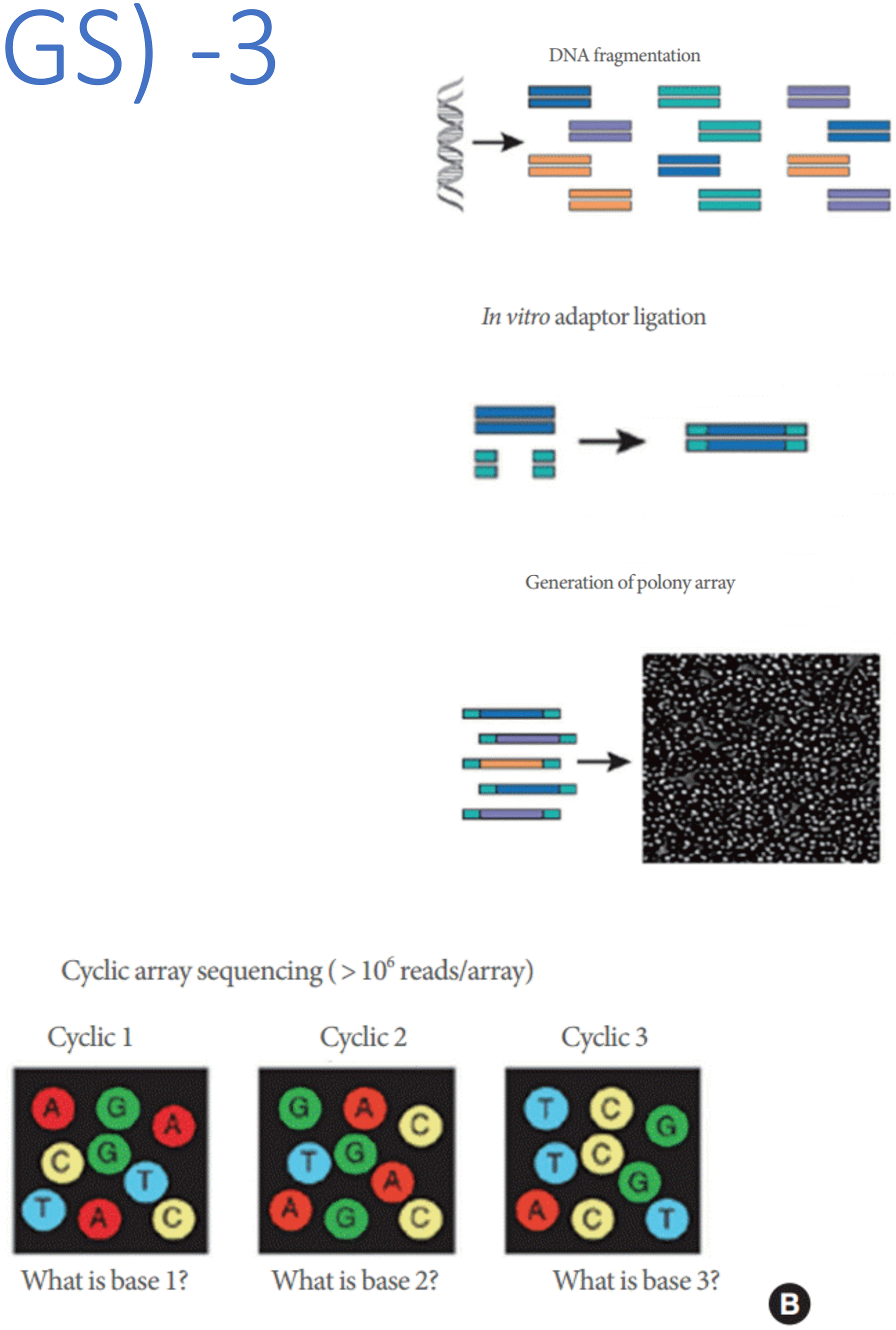
Next Generation Sequencing (NGS) -2

1) DNA is fragmented
2) **Adaptors** ligated to fragments
3) **Cluster** generation
4) Extension of fragments with fluorescently tagged nucleotides
5) Cyclic readout by imaging the array



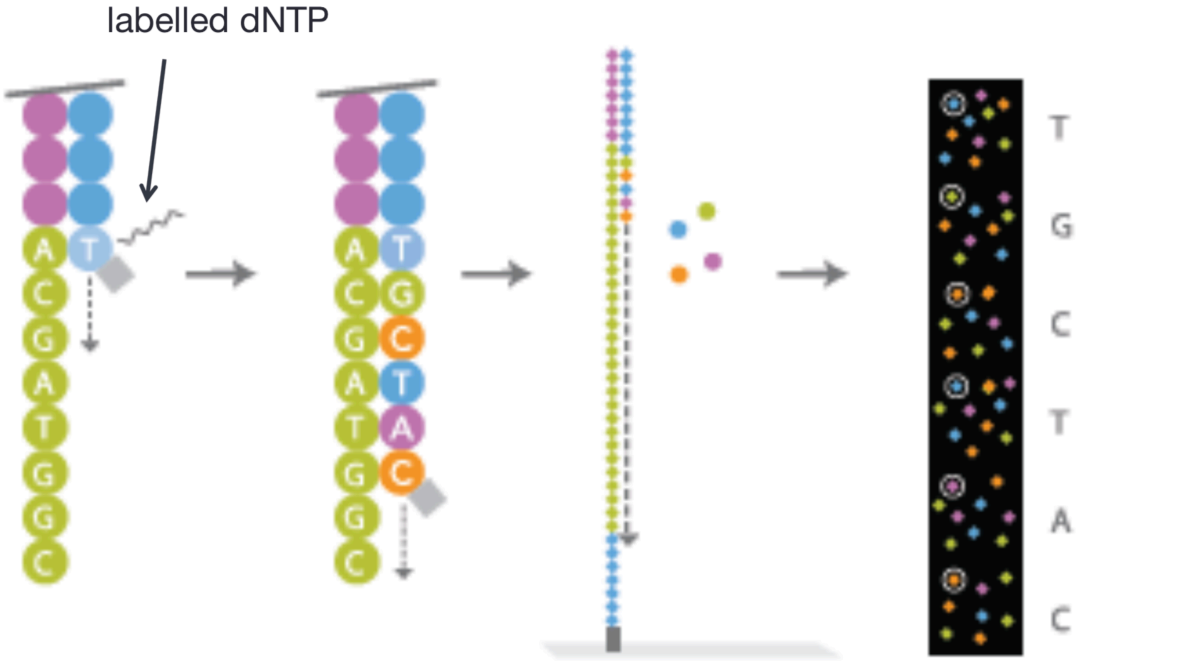
Next Generation Sequencing (NGS) -3

1) DNA is fragmented
2) **Adaptors** ligated to fragments
3) **Cluster** generation
4) Extension of fragments with fluorescently tagged nucleotides
5) Cyclic readout by imaging the array



Jay Shendure & Hanlee Ji, Nature Biotechnology 26, 1135 - 1145 (2008)

Next Generation Sequencing (NGS) -4



1. extend 1st base
2. read
3. deblock

repeat for 50 – 100 bp

generate base calls

Illumina Video!
https://www.illumina.com/science/technology/next-generation-sequencing/sequencing-technology.html

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**

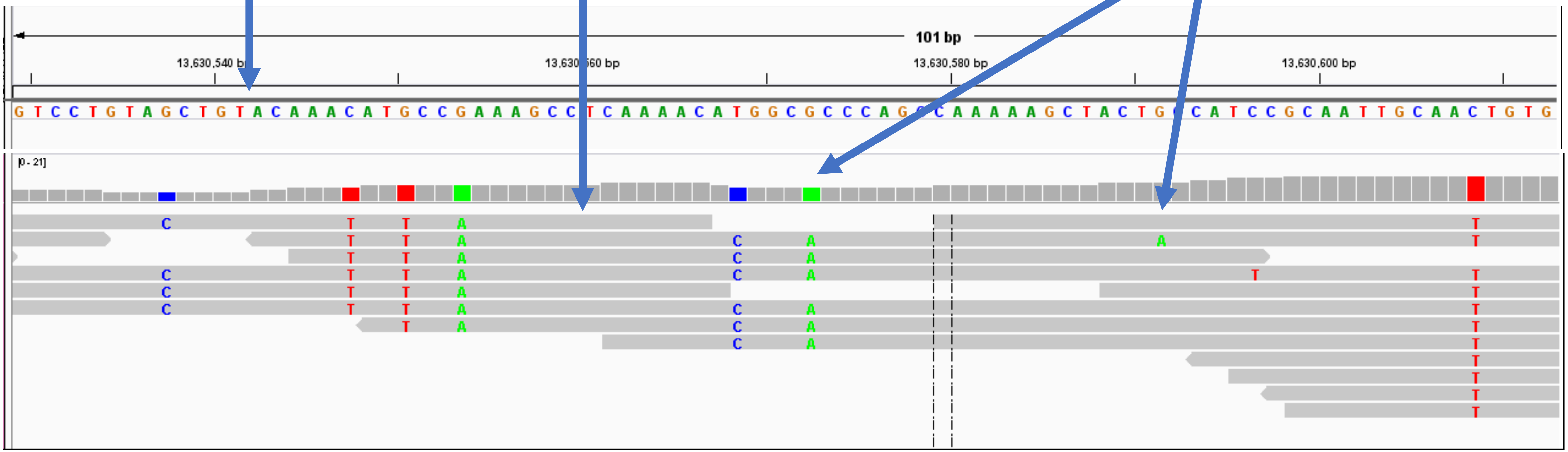
https://www.abmgood.com/marketing/knowledge_base/next_generation_sequencing_data_analysis.php

Variant Calling

- 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

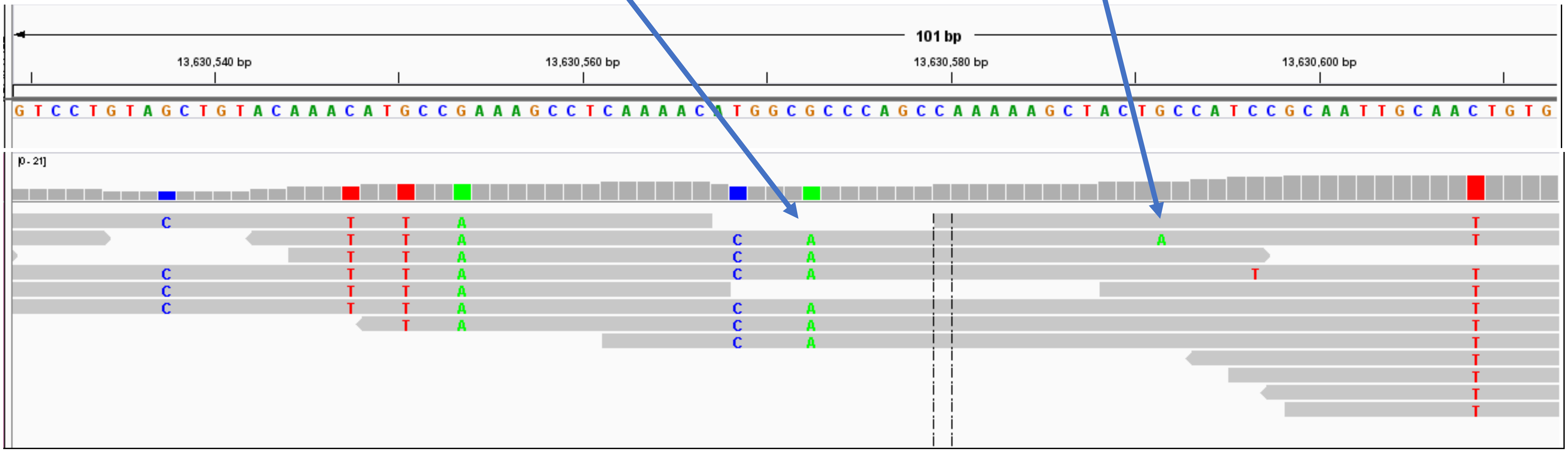

101 bp
13,630,540 bp
13,630,560 bp
13,630,580 bp
13,630,600 bp
G T C C T G T A G C T G T A C A A A C A T G C C G A A A G C C T C A A A A C A T G G C G C C C A G G C C A A A A A G C T A C T G C C A T C C G C A A T T G C A A C T G T G
[0 - 21]
C T T A C A T
C T T A C A T
C T T A C A T
T T A C A T
C T T A C A T
C T T A C A T
T T A C A T
C A T
T
T
T
T
A
T

Variant Calling

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

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

We would like a list of variants along with the confidence

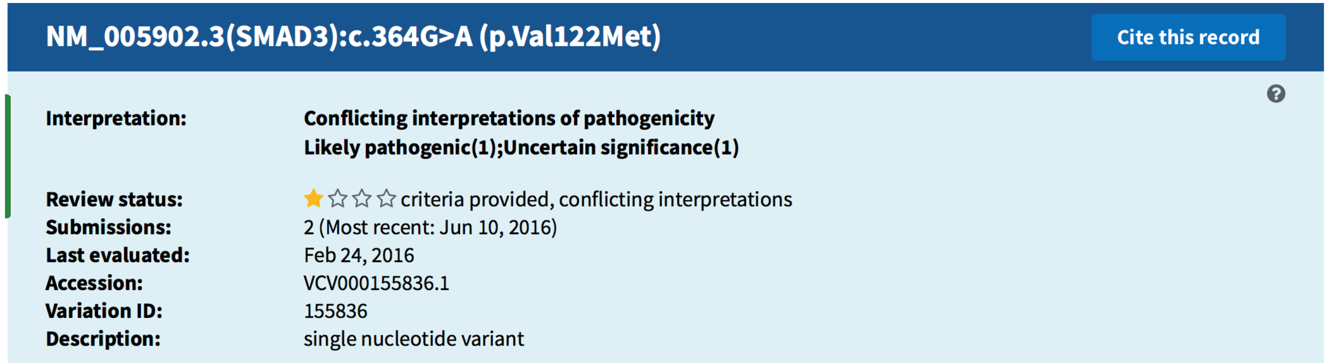


101 bp
13,630,540 bp
13,630,560 bp
13,630,580 bp
13,630,600 bp
G T C C T G T A G C T G T A C A A A C A T G C C G A A A G C C T C A A A A C A T G G C G C C C A G C C A A A A A G C T A C T G C C A T C C G C A A T T G C A A C T G T G
[0 - 21]

Interpretation

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

ClinVar: Database of variants in relation to human health

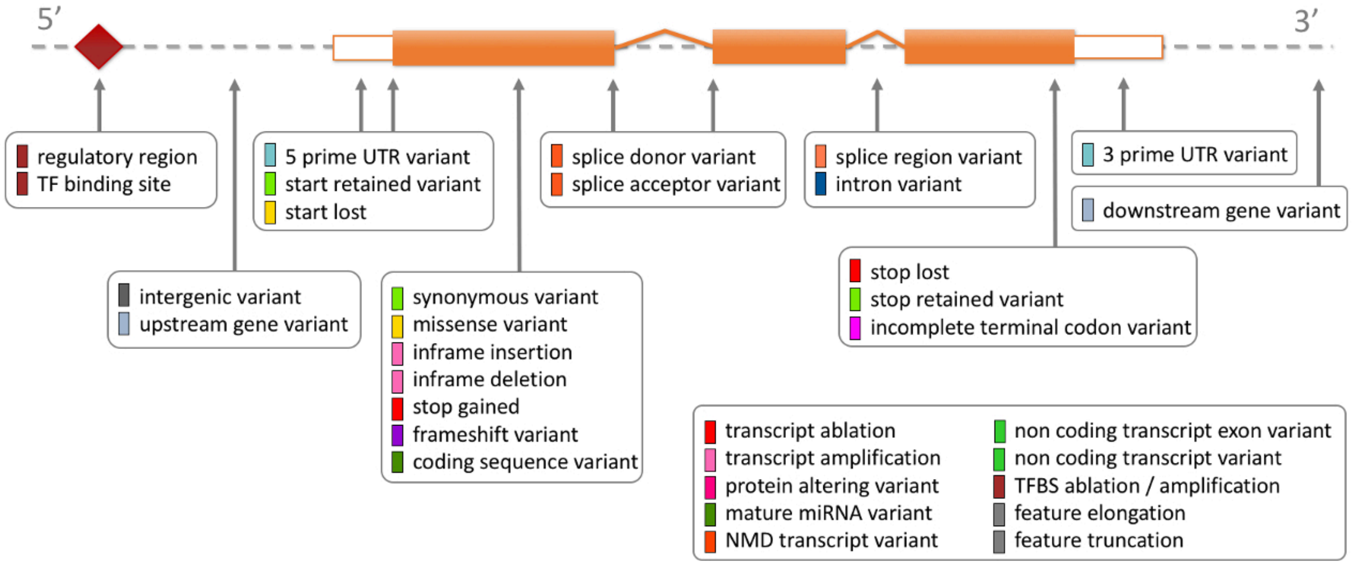


NM_005902.3(SMAD3):c.364G>A (p.Val122Met)

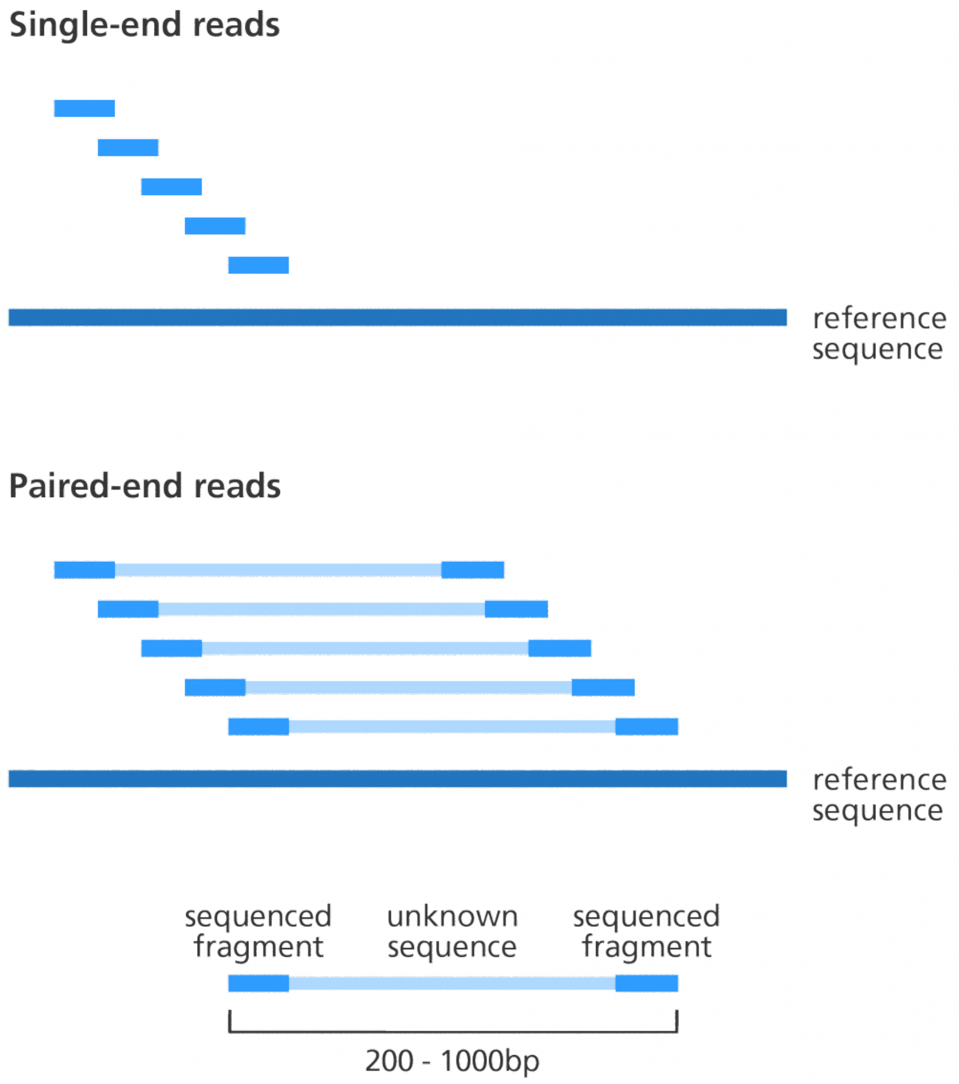
Interpretation: **Conflicting interpretations of pathogenicity**
Likely pathogenic(1);Uncertain significance(1)

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?

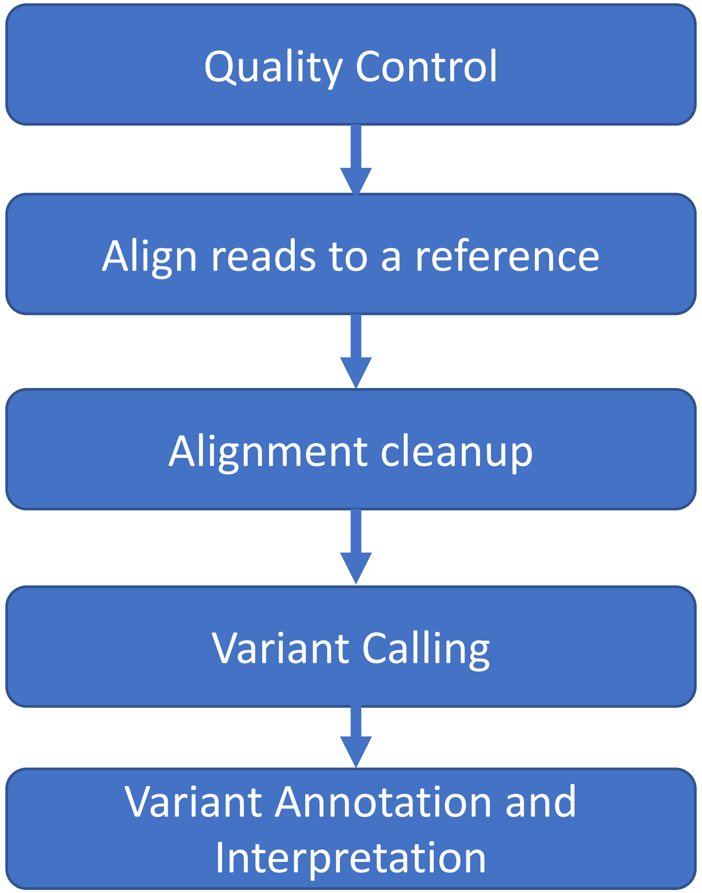


Paired end vs Single end reads



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

Variant Calling workflow



Quality Control

Align reads to a reference

Alignment cleanup

Variant Calling

Variant Annotation and Interpretation

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