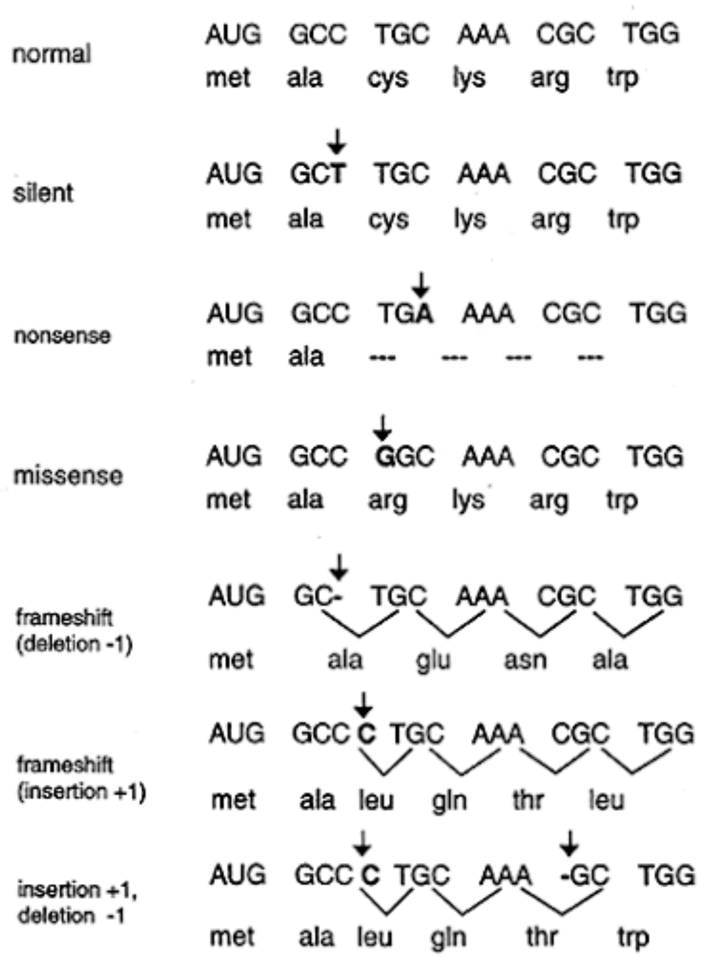
Variant Calling

BIOL 435/535: Bioinformatics

April 7th, 2022

Mutations – Inherited changes in nucleotide sequence



synonymous
nonsynonymous

1. How to confidently identify true variants
2. Low-frequency variants
3. Variant annotation

Class brainstorm:

What are some important considerations in variant calling?

How to confidently identify true variants

Primary considerations

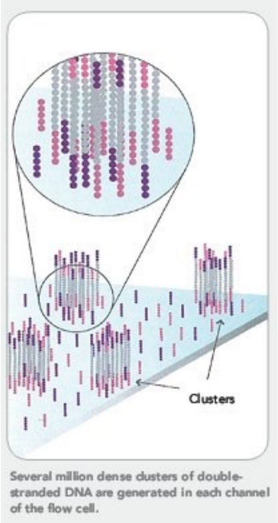
- Sequencing method/approach
- Depth of coverage
- Variant type
- Inheritance mode
- Reference quality
- Bioinformatic tool

How to confidently identify true variants

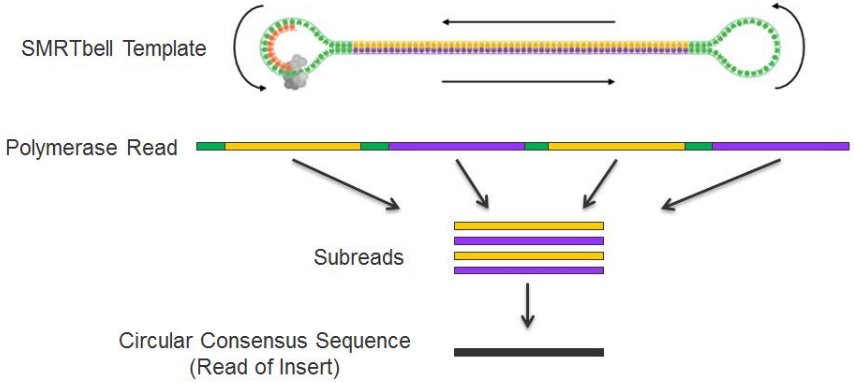
Primary considerations

- **Sequencing method/approach**
- Depth of coverage
- Variant type
- Inheritance mode
- Reference quality
- Bioinformatic tool

Short vs. long read:


Clusters
Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

Illumina



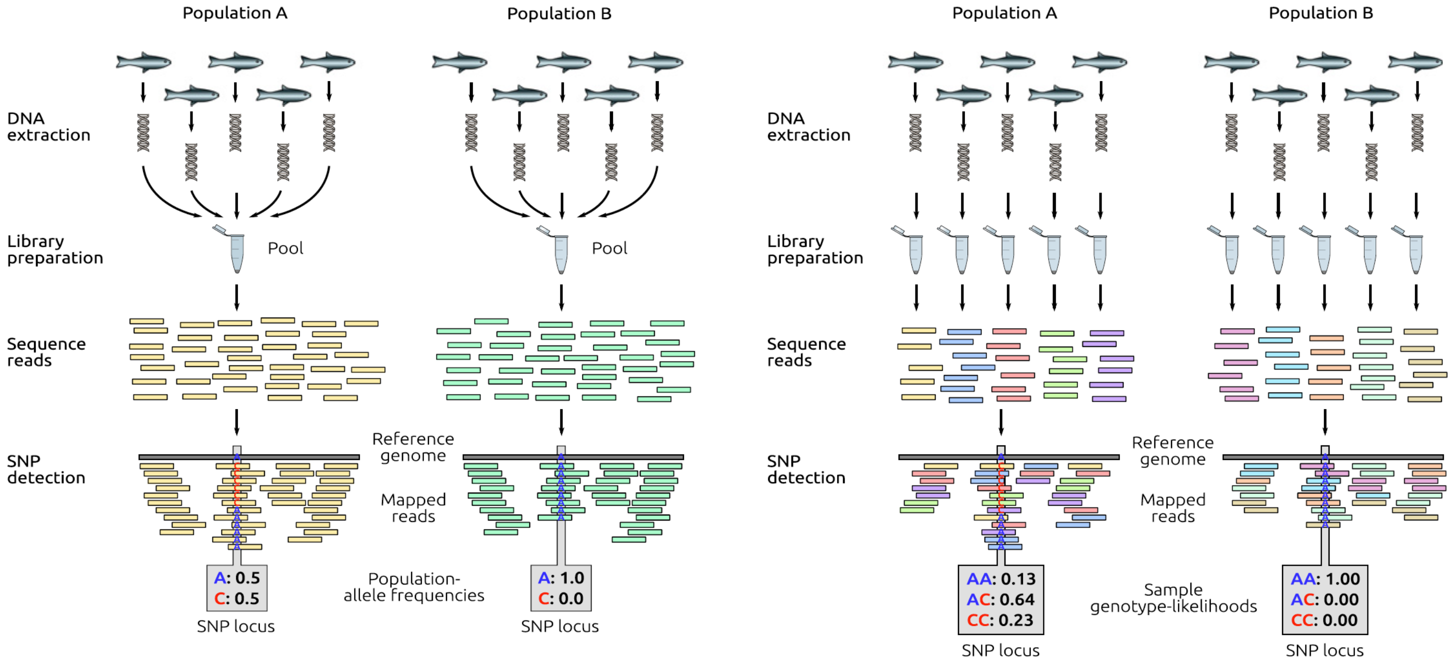
Pacbio

How to confidently identify true variants

Primary considerations

- **Sequencing method/approach**
- Depth of coverage
- Variant type
- Inheritance mode
- Reference quality
- Bioinformatic tool

Individual vs. Multiplex vs. PoolSeq



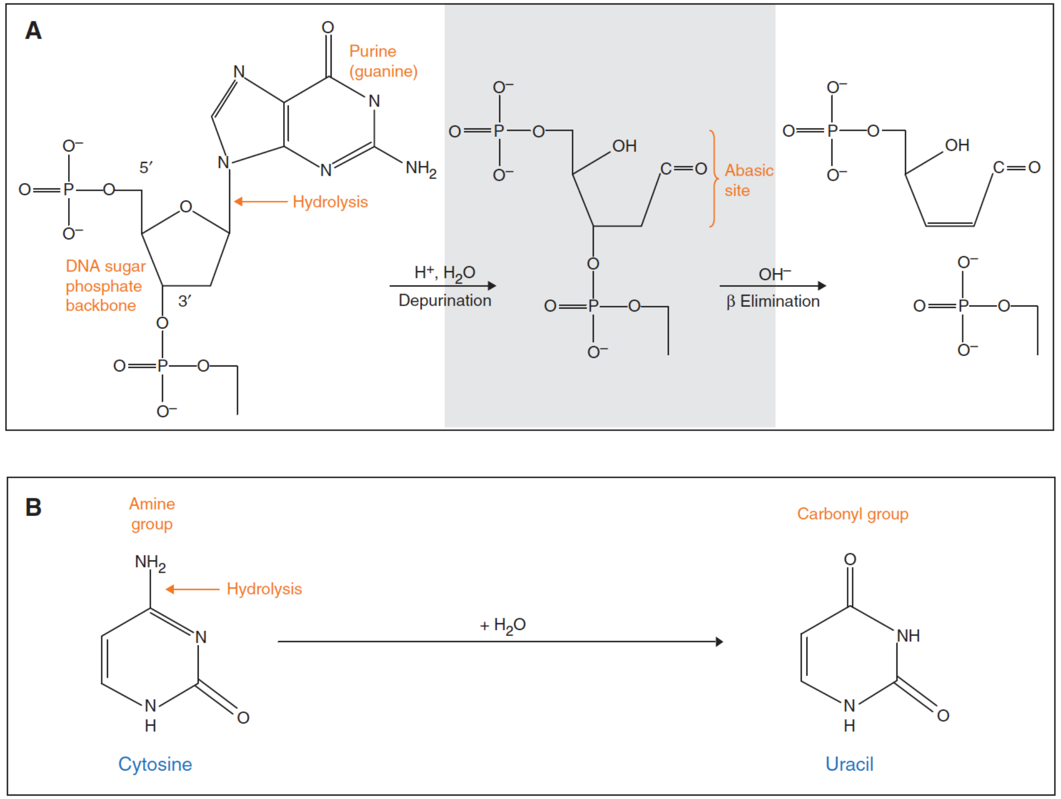
Pardo & Ruzzante. 2017, *Mol Ecol*

How to confidently identify true variants

Primary considerations

- **Sequencing method/approach**
- Depth of coverage
- Variant type
- Inheritance mode
- Reference quality
- Bioinformatic tool

DNA Damage (e.g., Ancient DNA)



Most C->T (and A->G) changes are the result of DNA Damage

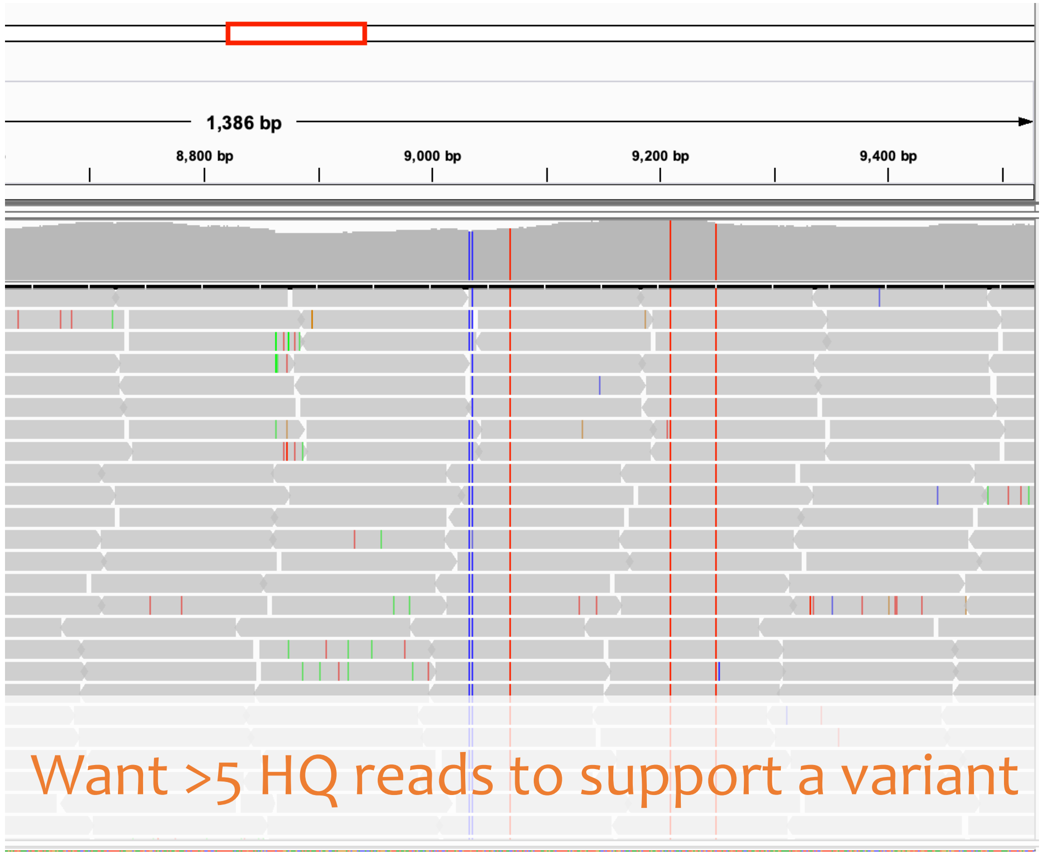
Depardy et al., 2013, *CSHP*

How to confidently identify true variants

Primary considerations

- Sequencing method/approach
- **Depth of coverage**
- Variant type
- Inheritance mode
- Reference quality
- Bioinformatic tool

Goldilocks Principle of Read Depth



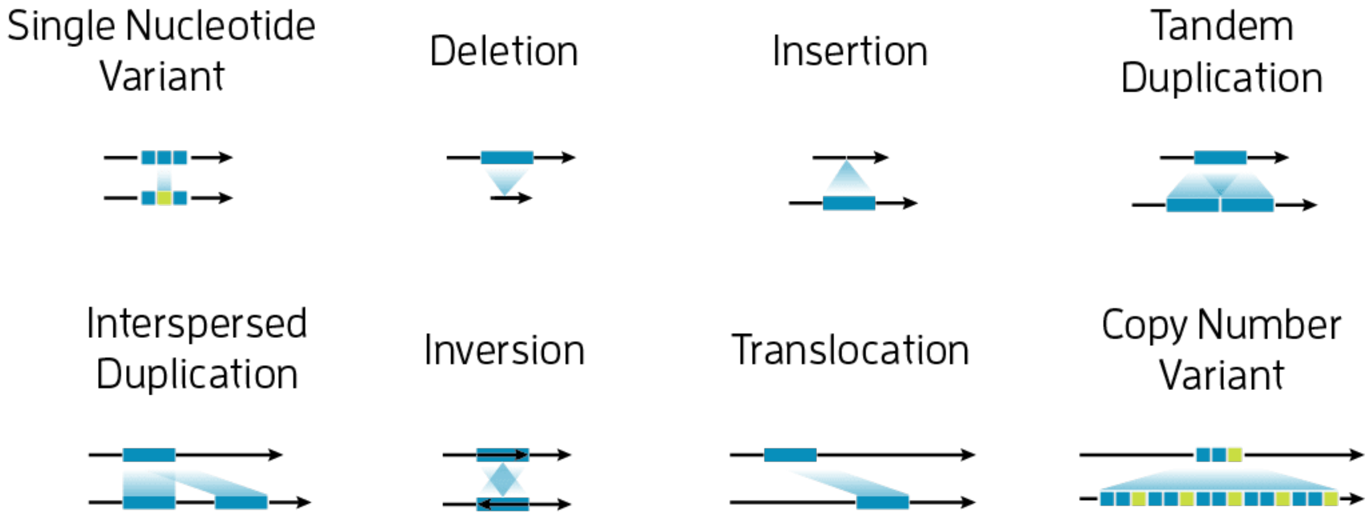
Want >5 HQ reads to support a variant

How to confidently identify true variants

Primary considerations

- Sequencing method/approach
- Depth of coverage
- **Variant type**
- Inheritance mode
- Reference quality
- Bioinformatic tool

SNV vs. In/Del vs. SV



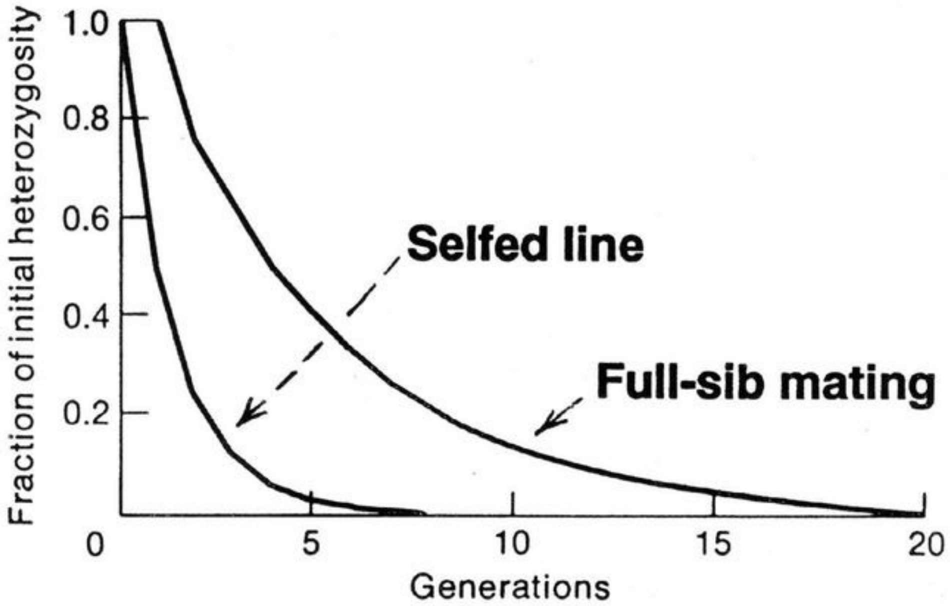
Types of Variants

How to confidently identify true variants

Primary considerations

- Sequencing method/approach
- Depth of coverage
- Variant type
- **Inheritance mode**
- Reference quality
- Bioinformatic tool

Selfing vs. Outcrossing

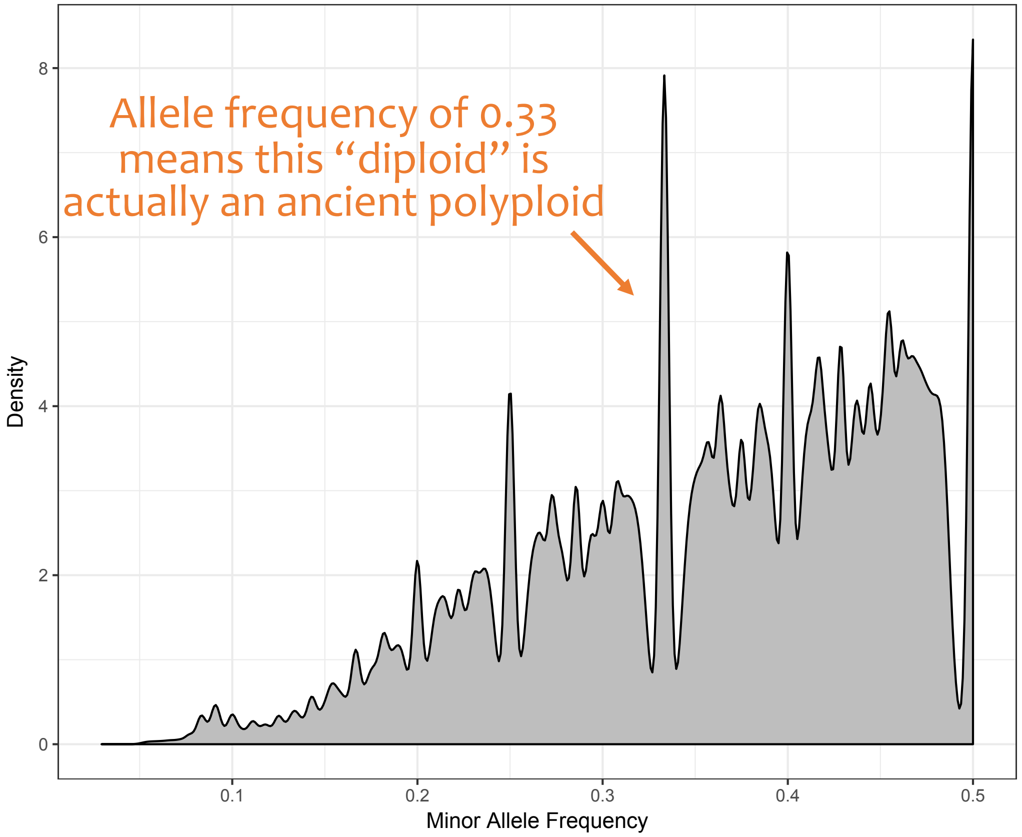


How to confidently identify true variants

Primary considerations

- Sequencing method/approach
- Depth of coverage
- Variant type
- Inheritance mode
- **Reference quality**
- Bioinformatic tool

Non-model reference genomes are often complex



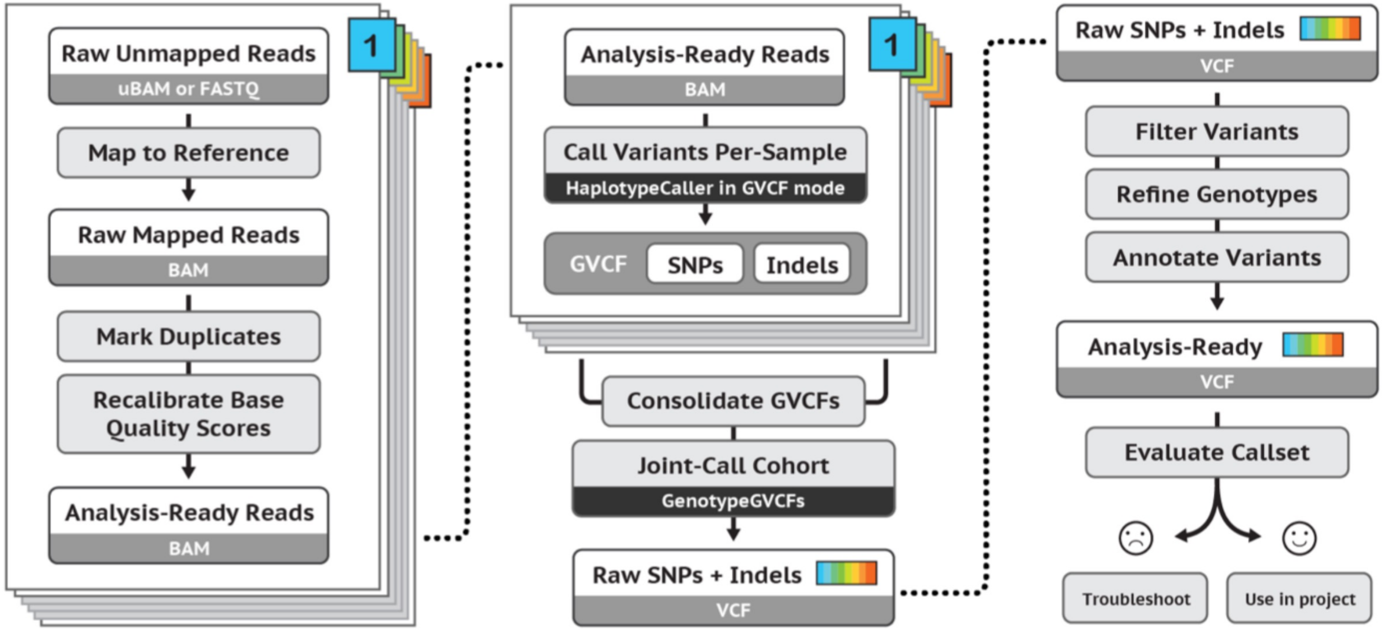
Allele frequency of 0.33 means this “diploid” is actually an ancient polyploid

How to confidently identify true variants

Primary considerations

- Sequencing method/approach
- Depth of coverage
- Variant type
- Inheritance mode
- Reference quality
- **Bioinformatic tool**

BCFTOOLS vs. GATK vs. FreeBayes vs. DeepVariant



Low-frequency variants:
Distinguishing between sequencing errors, DNA damage, and true mutations

Types of low-frequency variants:

- Somatic mutations
- Heteroplasmy
- Rare polymorphisms in PoolSeq
- DNA Damage

Illumina error rate ~ 10^{-3}

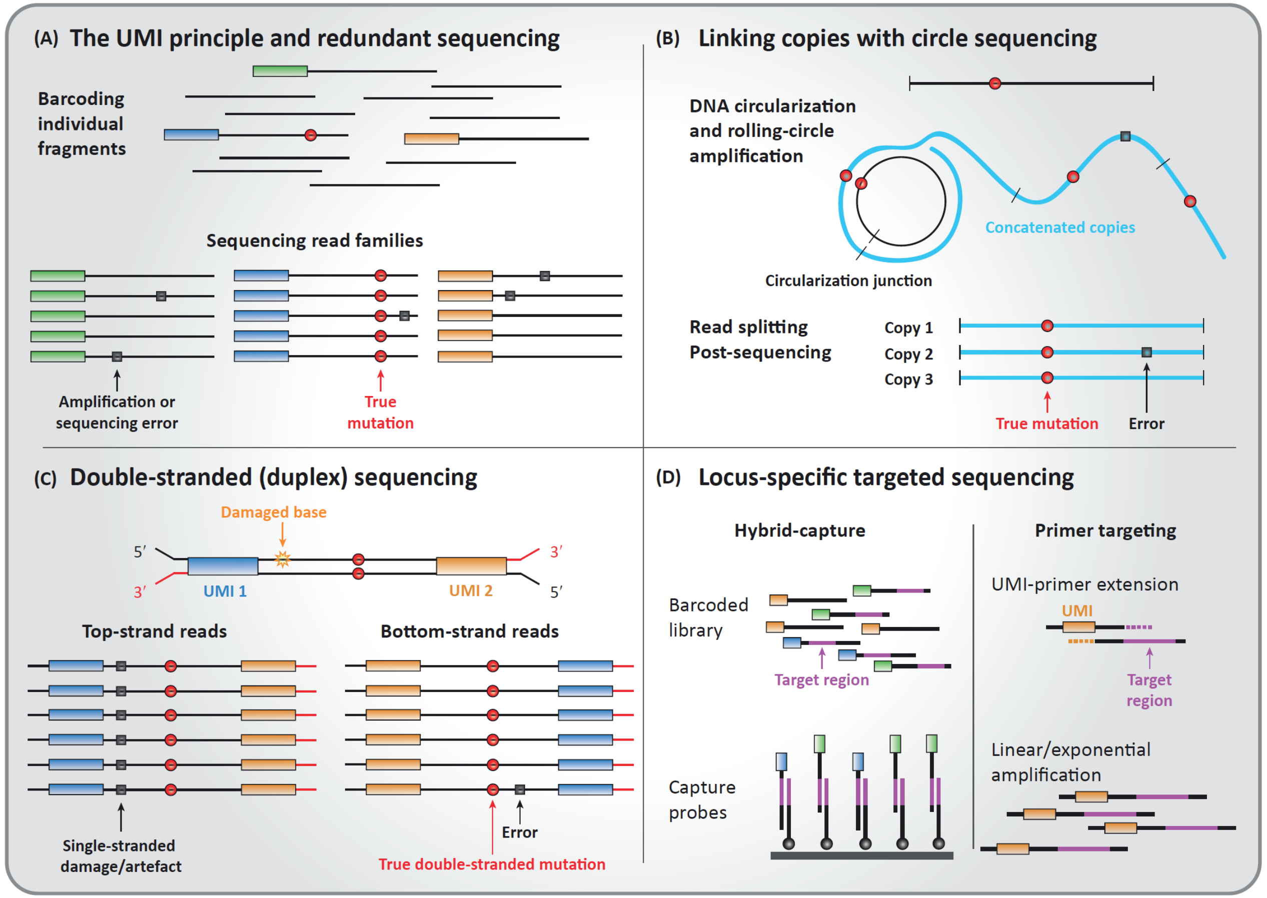
Coverage > 1000x means that you'll have an erroneous base call at **EVERY SITE**

How to mitigate?

Low-frequency variants:
Distinguishing between sequencing errors, DNA damage, and true mutations

Types of low-frequency variants:

- Somatic mutations
- Heteroplasmy
- Rare polymorphisms in PoolSeq
- DNA Damage


(A) The UMI principle and redundant sequencing
Barcoding individual fragments
Sequencing read families
Amplification or sequencing error
True mutation
(B) Linking copies with circle sequencing
DNA circularization and rolling-circle amplification
Concatenated copies
Circularization junction
Read splitting Post-sequencing
Copy 1
Copy 2
Copy 3
True mutation
Error
(C) Double-stranded (duplex) sequencing
Damaged base
5′
3′
3′
5′
UMI 1
UMI 2
Top-strand reads
Bottom-strand reads
Single-stranded damage/artefact
Error
True double-stranded mutation
(D) Locus-specific targeted sequencing
Hybrid-capture
Primer targeting
Barcoded library
Target region
UMI-primer extension
UMI
Target region
Capture probes
Linear/exponential amplification

Sloan et al., 2017, *TIBTECH*

Variant annotation:

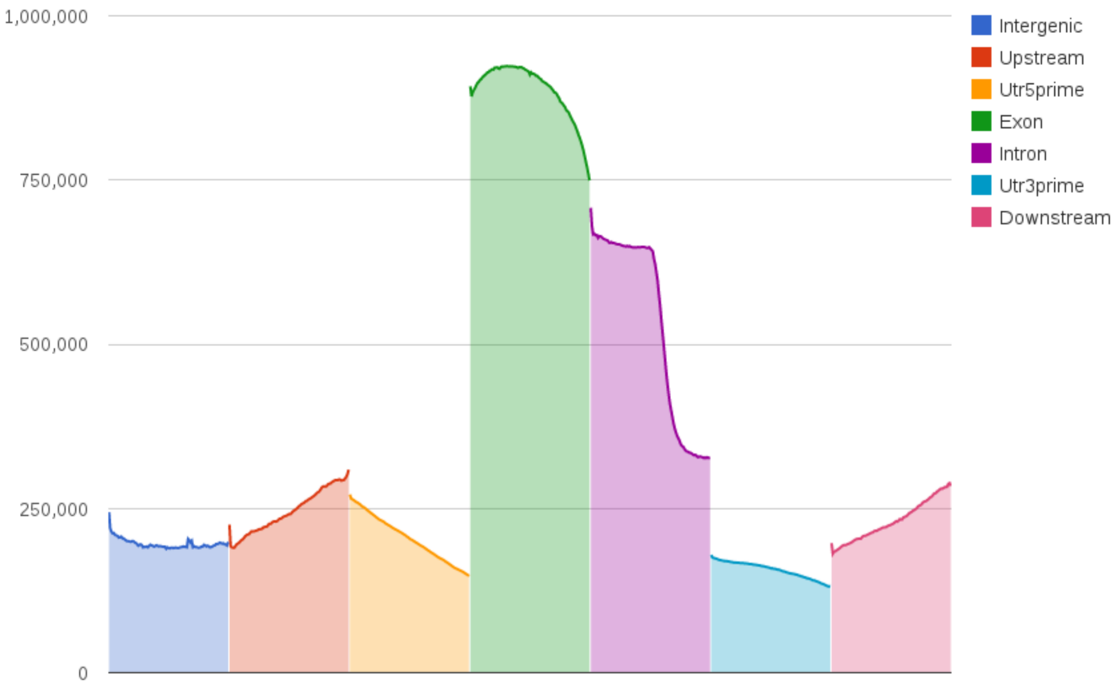
[SnpEff](), [VEP](), [ANNOVAR]()

Effect on function:

- Synonymous vs. Nonsynonymous
- Loss-of-Function

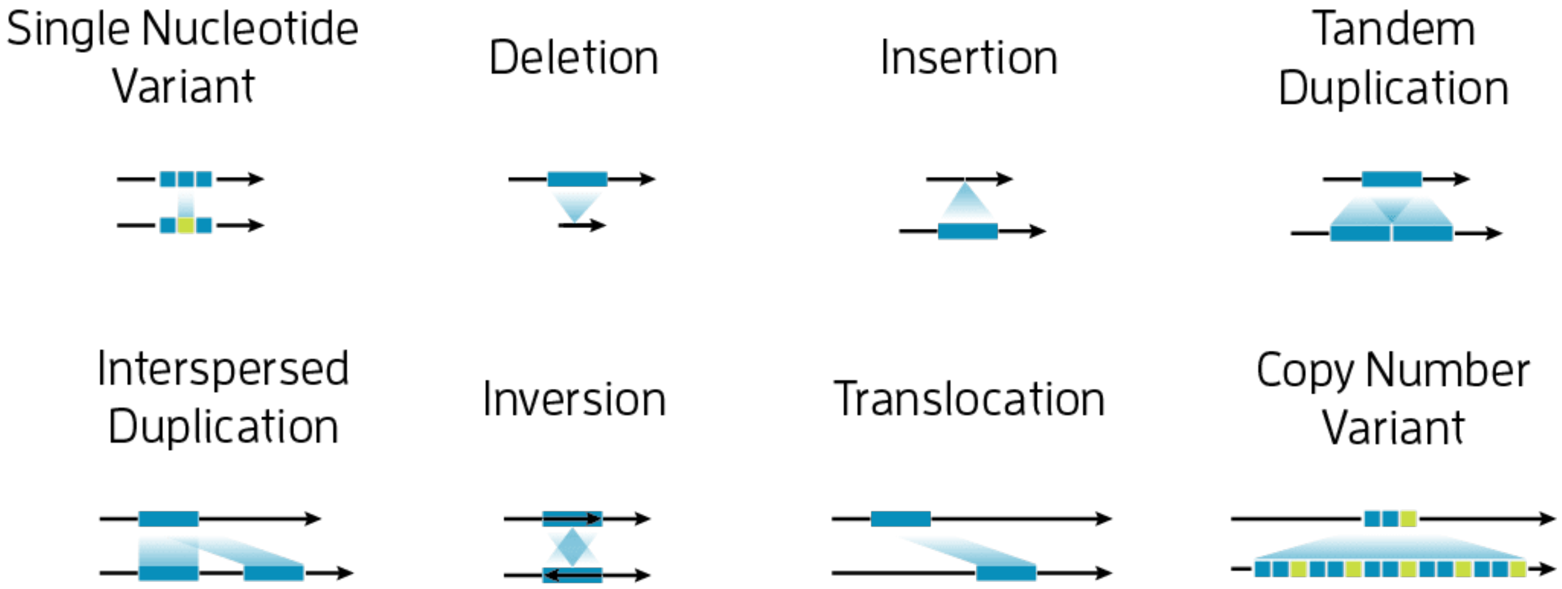
Genomic location:

- Gene of interest
- Position in gene (e.g., exon/intron/UTR)



Visualizing with IGV

Structural Variant Detection



Single Nucleotide Variant

Deletion

Insertion

Tandem Duplication

Interspersed Duplication

Inversion

Translocation

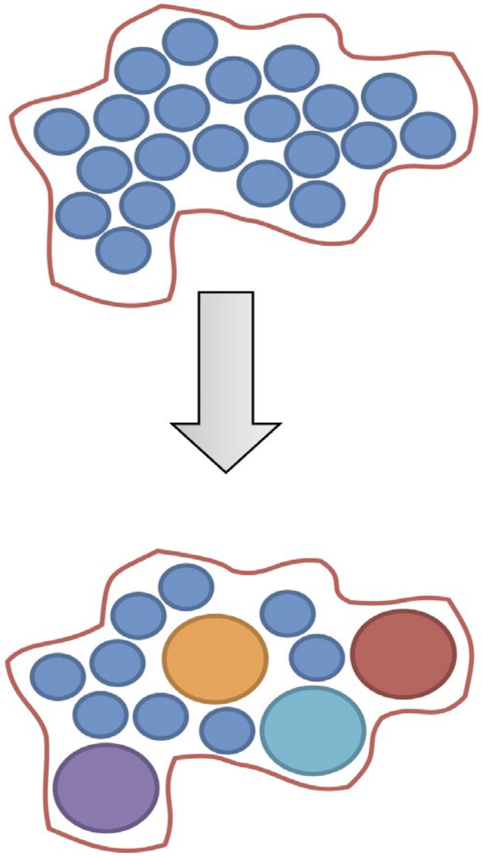
Copy Number Variant

Types of Variants

Structural variant detection depends upon context, experimental design

Sample heterogeneity

- Individual re-sequencing
- Somatic tissue
- Population-level sampling
- Tumor cells
- Plant mitochondria

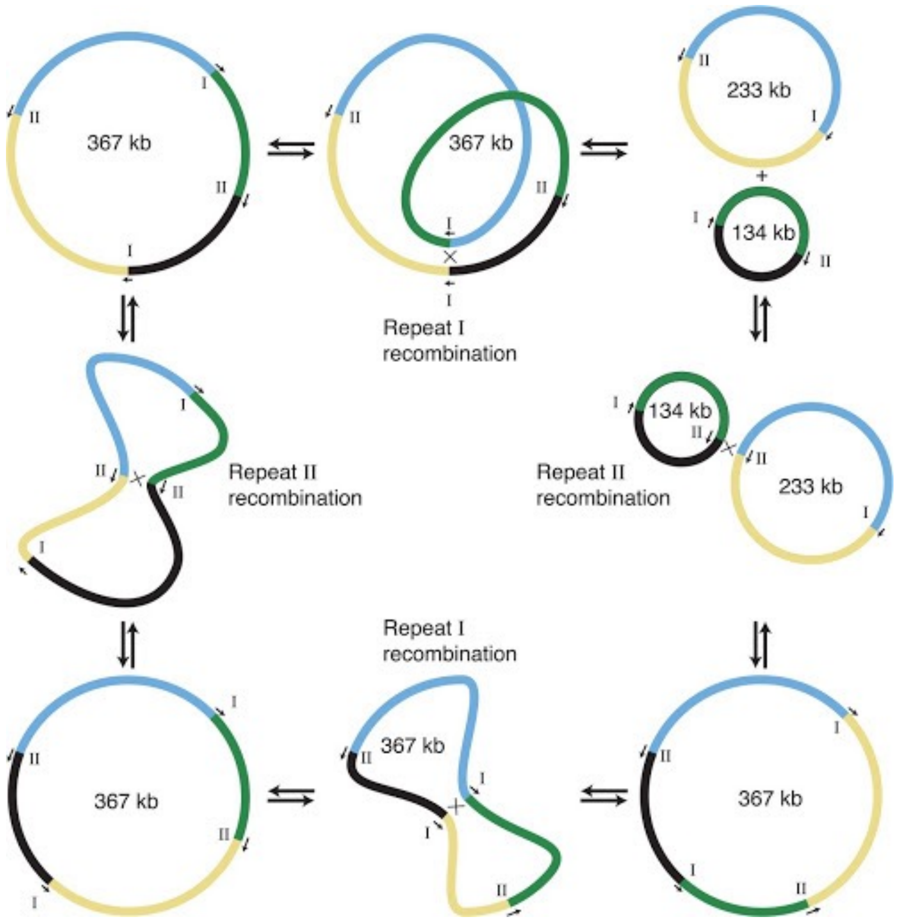


Schoenfeld & Fox 2013

Structural variant detection depends upon context, experimental design

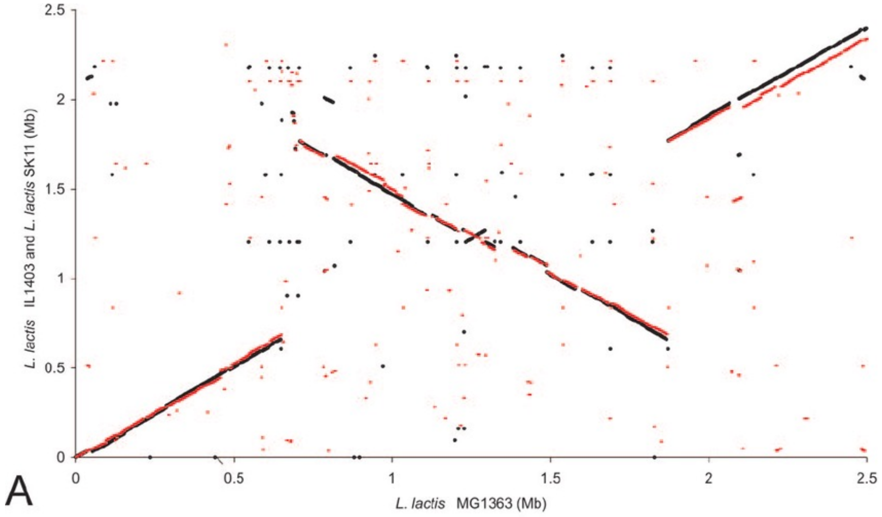
Sample heterogeneity

- Individual re-sequencing
- Somatic tissue
- Population-level sampling
- Tumor cells
- **Plant mitochondria**

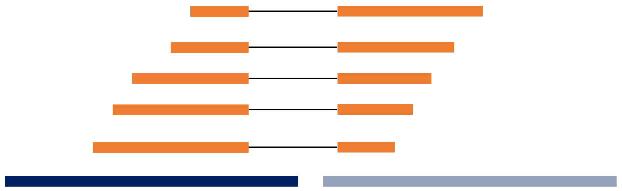


Unseld et al. 1997

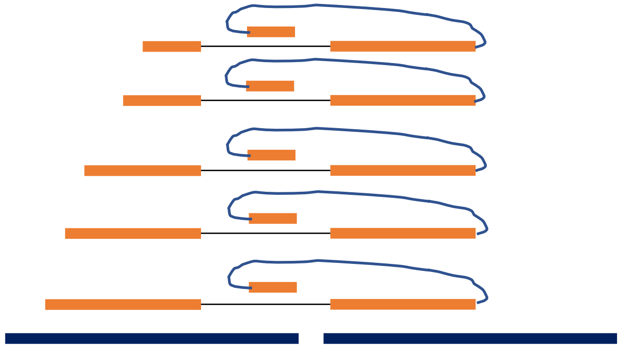
Three general approaches for structural variant calling



De novo assembly, alignment



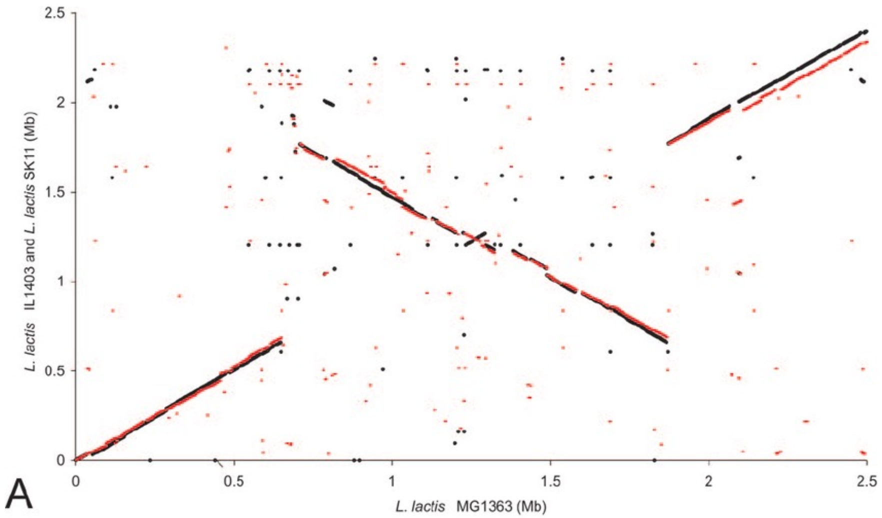
Short read mapping



Long read mapping

Mahmoud et al. 2019

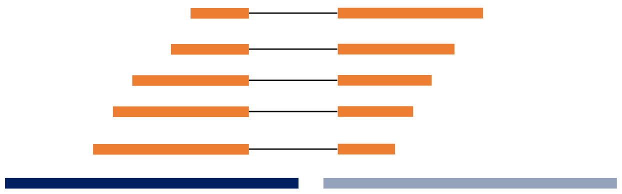
Assembly-to-assembly comparison



De novo assembly, alignment

- Genome-to-genome alignment
- Low sample heterogeneity
- Easy to visualize with a dot plot
- **Diploid assemblies**

Short-read mapping to identify structural rearrangements

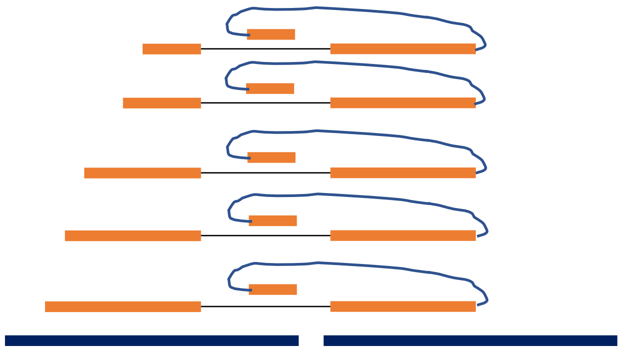


Short read mapping

- Allow reads to map to multiple locations
- Depth allows for high sample heterogeneity
- Often difficult to visualize, as reads don't often span structure
- **Good for finding breakpoints in heterogeneous samples**

Mahmoud et al. 2019

Long-read mapping to identify structural rearrangements

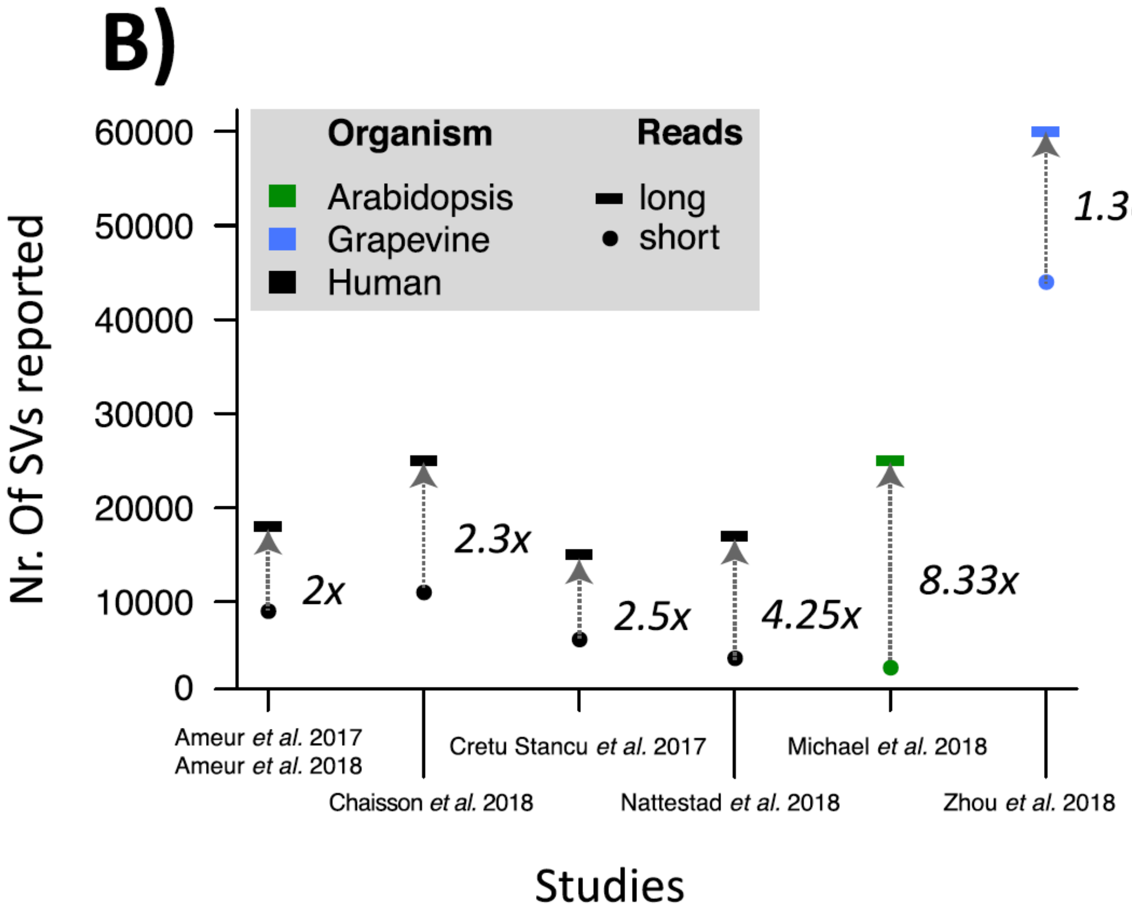


Long read mapping

- Allow reads to map to multiple locations
- Don’t need a ton of depth
- Reads are long enough to span across structural element
- Occasional library artifacts to watch out for
- **Good for homogeneous samples**

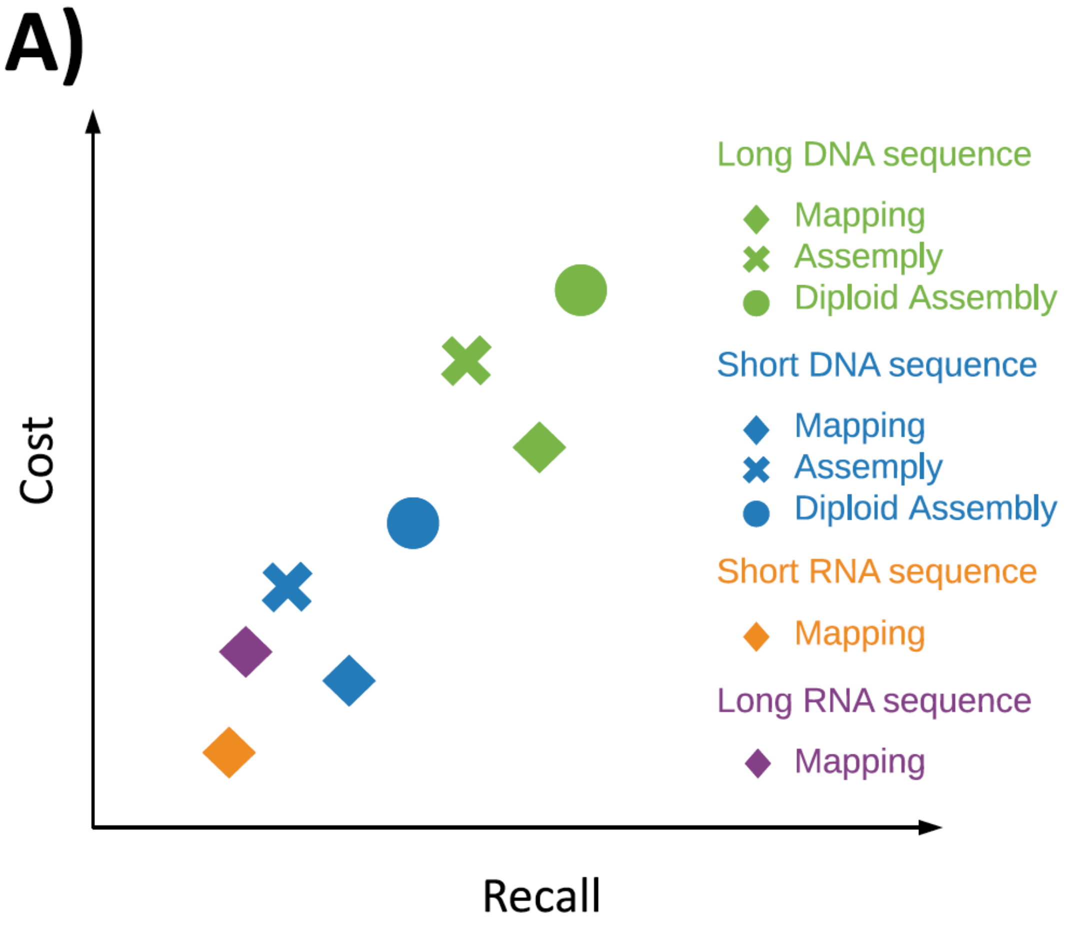
Mahmoud et al. 2019

Long-reads are better than short reads for identifying structural rearrangements



Mahmoud et al. 2019

Assembly-based methods are more sensitive than mapping-based methods



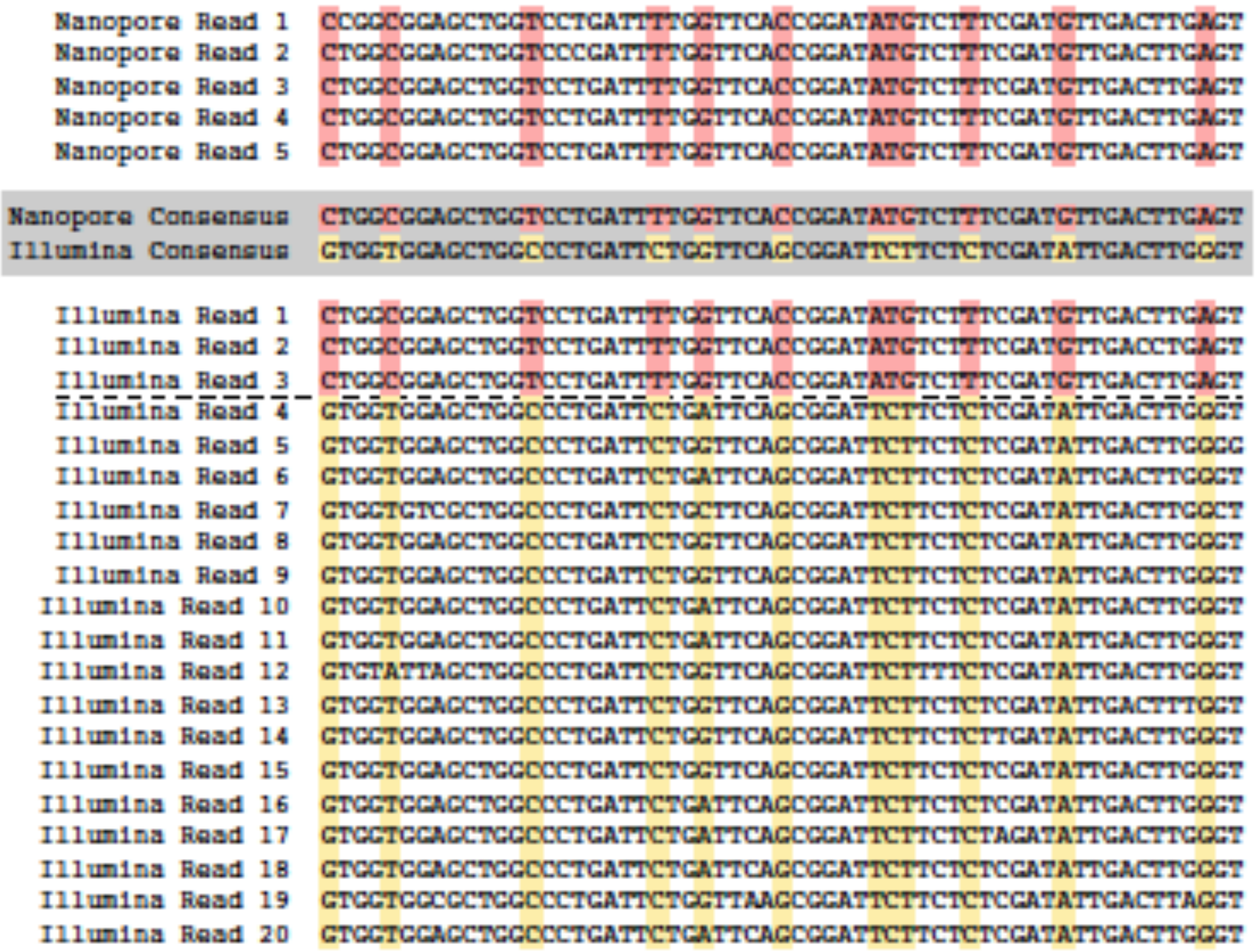
Mahmoud et al. 2019

Three general approaches for structural variant calling

Structural variant callers

Mahmoud et al. 2019

Reconstructing tandem duplications – problems with consensus



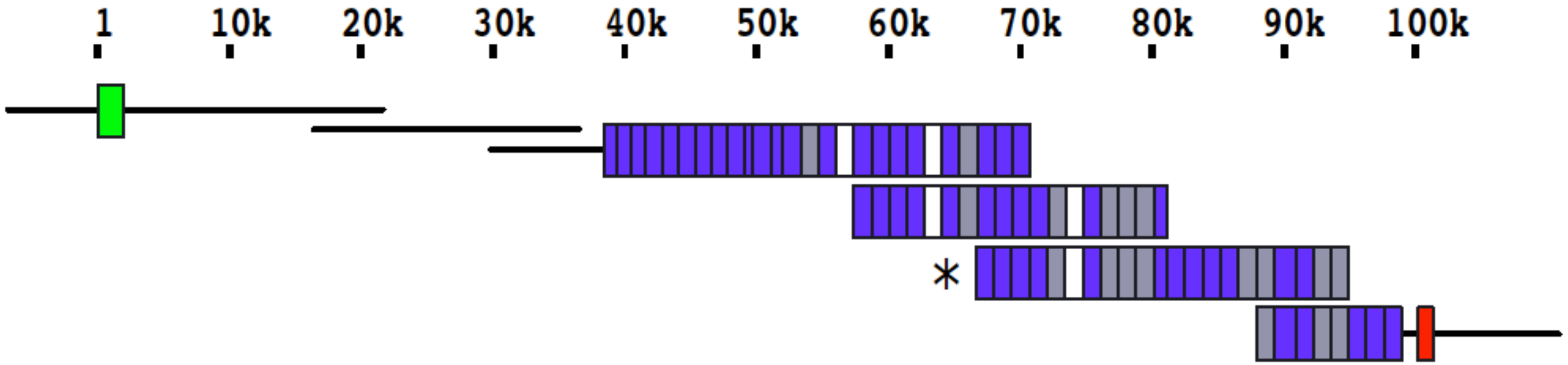
```
   Nanopore Read 1   CCGGCGGAGCTGGTCCTGATTTTGGTTCACCGGATATGTCTTTCGATGTTGACTTGAGT
   Nanopore Read 2   CTGGCGGAGCTGGTCCCGATTTTGGTTCACCGGATATGTCTTTCGATGTTGACTTGAGT
   Nanopore Read 3   CTGGCGGAGCTGGTCCTGATTTTGGTTCACCGGATATGTCTTTCGATGTTGACTTGAGT
   Nanopore Read 4   CTGGCGGAGCTGGTCCTGATTTTGGTTCACCGGATATGTCTTTCGATGTTGACTTGAGT
   Nanopore Read 5   CTGGCGGAGCTGGTCCTGATTTTGGTTCACCGGATATGTCTTTCGATGTTGACTTGAGT

Nanopore Consensus   CTGGCGGAGCTGGTCCTGATTTTGGTTCACCGGATATGTCTTTCGATGTTGACTTGAGT
Illumina Consensus   GTGGTGGAGCTGGCCCTGATTCTGGTTCAGCGGATTCTTCTCTCGATATTGACTTGGGT

   Illumina Read 1   CTGGCGGAGCTGGTCCTGATTTTGGTTCACCGGATATGTCTTTCGATGTTGACTTGAGT
   Illumina Read 2   CTGGCGGAGCTGGTCCTGATTTTGGTTCACCGGATATGTCTTTCGATGTTGACCTGAGT
   Illumina Read 3   CTGGCGGAGCTGGTCCTGATTTTGGTTCACCGGATATGTCTTTCGATGTTGACTTGAGT
------------------   - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
   Illumina Read 4   GTGGTGGAGCTGGCCCTGATTCTGATTCAGCGGATTCTTCTCTCGATATTGACTTGGGT
   Illumina Read 5   GTGGTGGAGCTGGCCCTGATTCTGGTTCAGCGGATTCTTCTCTCGATATTGACTTGGGG
   Illumina Read 6   GTGGTGGAGCTGGCCCTGATTCTGATTCAGCGGATTCTTCTCTCGATATTGACTTGGGT
   Illumina Read 7   GTGGTGTCGCTGGCCCTGATTCTGCTTCAGCGGATTCTTCTCTCGATATTGACTTGGCT
   Illumina Read 8   GTGGTGGAGCTGGCCCTGATTCTGGTTCAGCGGATTCTTCTCTCGATATTGACTTGGGT
   Illumina Read 9   GTGGTGGAGCTGGCCCTGATTCTGGTTCAGCGGATTCTTCTCTCGATATTGACTTGGGT
  Illumina Read 10   GTGGTGGAGCTGGCCCTGATTCTGATTCAGCGGATTCTTCTCTCGATATTGACTTGGGT
  Illumina Read 11   GTGGTGGAGCTGGCCCTGATTCTGATTCAGCGGATTCTTCTCTCGATATTGACTTGGGT
  Illumina Read 12   GTGTATTAGCTGGCCCTGATTCTGGTTCAGCGGATTCTTTTCTCGATATTGACTTGGGT
  Illumina Read 13   GTGGTGGAGCTGGCCCTGATTCTGGTTCAGCGGATTCTTCTCTCGATATTGACTTTGGT
  Illumina Read 14   GTGGTGGAGCTGGCCCTGATTCTGGTTCAGCGGATTCTTCTCTTGATATTGACTTGGGT
  Illumina Read 15   GTGGTGGAGCTGGCCCTGATTCTGGTTCAGCGGATTCTTCTCTCGATATTGACTTGGGT
  Illumina Read 16   GTGGTGGAGCTGGCCCTGATTCTGATTCAGCGGATTCTTCTCTCGATATTGACTTGGGT
  Illumina Read 17   GTGGTGGAGCTGGCCCTGATTCTGATTCAGCGGATTCTTCTCTAGATATTGACTTGGGT
  Illumina Read 18   GTGGTGGAGCTGGCCCTGATTCTGATTCAGCGGATTCTTCTCTCGATATTGACTTGGGT
  Illumina Read 19   GTGGTGGCGCTGGCCCTGATTCTGGTTAAGCGGATTCTTCTCTCGATATTGACTTAGGT
  Illumina Read 20   GTGGTGGAGCTGGCCCTGATTCTGATTCAGCGGATTCTTCTCTCGATATTGACTTGGGT
```

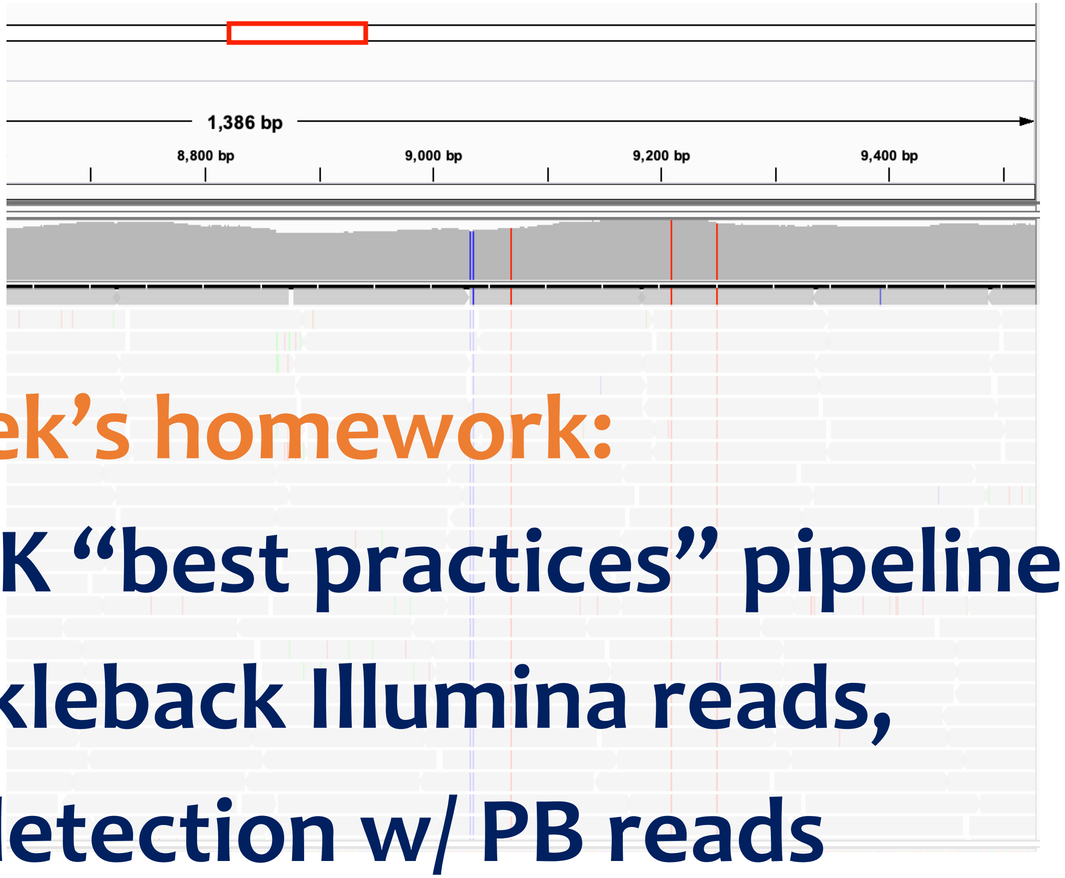
Stellwagen & Burns 2021

Reconstructing tandem duplications – problems with consensus



Stellwagen & Burns 2021

Next up:
Performing variant calling with GATK



Next week's homework:

1. GATK "best practices" pipeline on Stickleback Illumina reads,
2. SV detection w/ PB reads