Variant Calling
Resequencing-Based Genome Inference

Erik Garrison
University of Tennessee Health Science Center

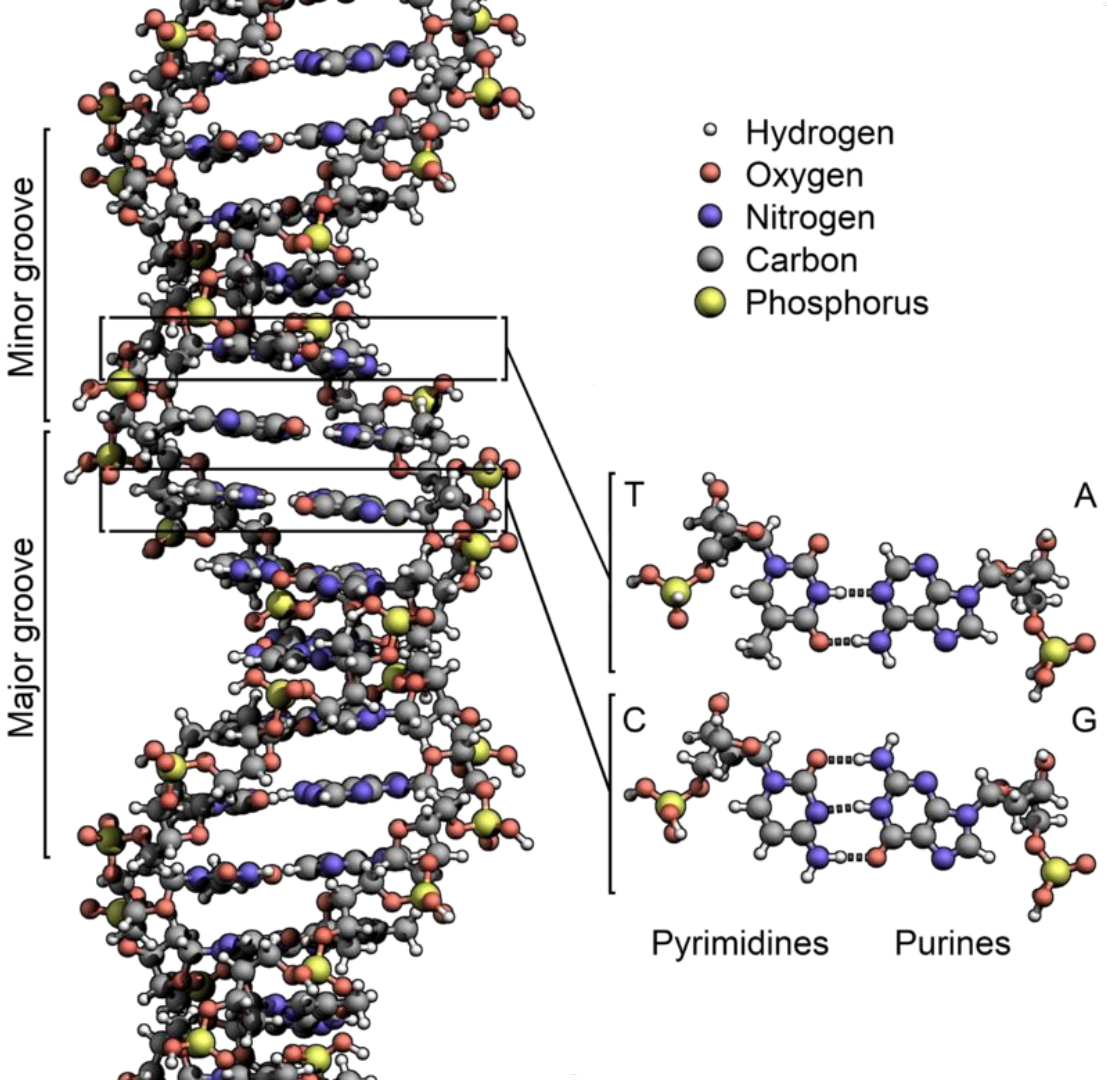
Workshop on Genomics - Český Krumlov
January 12, 2024

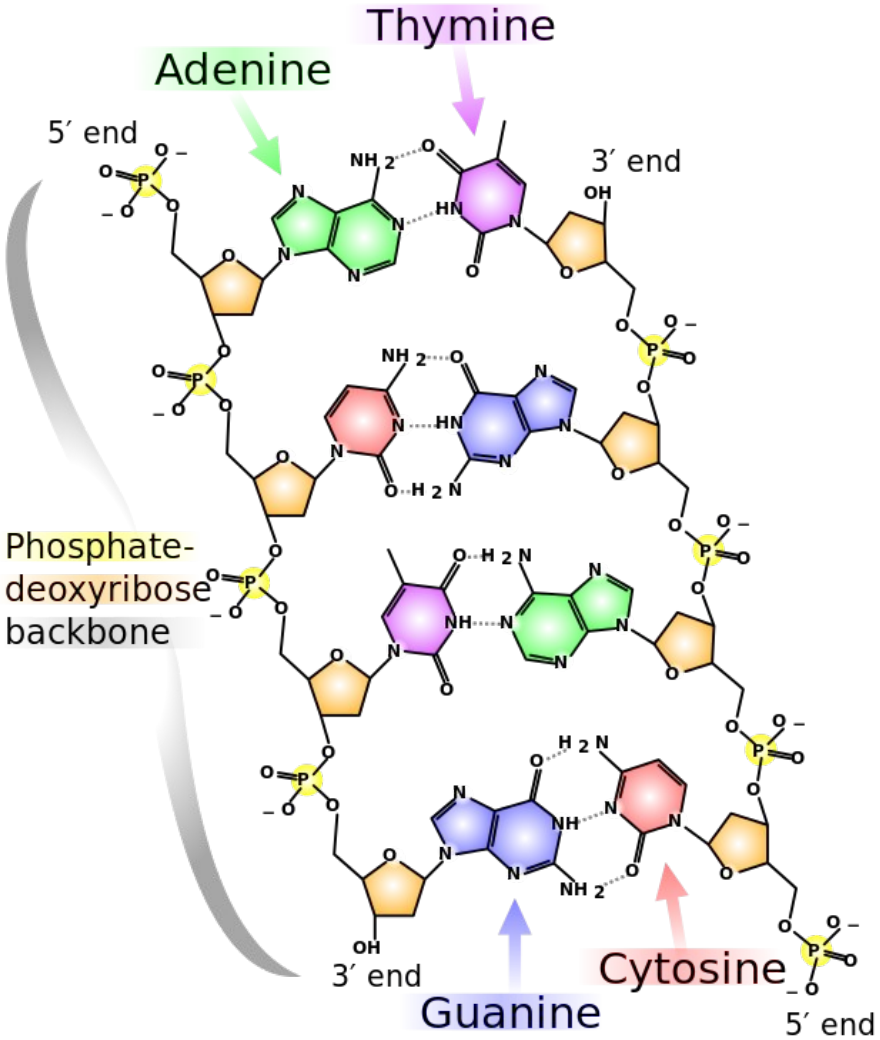
Overview

1. DNA variation and sequencing
2. Alignment to linear sequences
3. Error detection and genotyping
4. Learning to genotype
5. Practicals

(small) Genomic variation

DNA


Minor groove
Major groove
Hydrogen
Oxygen
Nitrogen
Carbon
Phosphorus
T
A
C
G
Pyrimidines
Purines



DNA

“twisted” view of ladder

A SNP

A point mutation in which one base is swapped for another.

AATTAG*C*CATTA

AATTAG*T*CATTA

An INDEL

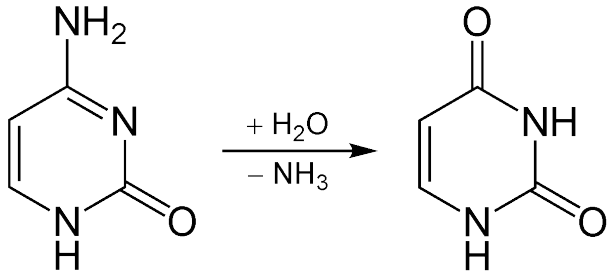
A mutation that results from the gain or loss of sequence.

AATTAGCCATTA

AATTA--CATTA

(some) causes of SNPs

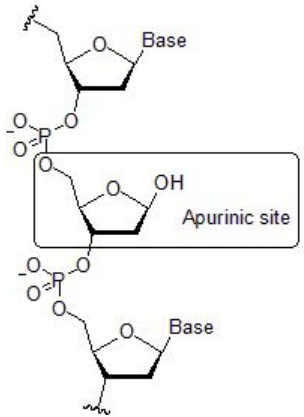
- Deamination
  - cytosine → uracil
  - 5-methylcytosine → thymine
  - guanine → xanthine (mispairs to A-T bp)
  - adenine → hypoxanthine (mispairs to G-C bp)

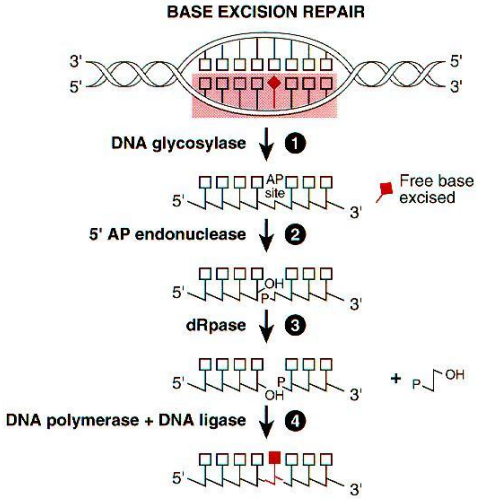


Deamination of Cytosine to Uracil
http://en.wikipedia.org/wiki/Deamination

(some) causes of SNPs

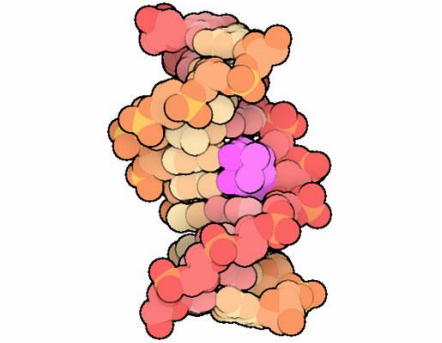
- Depurination
  - purines are cleaved from DNA sugar backbone (5000/cell/day, pyrimidines at much lower rate)
  - Base excision repair (BEP) can fail → mutation

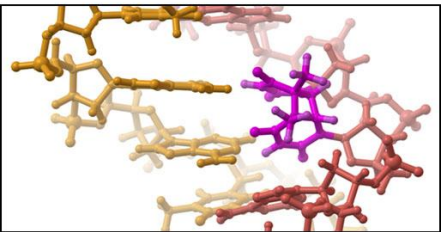


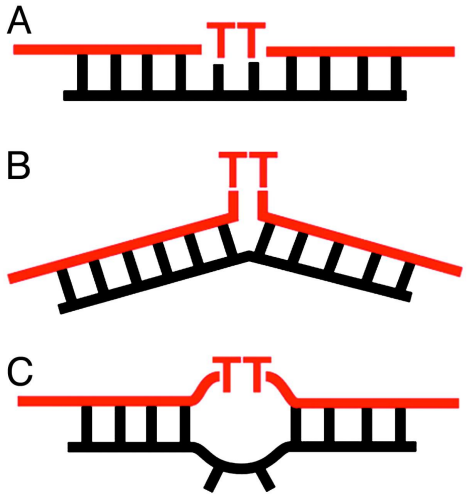


Multi-base events (MNPs)

- MNPs
  - thymine dimerization (UV induced)
  - other (e.g. oxidative stress induced)



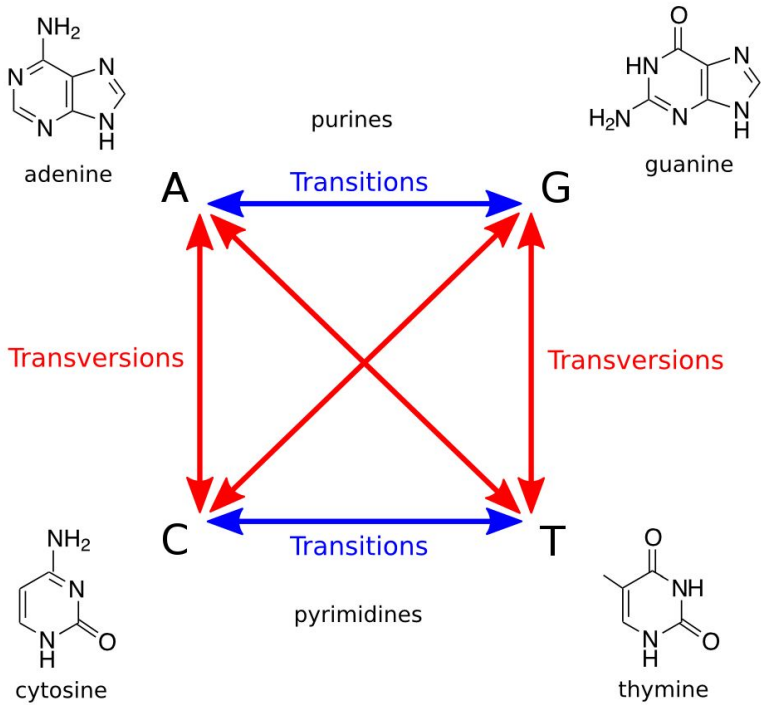




http://www.rcsb.org/pdb/101/motm.do?momID=91

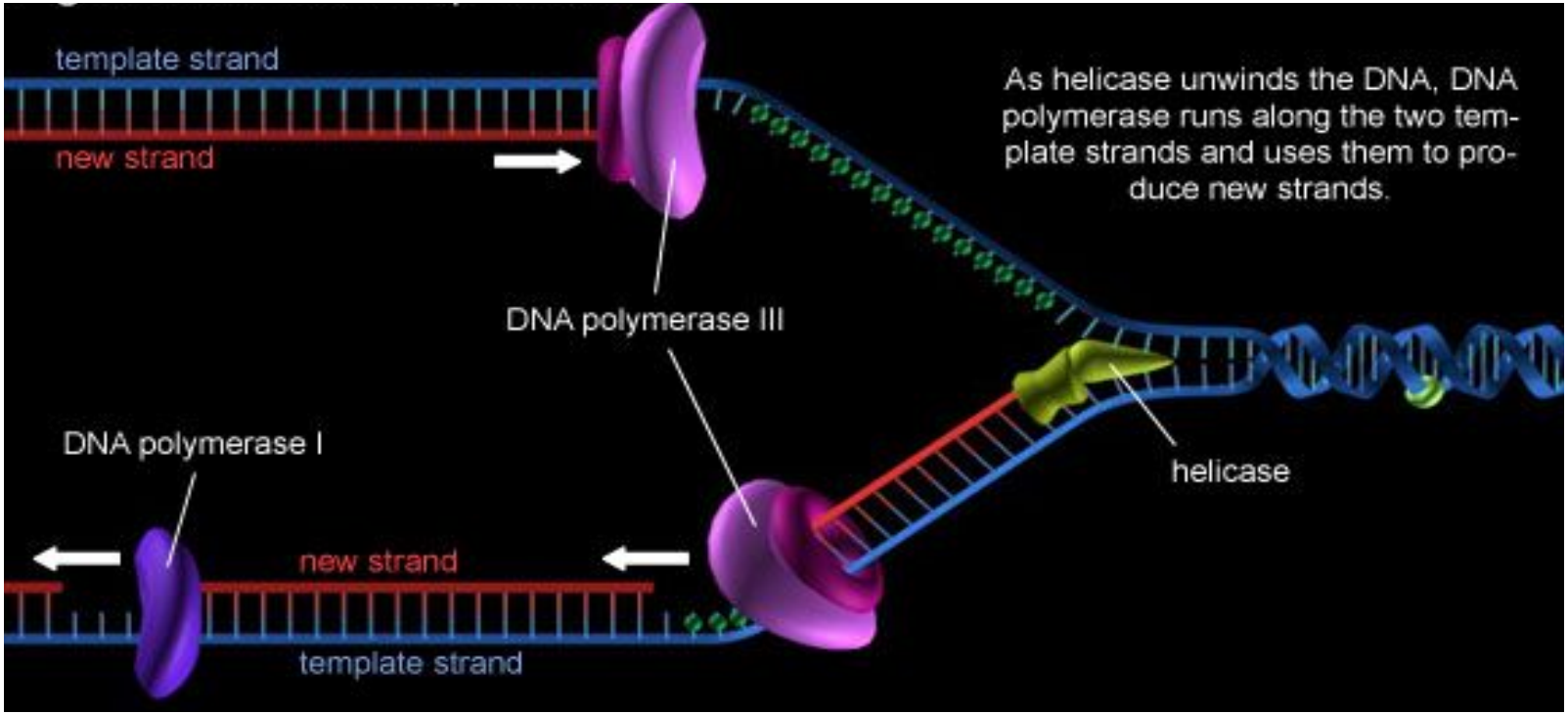
Transitions and transversions

In general **transitions are 2-3 times more common than transversions**. (But this depends on biological context.)



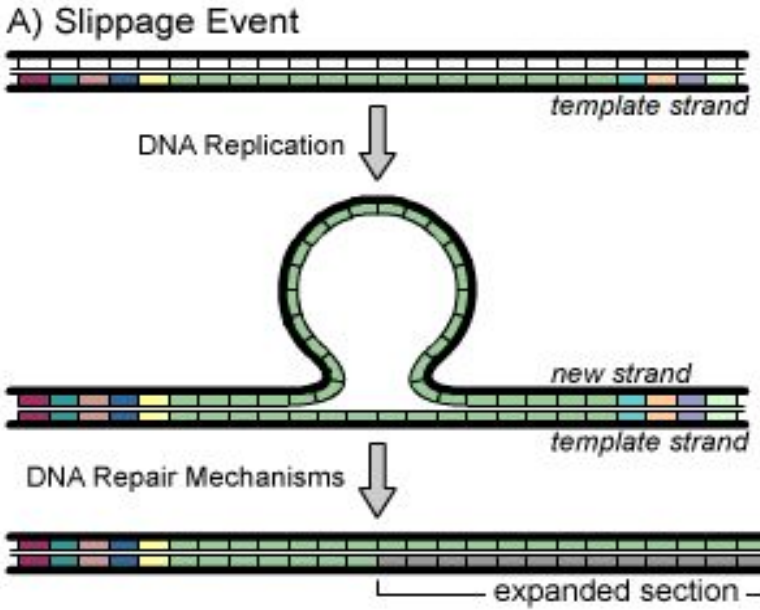
https://upload.wikimedia.org/wikipedia/commons/3/35/Transitions-transversions-v3.png

DNA replication



http://www.stanford.edu/group/hopes/cgi-bin/wordpress/2011/02/all-about-mutations/

Polymerase *slippage*

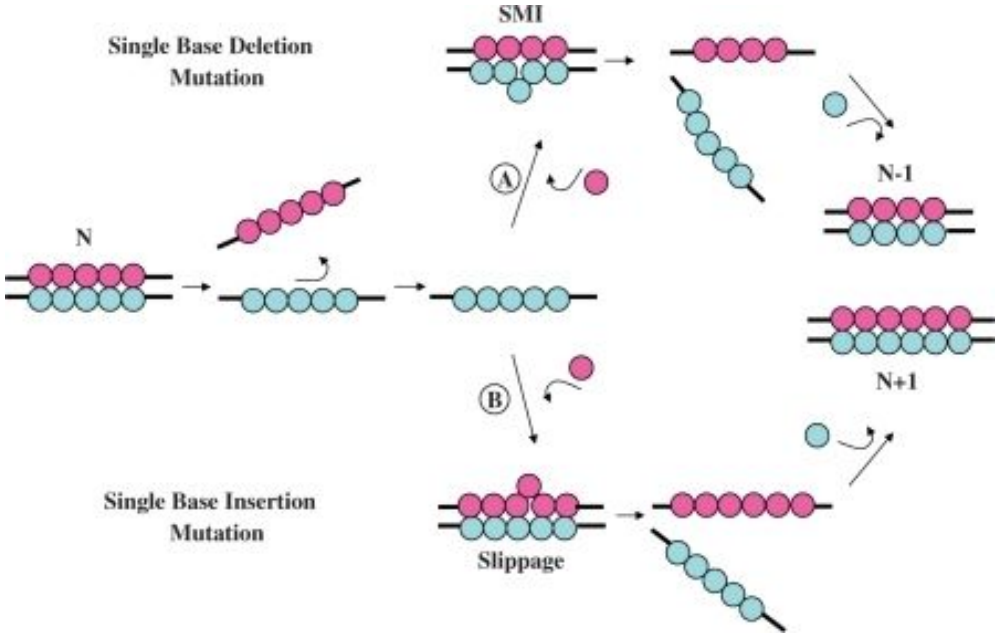


(A) During replication, polymerase slippage and subsequent reattachment may cause a bubble to form in the new strand. Slippage is thought to occur in sections of DNA with repeated patterns of bases (such as CAG), represented here by matching colors. Then, DNA repair mechanisms realign the template strand with the new strand and the bubble is straightened out. The resulting double helix is thus expanded.

(B) Polymerase slippage, as theorized, cannot occur in DNA without repeating patterns of bases.

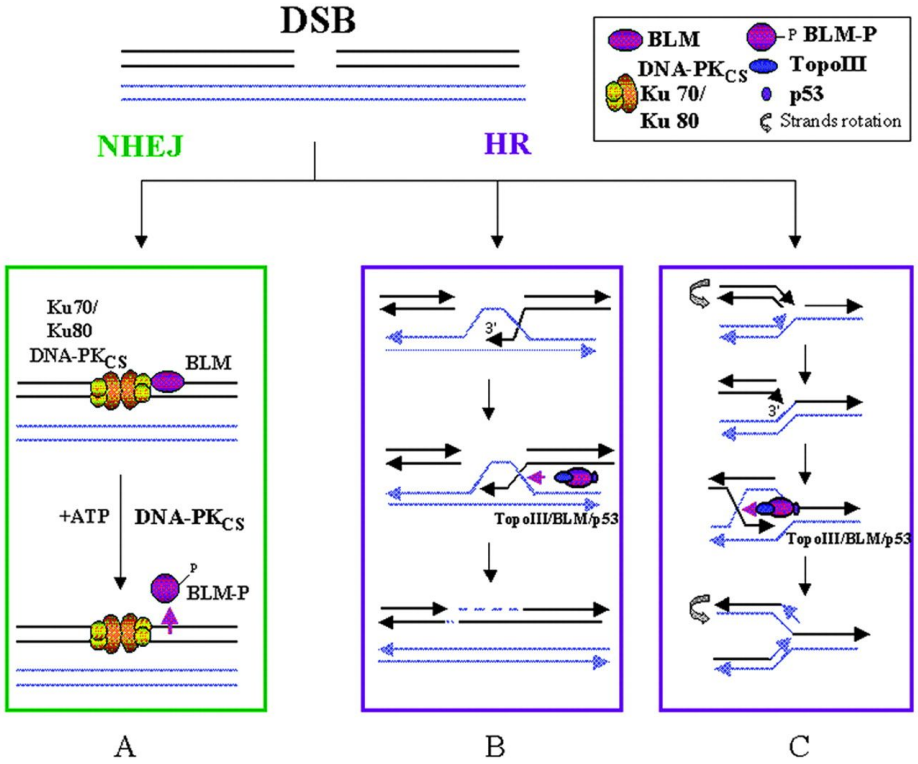
http://www.stanford.edu/group/hopes/cgi-bin/wordpress/2011/02/all-about-mutations/

Insertions and deletions via slippage



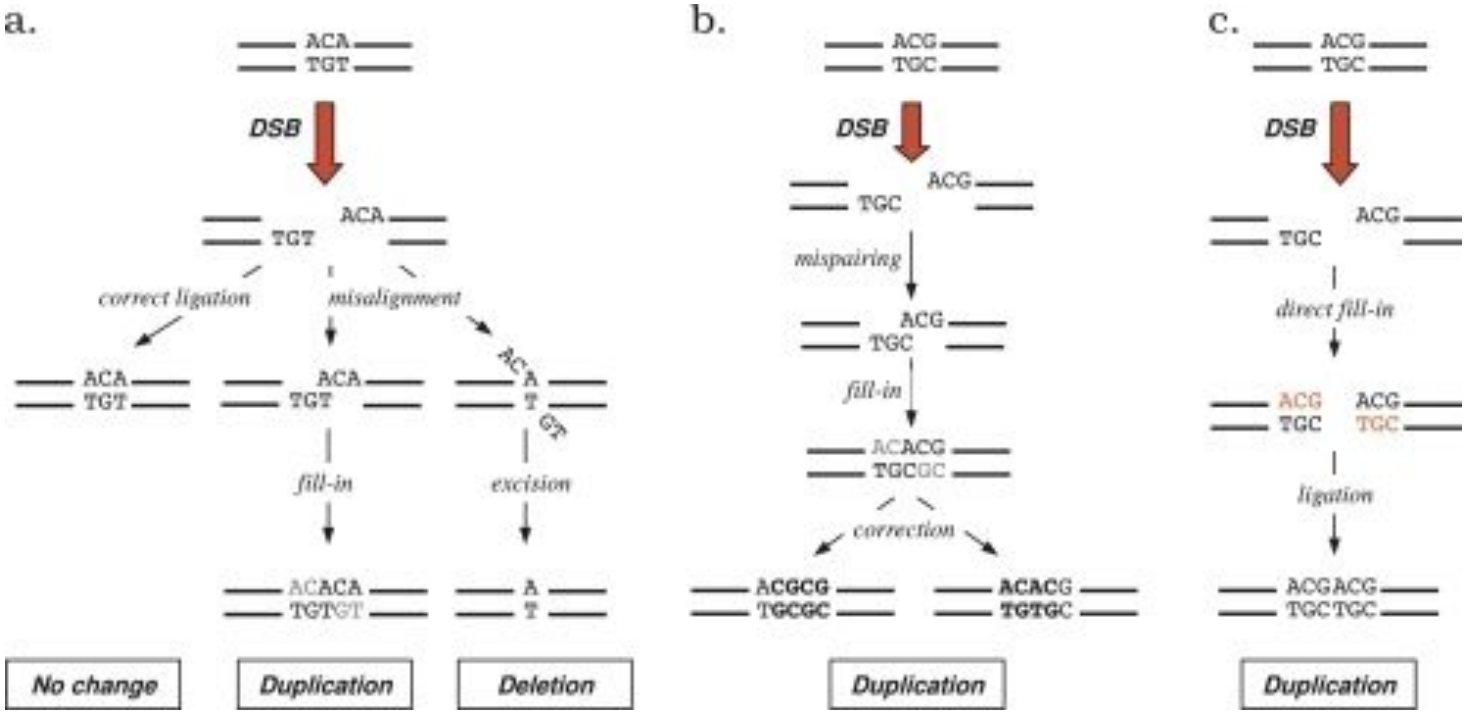
Energetic signatures of single base bulges: thermodynamic consequences and biological implications. Minetti CA, Remeta DP, Dickstein R, Breslauer KJ - Nucleic Acids Res. (2009)

Double-stranded break repair



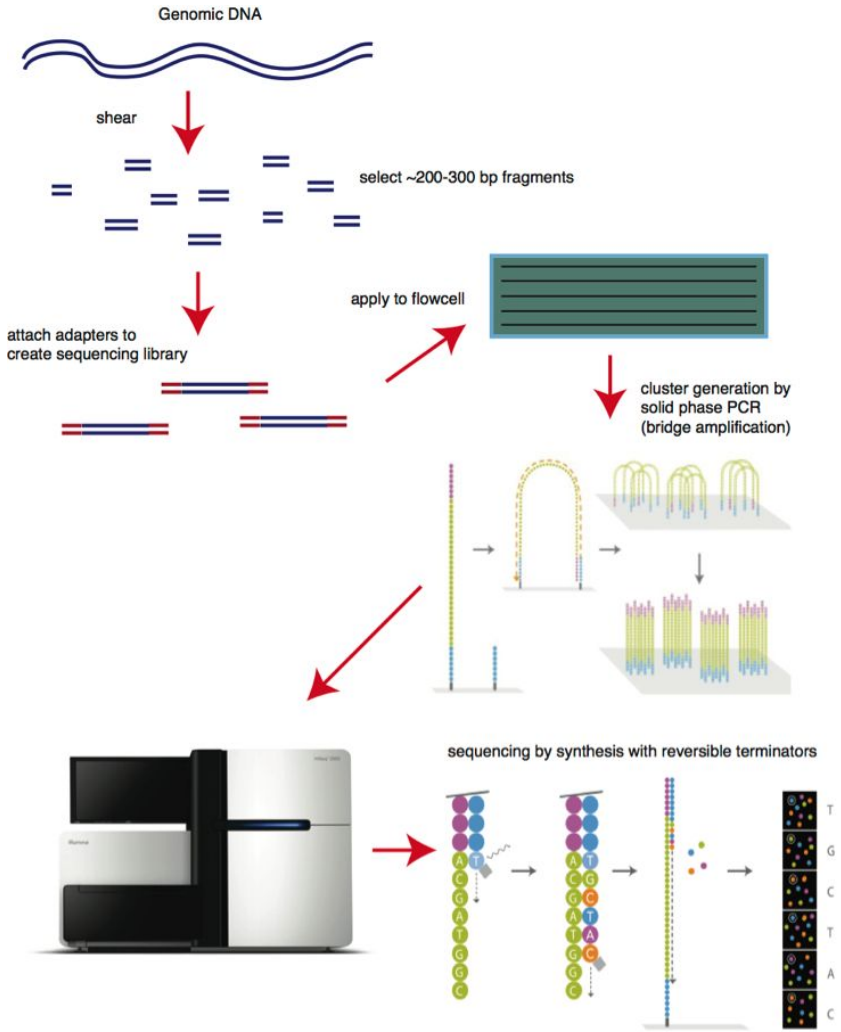
Possible anti-recombinogenic role of Bloom's syndrome helicase in double-strand break processing. doi: [10.1093/nar/gkg834](https://doi.org/10.1093/nar/gkg834)

NHEJ-derived indels



DNA Slippage Occurs at Microsatellite Loci without Minimal Threshold Length in Humans: A Comparative Genomic Approach. Leclercq S, Rivals E, Jarne P - Genome Biol Evol (2010)

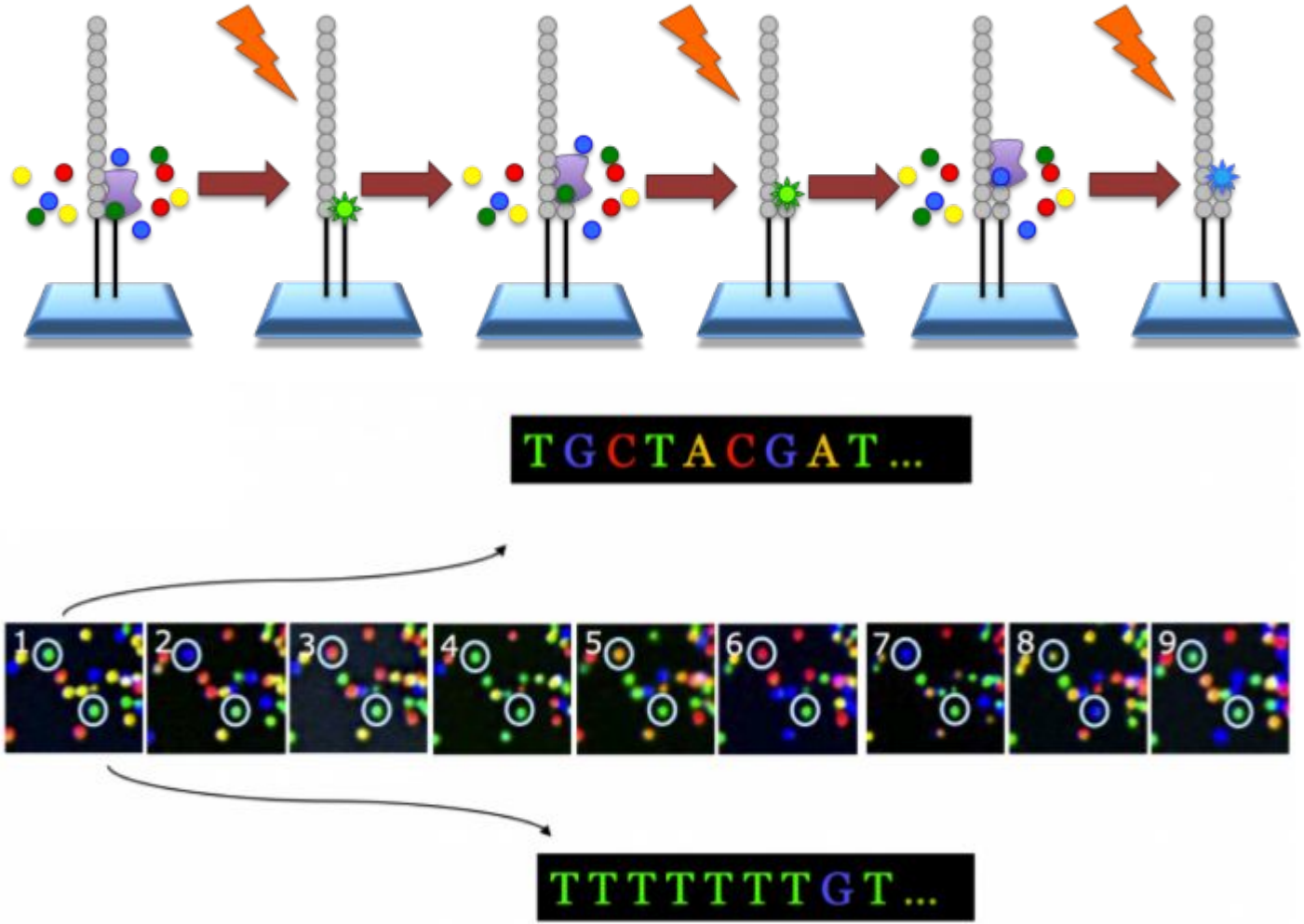
Genome sequencing recap



Sequencing by synthesis (Illumina)

n.b. Not exactly how this works on newer systems…

http://bitesizebio.com/13546/sequencing-by-synthesis-explaining-the-illumina-sequencing-technology/



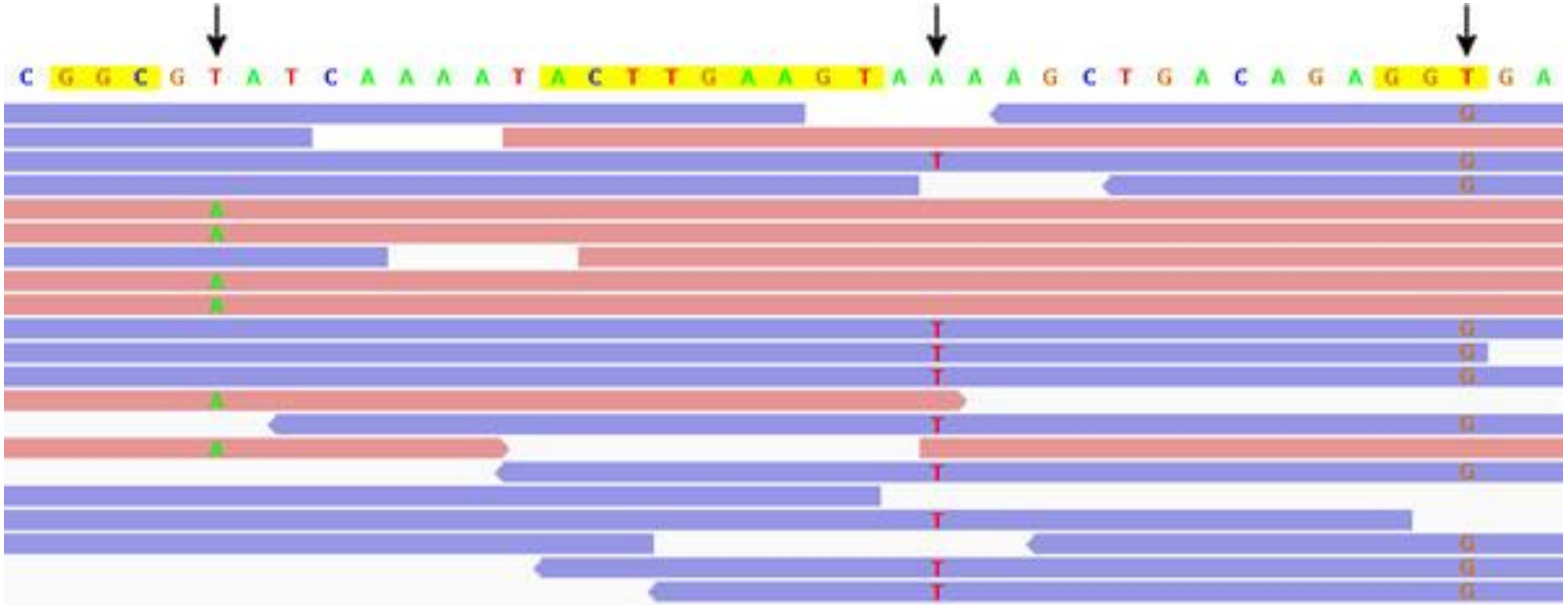
http://www.ebi.ac.uk/training/online/course/ebi-next-generation-sequencing-practical-course/what-next-generation-dna-sequencing/illumina-

What can go wrong?

1. Input artifacts, problems with library prep
    a. replication in PCR has no error-correction (→ SNPs)
    b. no quaternary structures (e.g. clamp) to prevent slippage (→ indels)
    c. chimeras…
    d. duplicates (worse if they are errors)
2. Sequencing-by-synthesis
    a. phasing of step
        i. synthesis reaction efficiency is not 100%
        ii. particularly bad in A/T homopolymers
    b. certain *context specific errors*
        i. vary by sequencing protocol, device
        ii. often strand-specific

Example: Context specific errors

Show up as strand-specific errors:



http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3622629/

Context specific errors (motifs)

Rank	Context	FER [%]	RER [%]	ERD [%]
1	ACGGCGGT	26.1	0.5	25.6
2	GTGGCGGT	25.1	0.7	24.4
3	GCGGCGGT	22.9	0.7	22.2
4	GTGGCTGT	22.4	0.6	21.8
5	ATGGCGGT	21.2	1.0	20.3
6	NCGGCGGT	20.0	0.7	19.3
7	GTGGCTTG	20.2	1.2	19.0
8	GNGGCGGT	19.2	0.7	18.5
9	GCGGCTGT	18.8	0.7	18.1
10	ACGGCTGT	18.6	0.8	17.7

← forward and reverse error rates for the ten most-common CSEs on a variety of illumina systems (in 2013)

Often GC-rich!

Changes in chemistry mean that this may not be such a big deal now, but this example is something to keep in mind!

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3622629/

Long read technologies

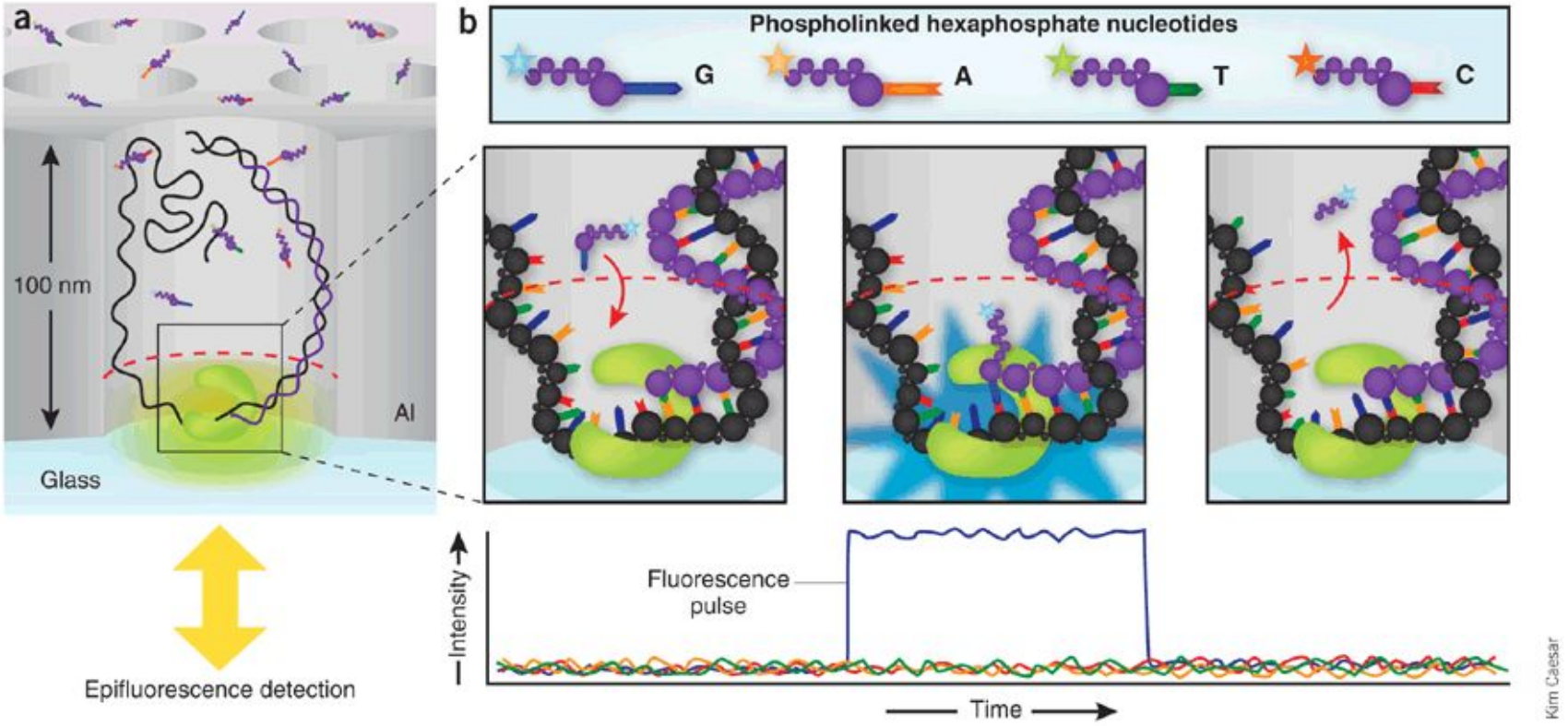
“3rd-gen” sequencing.

- Read single molecules (long too!)
- Have high error rates (10-15%)

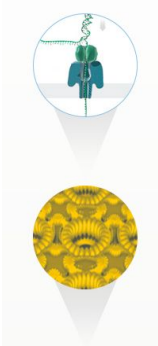
Pacific Biosciences

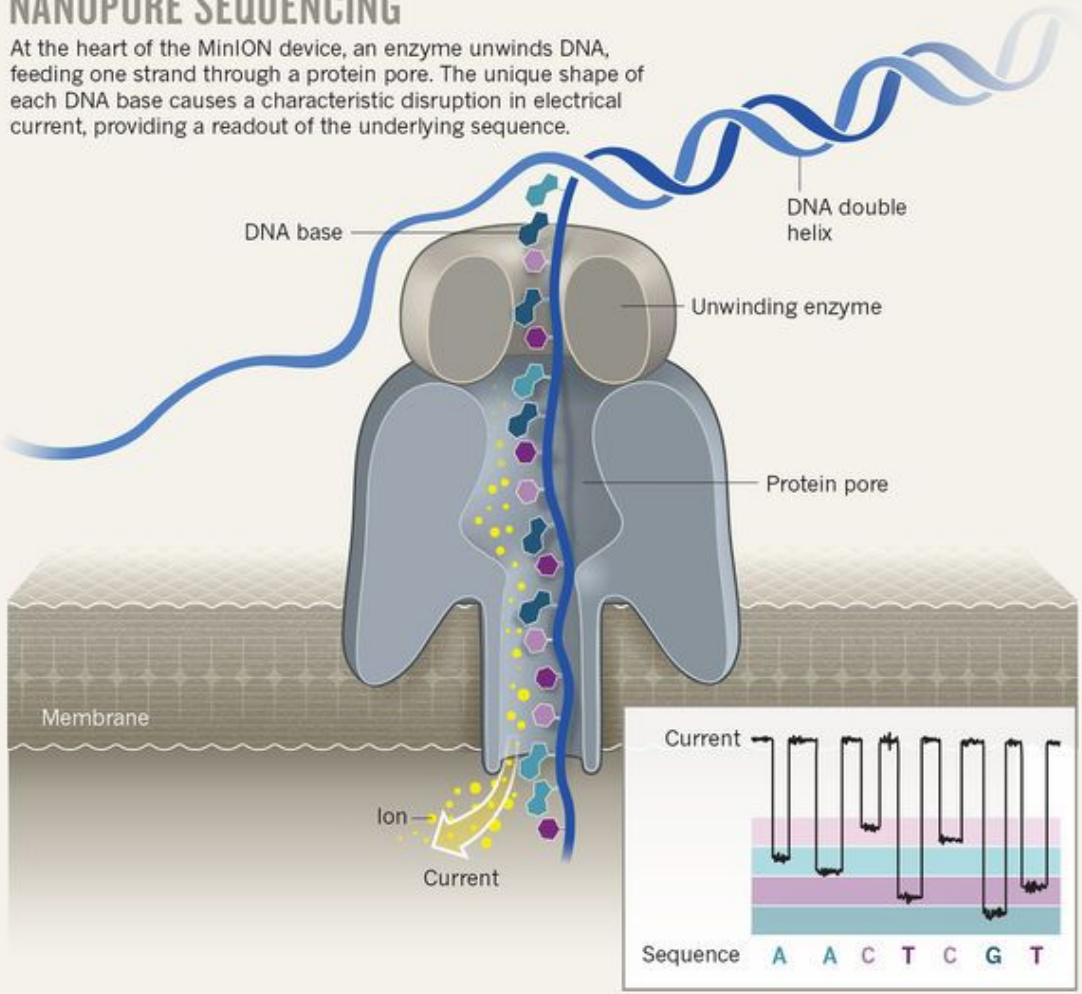
Oxford Nanopore

Pacific Biosciences sequencing



Oxford Nanopore sequencing





Alignment is interpretation

Alignment

Covered in our last session!

Key idea for variant calling:

alignment provides a kind of interpretation of variation.

Changes in parameters, alignments, or sequence context can lead to changes in called alleles.

Seeing variation

These sequences have mutations between them.

CAAATAAGGAAATTTTCTGGAGTTCTATTATA
CAAATAAGGTTTGCTATCTAGGTTATTATA

They are homologous but it’s not easy to see.

Pairwise alignment

One solution, assuming a particular set of alignment parameters, has 3 indels and a SNP:

```
CAAATAAGGAAATTT----TCTGGAGTTCTATTATA
CAAATAAGG---TTTGCTATCT--AGGT-TATTATA
```

But if we use a higher gap-open penalty, things look different:

```
CAAATAAGGAAATTT--TCTGGAGTTCTATTATA
CAAATAAGG---TTTGCTATCTAGGT-TATTATA
```

Alignment = interpretation

Different parameterizations can yield different results.

Different results suggest “different” variation.

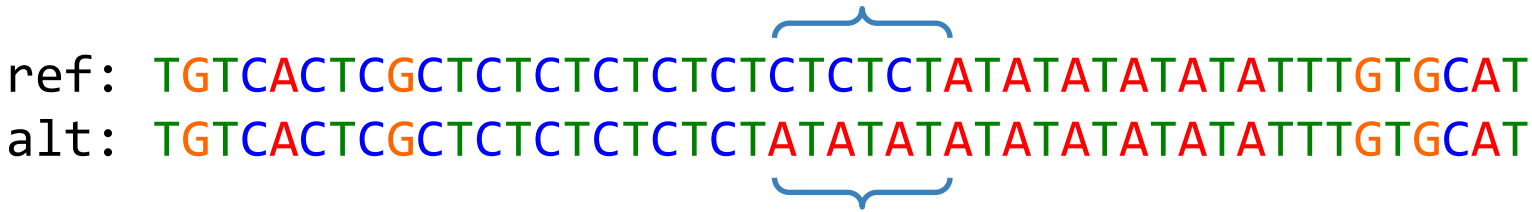
What kind of problems can this cause? (And how can we mitigate these issues?)

INDELs have multiple representations and require normalization for standard calling

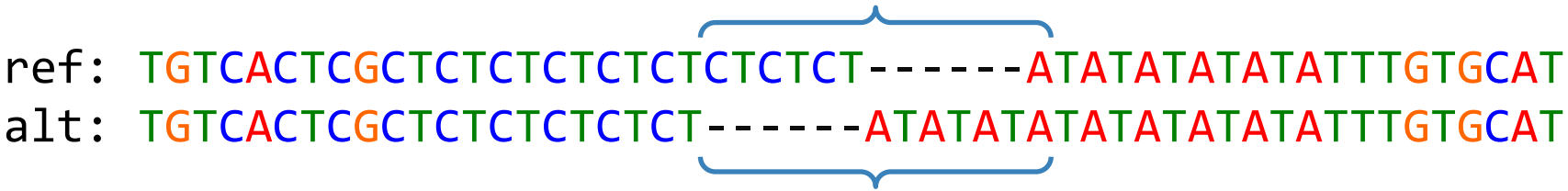
Left alignment allows us to ensure that our representation is consistent across alignments and also variant calls.

https://www.biostars.org/p/66843/ user sa9

example: 1000G PhaseI low coverage
chr15:81551110, ref:CTCTC alt:ATATA



Interpreted as 3 SNPs

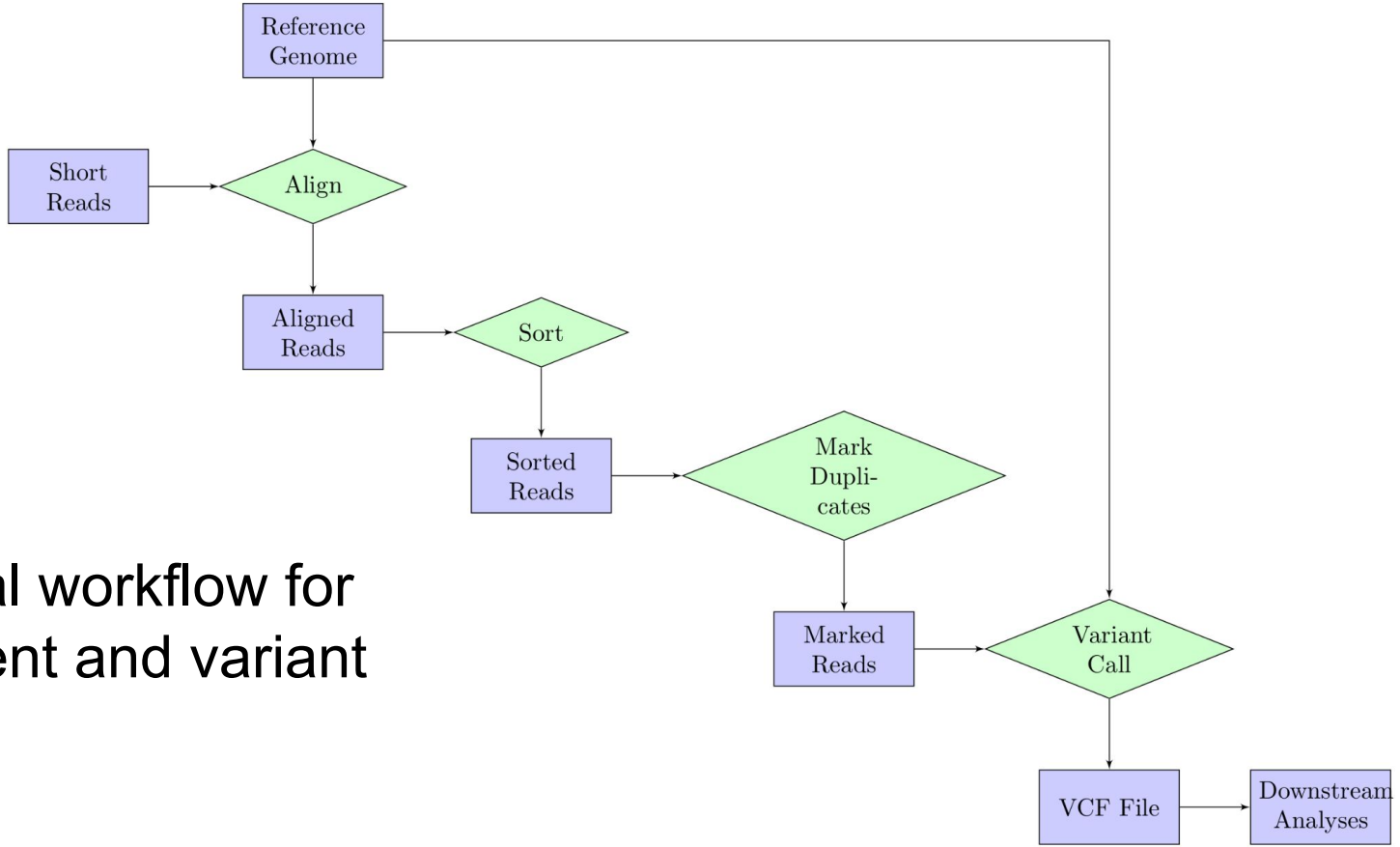


Interpreted as microsatellite expansion/contraction

example: 1000G Phase1 low coverage

chr20:708257, ref:AGC alt:CGA

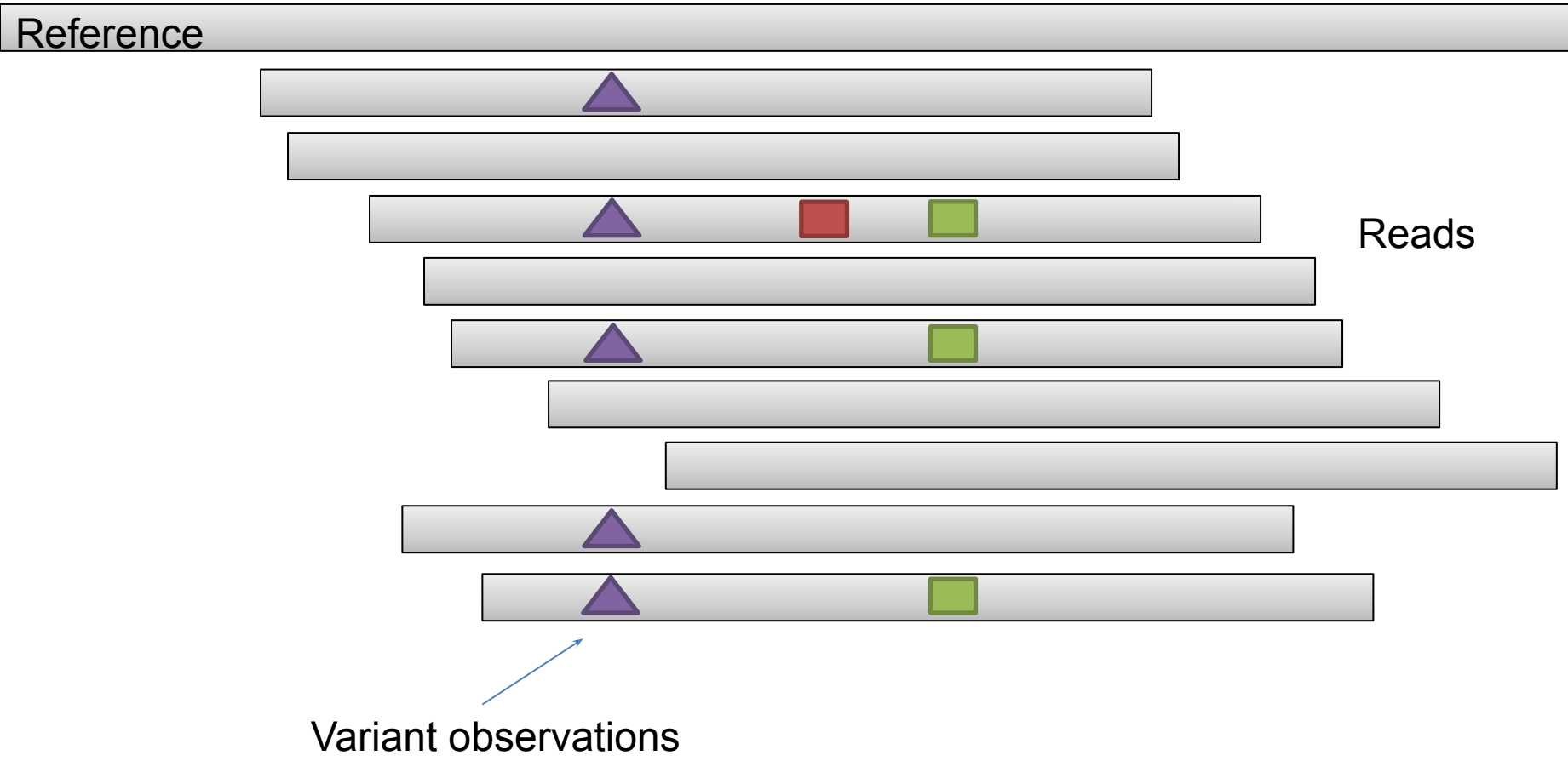
Processing alignments



A typical workflow for alignment and variant calling.

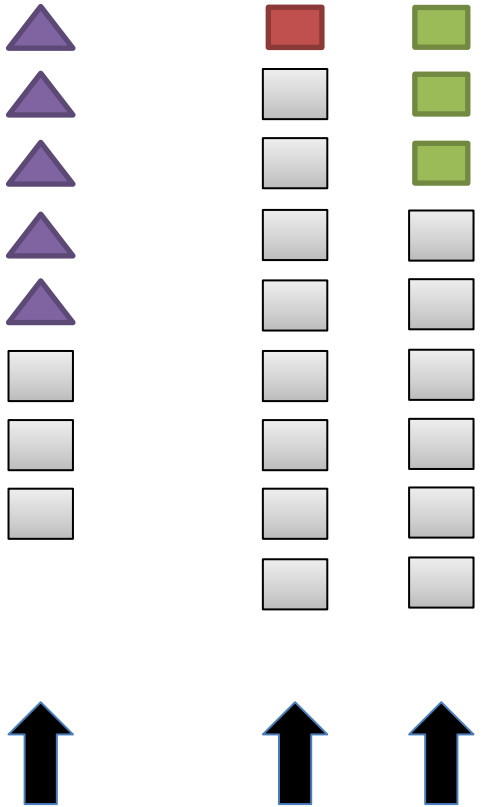
Variant (haplotype) detection

Alignments to candidates

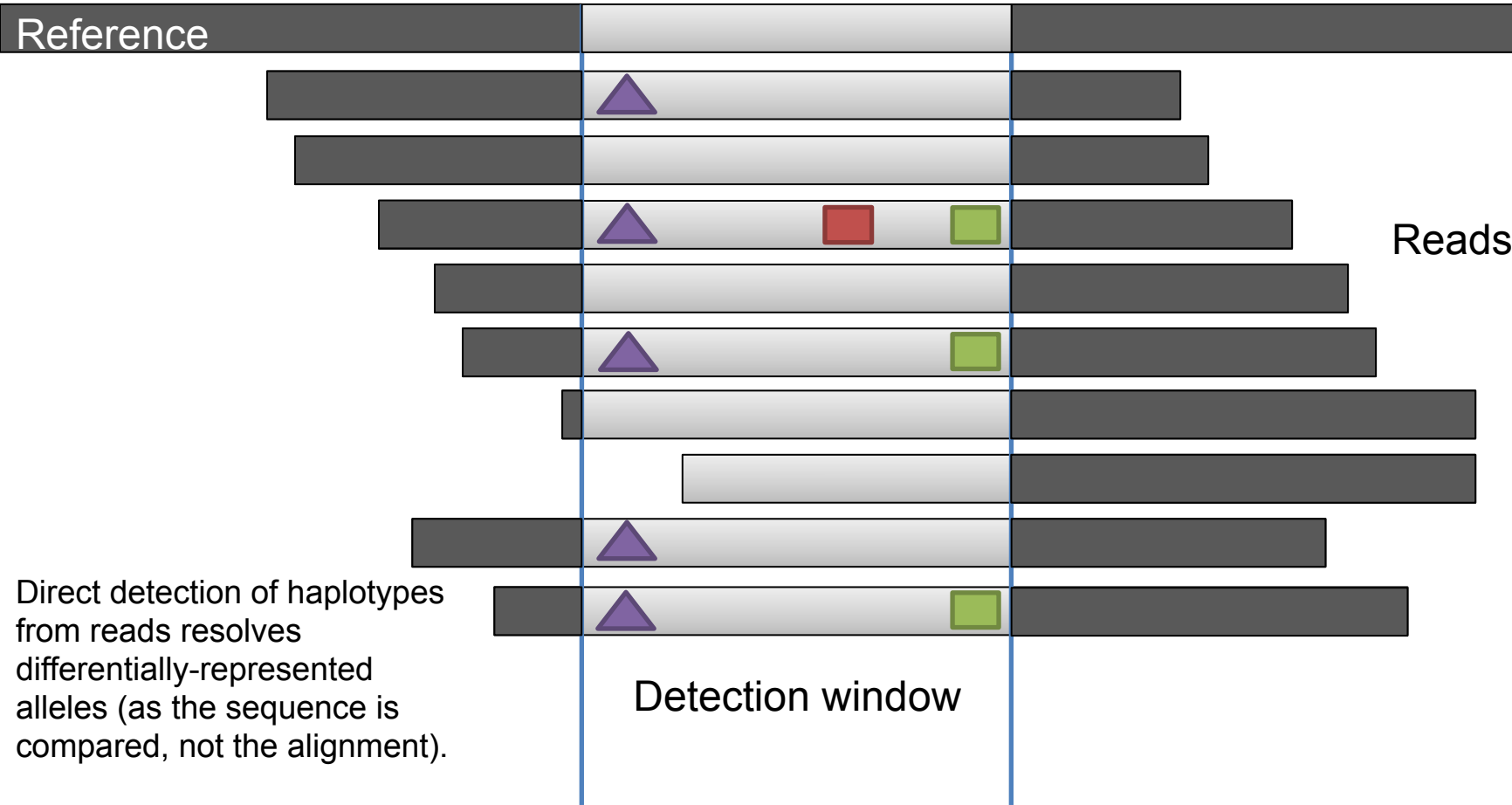


The data exposed to the caller

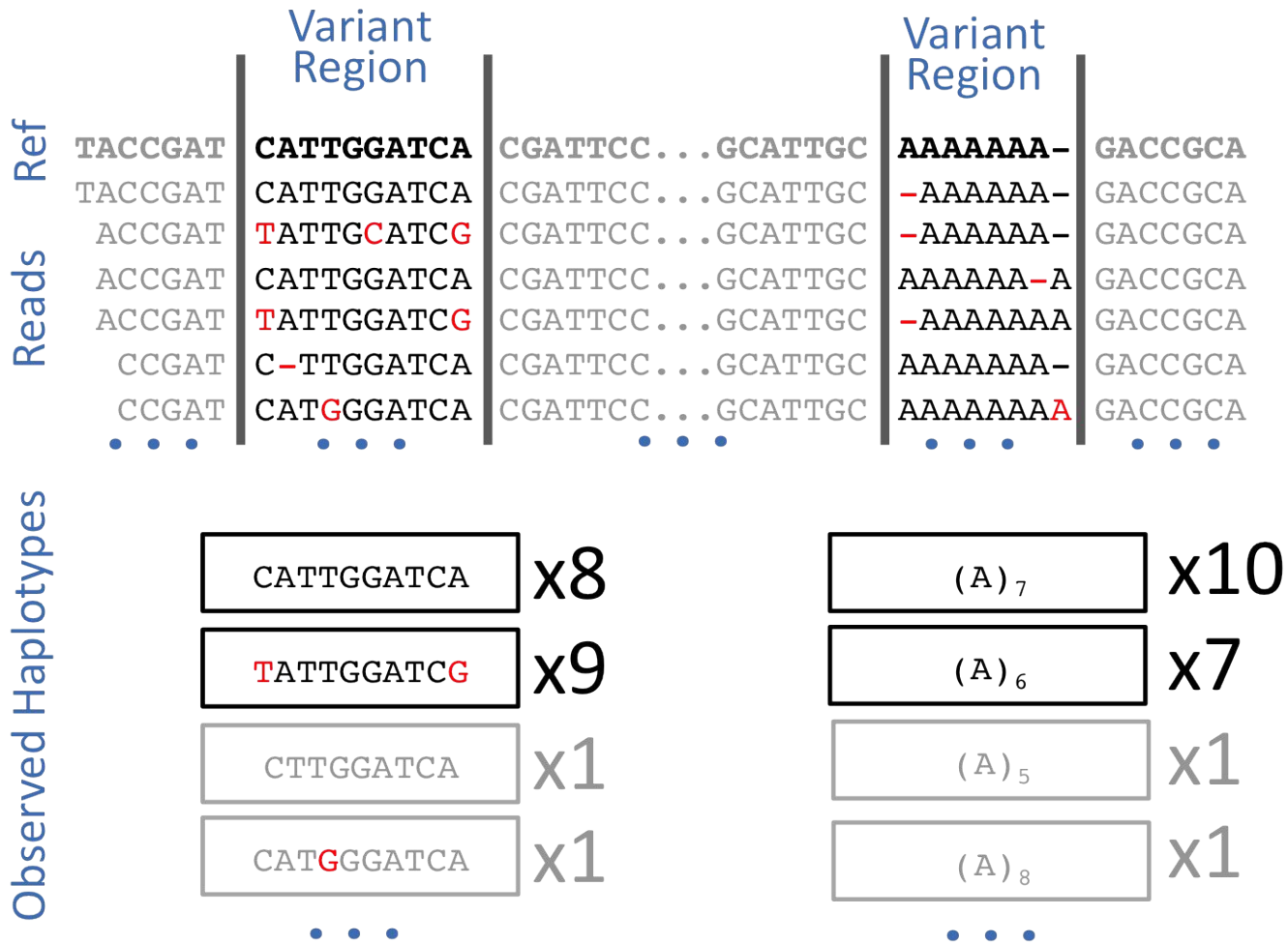
Reference



Direct detection of haplotypes



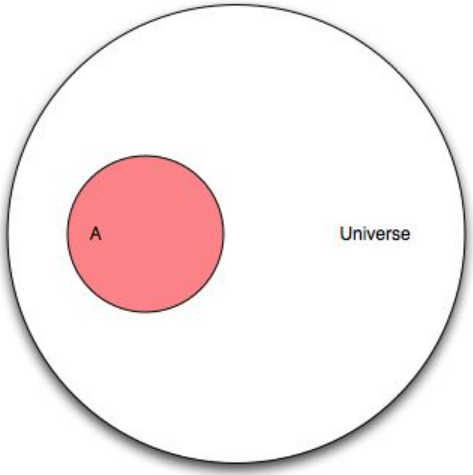
Direct detection of haplotypes from reads resolves differentially-represented alleles (as the sequence is compared, not the alignment).


Variant Region
Variant Region
Ref
Reads
TACCGAT CATTGGATCA CGATTCC...GCATTGC AAAAAAA- GACCGCA
TACCGAT CATTGGATCA CGATTCC...GCATTGC -AAAAAA- GACCGCA
ACCGAT TATTGCATCG CGATTCC...GCATTGC -AAAAAA- GACCGCA
ACCGAT CATTGGATCA CGATTCC...GCATTGC AAAAAA-A GACCGCA
ACCGAT TATTGGATCG CGATTCC...GCATTGC -AAAAAAA GACCGCA
CCGAT C-TTGGATCA CGATTCC...GCATTGC AAAAAAA- GACCGCA
CCGAT CATGGGATCA CGATTCC...GCATTGC AAAAAAAA GACCGCA
Observed Haplotypes
CATTGGATCA x8
TATTGGATCG x9
CTTGGATCA x1
CATGGGATCA x1
(A)7 x10
(A)6 x7
(A)5 x1
(A)8 x1

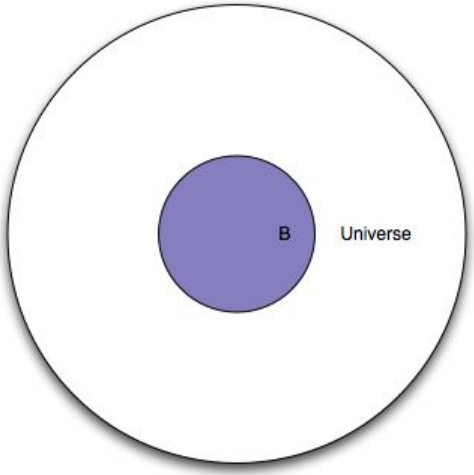
Genotyping and error detection

Bayesian (visual) intuition

We have a universe of individuals.



A = samples with a variant at some locus



B = putative observations of variant at some locus

Figures from http://oscarbonilla.com/2009/05/visualizing-bayes-theorem/

probability(A|B)

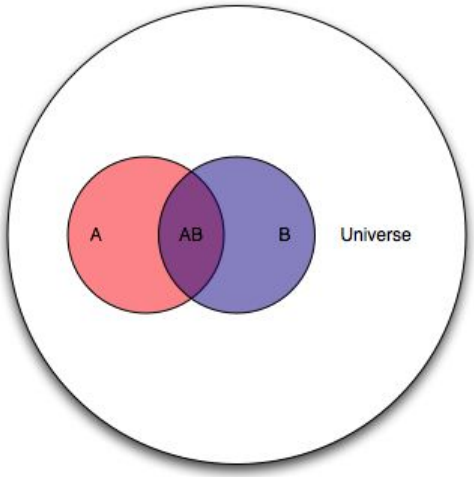
We want to estimate the probability that we have a real polymorphism "A" given "|" that we observed variants in our alignments "B".

$$P(A|B) = \frac{|AB|}{|B|}$$

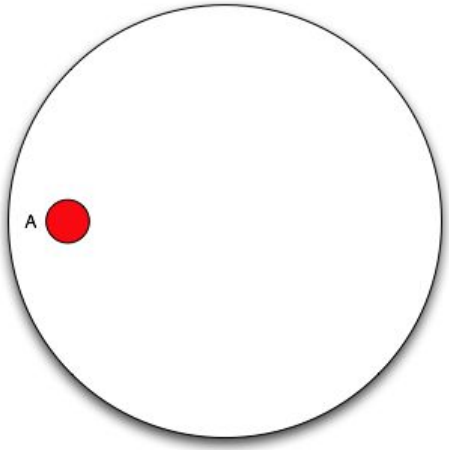
$$P(A|B) = \frac{P(AB)}{P(B)}$$

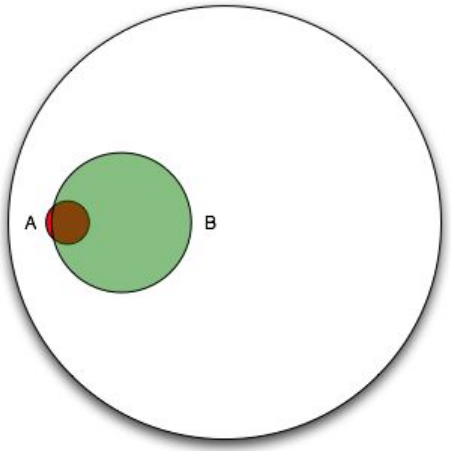
$$P(B|A) = \frac{P(AB)}{P(A)}$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$



In our case it's a bit more like this...





Observations (B) provide pretty good sensitivity, but poor specificity.

The model

- Bayesian model estimates the probability of polymorphism at a locus given input data and the population mutation rate (~pairwise heterozygosity) and assumption of “neutrality” (random mating).
- Following Bayes theorem, the probability of a specific set of genotypes over some number of samples is:
  - **P(G|R) = (P(R|G) P(G)) / P(R)**
- Which in FreeBayes we extend to:
  - **P(G,S|R) = (P(R|G,S) P(G)P(S)) / P(R)**
  - **G** = genotypes, **R** = reads, **S** = locus is well-characterized/mapped
  - **P(R|G,S)** is our data likelihood, **P(G)** is our prior estimate of the genotypes, **P(S)** is our prior estimate of the mappability of the locus, **P(R)** is a normalizer.

Handling non-biallelic/diploid cases

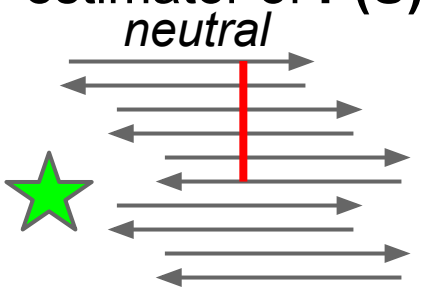
We compose our data likelihoods, **P(Reads|Genotype)** using a discrete multinomial sampling probability:

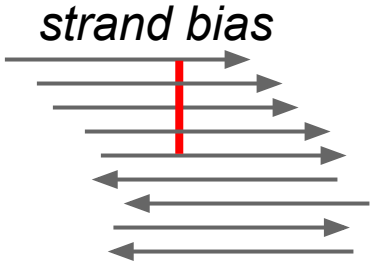
$$P(reads|genoytpe) = \binom{|reads|}{|reads = A|, |reads = B| \ldots} \times \prod_{\forall alleles \in genotype} freq(allele \in genotype) \times \prod_{\forall reads} P(correct(read))$$

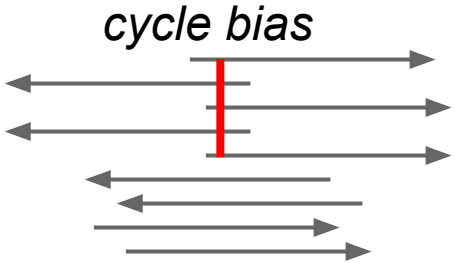
Our priors, **P(Genoypes)**, follow the Ewens Sampling Formula and the discrete sampling probability for genotypes.

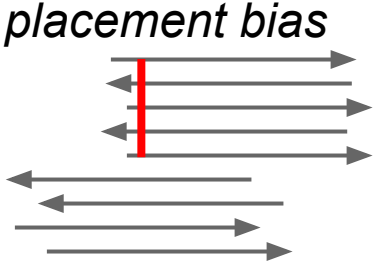
Are our locus and alleles sequenceable?

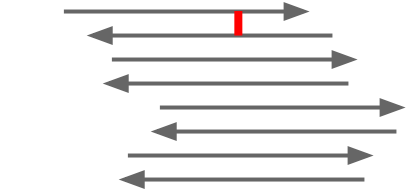
In WGS, biases in the way we observe an allele (placement, position, strand, cycle, or balance in heterozygotes) are often correlated with error. We include this in our posterior **P(G,S|R)**, and to do so we need an estimator of **P(S)**.





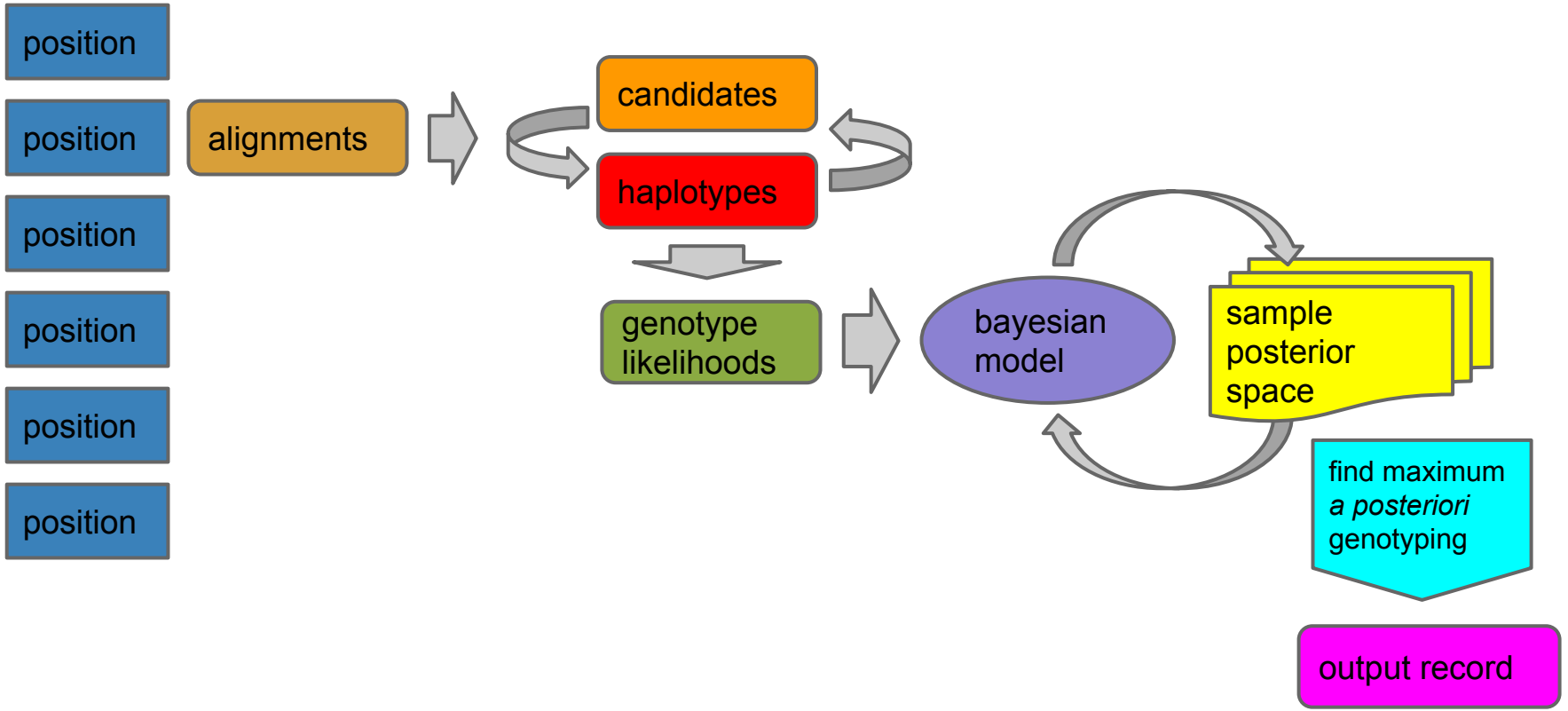






$$P(S) \propto \times \prod_{\forall b \in \{B\}} \begin{array}{l} multinom([|R \equiv b|\forall b_1, \ldots, b_K]; |\{R\}|, f_i, \ldots, f_K) \\ binom(|forwardStrand(\{R \equiv b\})|; |\{R \equiv b\}|, 1/2) \\ \times binom(|placedLeft(\{R \equiv b\})|; |\{R \equiv b\}|, 1/2) \\ \times binom(|placedRight(\{R \equiv b\})|; |\{R \equiv b\}|, 1/2) \end{array}$$

The detection process



Variant detector lineage

PolyBayes– original Bayesian variant detector (Gabor Marth, 1999); written in perl

GigaBayes– ported to C++

BamBayes– “modern” formats (BAM)

FreeBayes– 2010-present

FreeBayes-specific developments

FreeBayes model features (~in order of introduction):

- Multiple alleles
- Indels, SNPs, MNPs, complex alleles
- Local copy number variation (e.g. sex chromosomes)
- Global copy-number variation (e.g. species-level, genome ploidy)
- Pooled detection, both discrete and continuous
- Many, many samples (>30k exome-depth samples)
- Genotyping using known alleles (hints, haplotypes, or alleles)
- Genotyping using a reference panel of genotype likelihoods
- Direct detection of haplotypes from short-read sequencing
- Haplotype-based consensus generation (clumping)
- Allele-length-specific mapping bias
- Contamination-aware genotype likelihoods

Learning to genotype

Current best practices (in humans)

Lots of people use freebayes (not in human).

FYI: The current gold standard in human genomics is DeepVariant.

It *learns* how to genotype.

We won't use it in the practicals, but you should know how it works. It could help.

We can learn to genotype

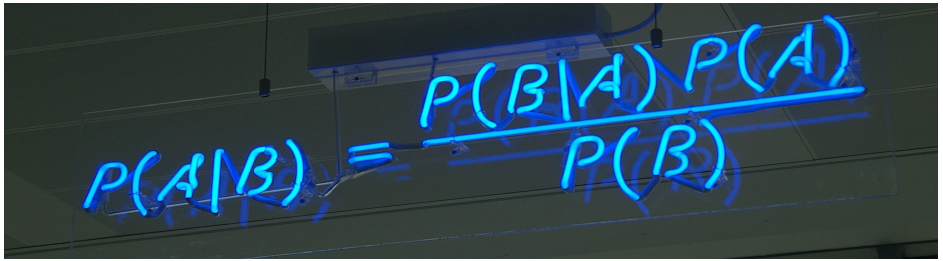
Bioinformaticians working with sequencing data can look at a visualization of alignments and make a good guess at the genotype.

“It looks wrong/right”

“It looks wrong/right”

Can machines learn to genotype?

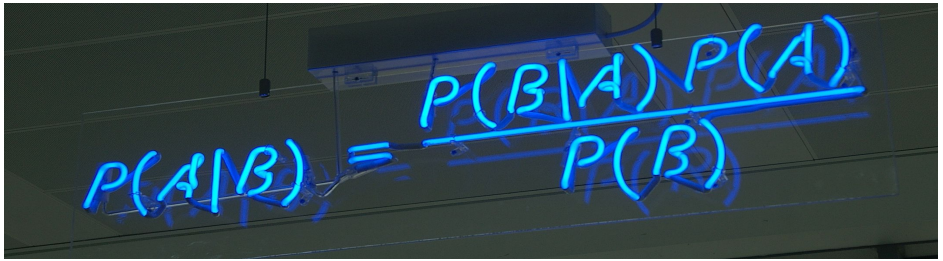
Thomas Bayes



Traditionally, we mix observations and *a priori* models using Bayesian statistics to find variants and estimate genotypes.

Can machines learn to genotype?

Marvin Minsky



But instead of building our model and prior from first principles, we could learn it (with machines).

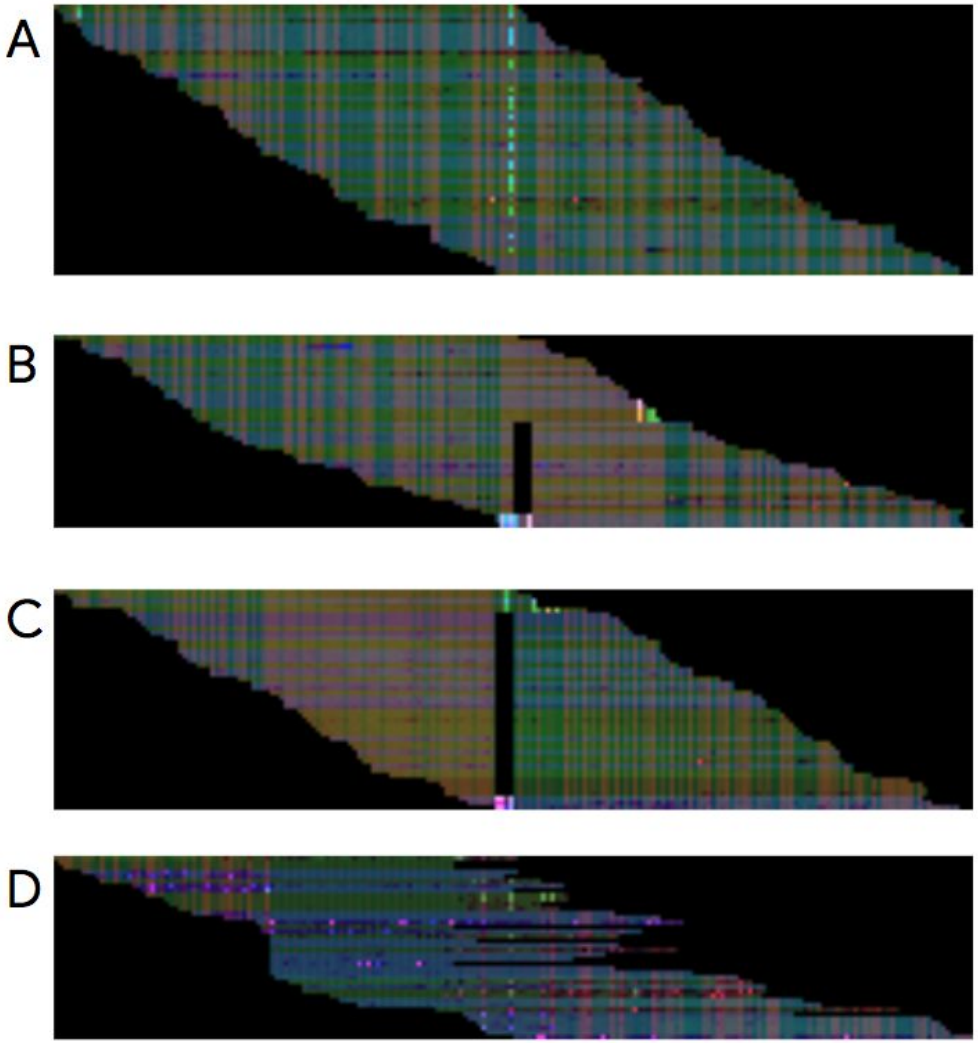
DeepVariant

Idea: you can leverage convolutional neural networks designed for images to *learn to genotype.*

https://doi.org/10.1038/nbt.4235

DeepVariant inputs

Idea: convert alignments to the reference into read pileups. Annotate various channels with useful things (like quality, read base, reference base, etc.)

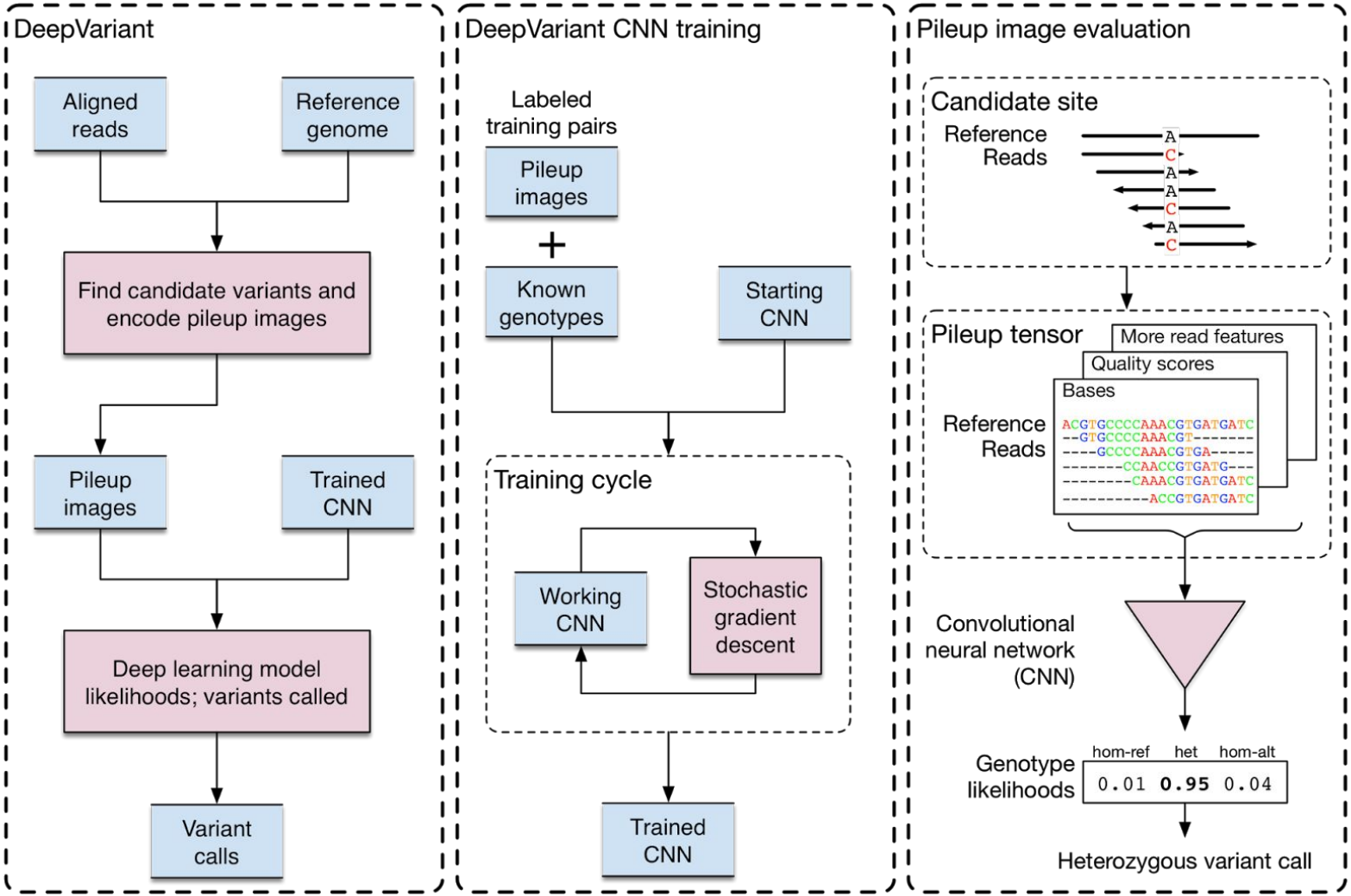


https://ai.googleblog.com/2017/12/deepvariant-highly-accurate-genomes.html

Channels are shown in greyscale below in the following order:

1. Read base: different intensities represent A, C, G, and T.
2. Base quality: set by the sequencing machine. White is higher quality.
3. Mapping quality: set by the aligner. White is higher quality.
4. Strand of alignment: Black is forward; white is reverse.
5. Read supports variant: White means the read supports the given alternate allele, grey means it does not.
6. Base differs from ref: White means the base is different from the reference, dark grey means the base matches the reference.

https://google.github.io/deepvariant/posts/2020-02-20-looking-through-deepvariants-eyes/


DeepVariant
Aligned reads
Reference genome
Find candidate variants and encode pileup images
Pileup images
Trained CNN
Deep learning model likelihoods; variants called
Variant calls
DeepVariant CNN training
Labeled training pairs
Pileup images
+
Known genotypes
Starting CNN
Training cycle
Working CNN
Stochastic gradient descent
Trained CNN
Pileup image evaluation
Candidate site
Reference Reads
Pileup tensor
More read features
Quality scores
Bases
Reference Reads
Convolutional neural network (CNN)
Genotype likelihoods
hom-ref het hom-alt
0.01 0.95 0.04
Heterozygous variant call

Words of caution

- Neural networks are universal approximators.
- But, they’re only guaranteed to model any pattern over the domain in which they’ve been trained.
- DeepVariant may work well (it wins all variant calling competitions), but it’s worth comparing it to other methods in the context of non-human genomes.

Practicals

Practical

Walkthrough:

https://github.com/ekg/alignment-and-variant-calling-tutorial/tree/evomics2024

Goal is to get our hand dirty with a variant calling workflow, and maybe to dig into interesting edge cases that arise.

Notes

- The workflow is pretty complete. You can copy-paste if you want, but please look at what you're ingesting and producing at each step.
- Data and time-consuming indexing results are in ~/workshop_materials/variant_calling
- If something is taking too long or you've got stuck, there is also a .results directory that contains most outputs.
- At the end we've got some open-ended mini-projects. Explore them!