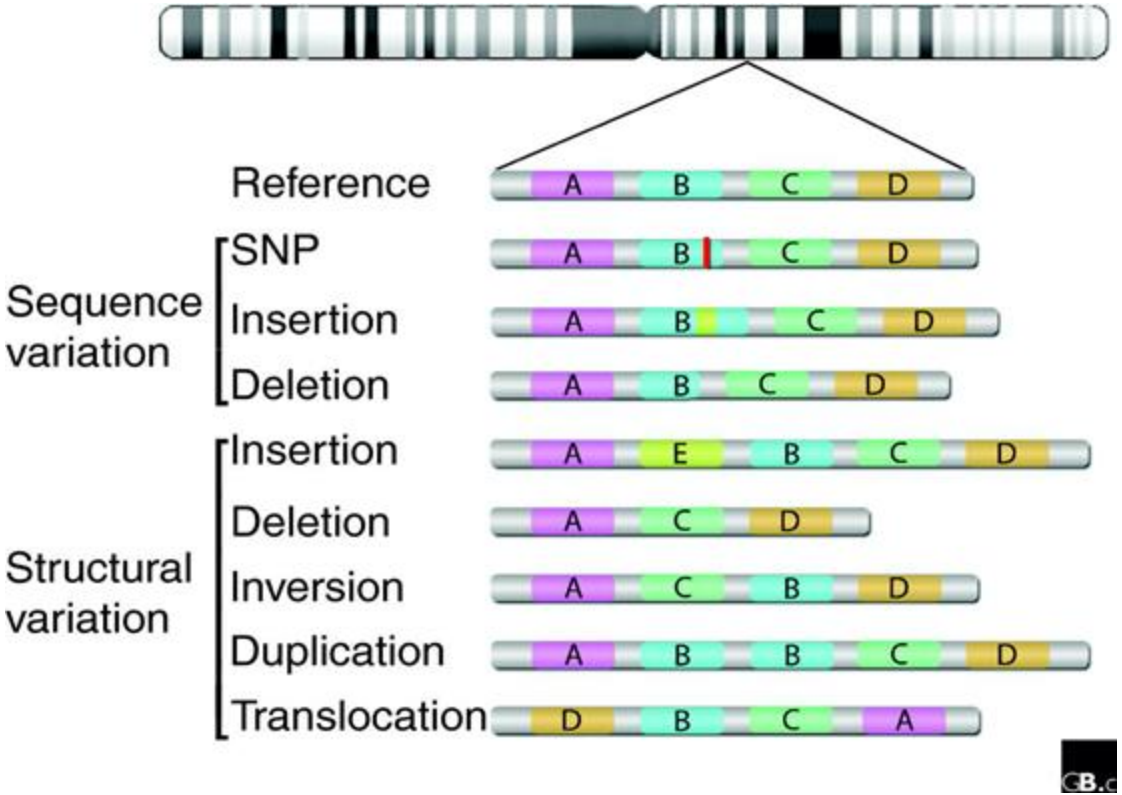
Variant calling

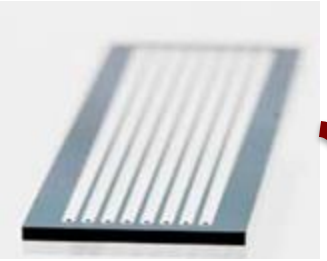
Chris Miller

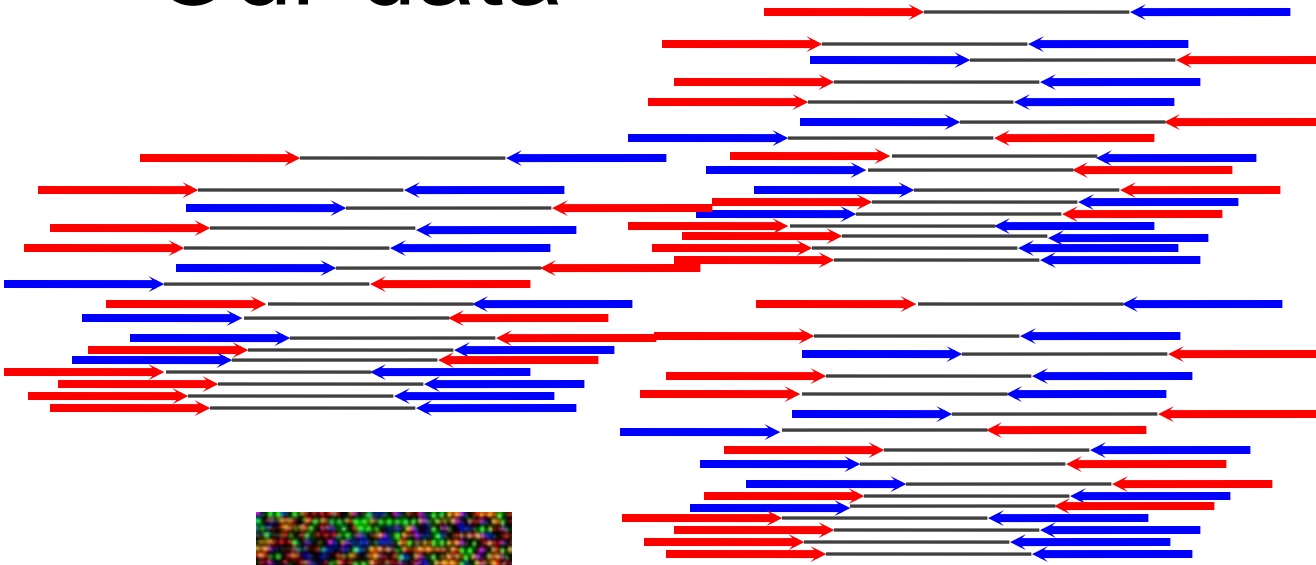
Some slides adapted from Dave Larson, Aaron Quinlan, Sam Peters, and Alex Paul

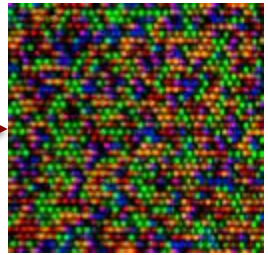
Small Variant Calling



Our data

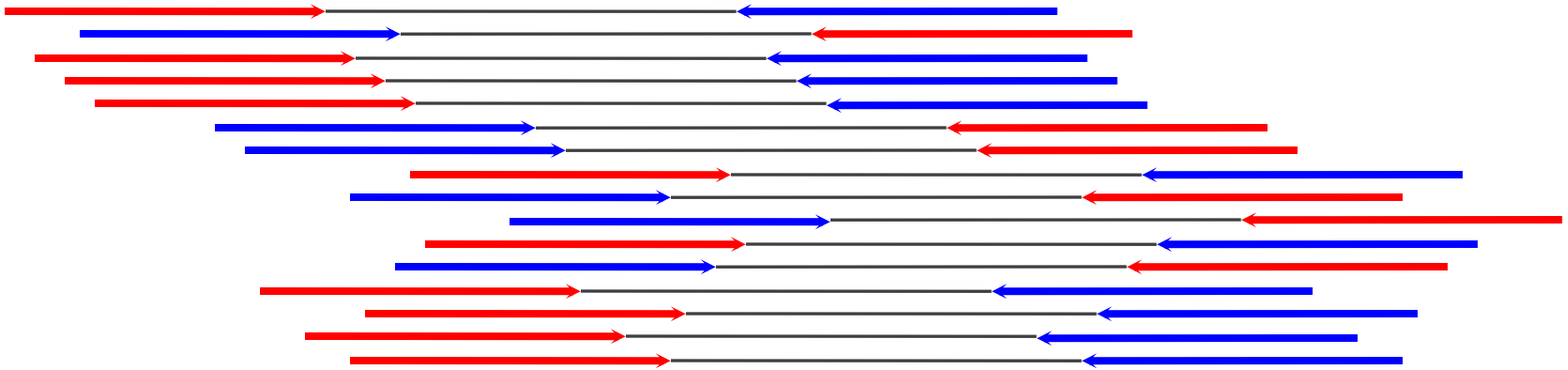






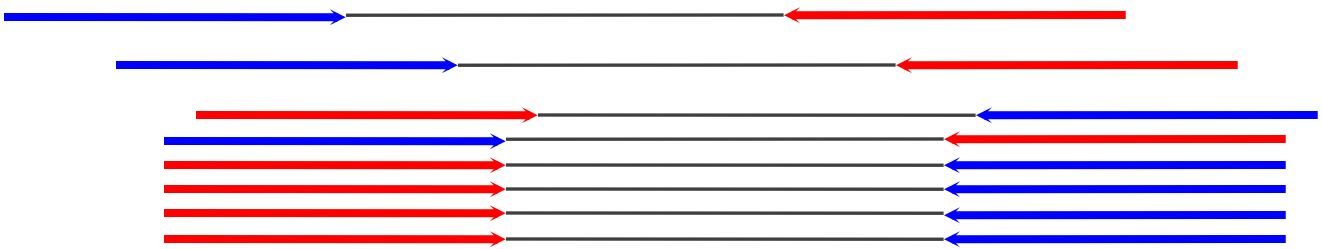
Mapping

Genome Reference Sequence

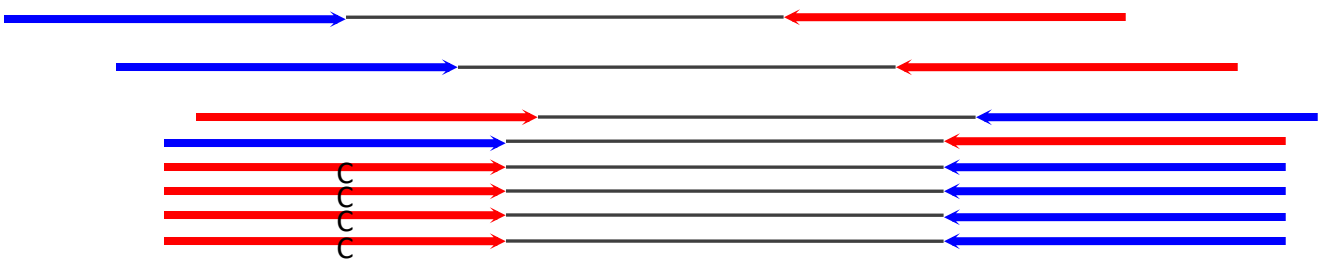


- Single-end reads can be longer, less unique depending on sequence context
- Paired-end reads can span repetitive regions, provide additional information
- Mapping has gotten quite fast, <24 hours for 120 Gbp of sequence
- Split-read alignments are the norm (BWA mem)

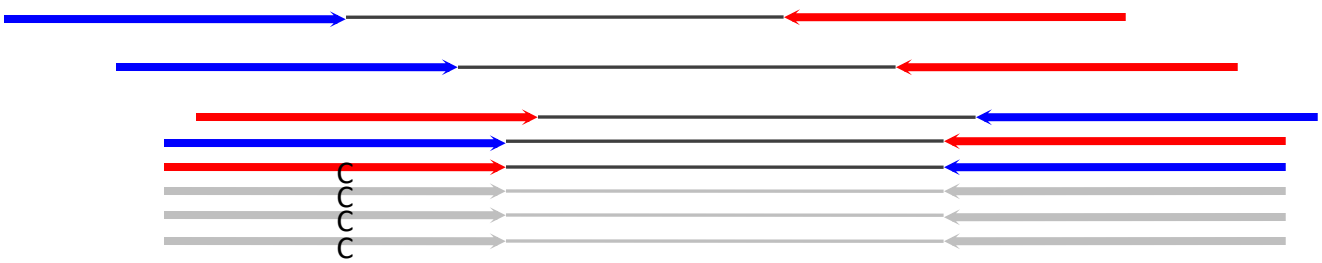
Duplication



Duplication

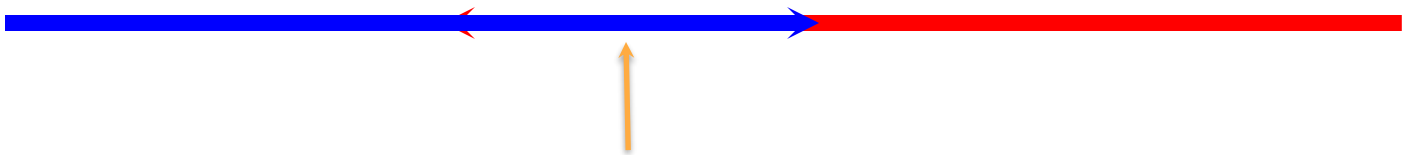


Duplication

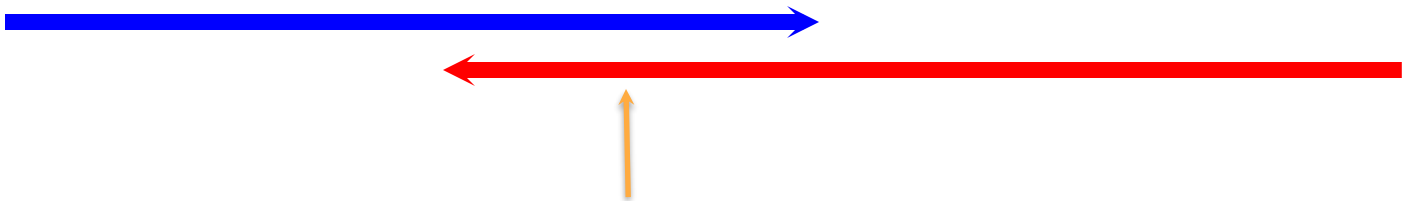


Overlapping reads

Overlapping reads



Overlapping reads



Overlapping reads

Every aspect of this process is fraught with error

- Base calling is not perfect: *0.5% error on average*
- Mapping is not perfect*: the reads are short*
- The reference sequence is not perfect

We have a little help

- Some uncertainty is encapsulated in quality scores
  - the rate at which the data is expected to be wrong
- Each base call (ACTGN) comes with a quality
  - Phred-scaled (-10 * $\log_{10}$ of quality)
  - A base call with quality of 20 is wrong 1 out of every 100 times.
- Read mapping has quality too
  - These are also Phred-scaled

Phred quality score calculation

$$Q = -10 * \log_{10}(P_{err})$$

Error probability (P_{err})	$\log_{10}(P_{err})$	Phred quality score
1	0	0
0.1	-1	10
0.01	-2	20
0.001	-3	30
0.0001	-4	40

Goals of a Variant Caller

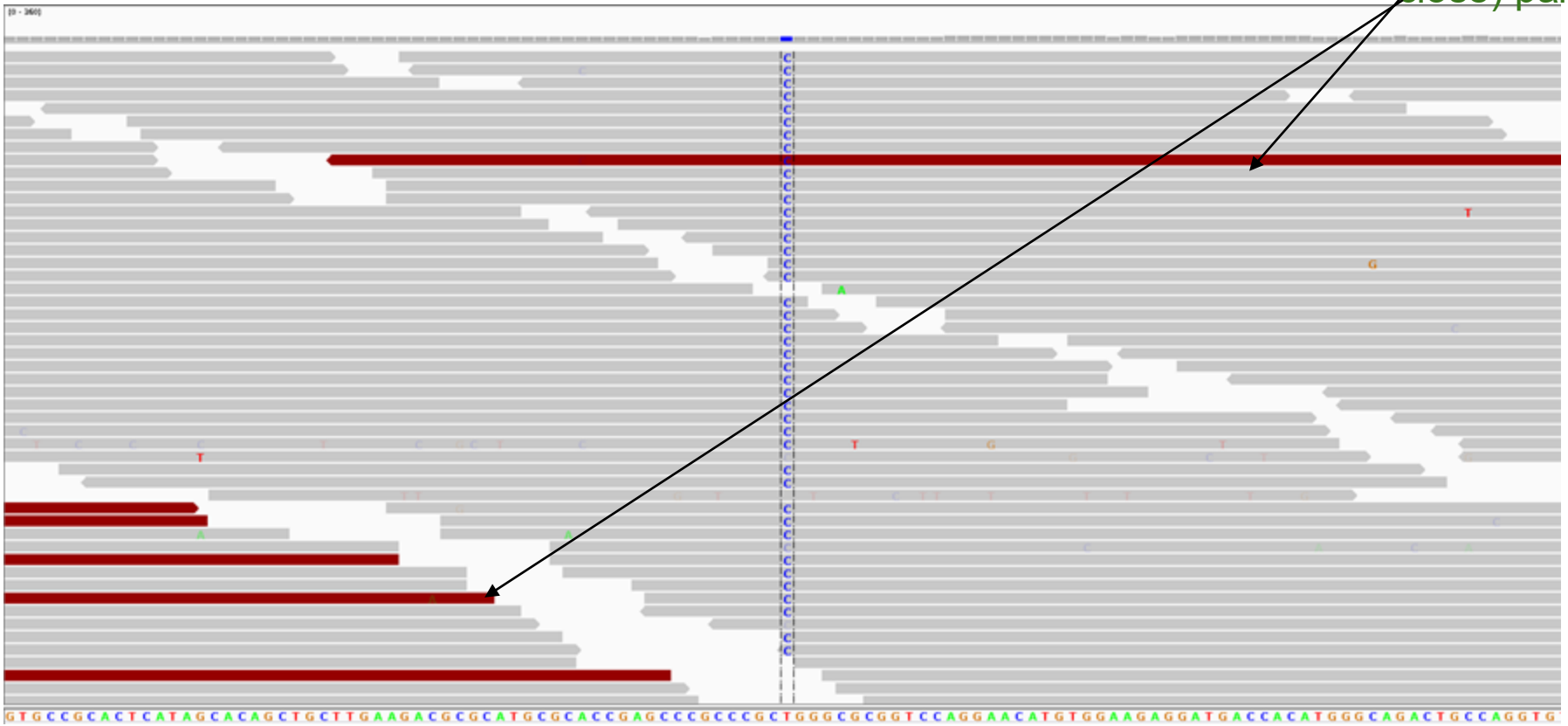
- Sensitively detect mutations
- Precisely detect mutations
  - Confounded by the error we just talked about
  - FDR must be very low as we're looking across a very large space!

Goals of a Variant Caller

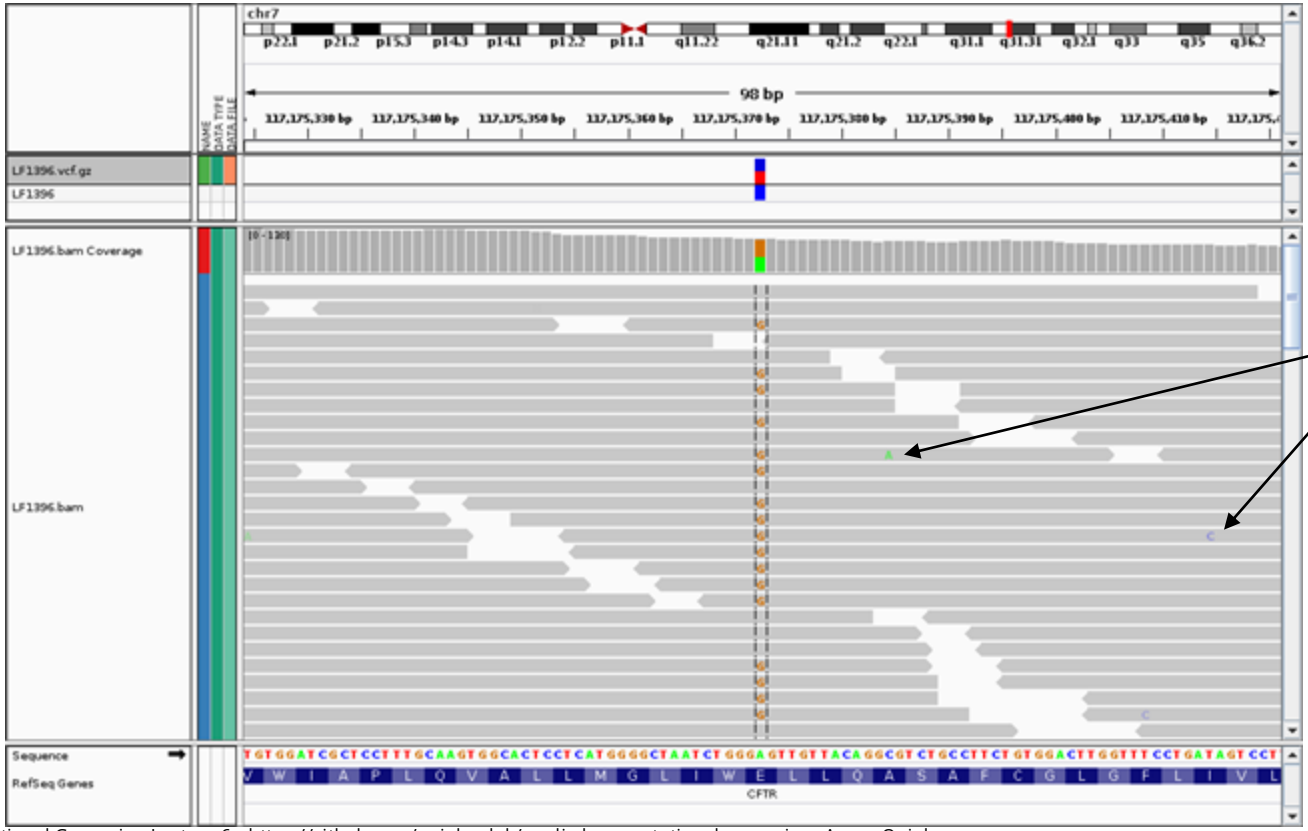
- Sensitively detect mutations
- Precisely detect mutations
  - Confounded by the error we just talked about
  - FDR must be very low as we're looking across a very large space!

    - **An FDR of 0.001 = 3.2 million false positives!**

Homozygous for the "C" allele



Sequencing errors fall out as noise (most of the time)



Sequencing errors

adapted from: Applied Computational Genomics, Lecture 6 - https://github.com/quinlan-lab/applied-computational-genomics - Aaron Quinlan

https://ichoigt.files.wordpress.com/2012/07/igv_e217g_snapshot.png

It is not always so easy

Random versus systematic error

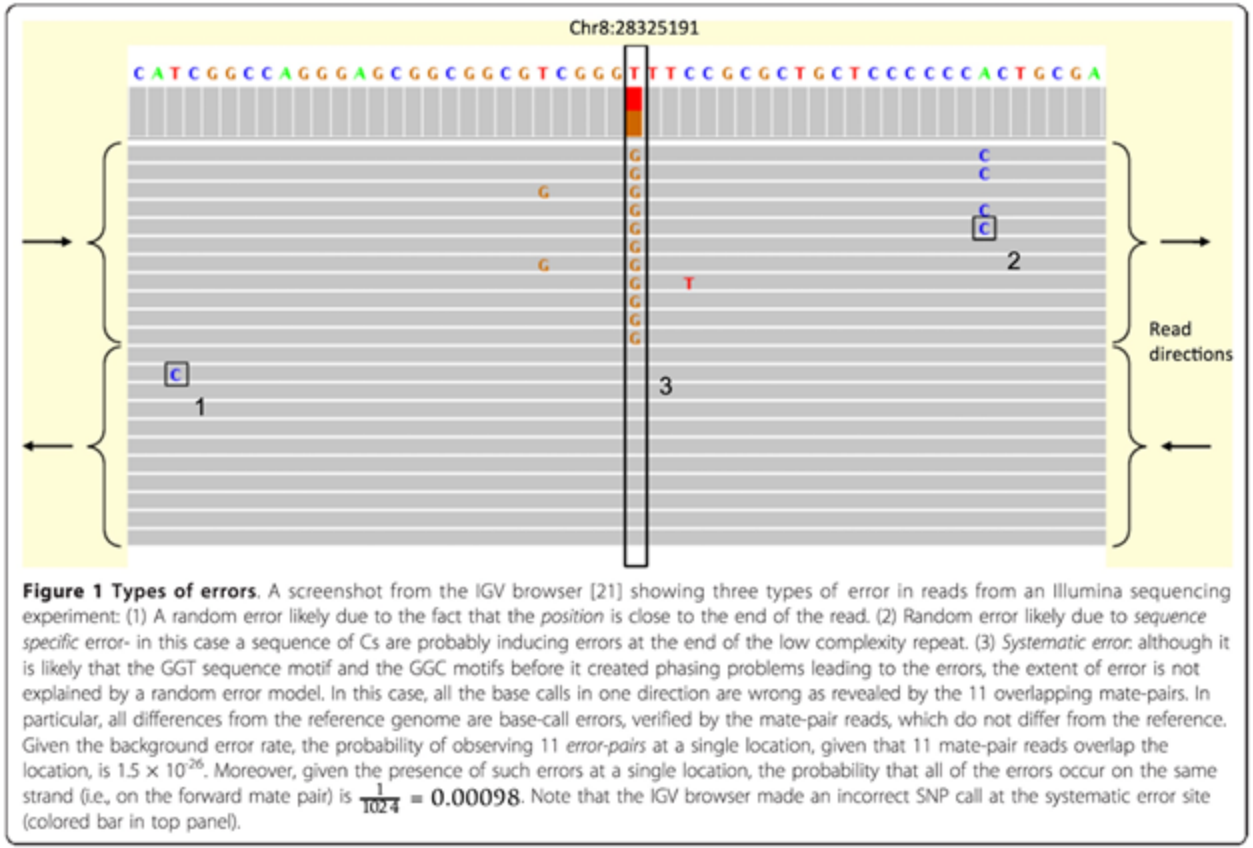


Figure 1 Types of errors. A screenshot from the IGV browser [21] showing three types of error in reads from an Illumina sequencing experiment: (1) A random error likely due to the fact that the *position* is close to the end of the read. (2) Random error likely due to *sequence specific* error- in this case a sequence of Cs are probably inducing errors at the end of the low complexity repeat. (3) *Systematic error*: although it is likely that the GGT sequence motif and the GGC motifs before it created phasing problems leading to the errors, the extent of error is not explained by a random error model. In this case, all the base calls in one direction are wrong as revealed by the 11 overlapping mate-pairs. In particular, all differences from the reference genome are base-call errors, verified by the mate-pair reads, which do not differ from the reference. Given the background error rate, the probability of observing 11 *error-pairs* at a single location, given that 11 mate-pair reads overlap the location, is 1.5×10^{-26}. Moreover, given the presence of such errors at a single location, the probability that all of the errors occur on the same strand (i.e., on the forward mate pair) is $\frac{1}{1024} = 0.00098$. Note that the IGV browser made an incorrect SNP call at the systematic error site (colored bar in top panel).

adapted from: Applied Computational Genomics, Lecture 6 - https://github.com/quinlan-lab/applied-computational-genomics - Aaron Quinlan

https://www.researchgate.net/publication/51814240_Identification_and_correction_of_systematic_error_in_high-throughput_sequence_data

Pileups of many differences from paralogy

RESEARCH ARTICLE | OPEN ACCESS

FLAGS, frequently mutated genes in public exomes

Casper Shyr, Maja Tarailo-Graovac, Michael Gottlieb, Jessica JY Lee, Clara van Karnebeek and Wyeth W Wasserman

BMC Medical Genomics 2014 7:64 | DOI: 10.1186/s12920-014-0064-y

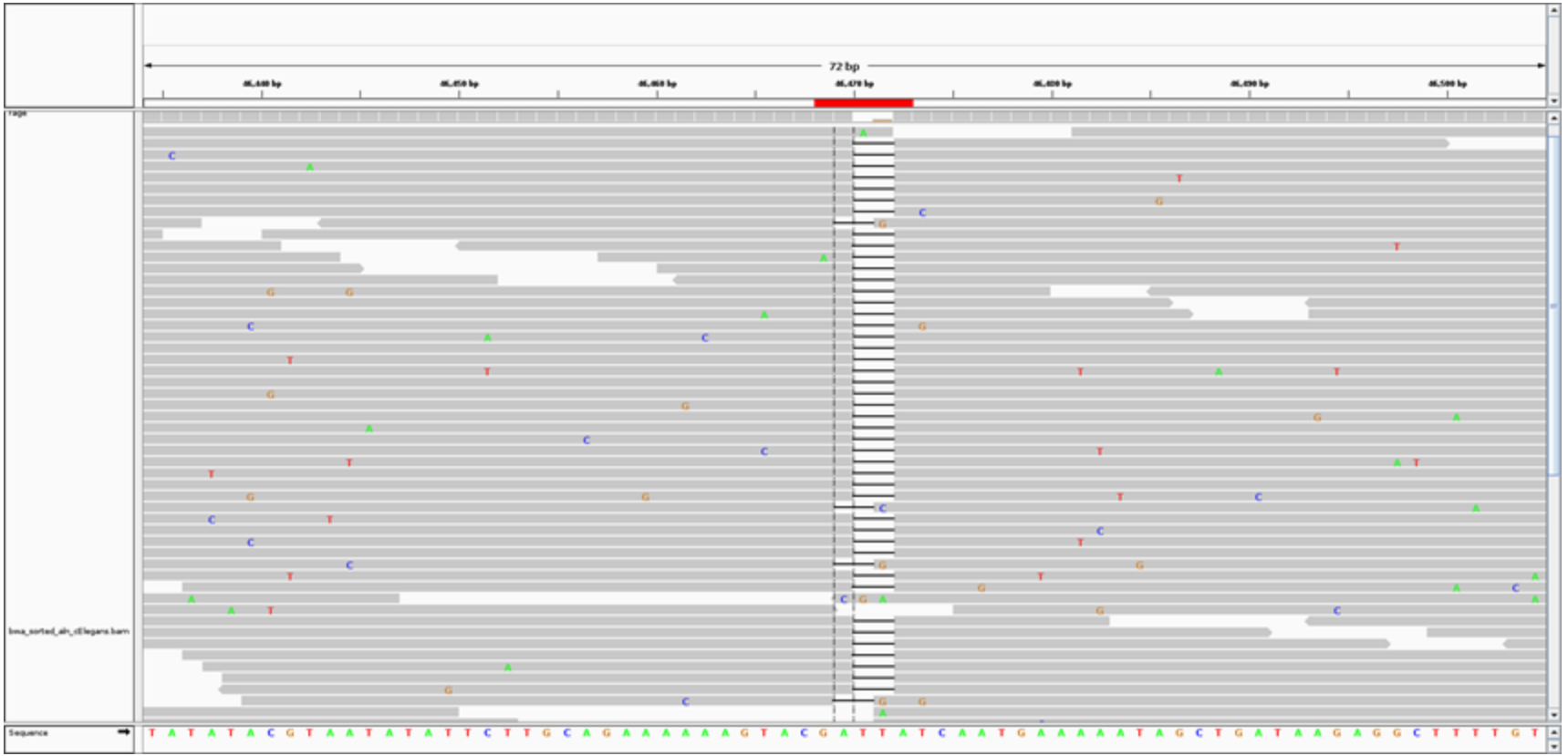
Received: 16 June 2014 | Accepted: 24 October 2014 | Published: 3 December 2014

Open Peer Review reports

adapted from: Applied Computational Genomics, Lecture 6 - https://github.com/quinlan-lab/applied-computational-genomics - Aaron Quinlan

http://massgenomics.org/2013/06/ngs-false-positives.html

Calling INDELs is *much* harder than SNPs

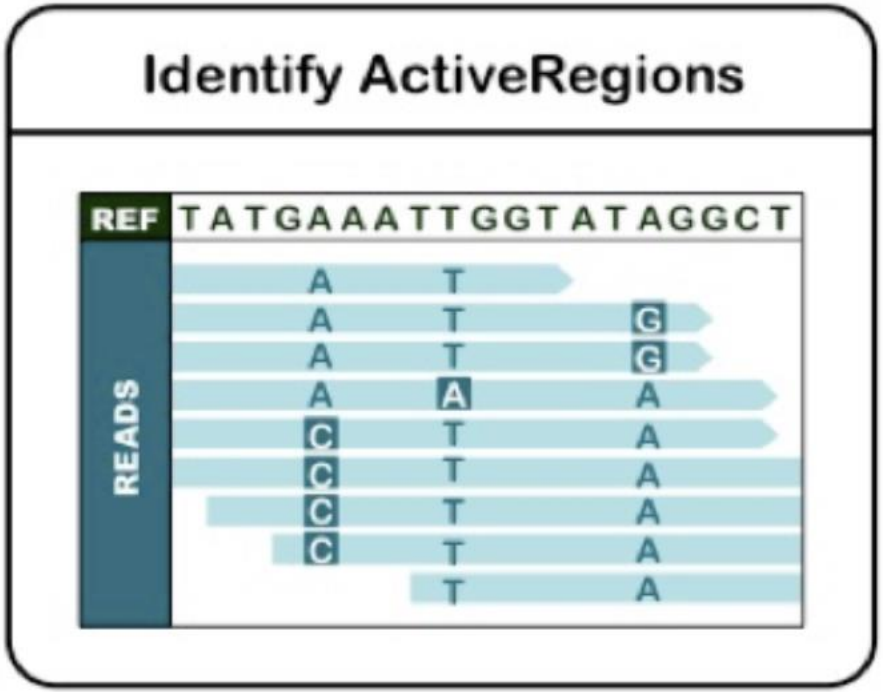


adapted from: Applied Computational Genomics, Lecture 6 - https://github.com/quinlan-lab/applied-computational-genomics - Aaron Quinlan

Begin Exercise

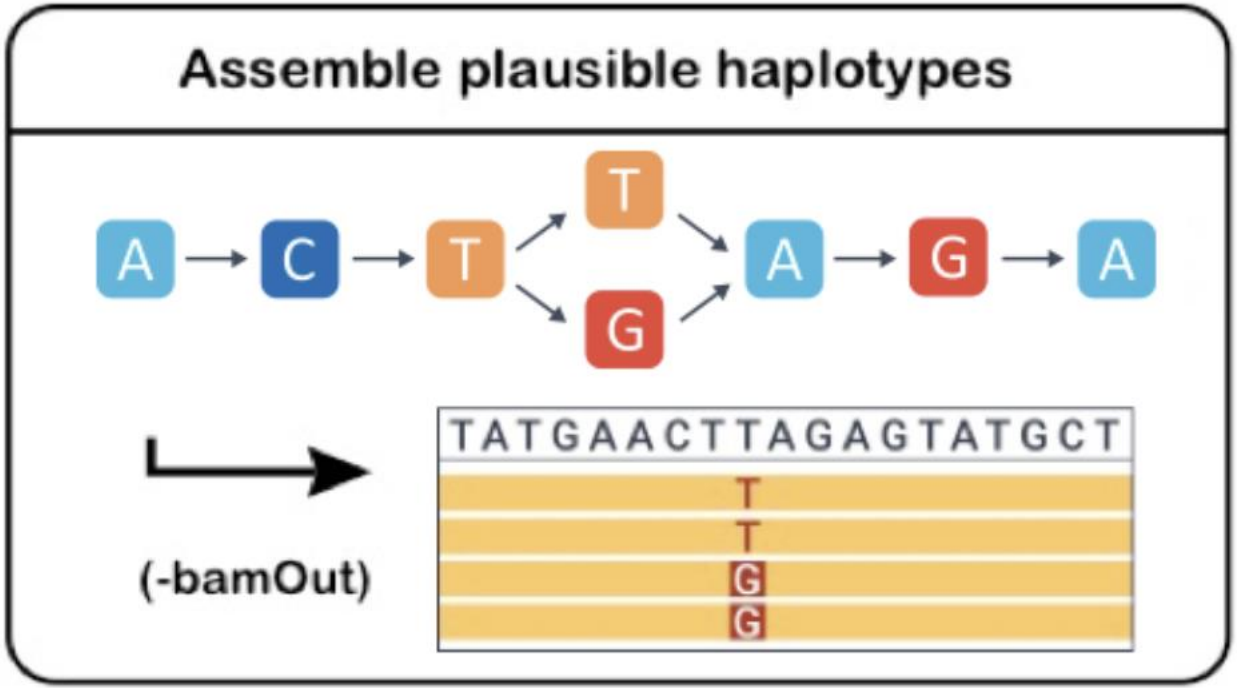
Germline SNV and Indel Calling

Call Genotypes Using GATK HaplotypeCaller



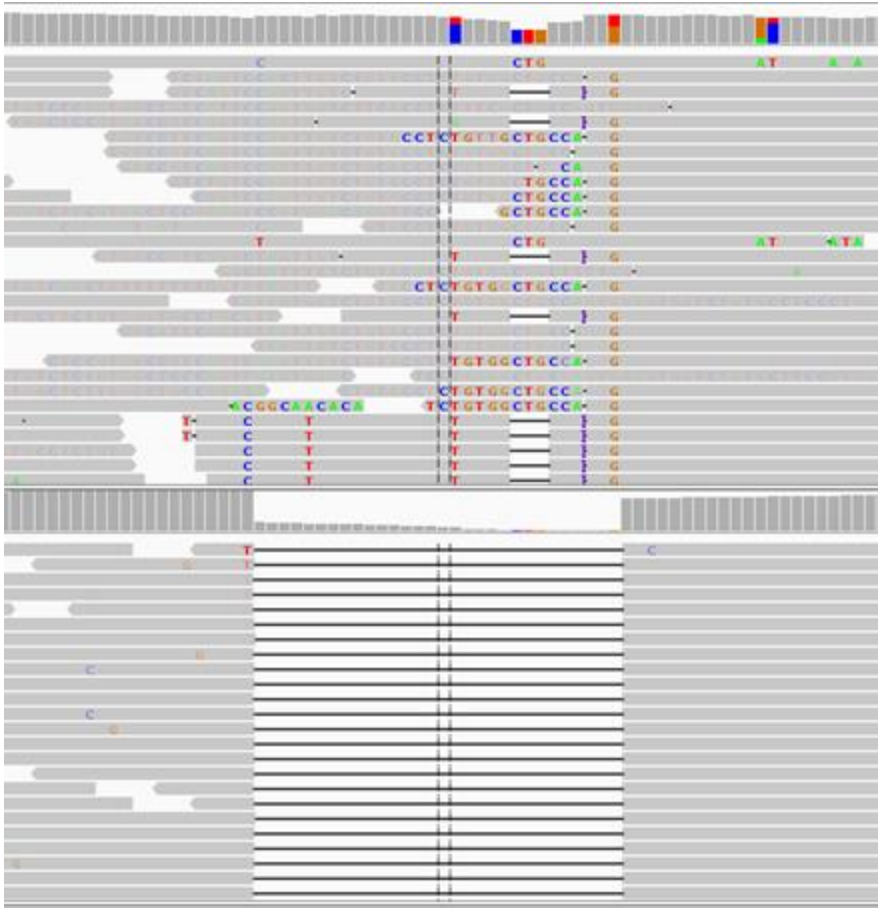
https://gatk.broadinstitute.org/hc/en-us/articles/360035531412-HaplotypeCaller-in-a-nutshell

Call Genotypes Using GATK HaplotypeCaller



https://gatk.broadinstitute.org/hc/en-us/articles/360035531412-HaplotypeCaller-in-a-nutshell

Indel "realignment"

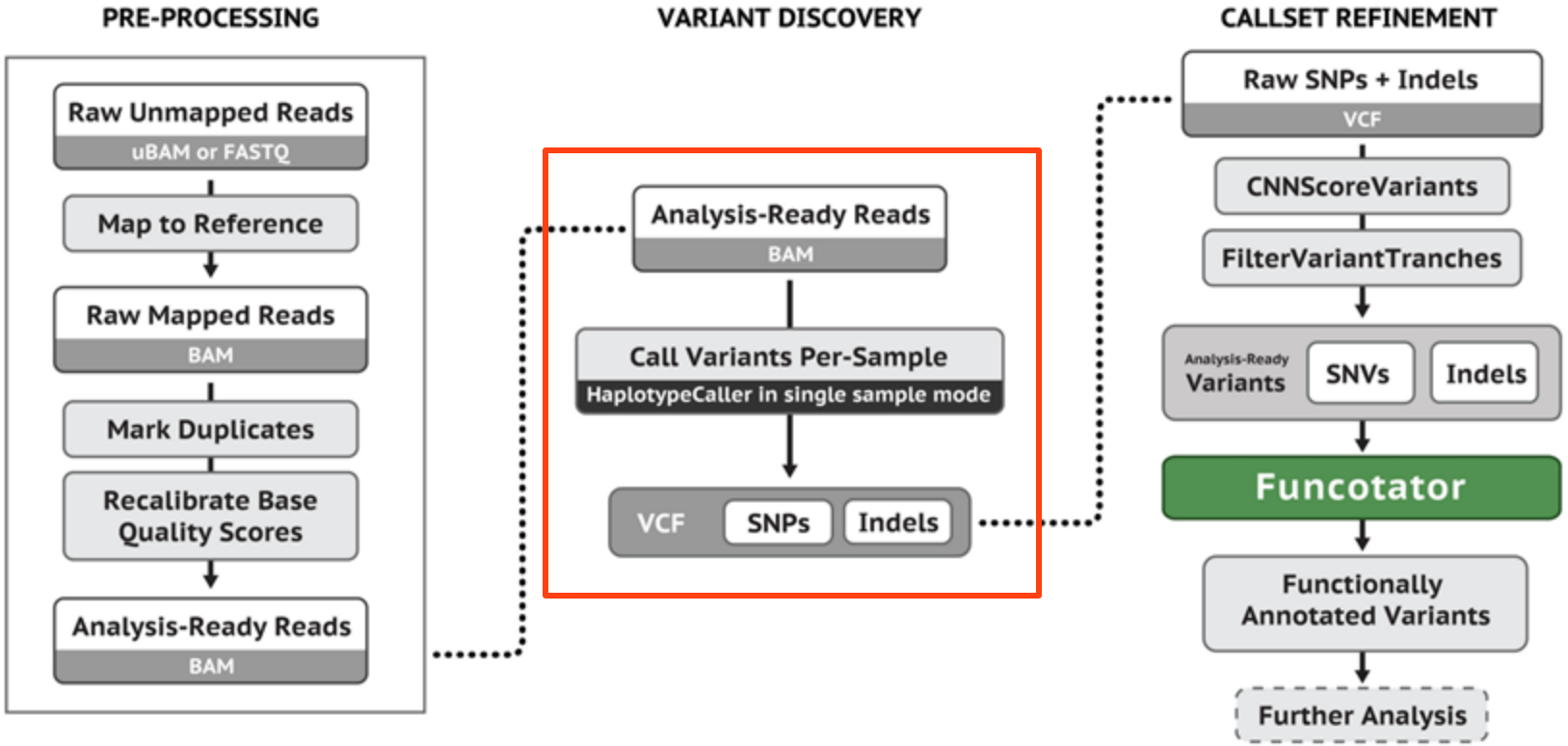


https://github.com/mozack/abra

Call Genotypes Using GATK HaplotypeCaller

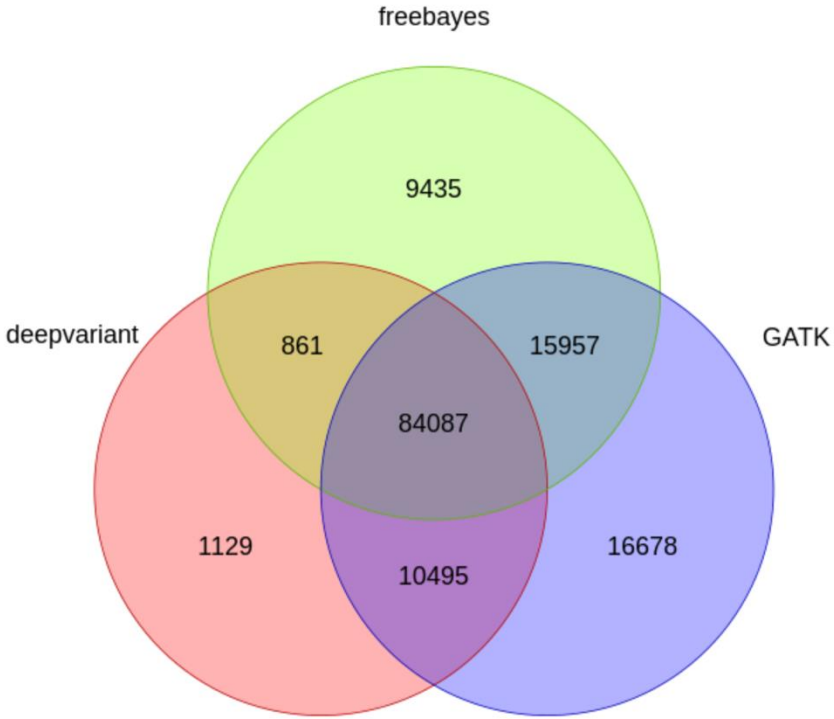
https://gatk.broadinstitute.org/hc/en-us/articles/360035531412-HaplotypeCaller-in-a-nutshell

Main steps for Germline Single-Sample Data



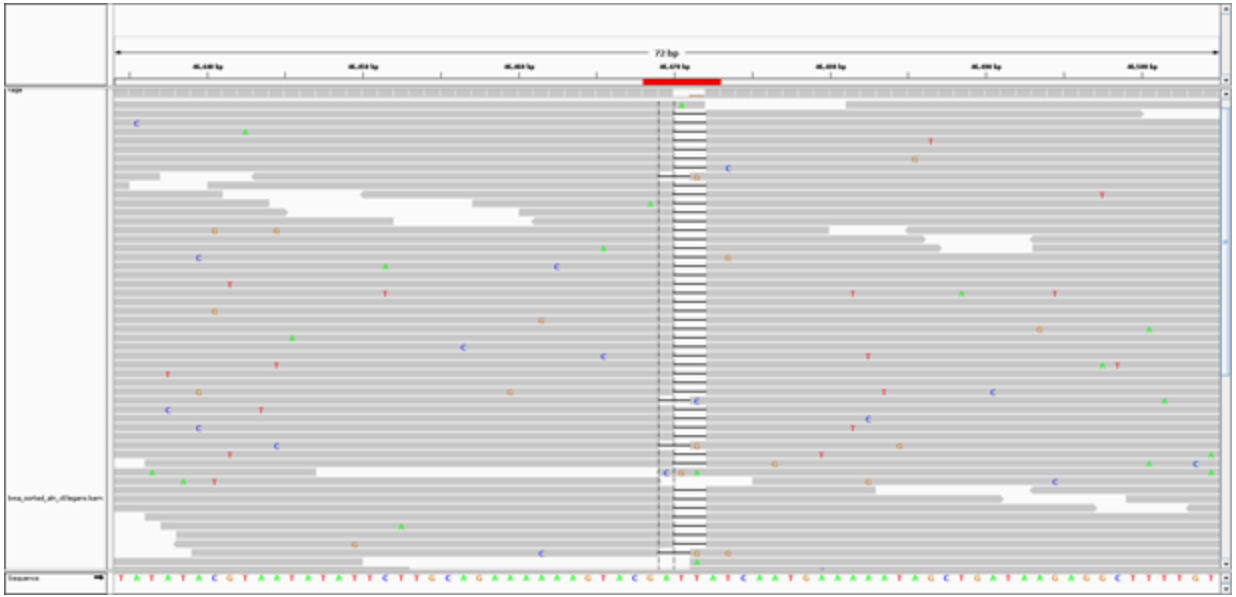
Other Germline Variant callers

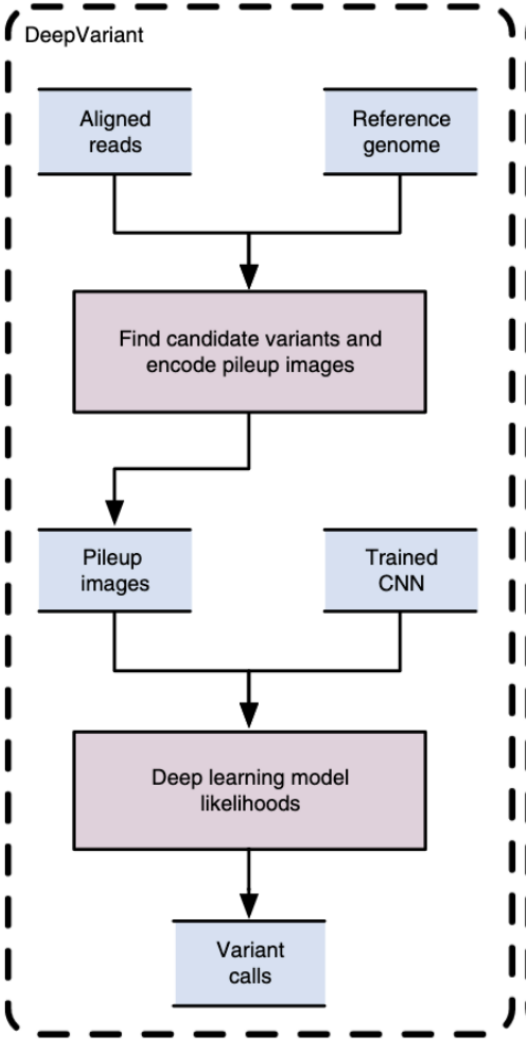
- Freebayes: lighter weight, faster
- DeepVariant: Neural-network based caller from Google ML



https://ucdavis-bioinformatics-training.github.io/2021-July-Genome-Wide-Association-Studies/data_reduction/comparison

DeepVariant

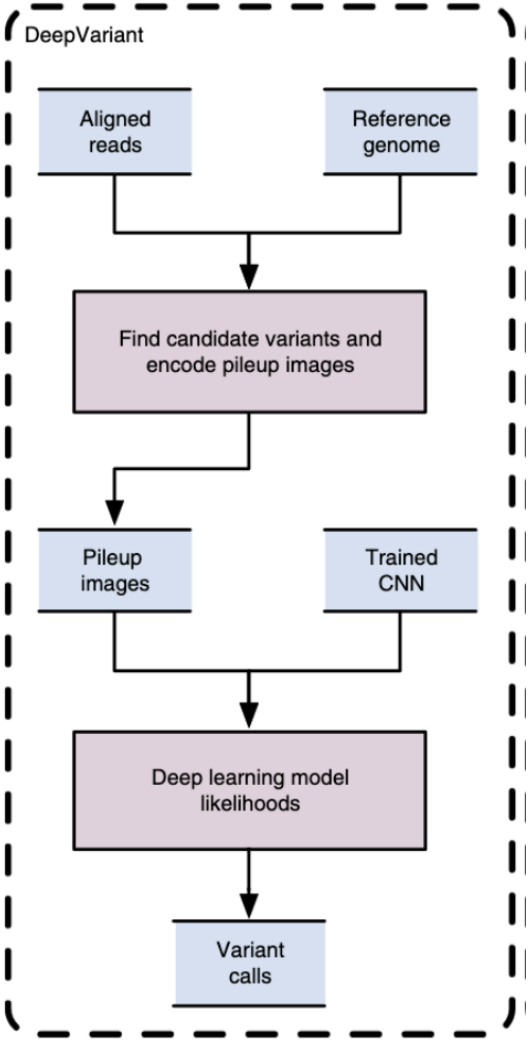




https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6450397/

https://www.nature.com/articles/nbt.4235

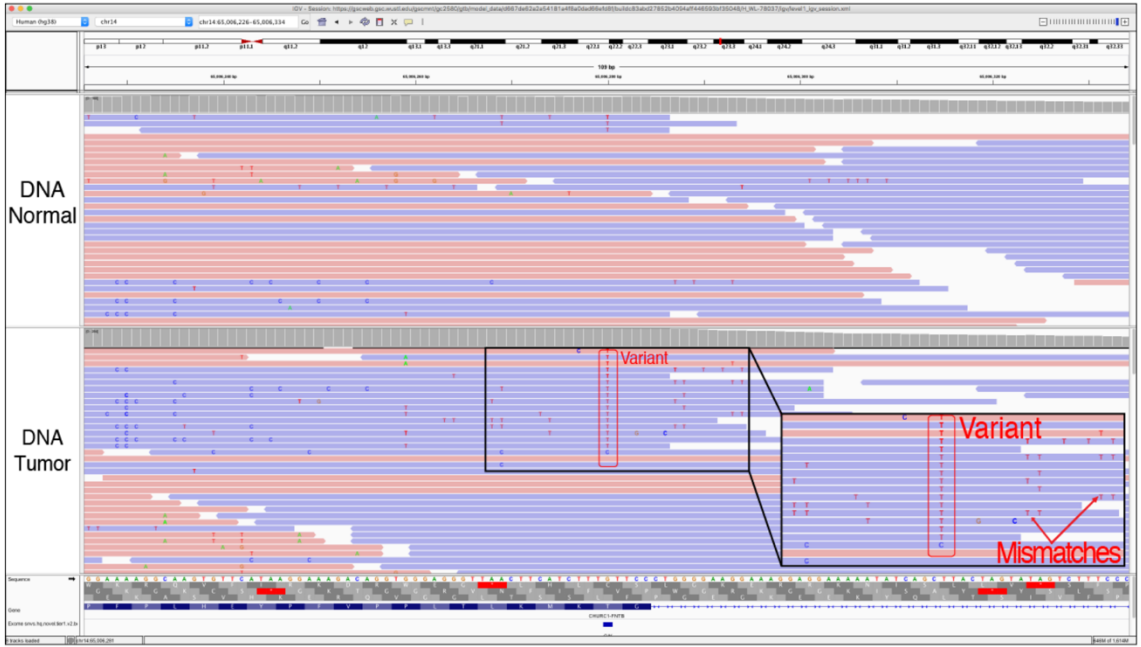
DeepVariant

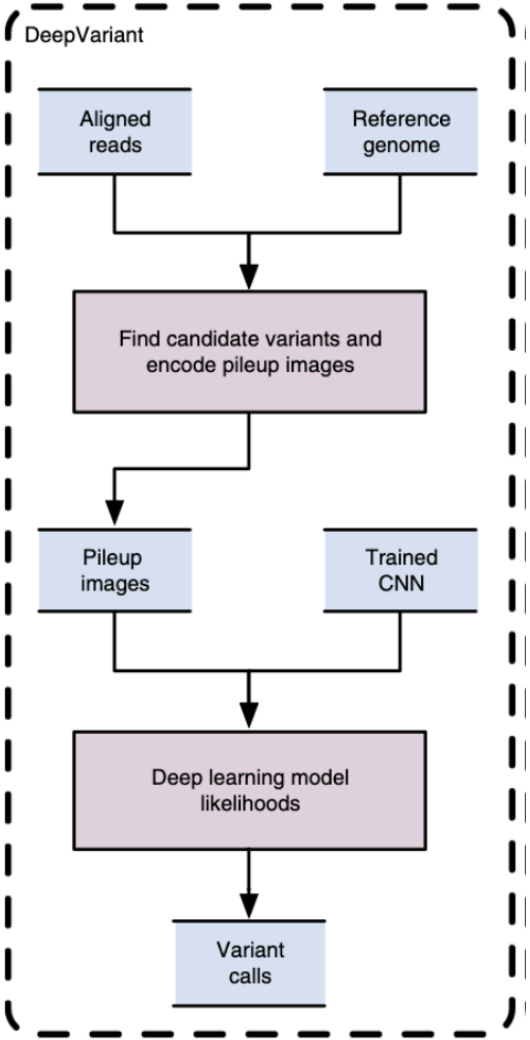


https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6450397/

https://www.nature.com/articles/nbt.4235

DeepVariant

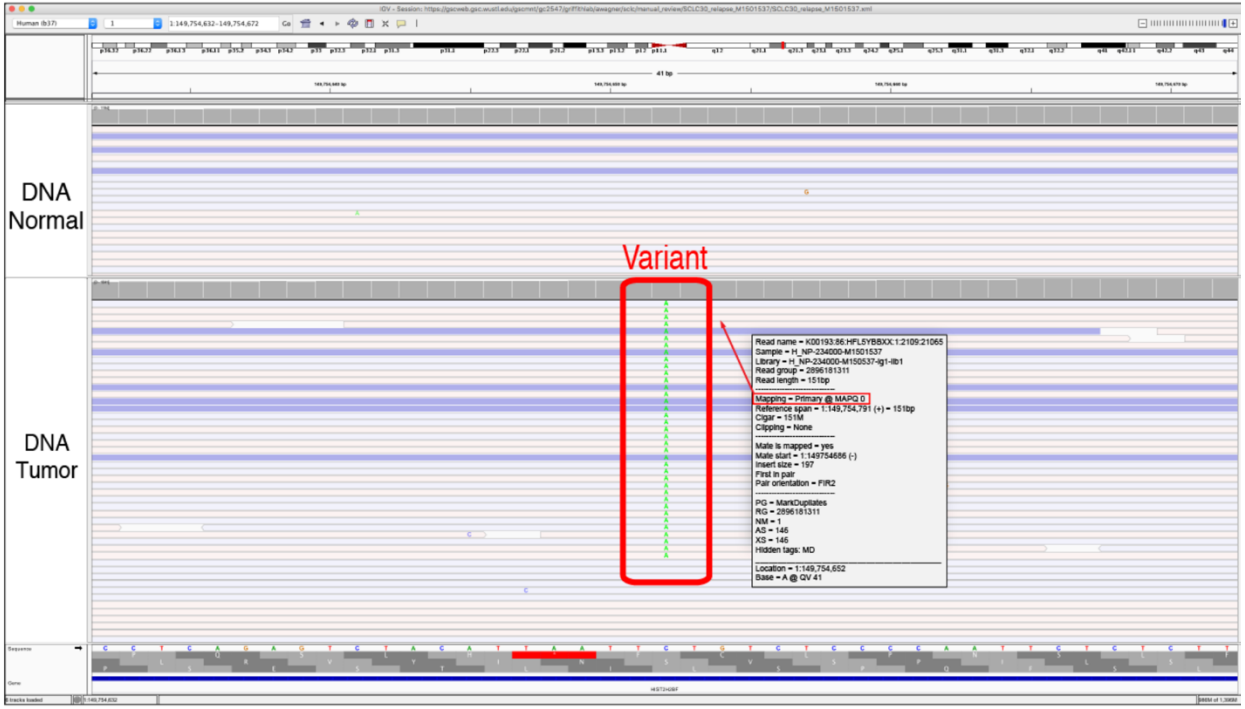


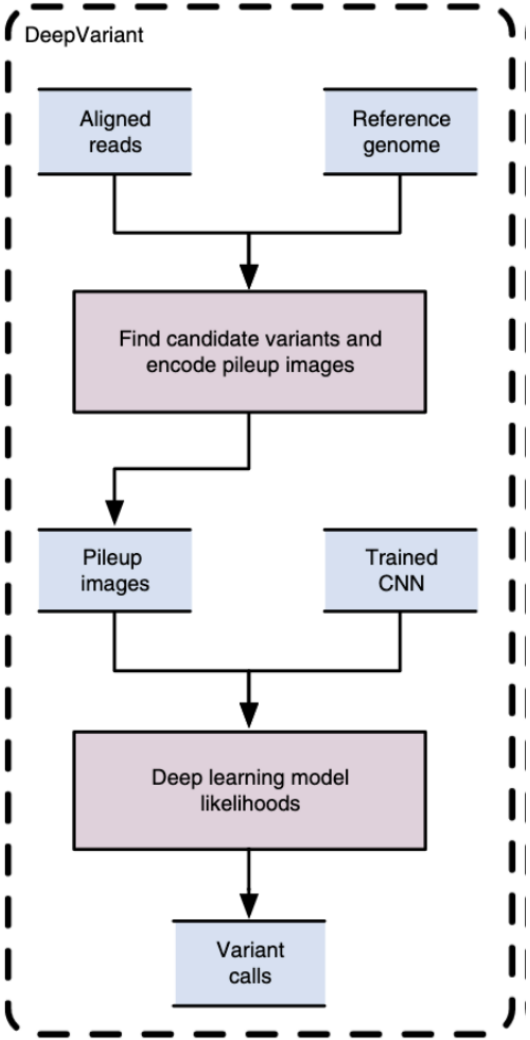


https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6450397/

https://www.nature.com/articles/nbt.4235

DeepVariant





https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6450397/

https://www.nature.com/articles/nbt.4235

DeepVariant

Here are 3 examples that we would consider canonical easy-to-classify loci, and that DeepVariant calls confidently and correctly:

The variant above is a “2”, which means both chromosomes match the variant allele, so this locus represents a homozygous alternate locus.

DeepVariant correctly classifies the variant above as a “1”, which means that one of the two alleles matches the variant allele, i.e. it is heterozygous.

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