Bioinformatics Workshop

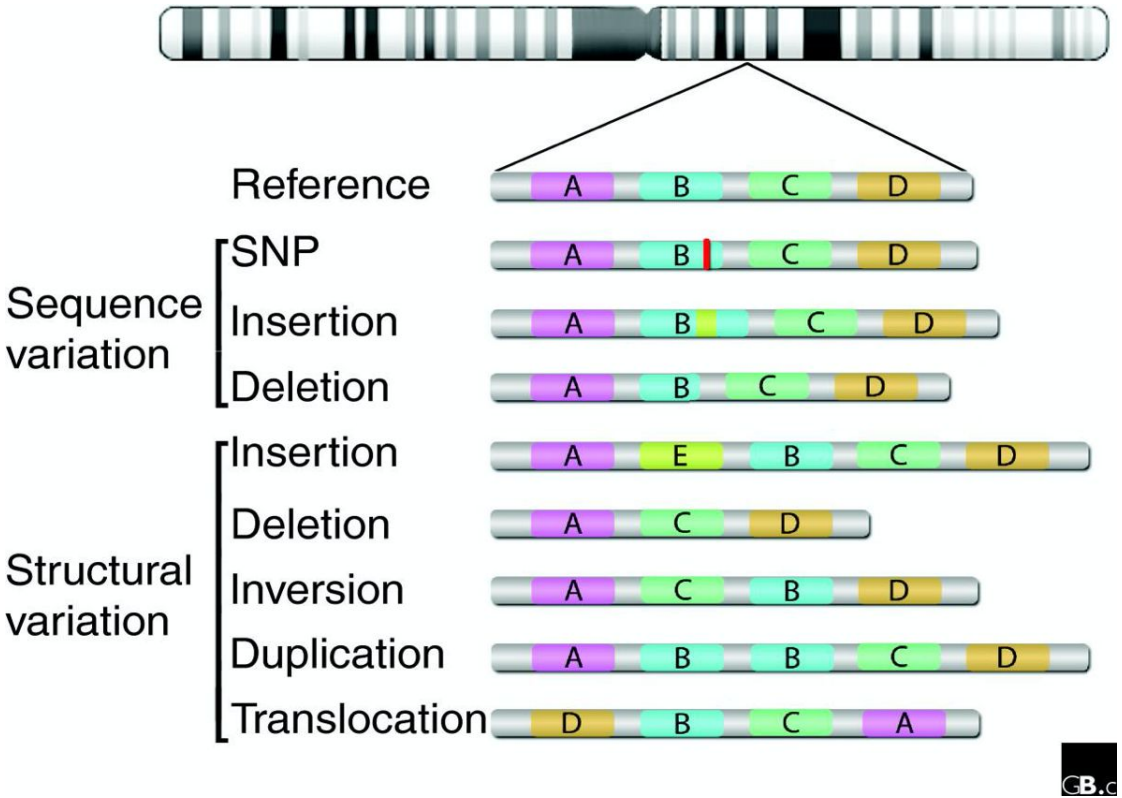
Week 06

Variant Calling

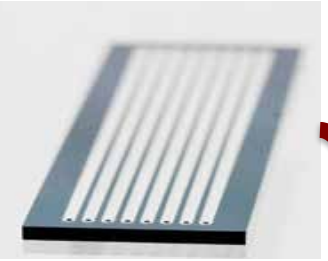
Chris Miller

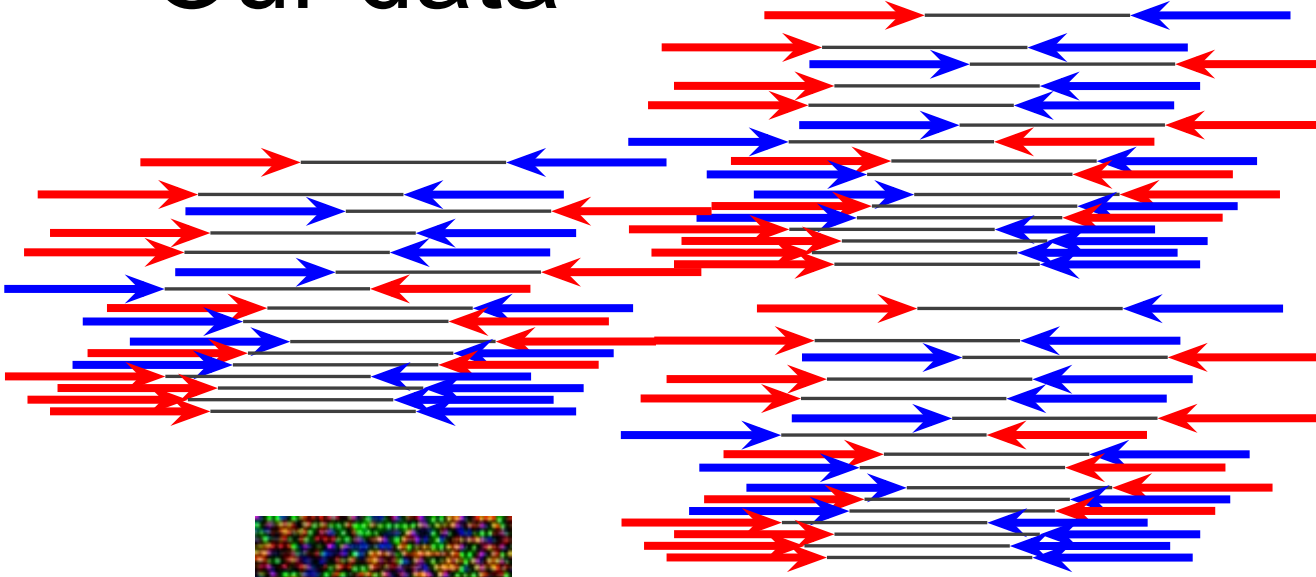
Some slides adapted from Dave Larson, Aaron Quinlan, Sam Peters, and Alex Paul

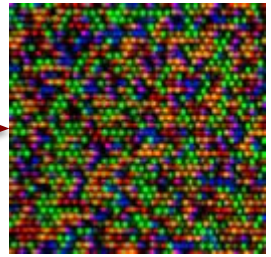
Small Variant Calling



Our data

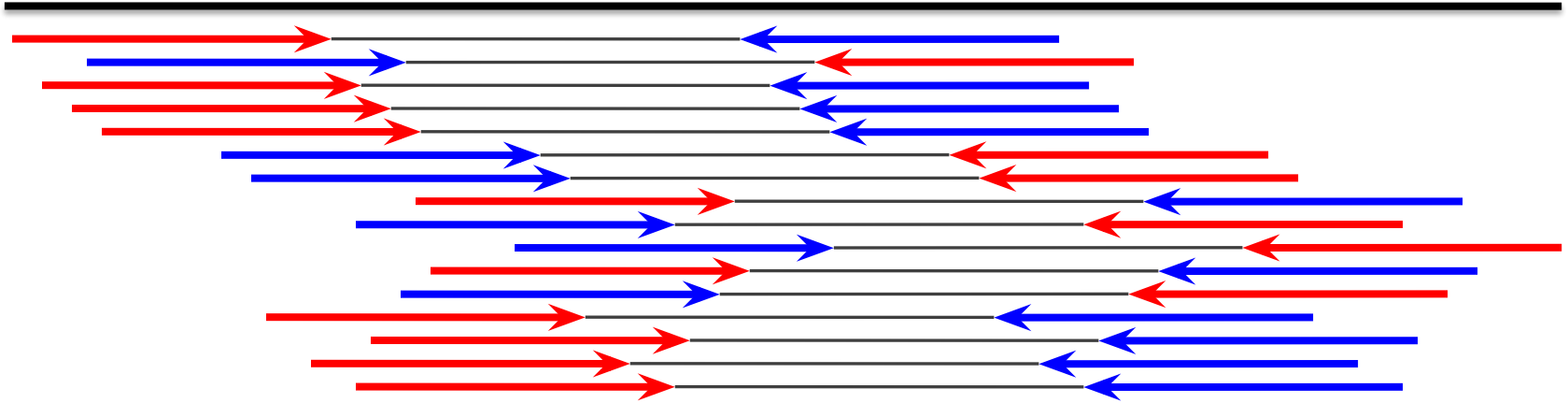






Mapping

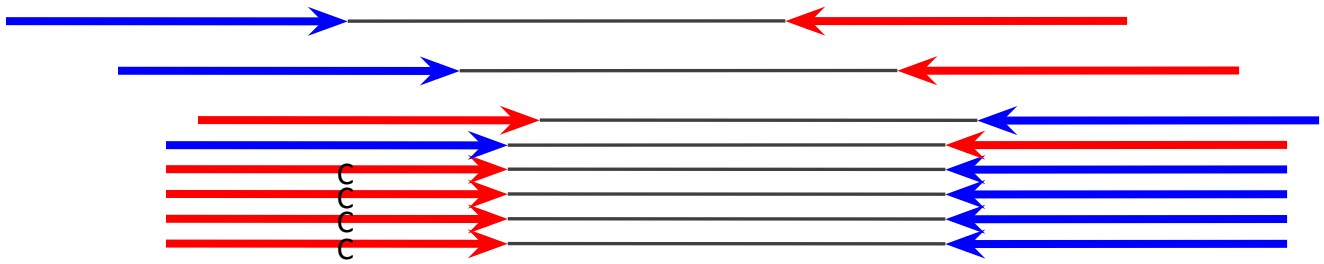
Genome Reference Sequence



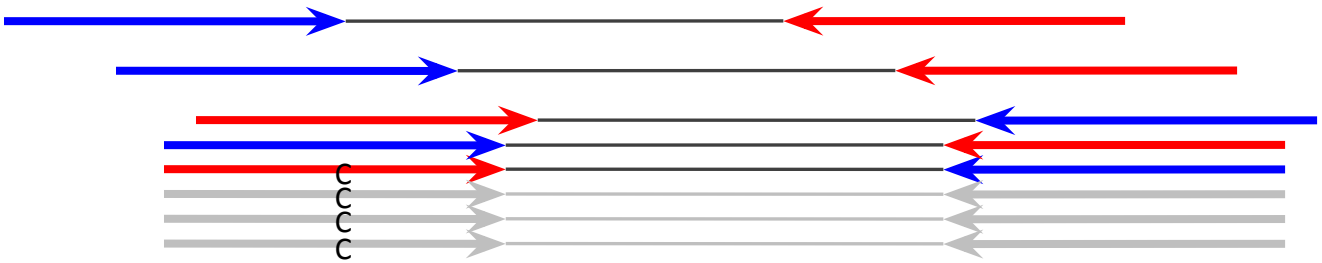
- Single-end reads can be longer, less unique depending on sequence context
- Paired-end reads can span repetitive regions, provide additional information
- Mapping has gotten quite fast, <24 hours for 120 Gbp of sequence
- Split-read alignments are the norm (BWA mem)

Duplication

Duplication

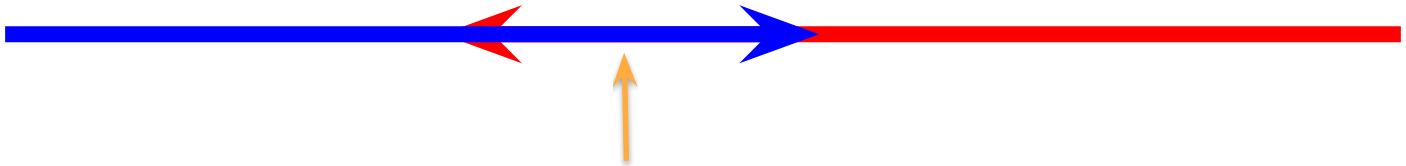


Duplication

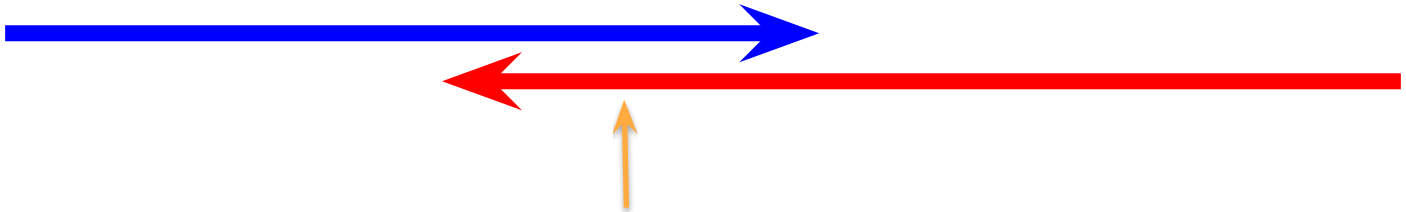


Overlapping reads

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

Goals of a Variant Caller

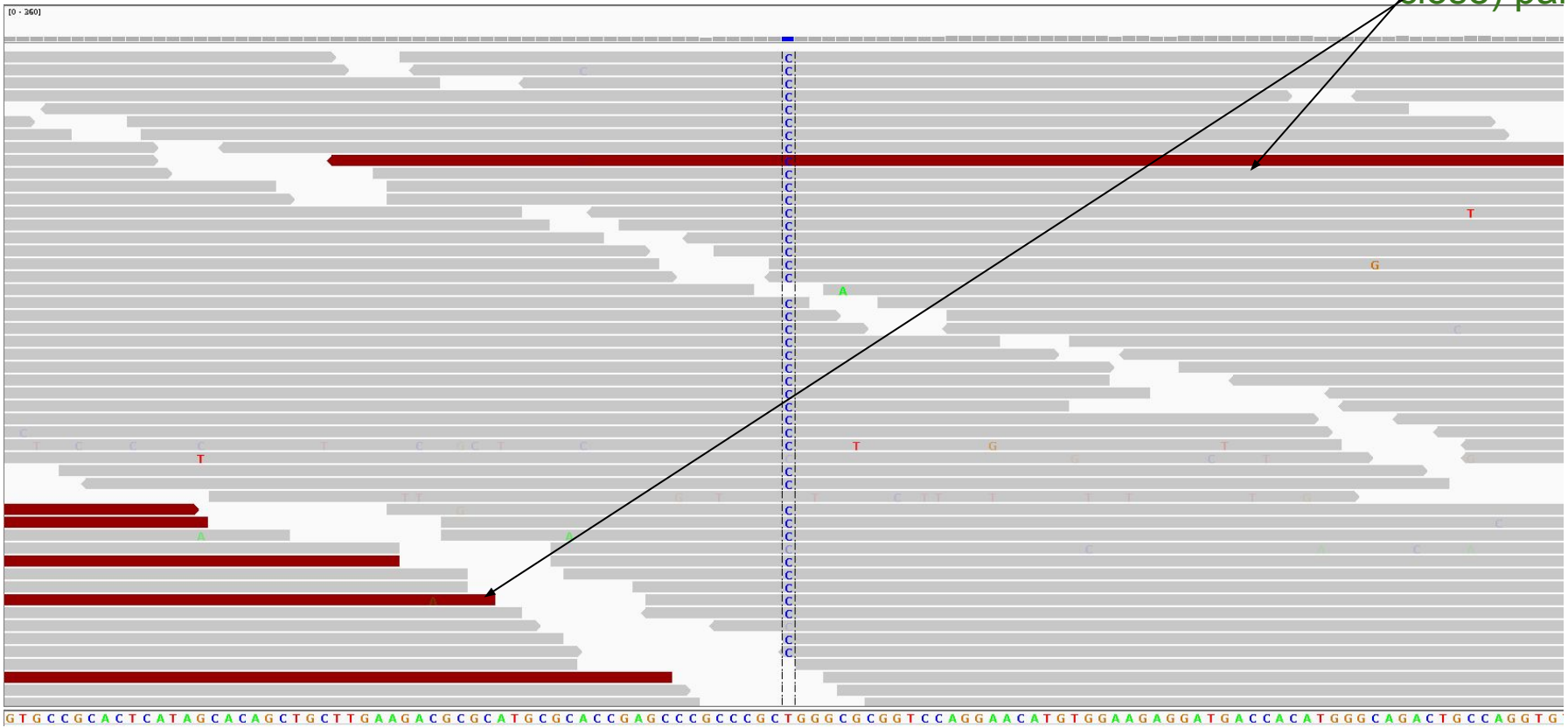
- Sensitively detect mutations
- Precisely detect mutations
  - Confounded by the error we just talked about
  - FDR must be very low as we're looking across a very large space!

Goals of a Variant Caller

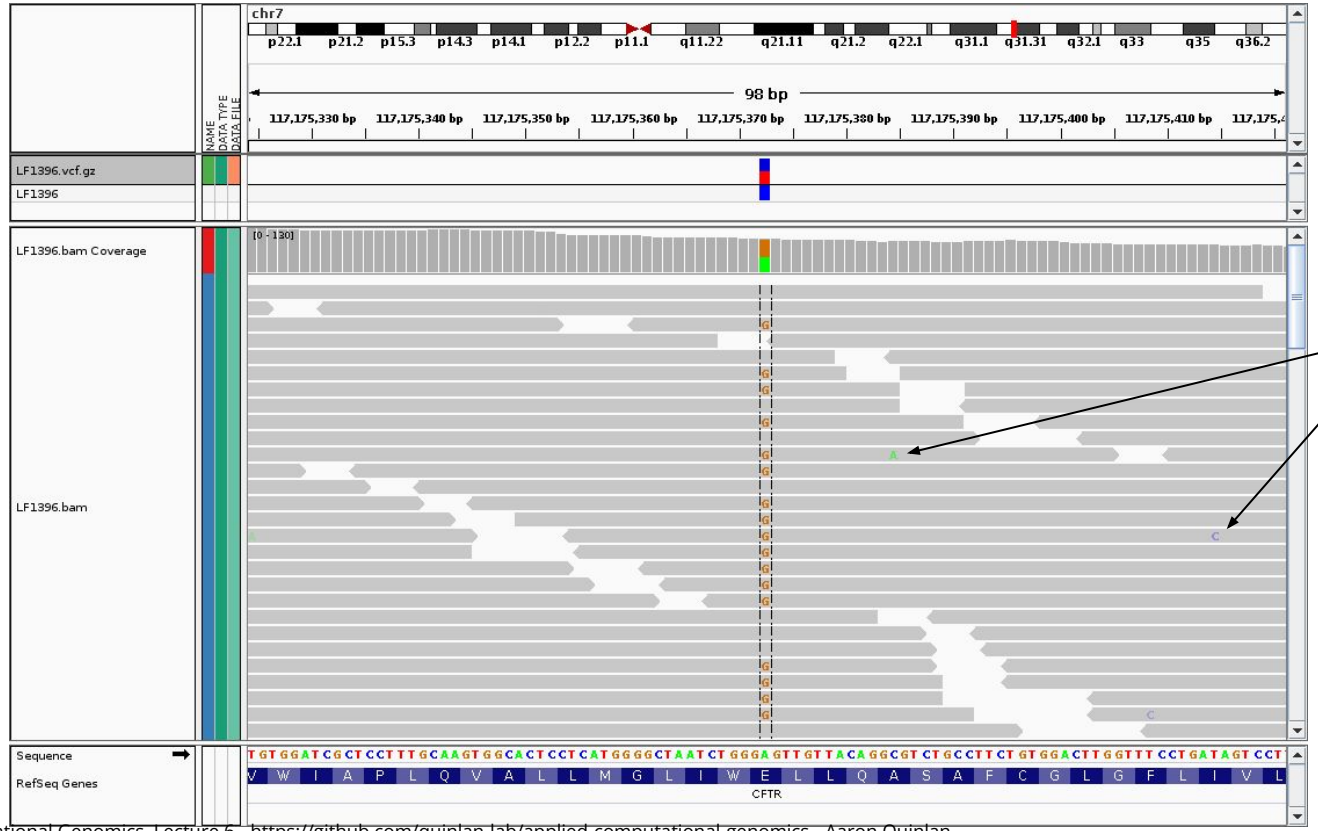
- Sensitively detect mutations
- Precisely detect mutations
  - Confounded by the error we just talked about
  - FDR must be very low as we're looking across a very large space!

    - **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://jchoigt.files.wordpress.com/2012/07/igv_e217g_snapshot.png

It is not always so easy

Random versus systematic error

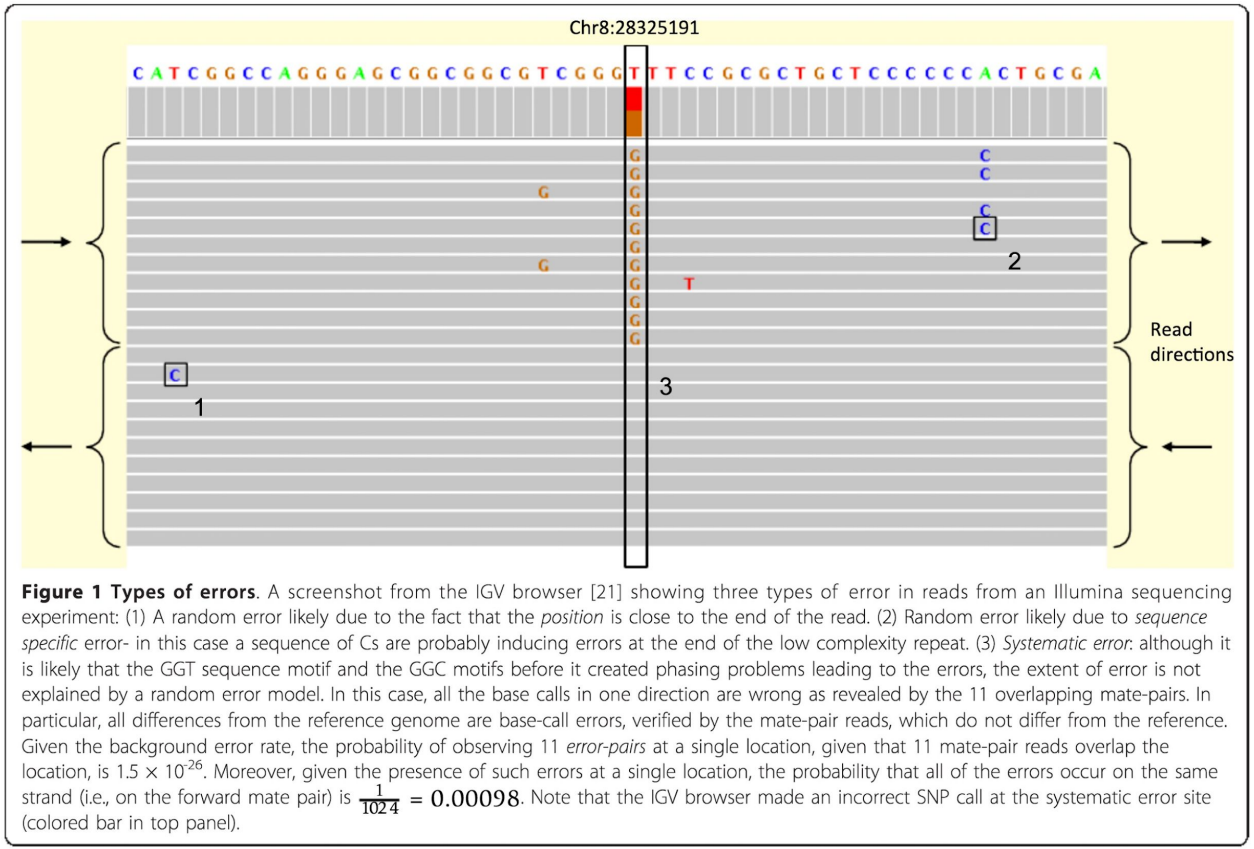
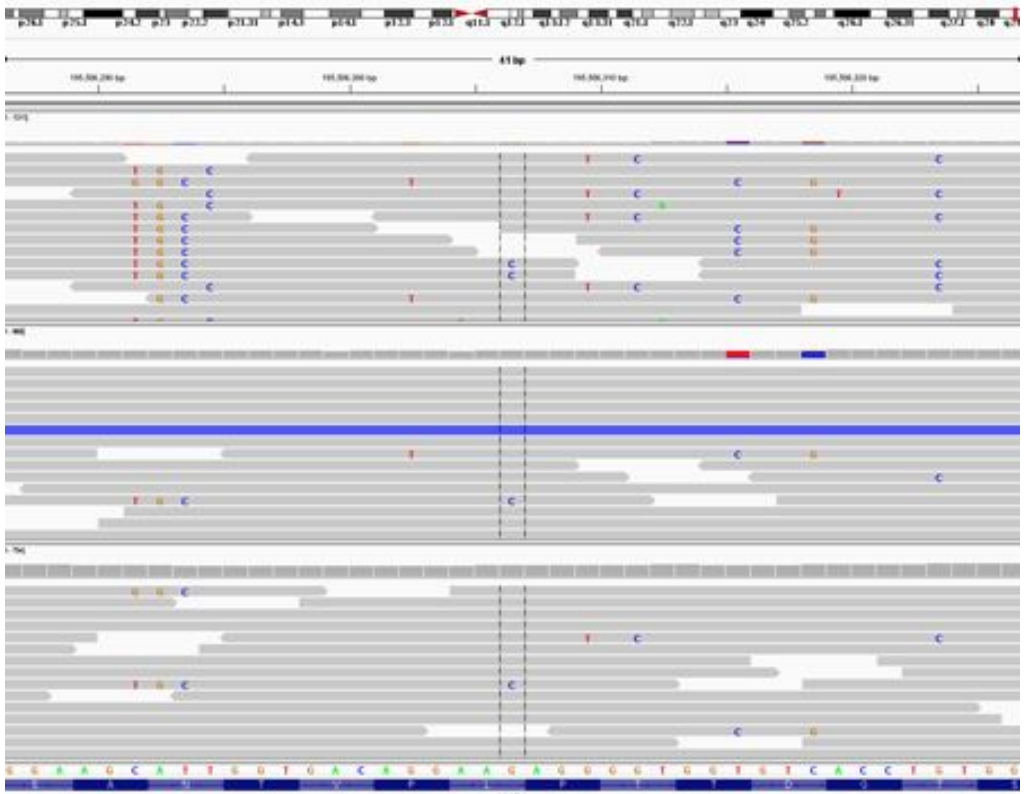


Figure 1 Types of errors. A screenshot from the IGV browser [21] showing three types of error in reads from an Illumina sequencing experiment: (1) A random error likely due to the fact that the *position* is close to the end of the read. (2) Random error likely due to *sequence specific* error- in this 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 | © Shyr et al.; licensee BioMed Central Ltd. 2014

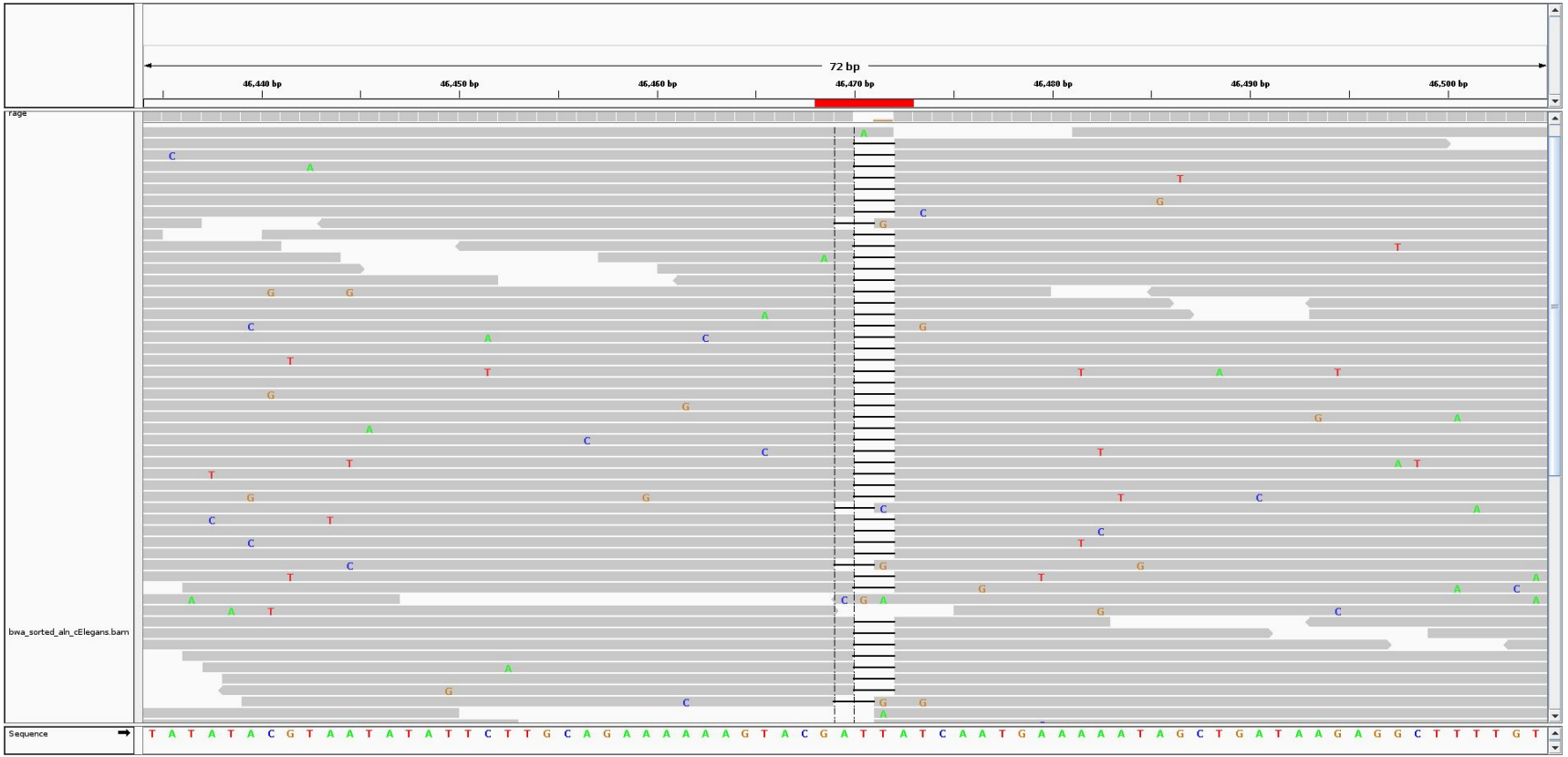
Received: 16 June 2014 | Accepted: 24 October 2014 | Published: 3 December 2014

Open Peer Review reports

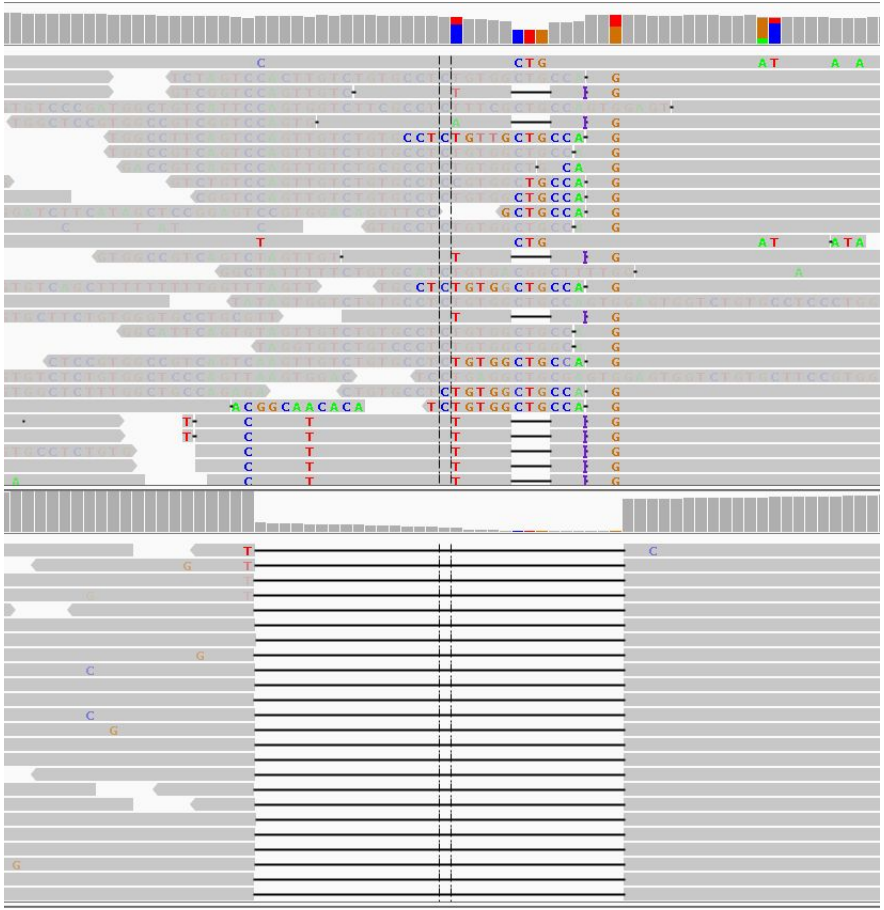
adapted from: Applied Computational Genomics, Lecture 6 - https://github.com/quinlan-lab/applied-computational-genomics - Aaron Quinlan

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

Calling INDELs is *much* harder than SNPs



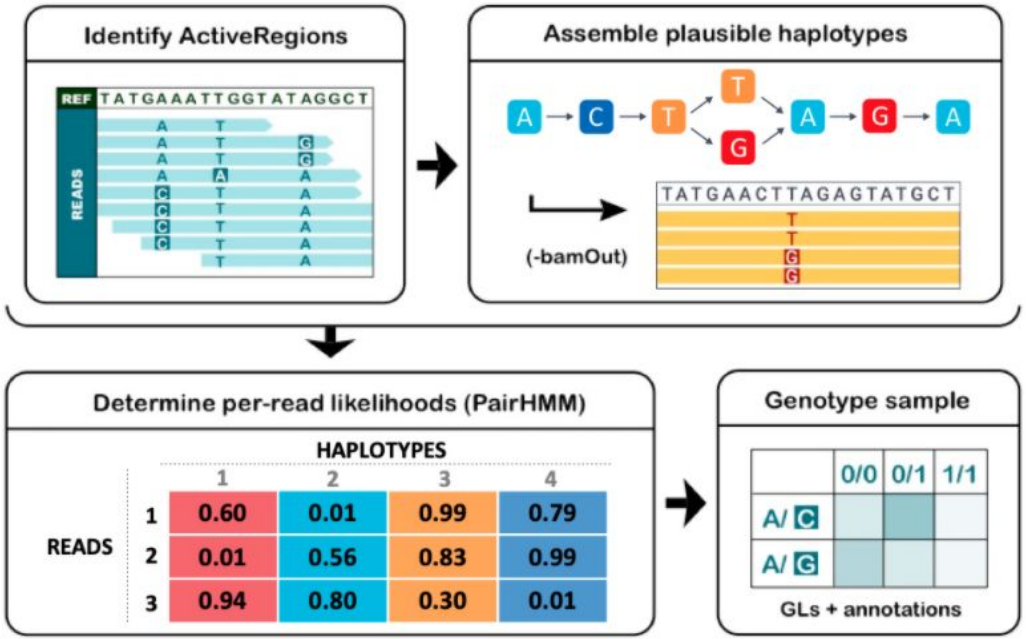
Indel "realignment"



https://github.com/mozack/abra

Germline SNV and Indel Calling

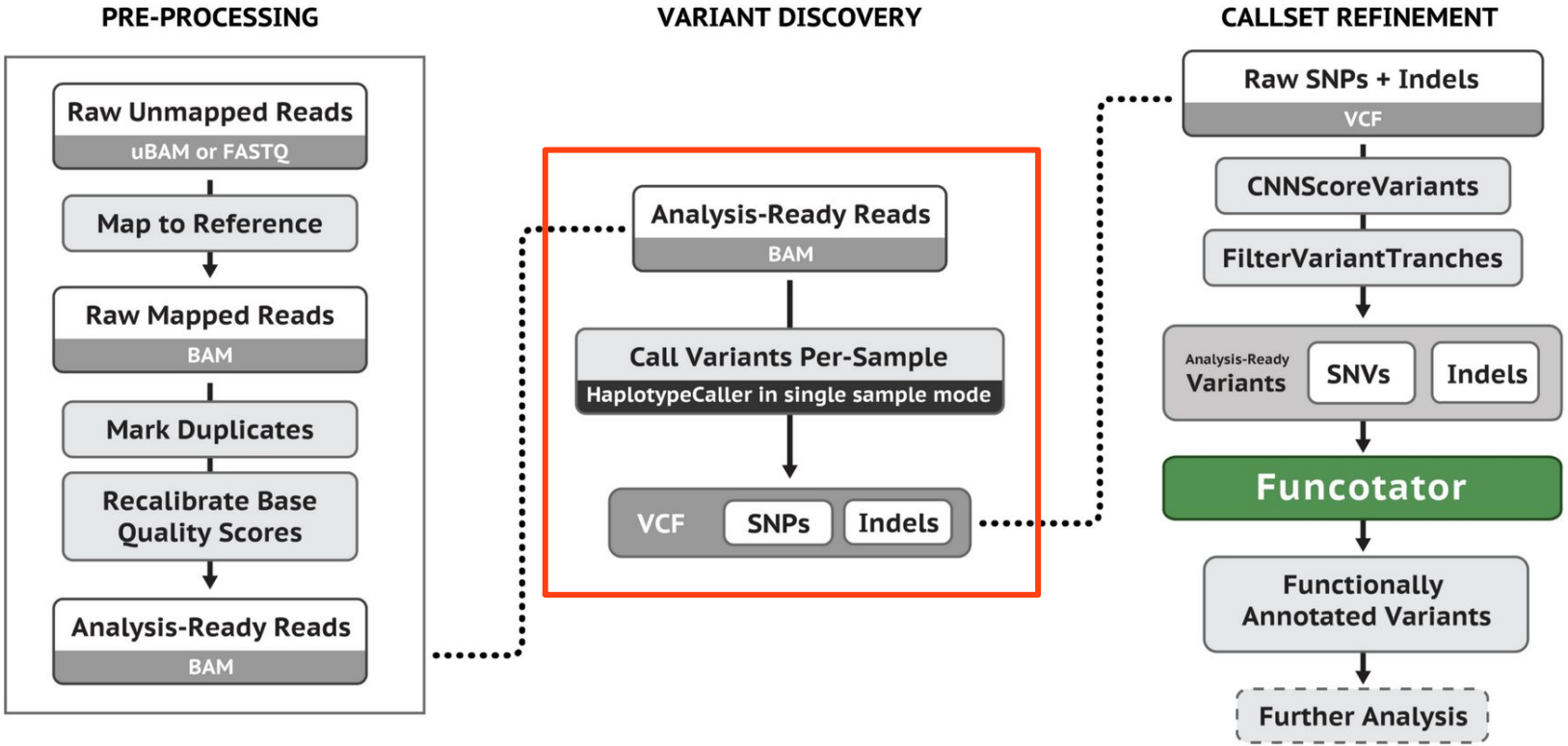
Call Genotypes Using GATK HaplotypeCaller



https://gatk.broadinstitute.org/hc/en-us/articles/360037225632-HaplotypeCaller

https://gatk.broadinstitute.org/hc/en-us/articles/360035531412

Main steps for Germline Single-Sample Data



VCF format

BIOINFORMATICS APPLICATIONS NOTE

Vol. 27 no. 15 2011, pages 2156–2158
doi:10.1093/bioinformatics/btr330

Sequence analysis

Advance Access publication June 7, 2011

The variant call format and VCFtools

Petr Danecek[1,†], Adam Auton[2,†], Goncalo Abecasis[3], Cornelis A. Albers[1], Eric Banks[4], Mark A. DePristo[4], Robert E. Handsaker[4], Gerton Lunter[2], Gabor T. Marth[5], Stephen T. Sherry[6], Gilean McVean[2,7], Richard Durbin[1,*] and 1000 Genomes Project Analysis Group[‡]

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge CB10 1SA, [2]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK, [3]Center for Statistical Genetics, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, [4]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02141, [5]Department of Biology, Boston College, MA 02467, [6]National Institutes of Health National Center for Biotechnology Information, MD 20894, USA and [7]Department of Statistics, University of Oxford, Oxford OX1 3TG, UK

Associate Editor: John Quackenbush

ABSTRACT

Summary: The variant call format (VCF) is a generic format for storing DNA polymorphism data such as SNPs, insertions, deletions and structural variants, together with rich annotations. VCF is usually stored in a compressed manner and can be indexed for fast data retrieval of variants from a range of positions on the reference genome. The format was developed for the 1000 Genomes Project, and has also been adopted by other projects such as UK10K, dbSNP and the NHLBI Exome Project. VCFtools is a software suite that implements various utilities for processing VCF files, including validation, merging, comparing and also provides a general Perl API.

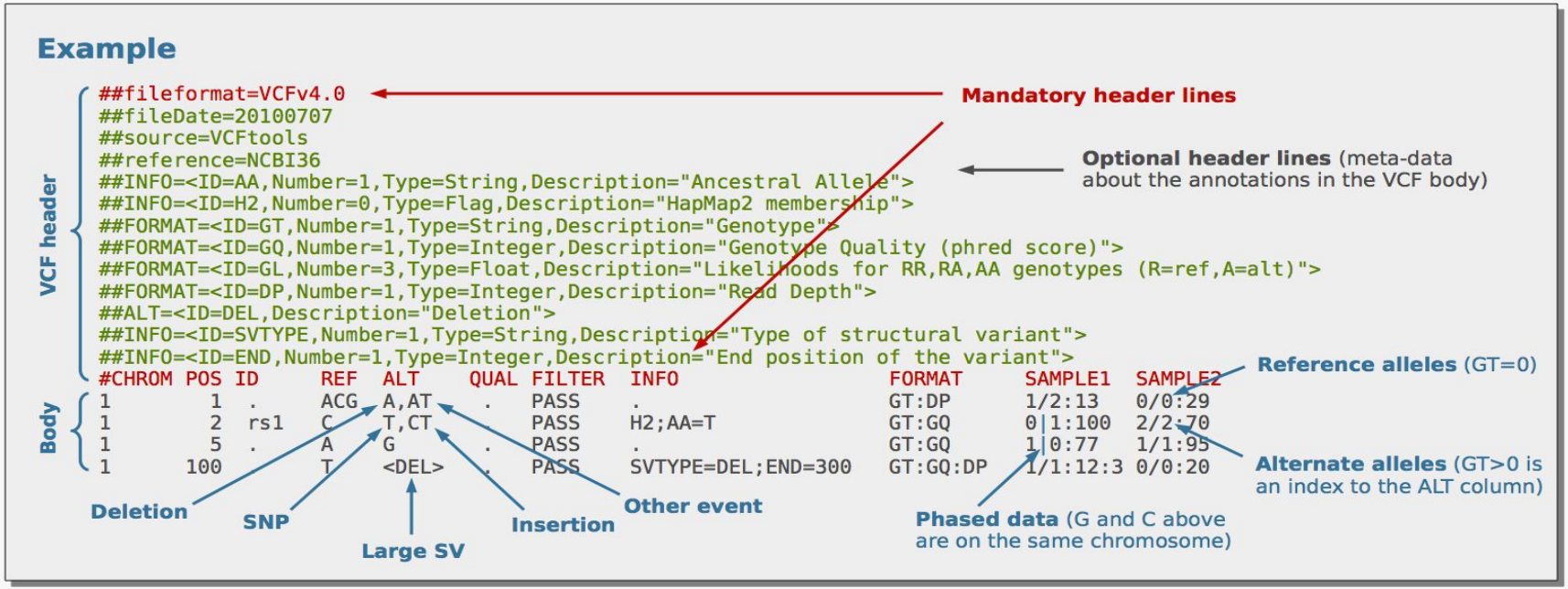
Availability: http://vcftools.sourceforge.net

Contact: rd@sanger.ac.uk

Although generic feature format (GFF) has recently been extended to standardize storage of variant information in genome variant format (GVF) (Reese *et al.*, 2010), this is not tailored for storing information across many samples. We have designed the VCF format to be scalable so as to encompass millions of sites with genotype data and annotations from thousands of samples. We have adopted a textual encoding, with complementary indexing, to allow easy generation of the files while maintaining fast data access. In this article, we present an overview of the VCF and briefly introduce the companion VCFtools software package. A detailed format specification and the complete documentation of VCFtools are available at the VCFtools web site.

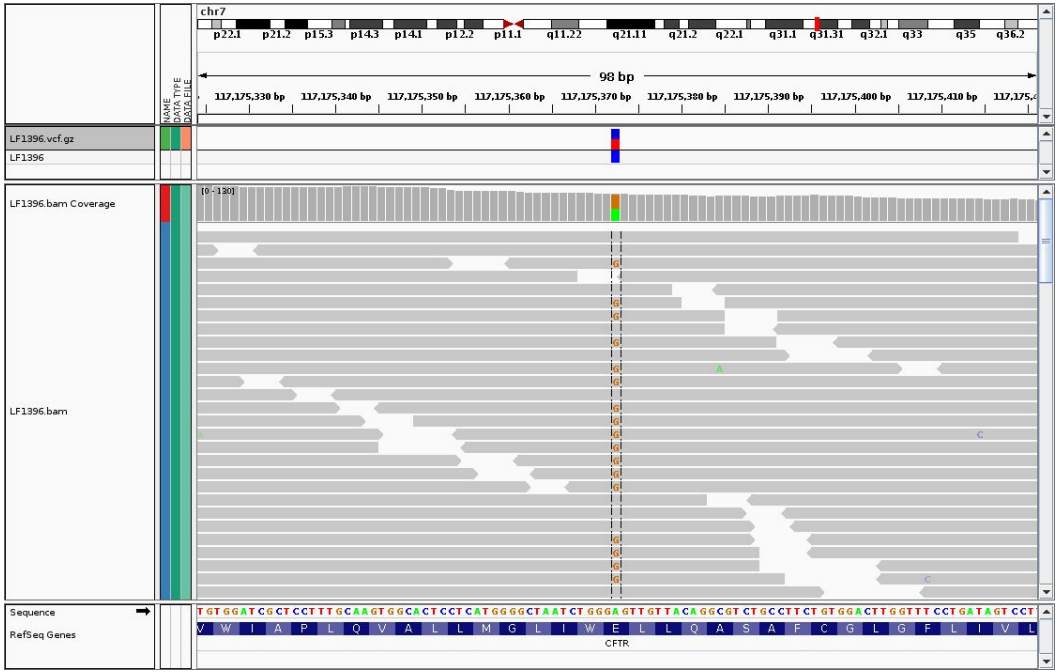
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3137218/pdf/btr330.pdf

VCF format



https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3137218/pdf/btr330.pdf

VCF format. A basic example



Heterozygous A/G.
REF allele is allele "0",
ALT is allele "1"

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	LF1396
chr7	117175373	.	A	G	90	PASS	AF=0.5	GT	0/1

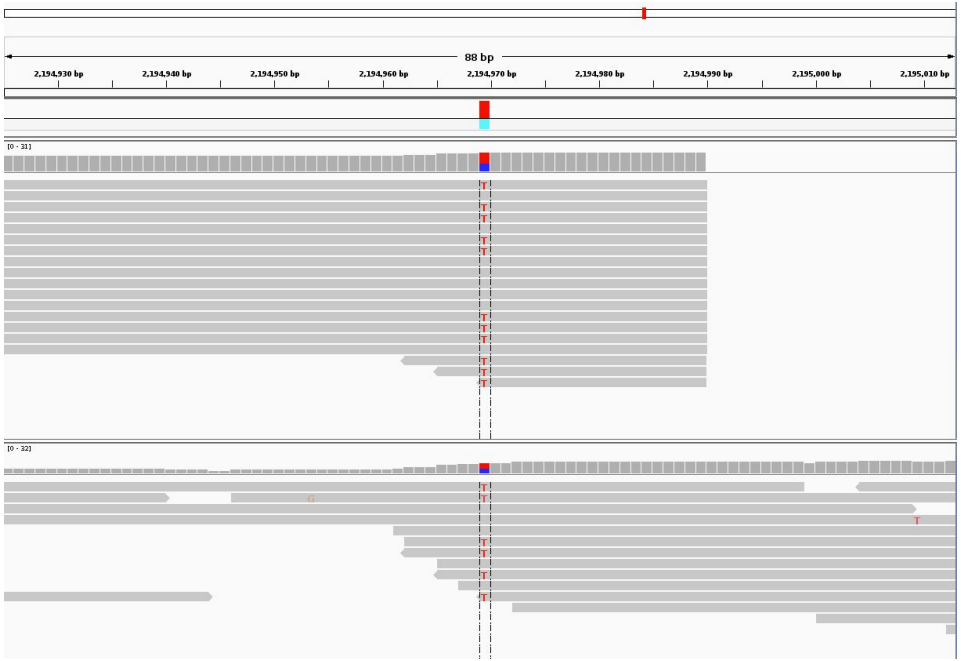
Genotypes

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	LF1396	
chr7	117175373	.	A	G	90	PASS	AF=0.0	GT	0/0	Hom. Ref.
chr7	117175373	.	A	G	90	PASS	AF=0.5	GT	0/1	Het.
chr7	117175373	.	A	G	90	PASS	AF=1.0	GT	1/1	Hom. Alt.
chr7	117175373	.	A	G	0	PASS	AF=0.0	GT	./.	Unknown

Multi-sample VCF

Mom

Kid



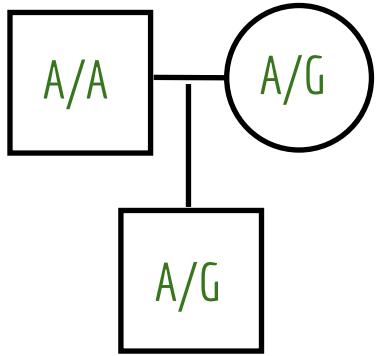
Heterozygous C/T

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	MOM	KID
chr7	2194169	.	C	T	210	PASS	AF=0.5	GT	0/1	0/1

Multi-sample VCF format example

```
##fileformat=VCFv4.3
##fileDate=20090805
##source=variantcallerXYZ
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
```

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	MOM	DAD	KID
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0|0:48:1:51,51	1|0:48:8:51,51	1/1:43:5:.,.
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017	GT:GQ:DP:HQ	0|0:49:3:58,50	0|1:3:5:65,3	0/0:41:3
20	1110696	rs6040355	A	G,T	67	PASS	NS=2;DP=10;AF=0.333,0.667;AA=T;DB	GT:GQ:DP:HQ	1|2:21:6:23,27	2|1:2:0:18,2	2/2:35:4
20	1234567	microsat1	GTC	G,GTCT	50	PASS	NS=3;DP=9;AA=G	GT:GQ:DP	0/1:35:4	0/2:17:2	1/1:40:3



https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3137218/pdf/btr330.pdf

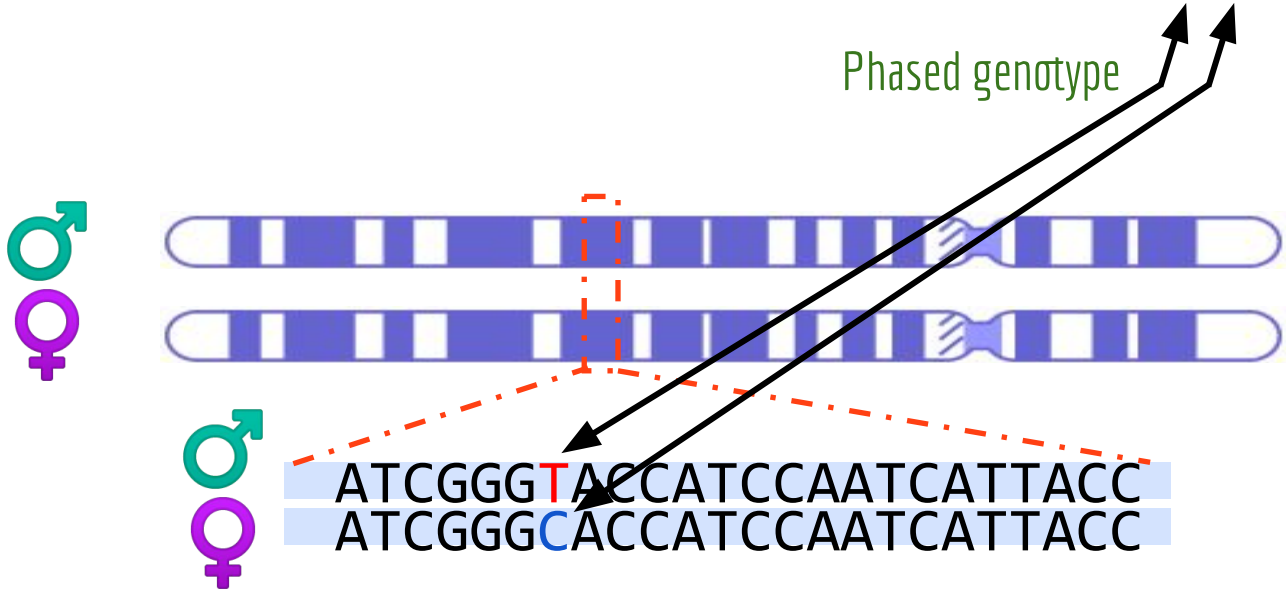
"Phased" genotypes

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	MOM	KID
chr7	2194169	.	C	T	210	PASS	AF=0.5	GT	0/1	0/1

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	MOM	KID
chr7	2194169	.	C	T	210	PASS	AF=0.5	GT	0|1	0|1

"Phased" genotypes distinguish haplotypes

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	MOM	KID
chr7	2194169	.	C	T	210	PASS	AF=0.5	GT	0|1	0|1

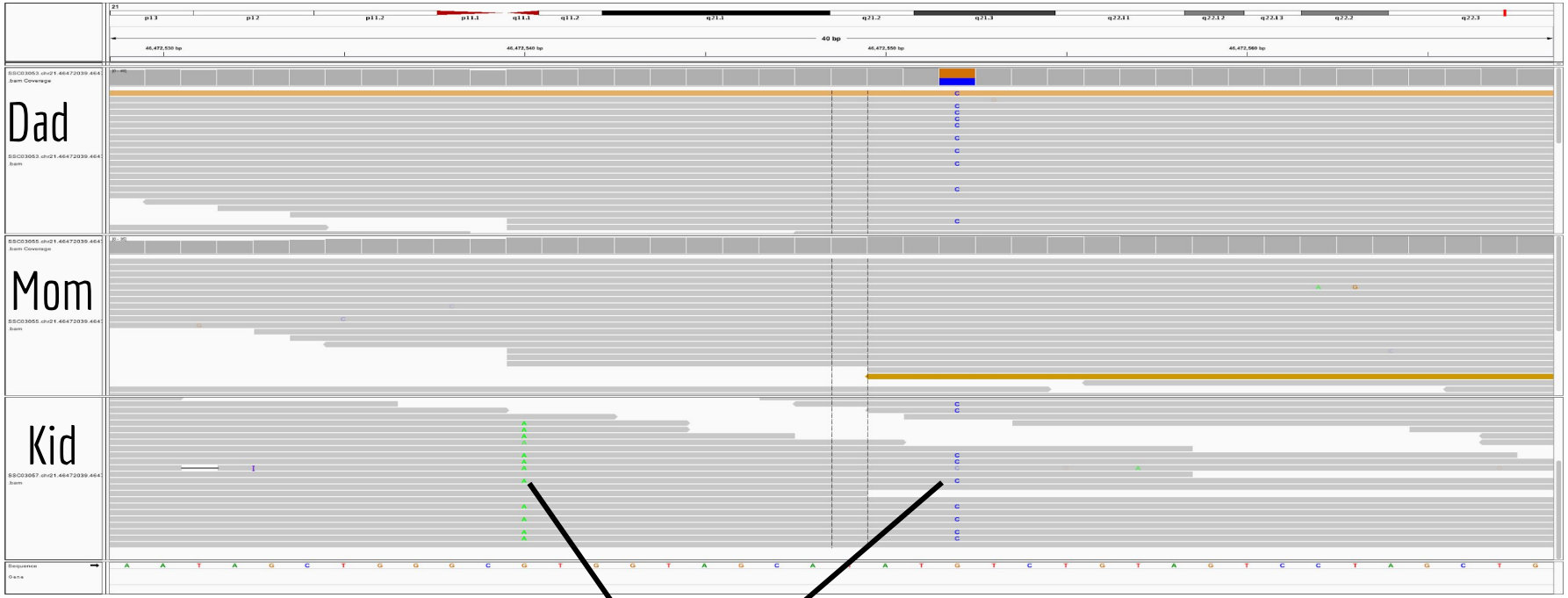


Phasing by inheritance

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	MOM	DAD	KID
chr1	100000	.	A	G	95	PASS	AF=0.33	GT	0/1	0/0	1|0
chr1	200000	.	T	C	99	PASS	AF=0.33	GT	0/0	0/1	0|1

This is an example of a compound heterozygote. Hets where the alt allele comes from different parents at two different loci

Read-backed phasing: evidence of haplotype in read or pair



Next Step: Filter Raw (Unfiltered) VCF file

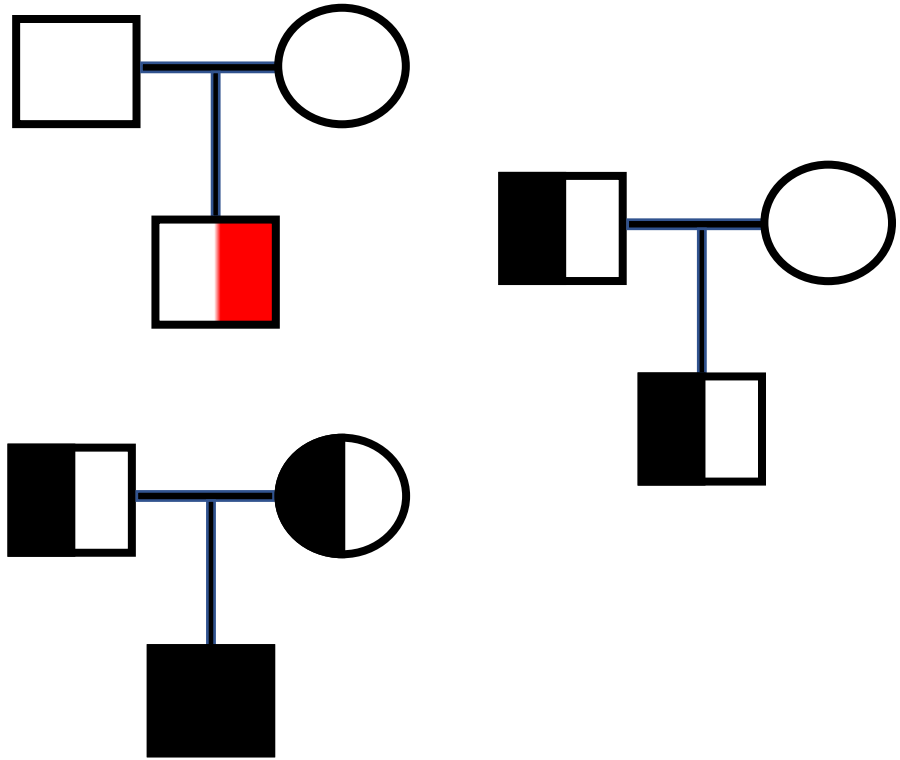
- **Problem**: **Too many Variants**
  - Non-related Human's differ by 0.1% --> 3 Million SNPs (haploid genome)

- Solution: Filter Low Quality and Common Variants
  - Common Variants > 5%
  - Low Quality
    - GQ < 20, DP < 8, MQ < 40 or GATK VQSR

VCF Visualization with IGV

Further Analysis after SNV Calling

- De-novo analysis with trios
- Rare transmitted analysis
- Compound het analysis



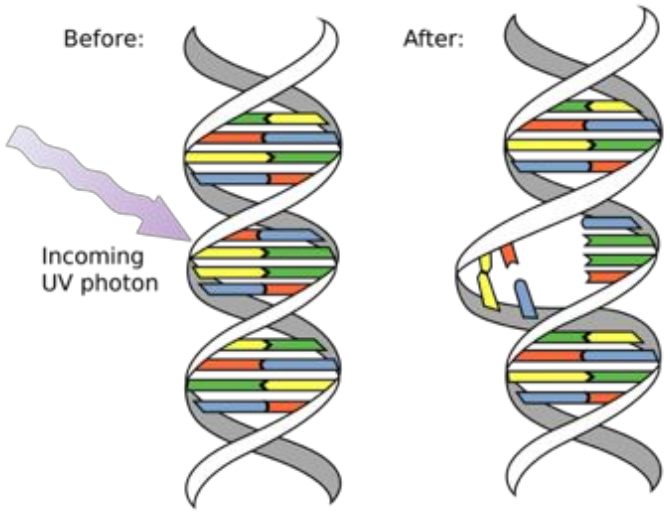
Somatic Mutation Calling

Cancer is a disease of the genome

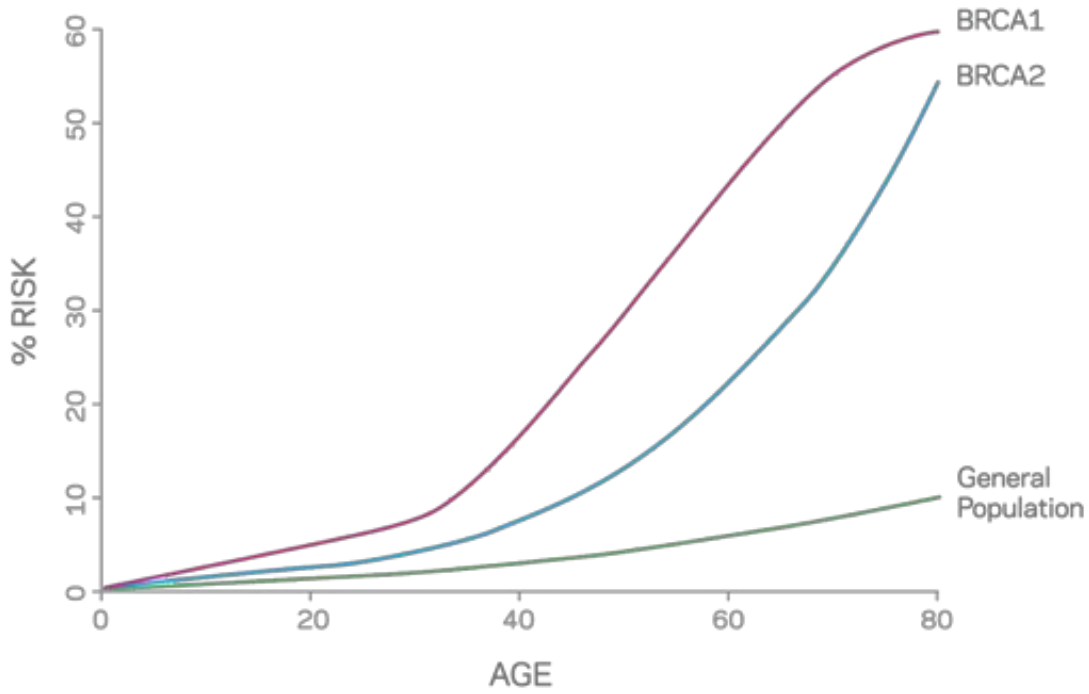
- Cancer is caused by **somatic** mutations

Cancer is a disease of the genome

- Cancer is caused by **somatic** mutations
- These mutations are introduced into the genome of a cell (errors in DNA copying, UV light, chemicals)
- Most cancers require around 3 driver mutations



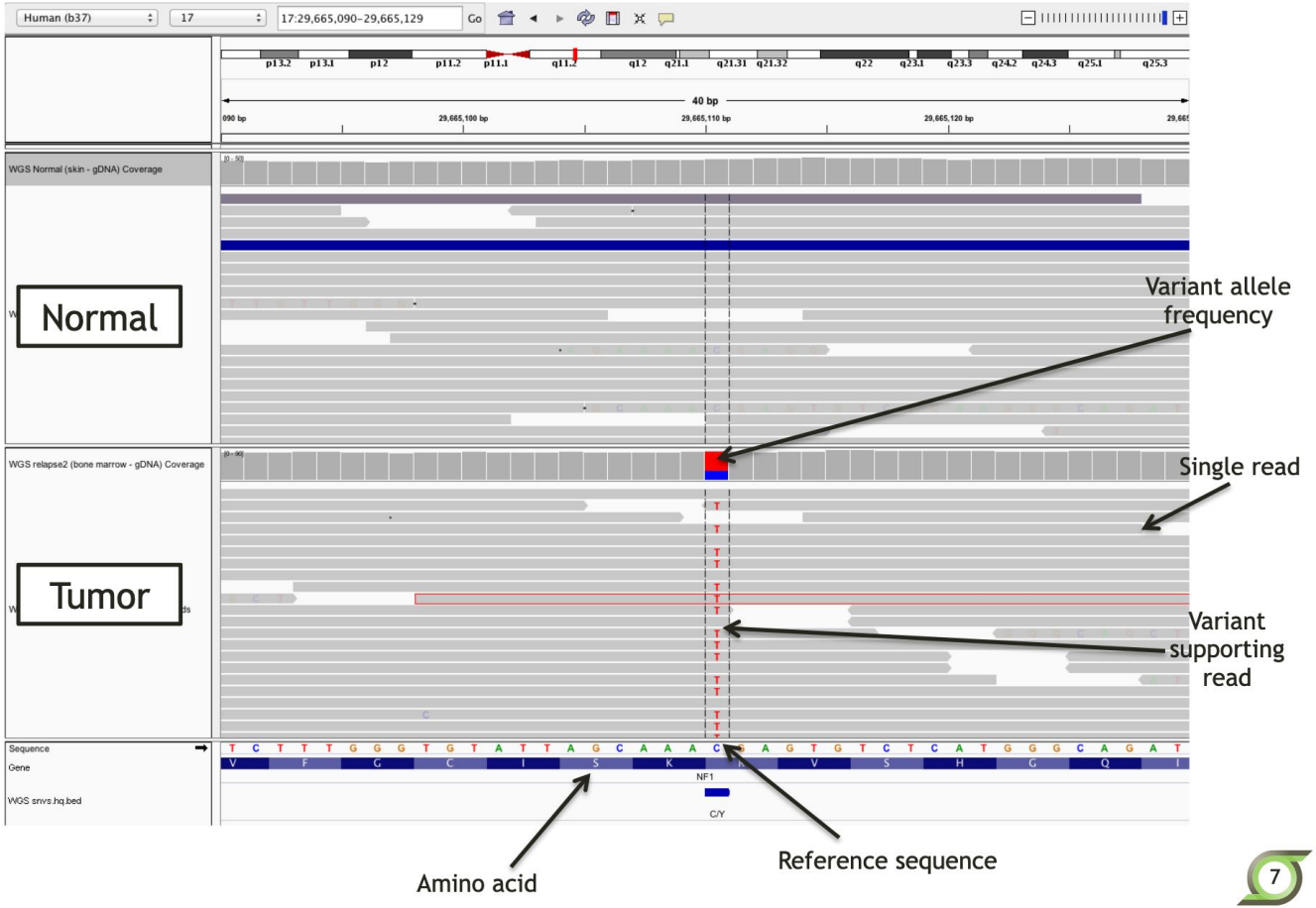
Germline Predisposition



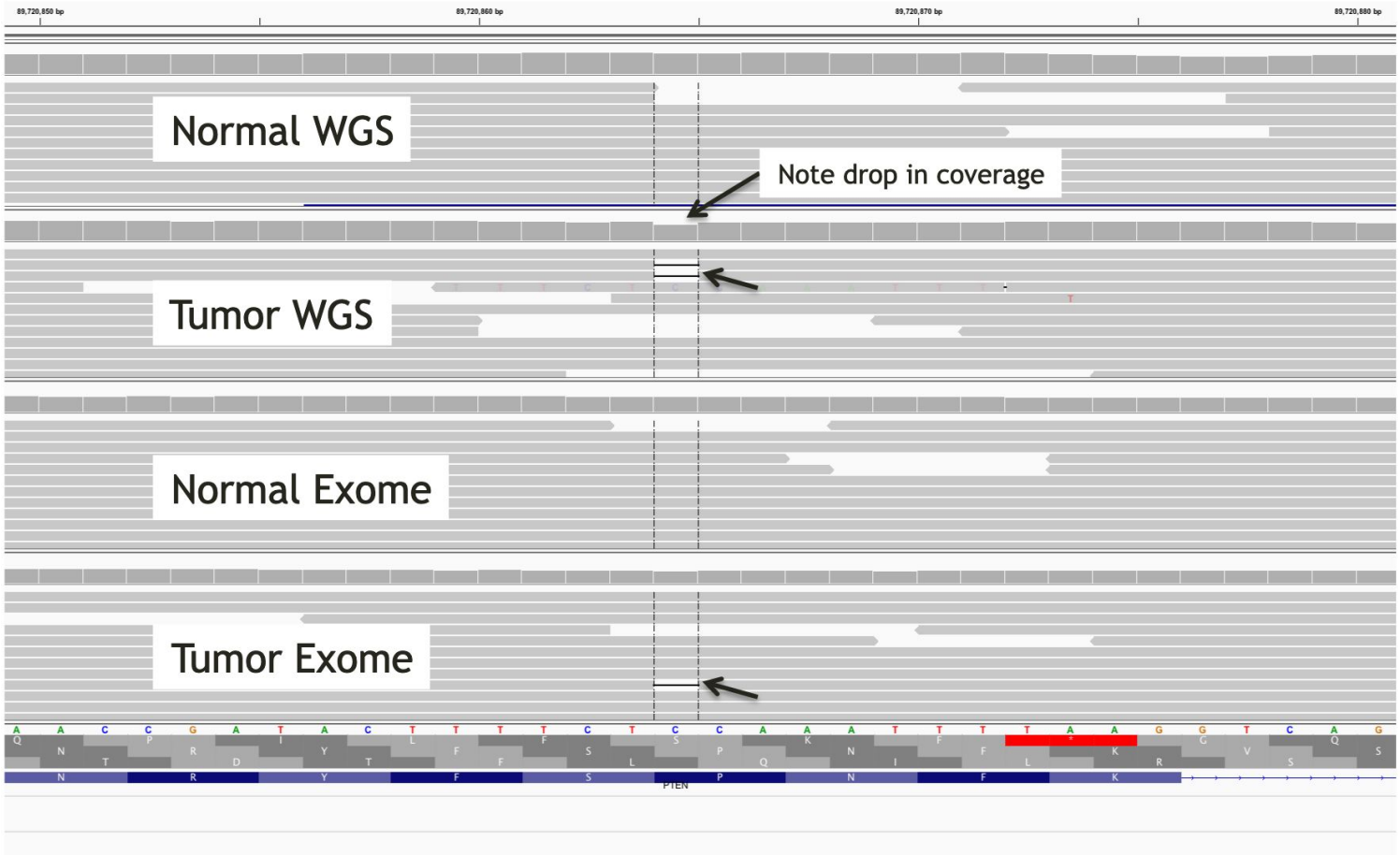
Cancer Sequencing

- In cancer, we have to (at least) double sequencing costs
- Uses both a tumor sample and a matched normal
- We compare them to find somatic mutations

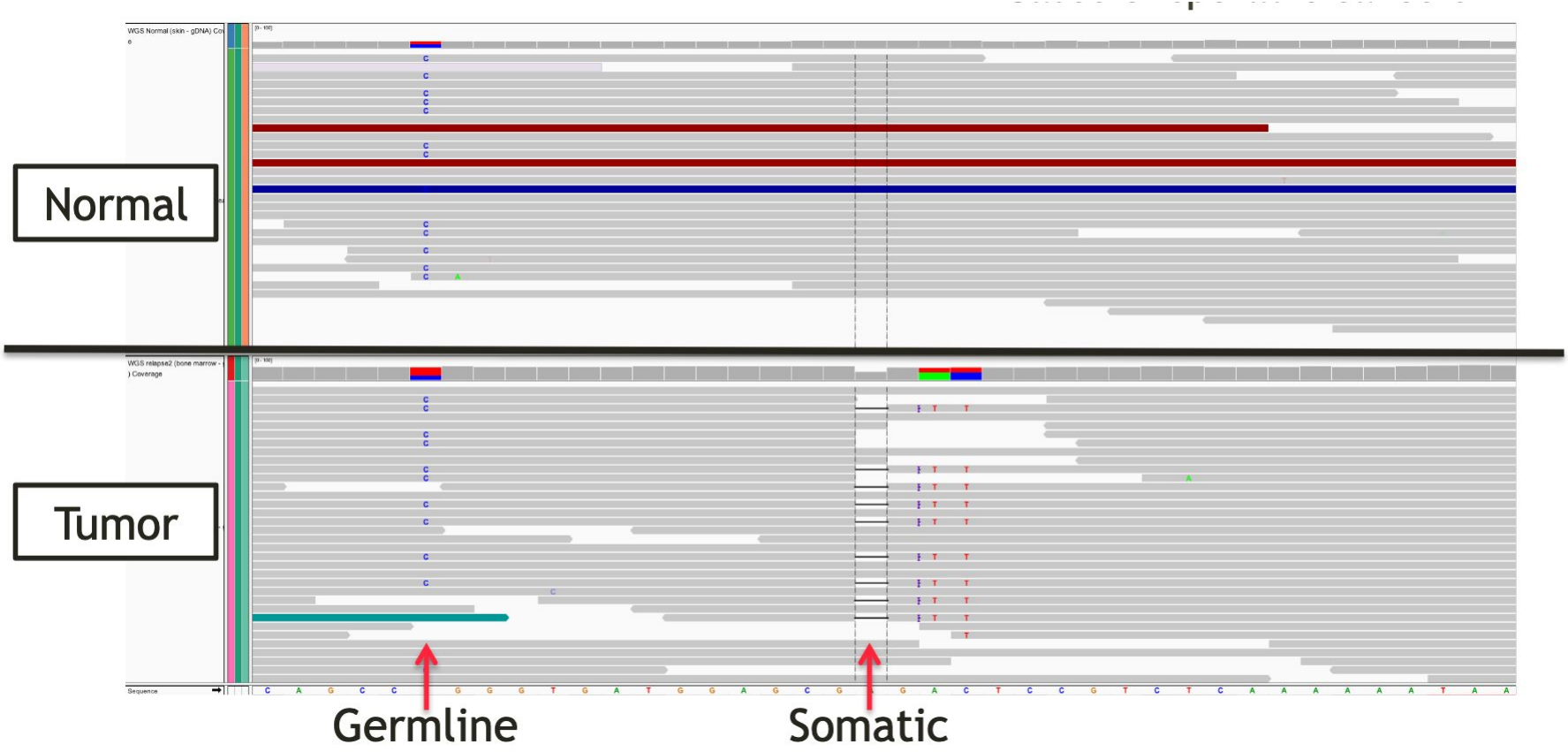
What do somatic variants look like?



Indels


89,720,850 bp
89,720,860 bp
89,720,870 bp
89,720,880 bp
Normal WGS
Note drop in coverage
Tumor WGS
Normal Exome
Tumor Exome
PTEN

Germline vs Somatic

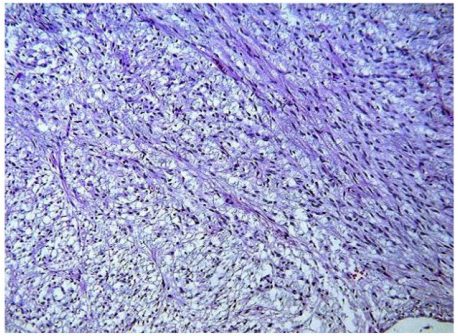


VAF = Variant reads / Total reads

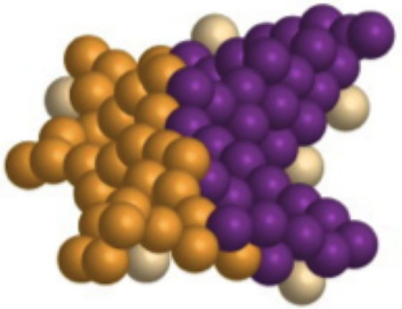
VAF = 0/20 = 0%

VAF = 14/20 = 70%

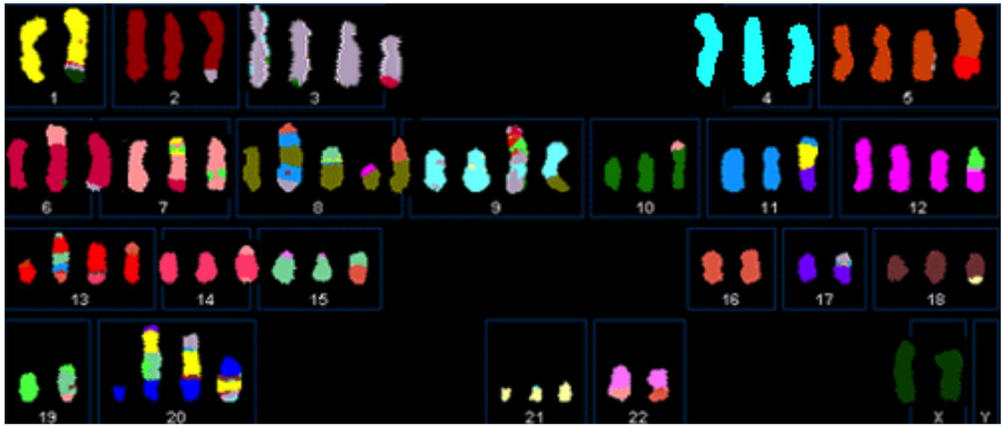
Tumors are often impure, heterogeneous, and aneuploid



Tumors are often impure
(contain normal cells)



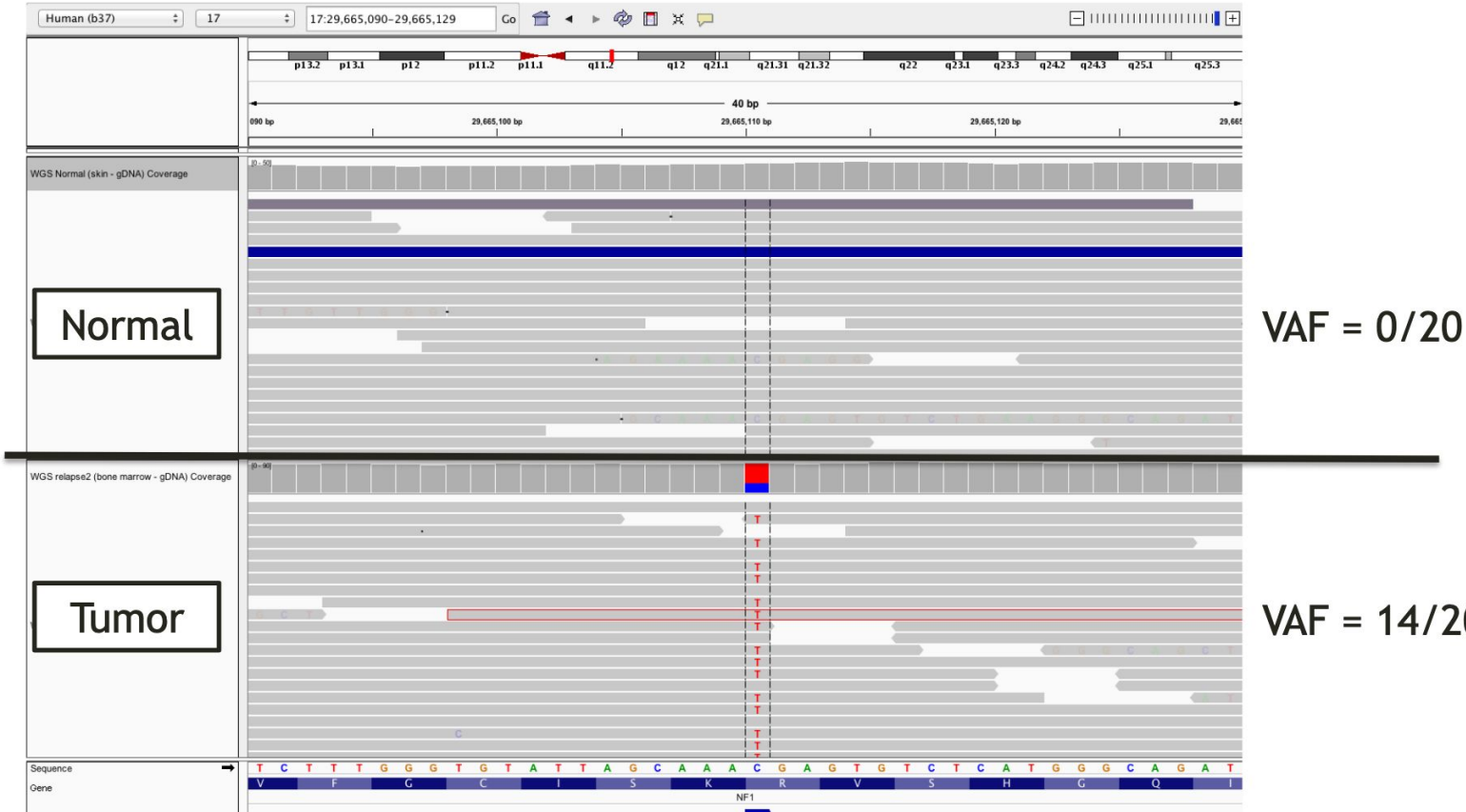
Tumors are often genetically
diverse collections of cells



Tumors may be aneuploid

How does purity influence VAF?

VAF = Variant reads / Total reads



VAF = 0/20 = 0%

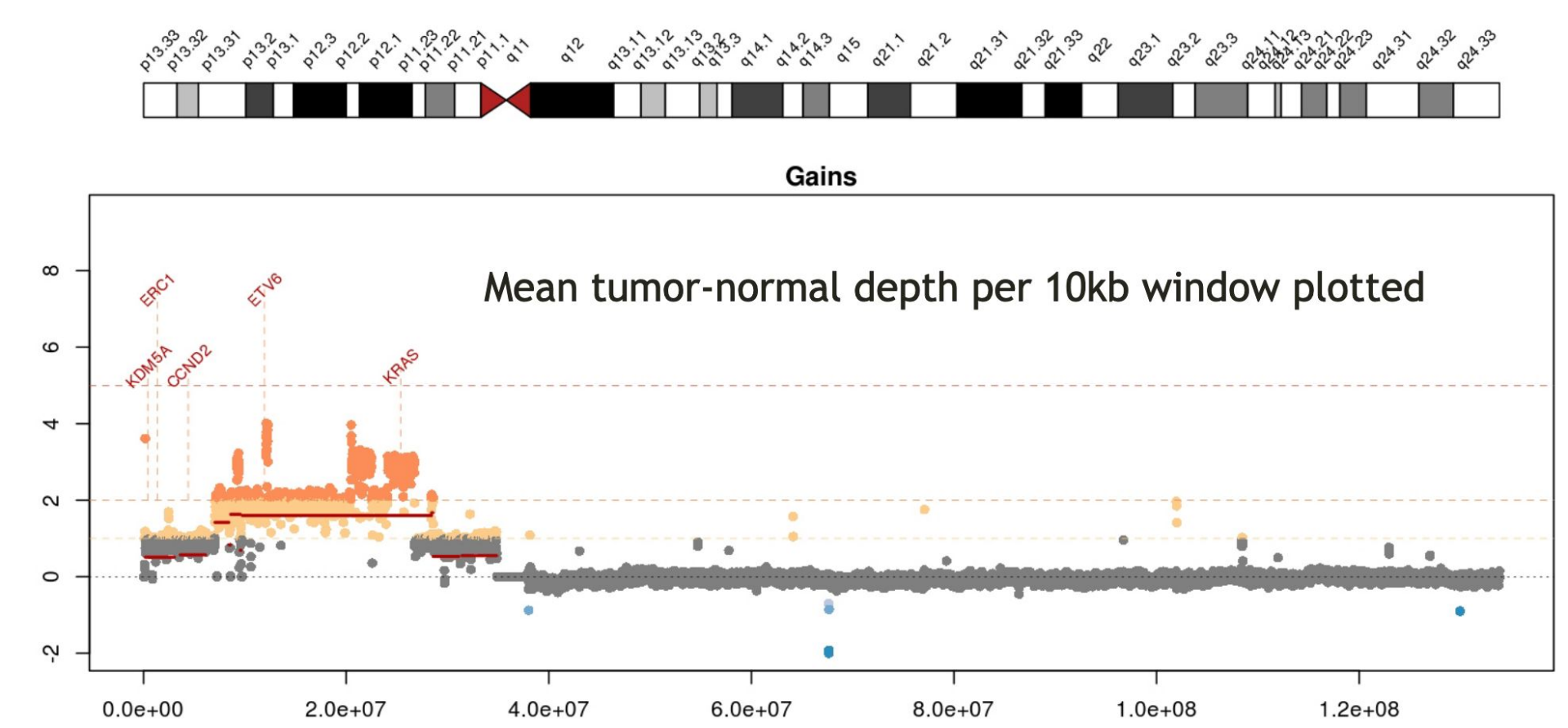
VAF = 14/20 = 70%

How does purity influence VAF?

VAF = Variant reads / Total reads

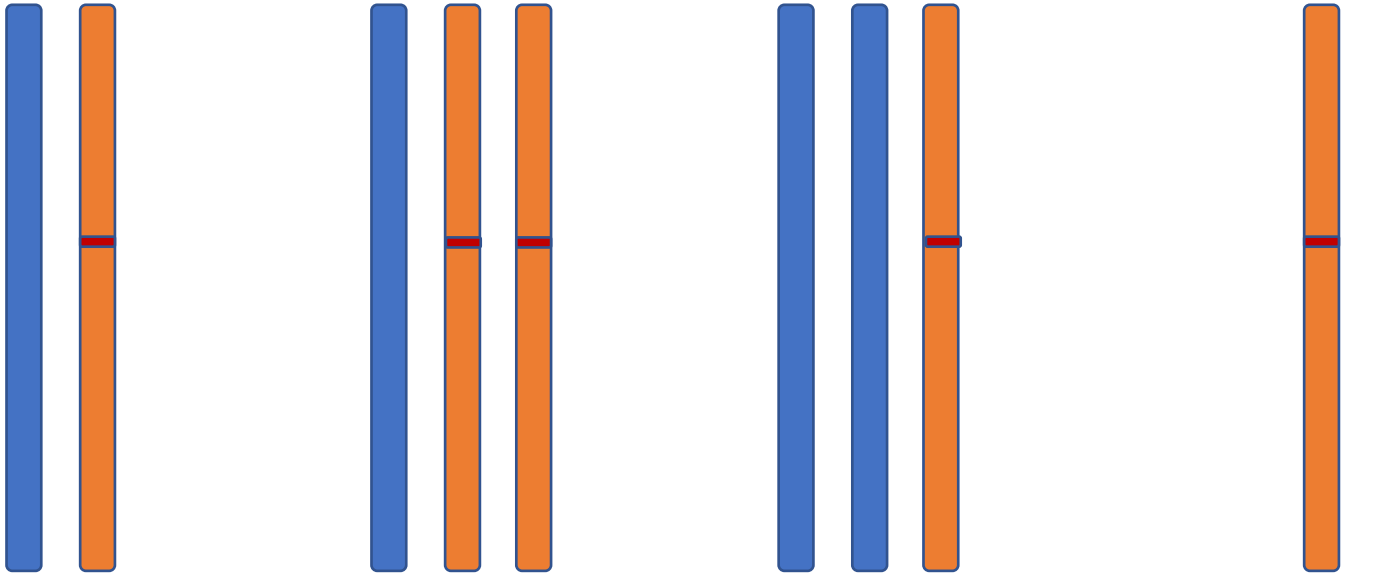
VAF = 0/20 = 0%

VAF = 5/20 = 25%



KRAS amplification in a metastatic breast cancer

How does copy number influence VAF?



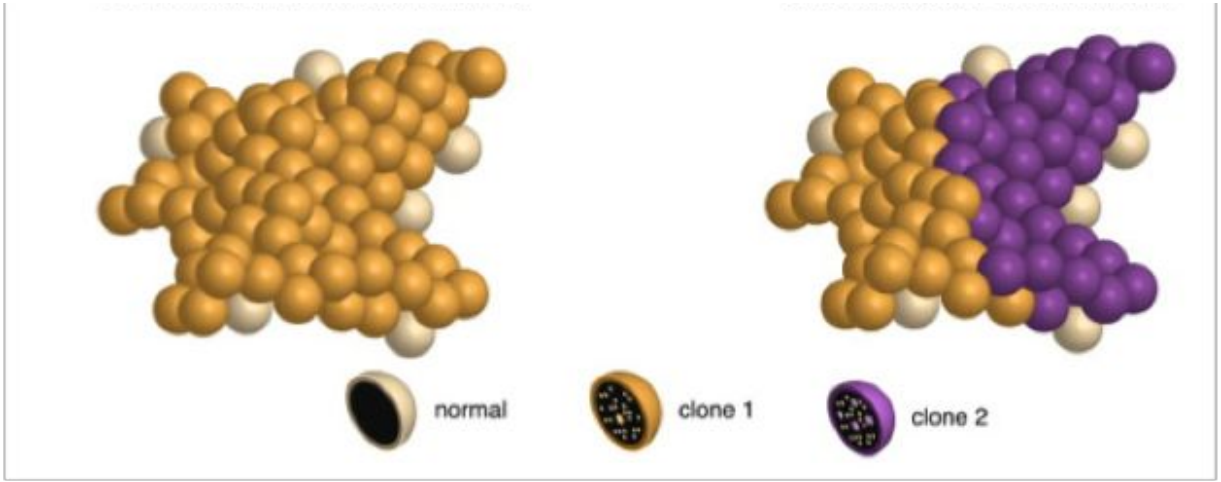
2 copies
50%
VAF

3 copies
66%
VAF

3 copies
33%
VAF

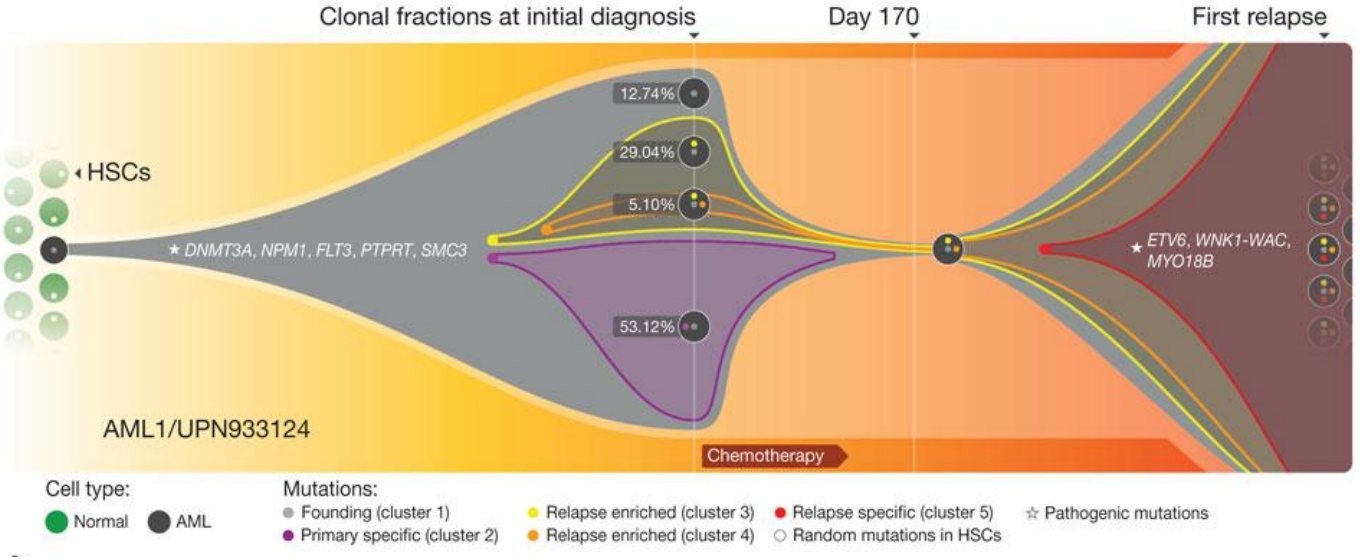
1 copy
100%
VAF

How does clonality influence VAF?



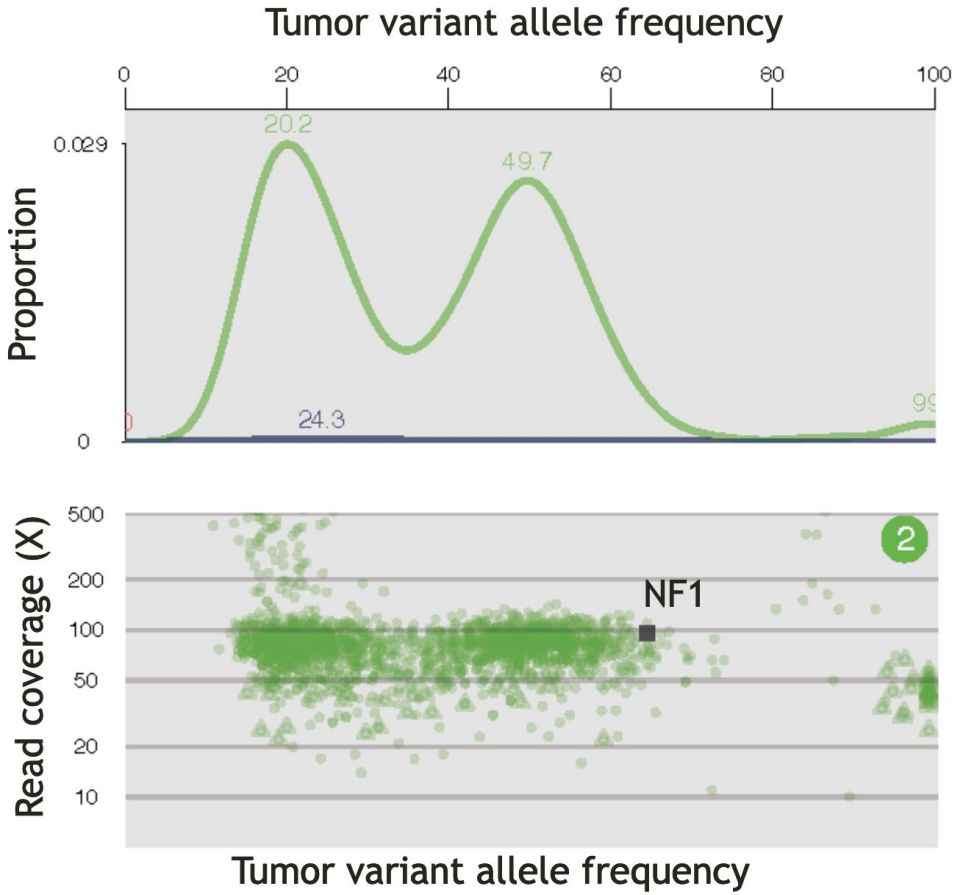
- Subclones contain genetically diverse populations of cells
- Evolution occurs at the molecular and cellular levels
- The growth rates for subclones are often different

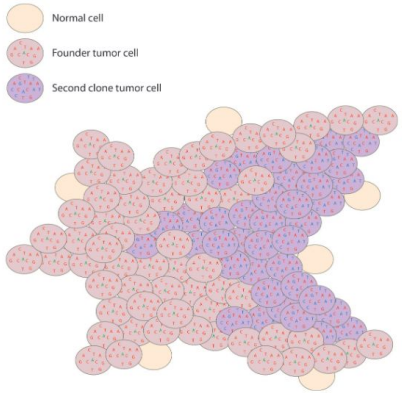
Clonal evolution in relapsed AML



Nature 481, 506–510
doi:10.1038/nature10738

Dominant clone vs. sub-clonal (and driver vs. passenger)



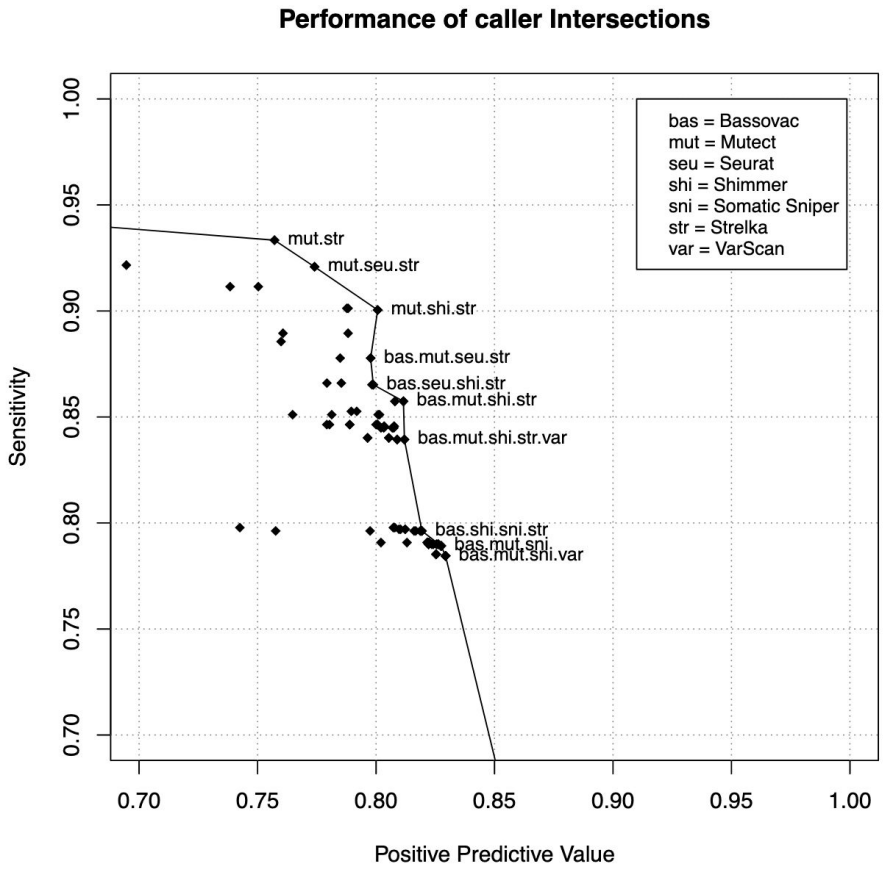


WGS data

Somatic variant calling is harder

- There are more factors to consider, a wider range of possibilities, and often, more sketchy samples

Use of multiple variant callers can improve sensitivity and accuracy



Somatic Variant detection workflow

- Step 1 – run the callers
  - Mutect, Strelka, Varscan, Pindel

Somatic Variant detection workflow

- Varscan
  - `java -jar VarScan.v2.3.6.jar somatic <(samtools mpileup … normal.bam) <(samtools mpileup … tumor.bam)`
- Mutect
  - `/gatk/gatk Mutect2 -O output.vcf -R ref.fa -I tumor.bam -tumor TUMORNAME -I nomal.bam -normal NORMALNAME -L interval.list`

Somatic Variant detection workflow

- Step 1 – run the callers
  - Mutect, Strelka, Varscan, Pindel

- Step 2 – Merge the VCFs
  - GATK - CombineVariants

Somatic Variant detection workflow

- Combine Variants

```
/usr/bin/java -jar /opt/GenomeAnalysisTK.jar -T
CombineVariants -genotypeMergeOptions PRIORITIZE
--rod_priority_list mutect,varscan -o combined.vcf.gz
--variant:mutect mutect_output.vcf --variant:varscan
varscan_output.vcf
```

Somatic Variant detection workflow

- Step 1 – run the callers
  - Mutect, Strelka, Varscan, Pindel
- Step 2 – Merge the VCFs
  - GATK - CombineVariants
- Step 3: Annotate and filter

Assignment

1) Take a normal bam file and call germline variants producing a VCF file

2) Take a tumor and matched normal bam file and identify somatic variants using multiple callers. Produce a merged VCF with your results