

# DNA Sequencing and Data Analysis

---

Prof Noam Shomron  
Hadas Volkov

Lecture 8, December 23, 2022

# DNA Sequencing and Data Analysis

---

Friday 8:45 AM to 11:15 AM  
Arazi-Ofer Building, C.L03

[nshomron@gmail.com](mailto:nshomron@gmail.com)

[hadas.volkov@post.runi.ac.il](mailto:hadas.volkov@post.runi.ac.il)

# DNA Sequencing and Data Analysis

---

Variant Calling  
SAM & VCF File Formats  
IGV

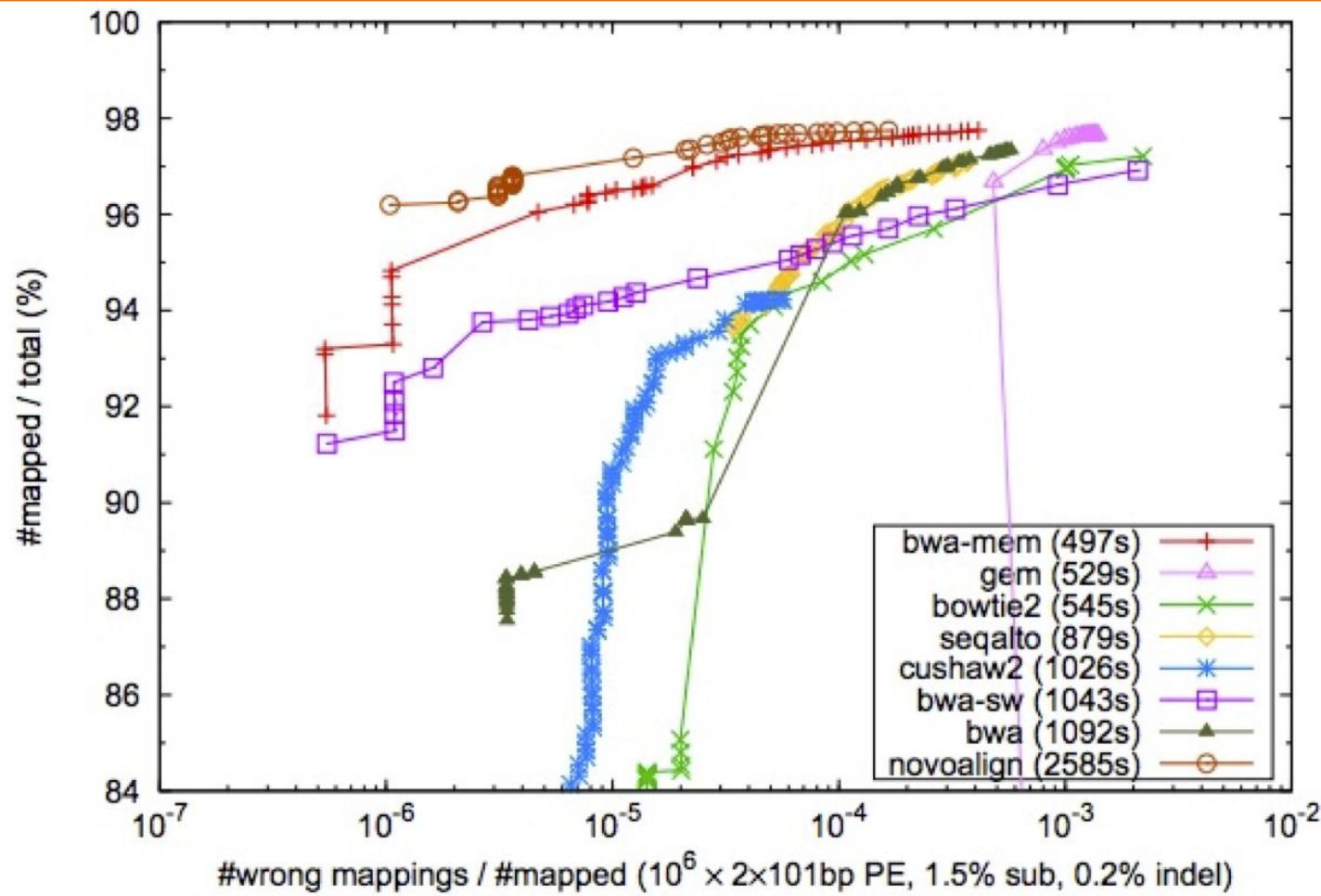
Class	Title	Content/assignments	Activity, location
1, 4.11	Introduction to Cells and DNA	Basic knowledge of biology	In the lecture hall, Noam
2, 11.11	DNA Sequencing past and present	Basic knowledge of molecular DNA	In the lecture hall, Noam
3, 18.11	Genomics technologies	DNA, RNA, technologies	In the lecture hall, Noam
4, 25.11	Introduction to Bioinformatics challenges in reading DNA	Focus on three methods: WES/WGS, RNA-seq, cell-free DNA	In the lecture hall, Noam
5, 2.12	Modern DNA Sequencing, 2nd wave File Formats, tools.	Analysis approaches for WES/WGS, RNA-seq, cell-free DNA	In the lecture hall, Hadas and Noam
6, 9.12	De novo Shotgun Assembly	The algorithms and methods behind the assembly problem	In computer class, Hadas and Noam
7, 16.12	Sequence Mapping and Alignment	The algorithms behind mapping and alignment, fast and heuristics	In computer class, Hadas and Noam
8, 23.12	<b>Variant Calling and Somatic Variant Analysis</b>	<b>The bioinformatics behind discovery of novel mutations in cancer</b>	<b>In computer class, Hadas and Noam</b>
9, 30.12	Nanopore data analysis introduction and class activity	The bioinformatics behind Nanopore analysis, activity on computers	In computer class, Hadas and Noam
10, 6.1	Practice molecular biology techniques	Pipetting, transferring small amounts of fluids, running a dry Nanopore experiment	In biology class, Meitar and Noam
11, 13.1	Nanopore DNA sequencing	Nanopore DNA sequencing, experimental run	In biology class, Meitar, Hadas, Assaf
12, 20.1	Nanopore data analysis	Nanopore DNA analysis, experimental run	In computer class, Hadas and Noam
13, 27.1	Nanopore data analysis and presentations	Groups present their results	In the lecture hall, Hadas and Noam

# Aligners Comparison

---

<u>Aligner</u>	<u>Index</u>	<u>Applications</u>	<u>Availability</u>
BWA-mem	Burrows-Wheeler	DNA, SE, PE, SV	open-source
Bowtie2	Burrows-Wheeler	DNA, SE, PE, SV	open-source
Novoalign	hash-based	DNA, SE, PE	proprietary
TopHat	Burrows-Wheeler	RNA-seq	open-source
STAR	hash-based (reads)	RNA-seq	open-source
GSNAP	hash-based (reads)	RNA-seq	open-source

# Aligners Comparison



# BWA-MEM Workflow

---

This takes a long time, but you do it once

Output is in SAM format.  
Use multiple threads if you have a computer with multiple CPUs.

Create BWT of reference genome.

```
$ bwa index grch38.fa
```



Align paired-end FASTQ to BWT index.

```
$ bwa mem -t 16 grch38.fa 1.fq 2.fq > sample.sam
```

# FASTQ to SAM

## Unaligned Sample Data (FASTQ)



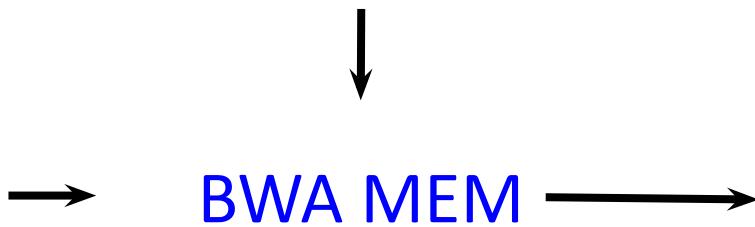
@seq1  
ATTCGAAACA...  
+  
DDED88(999...  
@seq2  
CCCCGTTCA...  
+  
AAC887BBAC...

## Reference genome (FASTA)

```
>chr1
TACCTCCAGGGGGCATCCTCCCCCAATTG
AAACACAATCGTAGCCCTGGCACTACCTATG
TGTGTCAATTGGAGAGAGAGAGATTACGAA
AAAAAAAGTCTGGACTCAAATAGGATAACACACA
TTCGGCTACAGATAACCAAAAAAAAAAAAAAAA
AAATTTCACCATTGAGGCACCACCTCTCGT
CGCTGCGTCGCTCTGCTCGCTCGCTAAAAA
TTCGCGCAATACATTGGCTACAGATAACCAAA
```

## Aligned Sample Data (SAM)

```
seq1 99 1 3666901 60 149M =
3666935 185 ATTGAAACA... DDED88(999
MC:Z:151M MD:Z:149 RG:Z:15-0017315_1
NM:i:0 MQ:i:60 AS:i:149 XS:i:44
seq2 147 1 3666935 60 151M =
3666901 -185 CCCC GTTCA...
AAC887BBAC... MC:Z:149M MD:Z:151
RG:Z:15-0017315_1 NM:i:0 MQ:i:60
AS:i:151 XS:i:59
```



# Sequence Alignment and Mapping (SAM)

---

## Sequence analysis

### The Sequence Alignment/Map format and SAMtools

Heng Li<sup>1,†</sup>, Bob Handsaker<sup>2,†</sup>, Alec Wysoker<sup>2</sup>, Tim Fennell<sup>2</sup>, Jue Ruan<sup>3</sup>, Nils Homer<sup>4</sup>, Gabor Marth<sup>5</sup>, Goncalo Abecasis<sup>6</sup>, Richard Durbin<sup>1,\*</sup> and 1000 Genome Project Data Processing Subgroup<sup>7</sup>

<sup>1</sup>Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, UK, <sup>2</sup>Broad Institute of MIT and Harvard, Cambridge, MA 02141, USA, <sup>3</sup>Beijing Institute of Genomics, Chinese Academy of Science, Beijing 100029, China, <sup>4</sup>Department of Computer Science, University of California Los Angeles, Los Angeles, CA 90095,

<sup>5</sup>Department of Biology, Boston College, Chestnut Hill, MA 02467, <sup>6</sup>Center for Statistical Genetics, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA and <sup>7</sup><http://1000genomes.org>

Received on April 28, 2009; revised on May 28, 2009; accepted on May 30, 2009

Advance Access publication June 8, 2009

Associate Editor: Alfonso Valencia

**Table 1.** Mandatory fields in the SAM format

No.	Name	Description
1	QNAME	Query NAME of the read or the read pair
2	FLAG	Bitwise FLAG (pairing, strand, mate strand, etc.)
3	RNAME	Reference sequence NAME
4	POS	1-Based leftmost POSition of clipped alignment
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIGAR	Extended CIGAR string (operations: M I D N S H P)
7	MRNM	Mate Reference NaMe ('=' if same as RNAME)
8	MPOS	1-Based leftmost Mate POSition
9	ISIZE	Inferred Insert SIZE
10	SEQ	Query SEQuence on the same strand as the reference
11	QUAL	Query QUALity (ASCII-33=Phred base quality)

# Sequence Alignment and Mapping (SAM)

---

What critical information do we need for sequence alignments?

# SAM Format

---

Col #	Name	Meaning	Example
1	QNAME	Read or Pair name	HWI:ST156_1:278:1:1058:4544:0
2	FLAG	Bitwise FLAG	<i>soon!</i>
3	RNAME	Reference sequence name	chr1
4	POS	1-based alignment start coordinate	8,724,005
5	MAPQ	Mapping quality	60
6	CIGAR	Extended CIGAR string	<i>soon!</i>
7	MRNM	If paired, the mate's reference seq.	chr1
8	MPOS	If paired, the mate's alignment start	8,724,505
9	ISIZE	If paired, the insert size	562
10	SEQ	The sequence of the query/mate	ACAAATTTCAG...
11	QUAL	The quality string for the query/mate	HHH\$^^%\$\$...\$
12	OPT	Optional Tags	XA:i:2, MD:Z:OT34G15

# SAM Format

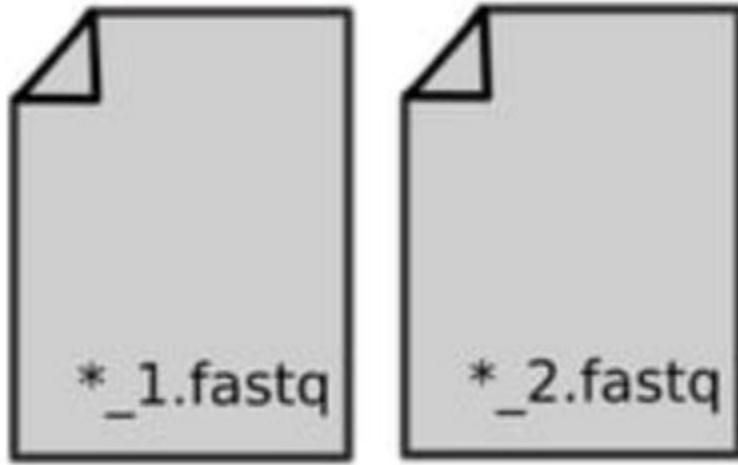
# SAM Flag

base2	base10	base16	Meaning	Applies to:
00000000001	1	0x0001	The read originated from a paired sequencing molecule	Both
00000000010	2	0x0002	The read is mapped in a <b>proper</b> pair	Pairs only
00000000100	4	0x0004	The query sequence itself is unmapped	Both
00000001000	8	0x0008	The query's mate is unmapped	Pairs only
00000010000	16	0x0010	Strand of the query (0 for forward; 1 for reverse strand)	Both
00000100000	32	0x0020	Strand of the query's mate	Pairs only
00001000000	64	0x0040	The query is the first read in the pair	Pairs only
00010000000	128	0x0080	The read is the second read in the pair	Pairs only
00100000000	256	0x0100	The alignment is not primary	Both
01000000000	512	0x0200	The read fails platform/vendor quality checks	Both
10000000000	1024	0x0400	The read is either a PCR duplicate or an optical duplicate	Both

# SAM Flag

---

**read paired**



**read paired**

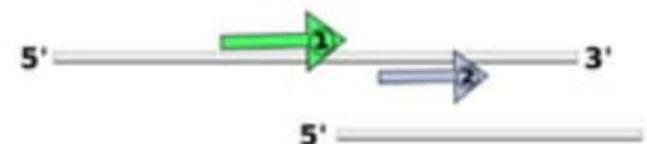
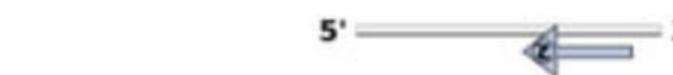


# SAM Flag

read mapped  
in proper pair



read mapped  
in proper pair



# SAM Flag

---

**read unmapped**  **read unmapped**



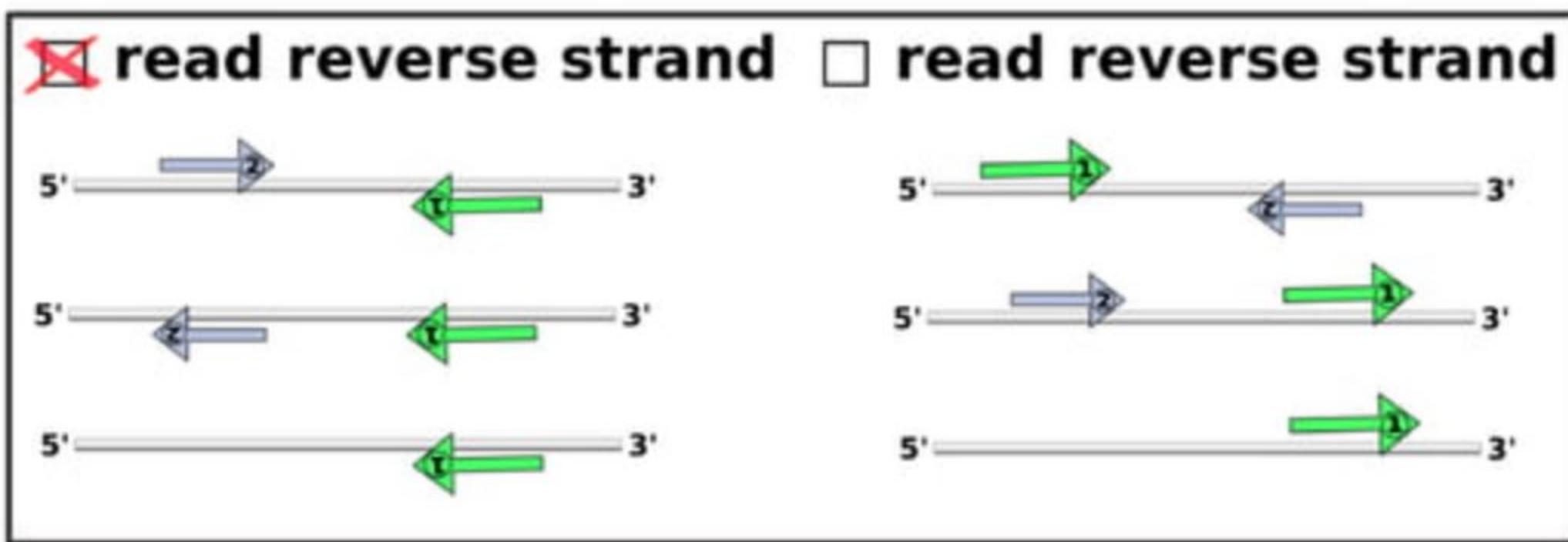
# SAM Flag

---

mate unmapped  mate unmapped

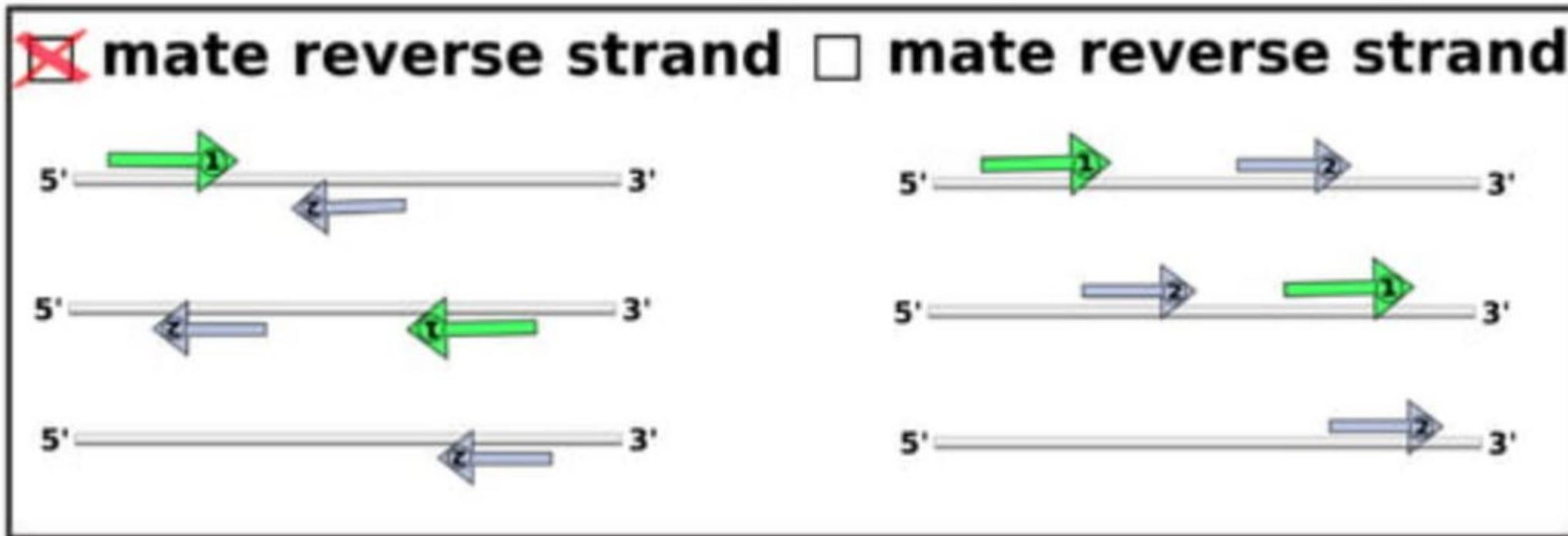


# SAM Flag



# SAM Flag

---



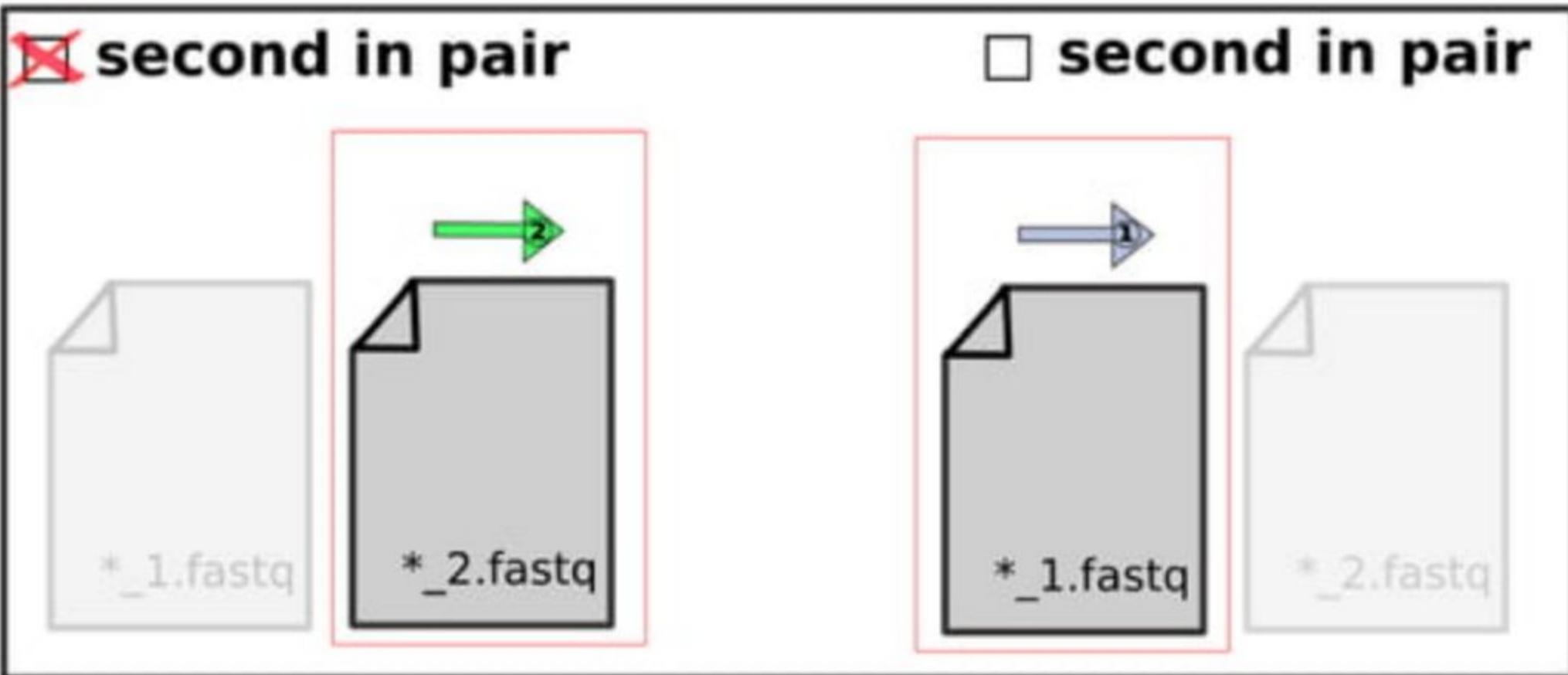
# SAM Flag

---

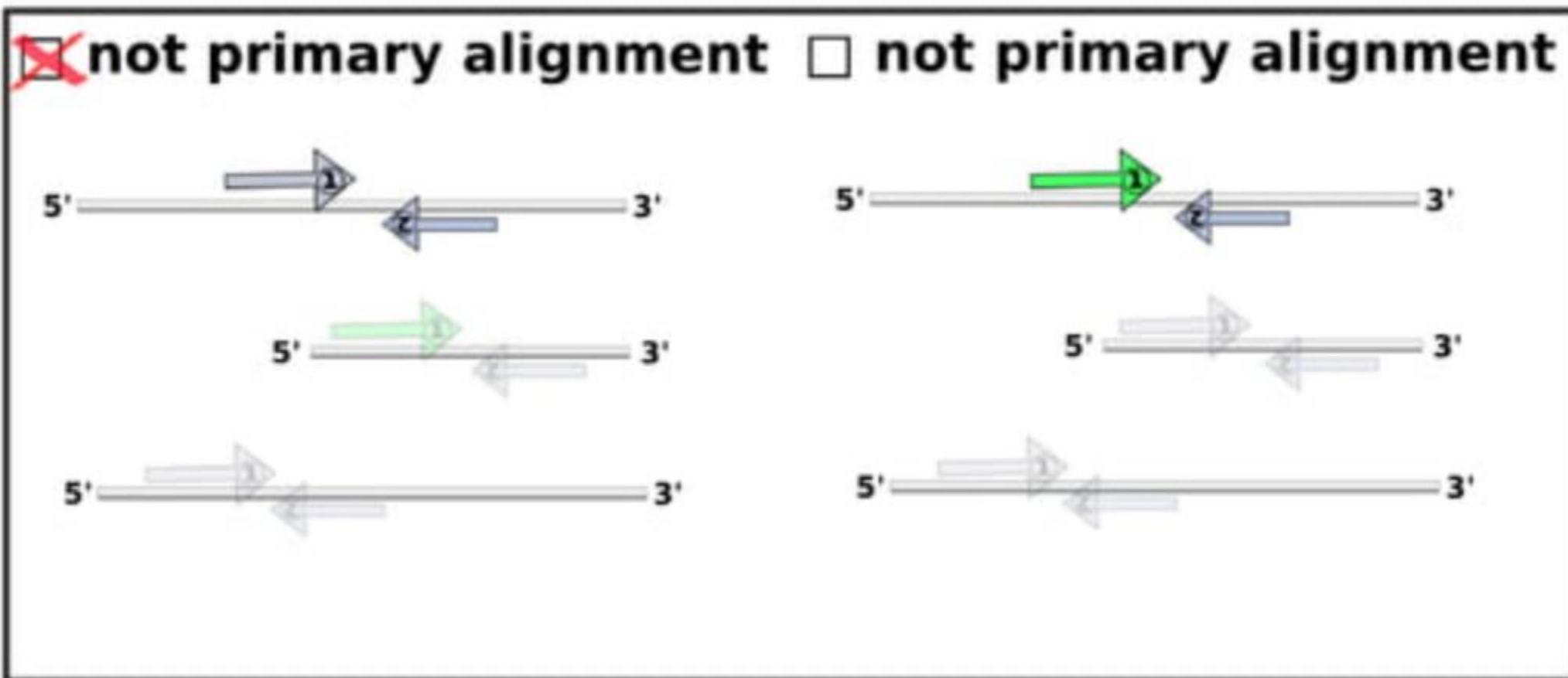


# SAM Flag

---



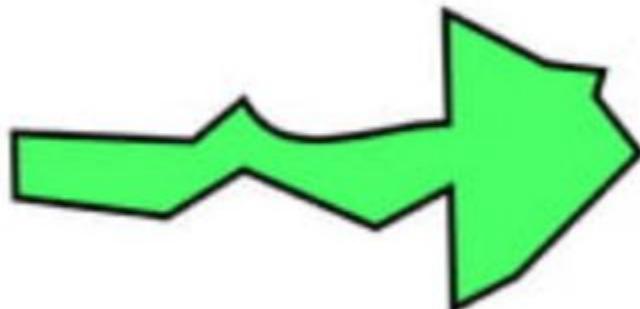
# SAM Flag



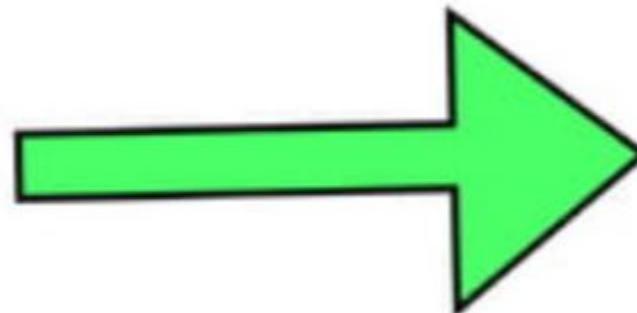
# SAM Flag

---

**read fails platform  
quality checks**



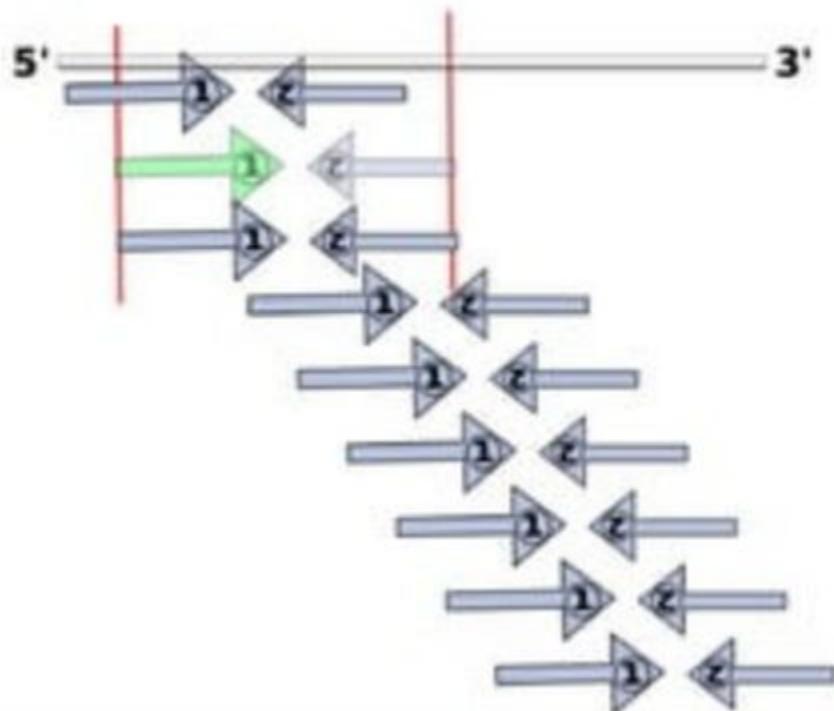
**read fails platform  
quality checks**



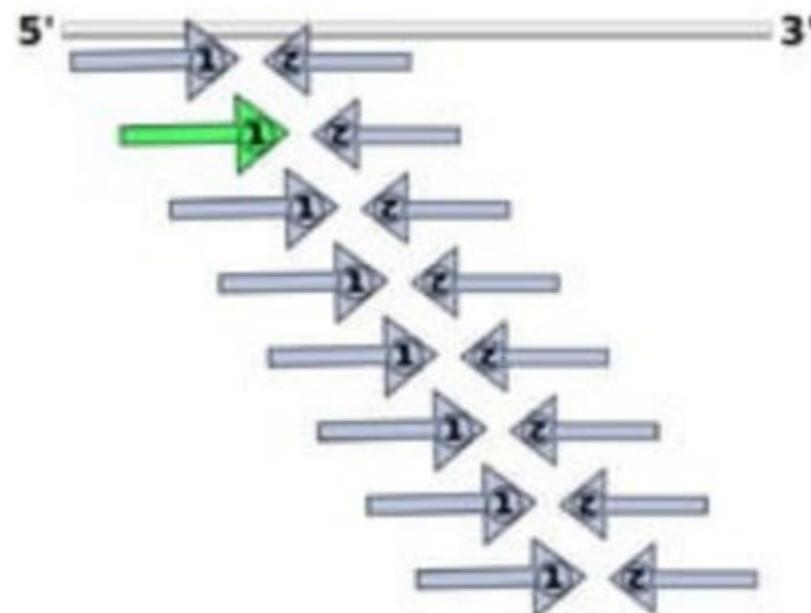
# SAM Flag

---

**read is duplicate**



**read is duplicate**





ST-E00223:32:H5J57CCXX:4:1220:14651:8868 99 1 10086

base2	base10	base16	Meaning	Applies to:
0000000001	1	0x0001	The read originated from a paired sequencing molecule	Both
0000000010	2	0x0002	The read is mapped in a <b>proper</b> pair	Pairs only
00000000100	4	0x0004	The query sequence itself is unmapped	Both
00000001000	8	0x0008	The query's mate is unmapped	Pairs only
00000010000	16	0x0010	Strand of the query (0 for forward; 1 for reverse strand)	Both
00000100000	32	0x0020	Strand of the query's mate	Pairs only
00001000000	64	0x0040	The query is the first read in the pair	Pairs only
00010000000	128	0x0080	The read is the second read in the pair	Pairs only
00100000000	256	0x0100	The alignment is not primary	Both
01000000000	512	0x0200	The read fails platform/vendor quality checks	Both
10000000000	1024	0x0400	The read is either a PCR duplicate or an optical duplicate	Both

00001100011

$$2^6 + 2^5 + 2^1 + 2^0 = 64 + 32 + 2 + 1 = 99$$

# CIGAR

# Encoding the details of the alignment

Operation	Meaning
M	Match*
D	Deletion w.r.t. reference
I	Insertion w.r.t. reference
N	Split or spliced alignment
S	Soft-clipping
H	Hard-clipping
P	Padding

## Reference:

## Experimental:

# CIGAR string:

ACCTGTC - - TAC**C**TTACG

ACCT-**TCCATAC****T**TTA**TC**

4M 1D 2M 2| 7M 2S

4M1D2M2I7M2S



## LENGTH/OPERATION

# CIGAR Extended

---

Operation	Meaning
=	Exact match
X	Mismatch
D	Deletion w.r.t. reference
I	Insertion w.r.t. reference
N	Split or spliced alignment
S	Soft-clipping
H	Hard-clipping
P	Padding

Reference:

ACCTGTC - - TAC**C**TTACG

Experimental:

ACCT - TCCATA**T**TTATC



4= 1D 2= 2I 3= 1X 3= 2S

CIGAR string:

4=1D2=2I3=1X3=2S

# SAM to BAM

---

**Do it once**

Create BWT of reference genome.

```
$ bwa index grch38.fa
```



**Output is in SAM format**

Align paired-end FASTQ to BWT index.

```
$ bwa mem -t 16 grch38.fa 1.fq 2.fq > sample.sam
```



**Output is in BAM format.**  
Unsorted!  
random genomic order as  
reads are randomly  
placed in FASTQ by  
sequencer.

Convert SAM to BAM

```
$ samtools view -Sb sample.sam > sample.bam
```

# SAMTOOLS

---

## Converting and manipulating SAM/BAM

Commands:

-- Indexing

dict	create a sequence dictionary file
faidx	index/extract FASTA
index	index alignment

-- Editing

calmd	recalculate MD/NM tags and '=' bases
fixmate	fix mate information
reheader	replace BAM header
rmdup	remove PCR duplicates
targetcut	cut fosmid regions (for fosmid pool only)
addreplacerg	adds or replaces RG tags

-- Viewing

flags	explain BAM flags
tview	text alignment viewer
view	SAM<->BAM<->CRAM conversion
depad	convert padded BAM to unpadded BAM

# SAMTOOLS

---

Commands:

-- File operations

collate	shuffle and group alignments by name
cat	concatenate BAMs
merge	merge sorted alignments
mpileup	multi-way pileup
sort	sort alignment file
split	splits a file by read group
quickcheck	quickly check if SAM/BAM/CRAM file appears intact
fastq	converts a BAM to a FASTQ
fasta	converts a BAM to a FASTA

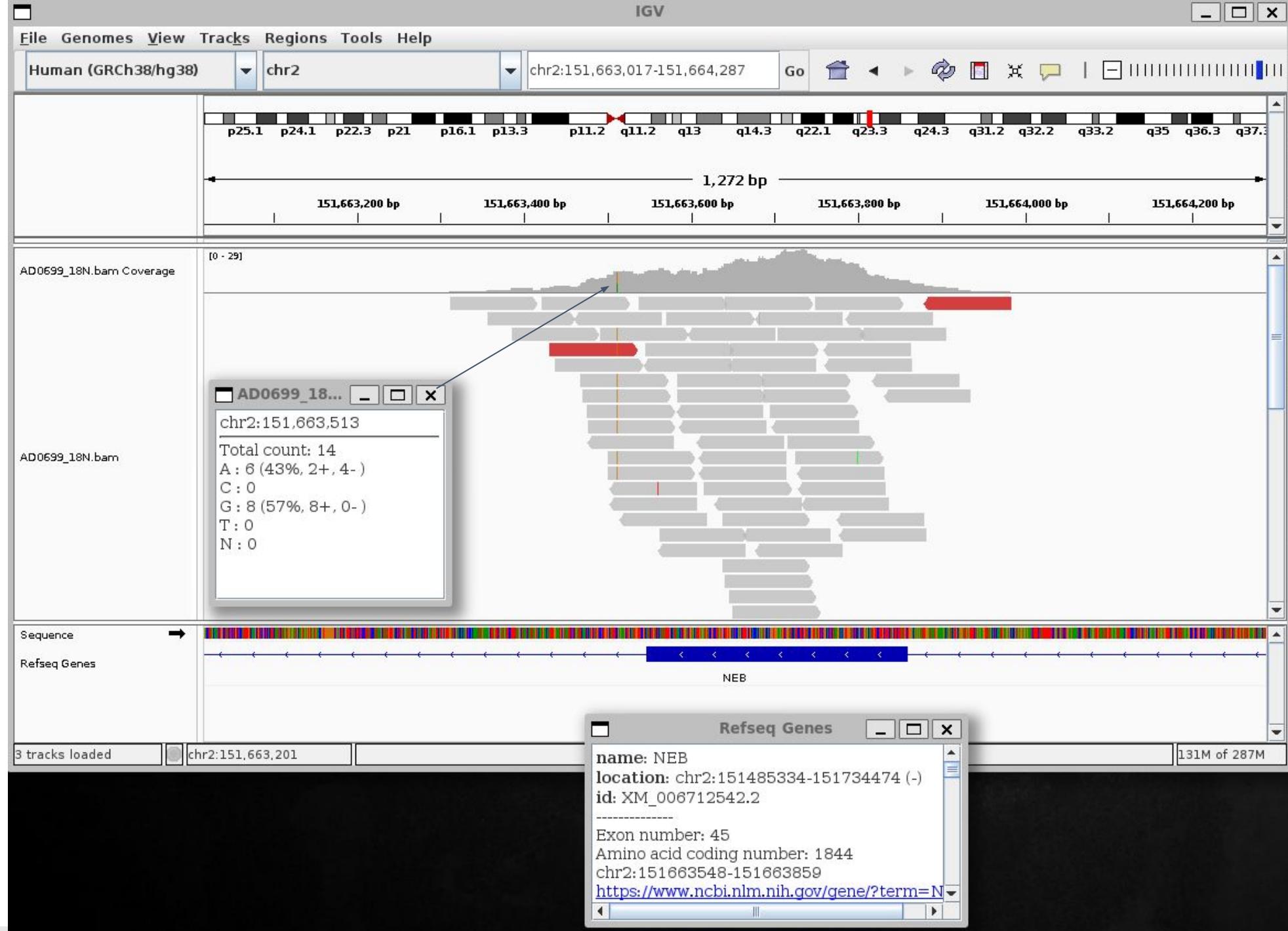
-- Statistics

bedcov	read depth per BED region
depth	compute the depth
flagstat	simple stats
idxstats	BAM index stats
phase	phase heterozygotes
stats	generate stats (former bamcheck)

# Integrative Genomics Viewer (IGV)

Visualization tool for exploring and analyzing genomic data





# Genetic Variation

---

Differences in DNA content or structure among individuals

- Any two individuals have ~99.5% identical DNA.

The human genome is big - each haploid set of 23 chromosomes has 3.1 billion nucleotides.

There are >100,000,000 known genetic variants in the human genome

~99.8% identical DNA  
(differ at 1/ 620 - 1/750 bp)



Drosophila ~ 1/180

99% identical DNA



# Types of Genetic Variation

---

ctc**c**gag  
ctc**t**gag

Single-nucleotide  
polymorphisms  
**(SNPs)**

*“DNA spelling  
mistakes”*

ctc--ag  
ctc**tg**ag

Insertion-deletion  
polymorphisms  
**(INDELs)**

*“extra or missing  
DNA”*

ctcaag  
ctcag

Structural  
variants  
**(SVs)**

*“Large blocks of extra,  
missing  
or rearranged  
DNA”*

# A Normal Human

---

"We find that a typical [human] genome differs from the reference human genome at **4.1 million to 5.0 million sites**. Although **>99.9% of variants consist of SNPs and short indels**, structural variants affect more bases: the typical genome contains an estimated **2,100 to 2,500 structural variants** (~1,000 large deletions, ~160 copy-number variants, ~915 Alu insertions, ~128 L1 insertions, ~51 SVA insertions, ~4 NUMTs, and ~10 inversions), **affecting ~20 million bases of sequence.** A global reference for human genetic variation

Nucleotide diversity (II):  
**1/756 bp to 1/620 bp**

The 1000 Genomes Project Consortium\*

The 1000 Genomes Project set out to provide a comprehensive description of common human genetic variation by applying whole-genome sequencing to a diverse set of individuals from multiple populations. Here we report completion of the project, having reconstructed the genomes of 2,504 individuals from 26 populations using a combination of low-coverage whole-genome sequencing, deep exome sequencing, and dense microarray genotyping. We characterized a broad spectrum of genetic variation, in total over 88 million variants (84.7 million single nucleotide polymorphisms (SNPs), 3.6 million short insertions/deletions (indels), and 60,000 structural variants), all phased onto high-quality haplotypes. This resource includes >99% of SNP variants with a frequency of >1% for a variety of ancestries. We describe the distribution of genetic variation across the global sample, and discuss the implications for common disease studies.

Mutation != Polymorphism (or SNP)

# Mutations

---

acctccgagta

a toy population of 10 identical chromosomes

# Mutations

---

Mutation creates genetic diversity

acctccgagta

acctccgagta

acctccgagta

acctccgagta

acctccgagta

acctccgagta

acctccgagta

acctccgagta

acctccgagta

acctc**T**gagta

mutation:

*private to this chromosome / individual*

# Mutations

---

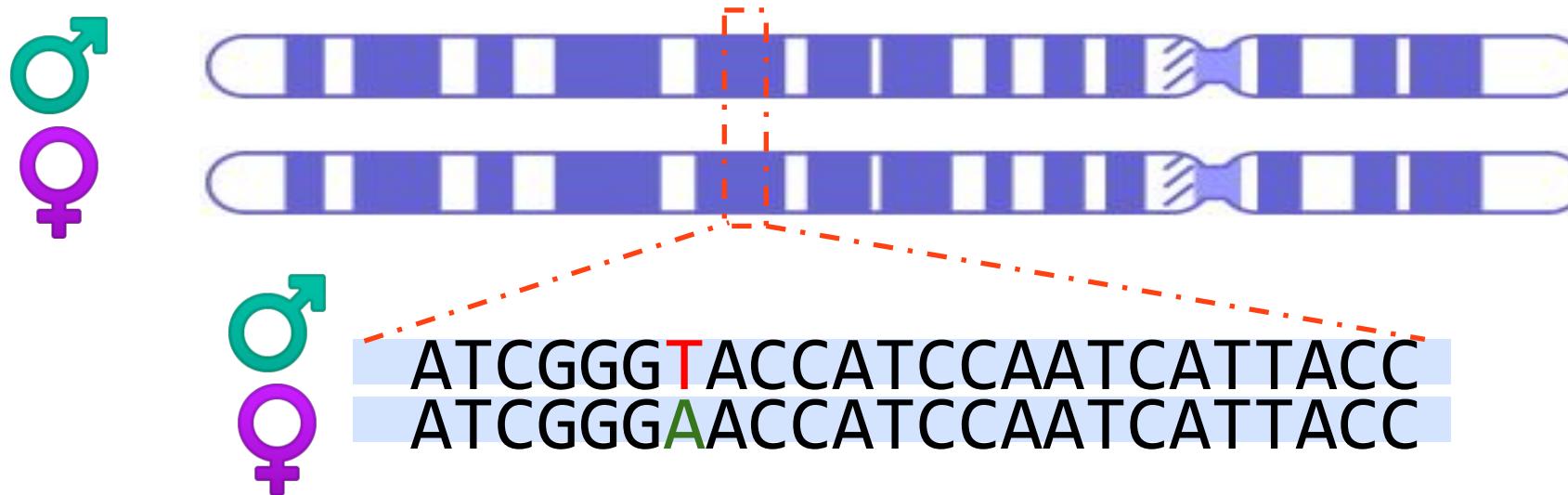
From mutation to polymorphism

acctccgagta  
acctccgagta  
acctccgagta  
acctc**T**gagta  
acctccgagta

acctc**T**gagta  
acctccgagta  
acctc**T**gagta  
acctccgagta  
acctc**T**gagta

# Diploid Genomes

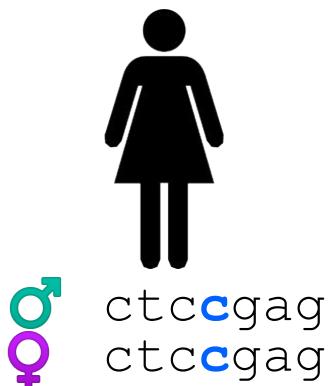
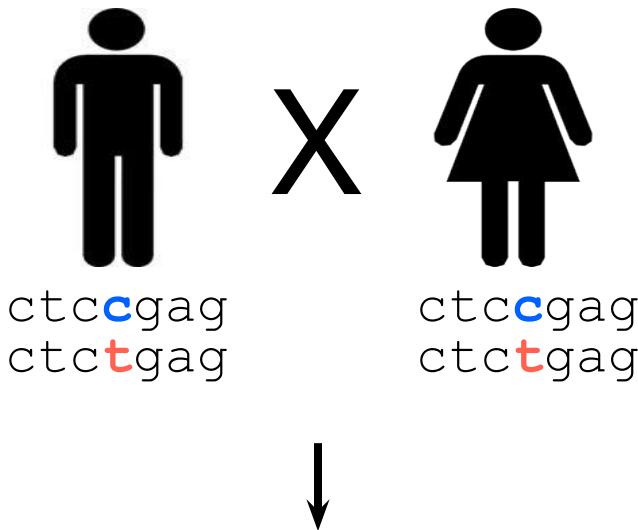
---



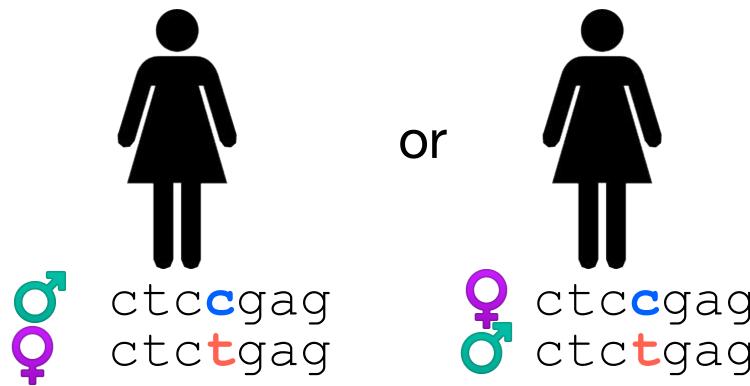
Our genome is comprised of a paternal and a maternal "haplotype". Together, they form our "genotype"

# Inherited Germline Variation

---

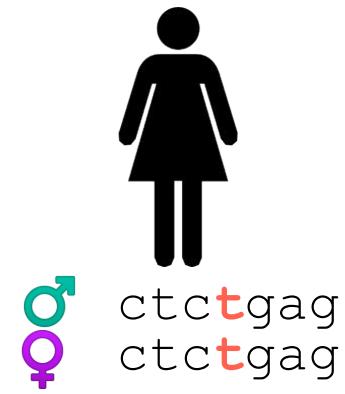


Kid is homozygous  
(C/C)



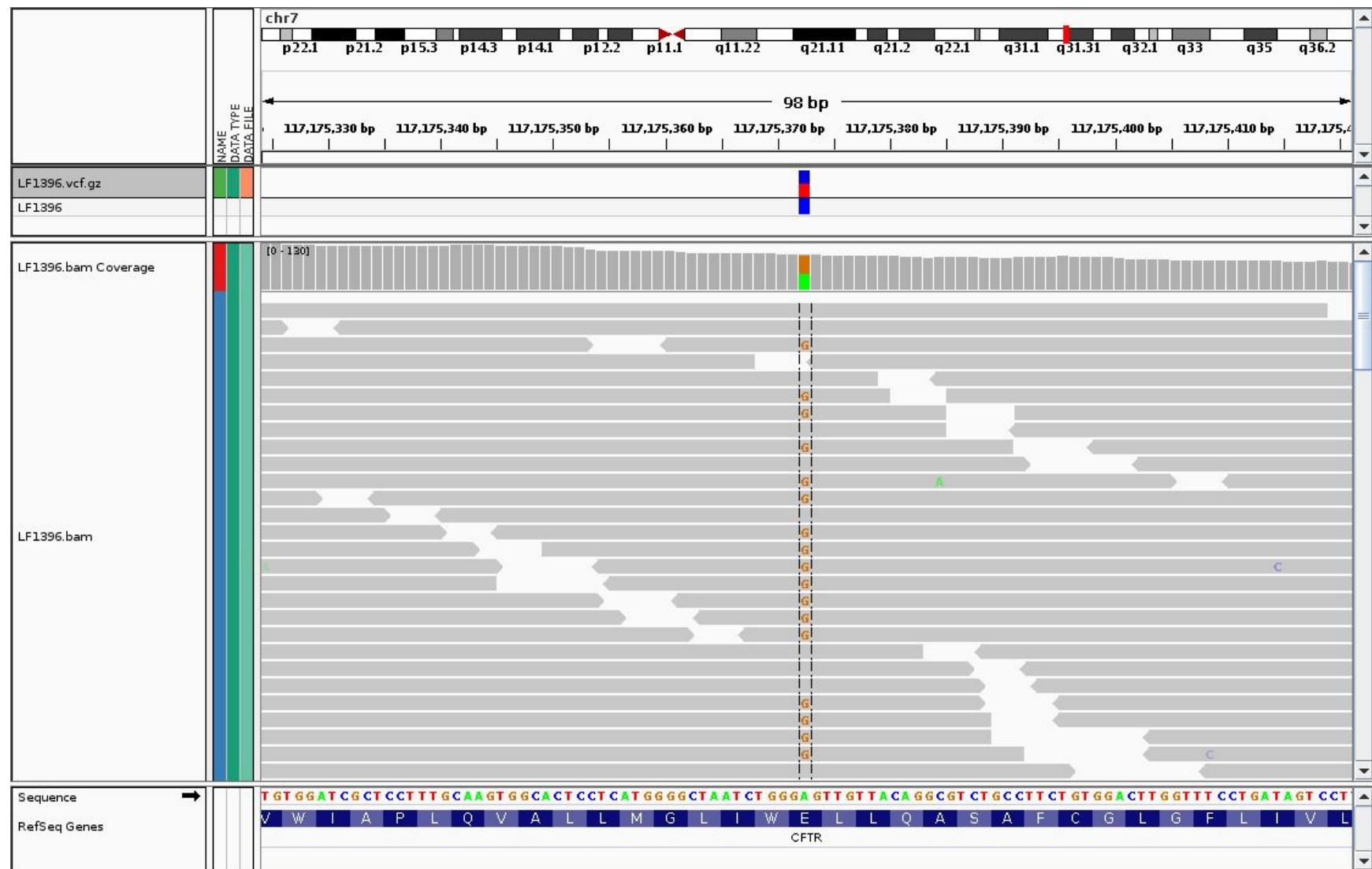
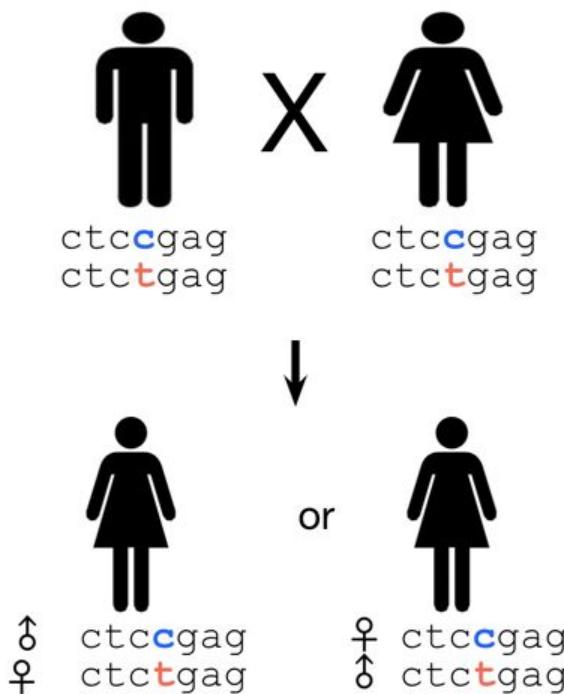
Kid is heterozygous  
(C/T)

**Example:** Mom and dad are heterozygous; that is, the zygote from which they developed was comprised of a sperm and egg with two different alleles



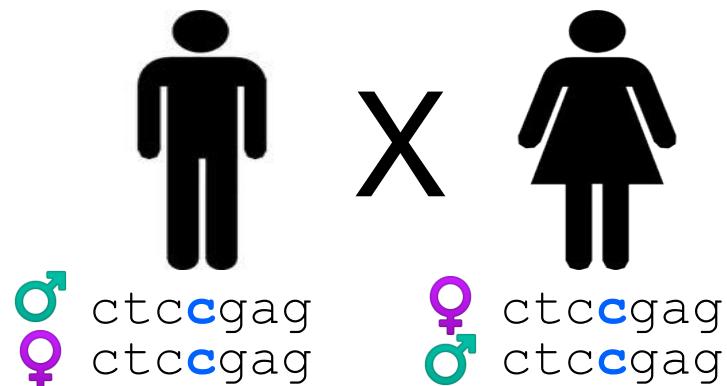
Kid is homozygous  
(T/T)

# Heterozygous Variation



# *De novo Mutation*

---

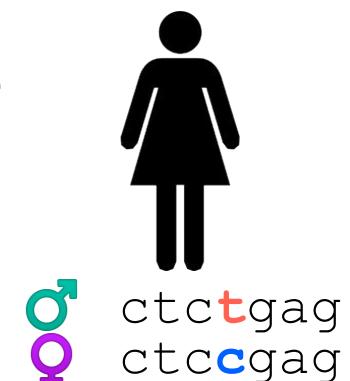


**Example:** Mom and dad are homozygous for the same alleles.

***New mutation occurs in father's or mother's germ cell***

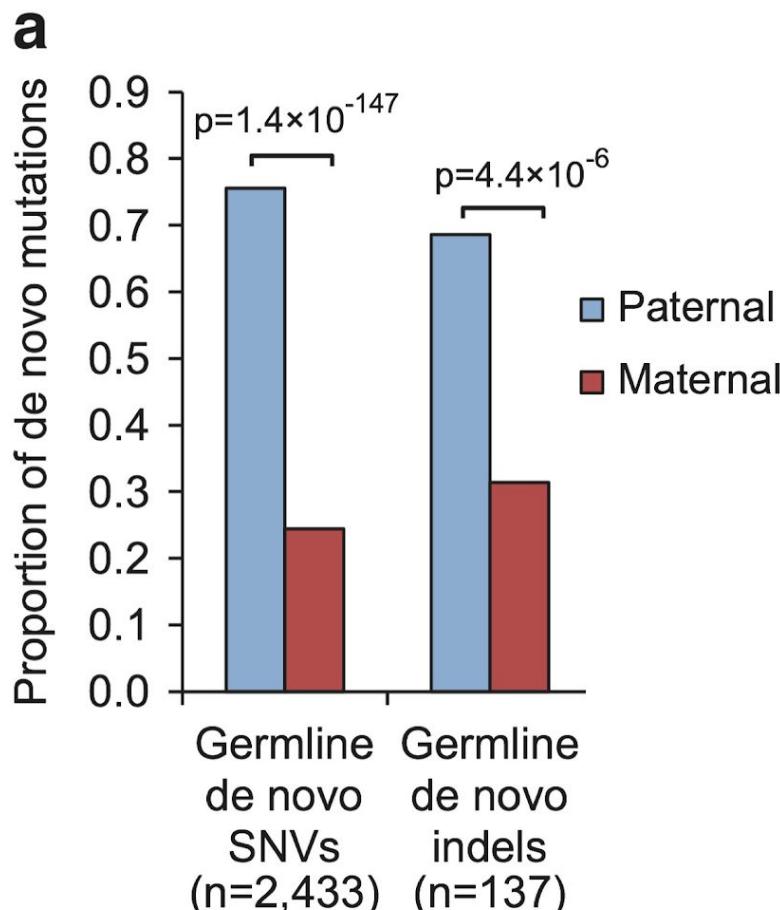


**Note:** This is a derivative chromosome of the one the father inherited from His parents



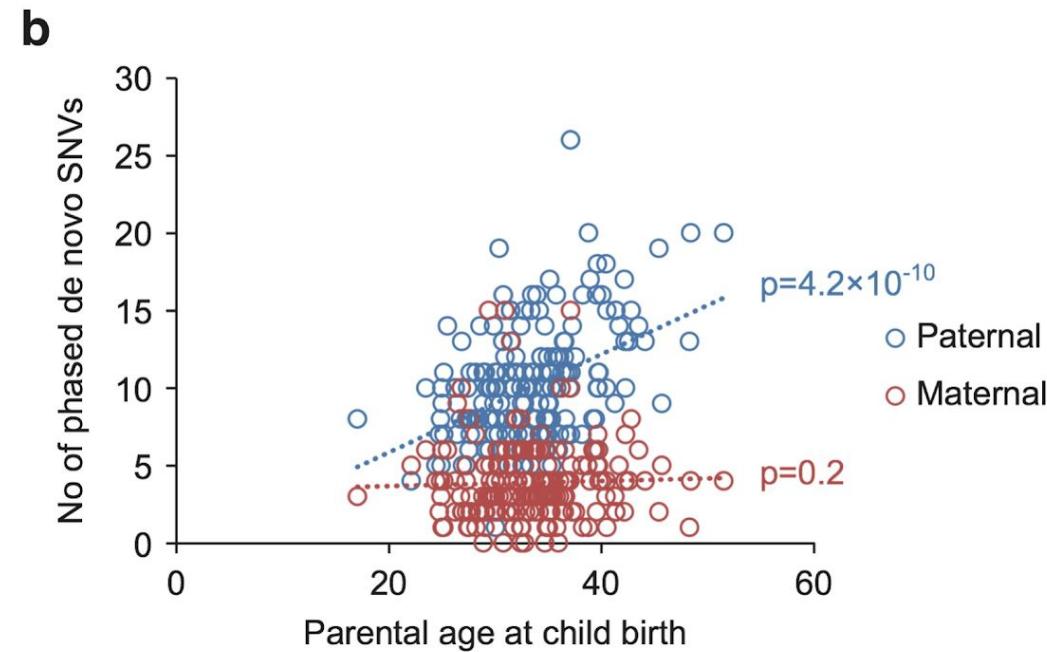
Kid is heterozygous owing to *de novo mutation*.  
(C/T)

# DNMs Frequency



(data from 200 ASD trios)

2 new DNMs per year of  
paternal age (Kong et al. 2012, *Nature*)

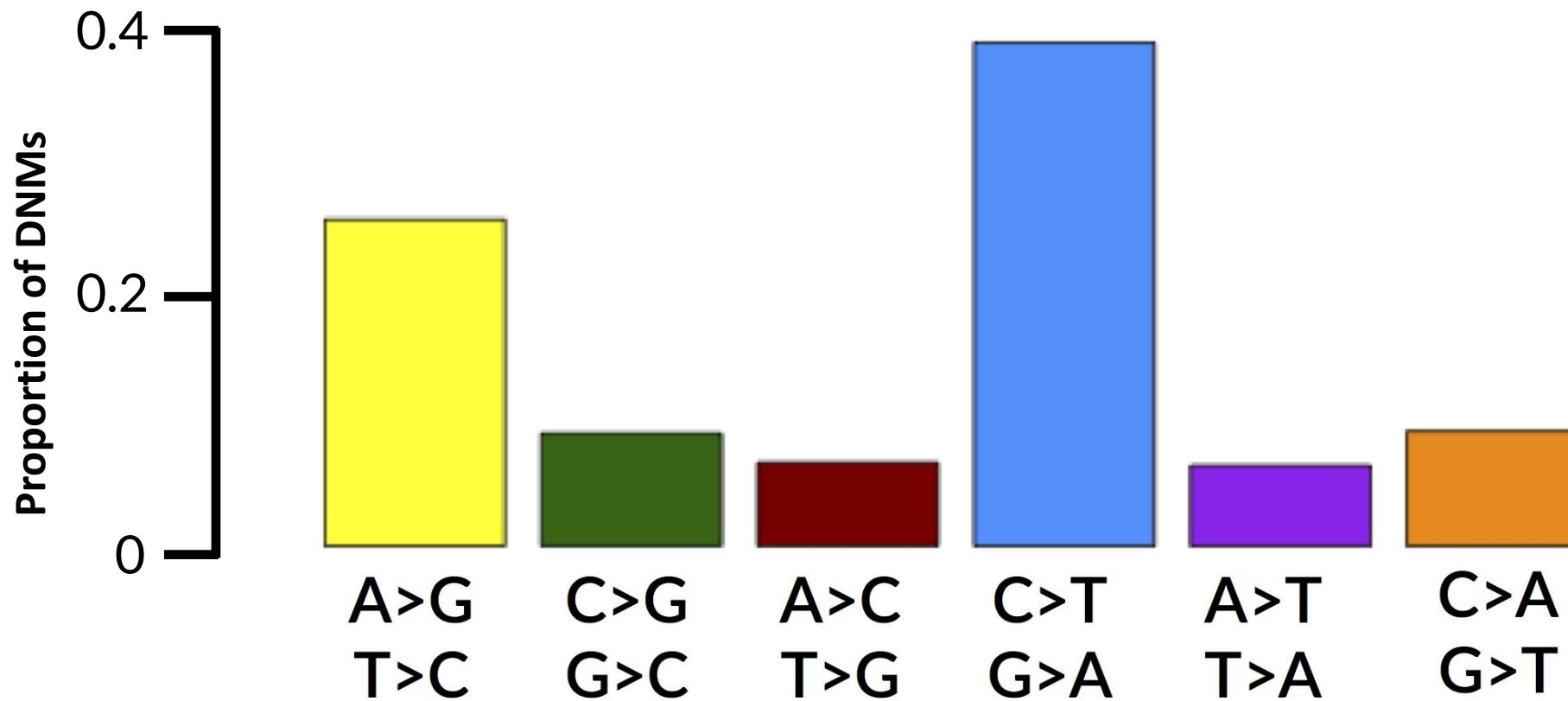


Yuen et al. (2016) *Nature Genomic Medicine*

# DNMs Frequency

---

Due to spontaneous deamination of methylated cytosines, C>T transitions predominate in DNMs



**Genome-wide DNM rates among SFARI families**

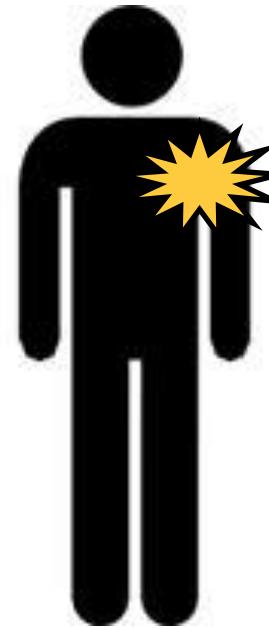
# Somatic Mutations

---



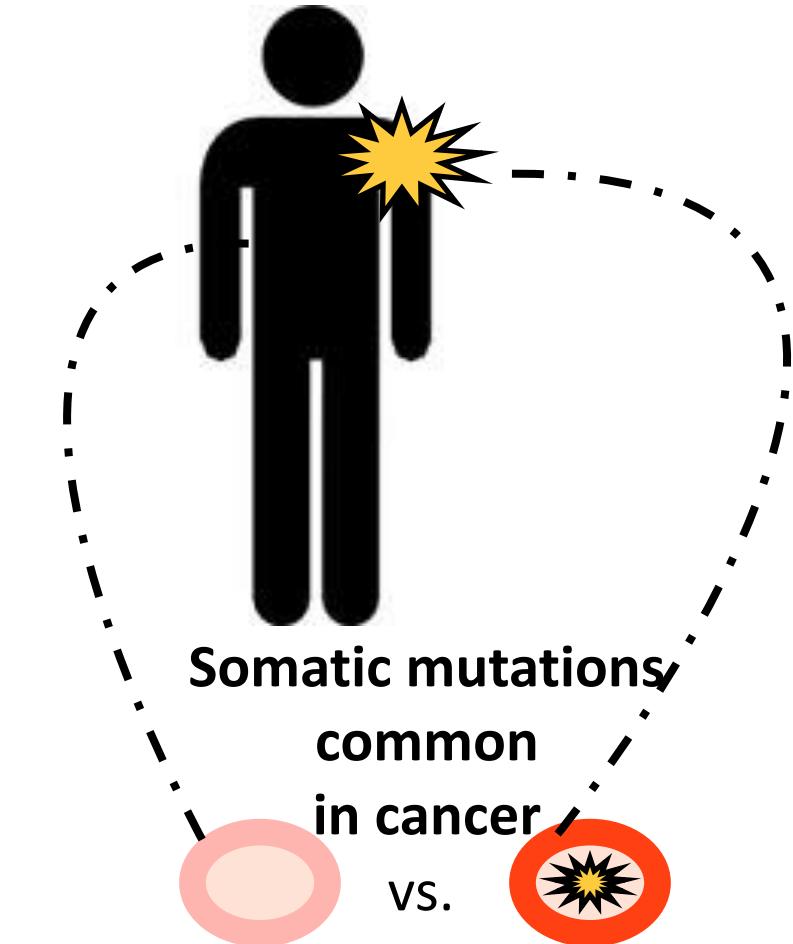
**Germline mutation**

- occur in sperm or egg.
- are heritable



**Somatic mutation**

- non-germline tissues.
- are not heritable



compare DNA from cancer cells to healthy cells from same individual

# The 1000 (2504) Genome Project

---

ARTICLE

---

---

OPEN

doi:10.1038/nature15393

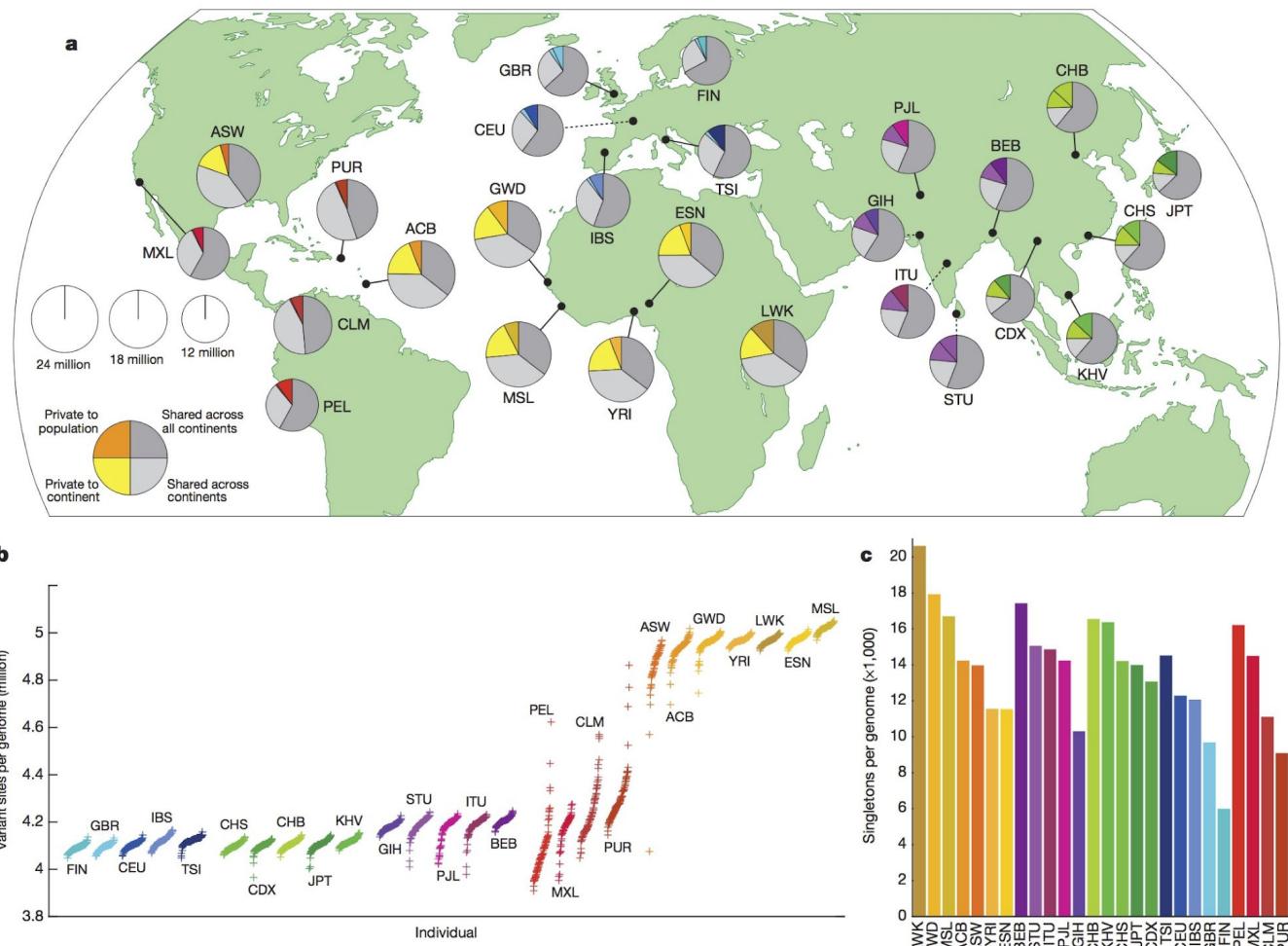
## A global reference for human genetic variation

The 1000 Genomes Project Consortium\*

The 1000 Genomes Project set out to provide a comprehensive description of common human genetic variation by applying whole-genome sequencing to a diverse set of individuals from multiple populations. Here we report completion of the project, having reconstructed the genomes of 2,504 individuals from 26 populations using a combination of low-coverage whole-genome sequencing, deep exome sequencing, and dense microarray genotyping. We characterized a broad spectrum of genetic variation, in total over 88 million variants (84.7 million single nucleotide polymorphisms (SNPs), 3.6 million short insertions/deletions (indels), and 60,000 structural variants), all phased onto high-quality haplotypes. This resource includes >99% of SNP variants with a frequency of >1% for a variety of ancestries. We describe the distribution of genetic variation across the global sample, and discuss the implications for common disease studies.

# The 1000 (2504) Genome Project

2,504 individuals  
from diverse  
ancestries



**Figure 1 | Population sampling.** a, Polymorphic variants within sampled populations. The area of each pie is proportional to the number of polymorphisms within a population. Pies are divided into four slices, representing variants private to a population (darker colour unique to population), private to a continental area (lighter colour shared across continental group), shared

across continental areas (light grey), and shared across all continents (dark grey). Dashed lines indicate populations sampled outside of their ancestral continental region. b, The number of variant sites per genome. c, The average number of singletons per genome.

# The extent of genetic variation by subpopulation

**Table 1 | Median autosomal variant sites per genome**

	AFR		AMR		EAS		EUR		SAS	
Samples	661		347		504		503		489	
Mean coverage	8.2		7.6		7.7		7.4		8.0	
	Var. sites	Singletons								
SNPs	4.31M	14.5k	3.64M	12.0k	3.55M	14.8k	3.53M	11.4k	3.60M	14.4k
Indels	625k	-	557k	-	546k	-	546k	-	556k	-
Large deletions	1.1k	5	949	5	940	7	939	5	947	5
CNVs	170	1	153	1	158	1	157	1	165	1
MEI (Alu)	1.03k	0	845	0	899	1	919	0	889	0
MEI (L1)	138	0	118	0	130	0	123	0	123	0
MEI (SVA)	52	0	44	0	56	0	53	0	44	0
MEI (MT)	5	0	5	0	4	0	4	0	4	0
Inversions	12	0	9	0	10	0	9	0	11	0
Nonsynon	12.2k	139	10.4k	121	10.2k	144	10.2k	116	10.3k	144
Synon	13.8k	78	11.4k	67	11.2k	79	11.2k	59	11.4k	78
Intron	2.06M	7.33k	1.72M	6.12k	1.68M	7.39k	1.68M	5.68k	1.72M	7.20k
UTR	37.2k	168	30.8k	136	30.0k	169	30.0k	129	30.7k	168
Promoter	102k	430	84.3k	332	81.6k	425	82.2k	336	84.0k	430
Insulator	70.9k	248	59.0k	199	57.7k	252	57.7k	189	59.1k	243
Enhancer	354k	1.32k	295k	1.05k	289k	1.34k	288k	1.02k	295k	1.31k
TFBSs	927	4	759	3	748	4	749	3	765	3
Filtered LoF	182	4	152	3	153	4	149	3	151	3
HGMD-DM	20	0	18	0	16	1	18	2	16	0
GWAS	2.00k	0	2.07k	0	1.99k	0	2.08k	0	2.06k	0
ClinVar	28	0	30	1	24	0	29	1	27	1

See Supplementary Table 1 for continental population groupings. CNVs, copy-number variants; HGMD-DM, Human Gene Mutation Database disease mutations; k, thousand; LoF, loss-of-function; M, million; MEI, mobile element insertions.

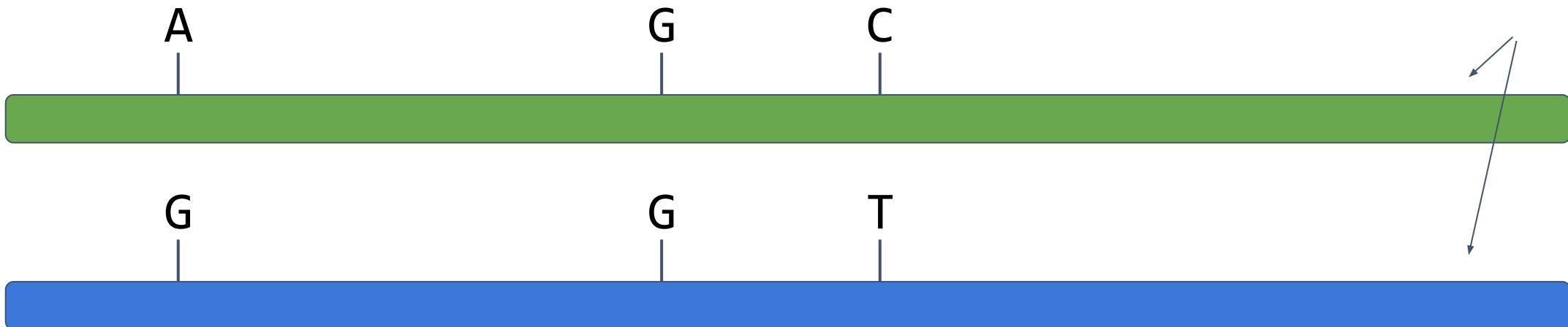
# Haplotypes

---

Each chromosome is a mosaic of alleles.

You inherit two haplotypes: one from mom, one from dad

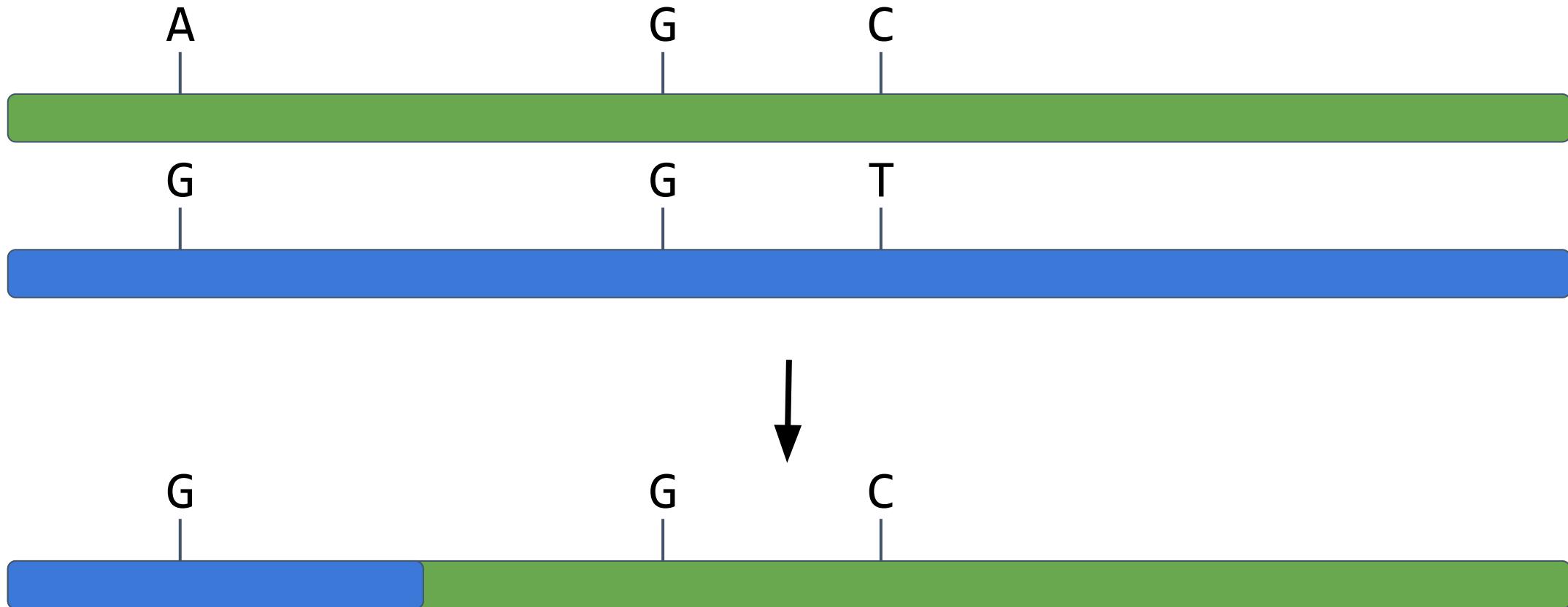
Haplotypes



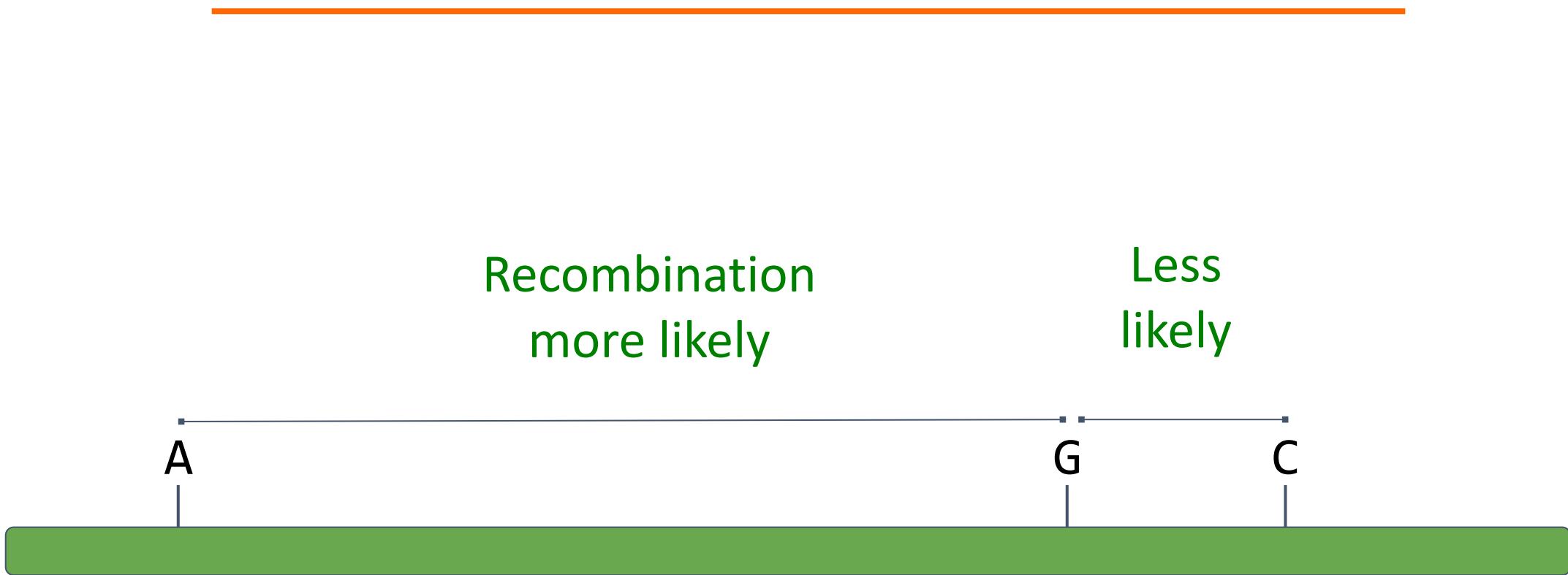
# Haplotypes

---

Meiotic recombination shuffles alleles and generates new haplotypes



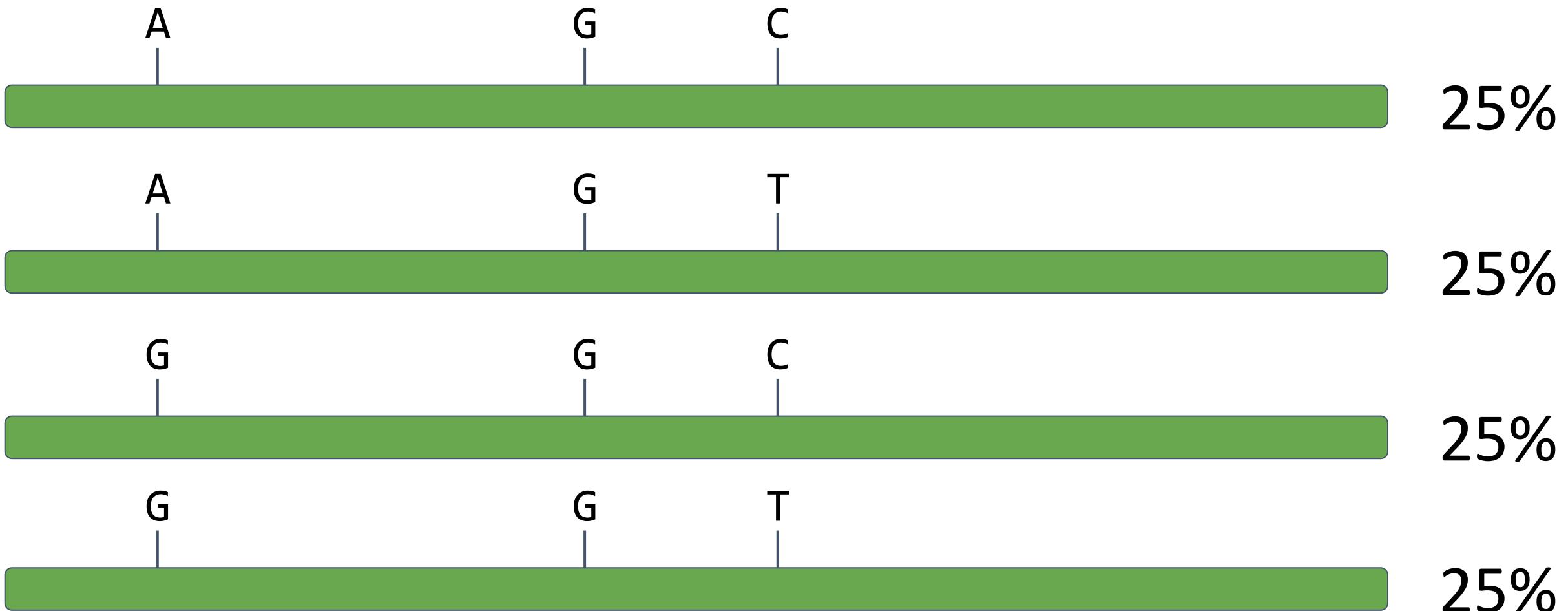
# Haplotypes



# Haplotypes

---

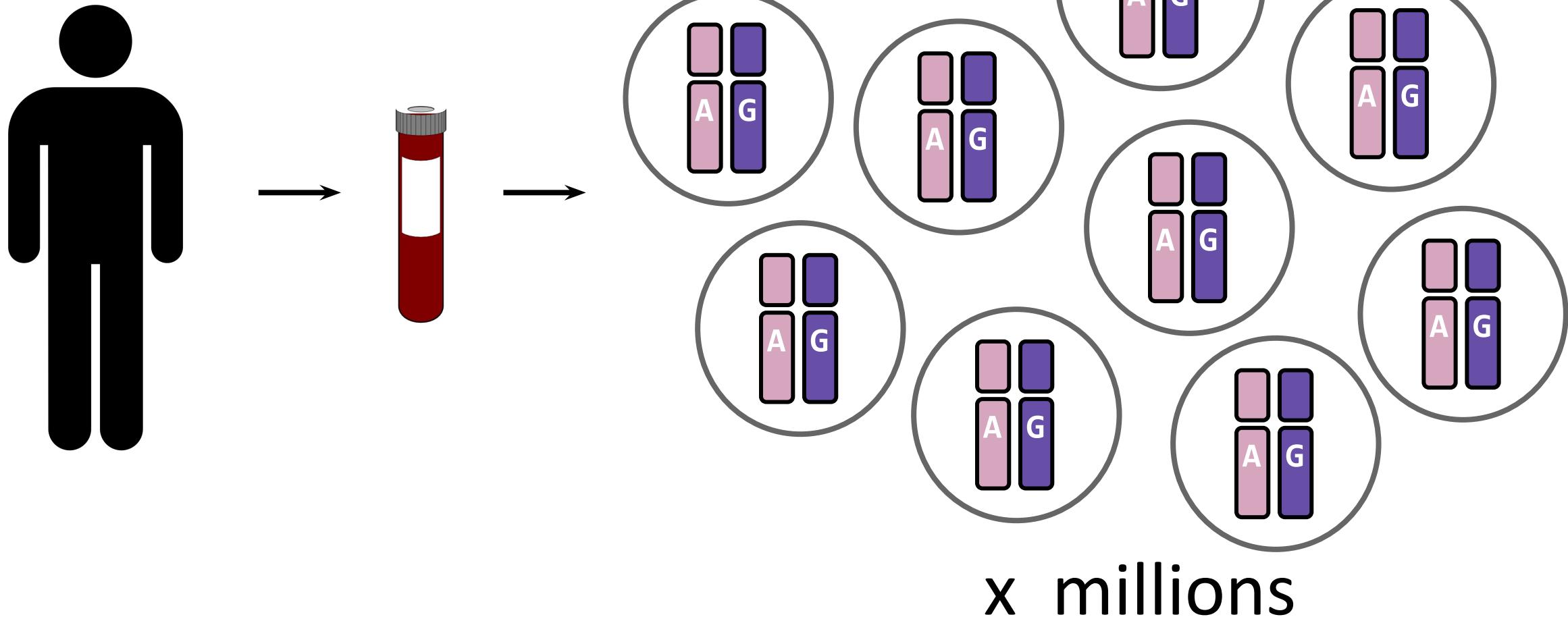
**Linkage equilibrium:** random association of alleles at different loci.



# Variant Calling

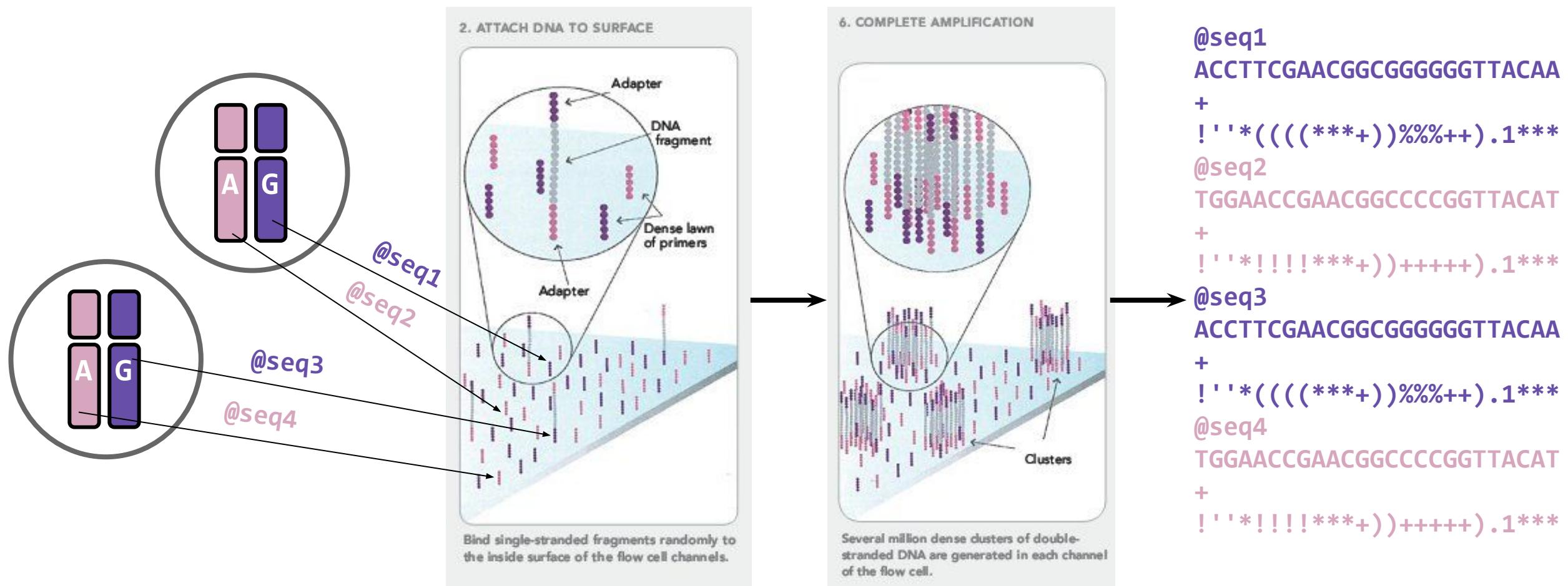
---

Find inherited genetic variation by sequencing DNA from millions of cells



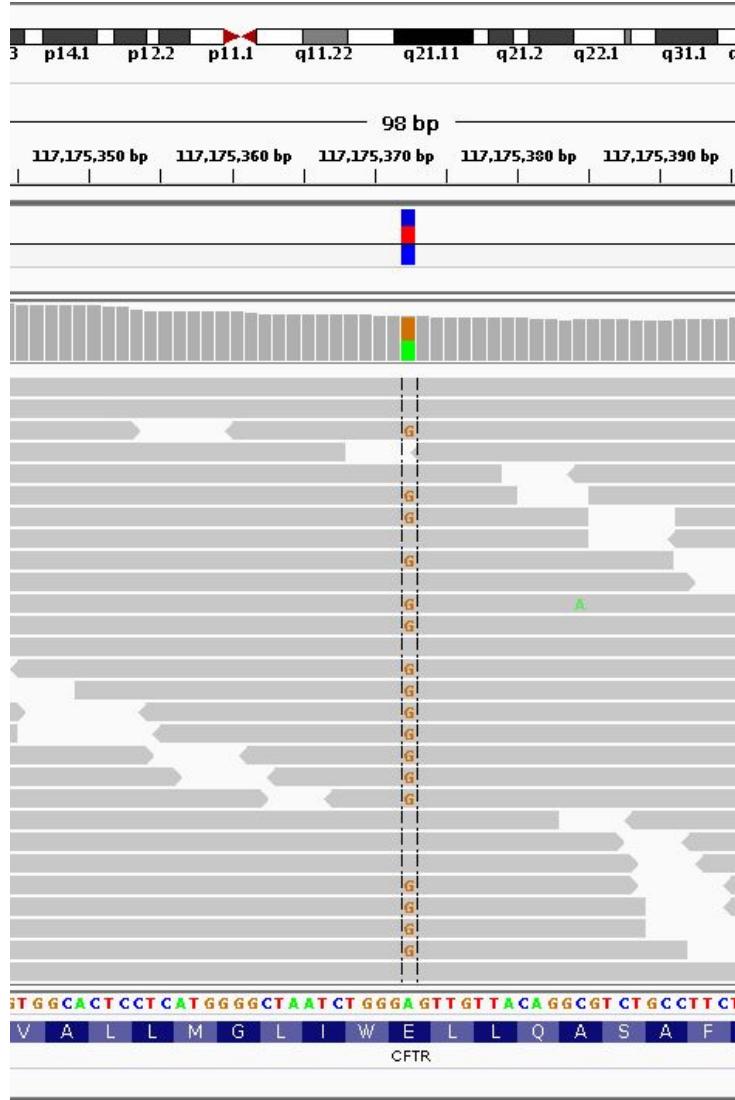
# Variant Calling

Each DNA cluster is amplified from a single strand from a single haploid chromosome from a single cell.



# Variant Calling

---



What information is needed to decide if a variant exists?

- Depth of coverage at the locus
- Bases observed at the locus
- The base qualities of each allele
- The strand composition
- Mapping qualities
- Proper pairs?
- Expected polymorphism rate

# PolyBayes

The first statistically rigorous variant detection tool.

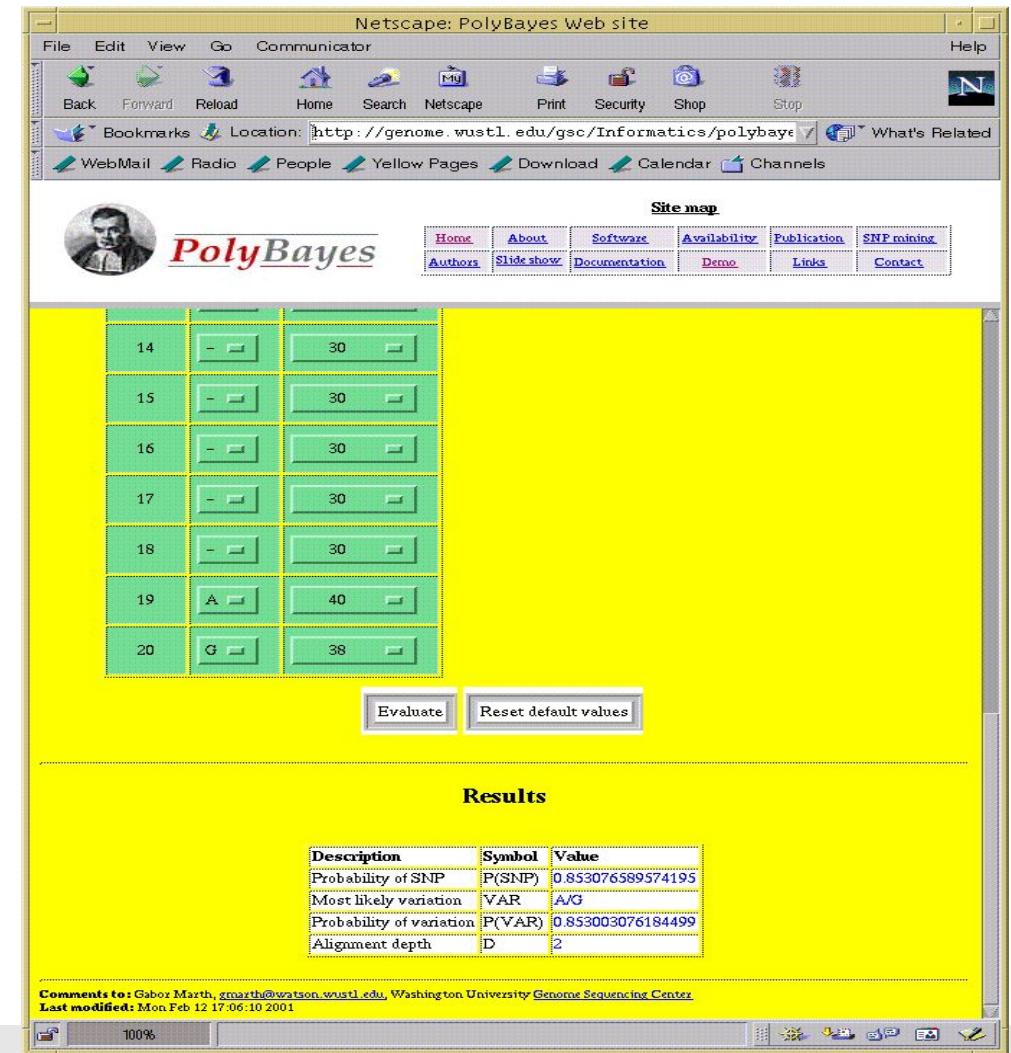
letter

© 1999 Nature America Inc. • <http://genetics.nature.com>

## A general approach to single-nucleotide polymorphism discovery

Gabor T. Marth<sup>1</sup>, Ian Korf<sup>1</sup>, Mark D. Yandell<sup>1</sup>, Raymond T. Yeh<sup>1</sup>, Zhijie Gu<sup>2</sup>, Hamideh Zakeri<sup>2</sup>, Nathan O. Stitzel<sup>1</sup>, LaDeana Hillier<sup>1</sup>, Pui-Yan Kwok<sup>2</sup> & Warren R. Gish<sup>1</sup>

Its main innovation was the use of Bayes's theorem



# Bayes' Theorem

<b>Thomas Bayes</b>	
	
Portrait used of Bayes in a 1936 book, <sup>[1]</sup> but it is doubtful whether the portrait is actually of him. <sup>[2]</sup> No earlier portrait or claimed portrait survives.	
<b>Born</b>	c. 1701 London, England
<b>Died</b>	7 April 1761 (aged 59) Tunbridge Wells, Kent, England
<b>Residence</b>	Tunbridge Wells, Kent, England
<b>Nationality</b>	English
<b>Known for</b>	Bayes' theorem Signature
	

$$\Pr(\text{spam}|\text{words}) = \frac{\Pr(\text{words}|\text{spam}) \Pr(\text{spam})}{\Pr(\text{words})}$$

## Statement of theorem [edit]

Bayes' theorem is stated mathematically as the following equation:<sup>[2]</sup>

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

where  $A$  and  $B$  are events and  $P(B) \neq 0$ .

- $P(A)$  and  $P(B)$  are the probabilities of observing  $A$  and  $B$  without regard to each other.
- $P(A | B)$ , a conditional probability, is the probability of observing event  $A$  given that  $B$  is true.
- $P(B | A)$  is the probability of observing event  $B$  given that  $A$  is true.

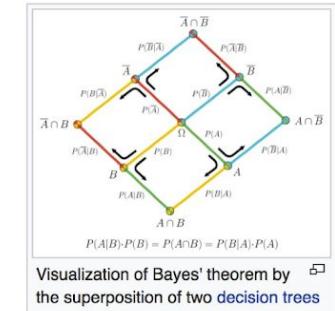
## History [edit]

Bayes' theorem was named after the Reverend Thomas Bayes (1701–1761), who studied how to compute a distribution for the probability parameter of a binomial distribution (in modern terminology). Bayes' unpublished manuscript was significantly edited by Richard Price before it was posthumously read at the Royal Society. Price edited<sup>[3]</sup> Bayes' major work "An Essay towards solving a Problem in the Doctrine of Chances" (1763), which appeared in "Philosophical Transactions,"<sup>[4]</sup> and contains Bayes' Theorem. Price wrote an introduction to the paper which provides some of the philosophical basis of Bayesian statistics. In 1765 he was elected a Fellow of the Royal Society in recognition of his work on the legacy of Bayes.<sup>[5][6]</sup>

The French mathematician Pierre-Simon Laplace reproduced and extended Bayes' results in 1774, apparently quite unaware of Bayes' work.<sup>[7][8]</sup> The Bayesian interpretation of probability was developed mainly by Laplace.<sup>[9]</sup>

Stephen Stigler suggested in 1983 that Bayes' theorem was discovered by Nicholas Saunderson, a blind English mathematician, some time before Bayes;<sup>[10][11]</sup> that interpretation, however, has been disputed.<sup>[12]</sup> Martyn Hooper<sup>[13]</sup> and Sharon McGayne<sup>[14]</sup> have argued that Richard Price's contribution was substantial:

By modern standards, we should refer to the Bayes–Price rule. Price discovered Bayes' work, recognized its importance, corrected it, contributed to the article, and found a use for it. The modern convention of employing Bayes' name alone is unfair but so entrenched that anything else makes little sense.<sup>[14]</sup>



# Bayes' Theorem

---

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



Conditional probability. That is,  
the probability of A occurring,  
given that B has occurred.

# Bayes' Theorem

---

$$P(H|D) = \frac{P(D|H) * P(H)}{P(D)}$$



Conditional probability. That is,  
the probability of Hypothesis  
being correct, given the  
observation of particular data (D)

# Bayes' Theorem

---

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

Posterior probability

Prior  
Probability  
Of A

# Bayes' Theorem Applications

---

Widely used in machine learning and finance.

Decision making in driverless cars

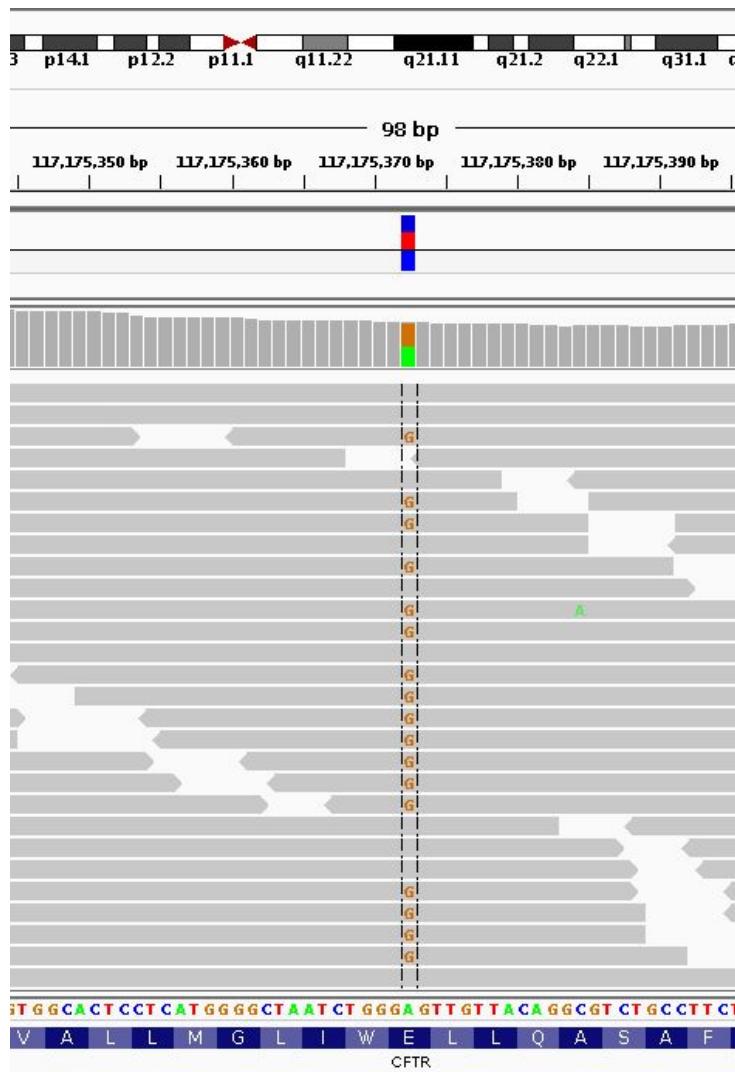
Email spam detection

Assess disease risk from test results

Voice recognition software

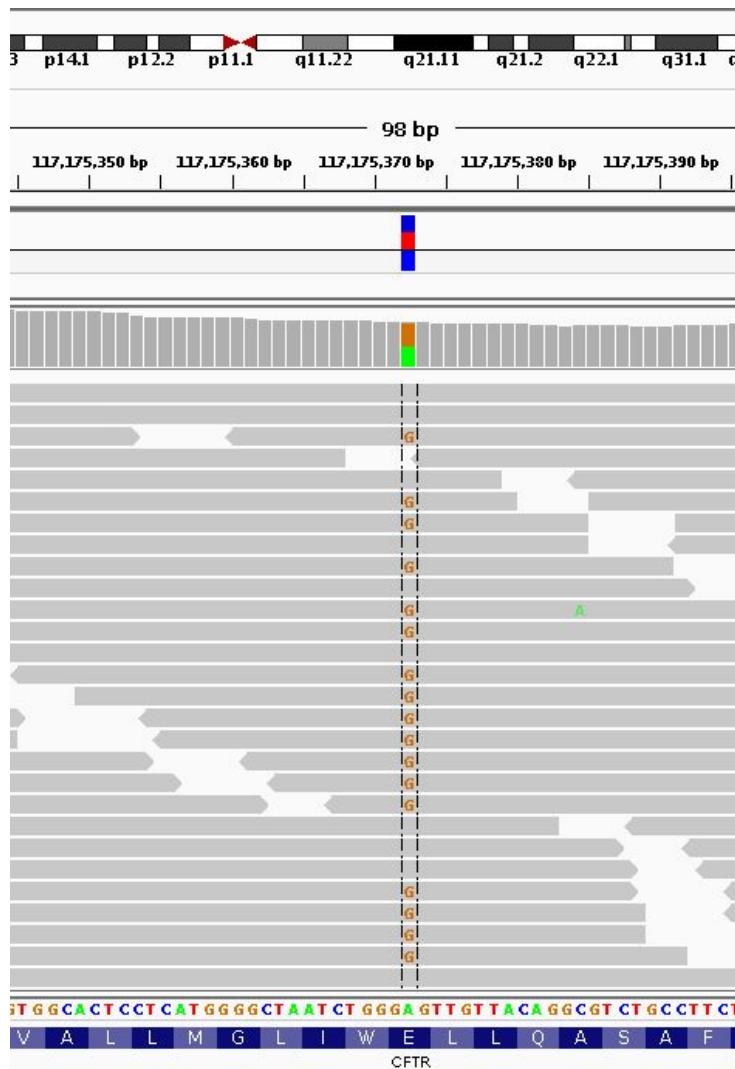
Text autocomplete...

# Bayesian SNP Calling



$$P(\text{SNP} \mid \text{Data}) = \frac{P(\text{Data} \mid \text{SNP}) * P(\text{SNP})}{P(\text{Data})}$$

# Bayesian SNP Calling



$$P(\text{SNP} \mid \text{Data}) = \frac{P(\text{Data} \mid \text{SNP}) * P(\text{SNP})}{P(\text{Data})}$$

- Depth of coverage at the locus
- Bases observed at the locus
- The base qualities of each allele
- The strand composition
- Mapping qualities
- Proper pairs?
- Expected polymorphism rate

# PolyBayes

letter

© 1999 Nature America Inc. • <http://genetics.nature.com>

## A general approach to single-nucleotide polymorphism discovery

Gabor T. Marth<sup>1</sup>, Ian Korf<sup>1</sup>, Mark D. Yandell<sup>1</sup>, Raymond T. Yeh<sup>1</sup>, Zhijie Gu<sup>2</sup>, Hamideh Zakeri<sup>2</sup>, Nathan O. Stitzel<sup>1</sup>, LaDeana Hillier<sup>1</sup>, Pui-Yan Kwok<sup>2</sup> & Warren R. Gish<sup>1</sup>

Bayesian posterior probability



$$P(\text{SNP}) = \boxed{\dots}$$

$$\sum_{\text{all variable } S} \frac{\frac{P(S_1 | R_1)}{P_{\text{Prior}}(S_1)} \cdot \dots \cdot \frac{P(S_N | R_N)}{P_{\text{Prior}}(S_N)} \cdot P_{\text{Prior}}(S_1, \dots, S_N)}{\sum_{S_{i_1} \in [A,C,G,T]} \dots \sum_{S_{i_N} \in [A,C,G,T]} \frac{P(S_{i_1} | R_1)}{P_{\text{Prior}}(S_{i_1})} \cdot \dots \cdot \frac{P(S_{i_N} | R_1)}{P_{\text{Prior}}(S_{i_N})} \cdot P_{\text{Prior}}(S_{i_1}, \dots, S_{i_N})}$$

Base call + Base quality



Expected (prior) polymorphism rate



Probability of observed base composition  
(should model sequencing error rate)

# Genome Analysis Toolkit (GATK)

---

NATURE GENETICS | TECHNICAL REPORT



日本語要約

## A framework for variation discovery and genotyping using next-generation DNA sequencing data

**Mark A DePristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, Guillermo del Angel, Manuel A Rivas, Matt Hanna, Aaron McKenna, Tim J Fennell, Andrew M Kernytsky, Andrey Y Sivachenko, Kristian Cibulskis, Stacey B Gabriel, David Altshuler & Mark J Daly**

[Affiliations](#) | [Contributions](#) | [Corresponding author](#)

*Nature Genetics* **43**, 491–498 (2011) | doi:[10.1038/ng.806](https://doi.org/10.1038/ng.806)

Received 27 August 2010 | Accepted 17 March 2011 | Published online 10 April 2011

# GATK's HaplotypeCaller

---

Calls SNPs and INDELs simultaneously

Performs local re-assembly to identify haplotypes

More accurate (esp. indels) compared to locus (site by site) callers

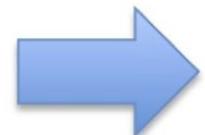
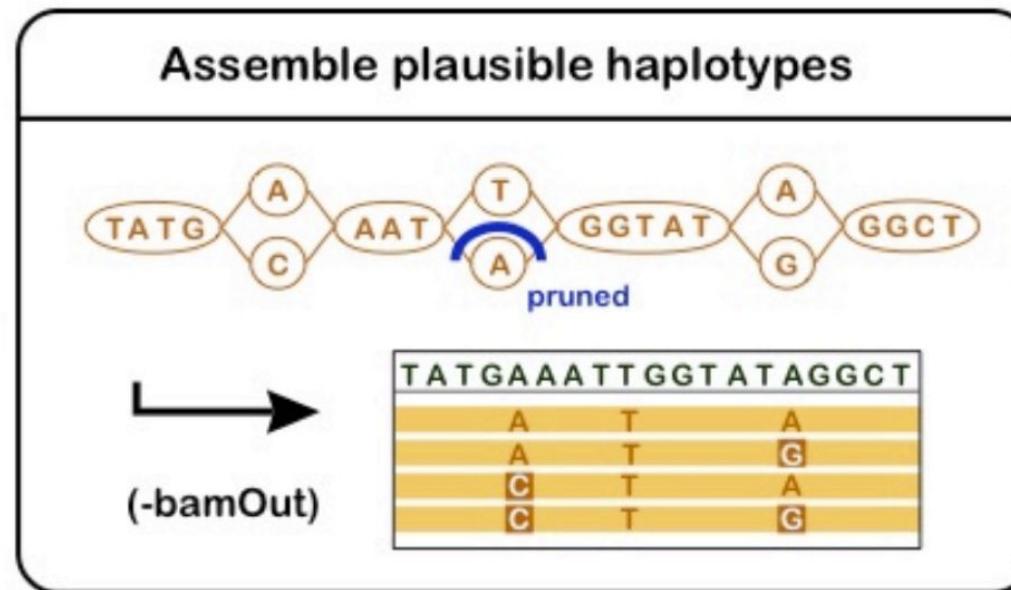
Reference confidence model enables joint-discovery workflow

Handles non-diploid organisms and pooled samples

# GATK's HaplotypeCaller

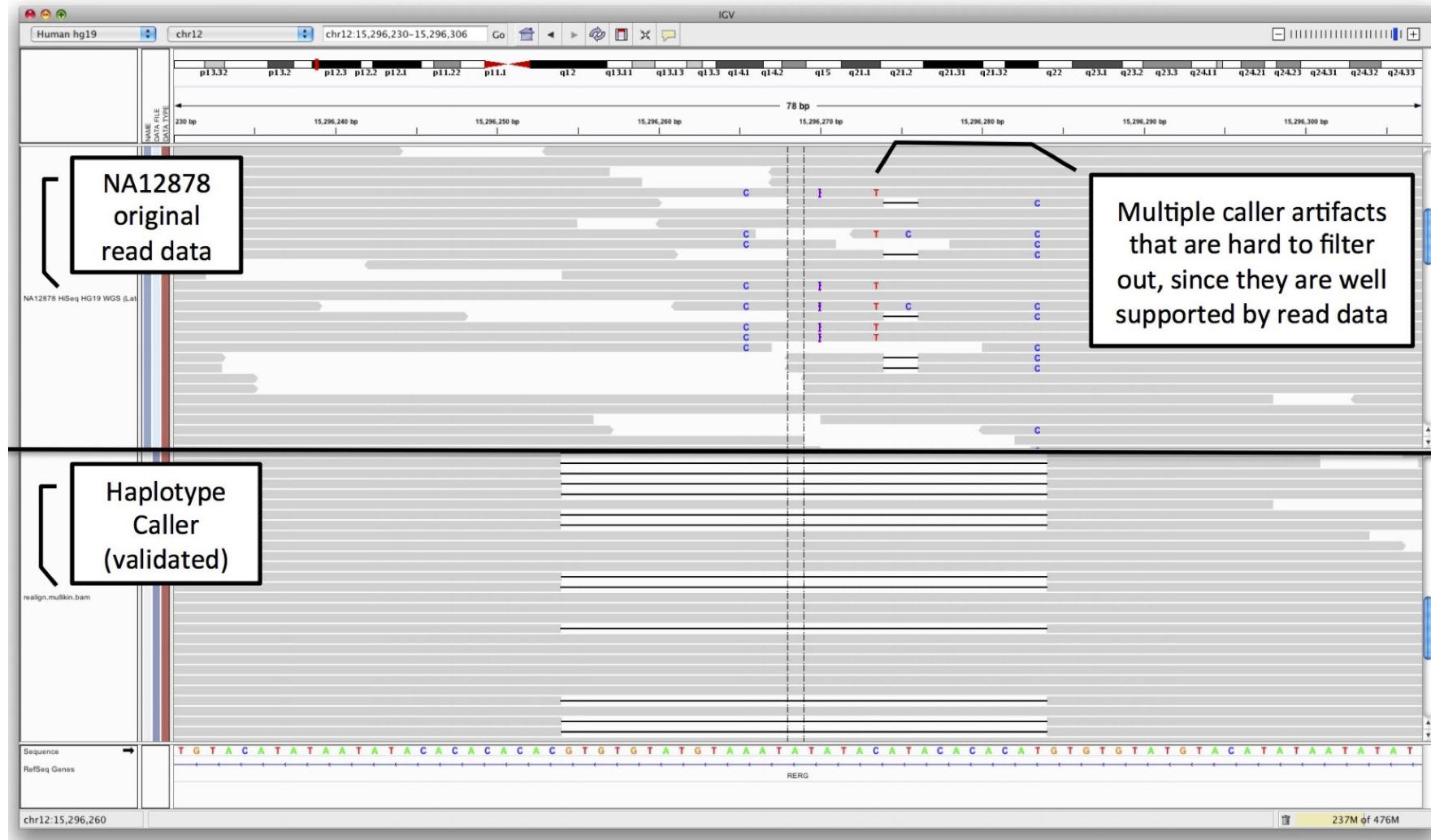
---

- Local re-assembly
- Traverse graph to collect most likely haplotypes
- Align haplotypes to ref using Smith-Waterman



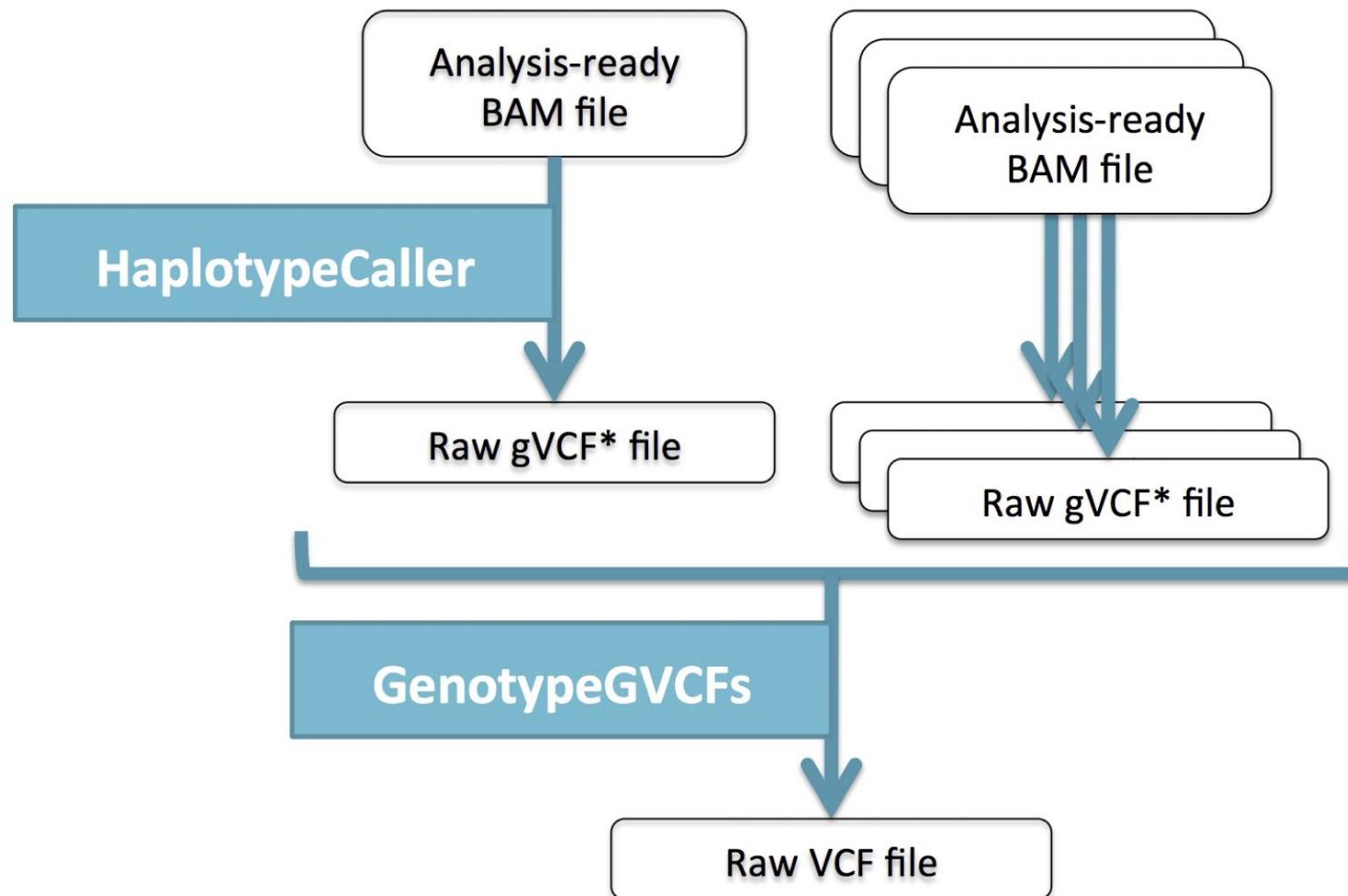
**Likely haplotypes + candidate variant sites**

# GATK's HaplotypeCaller



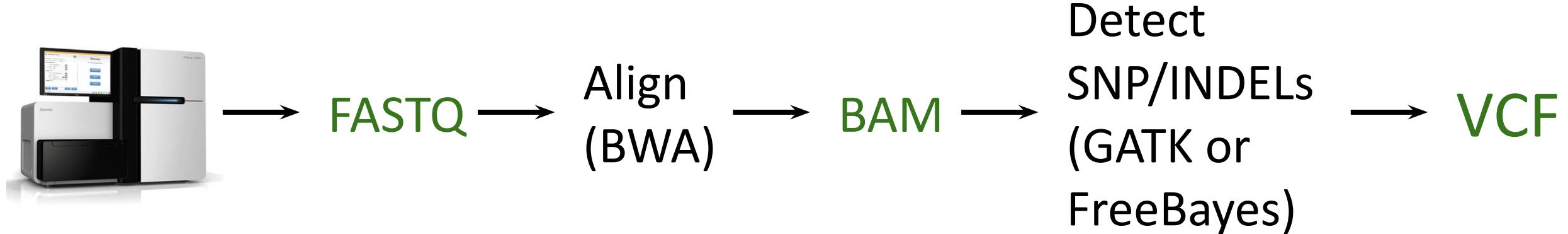
# GATK's HaplotypeCaller

---



# Variant Calling

---



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

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

# VCF Format

## Example

**VCF header**

```

##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1 SAMPLE2
  
```

**Body**

CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE1	SAMPLE2
1	1	.	ACG	A,AT	.	PASS	.	GT:DP	1/2:13	0/0:29
1	2	rs1	C	T,CT	.	PASS	H2;AA=T	GT:GQ	0 1:100	2/2:70
1	5	.	A	G	.	PASS	.	GT:GQ	1 0:77	1/1:95
1	100	.	T	<DEL>	.	PASS	SVTYPE=DEL;END=300	GT:GQ:DP	1/1:12:3	0/0:20

**Mandatory header lines**

**Optional header lines (meta-data about the annotations in the VCF body)**

**Reference alleles (GT=0)**

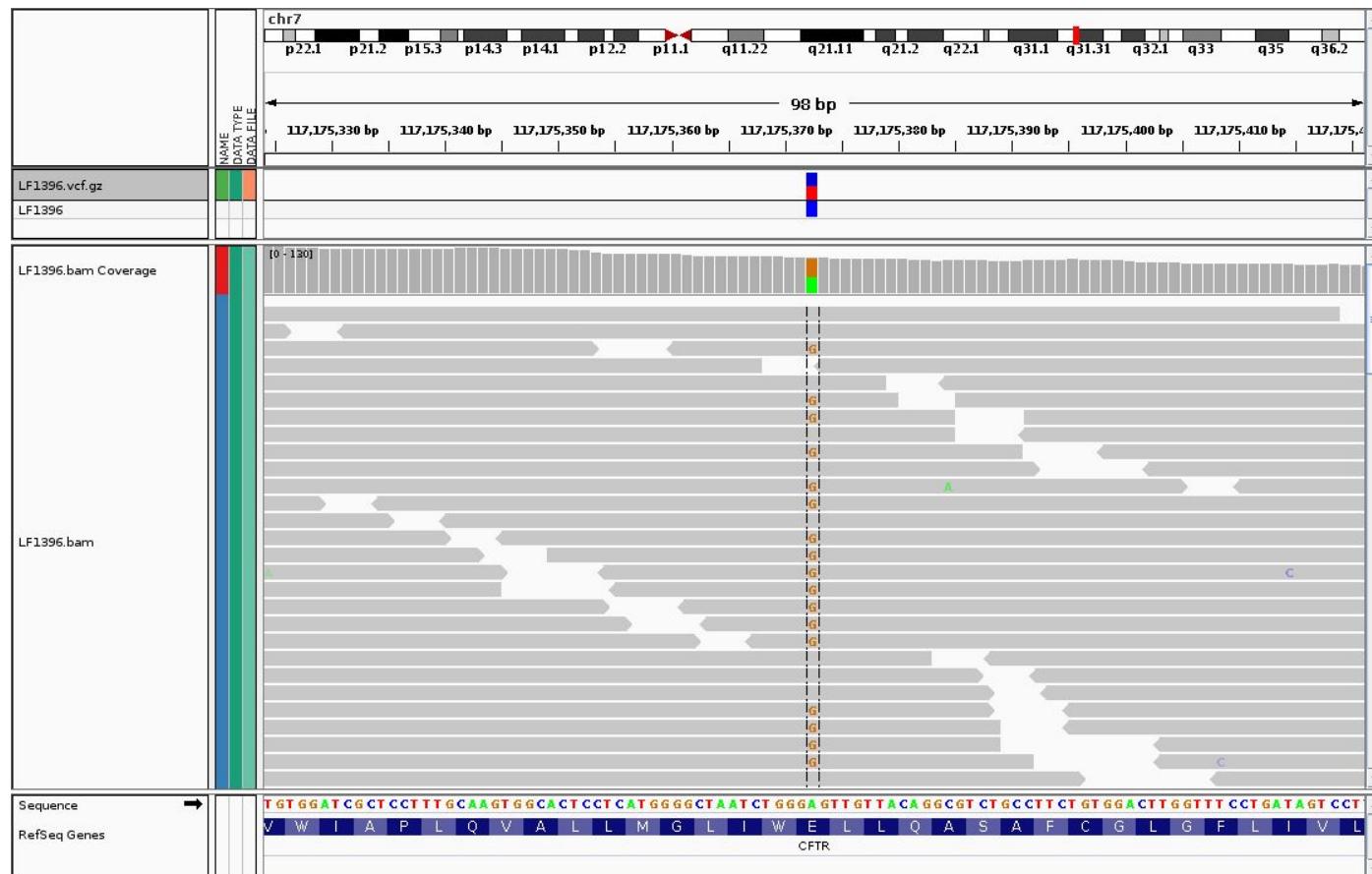
**Alternate alleles (GT>0 is an index to the ALT column)**

**Phased data (G and C above are on the same chromosome)**

**Annotations:**

- Deletion**: Variant at position 100.
- SNP**: Variants at positions 1, 2, and 5.
- Large SV**: Variant at position 100 labeled as a deletion (<DEL>).
- Insertion**: Variant at position 2 labeled as rs1.
- Other event**: Variant at position 5.

# VCF Format



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

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