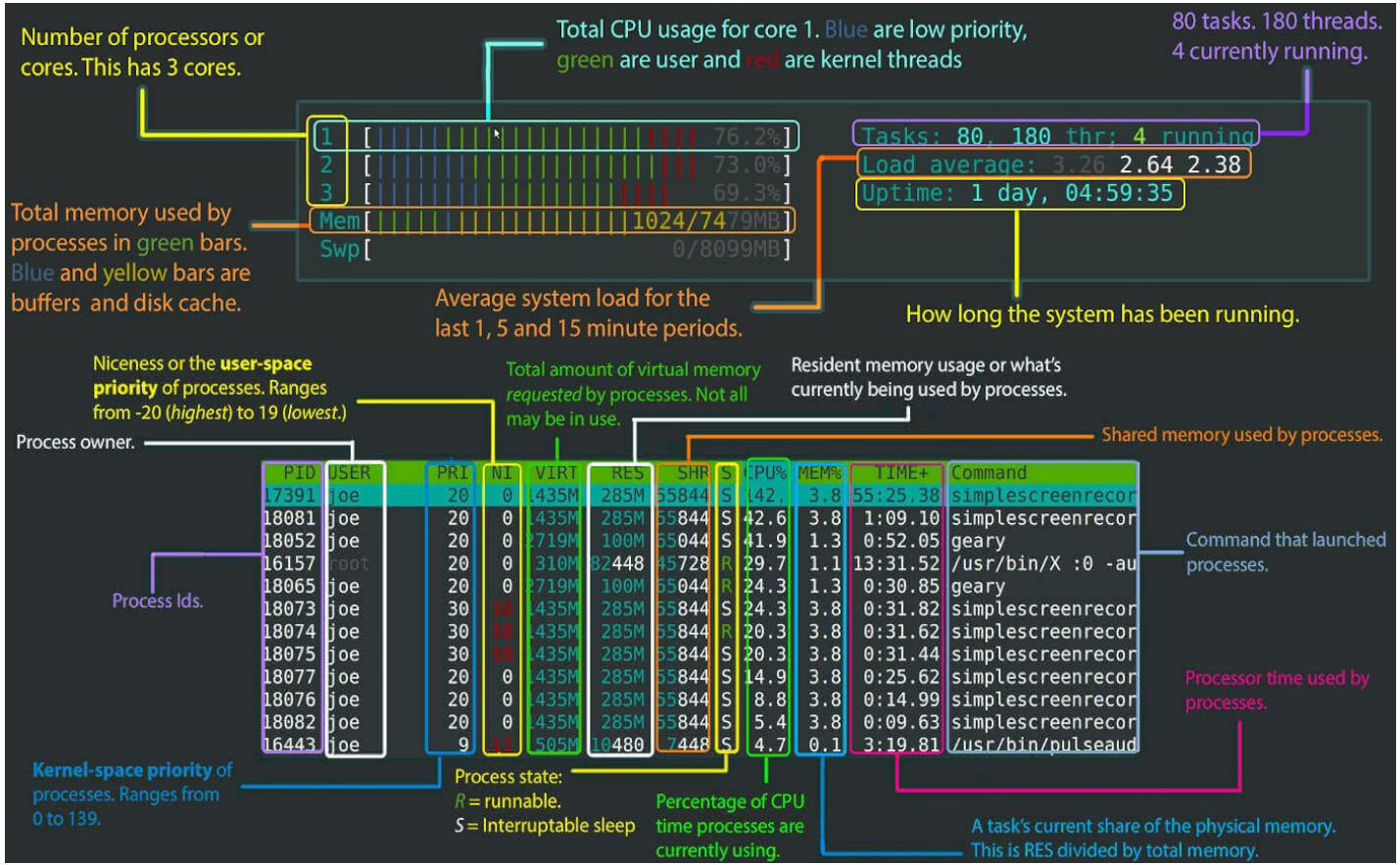
Lesson 6

Variant calling

htop/top


Number of processors or cores. This has 3 cores.
Total CPU usage for core 1. Blue are low priority, green are user and red are kernel threads
80 tasks. 180 threads. 4 currently running.
1 [76.2%]
2 [73.0%]
3 [69.3%]
Mem[1024/7479MB]
Swp[0/8099MB]
Tasks: 80, 180 thr; 4 running
Load average: 3.26 2.64 2.38
Uptime: 1 day, 04:59:35
Total memory used by processes in green bars. Blue and yellow bars are buffers and disk cache.
Average system load for the last 1, 5 and 15 minute periods.
How long the system has been running.
Niceness or the user-space priority of processes. Ranges from -20 (highest) to 19 (lowest.)
Total amount of virtual memory requested by processes. Not all may be in use.
Resident memory usage or what's currently being used by processes.
Process owner.
Shared memory used by processes.
PID USER PRI NI VIRT RES SHR S CPU% MEM% TIME+ Command
17391 joe 20 0 1435M 285M 55844 S 142. 3.8 55:25.38 simplescreenrecor
18081 joe 20 0 1435M 285M 55844 S 42.6 3.8 1:09.10 simplescreenrecor
18052 joe 20 0 2719M 100M 65044 S 41.9 1.3 0:52.05 geary
16157 root 20 0 310M 82448 45728 R 29.7 1.1 13:31.52 /usr/bin/X :0 -au
18065 joe 20 0 2719M 100M 65044 R 24.3 1.3 0:30.85 geary
18073 joe 30 10 1435M 285M 55844 S 24.3 3.8 0:31.82 simplescreenrecor
18074 joe 30 10 1435M 285M 55844 R 20.3 3.8 0:31.62 simplescreenrecor
18075 joe 30 10 1435M 285M 55844 S 20.3 3.8 0:31.44 simplescreenrecor
18077 joe 20 0 1435M 285M 55844 S 14.9 3.8 0:25.62 simplescreenrecor
18076 joe 20 0 1435M 285M 55844 S 8.8 3.8 0:14.99 simplescreenrecor
18082 joe 20 0 1435M 285M 55844 S 5.4 3.8 0:09.63 simplescreenrecor
16443 joe 9 -11 505M 10480 7448 S 4.7 0.1 3:19.81 /usr/bin/pulseaud
Command that launched processes.
Process Ids.
Processor time used by processes.
Kernel-space priority of processes. Ranges from 0 to 139.
Process state:
R = runnable.
S = Interruptable sleep
Percentage of CPU time processes are currently using.
A task's current share of the physical memory. This is RES divided by total memory.

By the end of this lesson you will...

- Understand the basic concepts of variant calling
  - Short variants
  - Structural variants
- Know how to perform short variant calling with bcftools / GATK
- Be able to interpret and work with the VCF format
- Know how to perform structural variant calling with Manta
- Know how to work with VCF files in IGV

How are they different?

Genetic Variation

Differences in DNA content or structure among individuals

- Any two individuals have ~99.8% identical DNA.

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

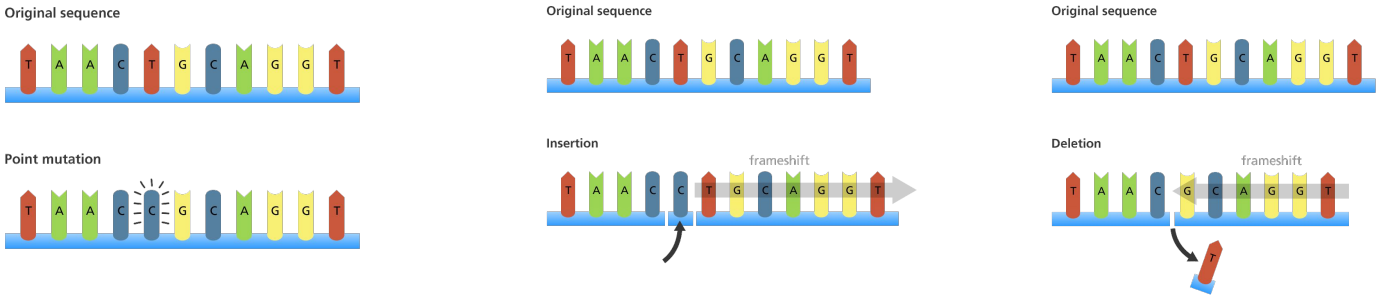
There are >100,000,000 know genetic variants in the human genome

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

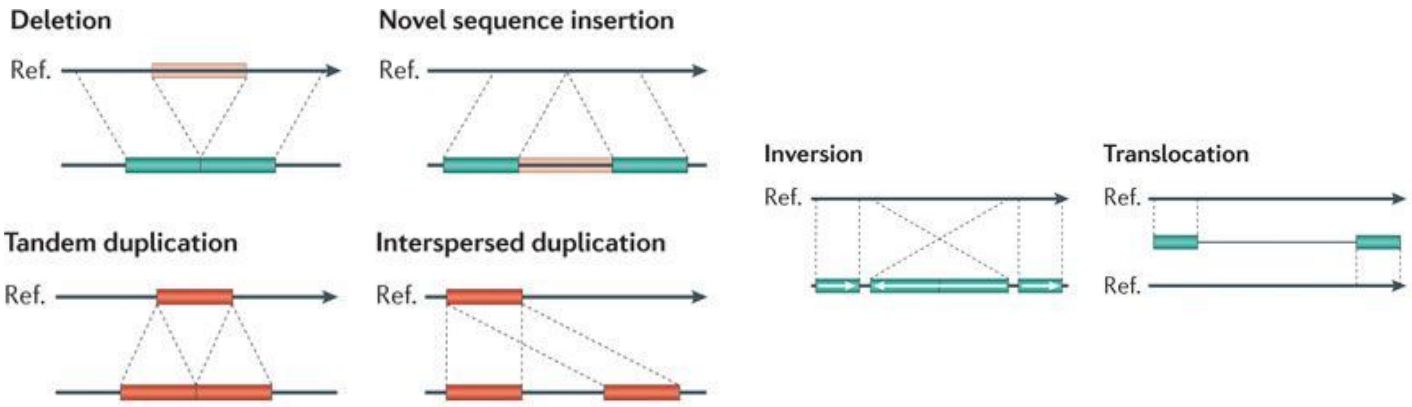
99% identical DNA

Types of genetic variation

Small-scale



Large scale (Structural)



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.

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

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        acctccgagta
acctccgagta        acctccgagta
acctccgagta        acctccgagta
acctccgagta        acctccgagta
acctccgagta        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

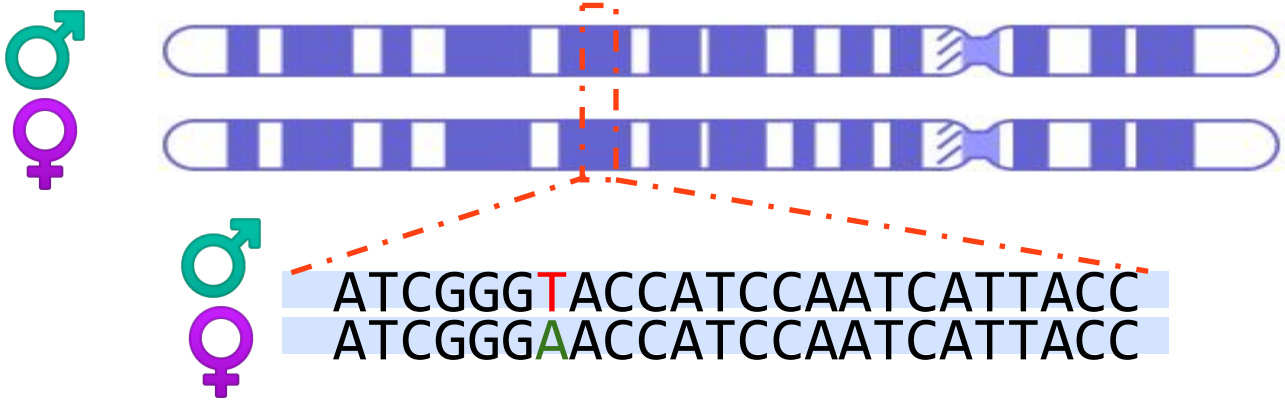
Mutations

From mutation to polymorphism

```
acctccgagta
acctccgagta
acctccgagta
acctcTgagta
acctccgagta
```

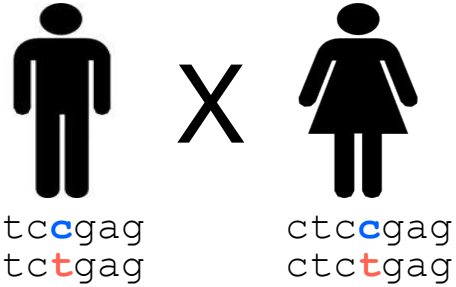
```
acctcTgagta
acctccgagta
acctcTgagta
acctccgagta
acctcTgagta
```

Diploid Genomes

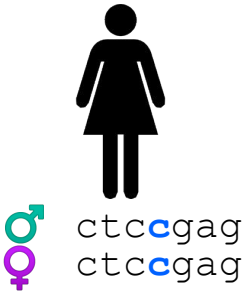


Our genome is comprised of a paternal and a maternal "haplotype".
Together, they form our "genotype"

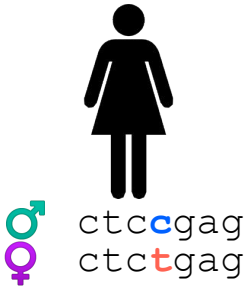
Inherited Germline Variation



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

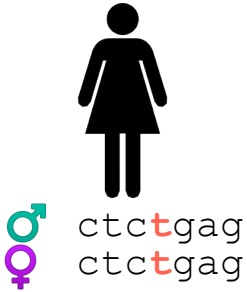


Kid is homozygous (C/C)



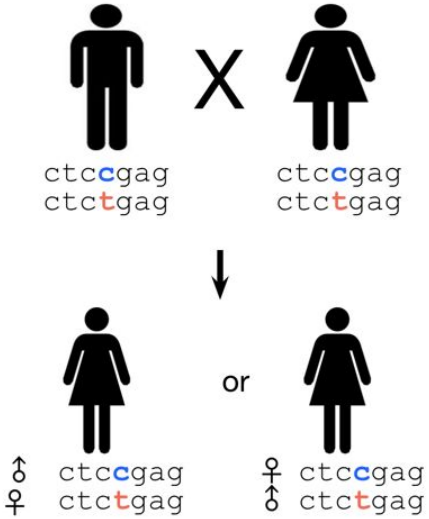
or

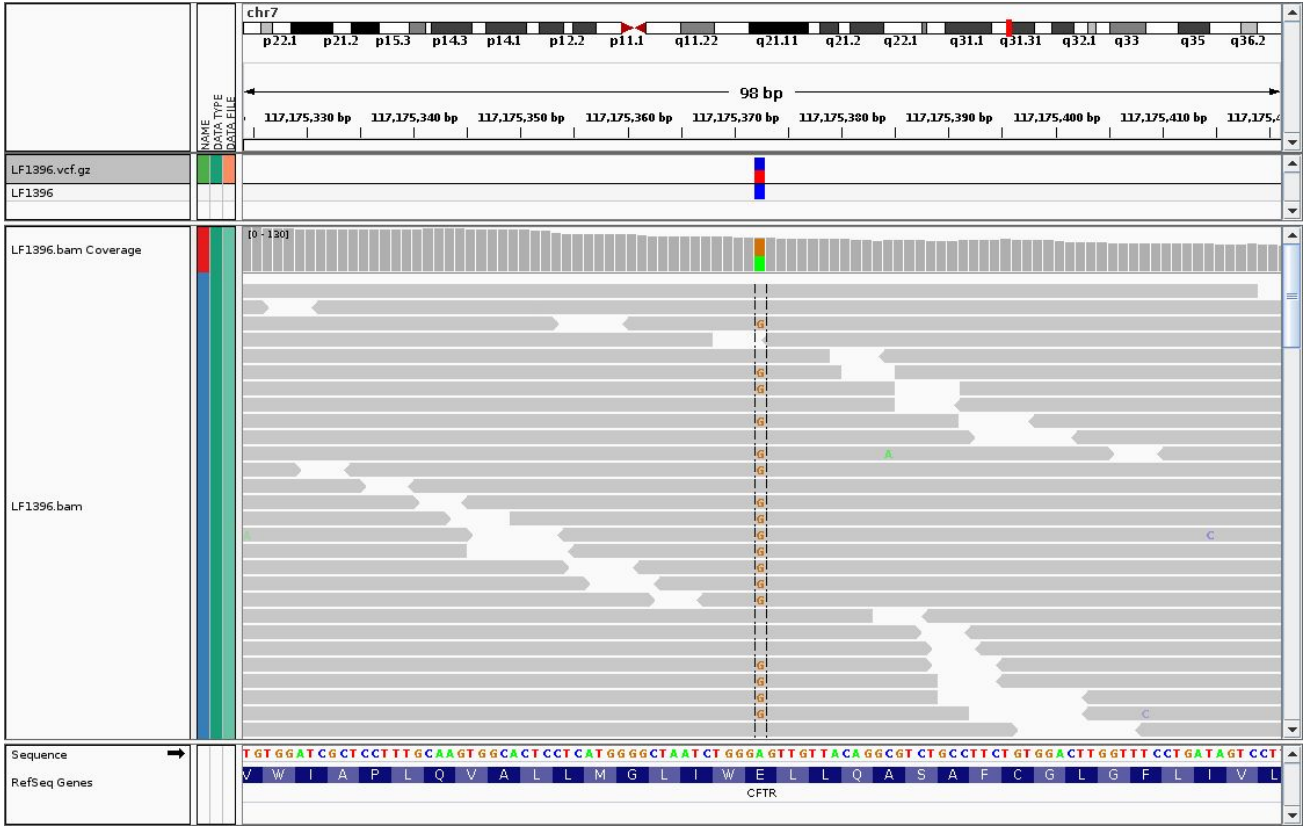
Kid is heterozygous (C/T)



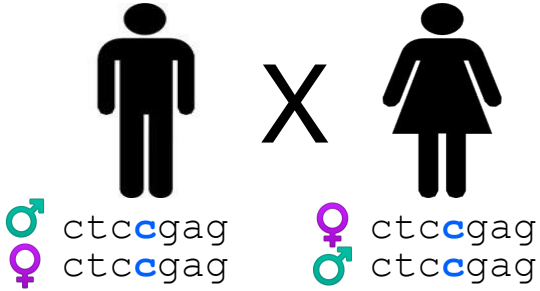
Kid is homozygous (T/T)

Heterozygous Variation





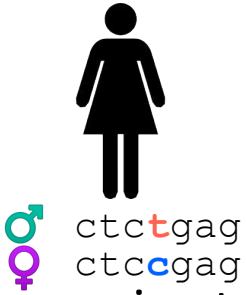
De novo Mutation



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

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

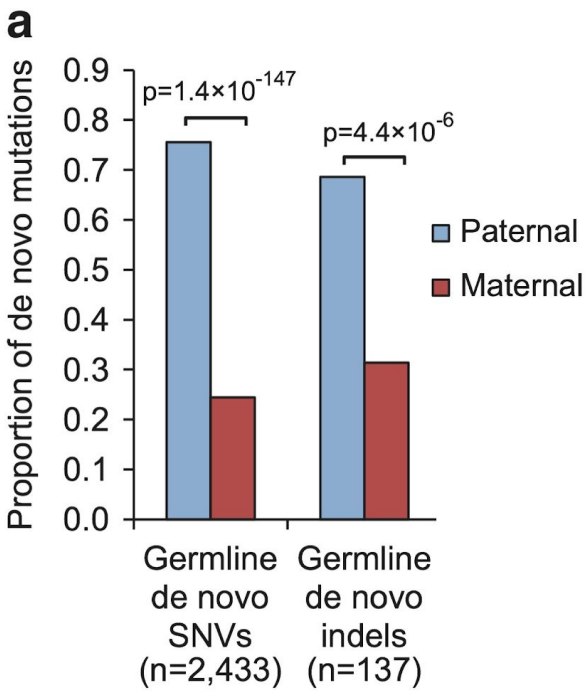
Note: This is a derivative chromosome of the one the father inherited from His parents



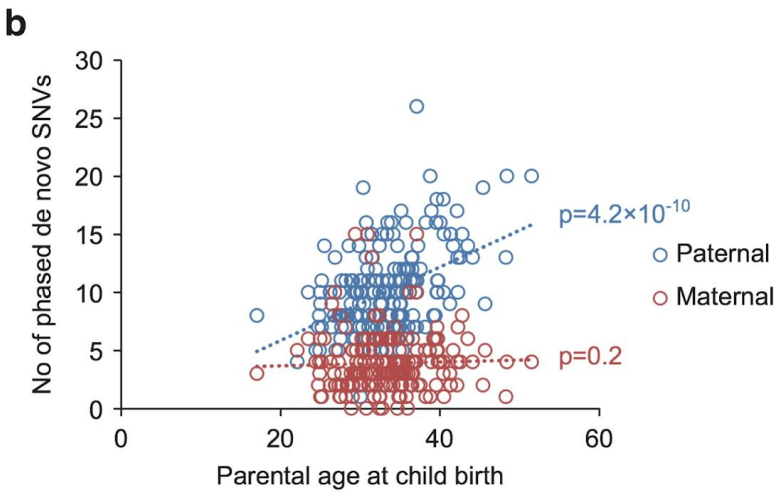
Kid is heterozygous owing to *de novo mutation.*

(C/T)

DNMs Frequency



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



(data from 200 ASD trios)

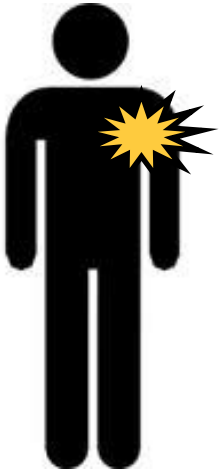
Yuen et al. (2016) *Nature Genomic Medicine*

Tom Sasani

Somatic Mutations

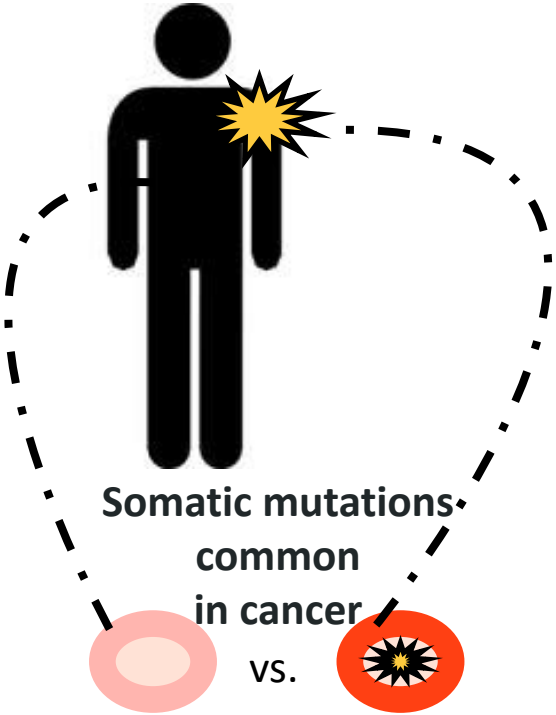
Germline mutation

- occur in sperm or egg.
- are heritable



Somatic mutation

- non-germline tissues.
- **<u>are not heritable</u>**



compare DNA from cancer cells to healthy cells from same individual

The 1000 (2504) Genome Project

2,504 individuals from diverse ancestries

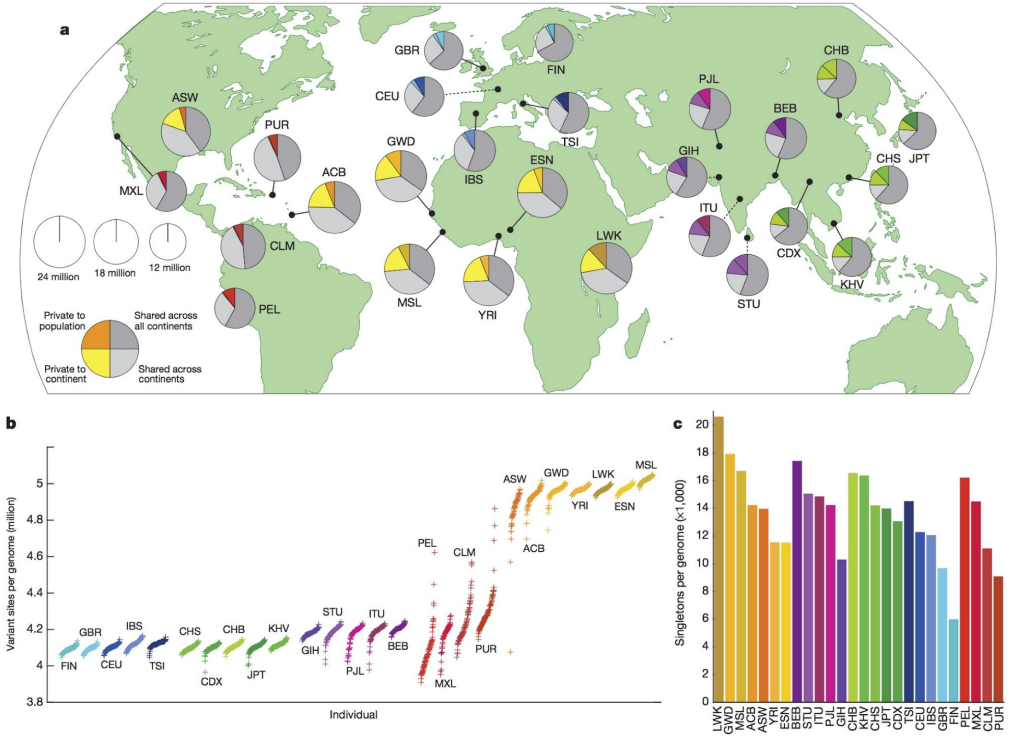


Figure 1 | Population sampling. a, Polymorphic variants within sampled populations. The area of each pie is proportional to the number of polymorphisms within a population. Pies are divided into four slices, representing variants private to a population (darker colour unique to population), private to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

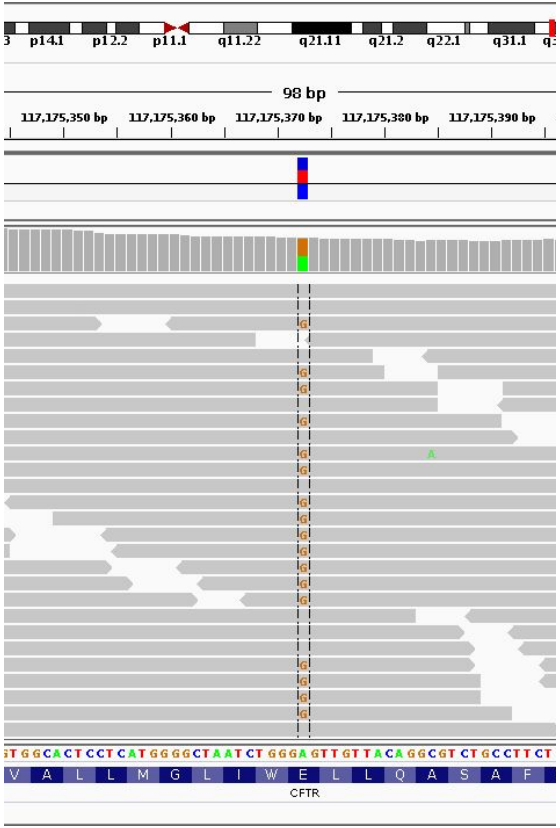
The extent of genetic variation by subpopulation

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	Var. sites	Singletons	Var. sites	Singletons	Var. sites	Singletons	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.

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

Why do variant calling?

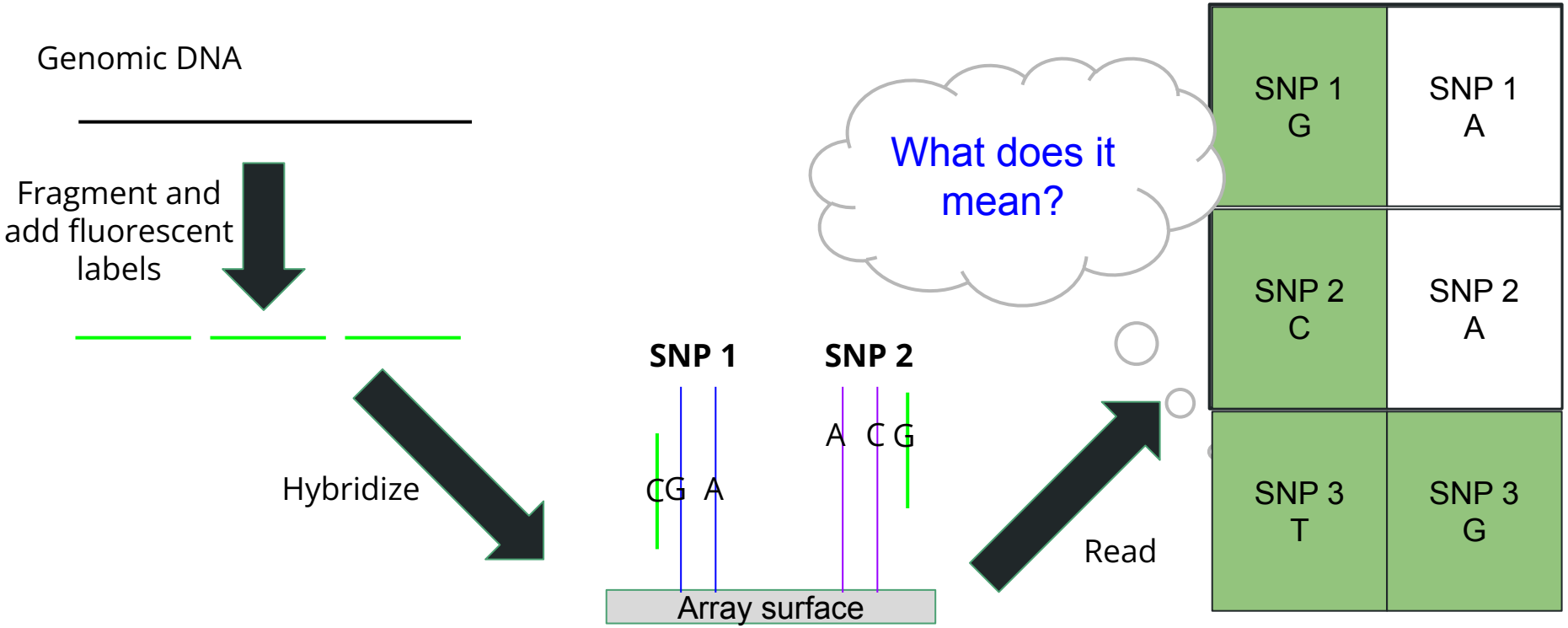
- Estimate variation within a population
- Correlate variation (and loci) with phenotypes - GWAS/QTL-mapping
  - Genetic diseases
  - Cancer genomics
  - Crop/animal breeding
- Study population structure and evolution
- Diagnostics - personalized medicine

Variant calling terminology

- Identification of variants from sequence data
- We usually define one sample as the **reference**
- **Reference / alternative allele**
- **Minor / major allele frequency**
- **Biallelic / multiallelic variation**

```
Reference     AGCTTAGCCAGGGATCGCTA
Sample 1      AGCATAGCCAGGGATAGCTA
Sample 2      AGCTTAGCCAGGGATAGCTA
Sample 3      AGCTTAGCCGGGGATAGCTA
Sample 4      AGCATAGCCTGGGATAGCTA
Sample 5      AGCTTAGCCAGGGATCGCTA
```

SNP arrays


Genomic DNA
Fragment and add fluorescent labels
Hybridize
SNP 1
SNP 2
CG A
A C G
Array surface
Read
What does it mean?
SNP 1 G
SNP 1 A
SNP 2 C
SNP 2 A
SNP 3 T
SNP 3 G

Short variant calling from NGS data

Inputs:

- Reference genome
- Sequencing data from one or more samples

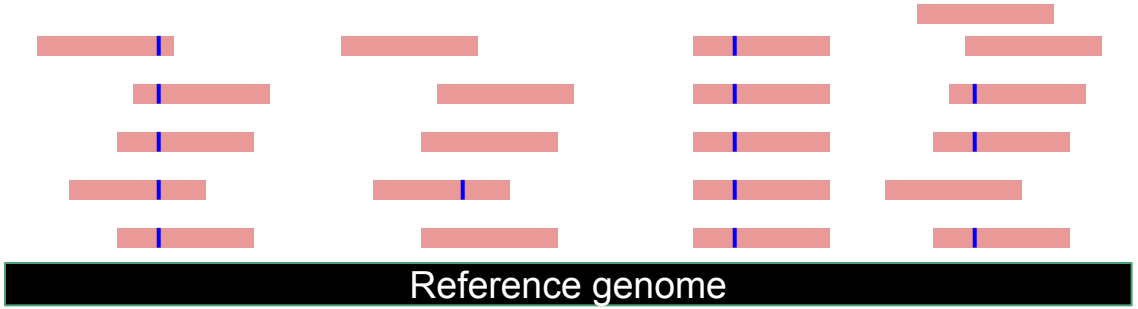
Step 1: map reads from each sample to the reference

Step 2: Find mismatches and call variants

Step 3: QA and filter variants

Do we believe every mismatch we find?

- Sequencing errors
- Mapping errors
- Need to consider:
  - Base quality (Phred)
  - Mapping quality (MAPQ)
  - **Number of reads (depth) supporting the variant**
  - Heterozygous variants
  - Degree of reliability



Simple variant calling with bcftools

- The `bcftools mpileup` summarizes base calls at each position
- Input:
  - A reference genome fasta (must be indexed with `samtools faidx`)
  - One or more bam files
- Output:
  - File in *special* (pileup) vcf format

Basic usage:

```
$ samtools faidx ref.fasta

$ bcftools mpileup -f ref.fasta sample1_vs_ref.bam -o
sample1_vs_ref.mpileup.vcf
```

Pileup - general idea

Chr	Position	Ref	A	T	G	C
Chr1	1	A	7	0	0	0
Chr1	2	G	0	0	12	1
Chr1	3	G	0	0	8	0
Chr1	4	T	0	5	0	6
Chr1	5	A	10	0	0	0
Chr1	6	C	0	0	0	11
Chr1	...					

bcftools pileup - more useful options

- Skip alignments with MAPQ lower than x

```
bcftools mpileup -q <x>...
```

- Skip bases with Phred score lower than x

```
bcftools mpileup -Q <x>...
```

- Only output positions on chromosome x

```
bcftools mpileup -r <Chr_x>...
```

Calling variants with bcftools call

- Input: mpileup VCF file
- Output: variants VCF file
- Important options:
  - -v - only print positions where variants exist
  - -m - the recommended method for variant calling
- Basic usage:

```
bcftools call -mv res.mpileup.vcf -o res.vcf
```

The VCF format - Variant Call Format

- A TSV file (with a special format)
- Consists of header lines and body lines
- Header lines start with #
- Body line consist of 9 mandatory fields (but more can be added)
- Each line represents a variant in a genomic position
- Additional fields are added per sample to describe genotypes

VCF fields

	Name	Description
1	CHROM	Reference chromosome
2	POS	Position on reference chromosome (starting from 1)
3	ID	Variant ID - usually empty (.)
4	REF	Reference allele
5	ALT	Alternative alleles, separated by ,
6	QUAL	Inference quality score of the variant
7	FILTER	List of filters the variant had passed - usually empty (.)
8	INFO	Additional information about the variant
9	FORMAT	Specification of the genotypes format

VCF example

```
##fileformat=VCFv4.0
###fileDate=20090805
###source=myImputationProgramV3.1
###reference=1000GenomesPilot-NCBI36
###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=.,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  NA00001 NA00002 NA00003
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
```

Sample genotypes

VCF - ID strings

- Different fields can use certain shortcuts to refer to some value
- Makes the file more compact
- IDs are defined in the VCF header
- Commonly used IDs are further explained in the VCF documentation

VCF - ID strings

```
##fileformat=VCFv4.0
###fileDate=20090805
###source=myImputationProgramV3.1
###reference=1000GenomesPilot-NCBI36
###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=.,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  NA00001 NA00002 NA00003
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
```

VCF sample genotypes

- One field per sample
- The FORMAT field defines how genotype fields look
- The genotype itself is stored in the **GT** ID
- It refers to the REF and ALT alleles by number
  - 0 - reference allele
  - 1,2,... - alternative alleles
- Can describe diploids:
  - 0|0 - REF homozygous
  - 0|1 - heterozygous
  - 1|1, 2|2, ... - ALT homozygous
  - 1|2 - ALT heterozygous
- Unknown genotype - ‘.’

VCF sample genotypes

```
##fileformat=VCFv4.0
###fileDate=20090805
###source=myImputationProgramV3.1
###reference=1000GenomesPilot-NCBI36
###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=.,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  NA00001 NA00002 NA00003
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
```

Test yourself…

```
##fileformat=VCFv4.0
###fileDate=20090805
###source=myImputationProgramV3.1
###reference=1000GenomesPilot-NCBI36
###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=.,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  NA00001 NA00002 NA00003
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
```

What is the genotype of sample NA00002 in position 1110696 on chr20 ?

VCF QUAL and FILTER

- QUAL field indicates reliability of variant existence
- QUAL = $-10 \log_{10} \Pr\{\text{ALT call is wrong}\}$
- There are other quality scores for each genotype

- FILTER field describes what filters a variant passed or failed
- Filters are listed in the header
- Listed filters are those that **failed**
- “PASS” means all filters were passed
- “.” means no filters were applied

What can we do with a VCF file?

How many variants:

```
$ bcftools view -H human.chr22.vcf | wc -l
1103547
```

How many samples:

```
$ bcftools query -l human.chr22.vcf | wc -l
2504
```

How many biallelic SNPs variants:

```
$ bcftools view -H -m2 -M2 -v snps human.chr22.vcf | wc -l
1055454
```

How many variants with minor allele frequency > 5%:

```
$ bcftools view -H -q 0.05:minor human.chr22.vcf | wc -l
111090
```

Visualizing VCF files in IGV

Allele frequencies

One track per sample

Grey = REF|REF ; Dark blue = REF|ALT ; Light blue = ALT|ALT

Variant filtration

- Not all output variant calls are reliable
- Look at the data! (load BAM + VCF to IGV)
- Manual filtration based on:
  - Quality
  - Depth
- Use `bcftools view -i`
- Filter by intersecting with a known set of variants
  - Use `bcftools isec`
- Use machine learning to differentiate real variants from noise
  - E.g. GATK CNNvariant module

Filtration examples

How many variants:

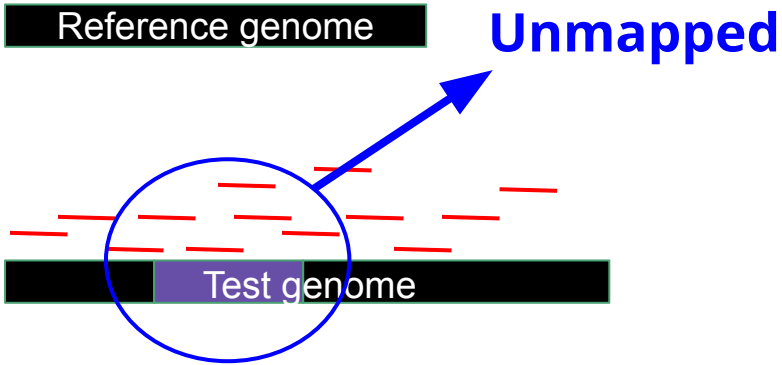
```
$ bcftools view -H human.chr22.vcf | wc -l
1103547
```

How many variants with QUAL > 30 and depth > 5

```
$ bcftools view -i "QUAL>30 & DP>5" human.chr22.vcf >
human.chr22.HQ.vcf
$ bcftools view -H human.chr22.HQ.vcf | wc -l
1101266
```

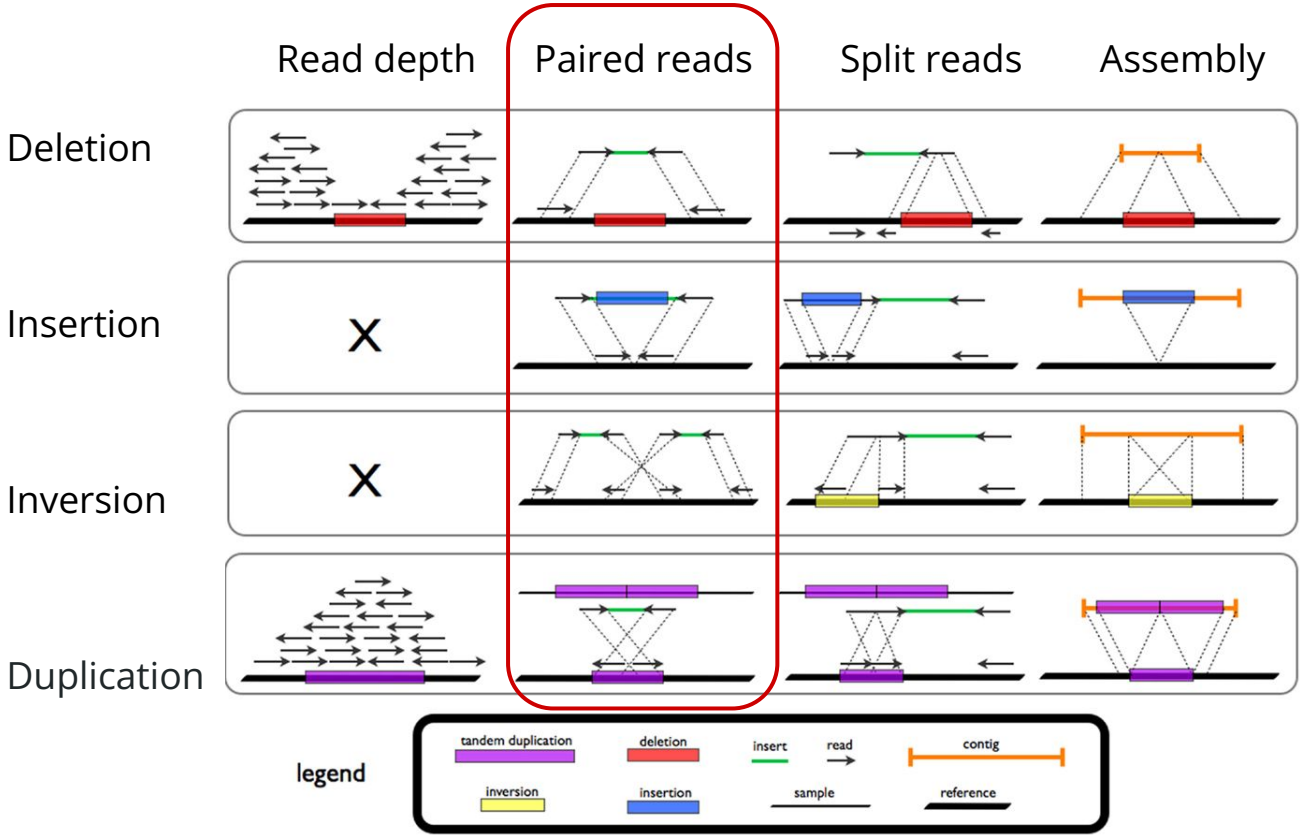
Calling structural variants

- The main challenge: short reads



- Use long reads
- Assemble short reads

SV calling from short reads



1. Tattini, L., D'Aurizio, R., & Magi, A. (2015). Detection of genomic structural variants from next-generation sequencing data. *Frontiers in bioengineering and biotechnology*, *3*, 92.

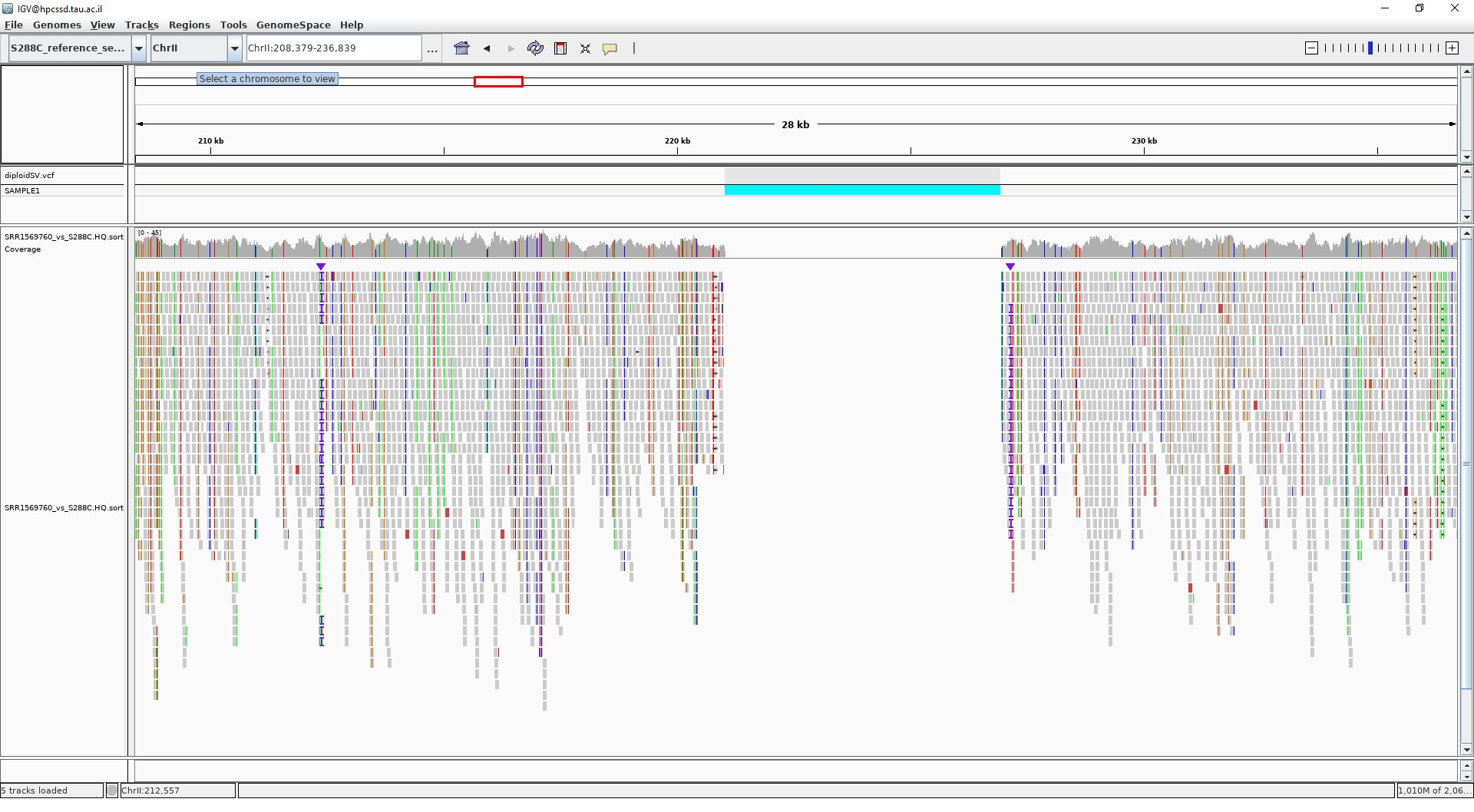
Calling SVs with Manta

- Developed and maintained by Illumina
- Combines paired-read and split-read evidence
- Detects all types of SVs
- Inputs:
  - sorted and indexed BAM file/s from a paired-end library
  - Reference genome - fasta
- Output: VCF file(s)
- Run includes two steps:

```
# 1) configure
$ configManta.py --bam sample1.bam --bam sample2.bam --bam sample3.bam
--referenceFasta ref_genomeme.fasta --runDir ./output
# 2) run
$ ./output/runWorkflow.py
```

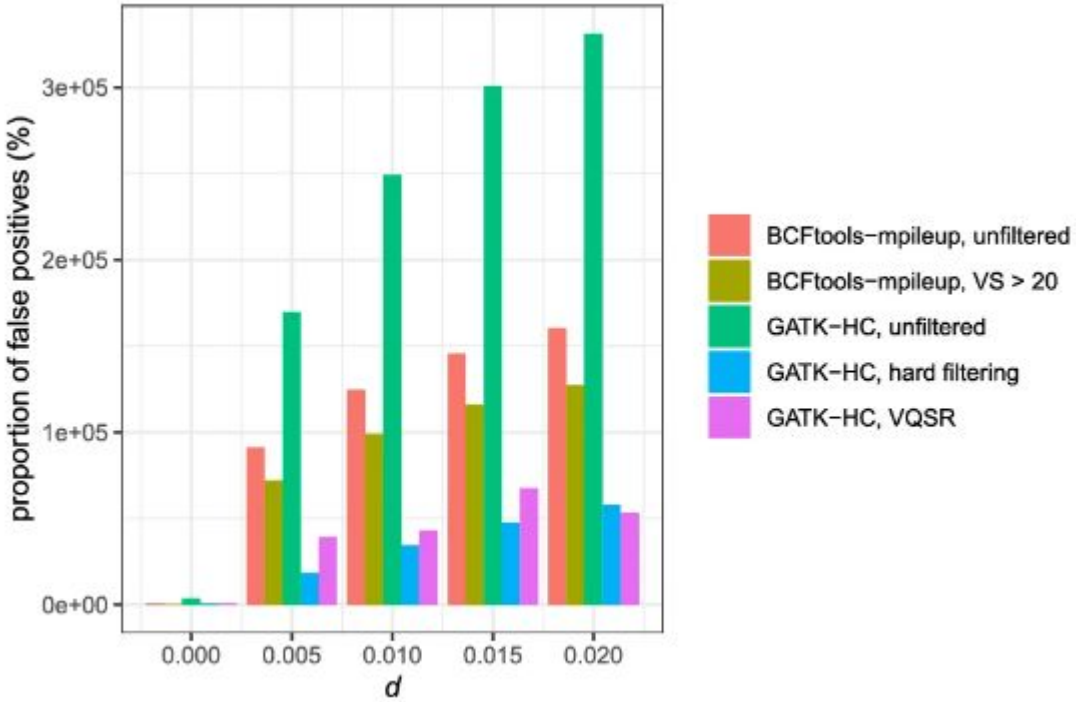
SV VCF

```
ChrII   9087    MantaDEL:49:0:1:0:0:0   TCCAATTGTTGGAATAAAAATCAACTATCATCTACTAACTAGTATTTACGTTACTAGTATATTATCATATACGGTGTTAGAAG
ATGACGCAAATGATGAGAAATAGTCATCTAAATTAGTGGAAGCTGAAACGCAAGGATTGATAATGTAATAGGATCAATGAATATTAACATATAAAACGATGATAATAATATTTATAGAATTGT
GTAGAATTGCAGATTCCCTTTTATGGATTCCTAAATCCTGGAGGAGAACTTCTAGTATATCTACATACCTAATATTATAGCCTTAATCACAATGGAATCCCAACAATTACATCAAAATCCACA
TTCTCTACA      T       698     PASS    END=9424;SVTYPE=DEL;SVLEN=-337;CIGAR=1M337D;CIPOS=0,5;HOMLEN=5;HOMSEQ=CCAAT     GT:F
T:GQ:PL:PR:SR       1/1:PASS:43:751,46,0:0,2:0,15
ChrII   35845   MantaDEL:45:0:1:0:0:0   TGTCTCTGTTGATAATTAGAGGTTAAAAATTAGTATTAATGAAGAAGTGAGTACTGATCTTCTTATACTAAATAAGAGAGGTA
TATAAAACACACGCCGATTGGTCATATTAATTATGACCAATATAATAGTGATTCCGGTAGTTACTATACATTGATGTGACGACTCATATTCCTCATATATGTACCTACCATAACATGTTCAAC
TAATAGGTCTTTAACACAGCTTCAGTATTGTCTGAGCTTCTCGTTTAACATTCCTTCTGCAATAGGCGCAATCACACTTAAACGTATACGAGTTGTACATTAATATACGATGTAAGCATTAGA
TTGTTACCATAGCAACTCATGTCACTATTAATTACTCTCGTTCCAACA       T       508     PASS    END=36221;SVTYPE=DEL;SVLEN=-376;CIGAR=1M376D
;CIPOS=0,5;HOMLEN=5;HOMSEQ=GTCTC    GT:FT:GQ:PL:PR:SR       1/1:PASS:35:561,38,0:0,3:0,12
ChrII   221031  MantaDEL:52:0:1:0:1:0   T       <DEL>   415     NoPairSupport   END=226952;SVTYPE=DEL;SVLEN=-5921;CIPOS=0,5
;CIEND=0,5;HOMLEN=5;HOMSEQ=GGAAT     GT:FT:GQ:PL:PR:SR       1/1:PASS:27:468,30,0:0,0:0,11
ChrII   259571  MantaDEL:58:0:1:0:0:0   A       <DEL>   396     NoPairSupport   END=265493;SVTYPE=DEL;SVLEN=-5922;CIPOS=0,7
;CIEND=0,7;HOMLEN=7;HOMSEQ=AGTAATT   GT:FT:GQ:PL:PR:SR       1/1:PASS:24:449,27,0:0,0:0,9
ChrII   363846  MantaINS:63:0:0:0:0:0   G       GGCAGATAGAAACCATACTGATTCGCAGATAGAAACCATACTGATTC  396
     PASS    END=363846;SVTYPE=INS;SVLEN=69;CIGAR=1M69I;CIPOS=0,25;HOMLEN=25;HOMSEQ=GCAGATAGAAACCATACTGATTCGC        GT:FT:
GQ:PL:PR:SR       1/1:PASS:34:449,37,0:0,0:0,14
ChrII   643488  MantaDEL:101:0:0:0:1:0  TGTTGATAATTAGAGGTTAAAAATTAGTATTAATGAAGAAATAATTACTGATCTTCTTATACTAAATAAGAGAGGTATATAAA
ACACACGCCGATTGGTCATATTAATCATGACCAATATAATAGTGATTCCGGTAGTTACTATACATTGATGTGACGACTCATATTCCTCATATATGTACCTACCATAACATGTTCAACTAATAG
GTCTTTAACACAGCTTCAGTATTGTCTGAGCTTCTCGTTTAACATTCCTTCTGCAATAGGCGCAATCACACTTAAACGTATACGAGTTGTACATTAATATACGATGTAAGCATTGAATTGTTA
CCATAGCAACTCATGTCACTATTAATTACTCTCGTTCCAACATAATATTA     TTT     932     PASS    END=643866;SVTYPE=DEL;SVLEN=-378;CIGAR=1M2I3
78D GT:FT:GQ:PL:PR:SR       1/1:PASS:67:985,70,0:0,12:0,22
ChrIII  2650    MantaINS:115:0:0:0:2:0  A       <INS>   532     PASS    END=2650;SVTYPE=INS;CIPOS=0,6;CIEND=0,6;HOMLEN=6;HO
MSEQ=GCTGGG;LEFT_SVINSSEQ=GCTGGGAAGTTAAATAATTATCTTTACATTTTCAATGATCTTACGCCTGTAGG;RIGHT_SVINSSEQ=GGAACTTTCCGCGTTCAAAGATCACAAA
ACGTACAGACTGGTGTCTCTTCCAAAGCAAAGAAGAGCTCTTGGATCGC GT:FT:GQ:PL:PR:SR       1/1:PASS:33:585,36,0:0,0:0,13
ChrIII  82694   MantaDEL:107:0:1:0:0:0  TAGTACTGTTGGAATAGAAATCAACTATCATCTACTAACTAGTATTTACATTACTAGTATATTATCATATACGGTGTTAGAAG
ATGACGCAAATGATGAGAAATAGTCATCTAAATTAGTGGAAGCTGAAACGCAAGGATTGATAATGTAATAGGATCAATGAATATAAACATATAAAACGGAATGAGGAATAATCGTAATATTAG
TATGTAGAAATATAGATTCCATTTTGAGGATTCCTATATCCTCGAGGAGAACTTCTAGTATATTCTGTATACCTAATATTATAGCCTTTATCAACAATGGAATCCCAACAATTATCTCAACAT
TCACCCATTTCTCA T       362     PASS    END=83036;SVTYPE=DEL;SVLEN=-342;CIGAR=1M342D;CIPOS=0,6;HOMLEN=6;HOMSEQ=AGTACT   GT:F
T:GQ:PL:PR:SR       1/1:PASS:26:415,29,0:0,5:0,8
```


IGV@hpcssd.tau.ac.il
File Genomes View Tracks Regions Tools GenomeSpace Help
S288C_reference_se...
ChrII
ChrII:208,379-236,839
Select a chromosome to view
28 kb
210 kb
220 kb
230 kb
diploidSV.vcf
SAMPLE1
SRR1569760_vs_S288C.HQ.sort
Coverage
[0 - 45]
SRR1569760_vs_S288C.HQ.sort
5 tracks loaded
ChrII:212,557
1,010M of 2,06...

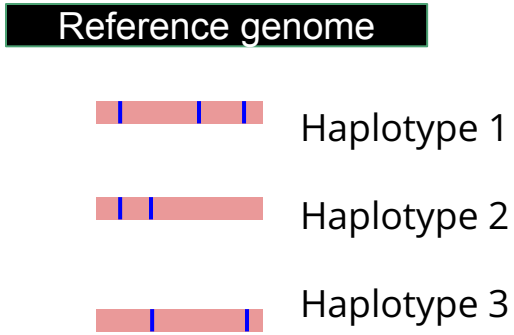
Haplotype-based variant calling with GATK

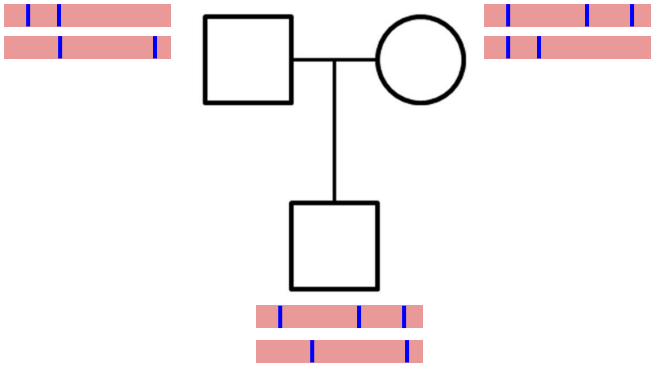
- GATK - Genome Analysis Tool Kit
- Developed by the Broad institute (MIT)



Haplotype-based variant calling with GATK

- GATK - Genome Analysis Tool Kit
- Developed by the Broad institute
- **Haplotype** - a block of genotypes inherited together from the same parent





How can haplotypes help us do variant calling?

Let’s say there are only two haplotypes in the population:

What can we learn from a read that looks like this?

GATK variant calling pipeline