DNA Sequencing and Data Analysis

Prof Noam Shomron

Amit Levon

Lecture 6, May 22, 2025

DNA Sequencing and Data Analysis

IGV
Variant Calling
VCF File Format
Long Read Technology

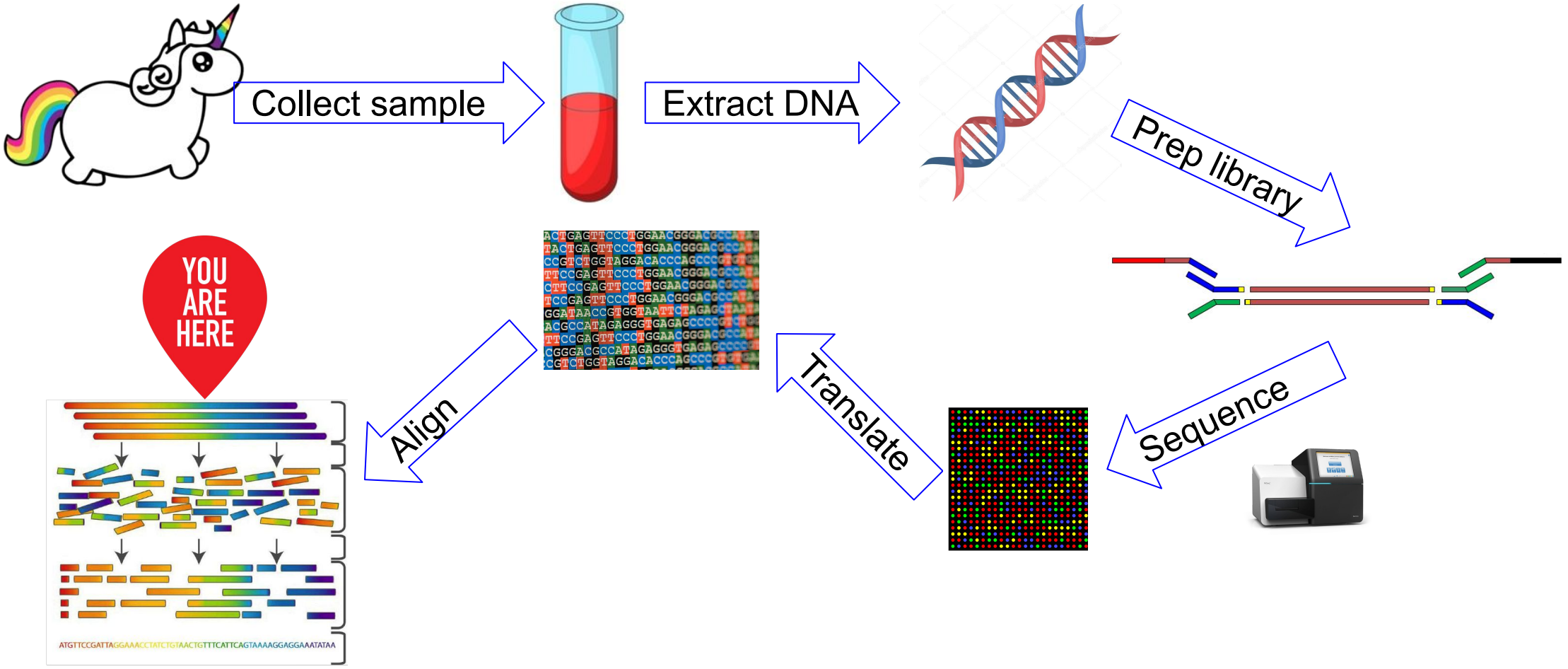
Thursday 18:30 to 21:00

C.L03

nshomron@gmail.com

amit.levon@post.runi.ac.il

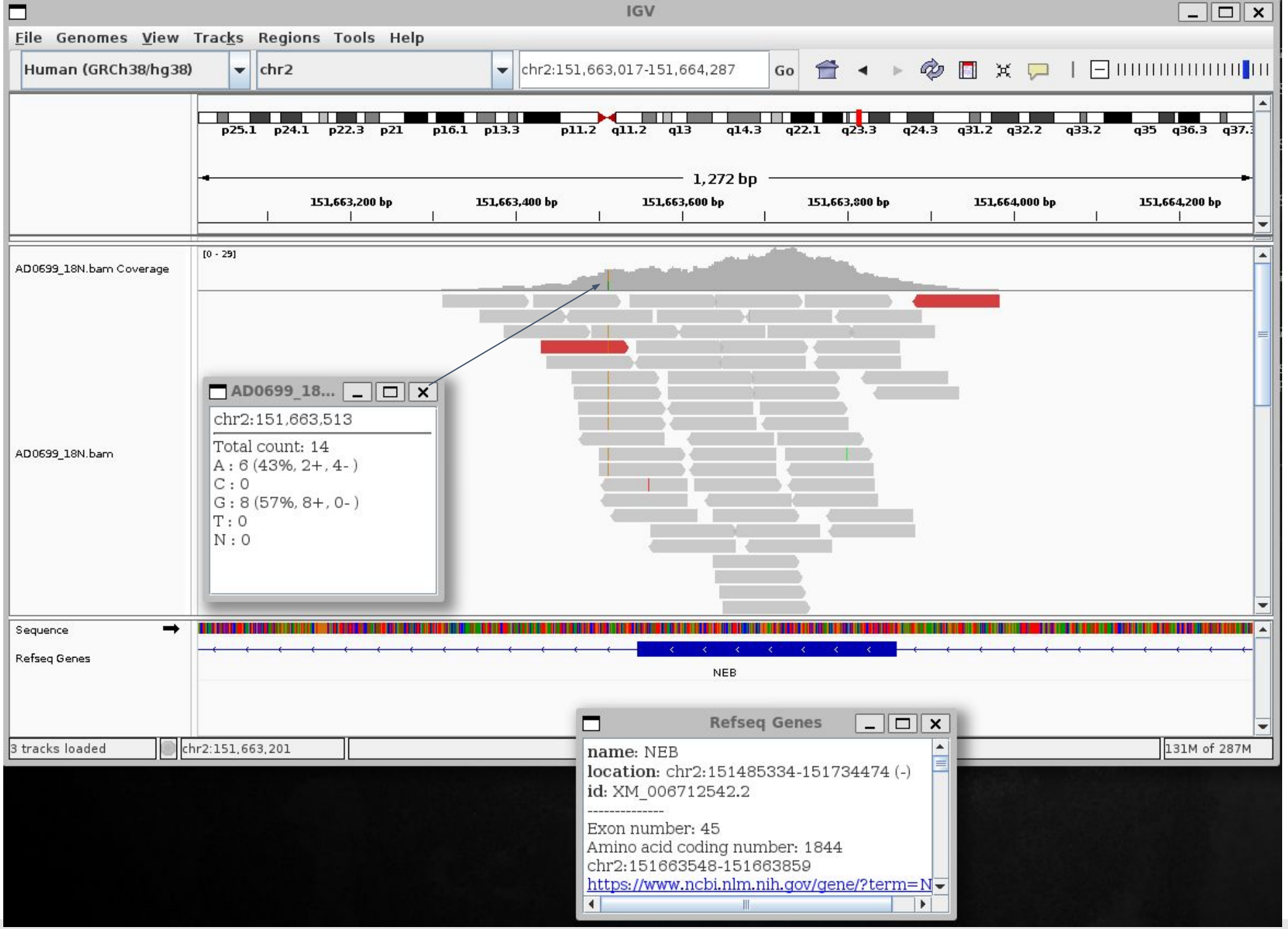
Recap


Collect sample
Extract DNA
Prep library
Sequence
Translate
Align
YOU ARE HERE

Integrative Genomics Viewer (IGV)

Visualization tool for exploring and analyzing genomic data

https://software.broadinstitute.org/software/igv/


IGV
File Genomes View Tracks Regions Tools Help
Human (GRCh38/hg38)
chr2
chr2:151,663,017-151,664,287
Go
p25.1 p24.1 p22.3 p21 p16.1 p13.3 p11.2 q11.2 q13 q14.3 q22.1 q23.3 q24.3 q31.2 q32.2 q33.2 q35 q36.3 q37.3
1,272 bp
151,663,200 bp
151,663,400 bp
151,663,600 bp
151,663,800 bp
151,664,000 bp
151,664,200 bp
AD0699_18N.bam Coverage
[0 - 29]
AD0699_18N.bam
AD0699_18...
chr2:151,663,513
Total count: 14
A : 6 (43%, 2+, 4-)
C : 0
G : 8 (57%, 8+, 0-)
T : 0
N : 0
Sequence
Refseq Genes
NEB
Refseq Genes
name: NEB
location: chr2:151485334-151734474 (-)
id: XM_006712542.2
Exon number: 45
Amino acid coding number: 1844
chr2:151663548-151663859
https://www.ncbi.nlm.nih.gov/gene/?term=N
3 tracks loaded
chr2:151,663,201
131M of 287M

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.

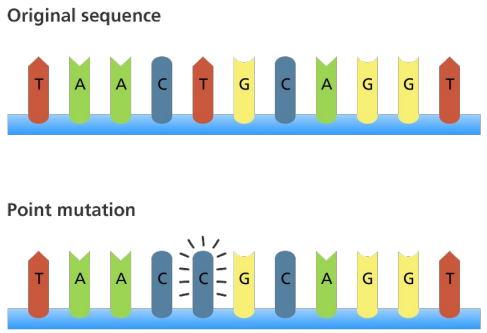
There are >100,000,000 know genetic variants in the human genome

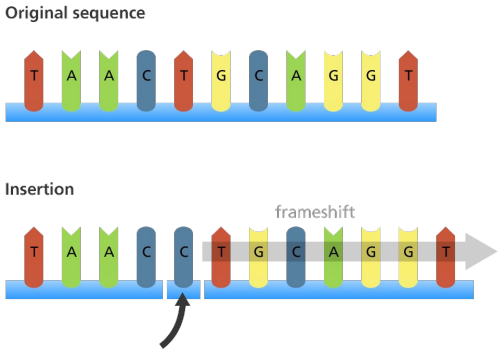
~99.8% identical DNA
(differ at 1/ 620 - 1/750 bp)

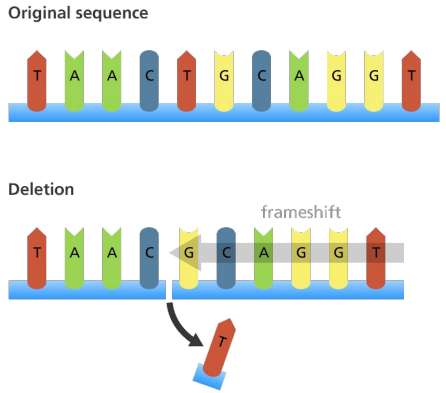
99% identical DNA

Types of Genetic Variation

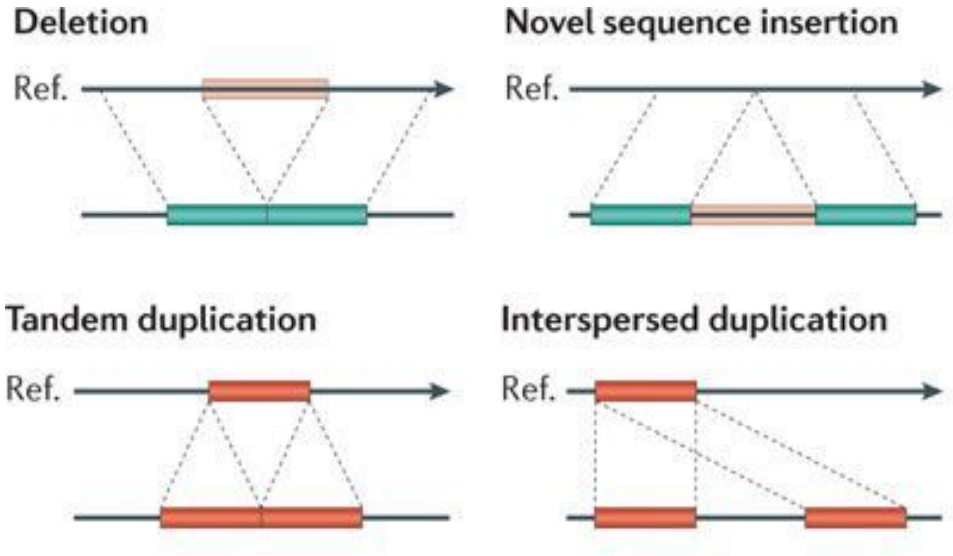
Small-scale

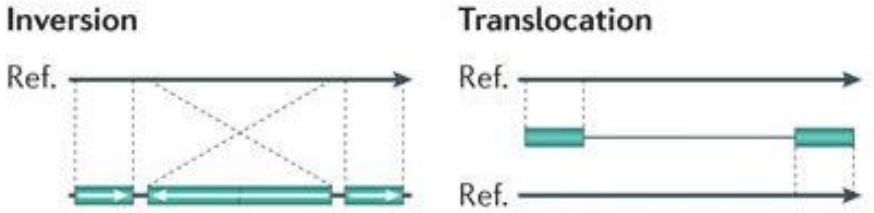






Large scale (Structural)





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
acctcTgagta
```

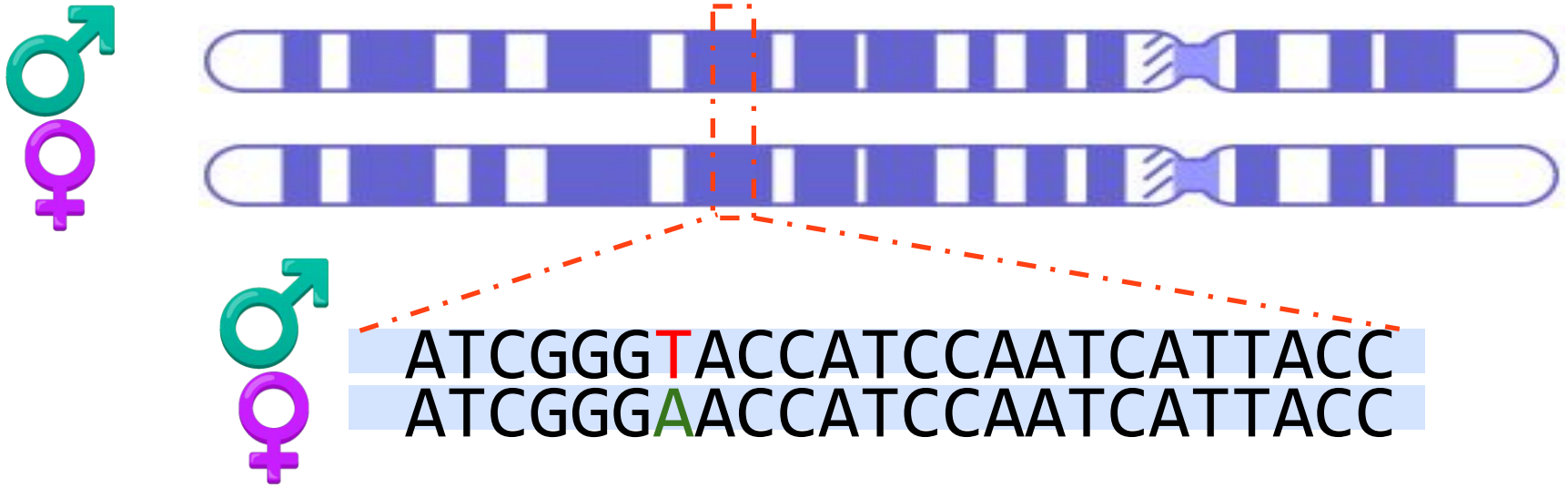
mutation:
private to this chromosome / individual

Mutations

From mutation to polymorphism

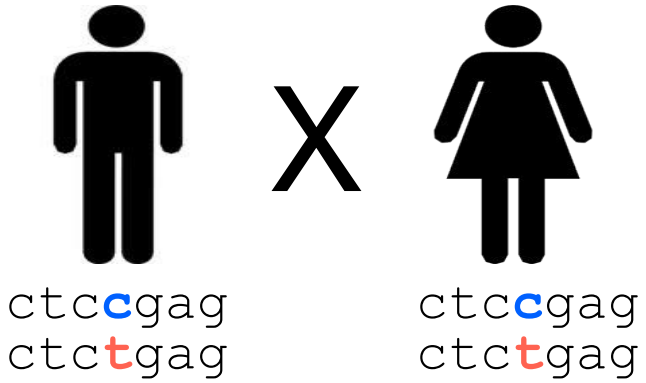
acctccgagta	acctcTgagta
acctccgagta	acctccgagta
acctccgagta	acctcTgagta
acctcTgagta	acctccgagta
acctccgagta	acctcTgagta

Diploid Genomes

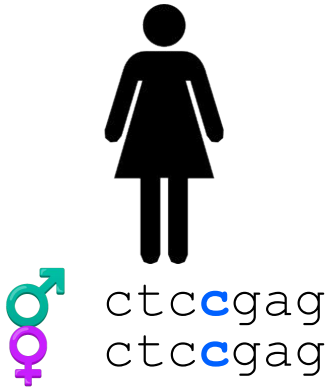


Our genome is comprised of a paternal and a maternal "haplotype".
Together, they form our "genotype"

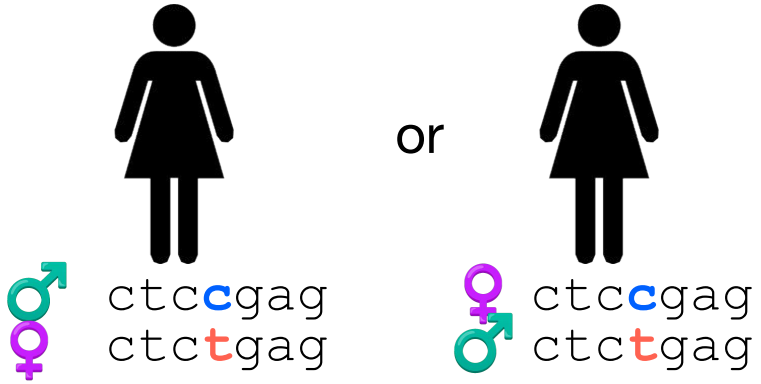
Inherited Germline Variation



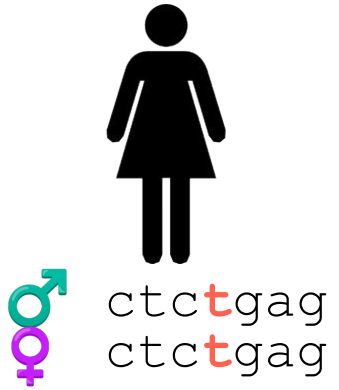
Example: Mom and dad are heterozygous; that is, the zygote from which they developed was comprised of a sperm and egg with two different alleles



Kid is homozygous (C/C)

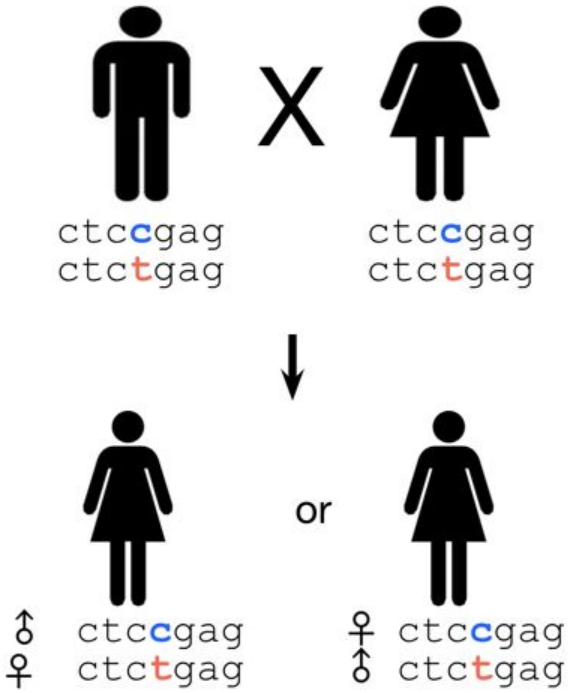


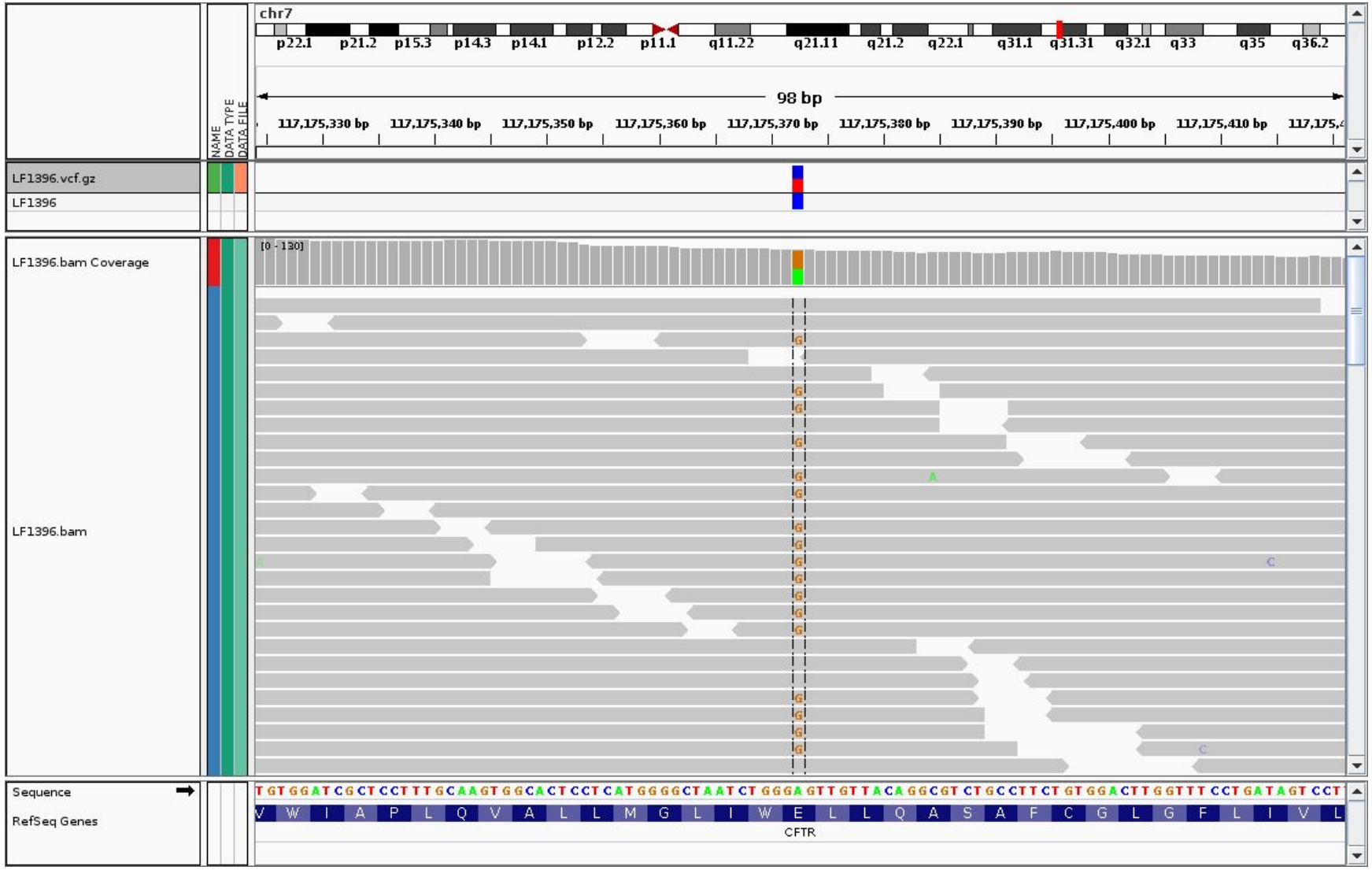
Kid is heterozygous (C/T)



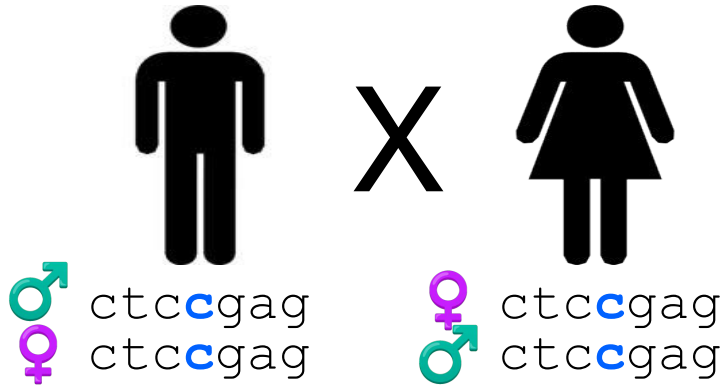
Kid is homozygous (T/T)

Heterozygous Variation





De novo Mutation

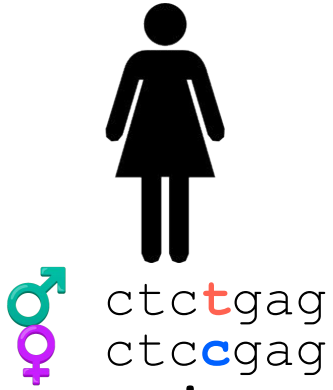


Example: Mom and dad are homozygous for the same alleles.

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

♂ ctccgag → ♂ ctctgag

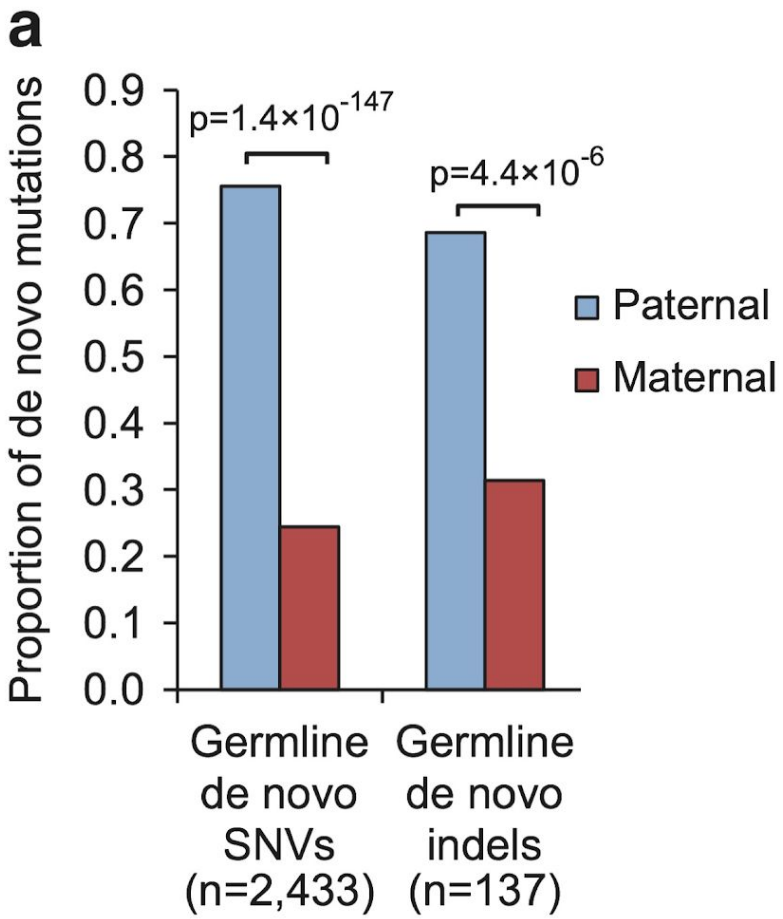
Note: This is a derivative chromosome of the one the father inherited from His parents



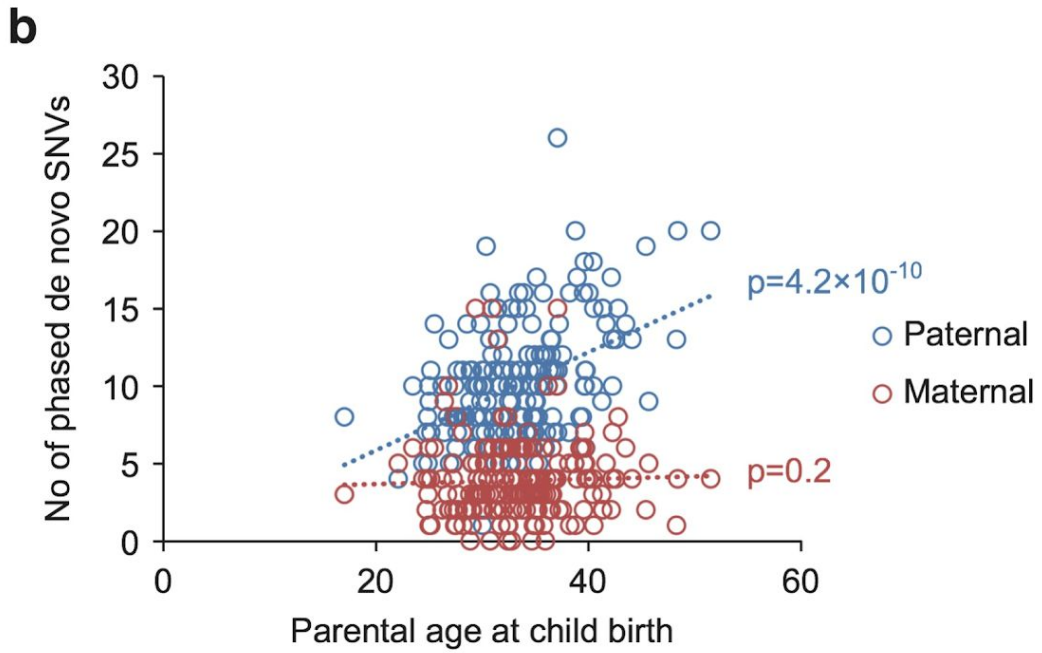
Kid is heterozygous owing to *de novo mutation.*

(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

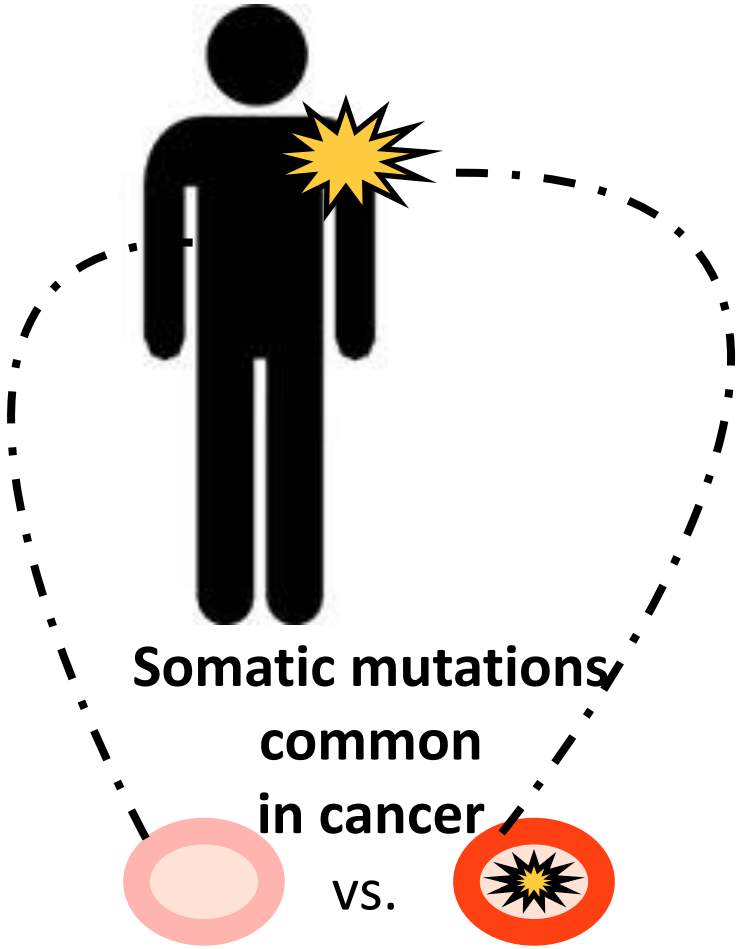
Somatic Mutations

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

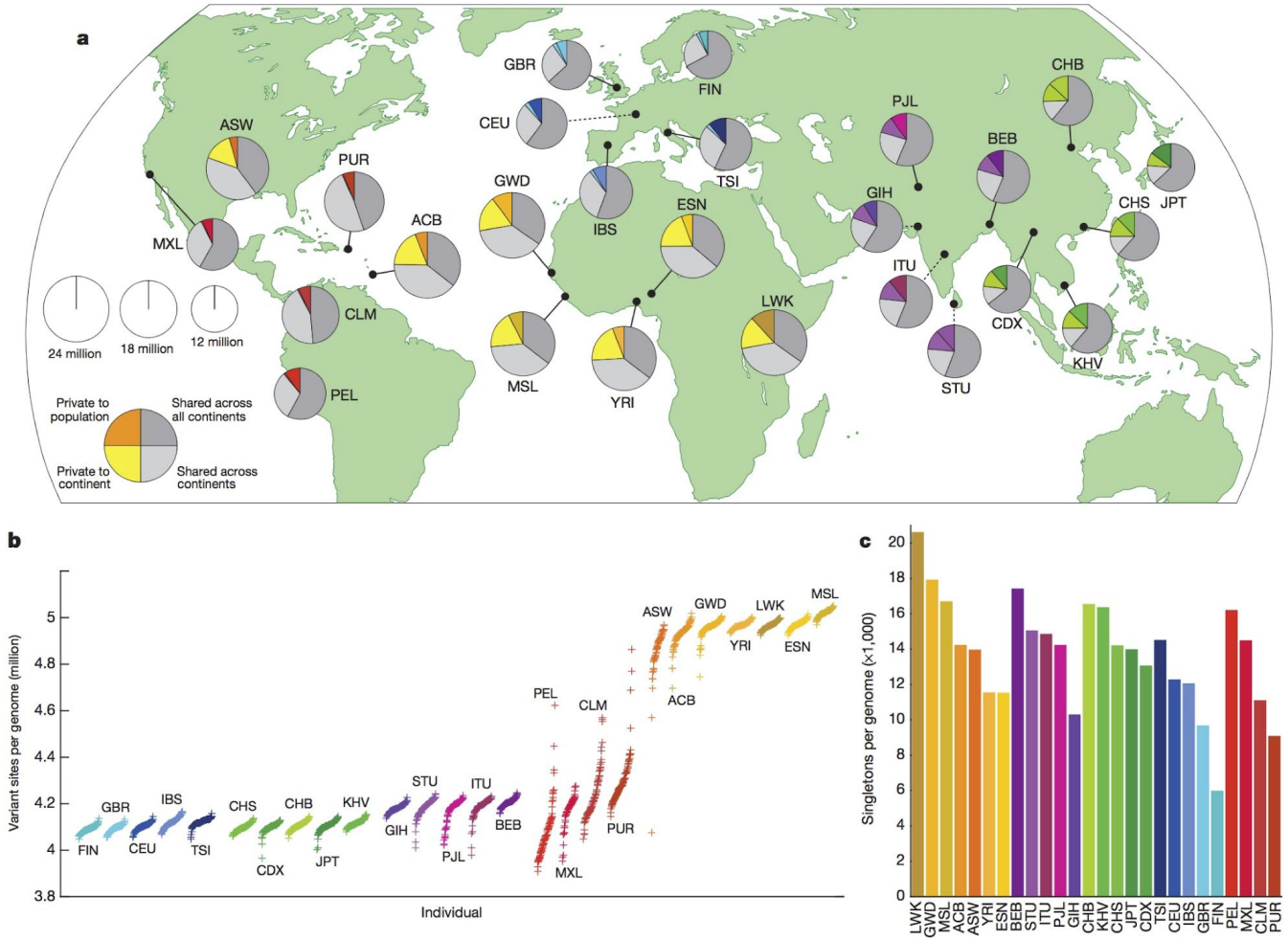


Figure 1 | Population sampling. **a**, Polymorphic variants within sampled populations. The area of each pie is proportional to the number of polymorphisms within a population. Pies are divided into four slices, representing variants private to a population (darker colour unique to population), private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

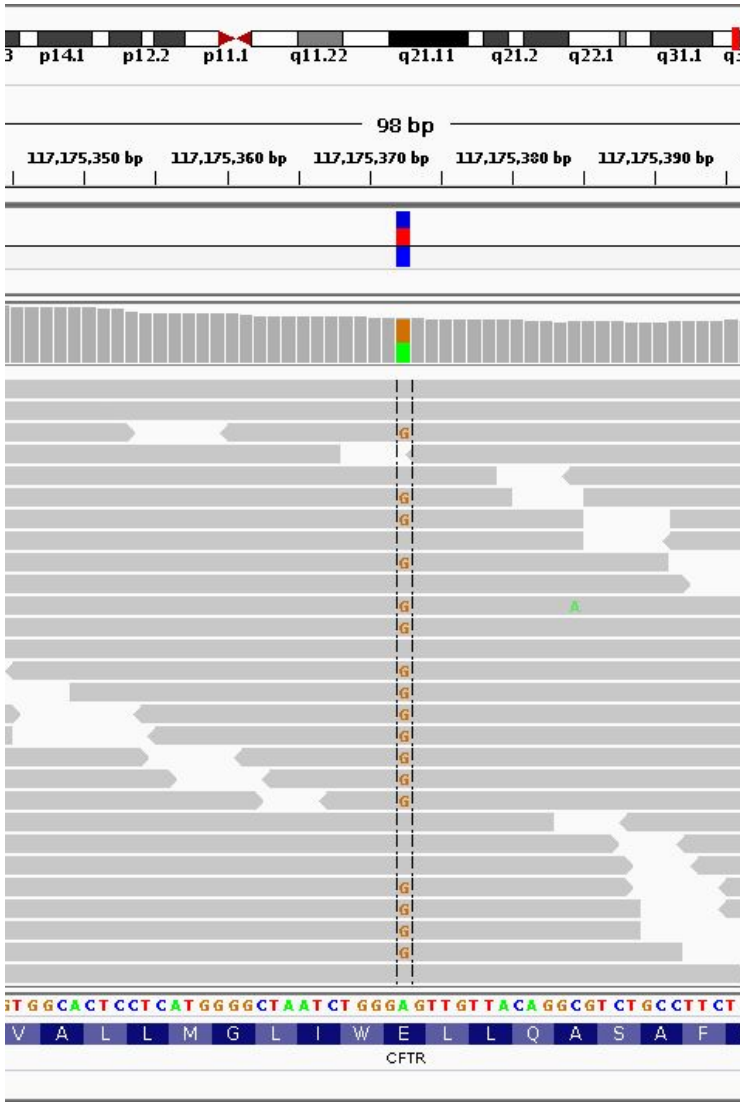
The extent of genetic variation by subpopulation

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

Bayes' Theorem

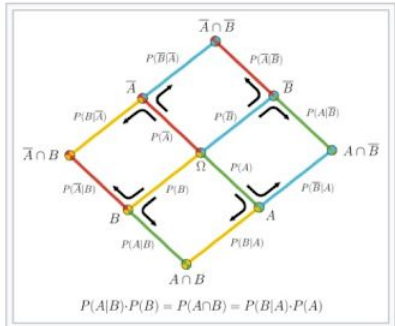
Statement of theorem [edit]

Bayes' theorem is stated mathematically as the following equation:[2]

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

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.
- $P(A \mid B)$, a conditional probability, is the probability of observing event A given that B is true.
- $P(B \mid A)$ is the probability of observing event B given that A is true.



Visualization of Bayes' theorem by the superposition of two decision trees

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[3] Bayes' major work "An Essay towards solving a Problem in the Doctrine of Chances" (1763), which appeared in "Philosophical Transactions,"[4] 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.[5][6]

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

Stephen Stigler suggested in 1983 that Bayes' theorem was discovered by Nicholas Saunderson, a blind English mathematician, some time before Bayes;[10][11] that interpretation, however, has been disputed.[12] Martyn Hooper[13] and Sharon McGrayne[14] 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.[14]

https://en.wikipedia.org/wiki/Thomas_Bayes

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…

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.

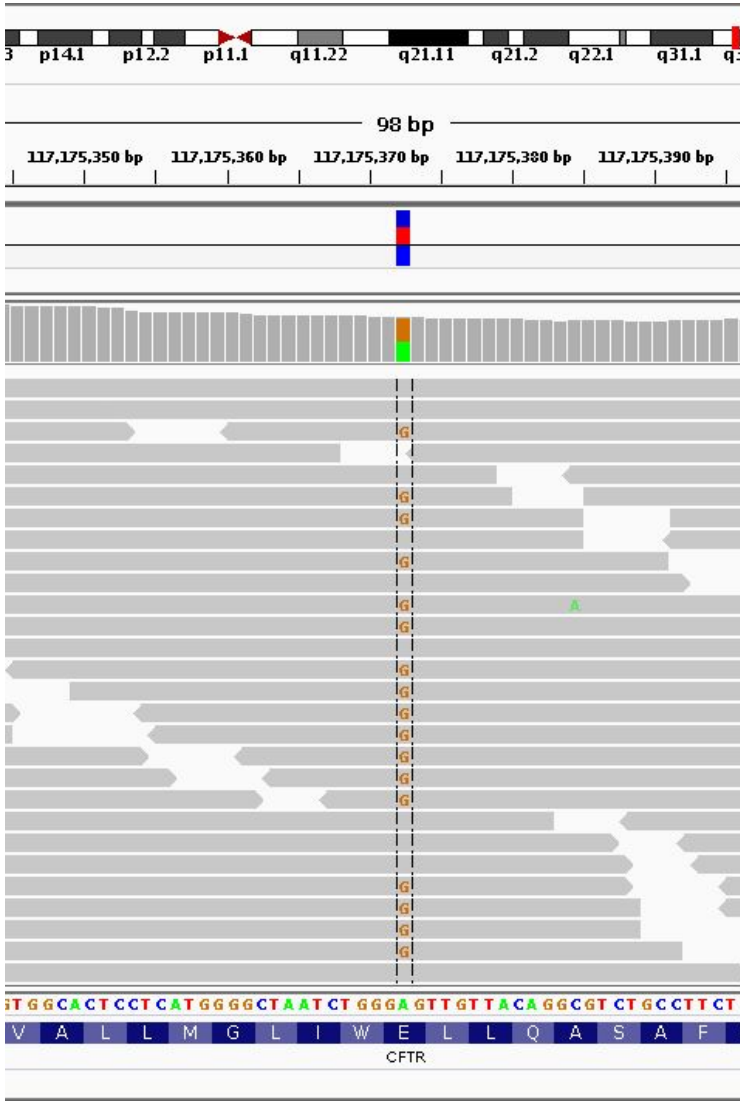
Bayes' Theorem

Posterior probability

Prior Probability Of A

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

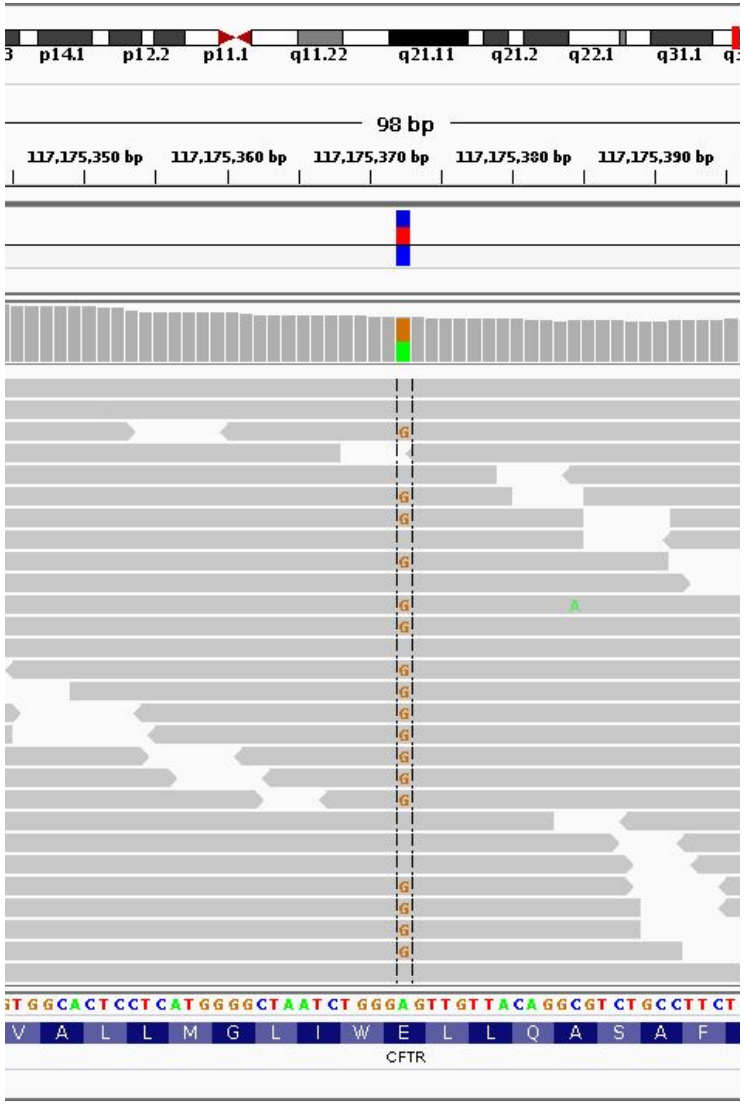
Bayesian SNP Calling



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

https://genomebiology.biomedcentral.com/articles/10.1186/gb-2010-11-10-r99

Bayesian SNP Calling



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

- 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

https://genomebiology.biomedcentral.com/articles/10.1186/gb-2010-11-10-r99

Bayesian SNP Calling

Bayesian posterior probability

Base call + Base quality

Expected (prior) polymorphism rate

$$P(SNP) = \sum_{all} \sum_{variable} S \frac{\frac{P(S_1 \mid R_1)}{P_{Prior}(S_1)} \cdot \ldots \cdot \frac{P(S_N \mid R_N)}{P_{Prior}(S_N)} \cdot P_{Prior}(S_1, \ldots, S_N)}{\sum_{S_{i_1} \in [A,C,G,T]} \cdots \sum_{S_{i_N} \in [A,C,G,T]} \frac{P(S_{i_1} \mid R_1)}{P_{Prior}(S_{i_1})} \cdot \ldots \cdot \frac{P(S_{i_N} \mid R_1)}{P_{Prior}(S_{i_N})} \cdot P_{Prior}(S_{i_1}, \ldots, S_{i_N})}$$

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

https://genomebiology.biomedcentral.com/articles/10.1186/gb-2010-11-10-r99

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

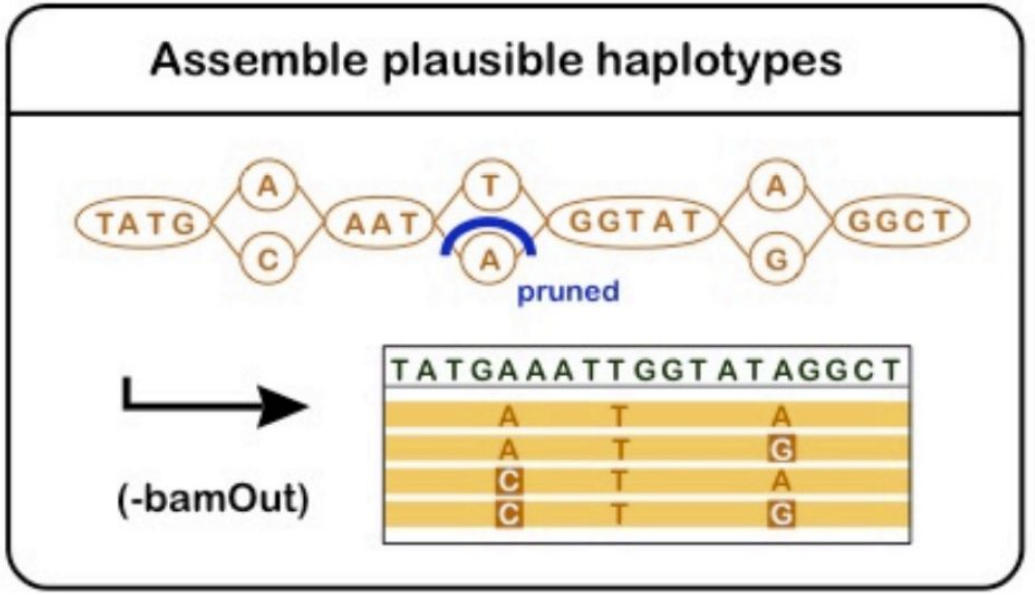
Nature Genetics **43**, 491–498 (2011) | doi:10.1038/ng.806

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

http://www.nature.com/ng/journal/v43/n5/full/ng.806.html

GATK's HaplotypeCaller

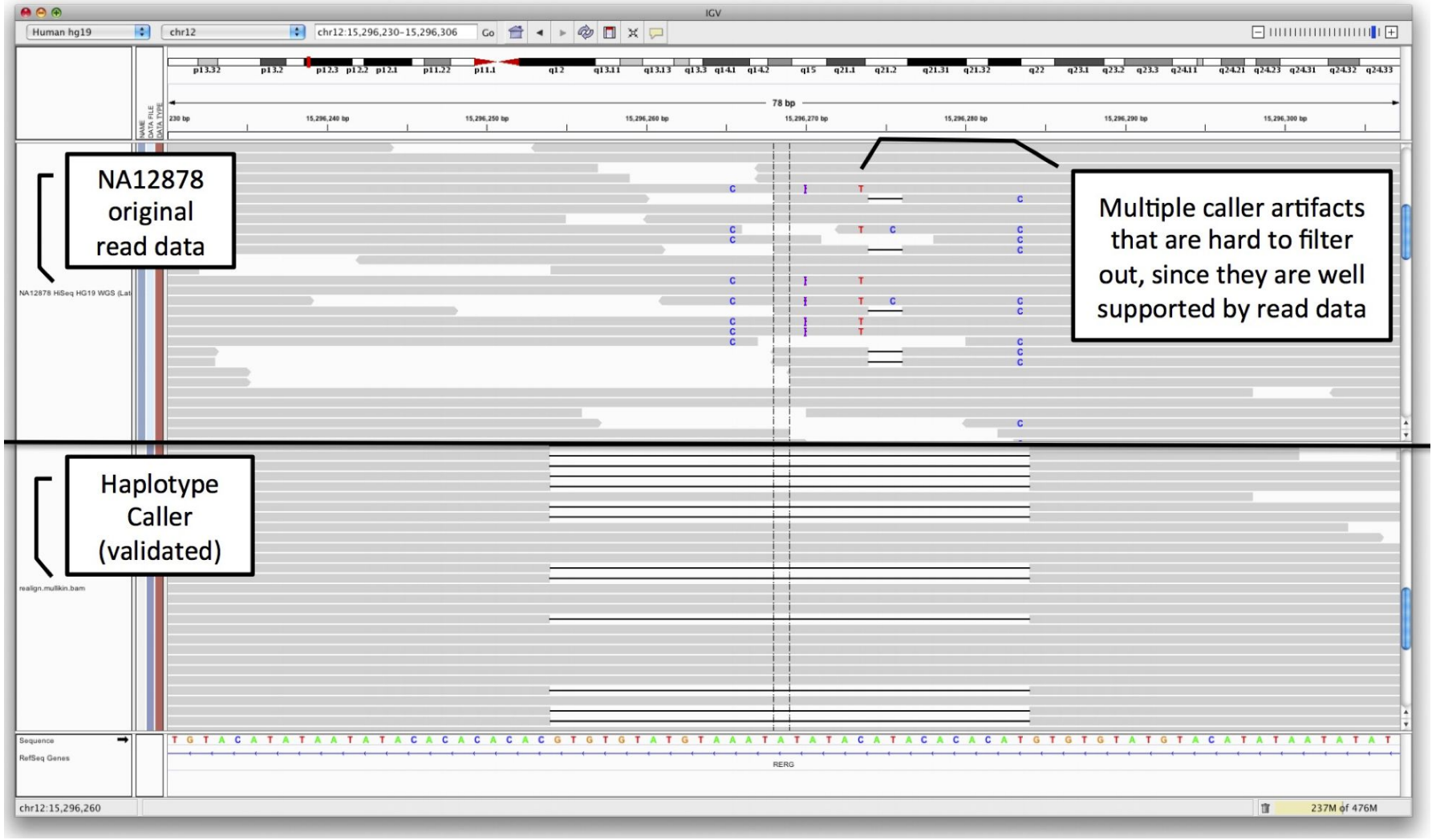
- Local re-assembly
- Traverse graph to collect most likely haplotypes
- Align haplotypes to ref using Smith-Waterman



Likely haplotypes + candidate variant sites

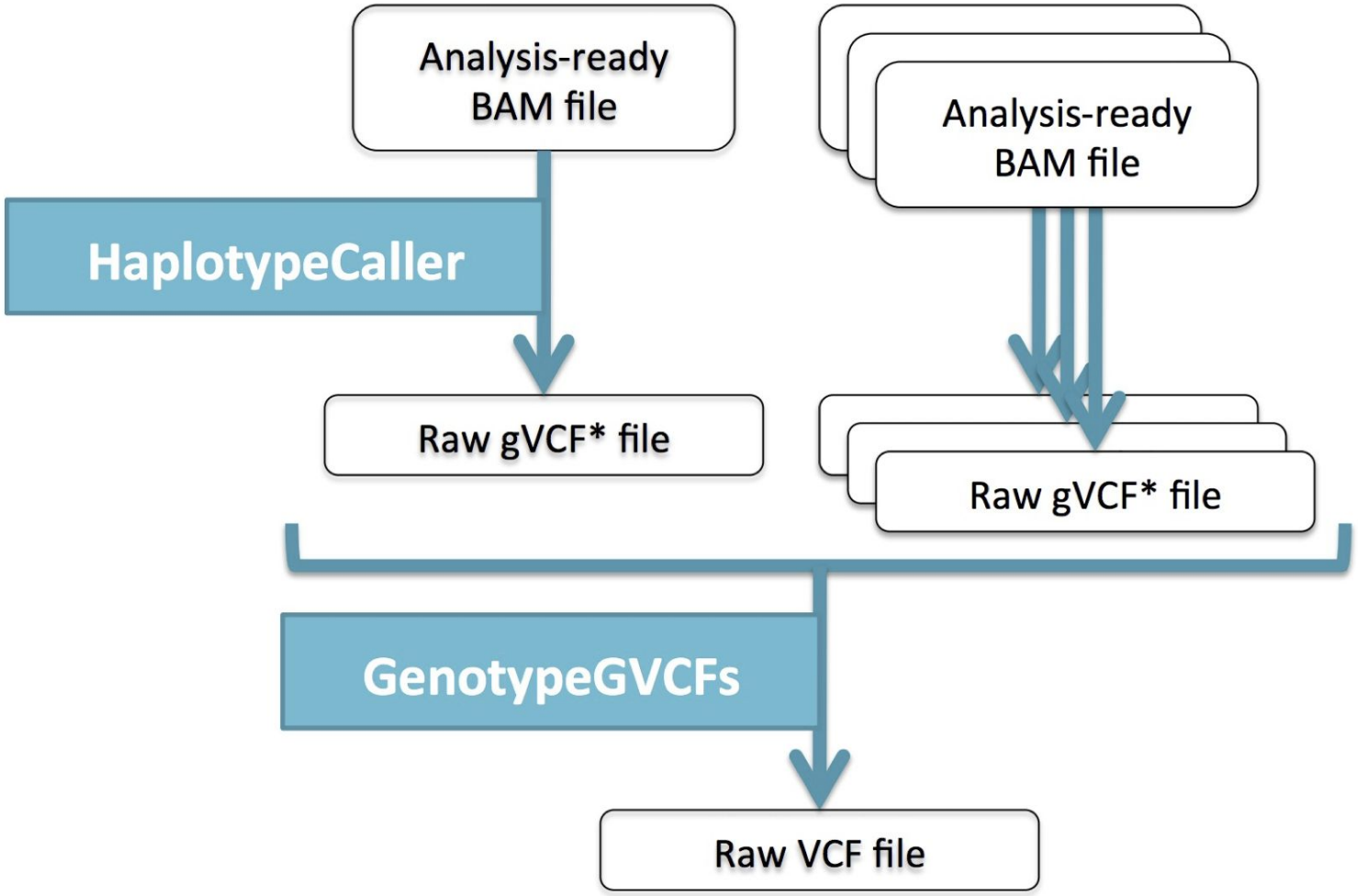
https://software.broadinstitute.org/gatk/events/slides/1506/GATKwr8-D-2-Variant_calling_and_joint_genotyping.pdf

GATK's HaplotypeCaller



https://software.broadinstitute.org/gatk/events/slides/1506/GATKwr8-D-2-Variant_calling_and_joint_genotyping.pdf

GATK's HaplotypeCaller



https://software.broadinstitute.org/gatk/events/slides/1506/GATKwr8-D-2-Variant_calling_and_joint_genotyping.pdf

Variant Calling

VCF 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 Genotyping

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 Genotyping

```
##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 Genotyping

```
##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

Long Read Technologies

Be familiar with the main 3rd generation sequencing technologies:

- PacBio SMRT sequencing
- ONT sequencing
- 10X linked reads

Understand various applications of long and linked reads

- Structural variant calling
- De novo assembly (*next lesson*)
- RNA-seq (*next lesson*)

What is 3rd Gen Sequencing

Sequencing technologies other than Illumina sequencing

Focus on producing **long-distance** information

- **Long reads**
- **Linked reads**

Developed or matured in the last decade

Actively being developed

Main technologies:

- Pacific Biosciences SMRT sequencing - **PacBio**
- Oxford Nanopore Technology - **ONT**
- 10X Genomics Chromium - **10X**

PacBio SMRT Sequencing

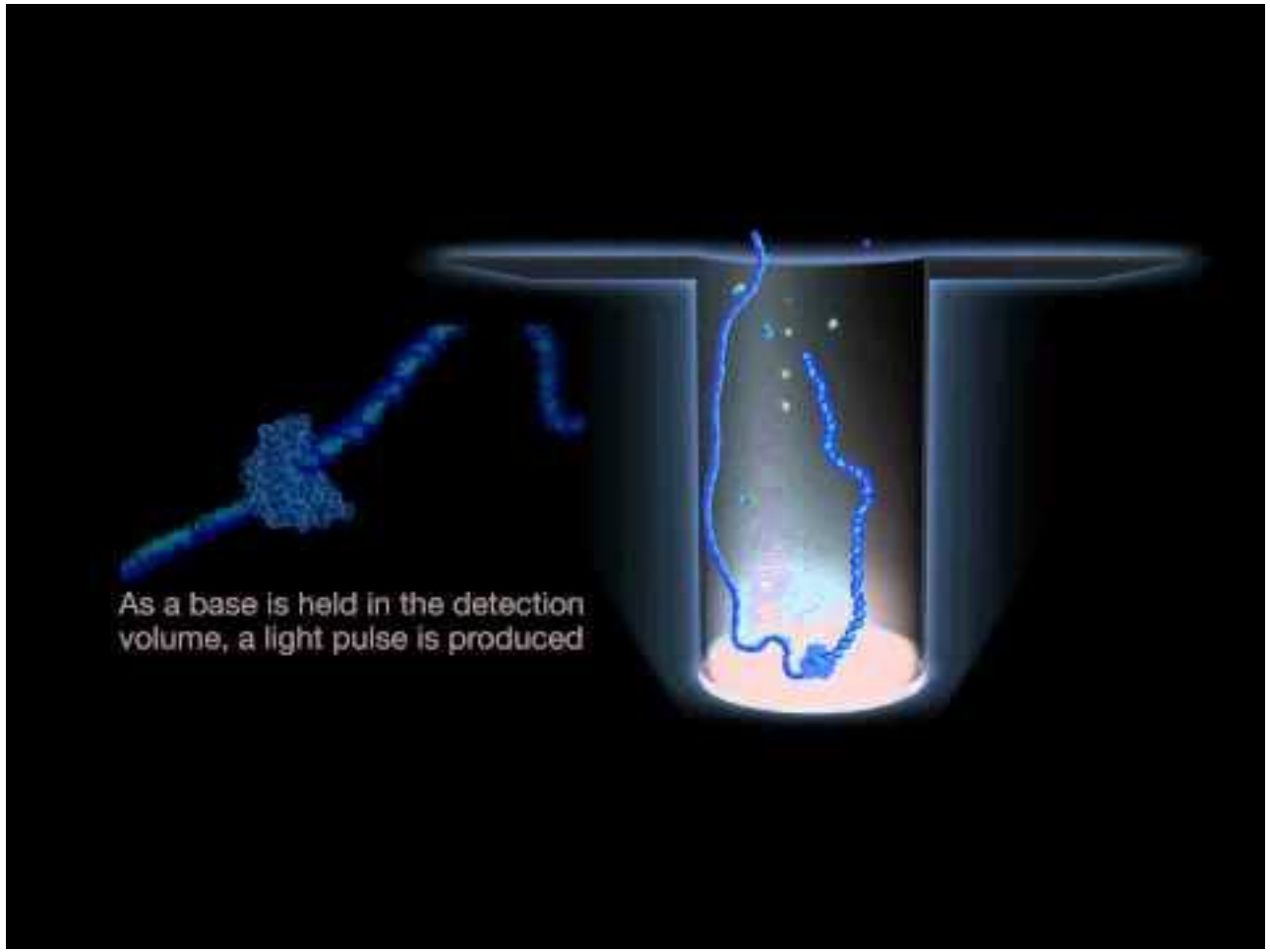
Single Molecule Real Time

No amplification step

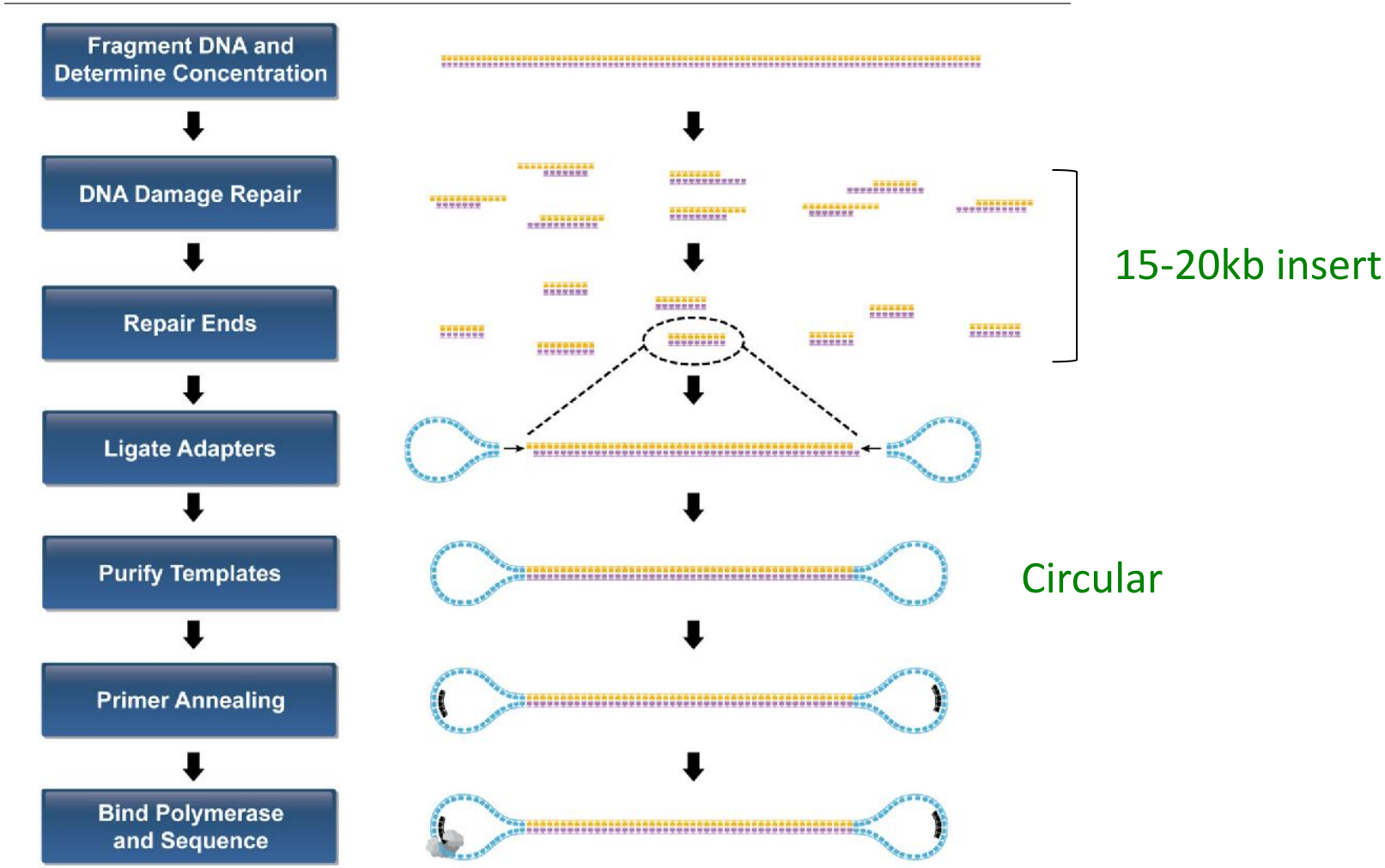
Based on the ability to analyze very small volumes

Sequencing by synthesis

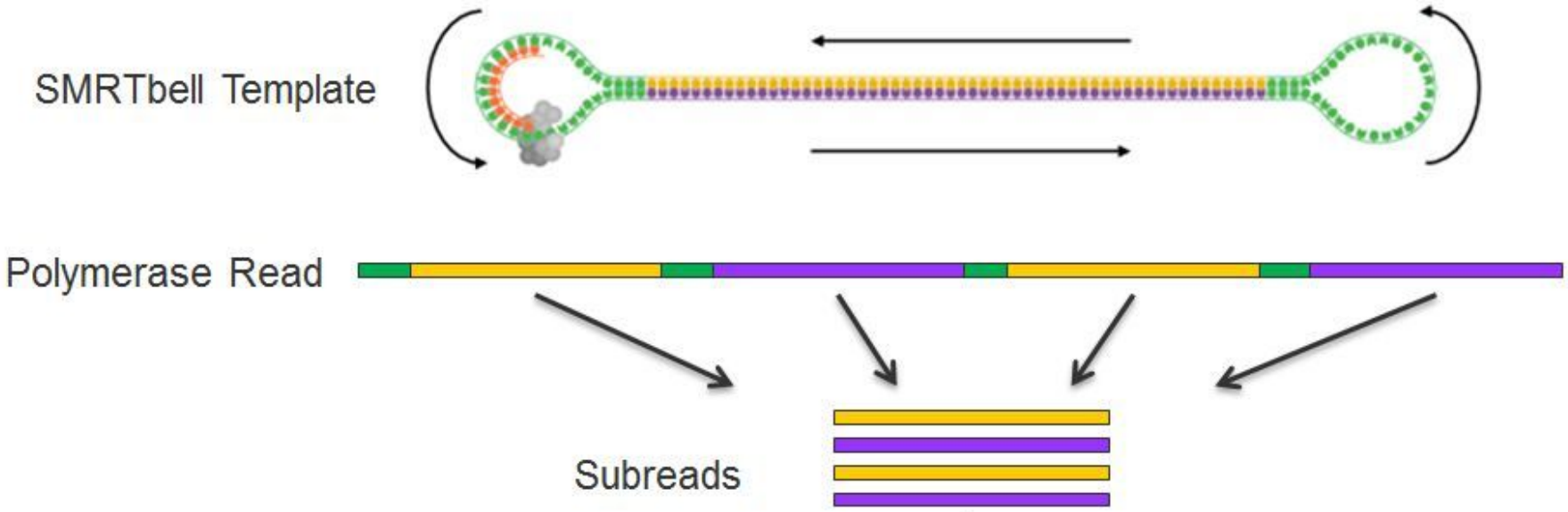
Sequel II



PacBio Library Prep



PacBio Sequencing



Properties of PacBio Sequencing

Read length

- Non-uniform
- Depends on selected insert size
- Usually 10-100kb

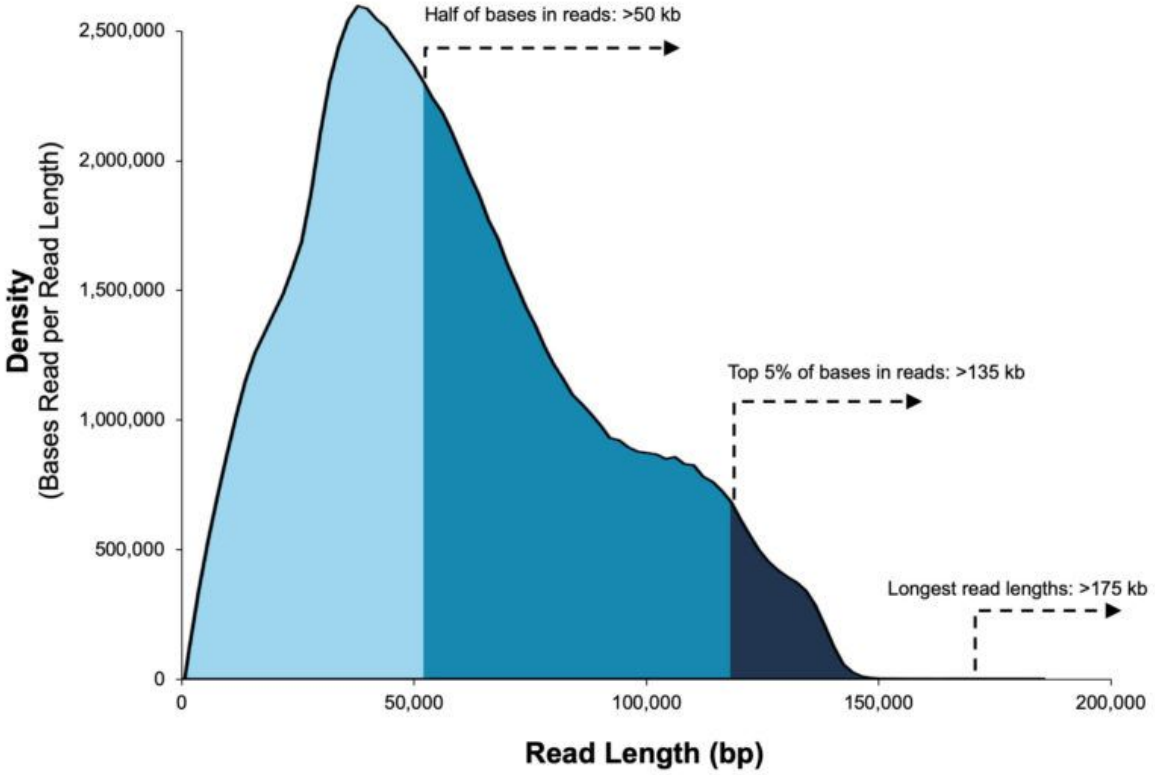
No paired-end option

One run can produce 4-5M reads - ~40Gb

Runs take several hours

Mostly uniform coverage - no GC-content bias

Raw reads error rate - **~10%**



Dealing With High Error Rates

Working with 10% error rate is impractical

Option 1:

Polymerase Read 

CLR - continuous long read

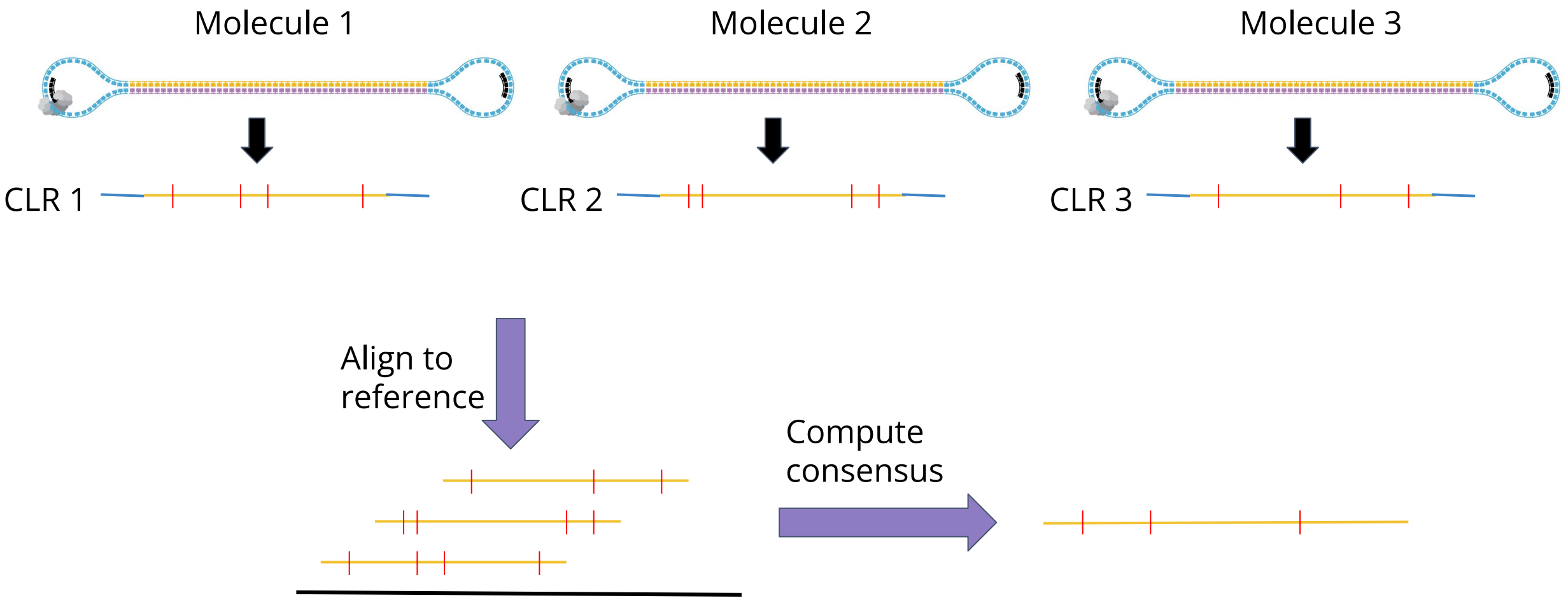
Polymerase read length ~= sub-read length

Align CLRs to a reference genome and correct errors

Find the consensus of multiple molecules

Accuracy increases with sequencing depth

CLR Error Correction



Dealing With High Error Rates

Option 2:

CCS - circular consensus read

Also called **HiFi reads**

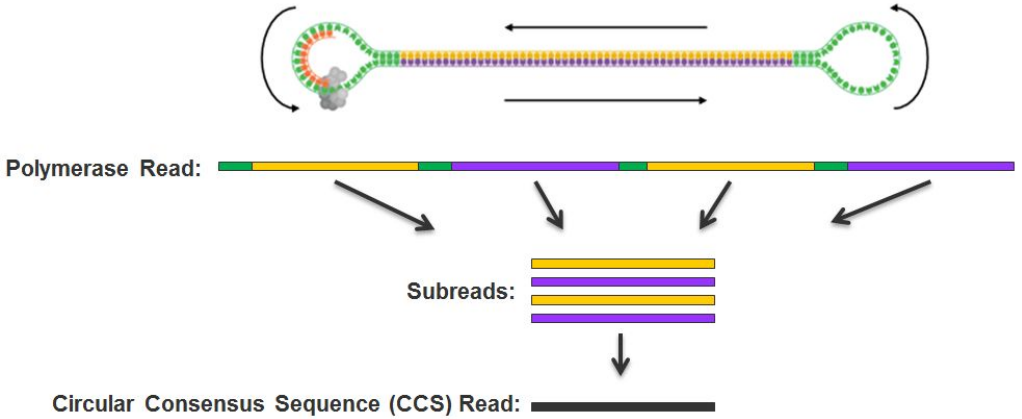
Polymerase read length > sub-read length

Align CCSs to one another and correct errors

Find the consensus of a single molecule

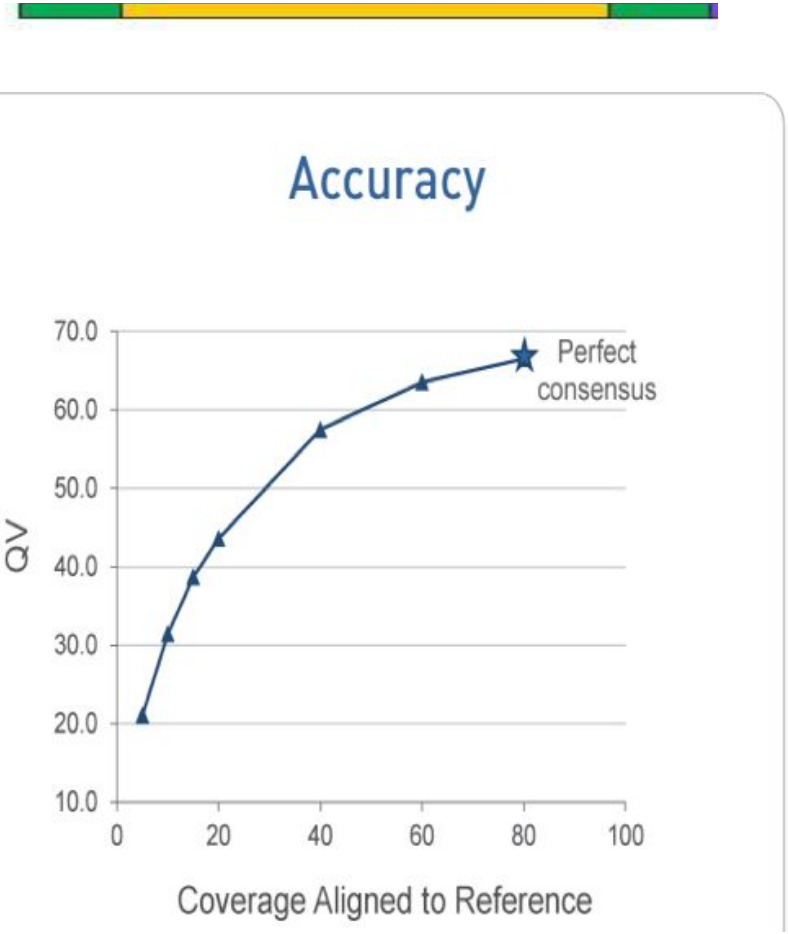
Accuracy >99%

Shorter reads (<20kb)

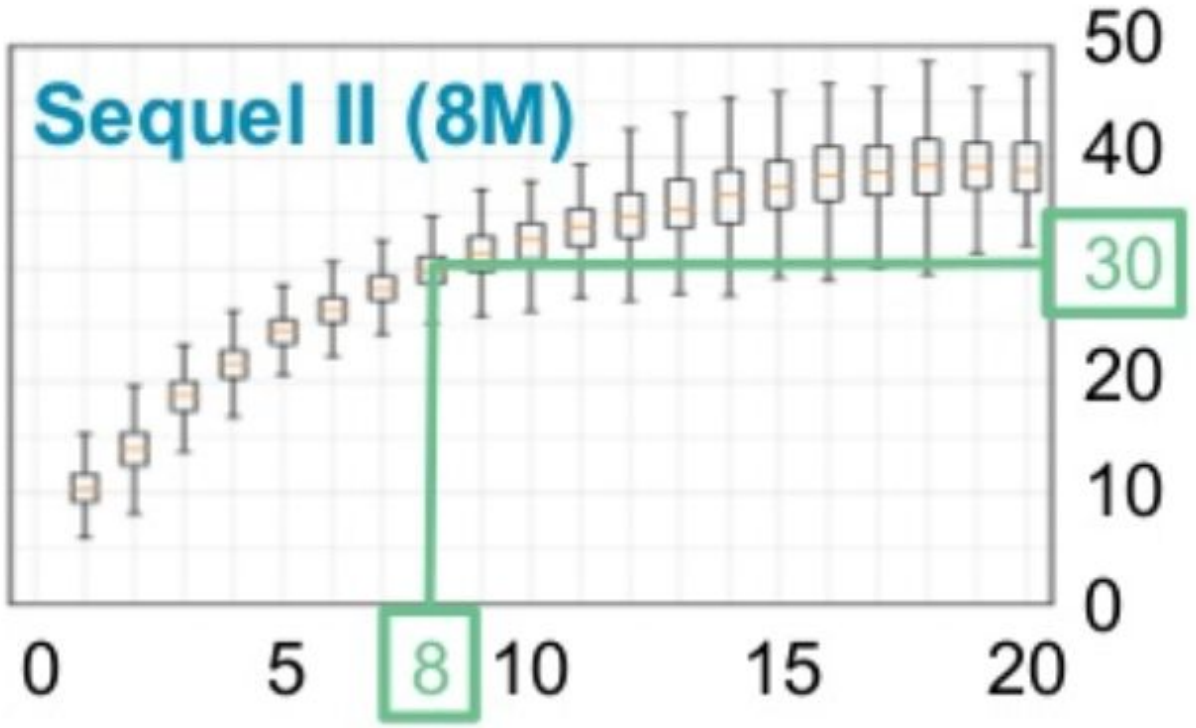


Accuracy CLR consensus Vs. CCS

CLR consensus



CCS



Oxford Nanopore Sequencing (ONT)

Single molecule

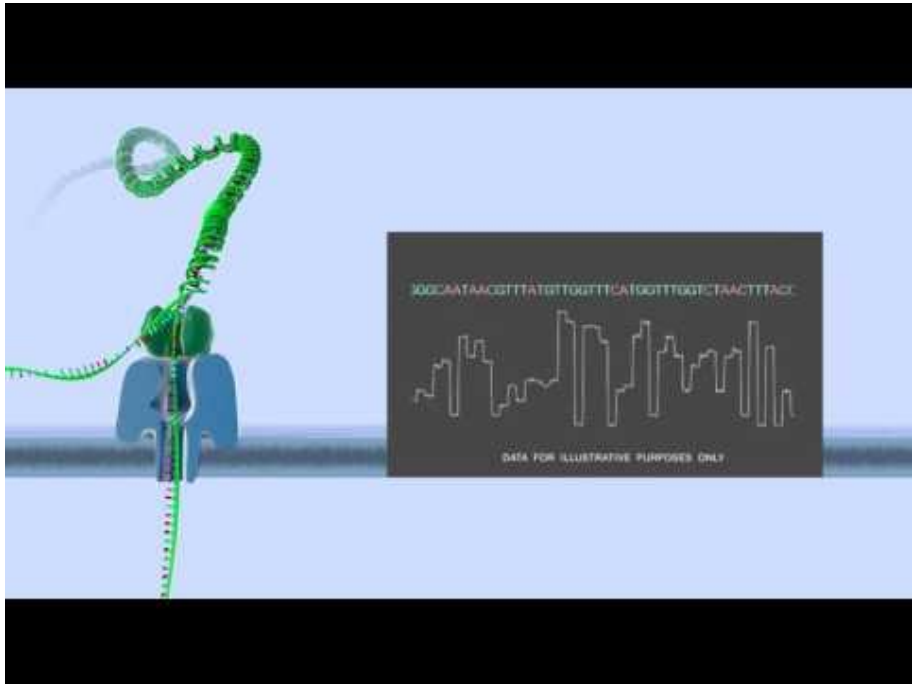
Real time

Not SBS

Palm-sized machine

MinION Mkl: portable, real time biological analyses

MinION



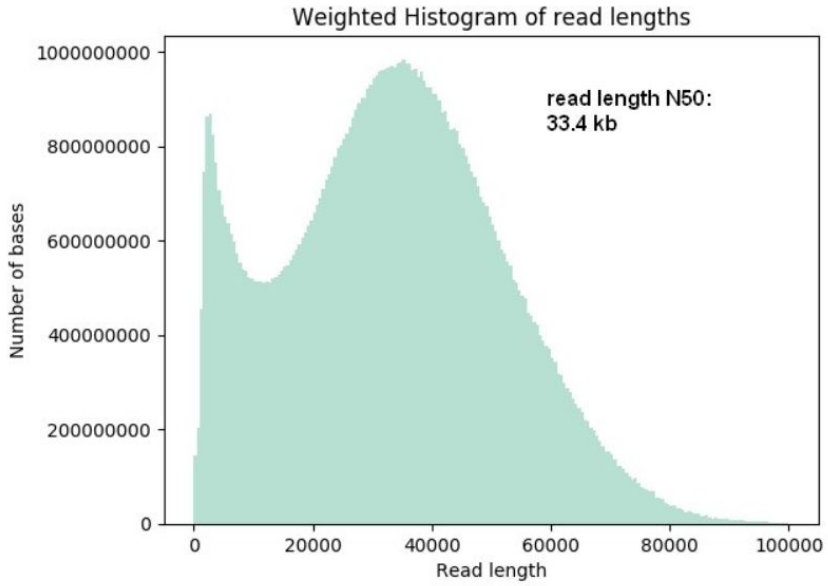
Properties of ONT Sequencing

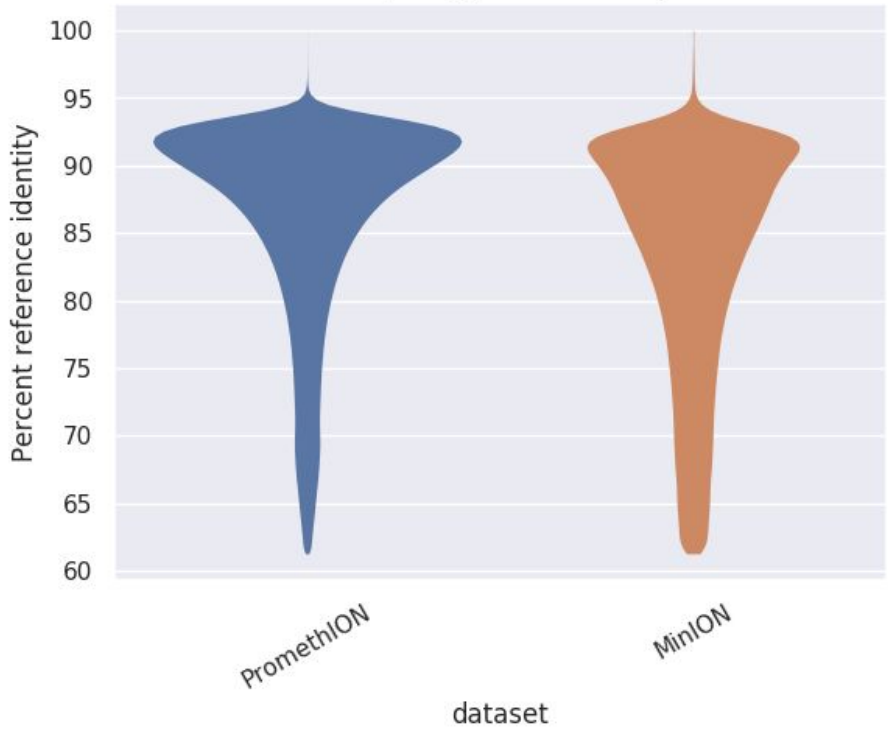
Read length - theoretically unlimited

In practice depends on DNA fragmentation - can produce reads > 2Mb

Yield - depends on machine model - 50Gb to 10Tb

Accuracy - ~10% error





Comparing Technologies

	Illumina	PacBio CLR	PacBio CCS	ONT
Read length	150-250 bp	50 kb	30 kb	10-30 kb
Overall error rate	0.1 %	10-15 %	<1 %	<5 %
Mismatch	~ 100 %	37 %	4 %	41 %
InDel	~ 0 %	63 %	96 %	59 %
Cost	$29/Gb	$85/Gb		$30/Gb*
Throughput	7 Gb/h	2.5 Gb/h		0.5 Gb/h*

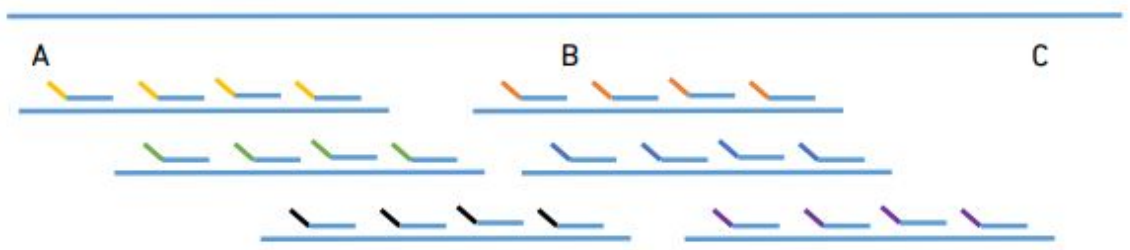
Amarasinghe, Shanika L., et al. "Opportunities and challenges in long-read sequencing data analysis." *Genome biology* 21.1 (2020): 1-16.

10X Genomics

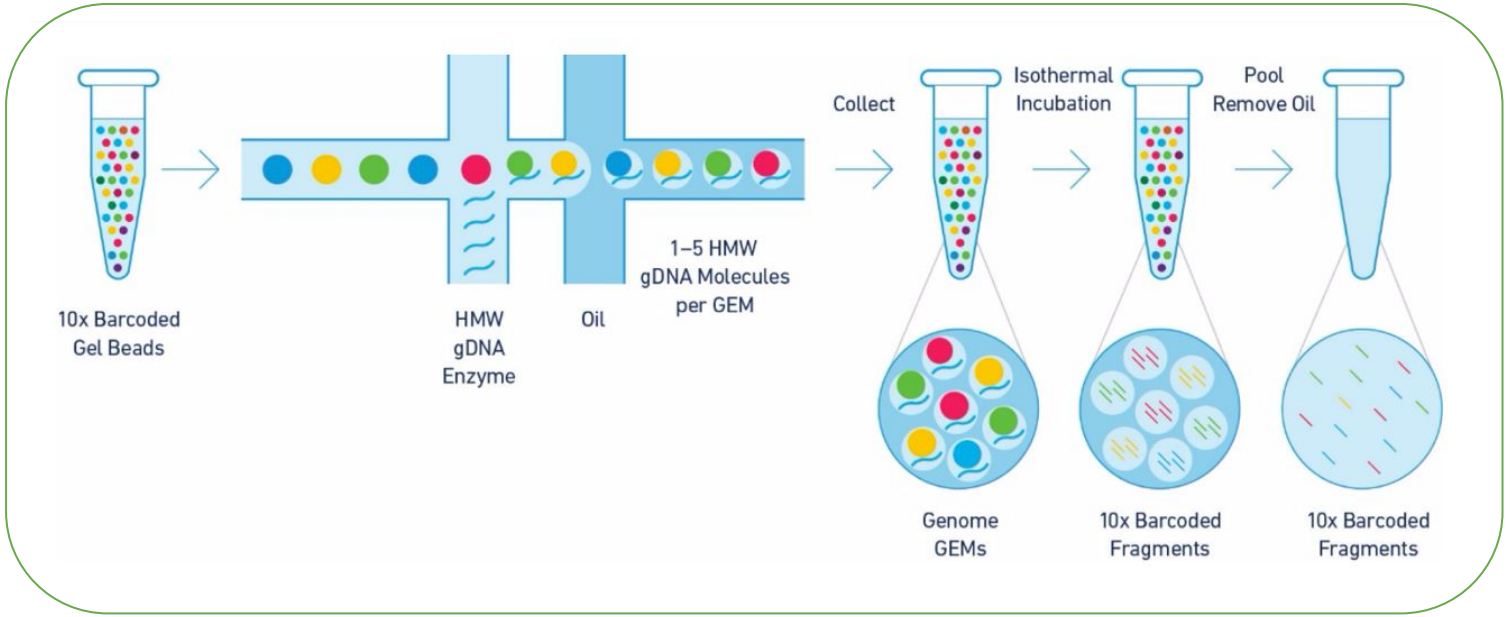
Not a long read technology

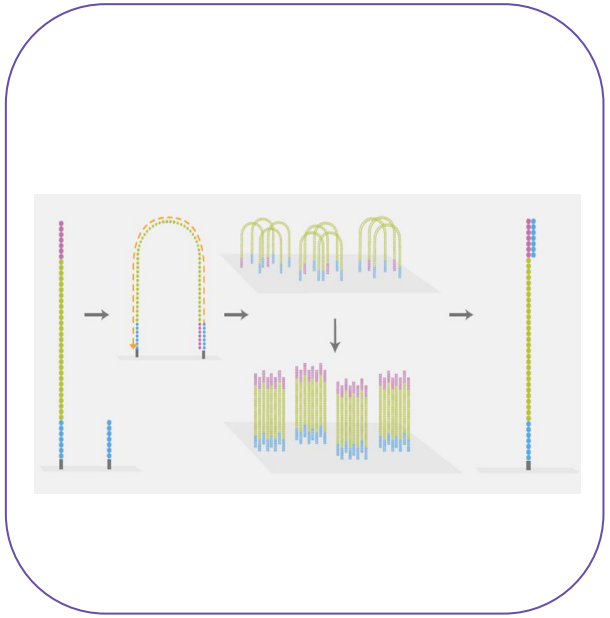
But provides long-range information through **linked reads**

Short reads originating from the same long molecule



Based on standard short read Illumina technology


10x Barcoded
Gel Beads
HMW
gDNA
Enzyme
Oil
1–5 HMW
gDNA Molecules
per GEM
Collect
Isothermal
Incubation
Pool
Remove Oil
Genome
GEMs
10x Barcoded
Fragments
10x Barcoded
Fragments




R1
Illumina
adaptor
10X
barcode
gDNA

Linked Reads

Reads with the same barcode likely come from the same gDNA fragment

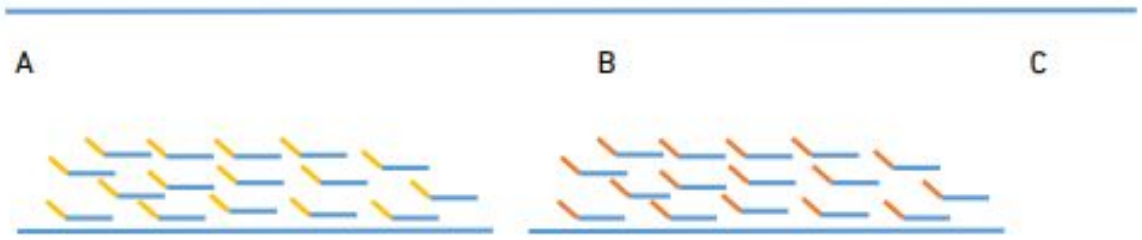
gDNA fragment size is usually 50-60kb

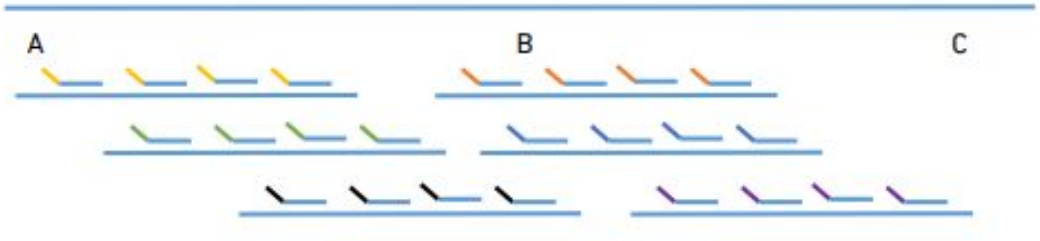
If ~x3 depth is used - we can produce “synthetic long reads”

Usually each molecule is sequenced at ~x0.2

We can still get useful long-range information

Non-trivial computational analysis is needed





Applications of 3rd Gen Sequencing

Structural variation detection

Genome assembly (*next lesson*)

Transcriptomics (*next lesson*)

Structural Variant Detection

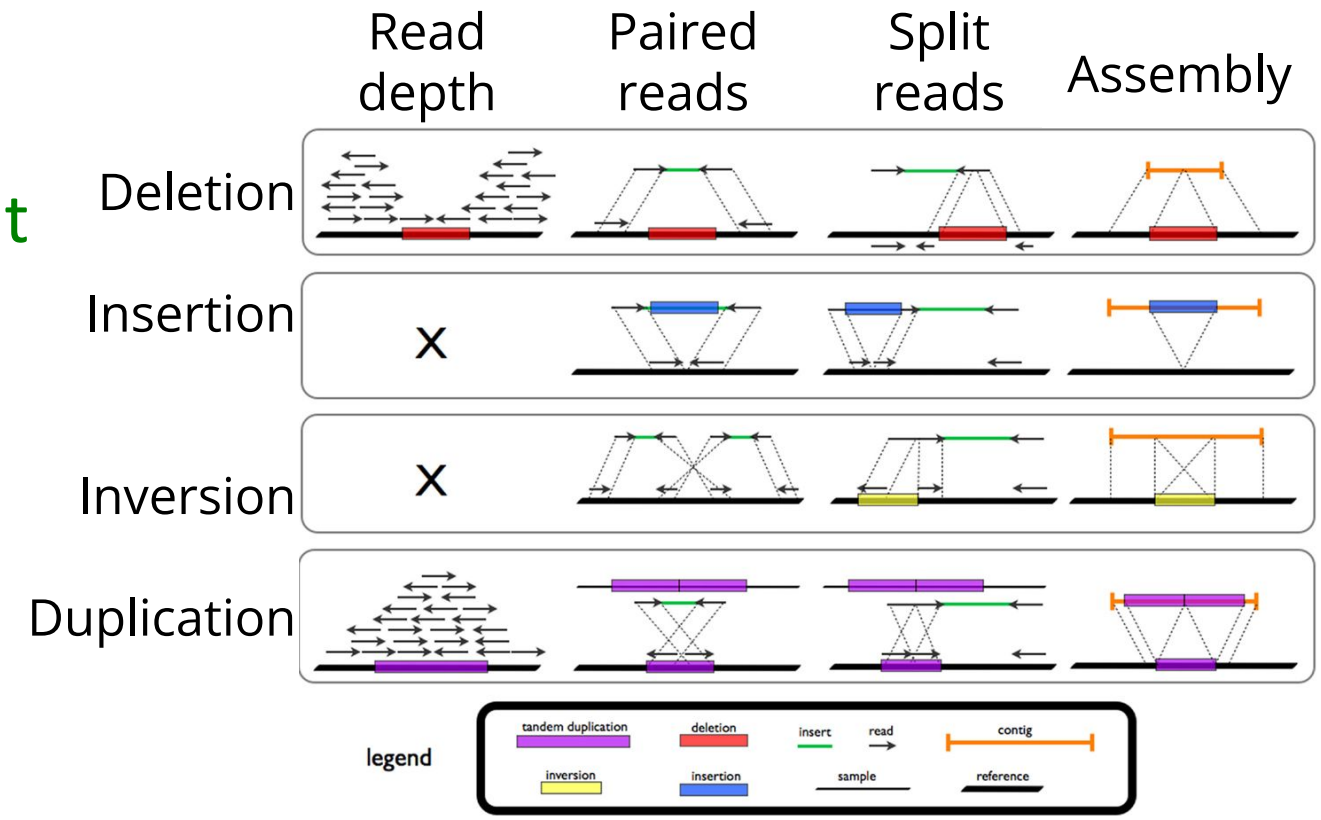
SVs are generally hard to detect with short reads

Many SVs are located in regions that are hard to sequence

SV detection is usually based on mapping reads to a reference

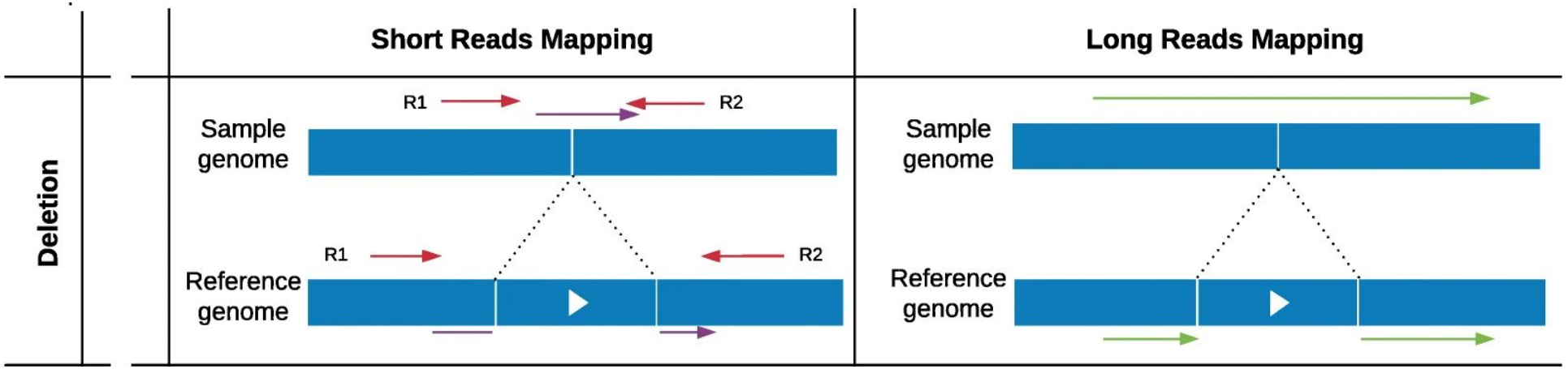
Long reads are useful because:

- They can cross long repeats
- They are not affected by GC-bias
- They can span large insertions



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.

How Do we Detect Variants



	Sequencing	Mapping	Variant calling
SNP	short reads	BWA	GATK
SV	short reads	BWA	Manta
	long reads	Minimap2	Sniffles

Read Mapping With Minimap2

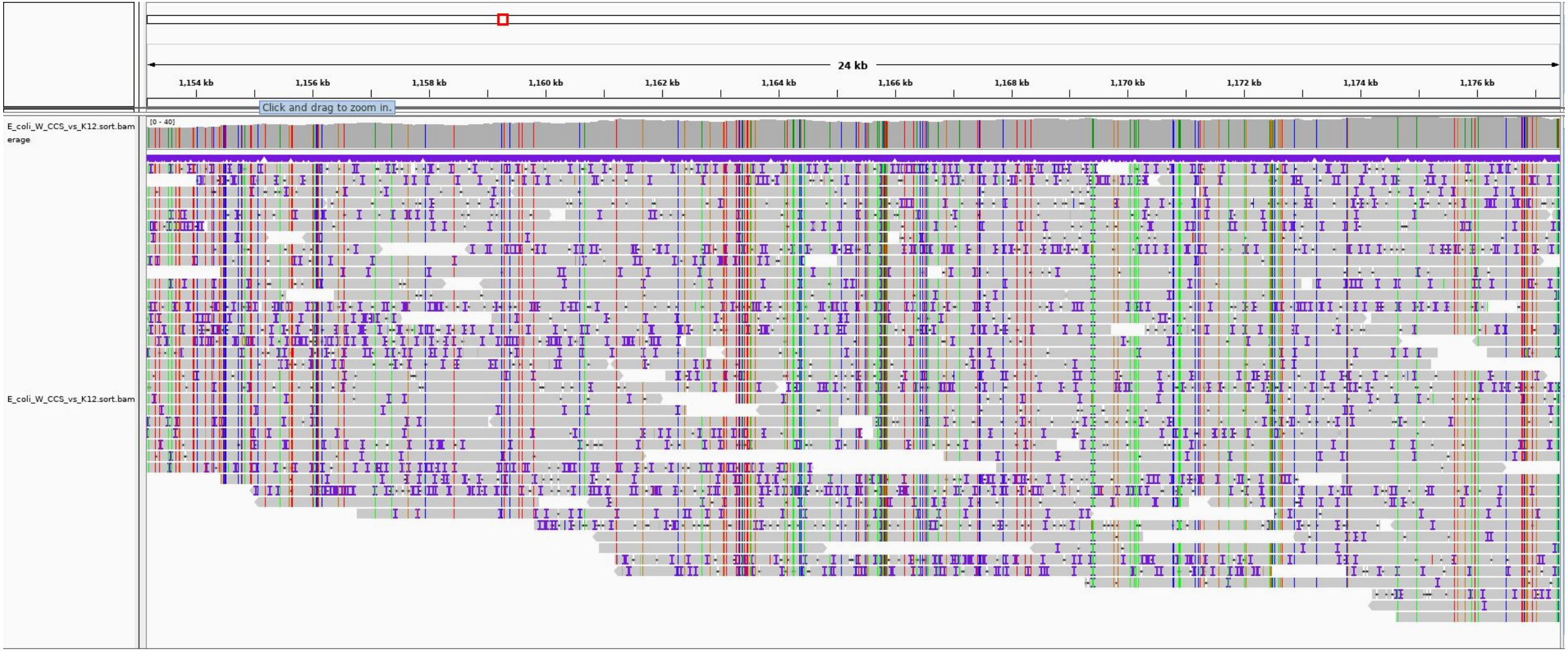
Minimap2 is a generic sequence mapping software

There are various mapping modes like:

- PacBio CLR to genome
- PacBio CCS to genome
- cDNA / PacBio Iso-Seq (transcripts) to genome
- ONT reads to genome
- PacBio reads to PacBio reads
- Short reads to genome (alternative to BWA)

Modes accounts for the specific biases of each technology

Input format is fasta/fastq


24 kb
1,154 kb
1,156 kb
1,158 kb
1,160 kb
1,162 kb
1,164 kb
1,166 kb
1,168 kb
1,170 kb
1,172 kb
1,174 kb
1,176 kb
Click and drag to zoom in.
E_coli_W_CCS_vs_K12.sort.bam
erage
[0 - 40]
E_coli_W_CCS_vs_K12.sort.bam

HYPE!

In recent years there have been **lots** of talk about long (an linked) reads

Many publications about data analysis and dedicated tools

Long reads are great! … **for some things**

Don’t trust everything you read

Always read the “small letters” (usually supplementary materials)

Vast majority of sequencing is still done with short reads

One technology can’t solve all problems in biology!