Introduction to variant calling

Bastiaan Star, Associate Professor
Centre for Ecological and Evolutionary Synthesis (CEES)
Archaeogenomics group
Department of Biosciences, University of Oslo (UiO)
Norway

@archaeogenomics

BIOS-IN 5K/9K
24th of Oct 2022

CEES
Centre for Ecological and Evolutionary Synthesis

UiO : University of Oslo

Evolutionary Biologist

specialize in ancient DNA

Archaeogenomics group
(10+ MSc, PhDs & Postdocs)

@archaeogenomics

Multidisciplinary research:
Archaeology
Biology
Ecology
Molecular methods/sequencing
Genomics
Bioinformatics

Today:

1) Introduction: variant calling, why do we want to do this, and what it is?

Today:

1) Introduction: variant calling, why do we want to do this, and what it is?
2) Variant calling pipelines/methods and limitations

Today:

1) Introduction: variant calling, why do we want to do this, and what it is?
2) Variant calling pipelines/methods and limitations
3) Practical session, going through (parts of) a SNP calling pipeline and interpret biological results

Introduction

Genetic variation (genomic differences between individuals) is everywhere

Genetic variation at different scales:

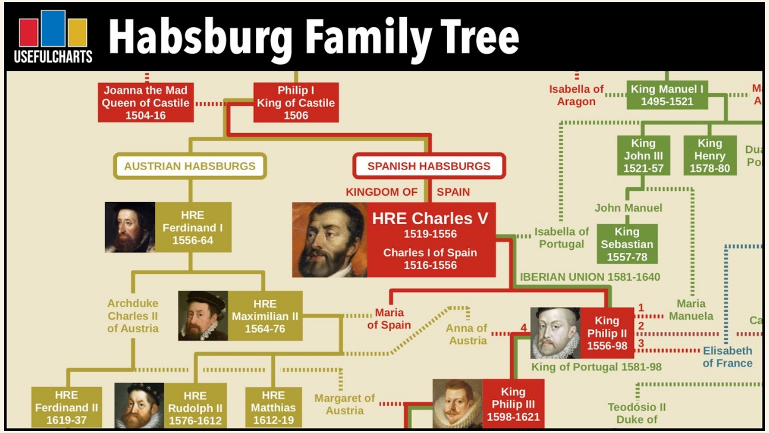
1) Biological differences (phenotypes) between species

Genetic variation at different scales :

1) Biological differences (phenotypes) between species
2) Biological differences within species

Genetic variation at different scales :

1) Biological differences (phenotypes) between species
2) Biological differences within species
3) Patterns of relatedness between individuals/ populations (23 and me)



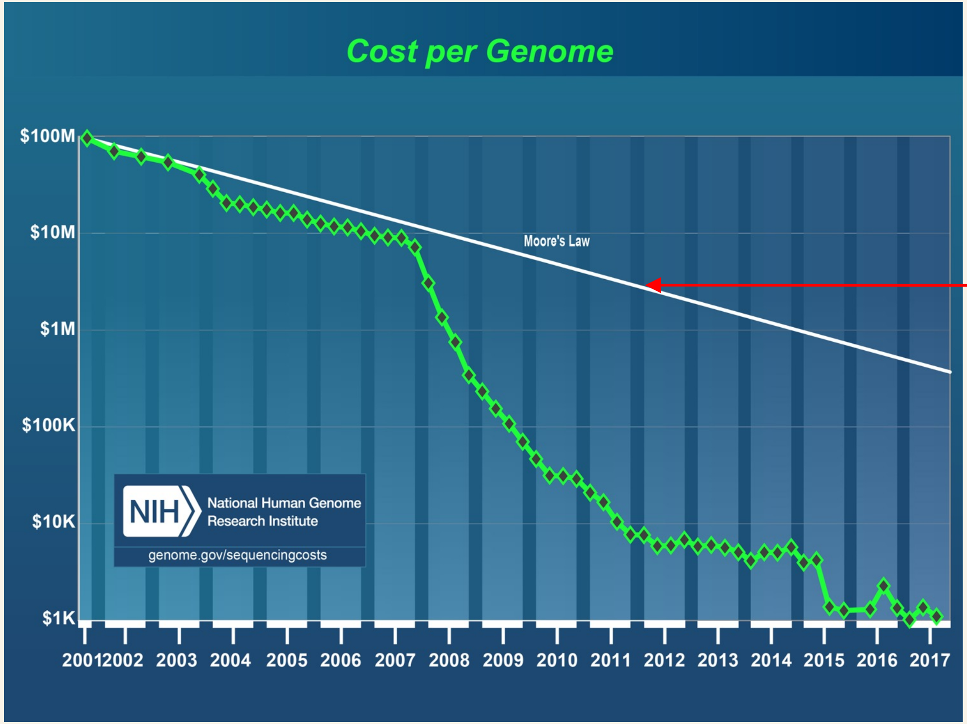
Genetic variation explains many observations within biology

Genetic variation explains many observations within biology

Knowing patterns of/quantifying genetic variation has enormous potential for a wide range of applications in society

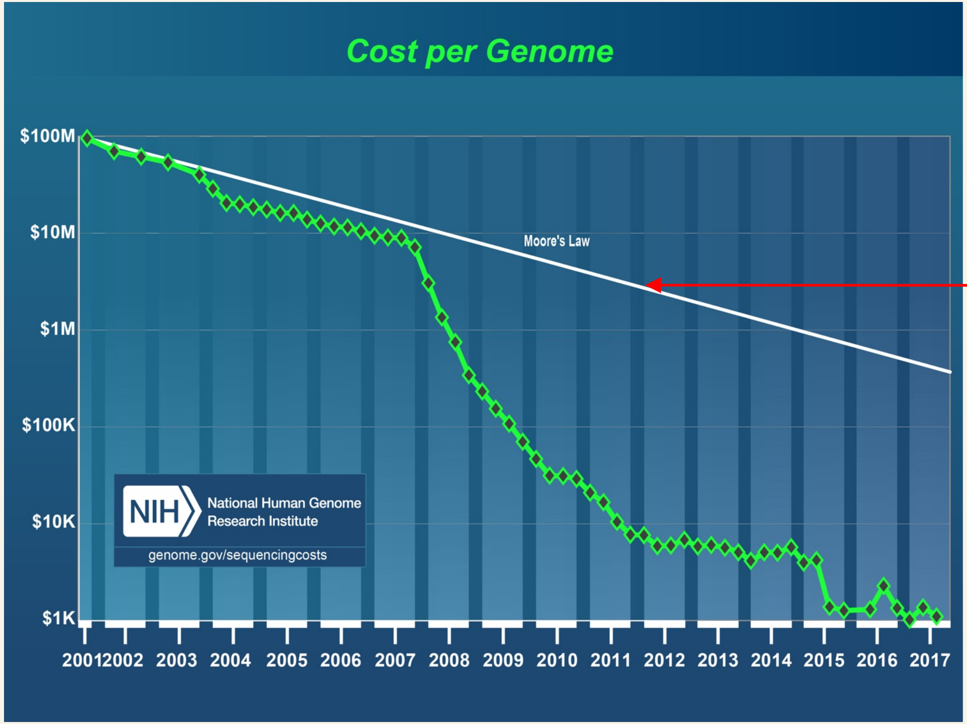
(e.g. personal medicine, forensic sciences, biodiversity assessments, crop improvement, animal breeding, conservation management, history & genealogy, etc etc)

Why are we here?



Moore's Law

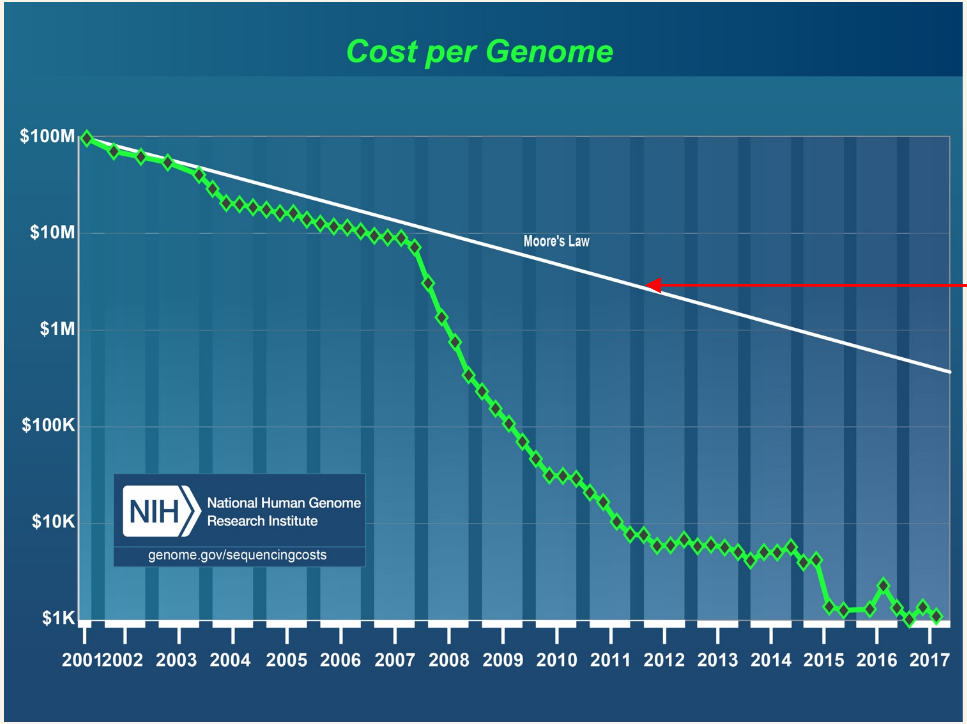
Why are we here?



2000

2017

Why are we here? ***Phenomenal*** technological advances

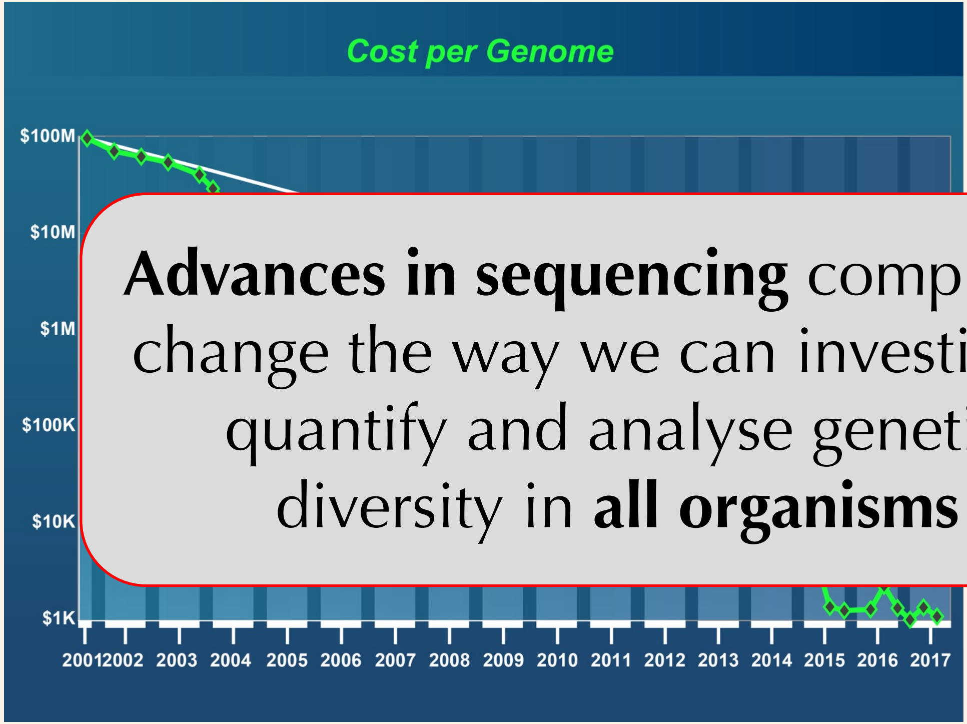


2000

2017

Technological revolution that has *fundamentally* changed the way we do biology

Why are we here? ***Phenomenal*** technological advances



Advances in sequencing completely change the way we can investigate, quantify and analyse genetic diversity in **all organisms**

2000

2017

Technological revolution that has *fundamentally* changed the way we do biology

How has sequencing changed and is changing the world?

Changed healthcare

Sequencing (genome and exome) funded solely by *healthcare* systems

2012

~1%

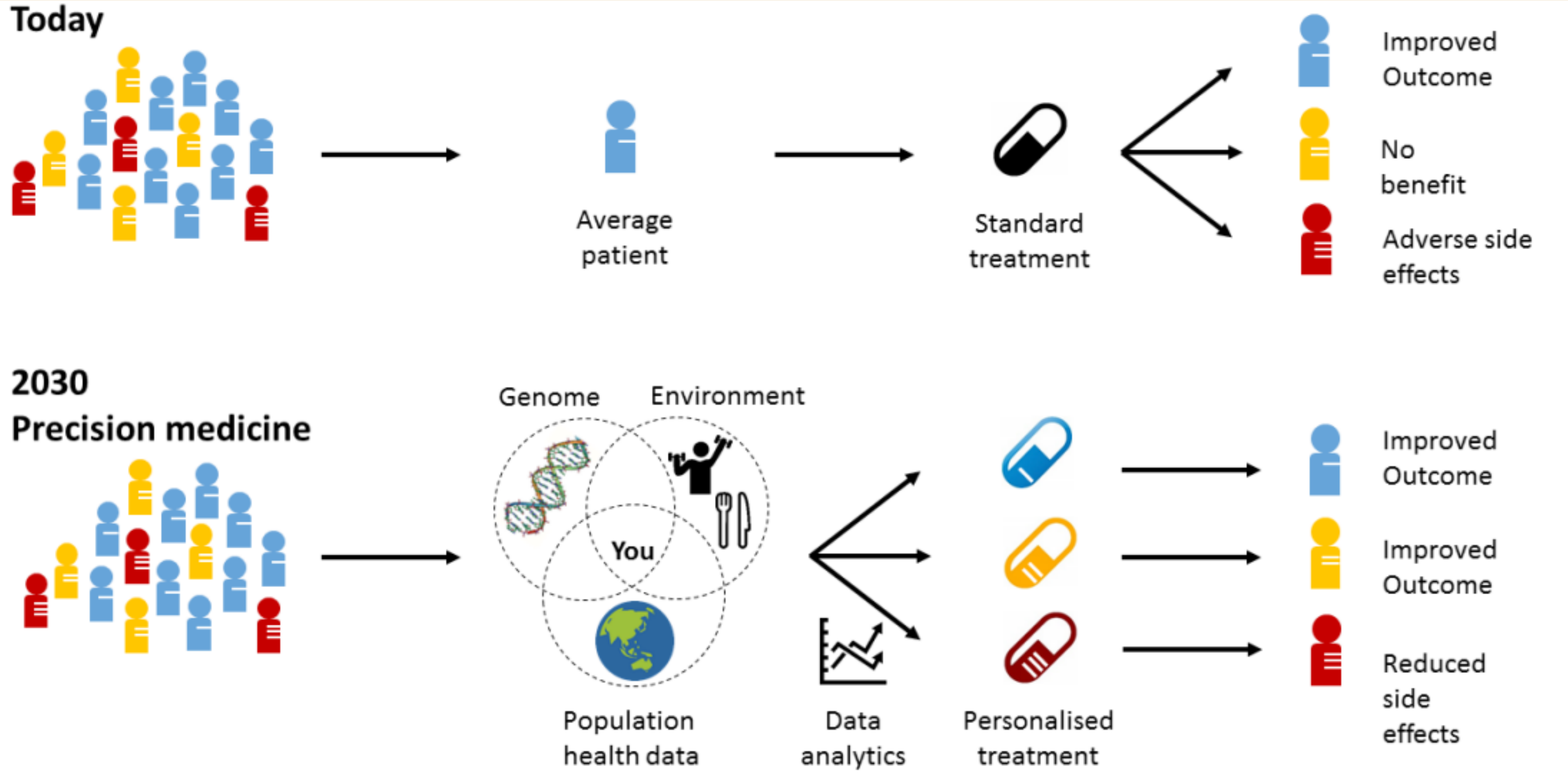
2017

~20%

2022

>80%

Dag Undlien (OUH)



Today

Average patient

Standard treatment

Improved Outcome

No benefit

Adverse side effects

2030
Precision medicine

Genome

Environment

You

Population health data

Data analytics

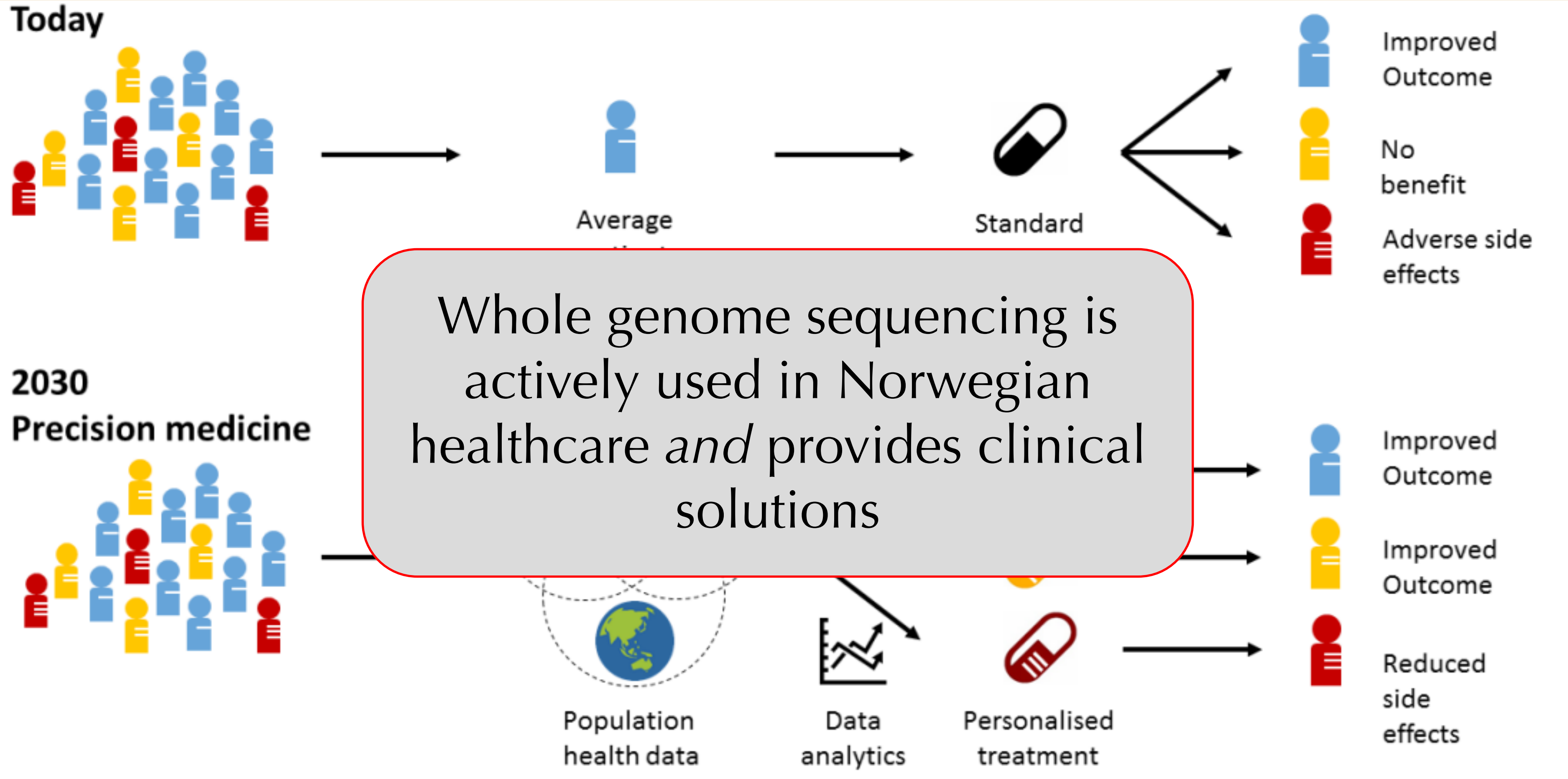
Personalised treatment

Improved Outcome

Improved Outcome

Reduced side effects

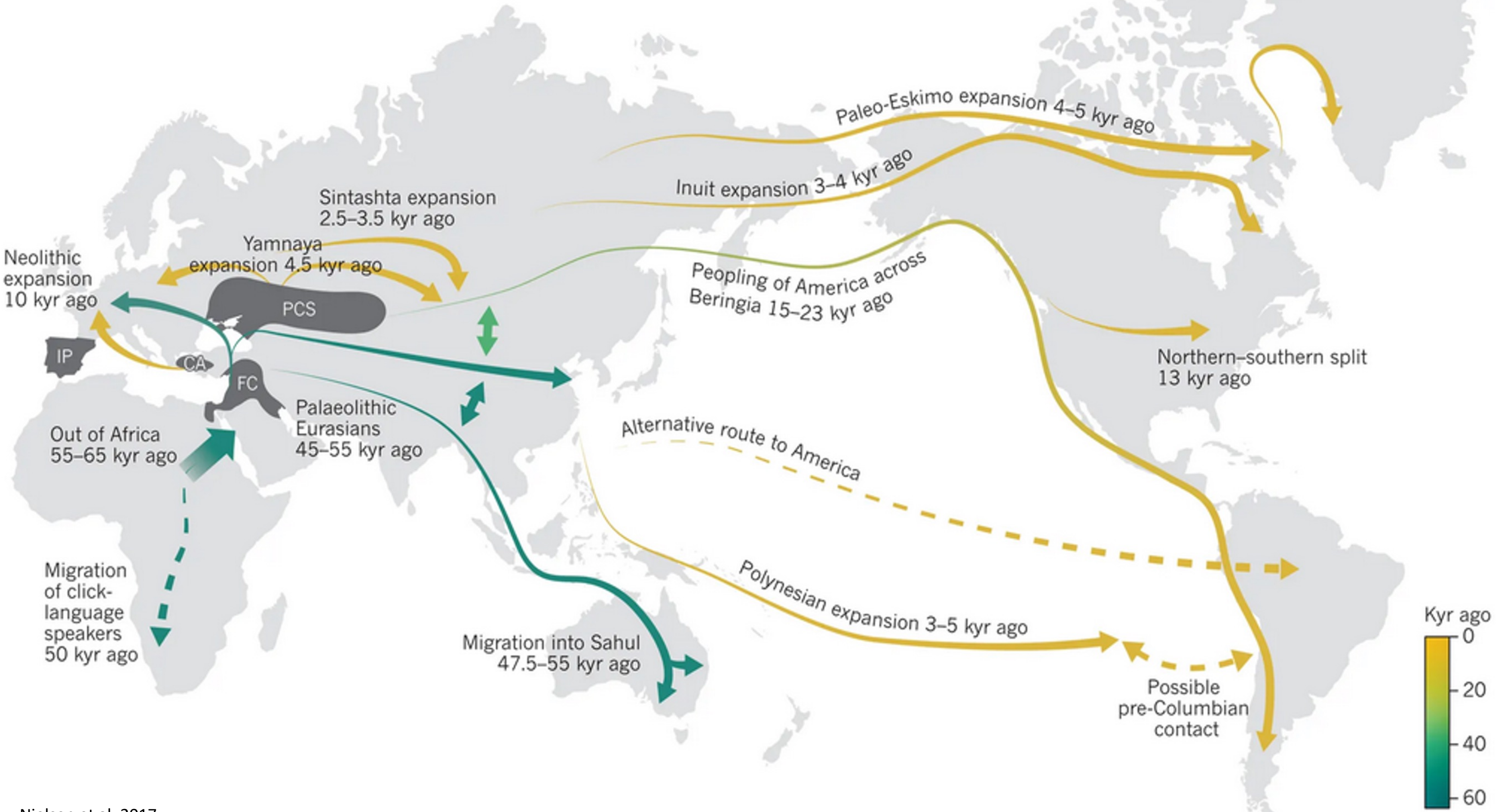
Dag Undlien (OUH)



Whole genome sequencing is actively used in Norwegian healthcare *and* provides clinical solutions

Dag Undlien (OUH)

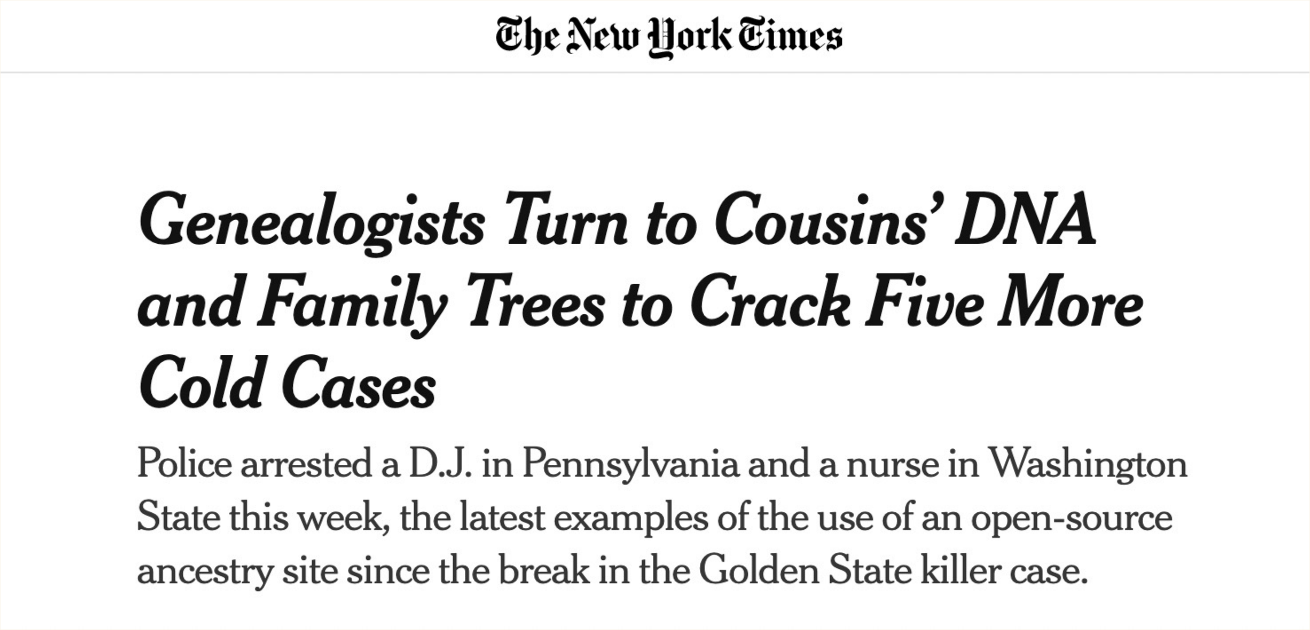
Changed our perspective of human history


Paleo-Eskimo expansion 4–5 kyr ago
Inuit expansion 3–4 kyr ago
Sintashta expansion 2.5–3.5 kyr ago
Yamnaya expansion 4.5 kyr ago
Neolithic expansion 10 kyr ago
PCS
IP
CA
FC
Peopling of America across Beringia 15–23 kyr ago
Northern–southern split 13 kyr ago
Palaeolithic Eurasians 45–55 kyr ago
Out of Africa 55–65 kyr ago
Alternative route to America
Migration of click-language speakers 50 kyr ago
Polynesian expansion 3–5 kyr ago
Migration into Sahul 47.5–55 kyr ago
Possible pre-Columbian contact
Kyr ago
0
20
40
60

Nielsen et al. 2017

Changed forensic capabilities

Using continuously expanding public genomic databases (e.g. 23 and me)…



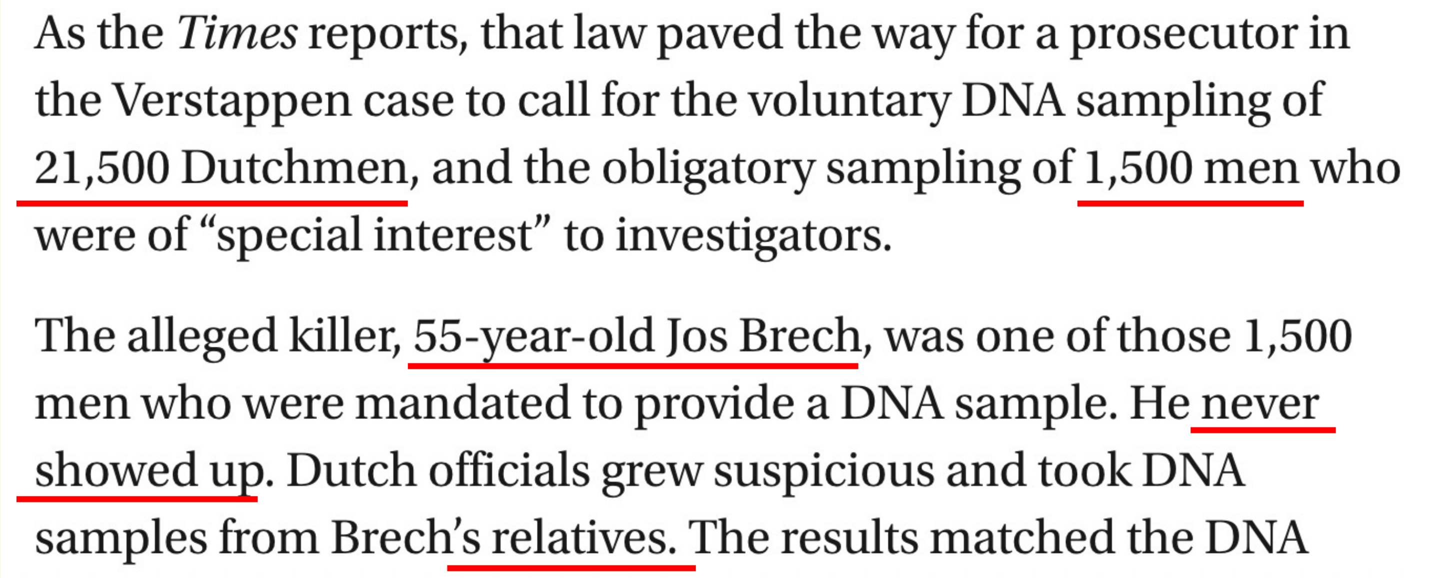
The New York Times

Genealogists Turn to Cousins' DNA and Family Trees to Crack Five More Cold Cases

Police arrested a D.J. in Pennsylvania and a nurse in Washington State this week, the latest examples of the use of an open-source ancestry site since the break in the Golden State killer case.

Changed forensic capabilities

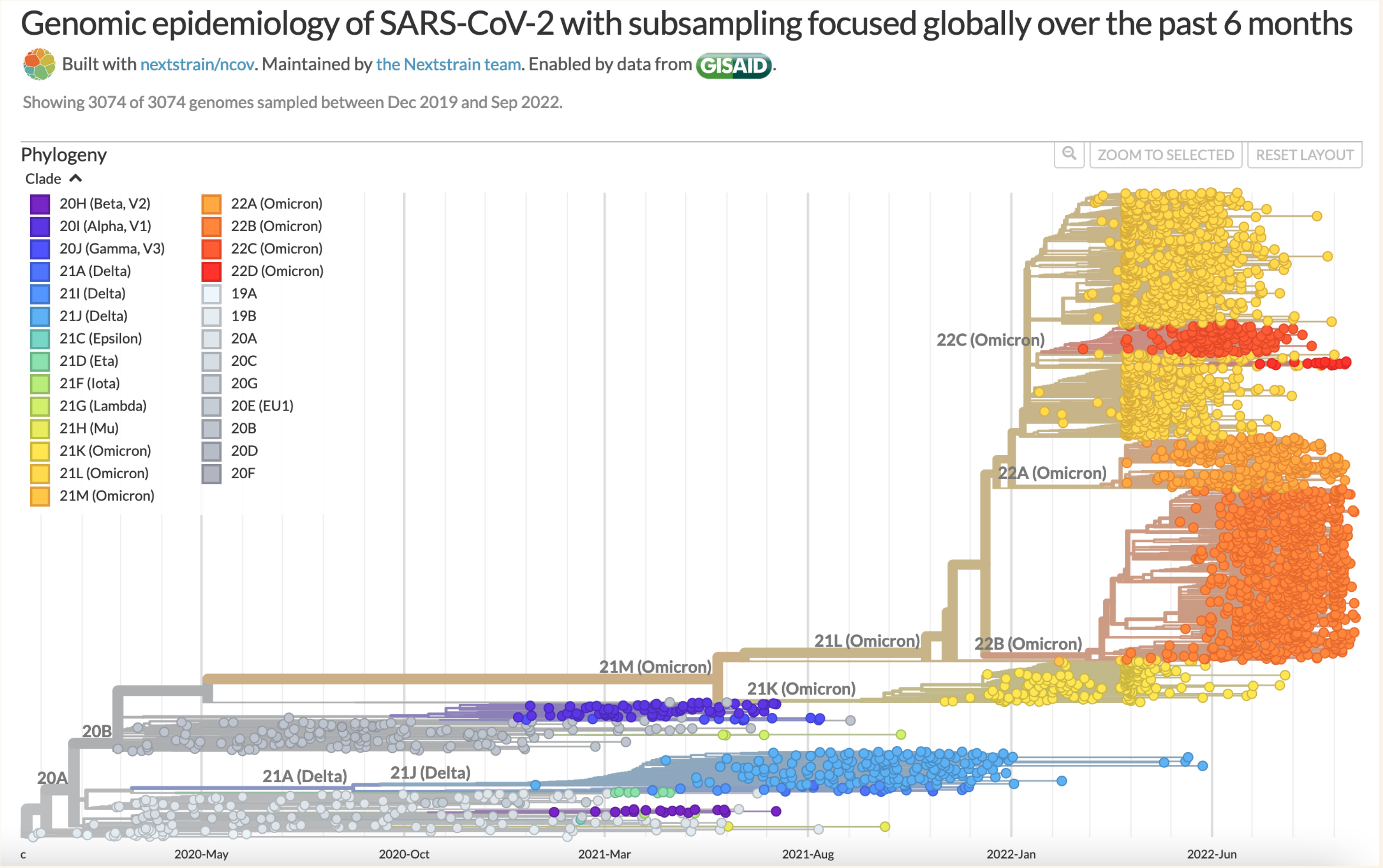
Or by the genetic testing of thousands of people!



As the *Times* reports, that law paved the way for a prosecutor in the Verstappen case to call for the voluntary DNA sampling of 21,500 Dutchmen, and the obligatory sampling of 1,500 men who were of "special interest" to investigators.

The alleged killer, 55-year-old Jos Brech, was one of those 1,500 men who were mandated to provide a DNA sample. He never showed up. Dutch officials grew suspicious and took DNA samples from Brech's relatives. The results matched the DNA

Changed vaccine development and disease tracking

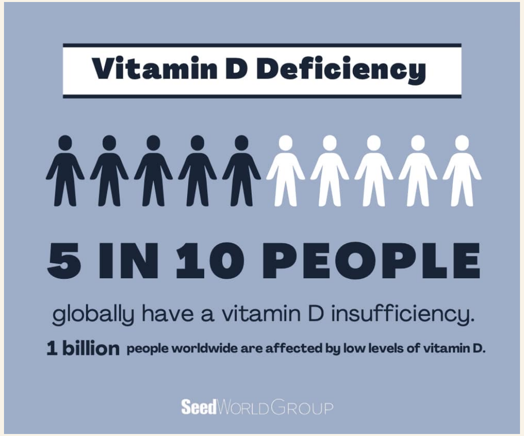


Genomic epidemiology of SARS-CoV-2 with subsampling focused globally over the past 6 months

Built with nextstrain/ncov. Maintained by the Nextstrain team. Enabled by data from GISAID.

Showing 3074 of 3074 genomes sampled between Dec 2019 and Sep 2022.

Changed improvement and selection of commercial crops



Vitamin D Deficiency

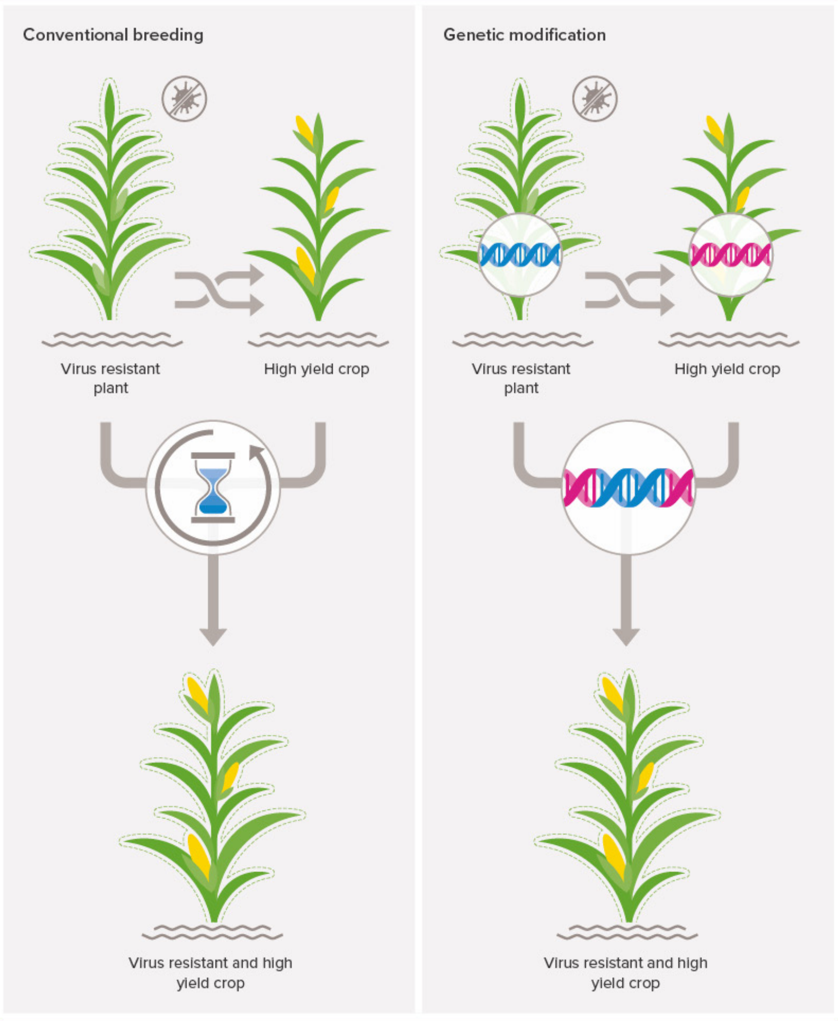
5 IN 10 PEOPLE

globally have a vitamin D insufficiency.

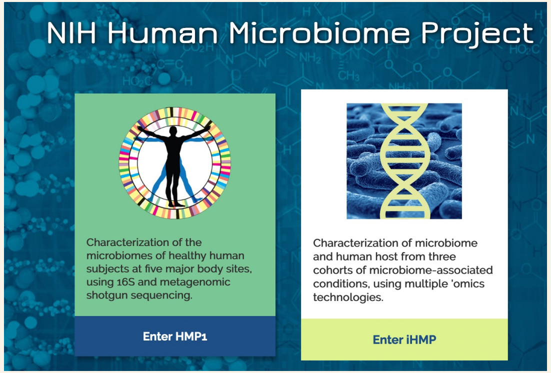
1 billion people worldwide are affected by low levels of vitamin D.

SeedWorldGroup

FIGURE 3 Differences between conventional breeding and GM



Changed our understanding of the human microbiome



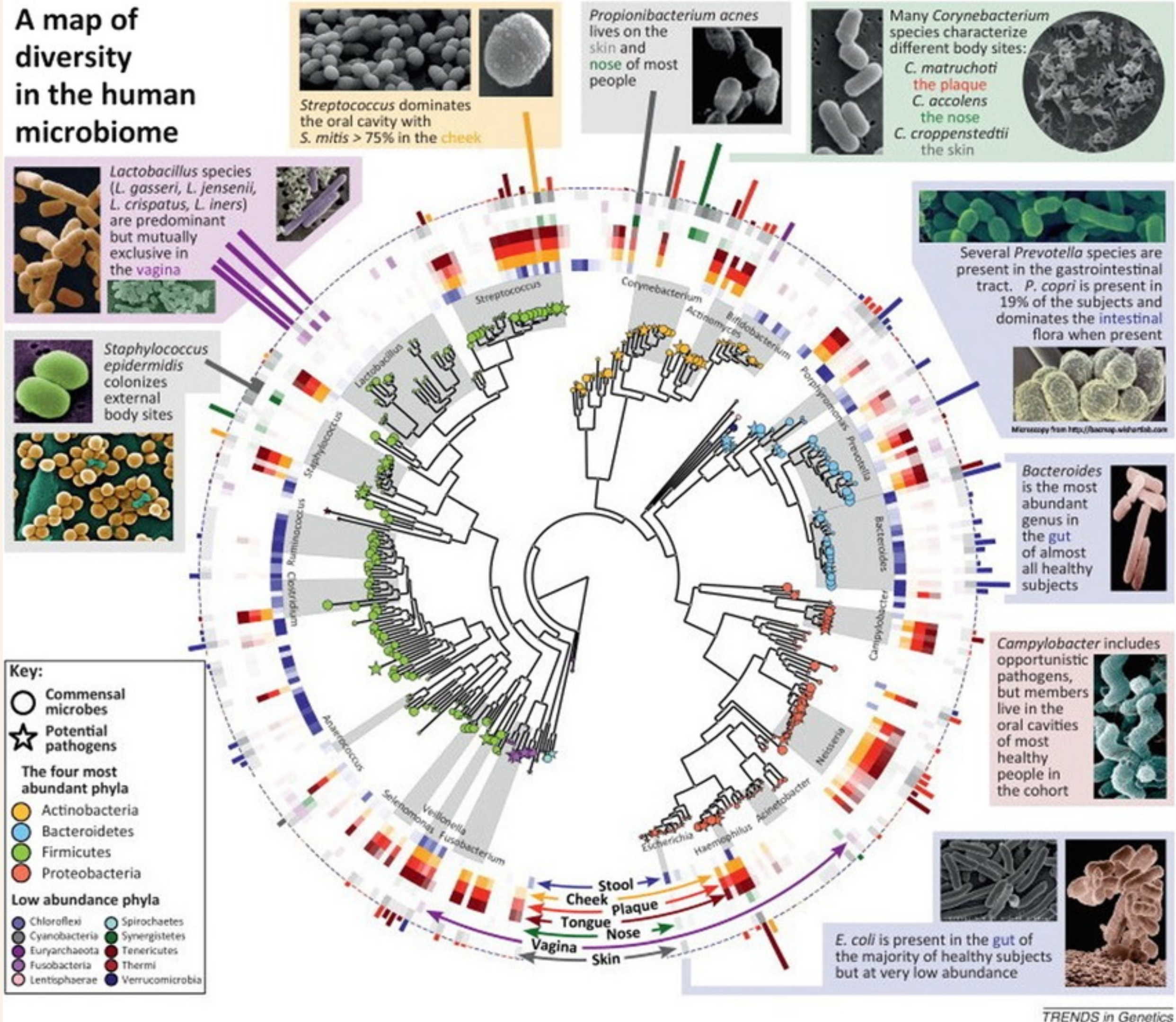
NIH Human Microbiome Project

Characterization of the microbiomes of healthy human subjects at five major body sites, using 16S and metagenomic shotgun sequencing.

Enter HMP1

Characterization of microbiome and human host from three cohorts of microbiome-associated conditions, using multiple 'omics technologies.

Enter iHMP



A map of diversity in the human microbiome

Streptococcus dominates the oral cavity with *S. mitis* > 75% in the cheek

Propionibacterium acnes lives on the skin and nose of most people

Many *Corynebacterium* species characterize different body sites:
C. matruchotii the plaque
C. accolens the nose
C. croppenstedtii the skin

Lactobacillus species (*L. gasseri, L. jensenii, L. crispatus, L. iners*) are predominant but mutually exclusive in the vagina

Staphylococcus epidermidis colonizes external body sites

Several *Prevotella* species are present in the gastrointestinal tract. *P. copri* is present in 19% of the subjects and dominates the intestinal flora when present

Microscopy from http://bacmap.wishartlab.com

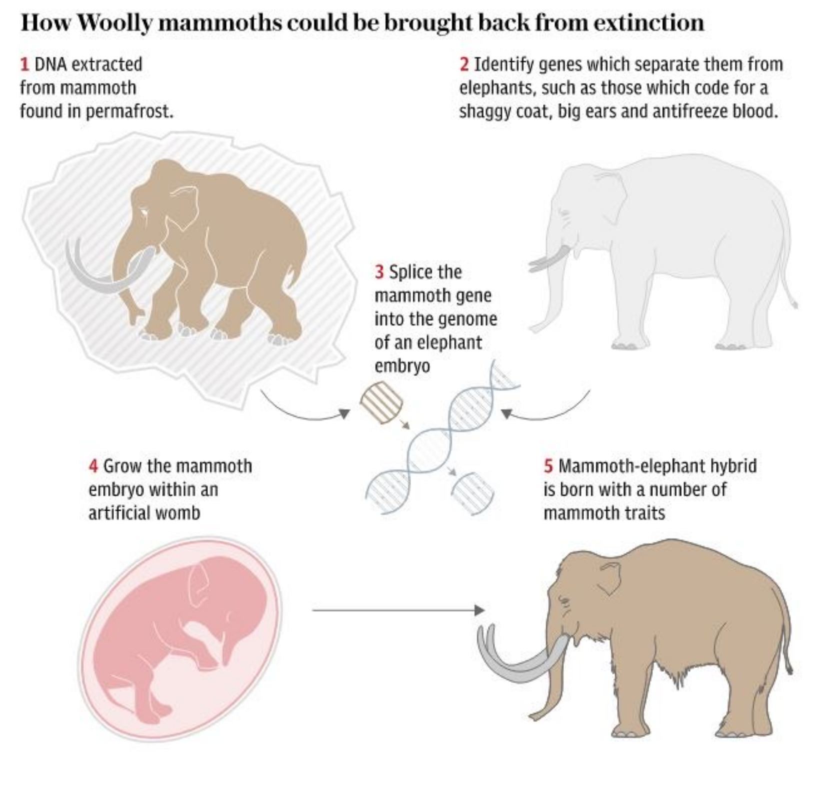
Bacteroides is the most abundant genus in the gut of almost all healthy subjects

Campylobacter includes opportunistic pathogens, but members live in the oral cavities of most healthy people in the cohort

E. coli is present in the gut of the majority of healthy subjects but at very low abundance

TRENDS in Genetics

Changing our perspective of extinct species



How Woolly mammoths could be brought back from extinction

1 DNA extracted from mammoth found in permafrost.

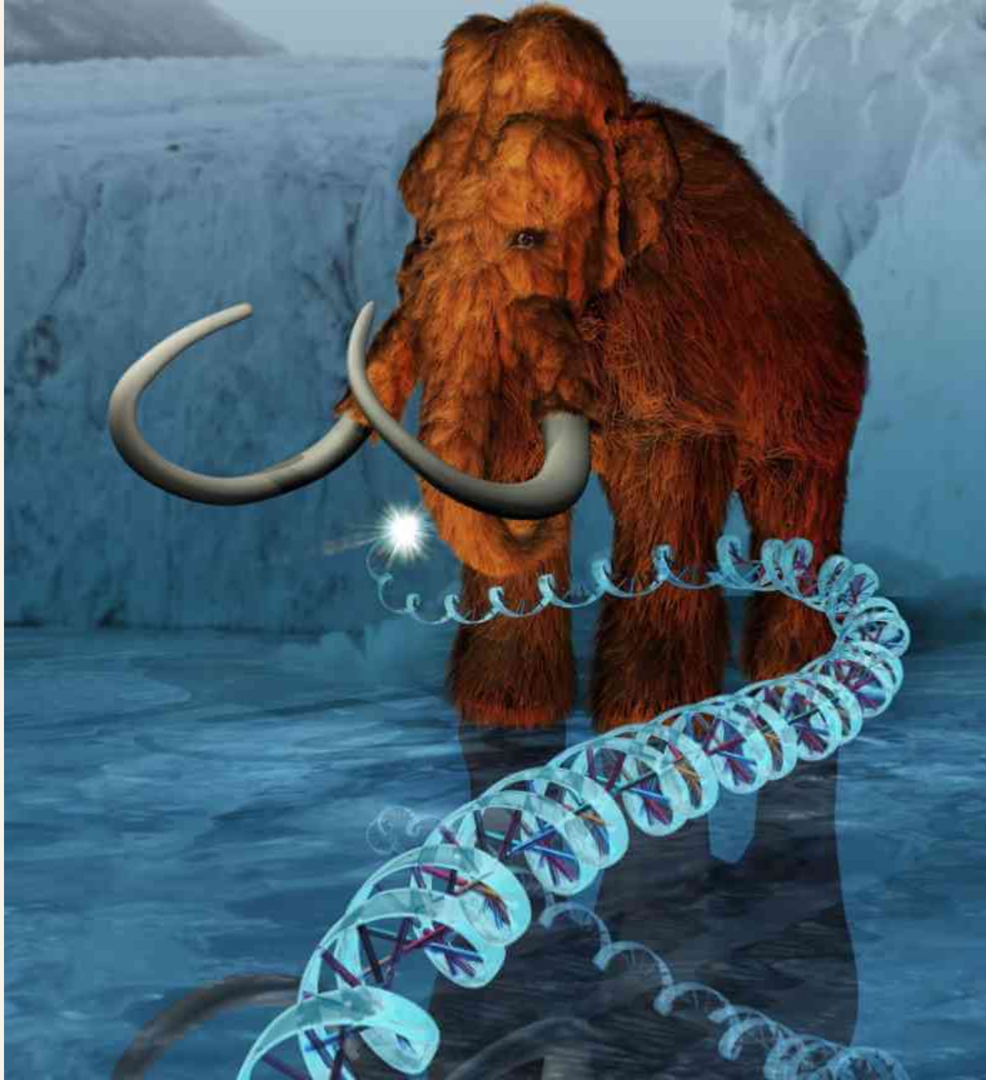
2 Identify genes which separate them from elephants, such as those which code for a shaggy coat, big ears and antifreeze blood.

3 Splice the mammoth gene into the genome of an elephant embryo

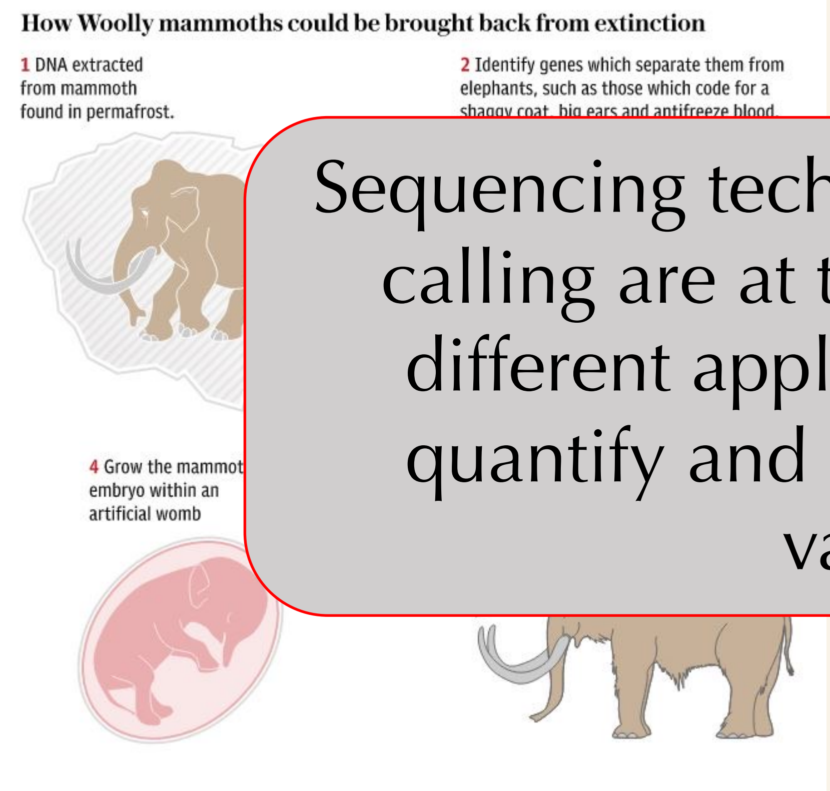
4 Grow the mammoth embryo within an artificial womb

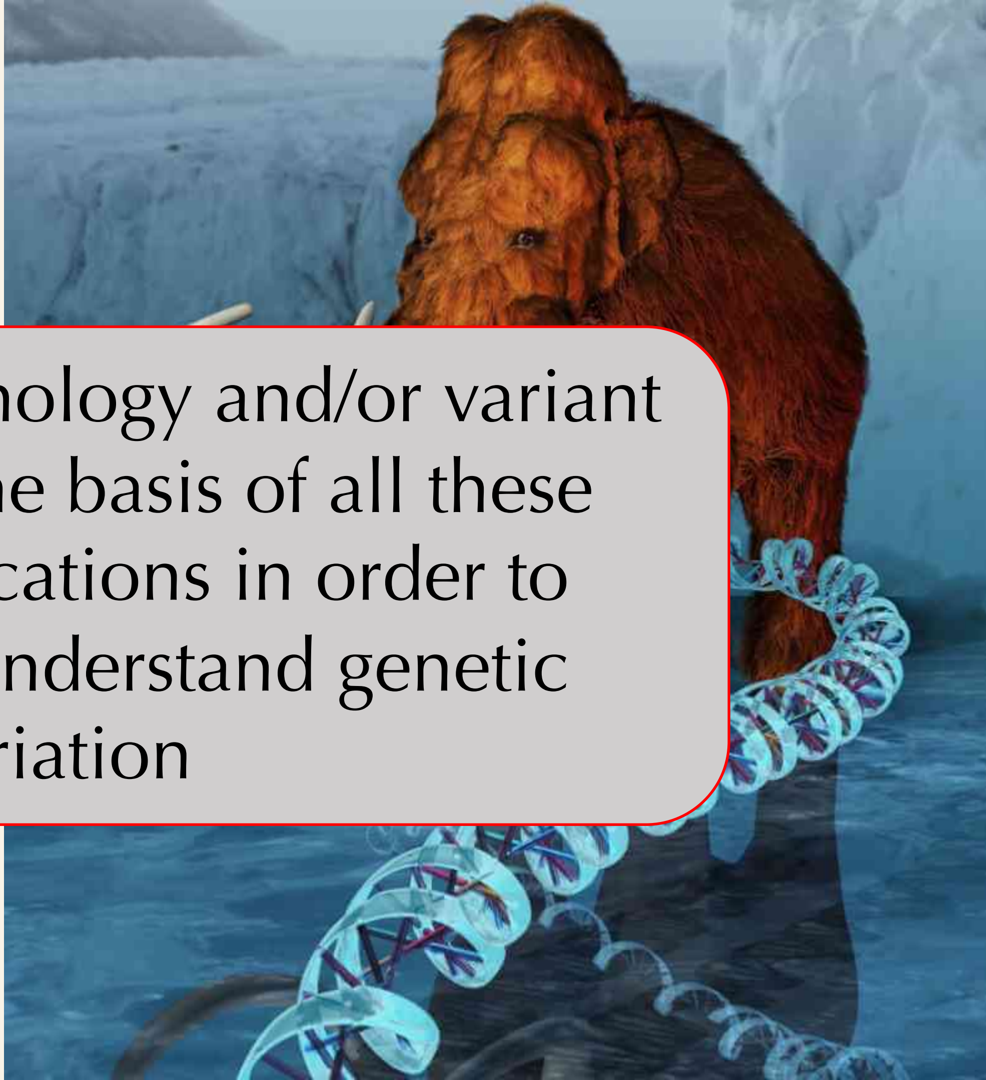
5 Mammoth-elephant hybrid is born with a number of mammoth traits

https://instinctforfilm.com/feed/cloning-mammoths-global-warming/



Changing our perspective of extinct species

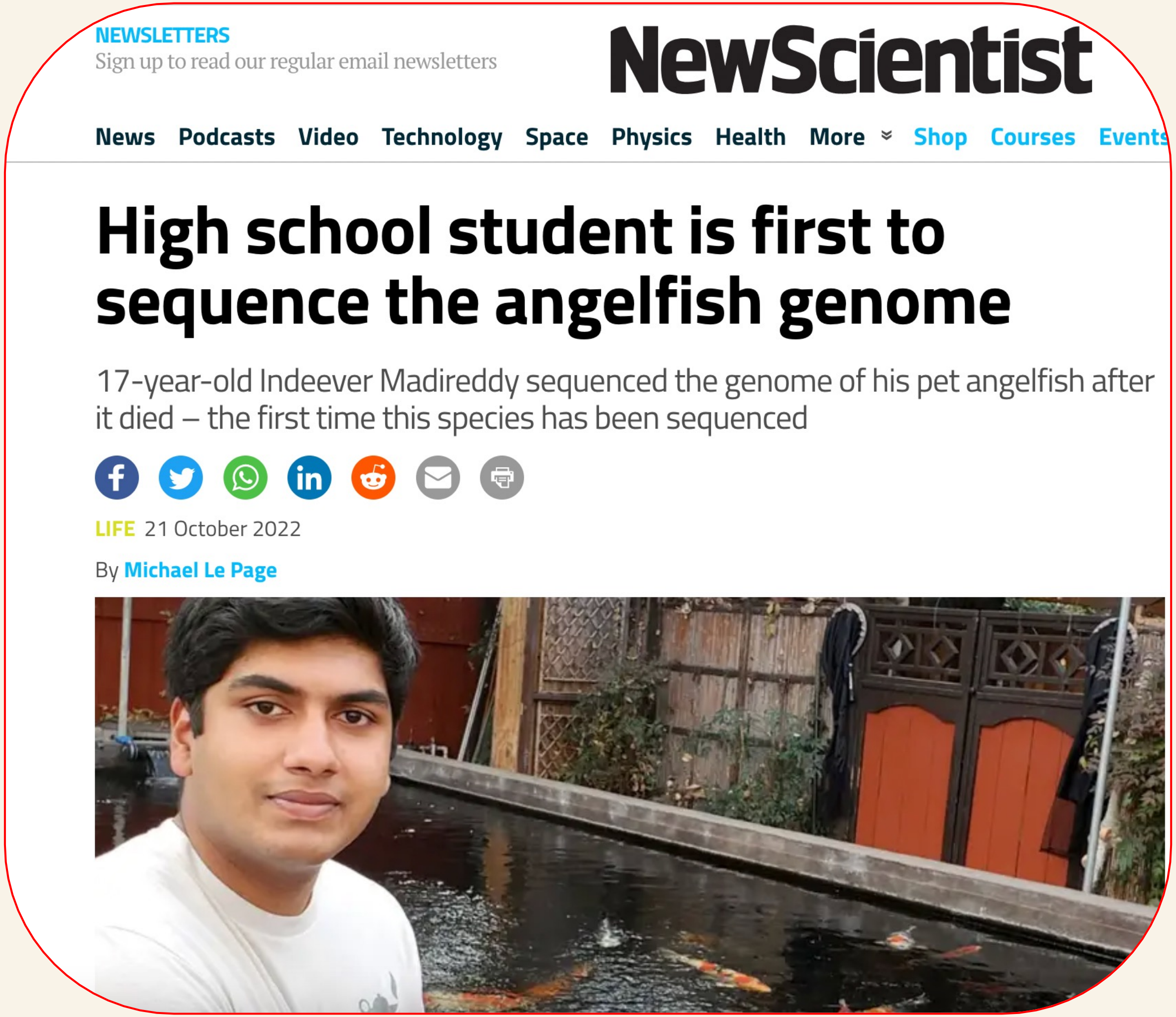

How Woolly mammoths could be brought back from extinction
1 DNA extracted from mammoth found in permafrost.
2 Identify genes which separate them from elephants, such as those which code for a shaggy coat, big ears and antifreeze blood
4 Grow the mammot
embryo within an artificial womb



Sequencing technology and/or variant calling are at the basis of all these different applications in order to quantify and understand genetic variation

https://instinctforfilm.com/feed/cloning-mammoths-global-warming/

Available to
high school
students!
21.10 2022



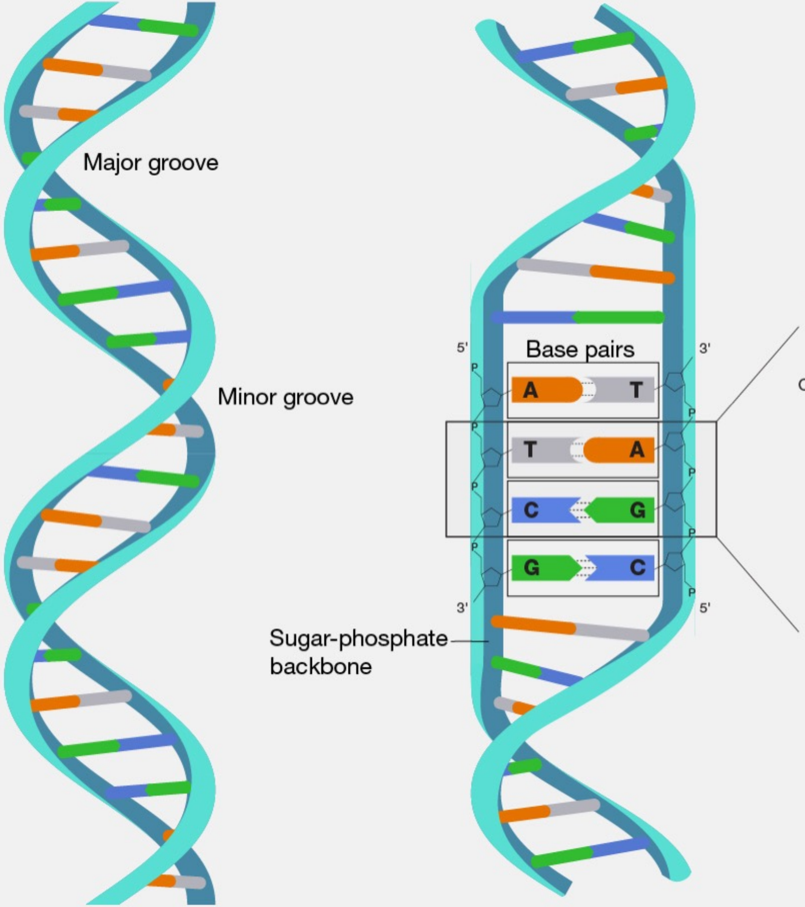
High school student is first to sequence the angelfish genome

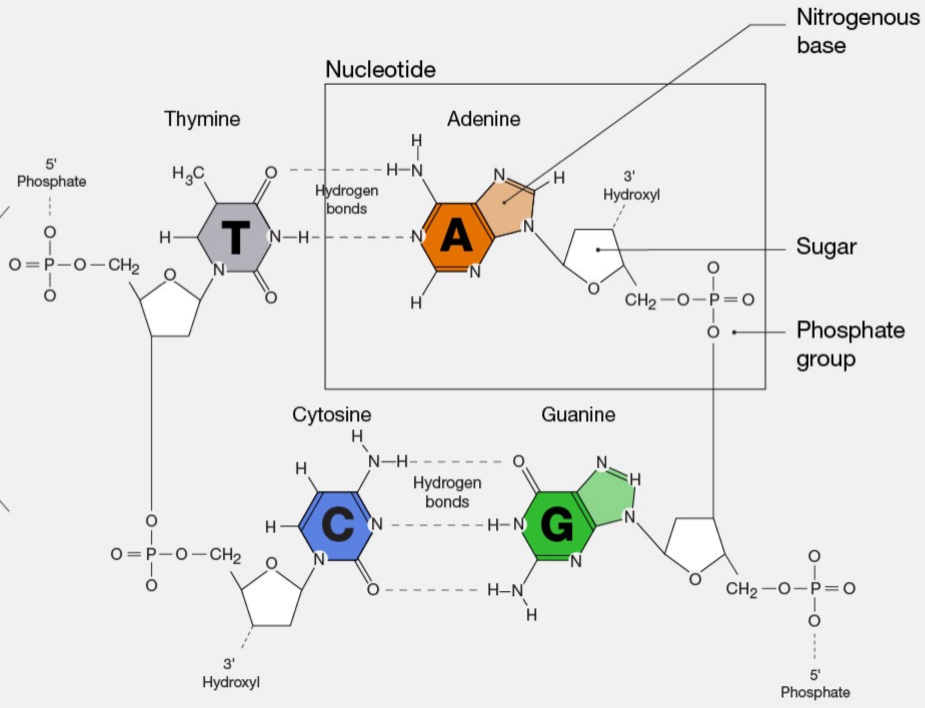
17-year-old Indeever Madireddy sequenced the genome of his pet angelfish after it died – the first time this species has been sequenced

LIFE 21 October 2022

By Michael Le Page

DNA

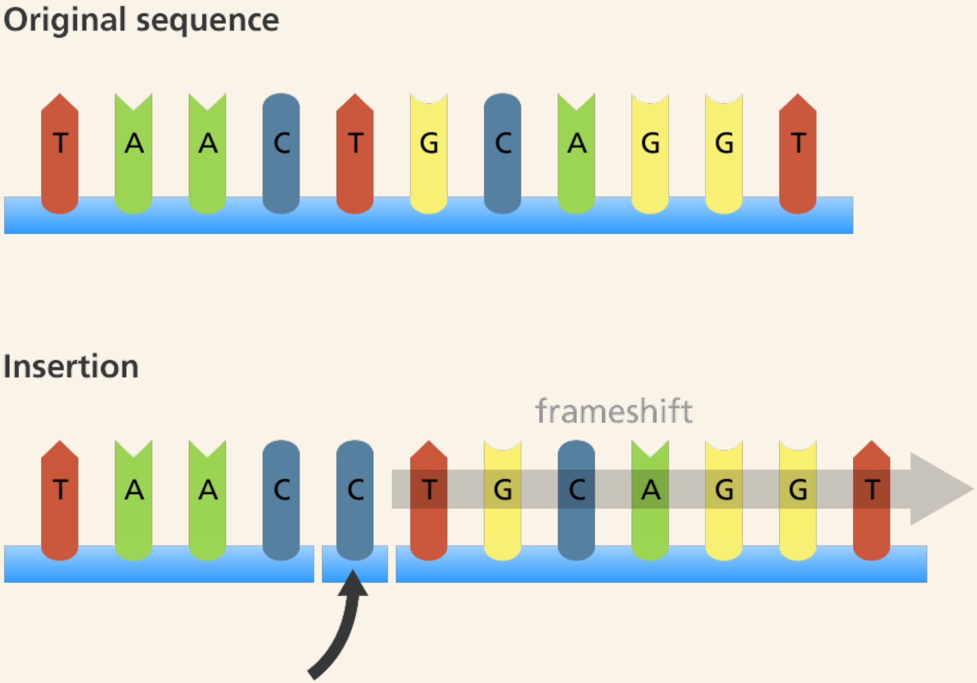

Major groove
Minor groove
Sugar-phosphate backbone
Base pairs
5'
3'
P
A
T
T
A
C
G
G
C
3'
5'


Nitrogenous base
Nucleotide
Thymine
Adenine
5' Phosphate
Hydrogen bonds
3' Hydroxyl
Sugar
Phosphate group
T
A
H_3C
O
H–N
H
N
N–H
CH₂
O=P–O–CH₂
Cytosine
Guanine
C
G
Hydrogen bonds
3' Hydroxyl
5' Phosphate

What does genetic variation look like?

1) DNA (nucleotides) can be inserted or deleted (*indels*).

1) Insertion/Deletion (Indel)



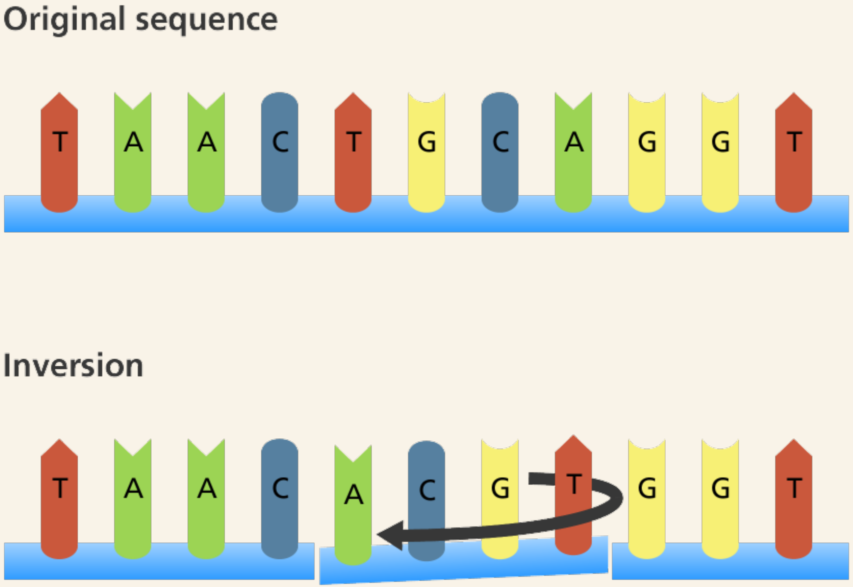
Can range from 1 base-pair (bp) to many bp

What does genetic variation look like?

1) DNA (nucleotides) can be inserted or deleted (*indels*).

2) DNA can be *structurally rearranged* (inversions/translocations)

2) Structural rearrangements (inversions/translocations)



Can be MILLIONS of bp long affecting the order of many genes simultaneously *(Supergenes)*

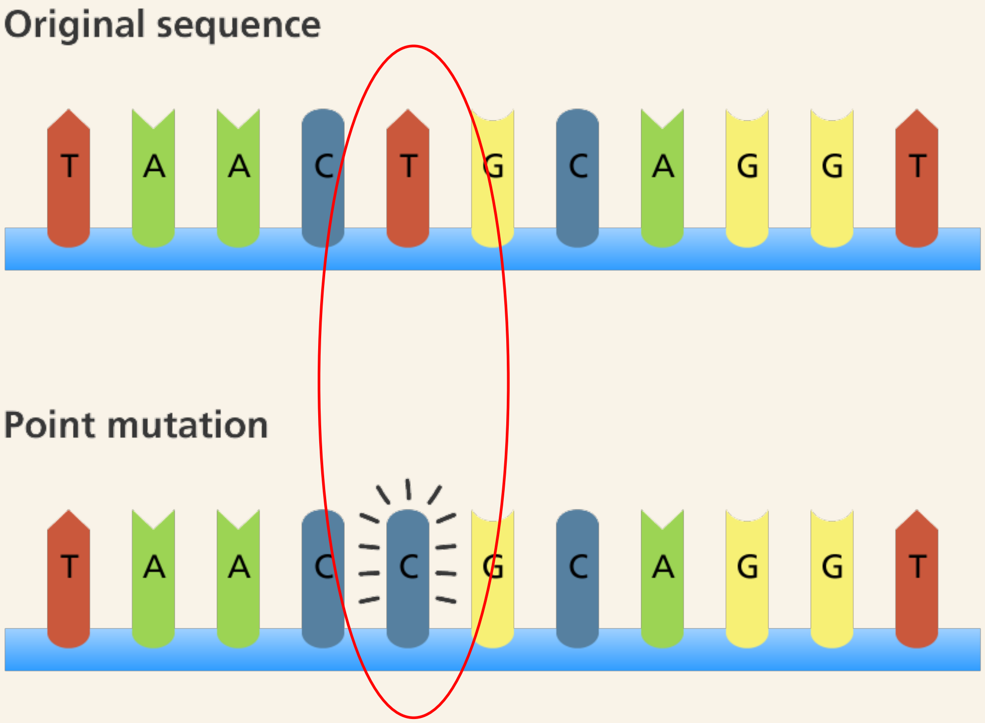
What does genetic variation look like?

1) DNA (nucleotides) can be inserted or deleted (*indels*).

2) DNA can be *structurally rearranged* (inversions/translocations)

3) DNA can be *altered* at a single base pair (Single Nucleotide Polymorphism or SNP)

3) Single Nucleotide Polymorphism (SNP)



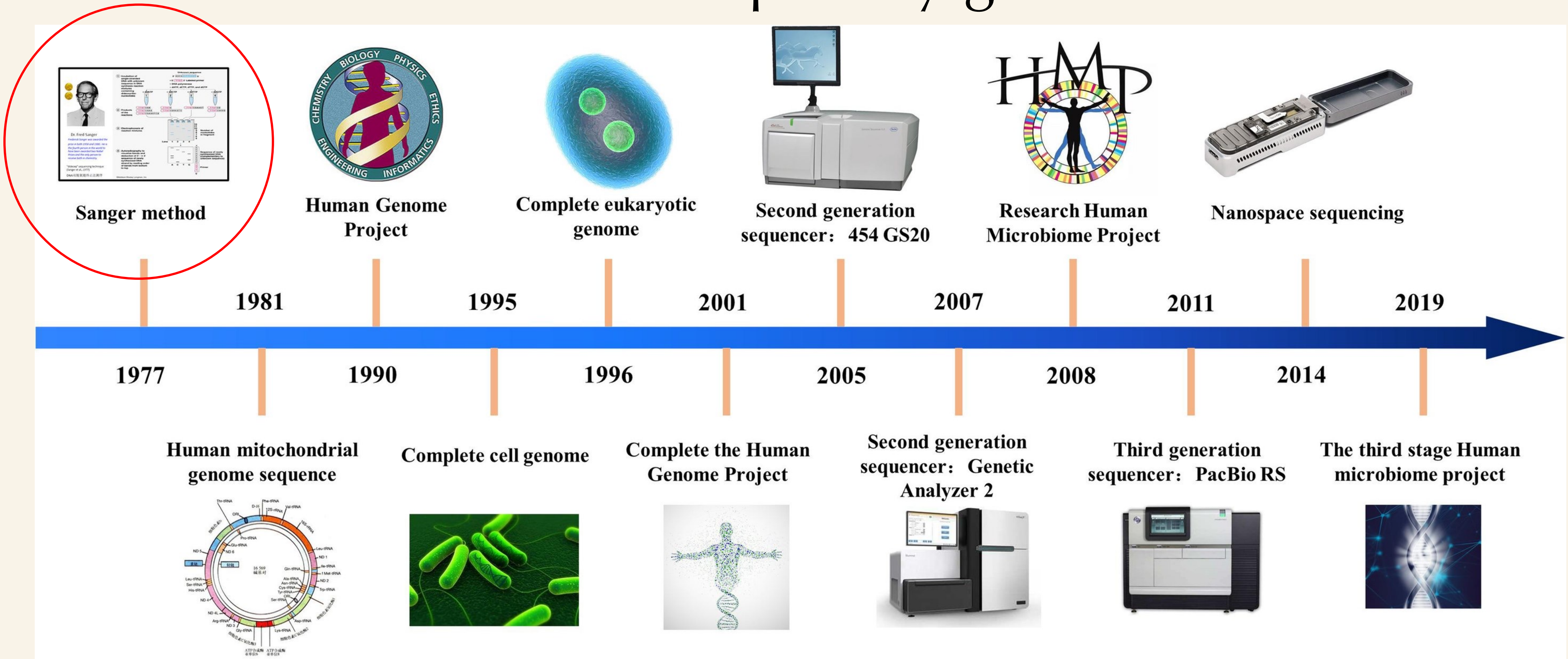
Extremely common

> 150 million known SNPs in humans (2015)

~100 SNPs unique in EVERY human

Putatively EVERY base in the human genome

How do we observe and quantify genetic variation?



1977

Sanger method

1981

Human mitochondrial genome sequence

1990

Human Genome Project

1995

Complete cell genome

1996

Complete eukaryotic genome

2001

Complete the Human Genome Project

2005

Second generation sequencer：454 GS20

2007

Second generation sequencer：Genetic Analyzer 2

2008

Research Human Microbiome Project

2011

Third generation sequencer：PacBio RS

2014

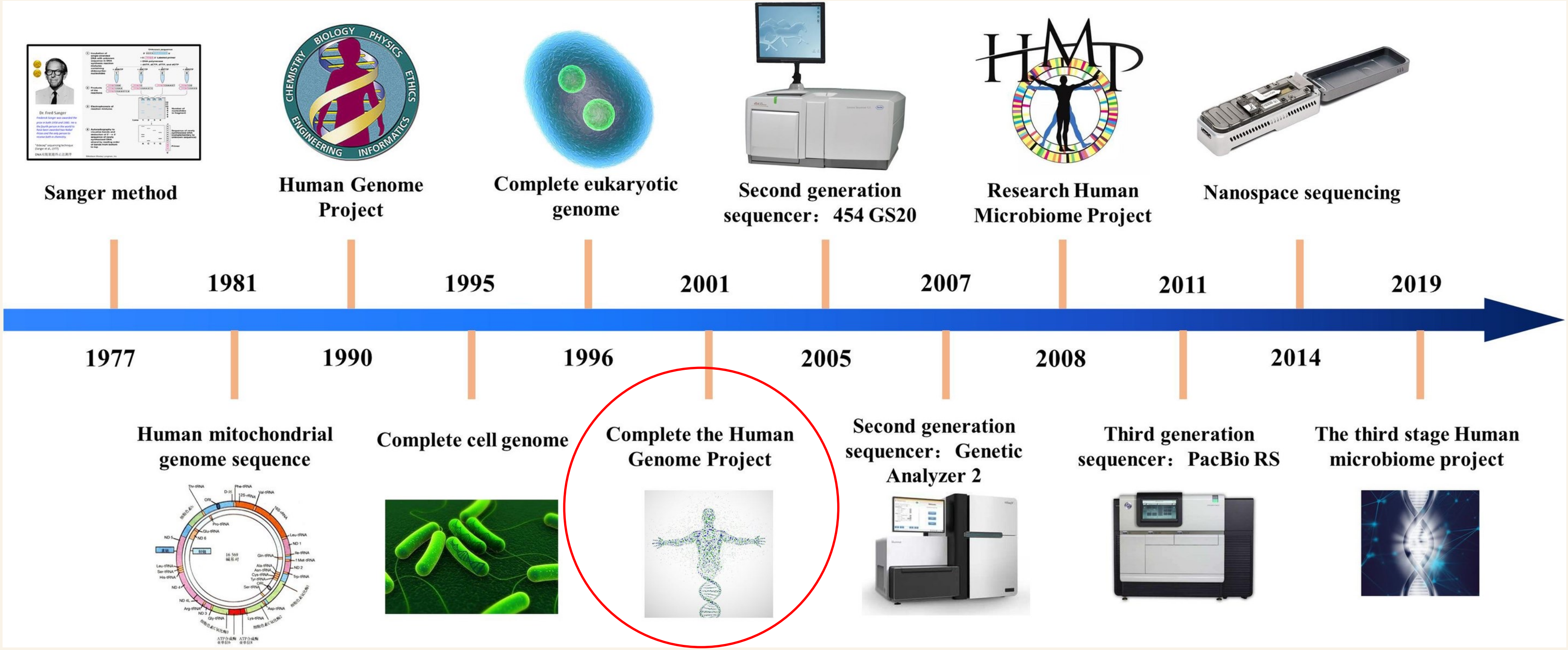
Nanospace sequencing

2019

The third stage Human microbiome project

Sanger sequencing – leading sequencing technology for decades

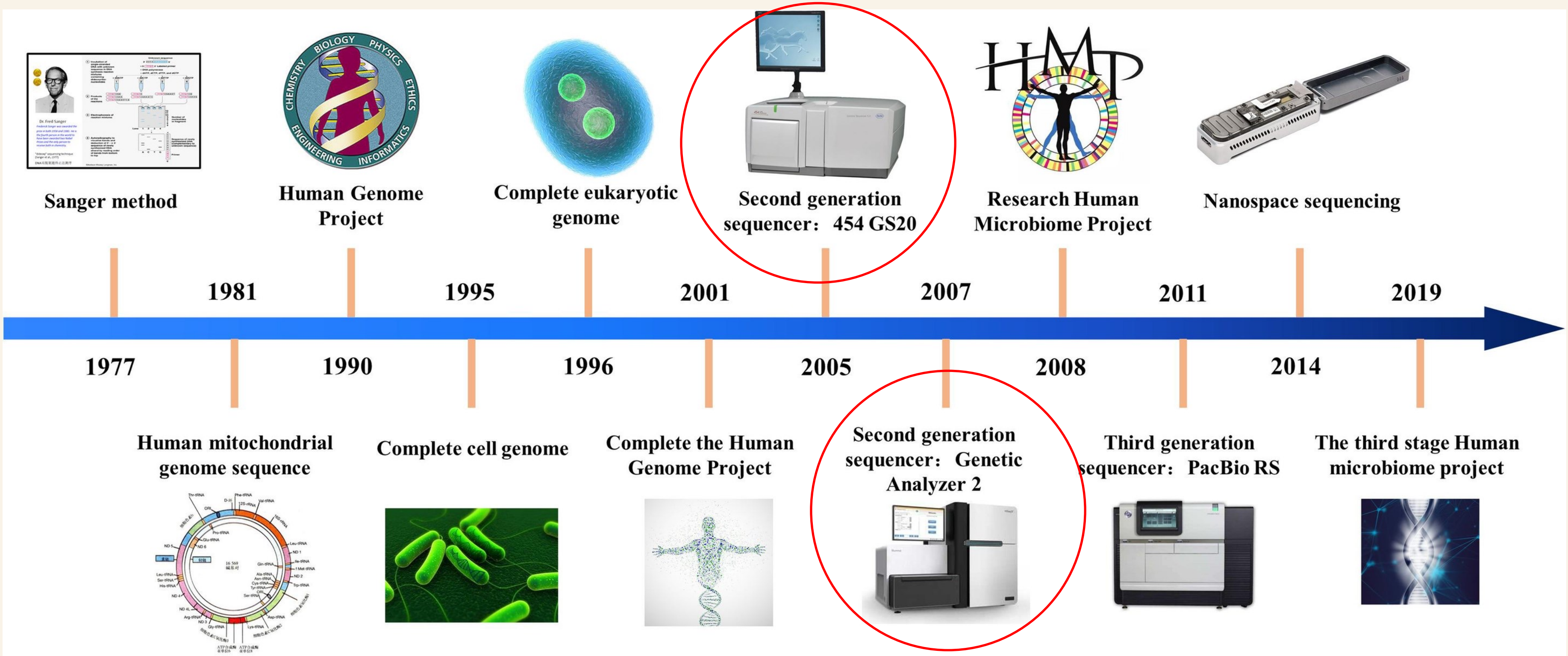
How do we observe and quantify genetic variation?



Sanger method
1977
1981
Human mitochondrial genome sequence
Human Genome Project
1990
1995
Complete cell genome
Complete eukaryotic genome
1996
2001
Complete the Human Genome Project
Second generation sequencer: 454 GS20
2005
2007
Second generation sequencer: Genetic Analyzer 2
Research Human Microbiome Project
2008
2011
Third generation sequencer: PacBio RS
Nanospace sequencing
2014
2019
The third stage Human microbiome project
BIOLOGY PHYSICS ETHICS INFORMATICS ENGINEERING CHEMISTRY
HMP

Human genome project: sparked a novel industry

How do we observe and quantify genetic variation?



1977 Sanger method

1981 Human mitochondrial genome sequence

1990 Human Genome Project

1995 Complete cell genome

1996 Complete eukaryotic genome

2001 Complete the Human Genome Project

2005 Second generation sequencer：454 GS20

2007 Second generation sequencer：Genetic Analyzer 2

2008 Research Human Microbiome Project

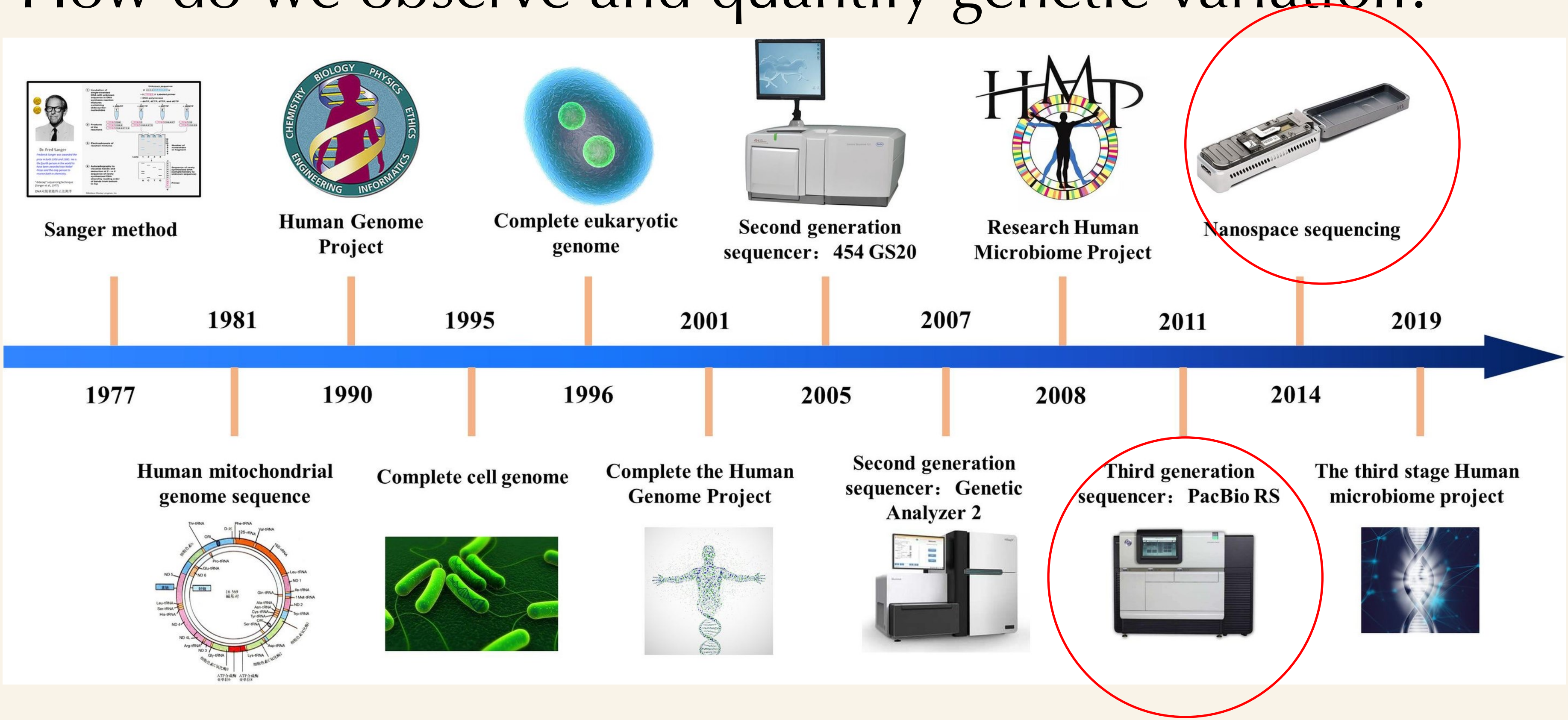
2011 Third generation sequencer：PacBio RS

2014 Nanospace sequencing

2019 The third stage Human microbiome project

“New” sequencing technologies (already outdated!)

How do we observe and quantify genetic variation?



1977 Sanger method

1981 Human mitochondrial genome sequence

1990 Human Genome Project

1995 Complete cell genome

1996 Complete eukaryotic genome

2001 Complete the Human Genome Project

2005 Second generation sequencer：454 GS20

2007 Second generation sequencer：Genetic Analyzer 2

2008 Research Human Microbiome Project

2011 Third generation sequencer：PacBio RS

2014 Nanospace sequencing

2019 The third stage Human microbiome project

Latest sequencing technologies that focus on long read sequencing

Two dominant technologies today

PacBio
Long read length (10k bp +)
More expensive
Specific applications

Two dominant technologies today

PacBio
Long read length (10k bp +)
More expensive
Specific applications

Illumina
Short read length (150-250 bp)
Cheap!
Workhorse of sequencing

Practical considerations: size matters!

PacBio
Long read length (10k bp +)
More expensive
Specific applications

Illumina
Short read length (150-250 bp)
Cheap!
Workhorse of sequencing

What variation can you assess with these different types of reads?

Type of variant	Short reads	Long reads
Indel	Only if small (~few bp)	Yes
Structural (inversion)	Difficult	Yes
SNP	Yes	Yes

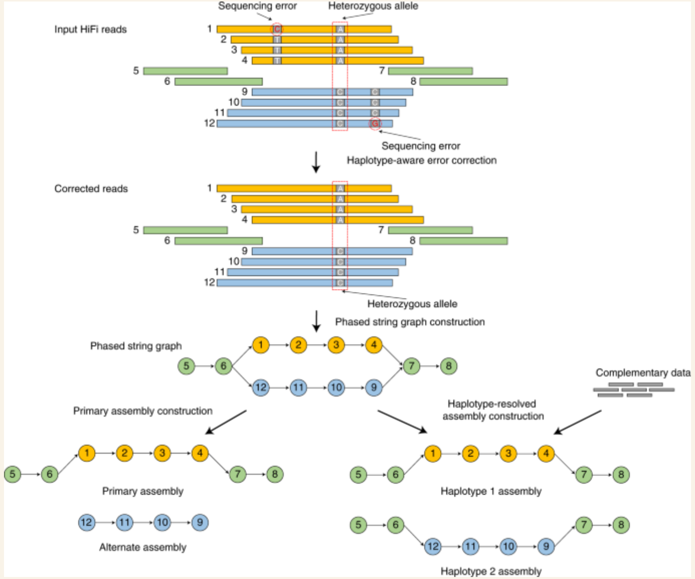
What variation can you assess with these different types of reads?

Type of variant	**Short reads**	**Long reads**
Indel	Only if small (~few bp)	Yes
Structural (inversion)	Difficult	Yes
SNP	Yes	Yes

Illumina ***re-sequencing*** domination means that SNPs are most reliably targeted and are most studied type of genetic variation

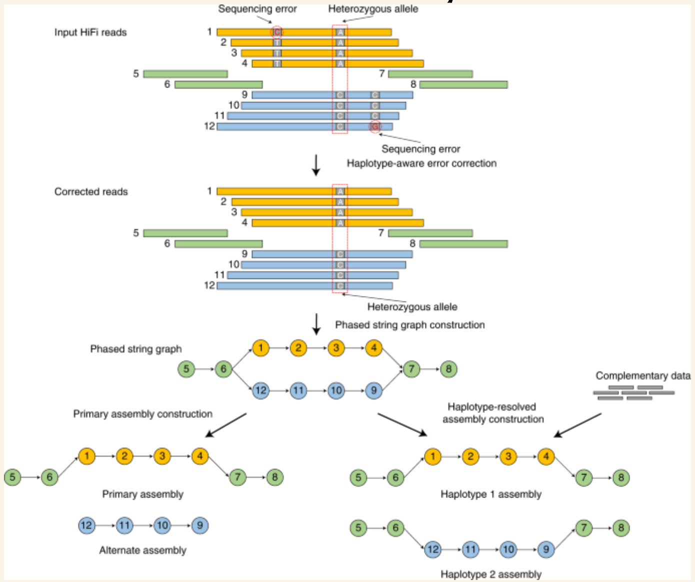
Yet there *are* different ways of assessing genetic variation

i.e. *de novo* haplotype aware assembly

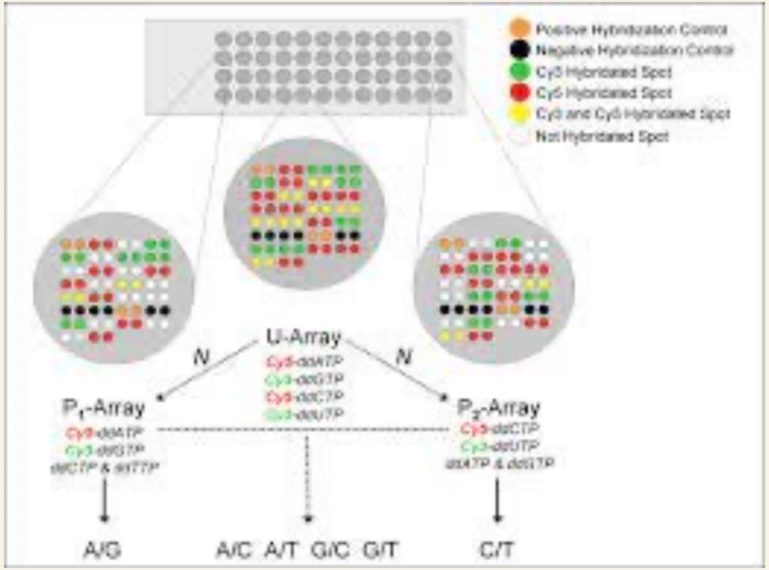


Yet there *are* different ways of assessing genetic variation

i.e. *de novo* haplotype aware assembly

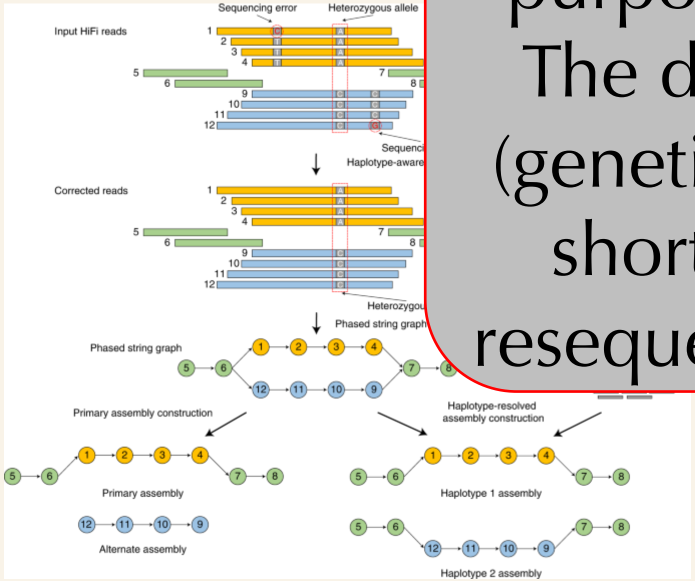


SNP Chip/DNA micro array genotyping

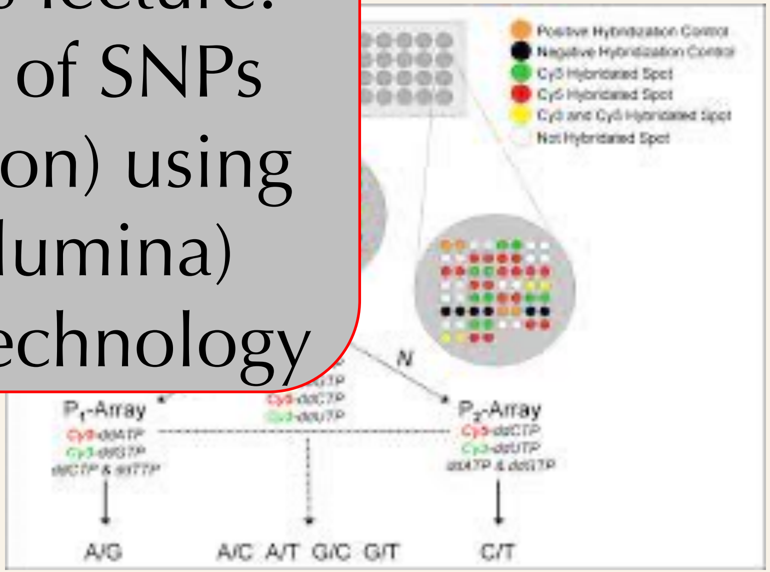


Yet there *are* different ways of assessing genetic variation

i.e. *de novo*
haplotype a
assemb



SNP Chip
typing



Variant calling for purpose of this lecture: The detection of SNPs (genetic variation) using short read (Illumina) resequencing technology

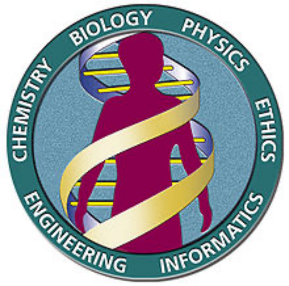
Questions?

HOW? WHEN?
WHO? WHERE?
WHEN? WHY? HOW?
HOW? WHAT? WHO?
WHERE?
WHO? WHERE? HOW? WHERE?
WHAT?
WHY? WHAT?
WHEN?
WHERE? WHO?
HOW? WHAT?
WHO? WHERE?
WHY?
WHO? WHEN?
WHEN? WHAT?
WHO? WHY? HOW?
HOW? WHERE?
WHY?
WHO? WHAT?
HOW?
WHERE?
WHAT? WHEN?
HOW? WHEN?
WHAT? WHO?
WHY? WHERE?
WHEN? HOW?
WHAT? WHEN?
WHERE? HOW?

2) Variant calling pipelines/methods and limitations

2) Variant calling pipelines/methods and limitations

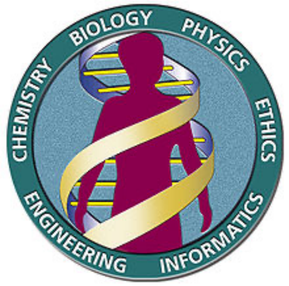
Variant calling **always** starts with a reference genome



Assembly of the first complex vertebrate genome
Human genome assembly project (2003)
Not easily repeated: it was massive task
Nowadays; much cheaper and faster

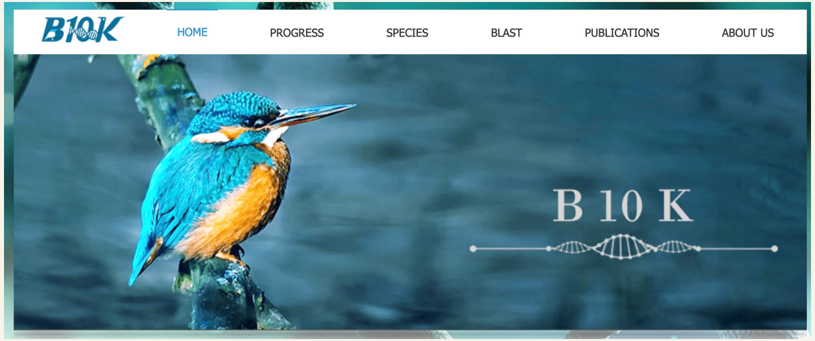
2) Variant calling pipelines/methods and limitations

Variant calling **always** starts with a reference genome



Assembly of the first complex vertebrate genome
Human genome assembly project (2003)
Not easily repeated: it was massive task
Nowadays; much cheaper and faster

Great push to provide reference genomes for many organisms!



B10 K:
10.000 bird genomes

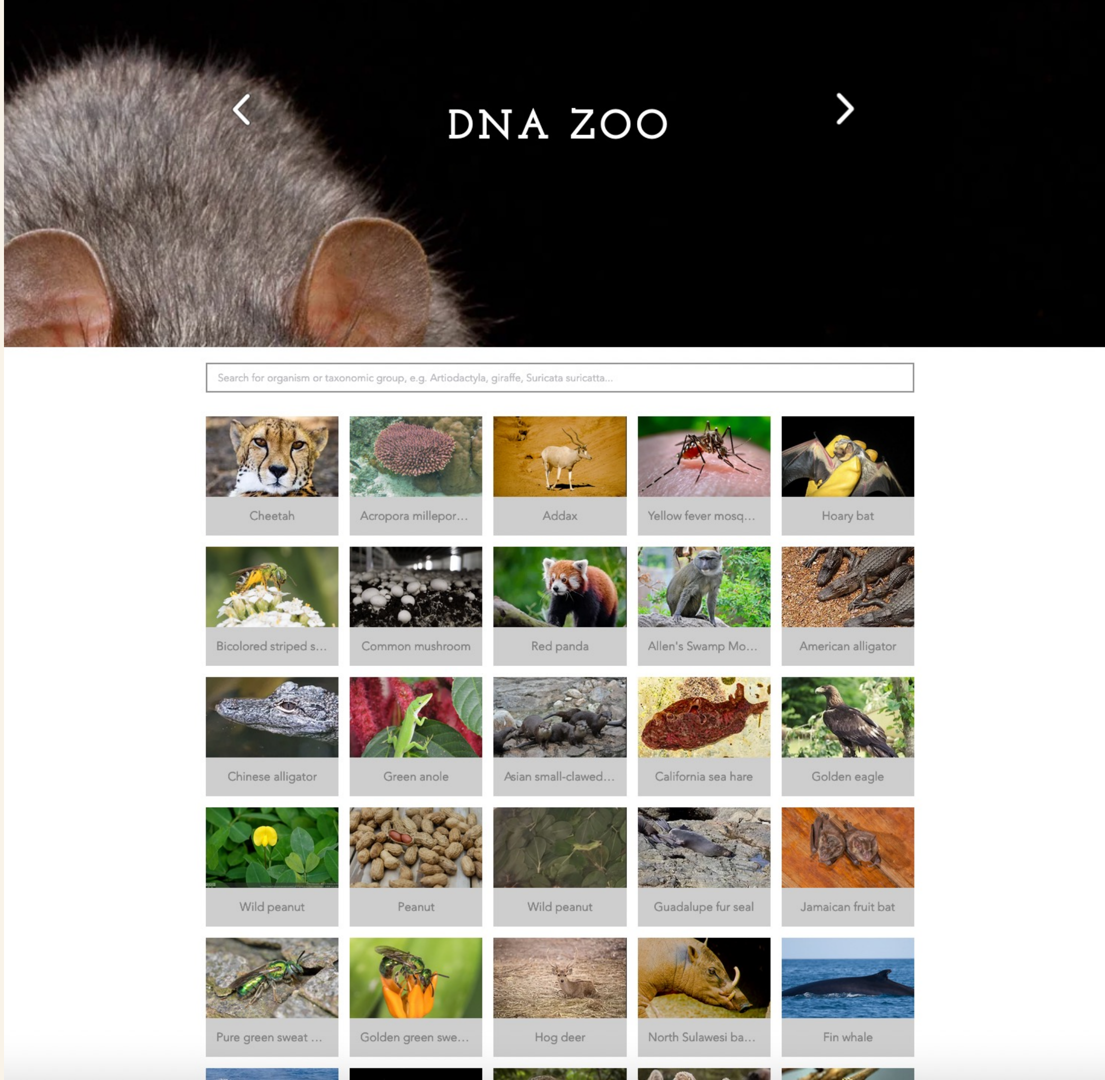
Deep evolutionary understanding of the entire living avian class

https://b10k.genomics.cn/

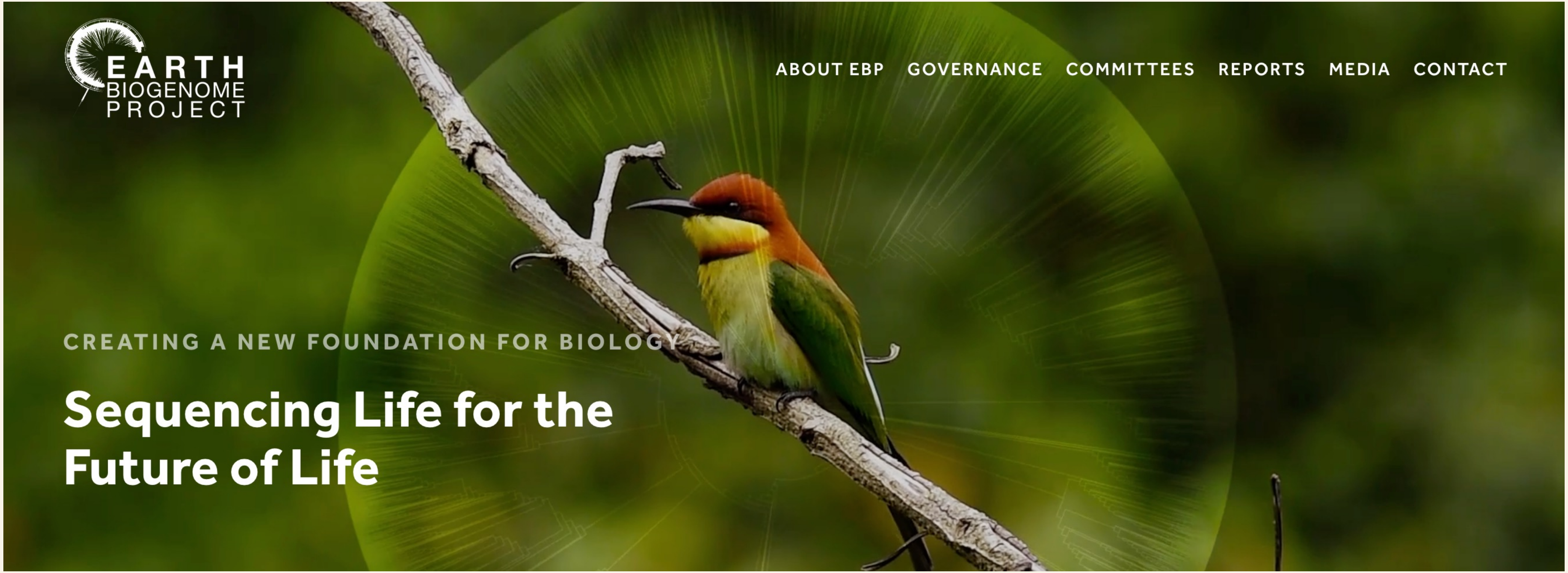
The *DNA Zoo*

facilitates conservation efforts by releasing high-quality genomics resources.

https://www.dnazoo.org



The most ambitious: Earth Biogenome Project



CREATING A NEW FOUNDATION FOR BIOLOGY

Sequencing Life for the Future of Life

https://www.earthbiogenome.org/

The most ambitious: Earth Biogenome Project

EBP: moonshot for biology, aims to characterize the genomes of all of Earth's eukaryotic biodiversity over a period of ten years.

https://www.earthbiogenome.org/

The most ambitious: Earth Biogenome Project

EBP: moonshot for biology, aims to characterize the genomes of all of Earth's eukaryotic biodiversity over a period of ten years.

The vision: to create a new foundation for biology, with new solutions for preserving biodiversity and sustaining human societies.

https://www.earthbiogenome.org/

But what is a reference genome?

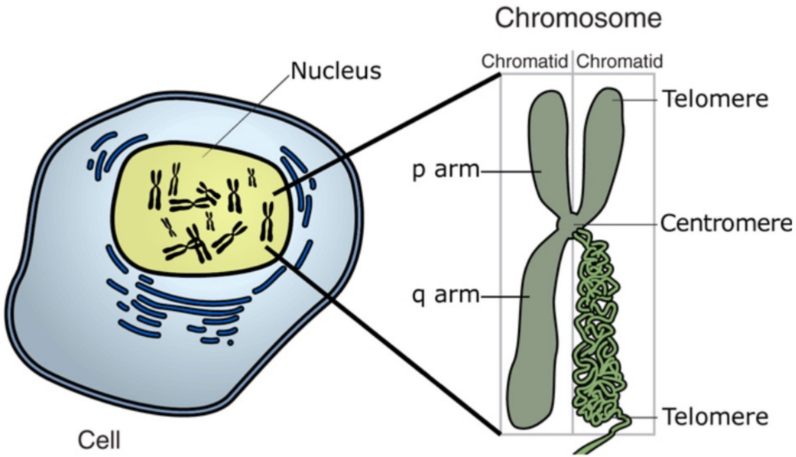
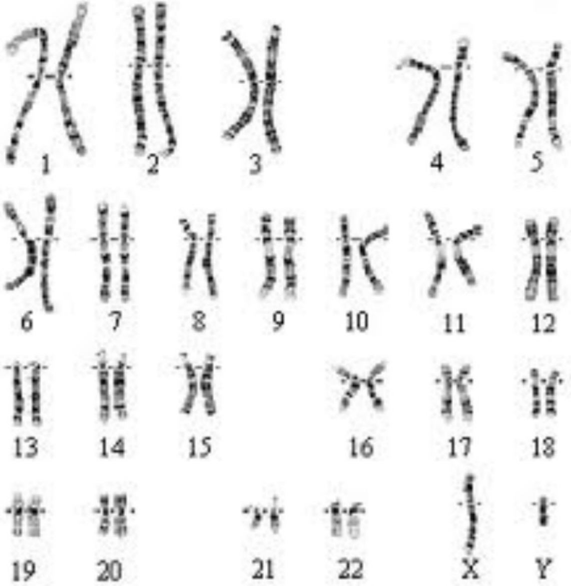
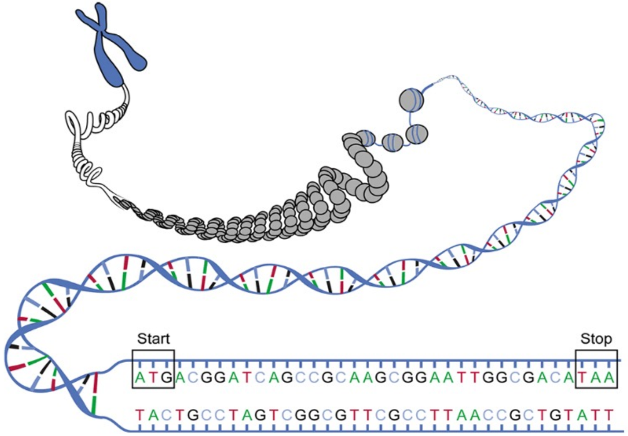


Image adapted from: National Human Genome Research Institute.




Start
Stop
ATGACGGATCAGCCGCAAGCGGAATTGGCGACATAA
TACTGCCTAGTCGGCGTTCGCCTTAACCGCTGTATT

But what is a reference genome?

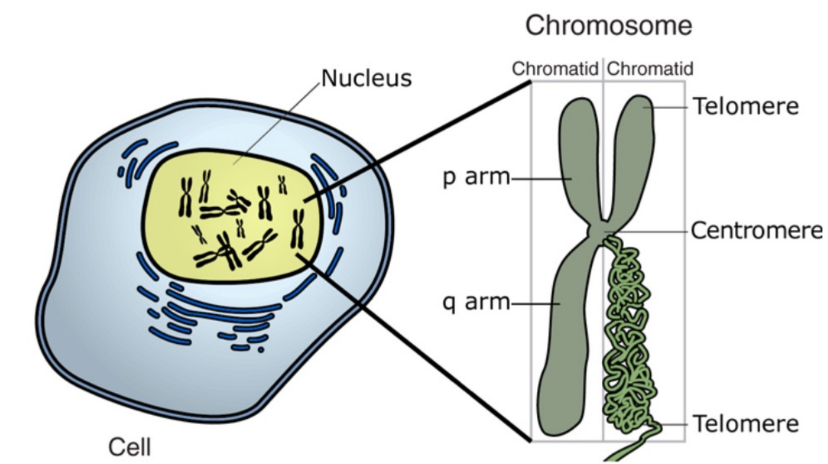
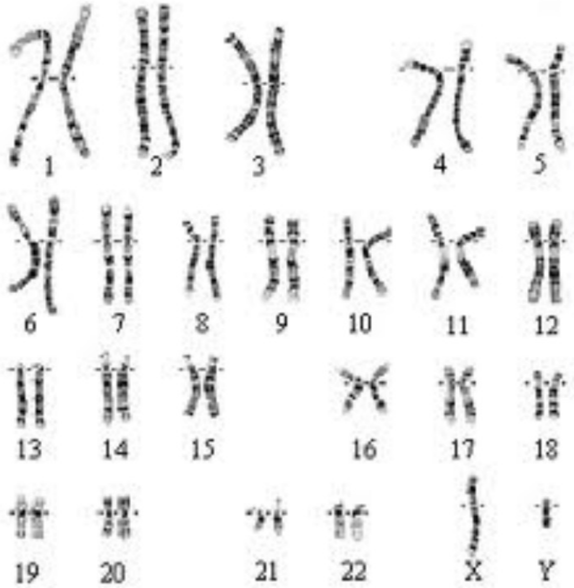
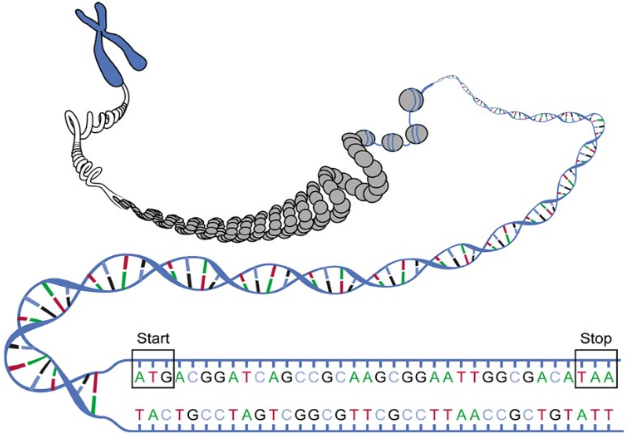


Image adapted from: National Human Genome Research Institute.




Start
Stop
ATGACGGATCAGCCGCAAGCGGAATTGGCGACATAA
TACTGCCTAGTCGGCGTTCGCCTTAACCGCTGTATT

Digital representation / abstraction of a physical, biological phenomenon

A reference genome is …

Usually from a single individual

Result of a genome assembly process -> errors are introduced

Of varying quality, that can vary from organism to organism

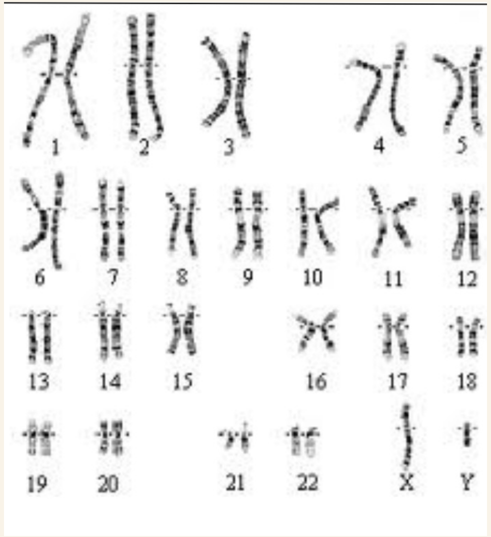
A reference genome is …

Usually from a single individual

Result of a genome assembly process -> errors are introduced

Of varying quality, that can vary from organism to organism

Digital version of the genome



Sequencing and assembly

```
>Chr01
ACTACGTATATAGCATGATCATGCATGATACATGGCTAGT…
>Chr02
ATCATGCATGATACATGGCTAGTACTACGTATATAGCATG…
>Chr03
ATGATCATGCATGATAACTACGTATATAGCCATGGCTAGT…
>CHr04
CGTATATAGCATGATCATGACTACATGATACATGGCTAGT…
…  …
```

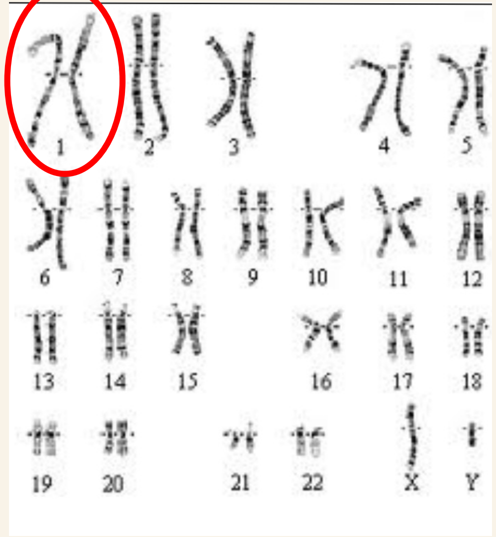
A reference genome is …

Usually from a single individual

Result of a genome assembly process -> errors are introduced

Of varying quality, that can vary from organism to organism

Digital version of the genome



Sequencing and assembly

```
>Chr01
ACTACGTATATAGCATGATCATGCATGATACATGGCTAGT…
>Chr02
ATCATGCATGATACATGGCTAGTACTACGTATATAGCATG…
>Chr03
ATGATCATGCATGATAACTACGTATATAGCCATGGCTAGT…
>CHr04
CGTATATAGCATGATCATGACTACATGATACATGGCTAGT…
…   …
```

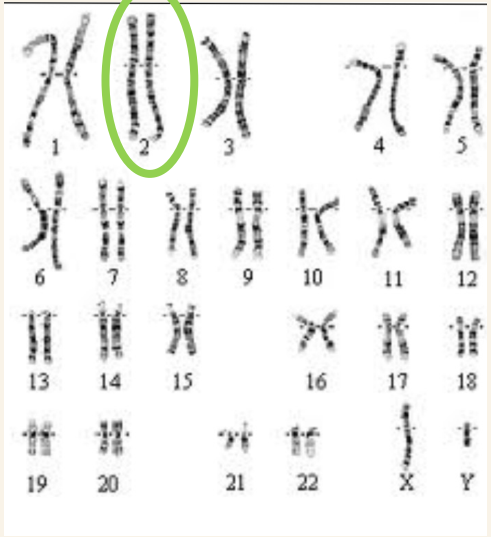
A reference genome is …

Usually from a single individual

Result of a genome assembly process -> errors are introduced

Of varying quality, that can vary from organism to organism

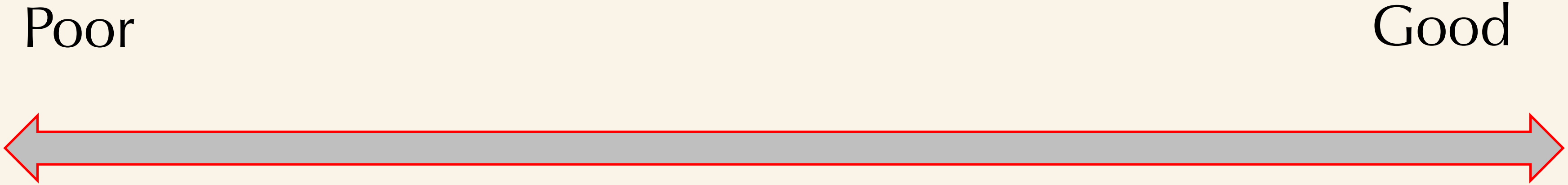
Digital version of the genome



Sequencing and assembly

```
>Chr01
ACTACGTATATAGCATGATCATGCATGATACATGGCTAGT…
>Chr02
ATCATGCATGATACATGGCTAGTACTACGTATATAGCATG…
>Chr03
ATGATCATGCATGATAACTACGTATATAGCCATGGCTAGT…
>CHr04
CGTATATAGCATGATCATGACTACATGATACATGGCTAGT…
… …
```

Quality scale of reference genomes



Poor

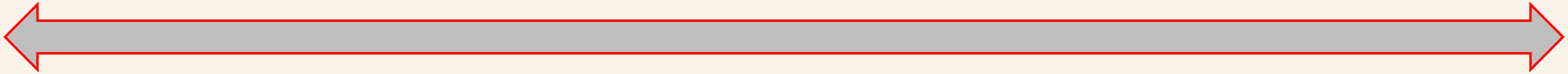
Good

Chromosomes unclear
Thousands of loose fragments
Gaps (*nnnnn*) in sequences
Missing nucleotides

Quality scale of reference genomes

Poor

Good



Chromosomes unclear
Thousands of loose fragments
Gaps (*nnnnn*) in sequences
Missing nucleotides

Chromosomes resolved
Continuous sequences
No gaps
Most nucleotides covered,
including centromeres and
repetitive regions

A reference genome has a 2D coordinate system

```
>Chr01
ACTACGTATATAGCATGATCATGCATGATGATCATGCATGATACATGGCTAGT...
123456789.........
```

→

Millions of nucleotides/bases

Note: some different coordinate systems exist (i.e. starting at 0 or 1)
or using the base or space as “location”

A reference genome has a 2D coordinate system

```
>Chr01
ACTACGTATATAGCATGATCATGCATGATGATCATGCATGATACATGGCTAGT...
123456789.........
```

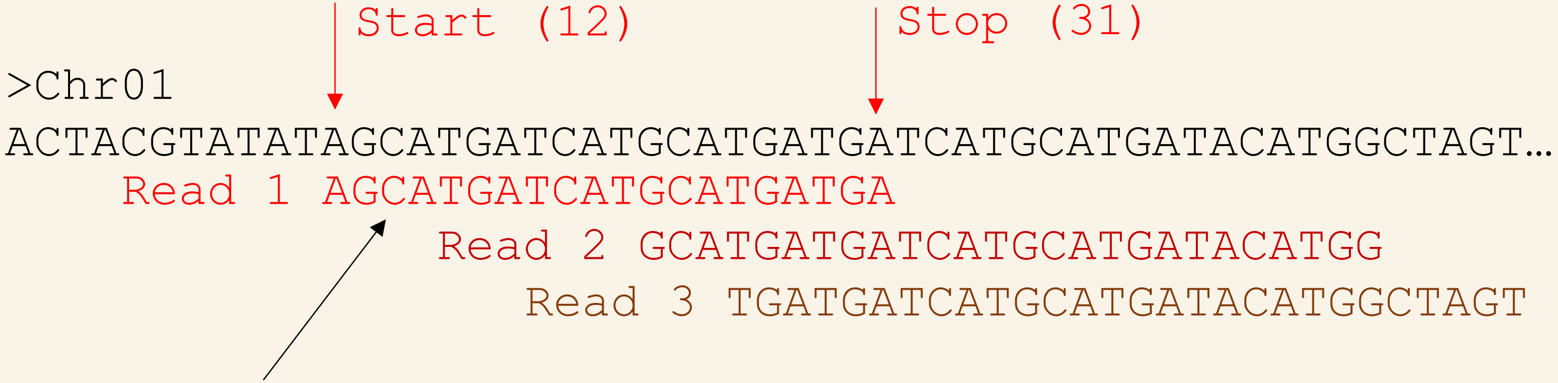
Note: some different coordinate systems exist (i.e. starting at 0 or 1)
or using the base or space as “location”

i.e.

```
A-C-T-A-C-G-T-A
1 2 3 4 5 6 7 8
 1 2 3 4 5 6 7
```

Such different systems are usually automatically recognized by different software

A multiple alignment towards a reference

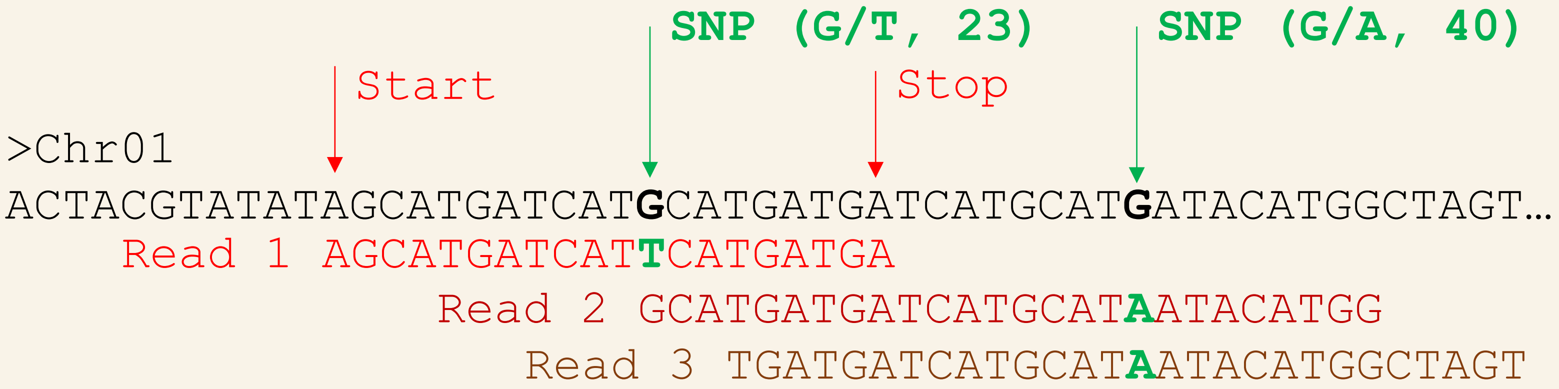


```
            Start (12)               Stop (31)
>Chr01
ACTACGTATATAGCATGATCATGCATGATGATCATGCATGATACATGGCTAGT…
    Read 1 AGCATGATCATGCATGATGA
               Read 2 GCATGATGATCATGCATGATACATGG
                  Read 3 TGATGATCATGCATGATACATGGCTAGT
```

Short read sequencing data is compared to the reference (looking for a ″match″)

We first need such alignment before we can analyse variation

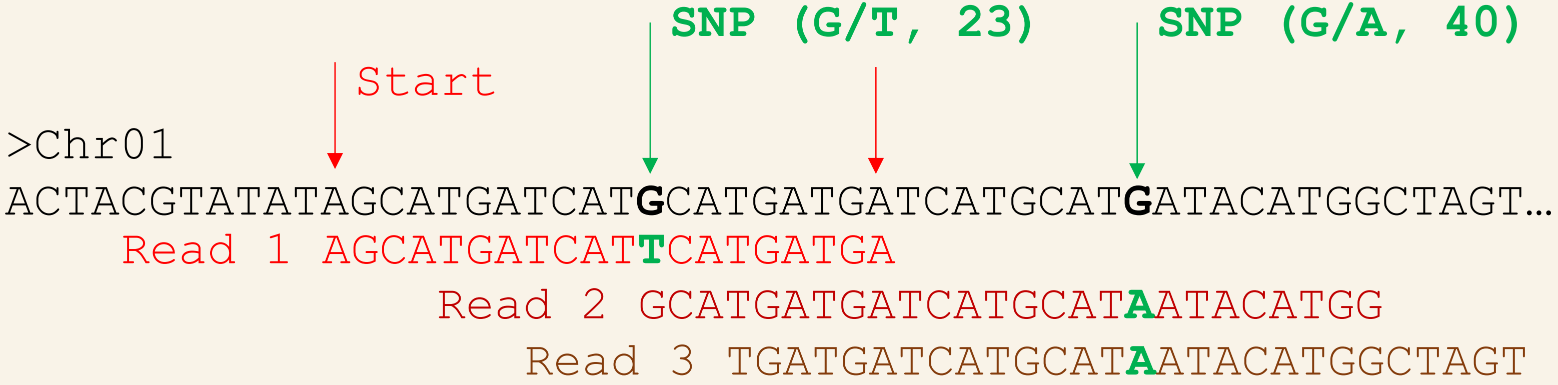
Read variation is analysed within this alignment context



```
>Chr01
ACTACGTATATAGCATGATCATGCATGATGATCATGCATGATACATGGCTAGT…
    Read 1 AGCATGATCATTCATGATGA
               Read 2 GCATGATGATCATGCATAATACATGG
                  Read 3 TGATGATCATGCATAATACATGGCTAGT
```

An accurate alignment is *essential* before we can trust any variant

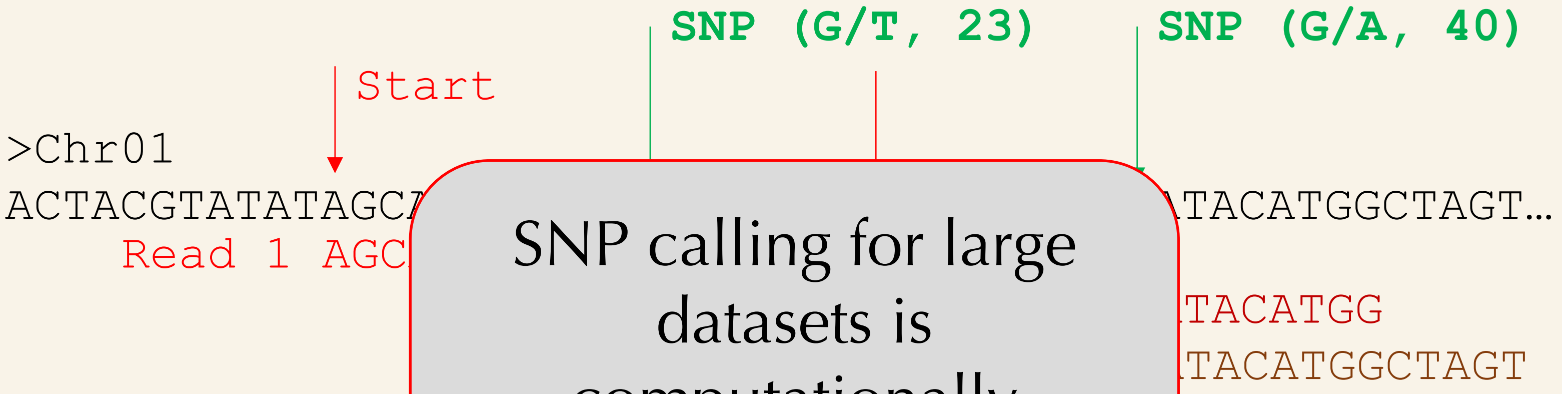
Read variation is analysed within this alignment context



```
                       SNP (G/T, 23)    SNP (G/A, 40)
            Start
>Chr01
ACTACGTATATAGCATGATCATGCATGATGATCATGCATGATACATGGCTAGT...
    Read 1 AGCATGATCATTCATGATGA
               Read 2 GCATGATGATCATGCATAATACATGG
                  Read 3 TGATGATCATGCATAATACATGGCTAGT
```

We usually analyse ***millions to billions*** of reads and compare these to reference genomes that consist of **billions** of nucleotides/bases (Human genome ~3Gb)

Read variation is analysed within this alignment context

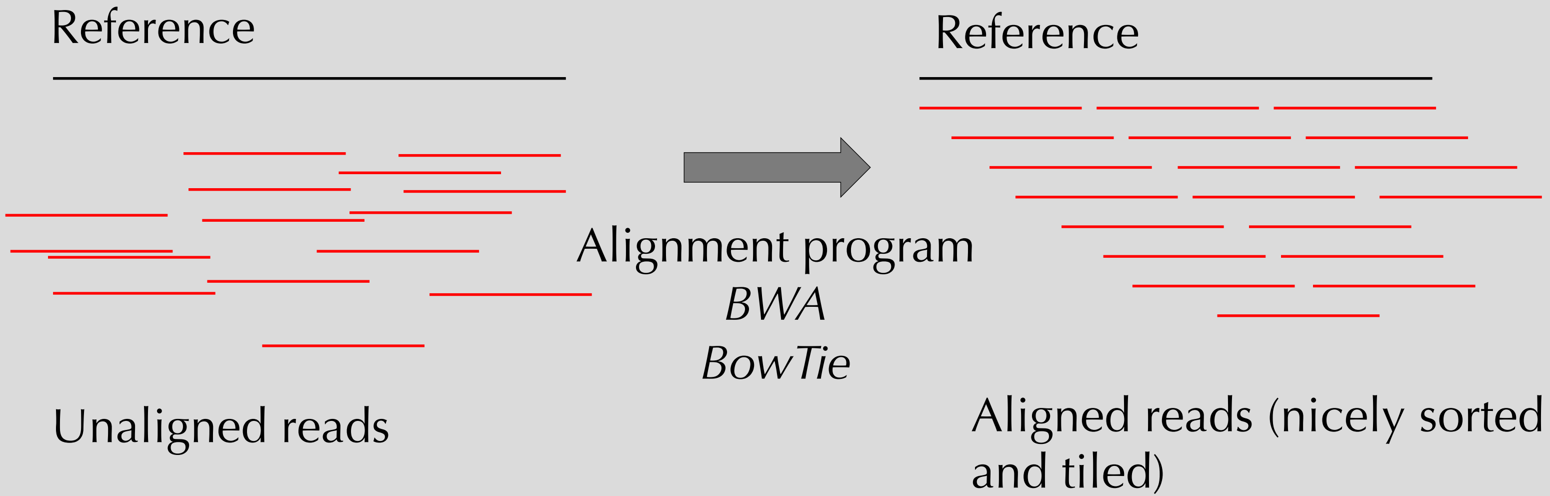


SNP calling for large datasets is computationally intensive -> work on remote HPC clusters

We usually … of reads and compare th… that consist of **billions** of nucleotides/bases (Human genome ~3Gb)

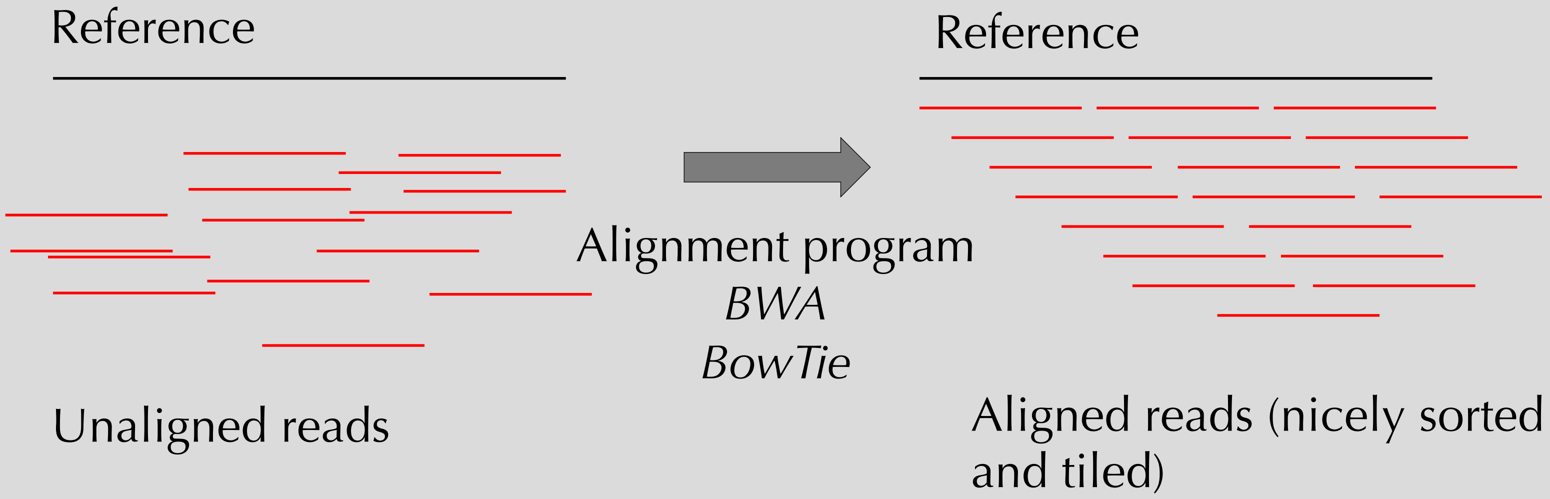
Read variation is analysed within this alignment context

Incredibly efficient software has been designed to take care of this task!



Read variation is analysed within this alignment context

Incredibly efficient software has been designed to take care of this task!



Reference

Unaligned reads

Alignment program

BWA

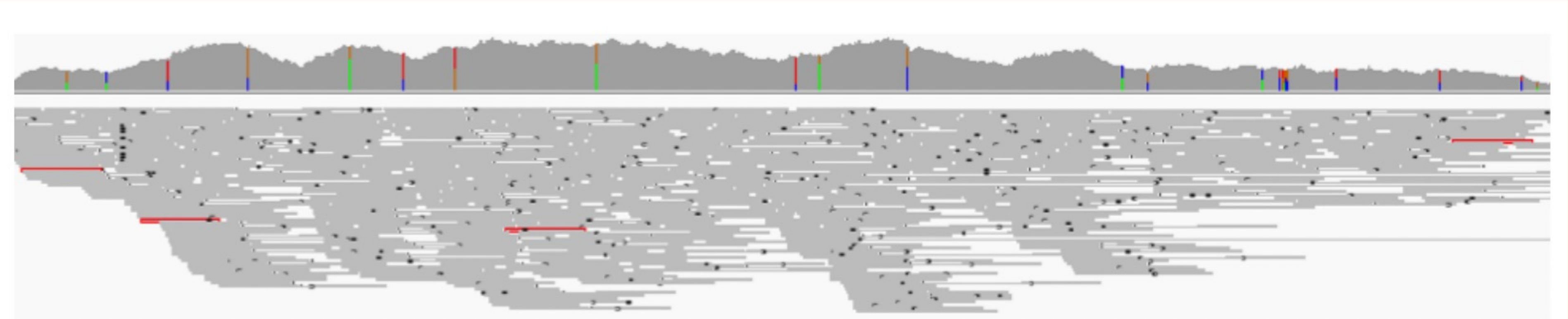
BowTie

Reference

Aligned reads (nicely sorted and tiled)

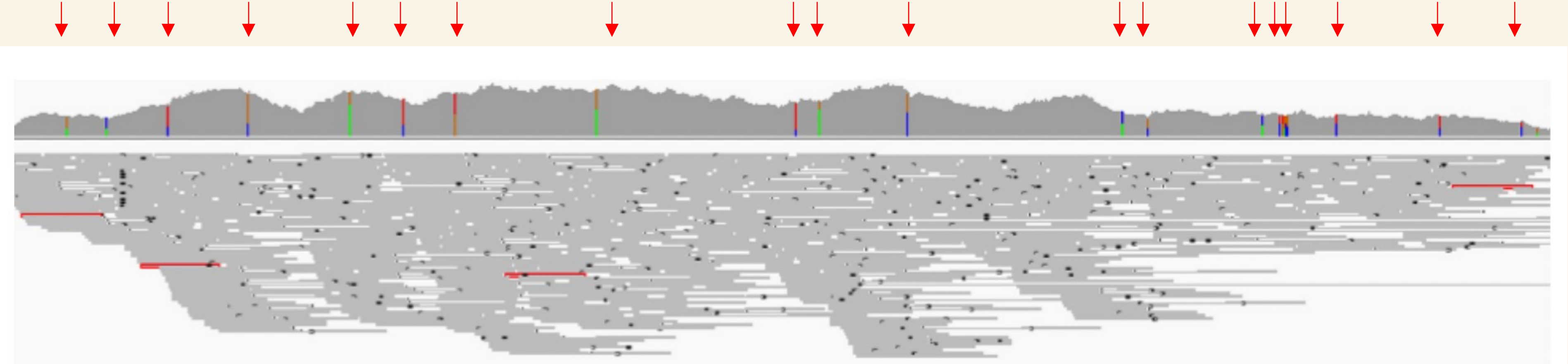
Standard program settings are usually sufficient

Visualisation of thousands of reads



Visualisation of thousands of reads

Genetic variation (SNPs) colours reflect which bases are variable (A-C, A-T, G-C, etc)



After aligning, we need another program to determine which bases are variable:

A SNP caller

SNP calling programs

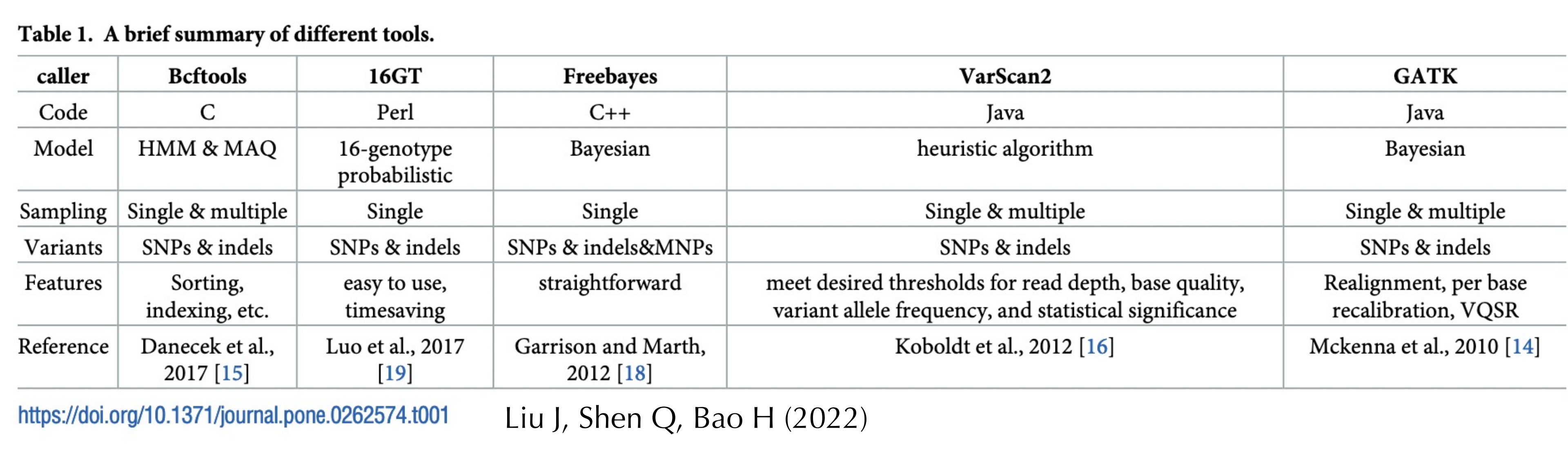


Table 1. A brief summary of different tools.

caller	Bcftools	16GT	Freebayes	VarScan2	GATK
Code	C	Perl	C++	Java	Java
Model	HMM & MAQ	16-genotype probabilistic	Bayesian	heuristic algorithm	Bayesian
Sampling	Single & multiple	Single	Single	Single & multiple	Single & multiple
Variants	SNPs & indels	SNPs & indels	SNPs & indels&MNPs	SNPs & indels	SNPs & indels
Features	Sorting, indexing, etc.	easy to use, timesaving	straightforward	meet desired thresholds for read depth, base quality, variant allele frequency, and statistical significance	Realignment, per base recalibration, VQSR
Reference	Danecek et al., 2017 [15]	Luo et al., 2017 [19]	Garrison and Marth, 2012 [18]	Koboldt et al., 2012 [16]	Mckenna et al., 2010 [14]

https://doi.org/10.1371/journal.pone.0262574.t001

Liu J, Shen Q, Bao H (2022)

Many programs exist, and there is *continuous* development
For instance Bcftools/16GT are now recommended
Yet use of GATK is wide-spread (oldest, developed by Broad institute, good documentation)

What does a variant caller do?

Aims to provide statistical confidence in observing TRUE genetic variation

```
>Chr01
ACTACGTATATAGCATGATCAT**G**CATGATGATCATGCATGATACATGGCTAGT…
    Read 1 AGCATGATCAT**T**CATGATGA
```

Is this real or not?

What does a variant caller do?

Aims to provide statistical confidence in observing TRUE genetic variation

```
>Chr01
ACTACGTATATAGCATGATCATGCATGATGATCATGCATGATACATGGCTAGT…
    Read 1 AGCATGATCATTCATGATGA
```

Is this real or not?

Sequencing data (as any type of data) comes with errors (wrong bases called) and/or uncertainty (low quality of bases) in the call

Solution? Generate LOTS more data!

What does a variant caller do?

With more data (read), more certainty is obtained: ***fold coverage***

```
>Chr01
ACTACGTATATAGCATGATCATGCATGATGATCATGCATGATACATGGCTAGT…
    Read 1 AGCATGATCATTCATGATGA
       Read 2 ATGATCATTCATGATGATCAT
         Read 3 GATCATTCATGATGATCATGCATGAT
           Read 4 TCATTCATGATGATCATGCAT
            Read 5 CATTCATGATGATCATGCATGATACATGG
```

5-fold coverage, all the same, we are pretty certain about this call (note: we usually strive for ~20 fold coverage)

What does a variant caller do?

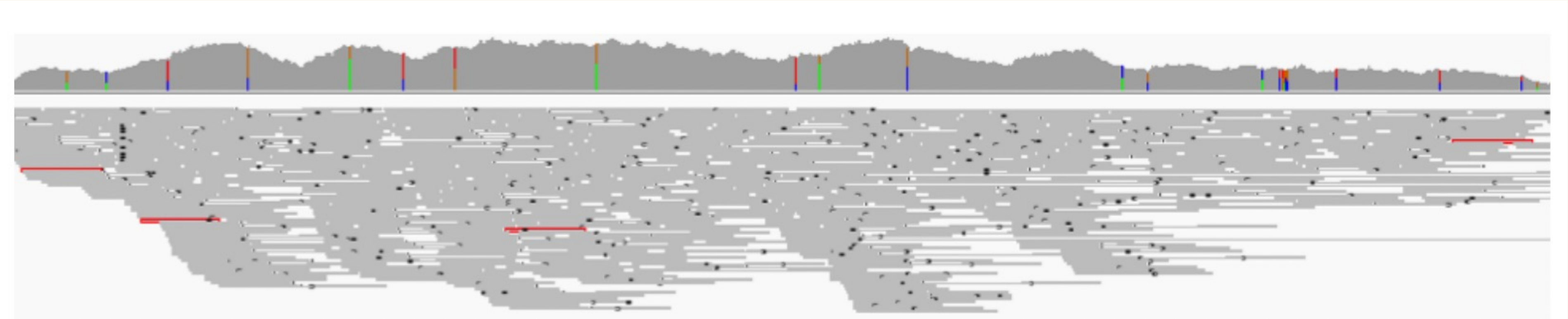
Another example

```
>Chr01
ACTACGTATATAGCATGATCATGCATGATGATCATGCATGATACATGGCTAGT…
    Read 1 AGCATGATCATTCATGATGA
       Read 2 ATGATCATTCATGATGATCAT
         Read 3 GATCATTCATGATGATCATGCATGAT
           Read 4 TCATACATGATGATCATGCAT
            Read 5 CATTCATGATGATCATGCATGATACATGG
```

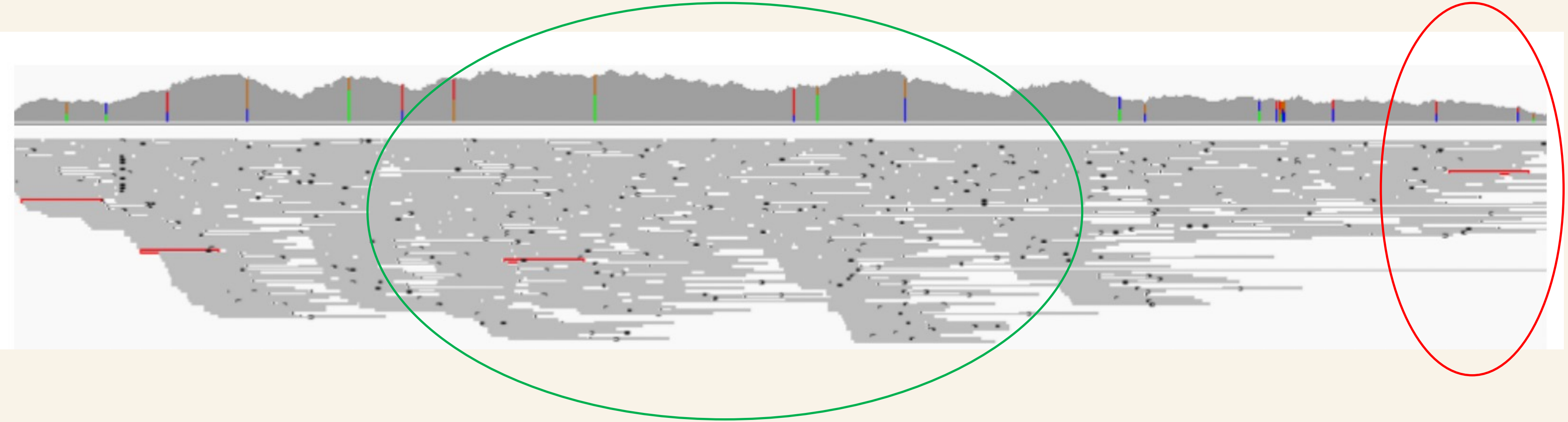
We cannot be so certain about the A, until we get more data

Coverage is the most important determinant for the quality of your data

Yet along a reference, you'll obtain variable coverage due to random processes, assembly quality, or genomic complexity



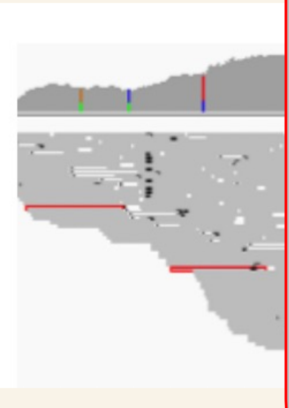
Yet along a reference, you'll obtain variable coverage due to random processes, assembly quality, or genomic complexity



Higher coverage

Lower coverage

Yet along a reference, you'll obtain variable coverage due to random plexity



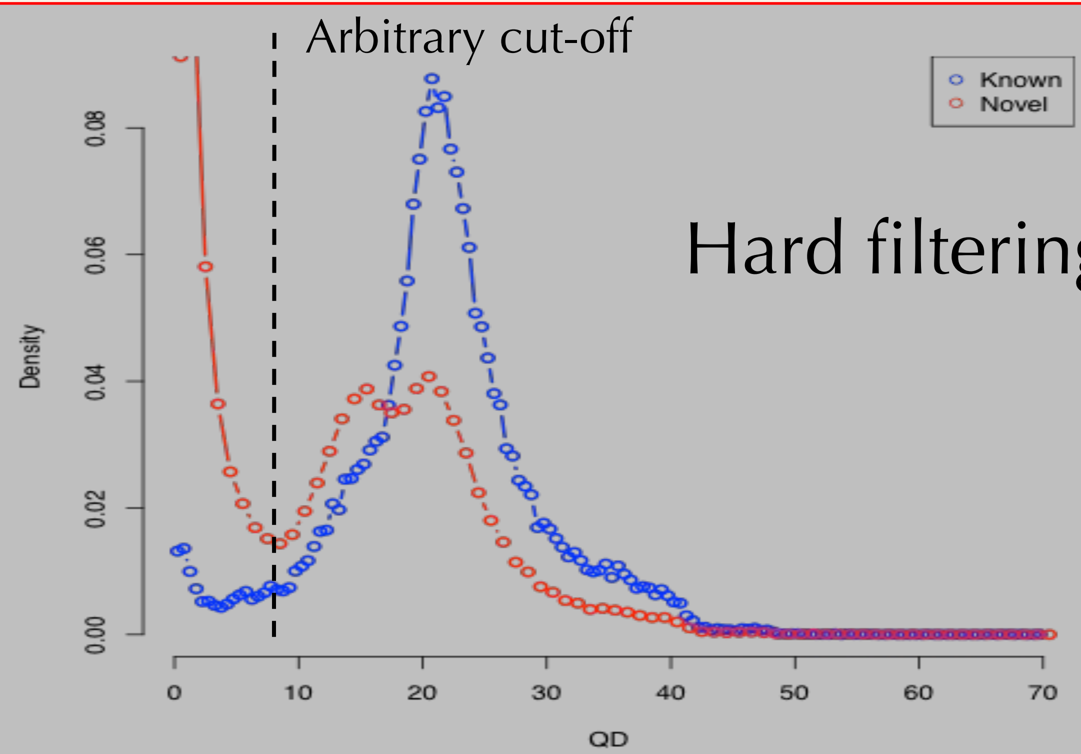
SNP callers run complex statistical models (e.g. Bayesian or HMM models) to provide confidence in SNP calls and if they are "TRUE". They often assume correct read alignment **and** require sufficient read coverage in order to provide high-quality calls

Higher coverage

Lower coverage

SNP callers will ALSO yield a large numbers of SNPs of which many will NOT be true (false positives)

We need to ***filter*** our data to only retain the high quality part of the data

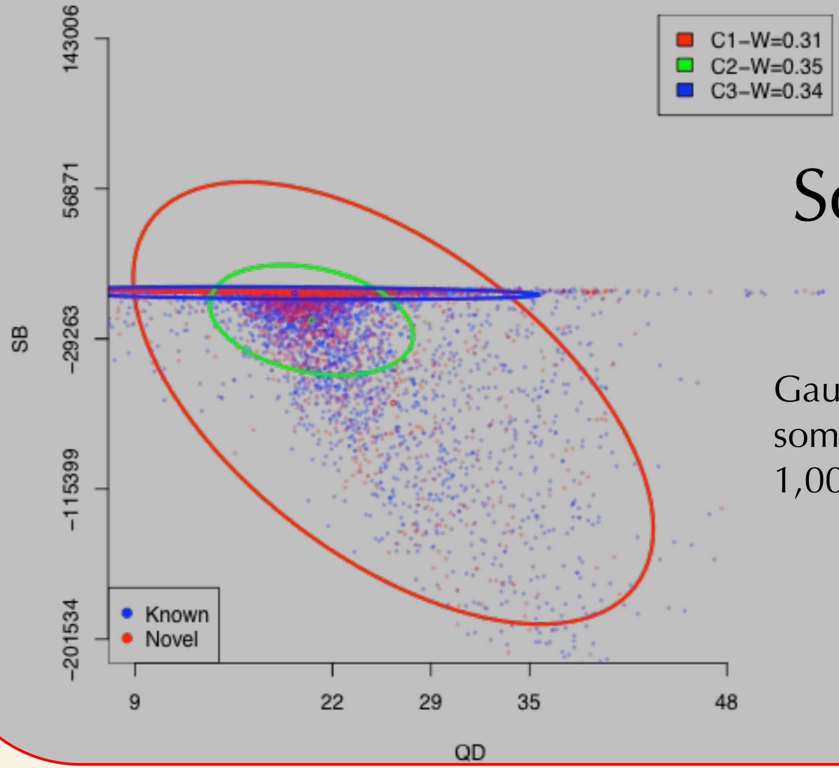


Hard filtering

Mark de Pristo 2010

SNP callers will ALSO yield a large numbers of SNPs of which many will NOT be true (false positives)

We need to ***filter*** our data to only retain the high quality part of the data

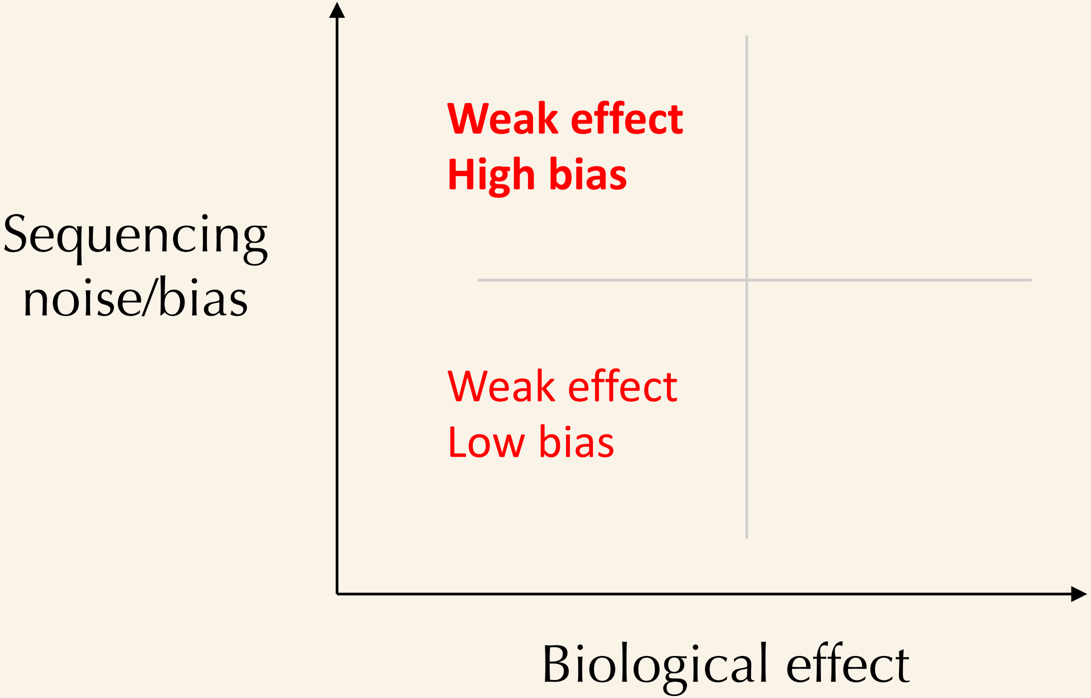


Soft filtering

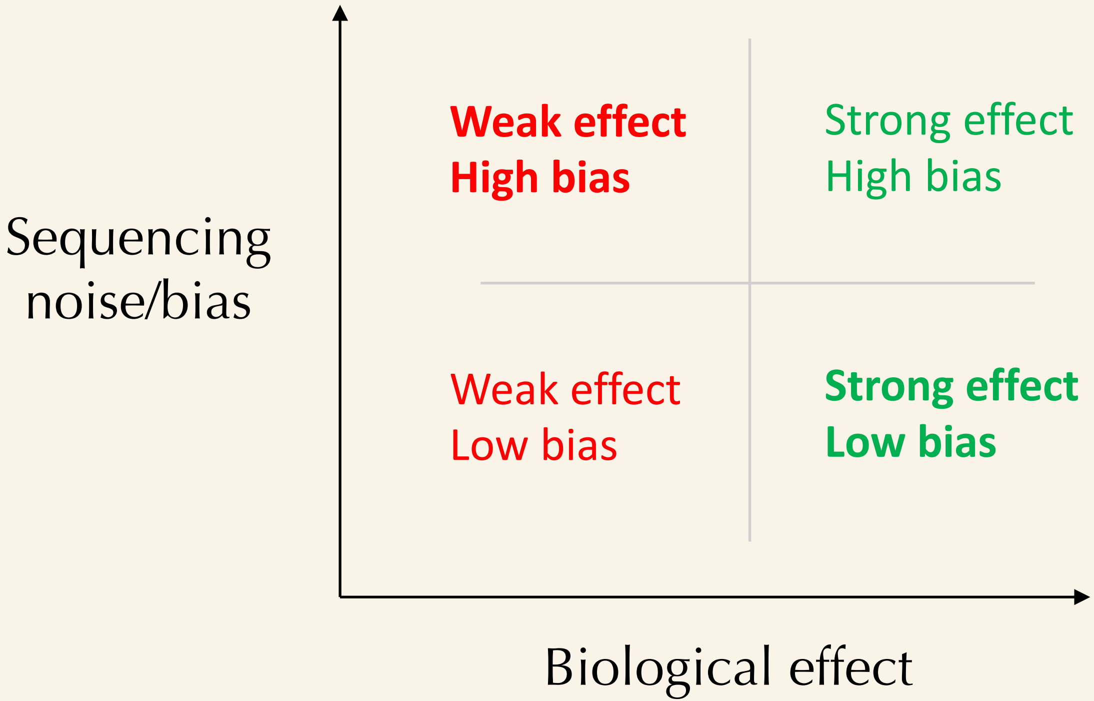
Gaussian Mixture Model fit to some training data (e.g. dbSNP, 1,000 Genomes, etc)

Mark de Pristo 2010

Yet there is no “fixed” approach to filtering your data



Yet there is no “fixed” approach to filtering your data



Sequencing noise/bias

Weak effect
High bias

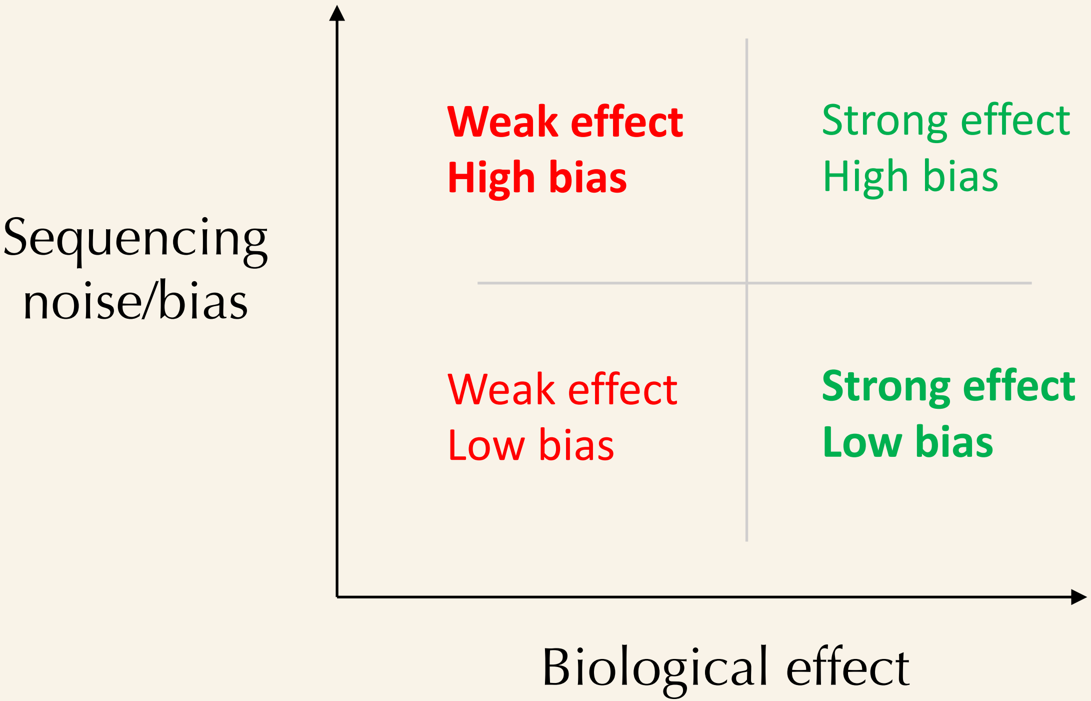
Strong effect
High bias

Weak effect
Low bias

Strong effect
Low bias

Biological effect

Yet there is no “fixed” approach to filtering your data



Sequencing noise/bias

Weak effect
High bias

Strong effect
High bias

Weak effect
Low bias

Strong effect
Low bias

Biological effect

It is not always clear from the outset where you are! You need to explore your data and use preliminary analyses

Questions?

HOW? WHEN?
WHO? WHERE?
WHEN? WHY? WHAT?
WHERE? HOW?
HOW? WHAT? WHO?
WHERE?
WHO? WHERE? HOW?
HOW? WHERE?
WHAT?
WHO?
WHY? WHAT?
WHERE?
WHAT? WHEN?
WHERE? WHO?
HOW? WHAT?
WHO? WHERE?
WHY? WHAT? HOW?
WHO? WHEN?
WHERE?
WHEN? WHAT?
WHO? WHY? HOW?
HOW? WHERE?
WHAT?
WHEN? WHY?
WHO? WHAT?
HOW?
WHO? WHY? WHERE?
WHAT? WHEN?
WHO? WHAT?
WHERE? WHY?
HOW? WHEN?
WHAT? WHO?
WHEN?
WHY? WHERE?
WHEN? HOW?
WHAT? WHEN?
WHERE? HOW?
WHEN? WHO?

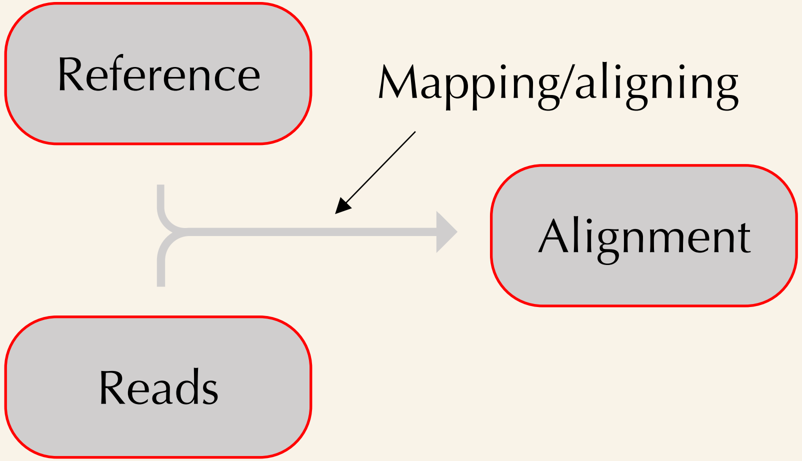
After all this, what does a variant calling pipeline look like?

Reference

Reads

e.g.
population
data

After all this, what does a variant calling pipeline look like?



Reference

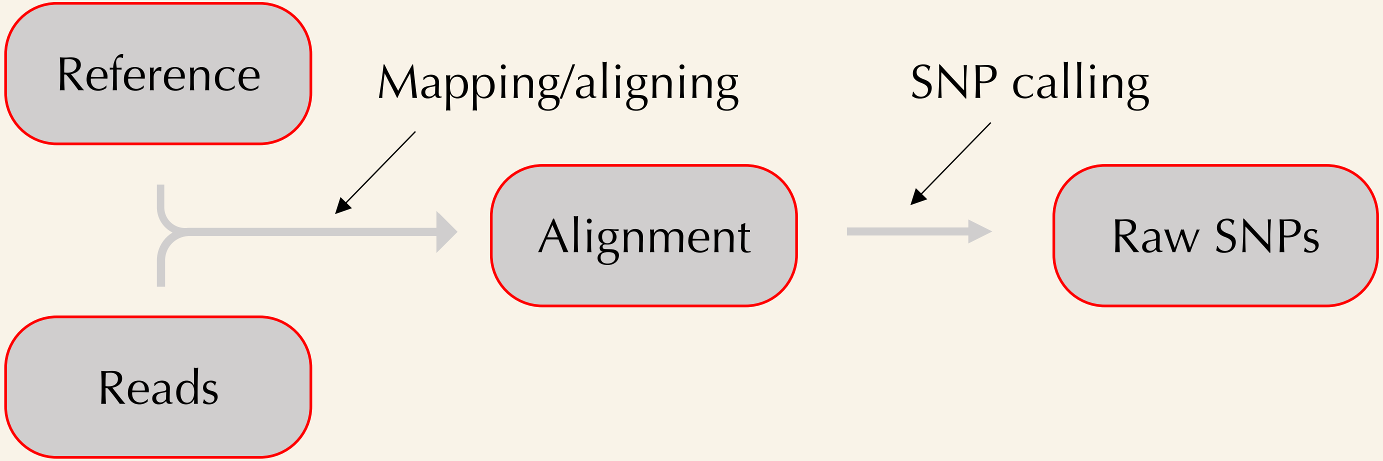
Mapping/aligning

Alignment

Reads

e.g.
population
data

After all this, what does a variant calling pipeline look like?



Reference

Mapping/aligning

SNP calling

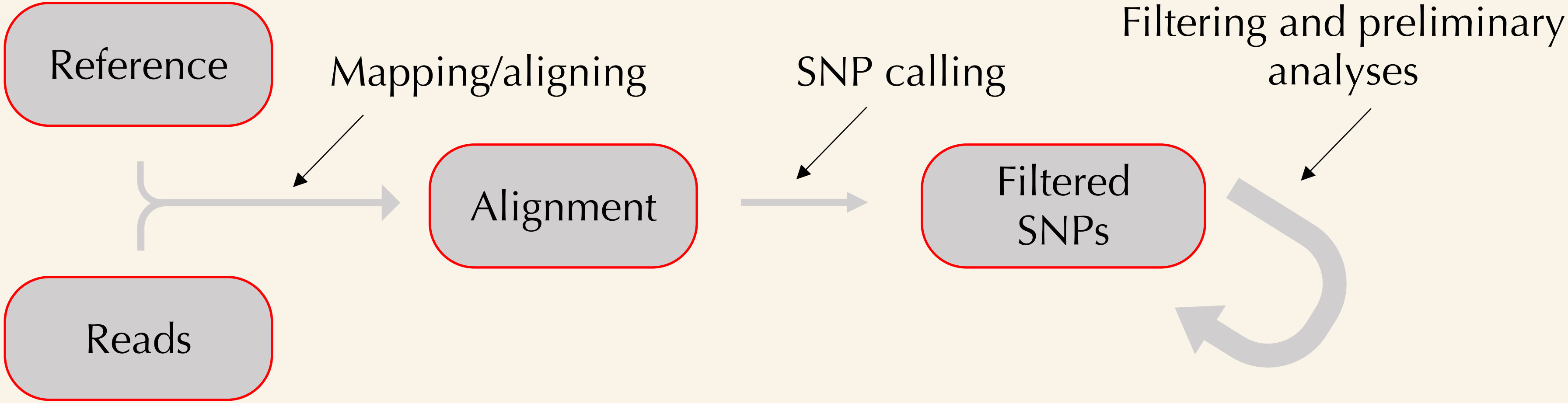
Alignment

Raw SNPs

Reads

e.g.
population
data

After all this, what does a variant calling pipeline look like?



Reference

Mapping/aligning

SNP calling

Filtering and preliminary analyses

Alignment

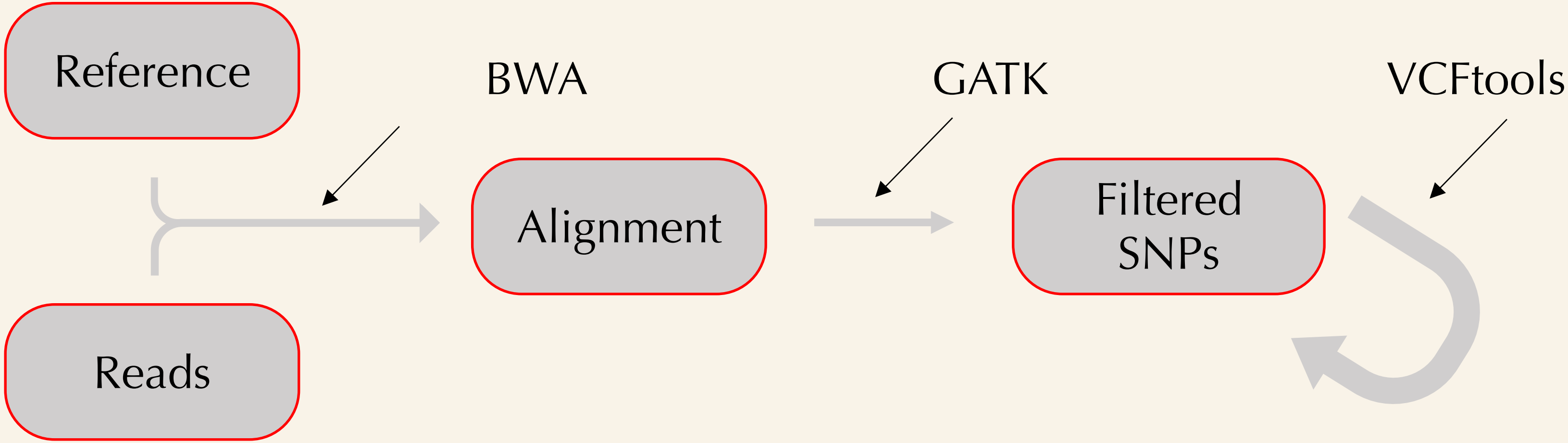
Filtered SNPs

Reads

e.g.
population
data

After all this, what does a variant calling pipeline look like?

A selection of programs that can be used



Reference

Reads

BWA

Alignment

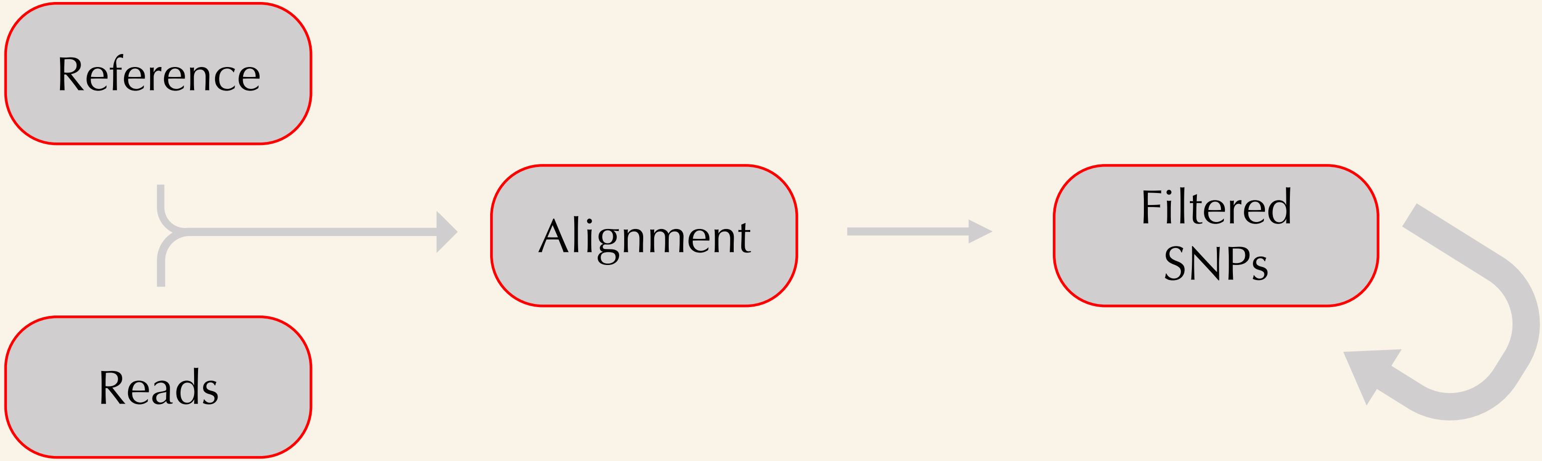
GATK

Filtered SNPs

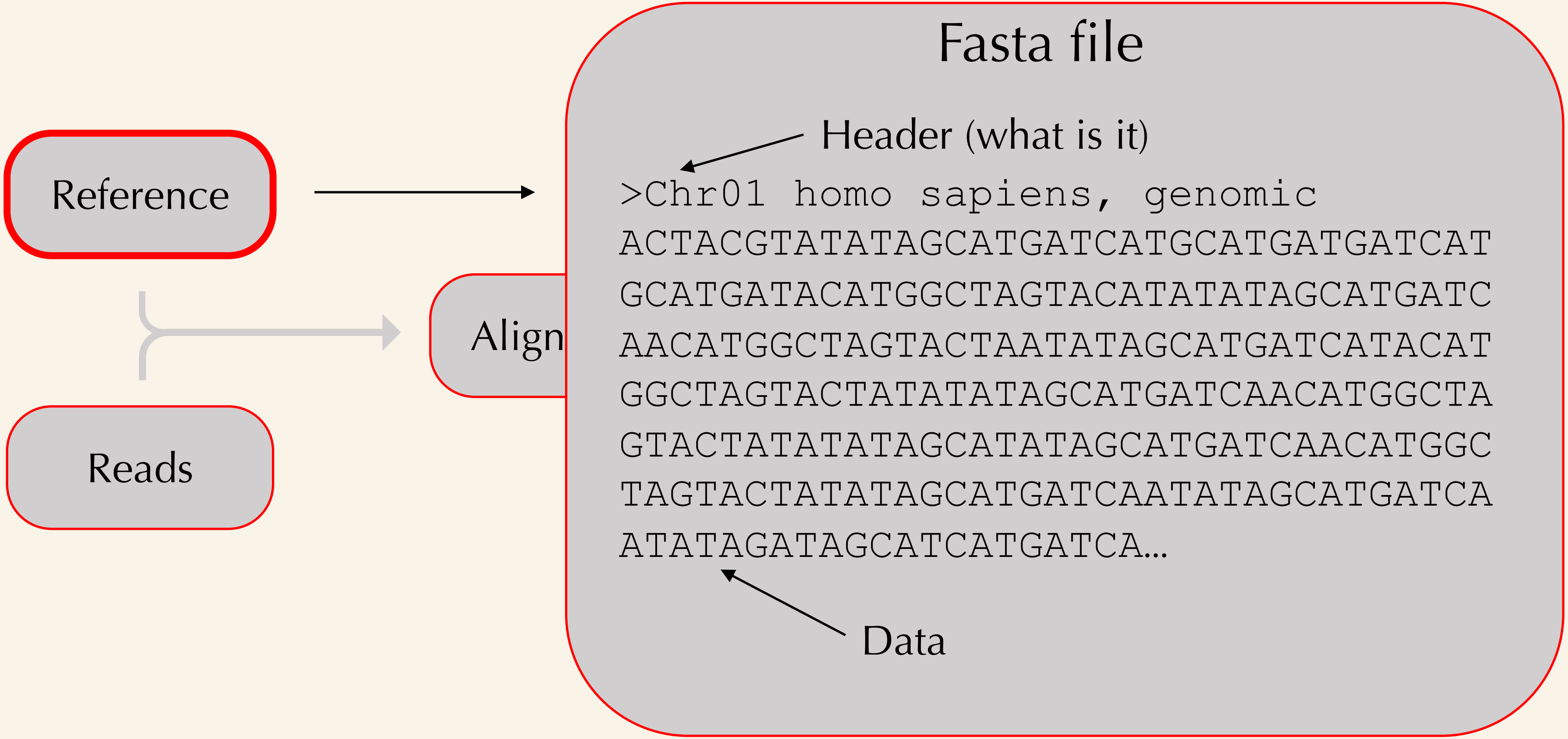
VCFtools

e.g.
population
data

Each of these steps requires specific files to work with!



Each of these steps requires specific files to work with!



Reference

Reads

Align

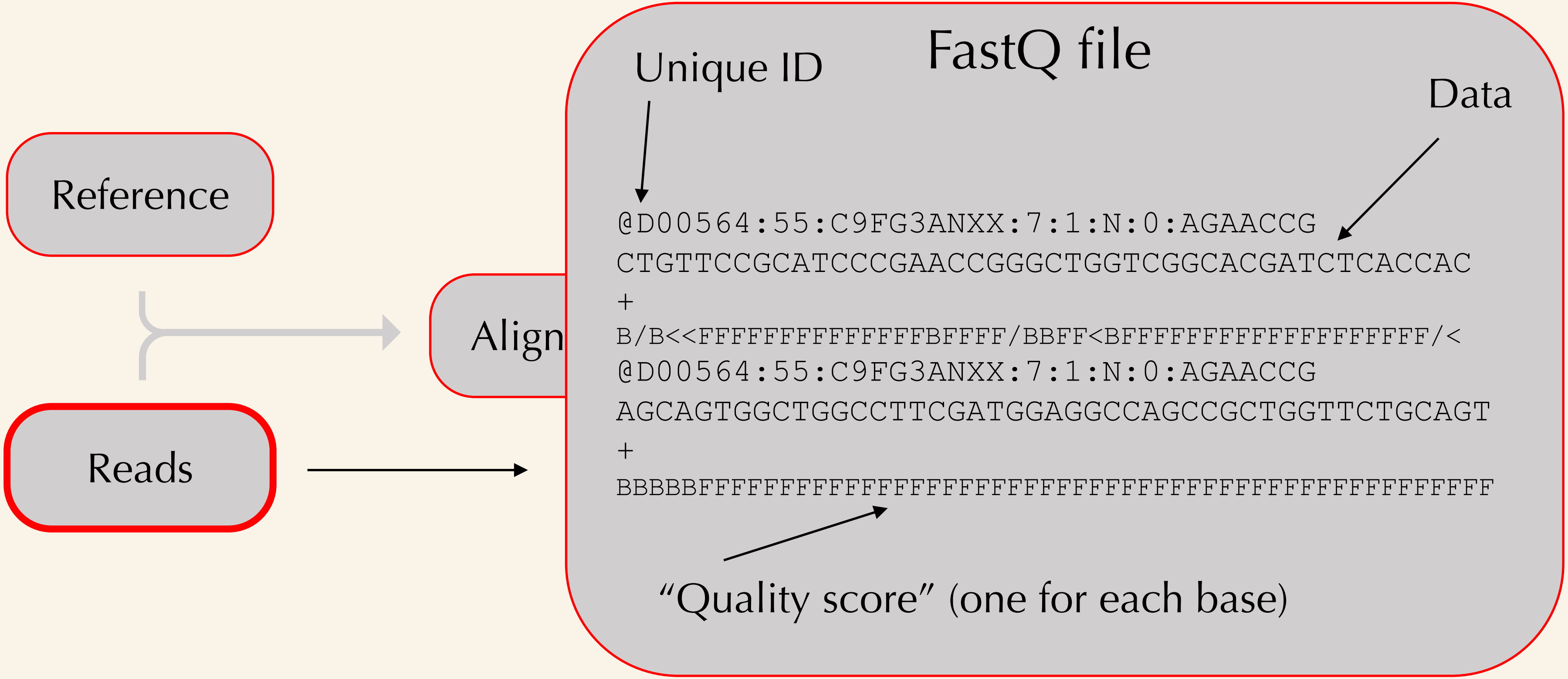
Fasta file

Header (what is it)

```
>Chr01 homo sapiens, genomic
ACTACGTATATAGCATGATCATGCATGATGATCAT
GCATGATACATGGCTAGTACATATATAGCATGATC
AACATGGCTAGTACTAATATAGCATGATCATACAT
GGCTAGTACTATATATAGCATGATCAACATGGCTA
GTACTATATATAGCATATAGCATGATCAACATGGC
TAGTACTATATAGCATGATCAATATAGCATGATCA
ATATAGATAGCATCATGATCA...
```

Data

Each of these steps requires specific files to work with!



FastQ file

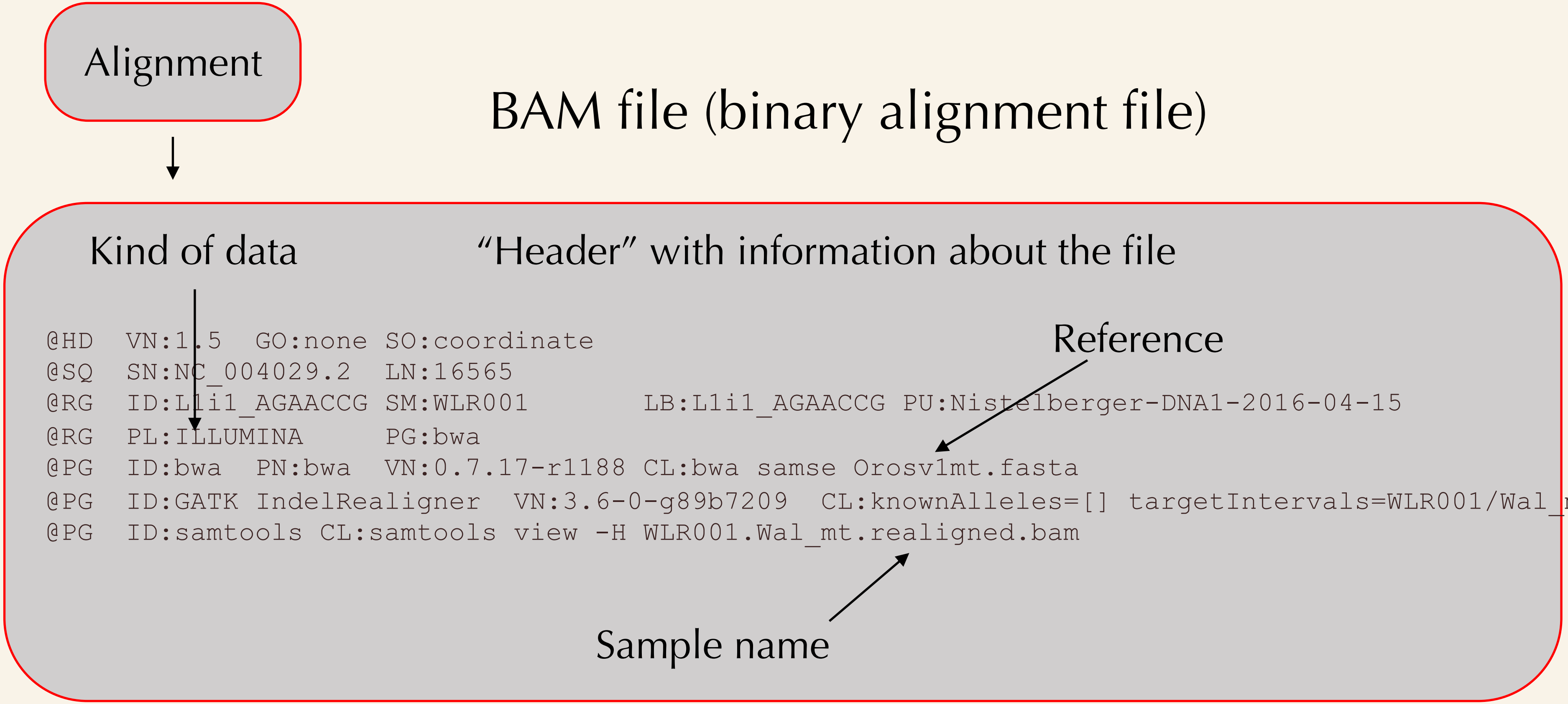
Unique ID

Data

```
@D00564:55:C9FG3ANXX:7:1:N:0:AGAACCG
CTGTTCCGCATCCCGAACCGGGCTGGTCGGCACGATCTCACCAC
+
B/B<<FFFFFFFFFFFFFFBFFFF/BBFF<BFFFFFFFFFFFFFFFFFFF/<
@D00564:55:C9FG3ANXX:7:1:N:0:AGAACCG
AGCAGTGGCTGGCCTTCGATGGAGGCCAGCCGCTGGTTCTGCAGT
+
BBBBBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
```

“Quality score” (one for each base)

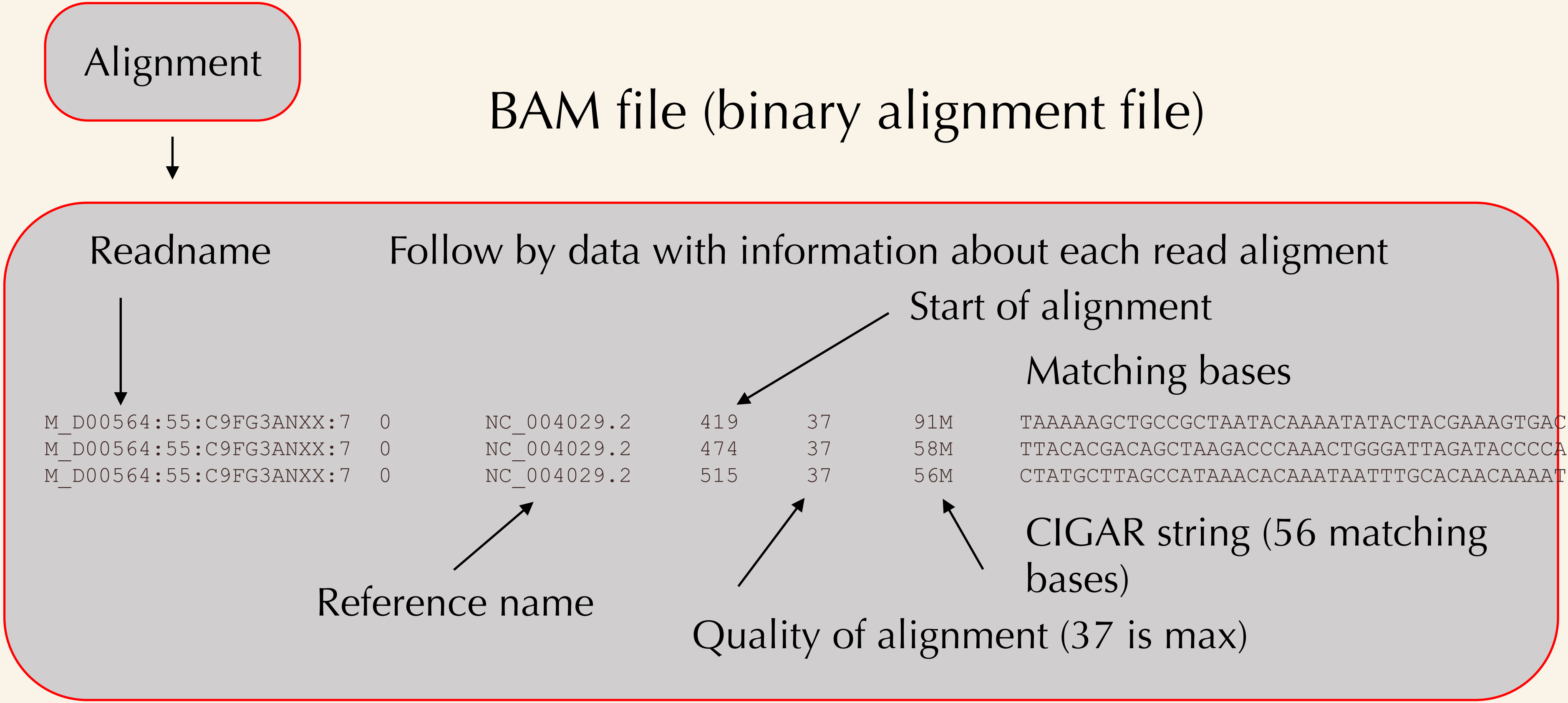
Each of these steps requires specific files to work with!



BAM file (binary alignment file)

```
@HD   VN:1.5  GO:none SO:coordinate
@SQ   SN:NC_004029.2  LN:16565
@RG   ID:L1i1_AGAACCG SM:WLR001       LB:L1i1_AGAACCG PU:Nistelberger-DNA1-2016-04-15
@RG   PL:ILLUMINA     PG:bwa
@PG   ID:bwa  PN:bwa  VN:0.7.17-r1188 CL:bwa samse Orosv1mt.fasta
@PG   ID:GATK IndelRealigner  VN:3.6-0-g89b7209  CL:knownAlleles=[] targetIntervals=WLR001/Wal_m
@PG   ID:samtools CL:samtools view -H WLR001.Wal_mt.realigned.bam
```

Each of these steps requires specific files to work with!



Alignment

BAM file (binary alignment file)

Readname

Follow by data with information about each read aligment

Start of alignment

Matching bases

```
M_D00564:55:C9FG3ANXX:7  0        NC_004029.2     419      37       91M      TAAAAAGCTGCCGCTAATACAAAATATACTACGAAAGTGACT
M_D00564:55:C9FG3ANXX:7  0        NC_004029.2     474      37       58M      TTACACGACAGCTAAGACCCAAACTGGGATTAGATACCCCAC
M_D00564:55:C9FG3ANXX:7  0        NC_004029.2     515      37       56M      CTATGCTTAGCCATAAACACAAATAATTTGCACAACAAAATT
```

Reference name

Quality of alignment (37 is max)

CIGAR string (56 matching bases)

Each of these steps requires specific files to work with!

SNP data

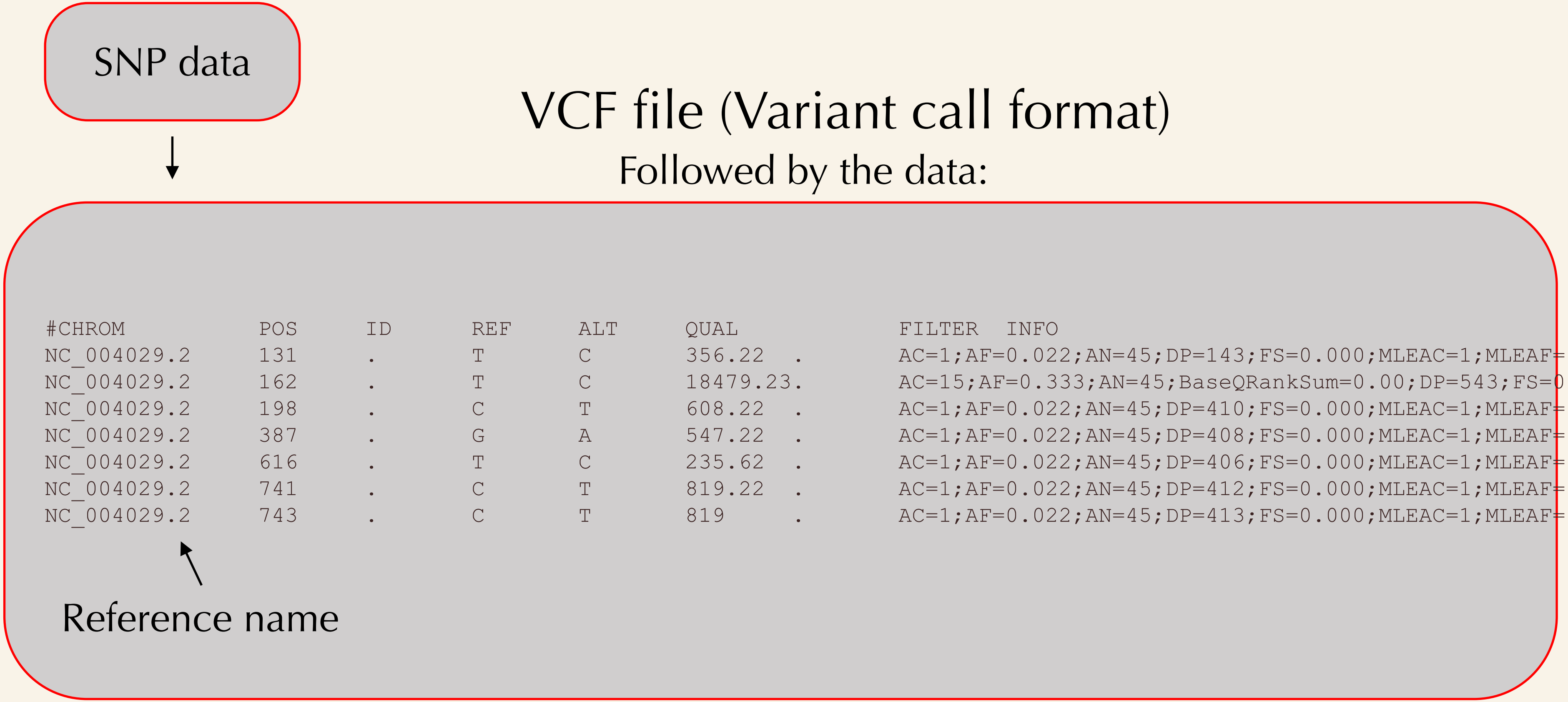
↓

VCF file (Variant call format)

Again, a “Header” with lots of information about the file

```
##fileformat=VCFv4.2
##ALT=<ID=NON_REF,Description="Represents any possible alternative allele not already represented at this location
##FILTER=<ID=LowQual,Description="Low quality">
##FILTER=<ID=PASS,Description="All filters passed">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth (reads with MQ=255 or with bad mates are
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=MIN_DP,Number=1,Type=Integer,Description="Minimum DP observed within the GVCF block">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="Normalized, Phred-scaled likelihoods for genotypes as defined in
##FORMAT=<ID=RGQ,Number=1,Type=Integer,Description="Unconditional reference genotype confidence, encoded as a phred
##FORMAT=<ID=SB,Number=4,Type=Integer,Description="Per-sample component statistics which comprise the Fisher's Exac
##GATKCommandLine=<ID=GenomicsDBImport,CommandLine="GenomicsDBImport --genomicsdb-workspace-path Walrus_DB --varian
##GATKCommandLine=<ID=GenotypeGVCFs,CommandLine="GenotypeGVCFs --output Walrus_MT.vcf.gz --variant gendb://Walrus_D
##GATKCommandLine=<ID=HaplotypeCaller,CommandLine="HaplotypeCaller --sample-ploidy 1 --emit-ref-confidence GVCF --
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the same order
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency, for each ALT allele, in the same order as listed">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=BaseQRankSum,Number=1,Type=Float,Description="Z-score from Wilcoxon rank sum test of Alt Vs. Ref base qu
```

Each of these steps requires specific files to work with!



SNP data

VCF file (Variant call format)

Followed by the data:

```
#CHROM          POS      ID       REF      ALT      QUAL             FILTER  INFO
NC_004029.2     131      .        T        C        356.22   .       AC=1;AF=0.022;AN=45;DP=143;FS=0.000;MLEAC=1;MLEAF=0
NC_004029.2     162      .        T        C        18479.23.        AC=15;AF=0.333;AN=45;BaseQRankSum=0.00;DP=543;FS=0.
NC_004029.2     198      .        C        T        608.22   .       AC=1;AF=0.022;AN=45;DP=410;FS=0.000;MLEAC=1;MLEAF=0
NC_004029.2     387      .        G        A        547.22   .       AC=1;AF=0.022;AN=45;DP=408;FS=0.000;MLEAC=1;MLEAF=0
NC_004029.2     616      .        T        C        235.62   .       AC=1;AF=0.022;AN=45;DP=406;FS=0.000;MLEAC=1;MLEAF=0
NC_004029.2     741      .        C        T        819.22   .       AC=1;AF=0.022;AN=45;DP=412;FS=0.000;MLEAC=1;MLEAF=0
NC_004029.2     743      .        C        T        819      .       AC=1;AF=0.022;AN=45;DP=413;FS=0.000;MLEAC=1;MLEAF=0
```

Reference name

Each of these steps requires specific files to work with!

SNP data

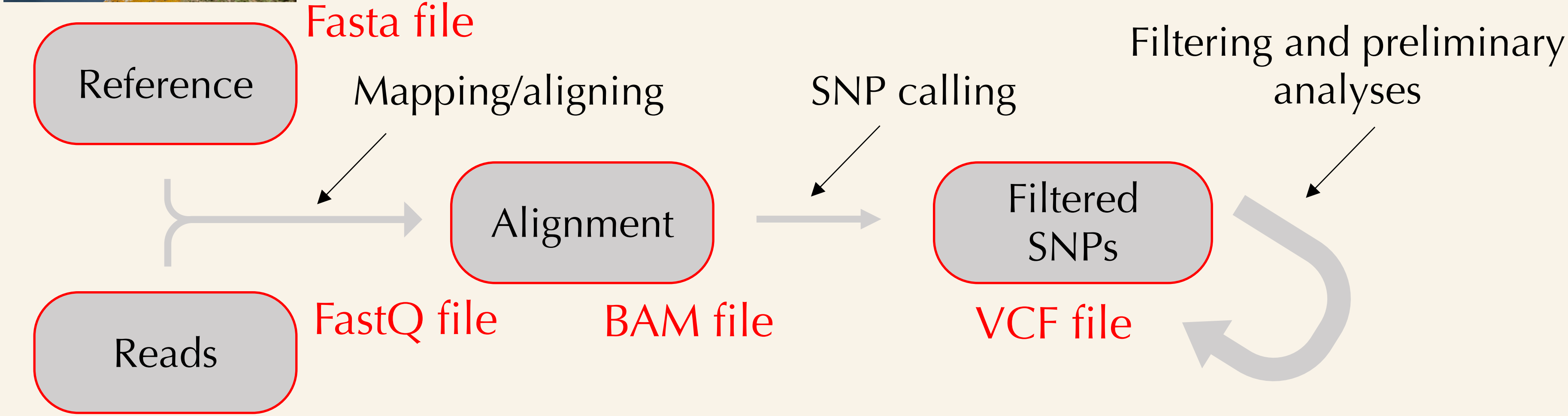
VCF file (Variant call format)

Followed by the data:

Geno**T**ype: **A**llele **D**epth: Read Depth (DP): **G**enotype **Q**uality: **P**hred-scaled **L**ikelihood

```
FORMAT          WLR001            WLR002             WLR003            WLR004            WL
GT:AD:DP:GQ:PL  0:0,0:0:0:0,0     0:2,0:2:90:0,90    0:0,0:0:0:0,0     0:0,0:0:0:0,0
GT:AD:DP:GQ:PL  0:0,0:0:0:0,0     0:4,0:4:99:0,135   0:1,0:1:0:0,0     0:1,0:1:45:0,45
GT:AD:DP:GQ:PL  0:0,0:0:0:0,0     0:5,0:5:46:0,46    0:0,0:0:0:0,0     0:2,0:2:90:0,90
GT:AD:DP:GQ:PL  0:0,0:0:0:0,0     0:3,0:3:99:0,135   0:0,0:0:0:0,0     0:2,0:2:45:0,45
GT:AD:DP:GQ:PL  0:0,0:0:0:0,0     0:0,0:0:0:0,0      0:0,0:0:0:0,0     0:0,0:0:0:0,0
GT:AD:DP:GQ:PL  0:0,0:0:0:0,0     0:3,0:3:99:0,128   0:0,0:0:0:0,0     0:1,0:1:45:0,45
GT:AD:DP:GQ:PL  0:0,0:0:0:0,0     0:3,0:3:99:0,128   0:0,0:0:0:0,0     0:1,0:1:45:0,45
GT:AD:DP:GQ:PL  0:0,0:0:0:0,0     0:3,0:3:99:0,128   0:0,0:0:0:0,0     0:1,0:1:45:0,45
GT:AD:DP:GQ:PL  0:0,0:0:0:0,0     0:3,0:3:99:0,135   0:0,0:0:0:0,0     0:1,0:1:42:0,42
GT:AD:DP:GQ:PL  0:0,0:0:0:0,0     0:1,0:1:45:0,45    0:0,0:0:0:0,0     0:1,0:1:42:0,42
GT:AD:DP:GQ:PL  0:1,0:1:45:0,45   0:1,0:1:45:0,45    0:0,0:0:0:0,0     0:3,0:3:99:0,119
GT:AD:DP:GQ:PL  0:1,0:1:45:0,45   0:1,0:1:45:0,45    0:0,0:0:0:0,0     0:3,0:3:99:0,119
GT:AD:DP:GQ:PL  0:1,0:1:0:0,0     0:1,0:1:0:0,0      0:0,0:0:0:0,0     0:0,0:0:0:0,0
GT:AD:DP:GQ:PL  0:1,0:1:0:0,0     0:1,0:1:0:0,0      0:0,0:0:0:0,0     0:0,0:0:0:0,0
GT:AD:DP:GQ:PL  0:0,0:0:0:0,0     0:1,0:1:45:0,45    0:0,0:0:0:0,0     0:0,0:0:0:0,0
GT:AD:DP:GQ:PL  0:0,0:0:0:0,0     0:1,0:1:45:0,45    0:0,0:0:0:0,0     0:0,0:0:0:0,0
```

After all this, what does a variant calling pipeline look like?



Fasta file

Reference

Mapping/aligning

SNP calling

Filtering and preliminary analyses

Alignment

Filtered SNPs

Reads

FastQ file

BAM file

VCF file

e.g.
population
data

Questions?

HOW? WHEN?
WHO? WHERE?
WHEN? WHY? WHAT? HOW?
WHERE?
HOW? WHAT? WHO? WHERE?
WHO? WHERE? HOW? HOW? WHERE?
WHAT?
WHO?
WHY? WHAT?
WHERE? WHEN?
WHAT? WHEN?
WHERE? WHO?
HOW? WHAT?
WHO? WHERE?
WHY? WHAT? HOW?
WHO? WHEN? WHERE?
WHAT?
WHY? WHEN? WHAT?
WHO? WHY? HOW?
HOW? WHERE?
WHAT? WHY?
WHEN?
WHO? WHAT?
WHAT? HOW?
WHO? WHY? WHERE?
WHAT? WHEN?
WHO? WHAT?
WHERE? WHY?
HOW? WHEN?
WHAT? WHO? WHEN?
WHY? WHERE?
WHEN? HOW?
HOW? WHO? WHAT?
WHAT? WHEN?
WHERE? HOW?
WHEN? WHO?

Today:

1) Introduction: variant calling, why do we want to do this, and what it is?
2) Variant calling pipelines/methods and pitfalls
3) Practical session, going through (parts of) a SNP calling pipeline and interpret biological results