Mapping, SNP calling and the study of genetic variation in ecology and evolution

Formation CNRS 2023

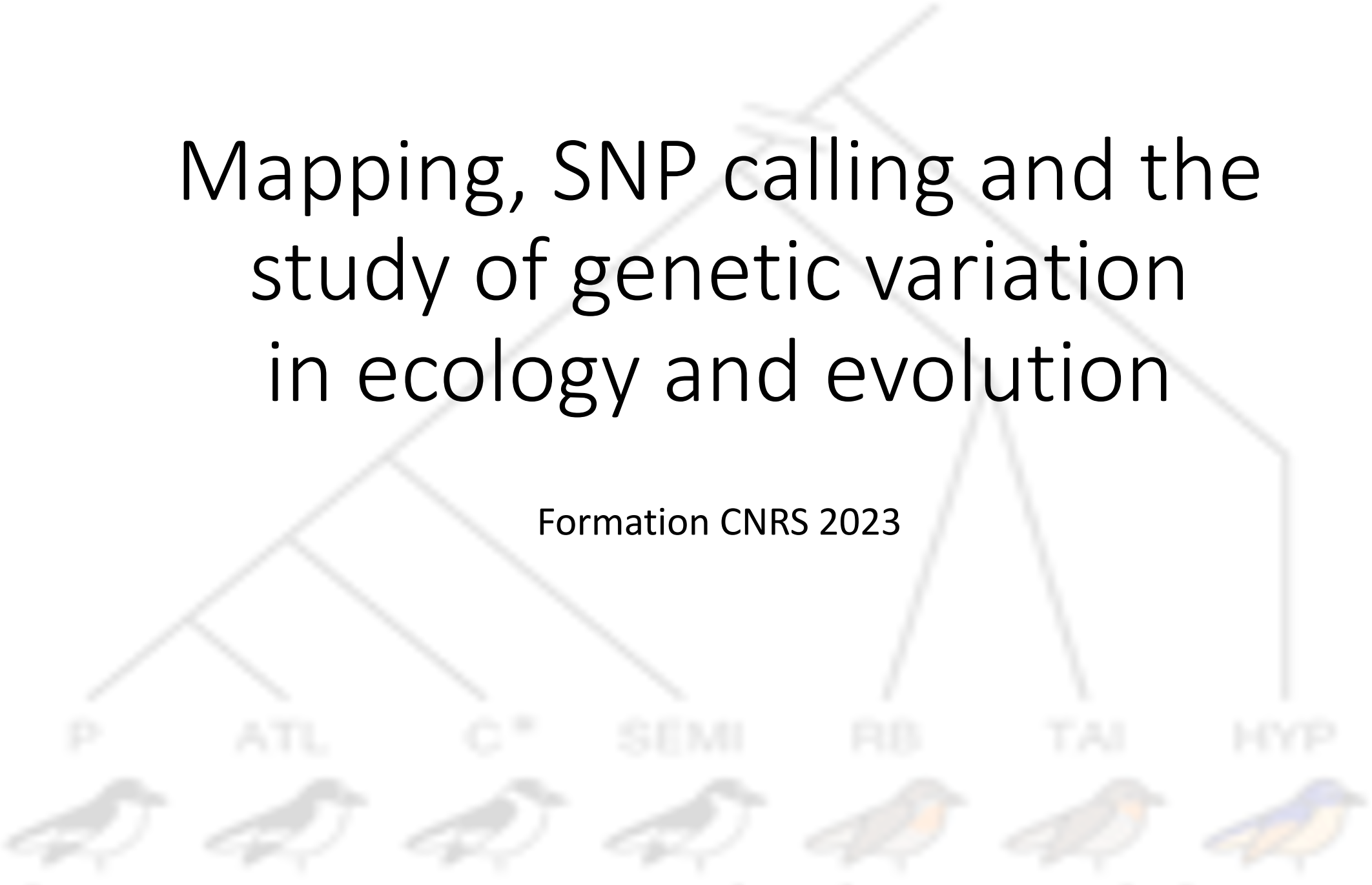
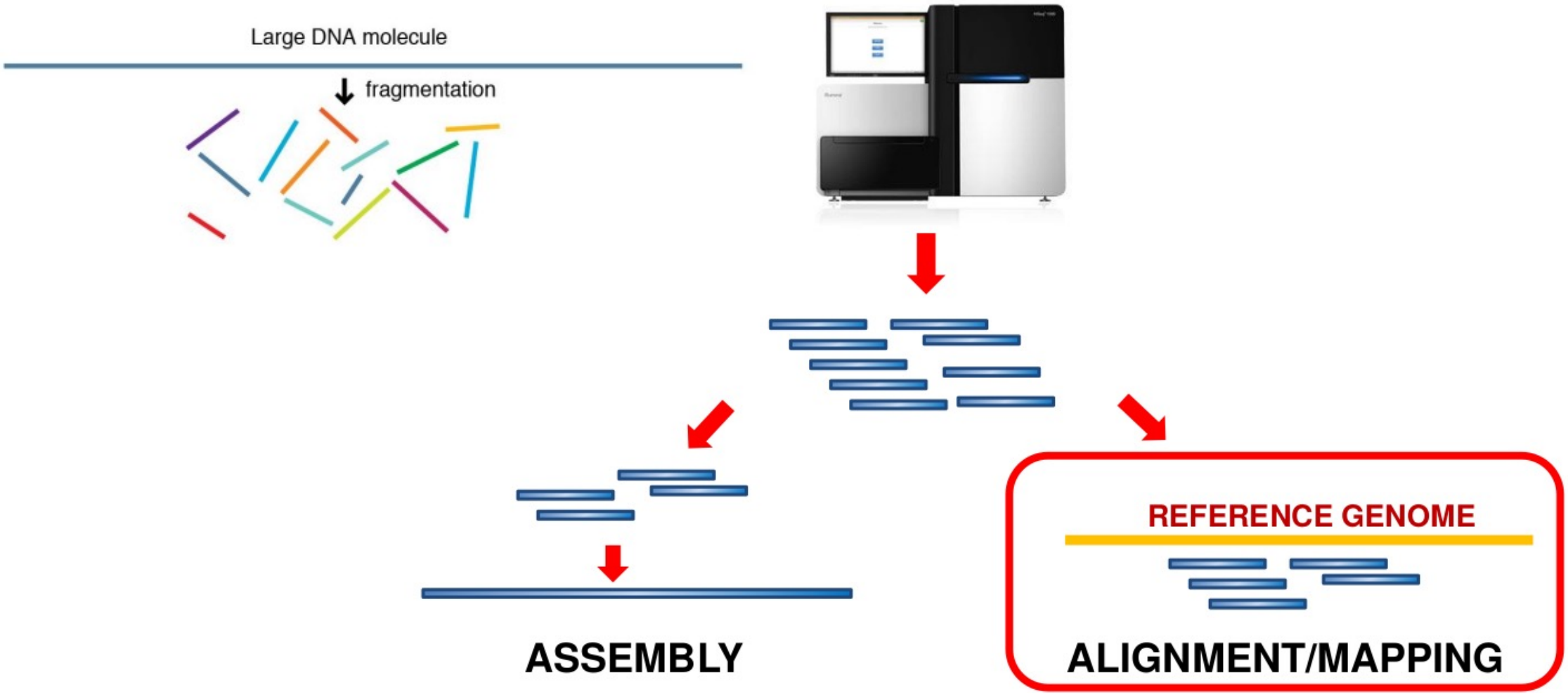


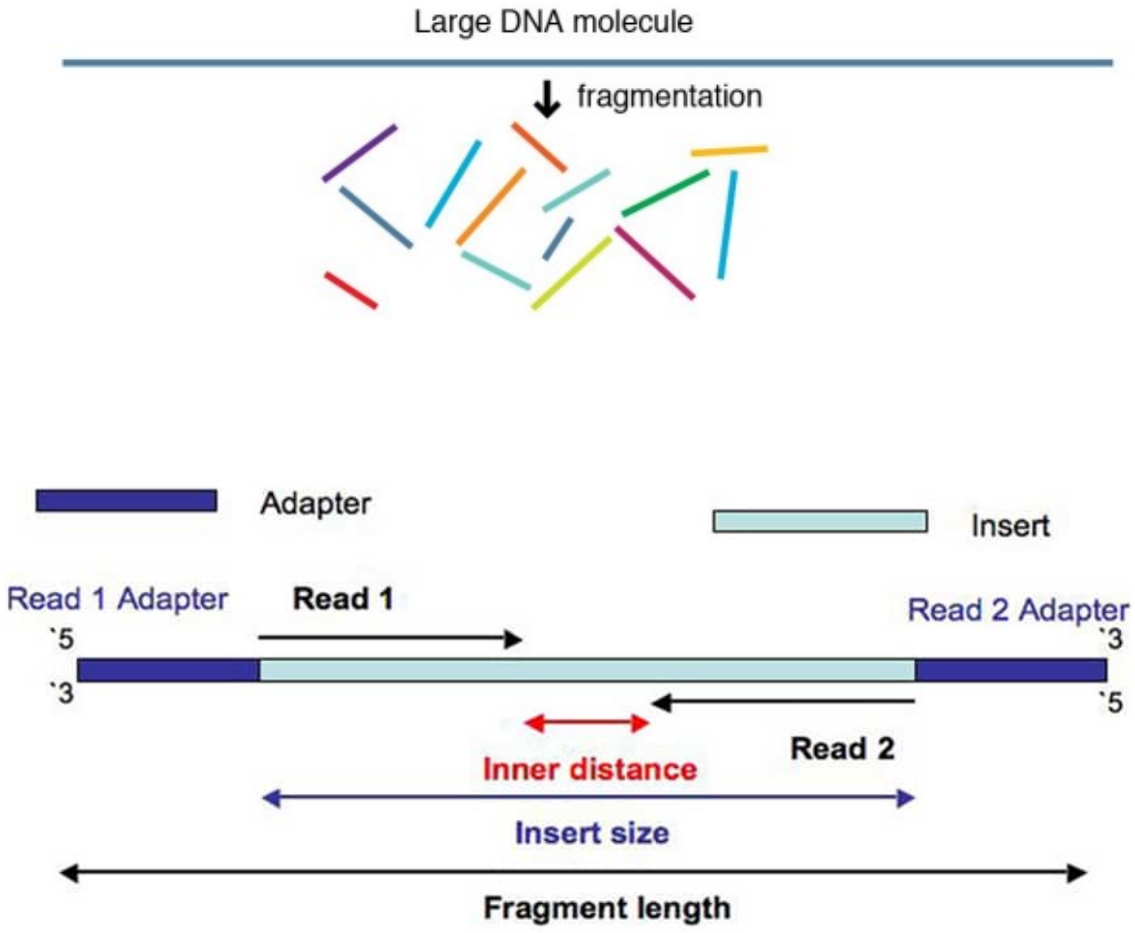
Table of contents

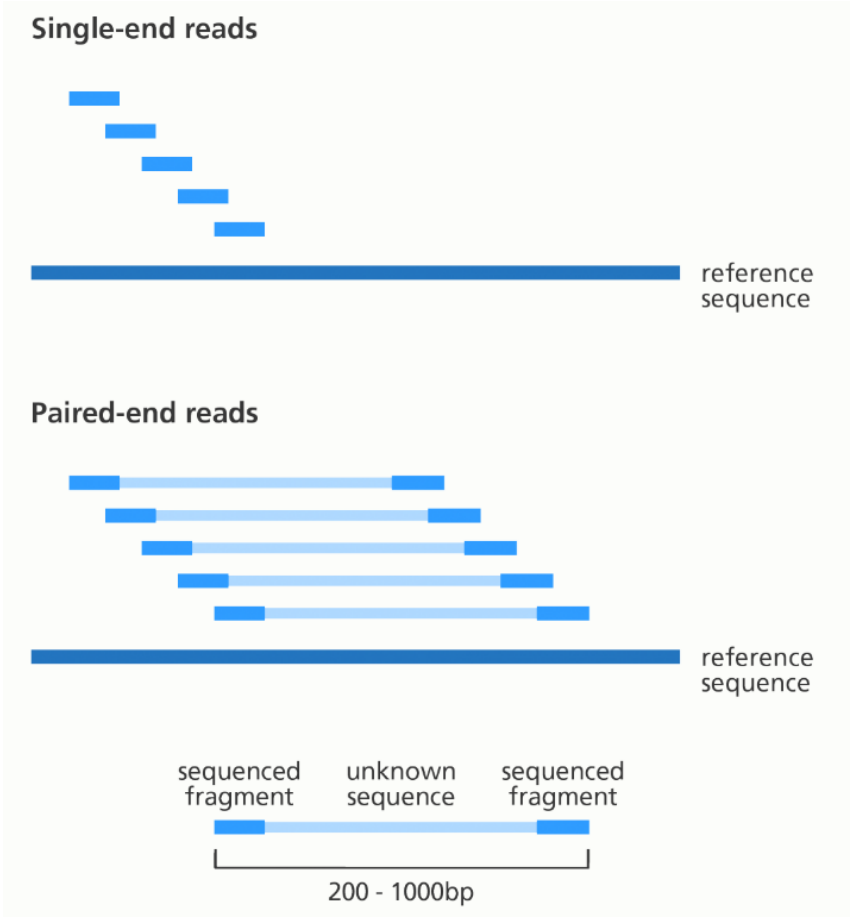
Table of contents

NGS principle

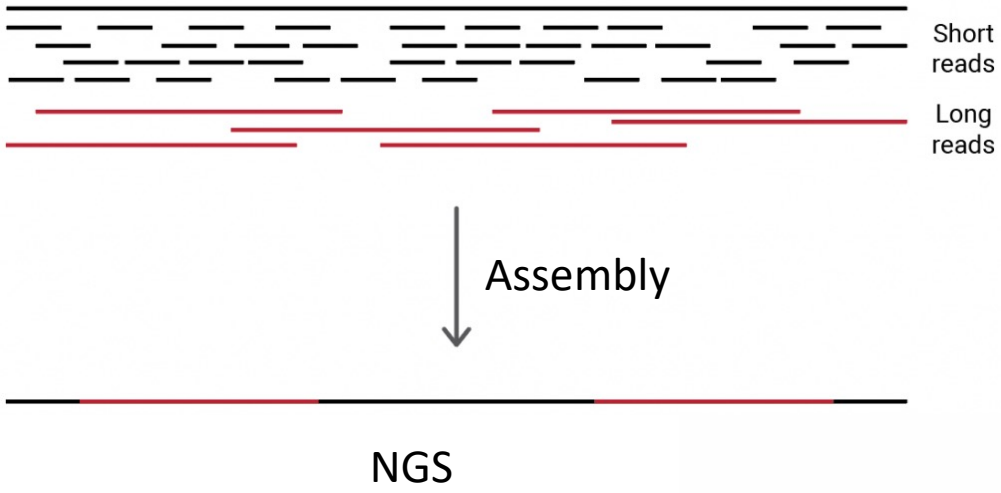


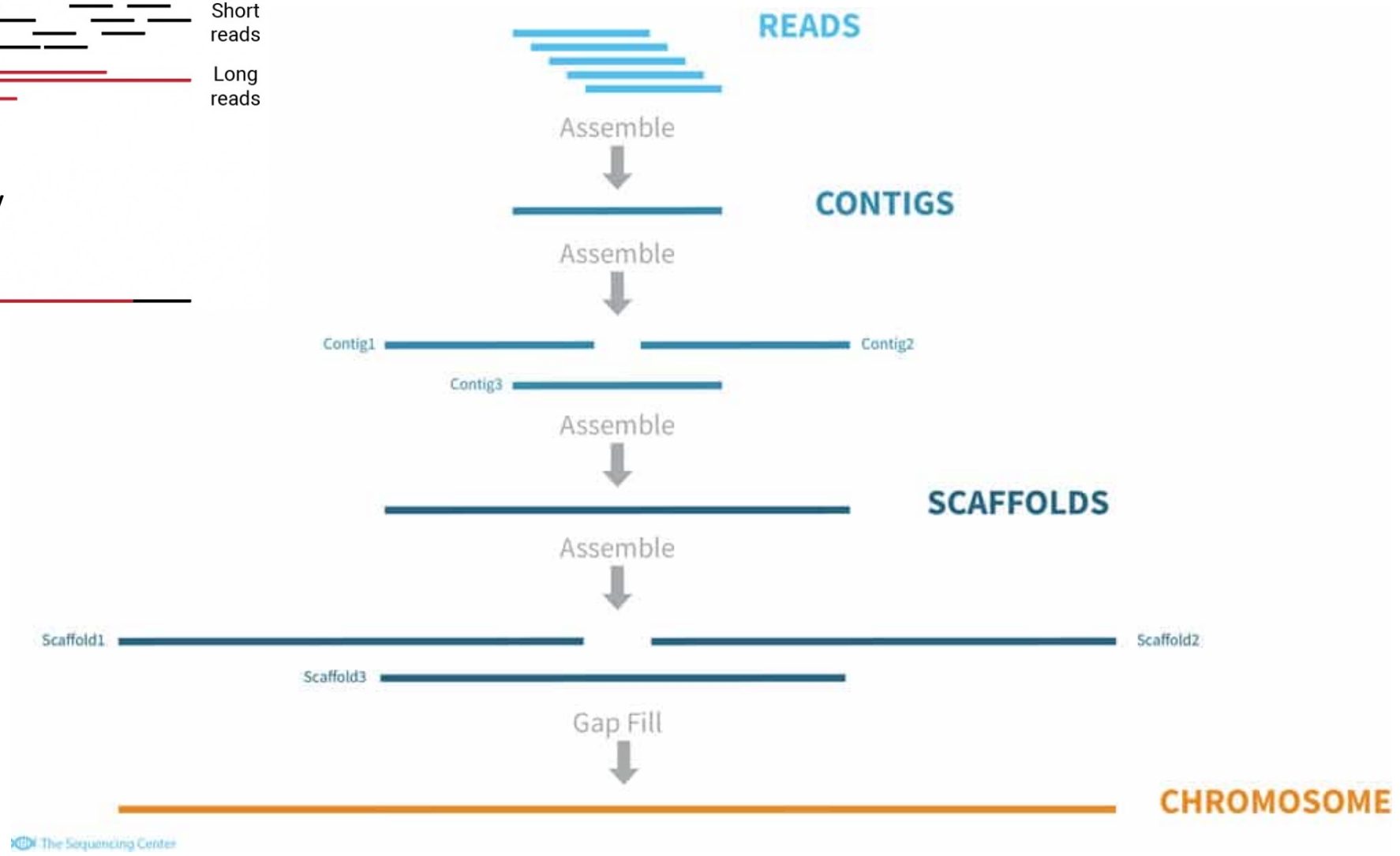
NGS Single-end vs paired-end reads



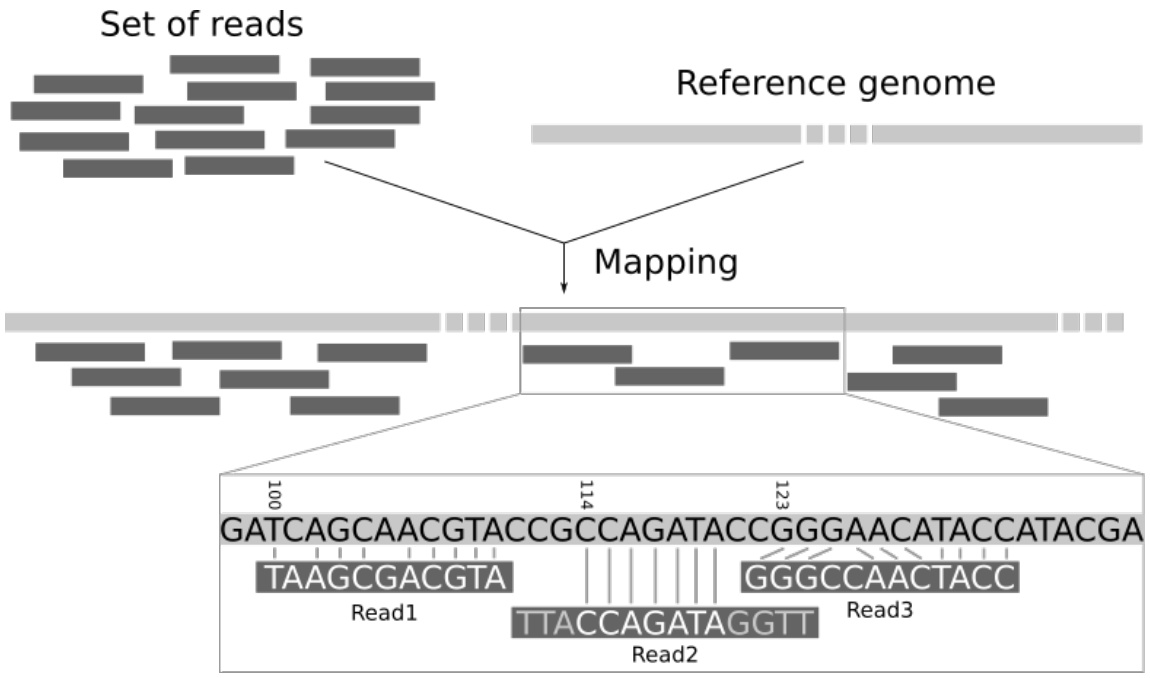


De novo assembly





Mapping / SNP calling principle



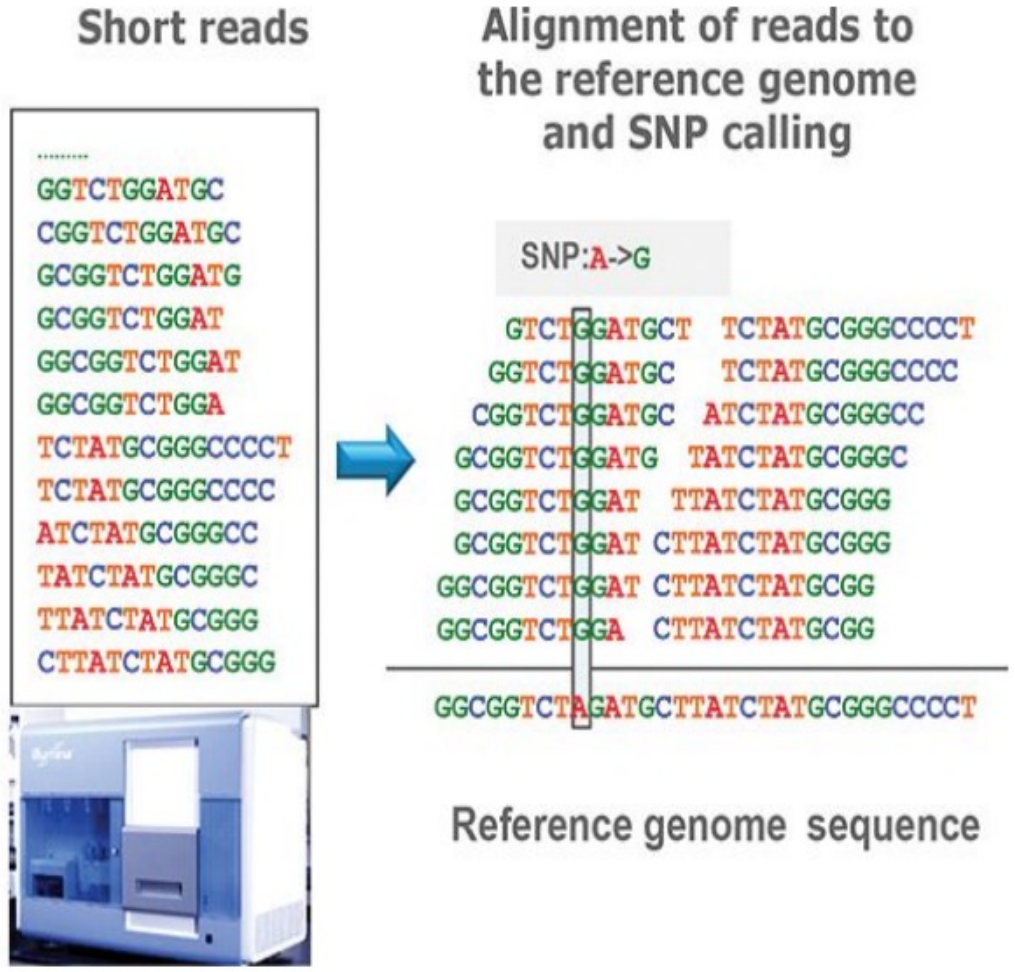


Table of contents

Genetic variation

- Genetic variation (polymorphisms) refers to the differences found in the genome sequence between individuals within a population or species

(here, haploid)

Ind1	A	C	T	T	A	G	A	T
Ind2	G	C	T	C	A	G	T	C
Ind3	G	C	G	C	A	G	T	C
Ind4	A	C	T	T	-	G	T	A
Ind5	G	C	T	C	A	G	T	C

Genetic variation

- Genetic variation (polymorphisms) refers to the differences found in the genome sequence between individuals within a population or species

(here, haploid)

Ind1	A	C	T	T	A	G	A	T
Ind2	G	C	T	C	A	G	T	C
Ind3	G	C	G	C	A	G	T	C
Ind4	A	C	T	T	-	G	T	A
Ind5	G	C	T	C	A	G	T	C

How many:

- Sites
- Polymorphisms
- SNPs
- Alleles
- Haplotypes

Genetic variation

- Genetic variation (polymorphisms) refers to the differences found in the genome sequence between individuals within a population or species

(here, haploid)

Ind1	A	C	T	T	A	G	A	T
Ind2	G	C	T	C	A	G	T	C
Ind3	G	C	G	C	A	G	T	C
Ind4	A	C	T	T	-	G	T	A
Ind5	G	C	T	C	A	G	T	C

How many:

- Sites / Positions: 8
(Polymorphic + monomorphic)
- Polymorphisms:6
- SNPs: 5
(6 polymorphisms: 5 SNPs + 1 indel)
- Alleles: 2, 1, 2, 2, 2, 1, 2, 3
(5 sites bi-allelic + 1 tri-allelic)
- Haplotypes: 4

Genetic variation

- Genetic variation (polymorphisms) refers to the differences found in the genome sequence between individuals within a population or species

(here, haploid)

Ind1	A	C	T	T	A	G	A	T
Ind2	G	C	T	C	A	G	T	C
Ind3	G	C	G	C	A	G	T	C
Ind4	A	C	T	T	-	G	T	A
Ind5	G	C	T	C	A	G	T	C
Sp2	G	T	T	C	A	A	T	C

How many:

- Sites / Positions: 8
(Polymorphic + monomorphic)
- Polymorphisms:6
- SNPs: 5
(6 polymorphisms: 5 SNPs + 1 indel)
- Alleles: 2, 1, 2, 2, 2, 1, 2, 3
(5 sites bi-allelic + 1 tri-allelic)
- Haplotypes: 4

- Genetic differences fixed between species = divergent sites / substitutions

Genetic variation

- Genetic variation (polymorphisms) refers to the differences found in the genome sequence between individuals within a population or species

(here, haploid)

Ind1	A	C	T	T	A	G	A	T
Ind2	G	C	T	C	A	G	T	C
Ind3	G	C	G	C	A	G	T	C
Ind4	A	C	T	T	-	G	T	A
Ind5	G	C	T	C	A	G	T	C
Sp2	G	T	T	C	A	A	T	C

In evolution, SNPs + divergent sites are sometimes grouped under the term SNV.

But different meaning in medicine ! (SNV = 0-1%; SNP = >1%)

- Genetic differences fixed between species = divergent sites / substitutions

Genetic variation

- The ultimate source of genetic variation is **mutation**

(here, haploid)

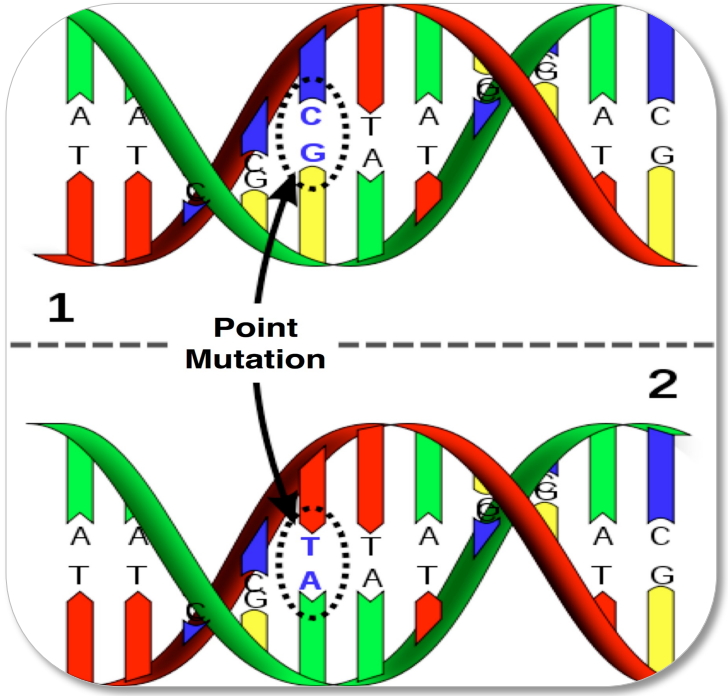
Ind1	A	C	T	T	A	G	A	T
Ind2	G	C	T	C	A	G	T	C
Ind3	G	C	G	C	A	G	T	C
Ind4	A	C	T	T	-	G	T	A
Ind5	G	C	T	C	A	G	T	C
Sp2	G	T	T	C	A	A	T	C

Different types of mutations that cause genetic variation

- point mutations (SNPs)
- small insertions/deletions (indels) – STRs (microsatellites)
- segmental mutations
- transpositions
- chromosome fission/fusion
- aneuploidy and whole genome duplications

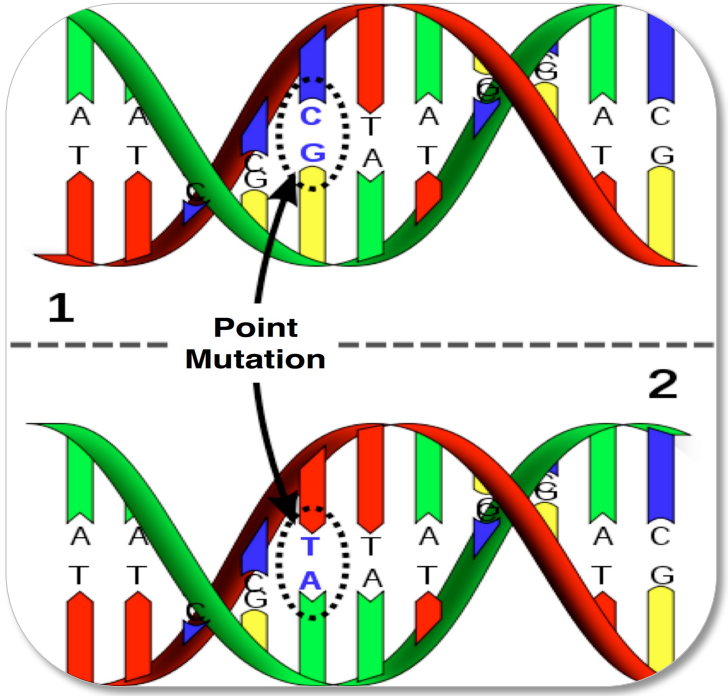
Point mutations (SNPs)

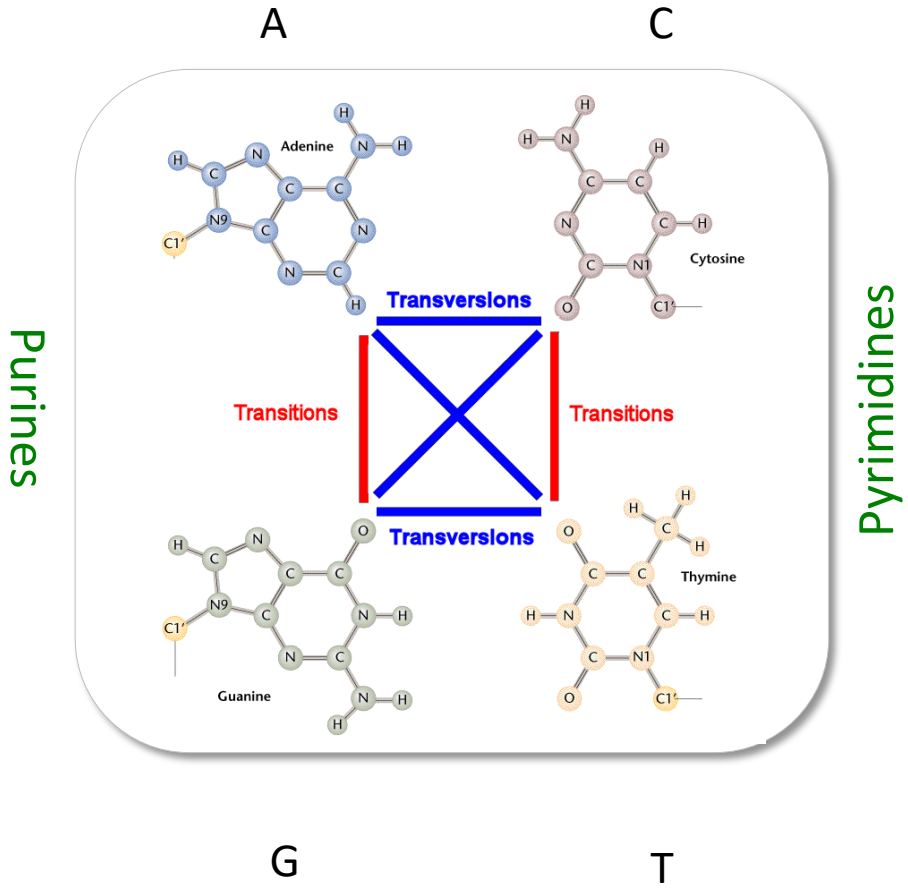
- Replacement of nucleotides without changing the number of nucleotides
- Arise by replication errors and spontaneous mutations (due to the instability of the DNA or induced by exogenous/endogenous sources)



Point mutations (SNPs)

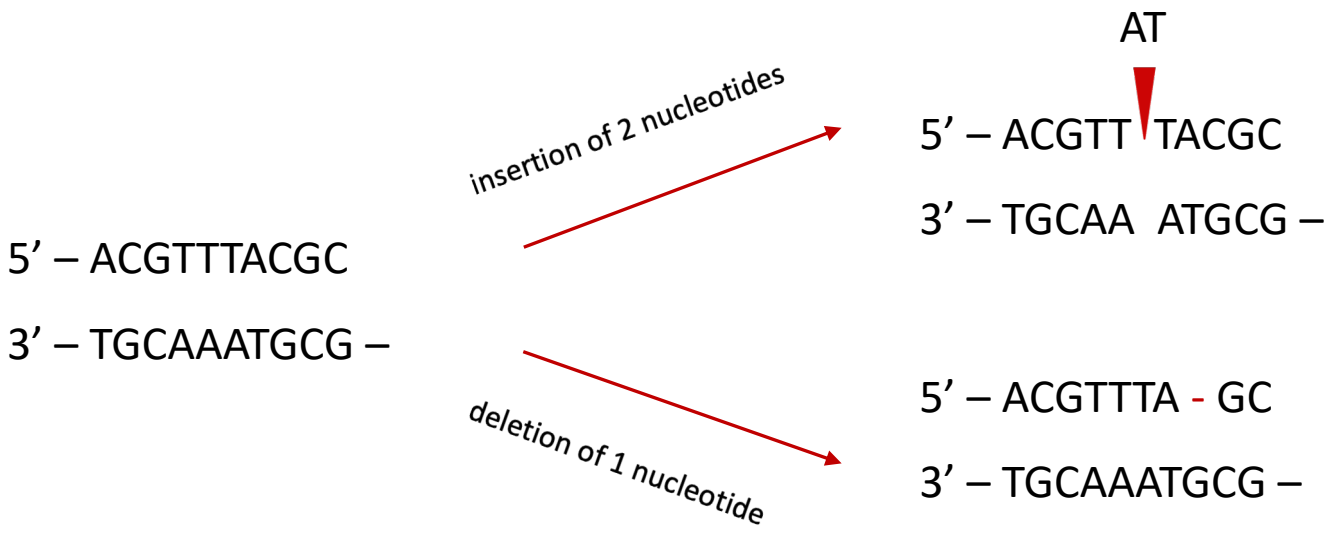
- Replacement of nucleotides without changing the number of nucleotides
- Arise by replication errors and spontaneous mutations (due to the instability of the DNA or induced by exogenous/endogenous sources)





Small Insertions/Deletions (Indels)

- small insertion and deletion of nucleotides from a sequence causing a change in the number of nucleotides ($\geq$ 1 nucleotides)



STRs (microsatellites)

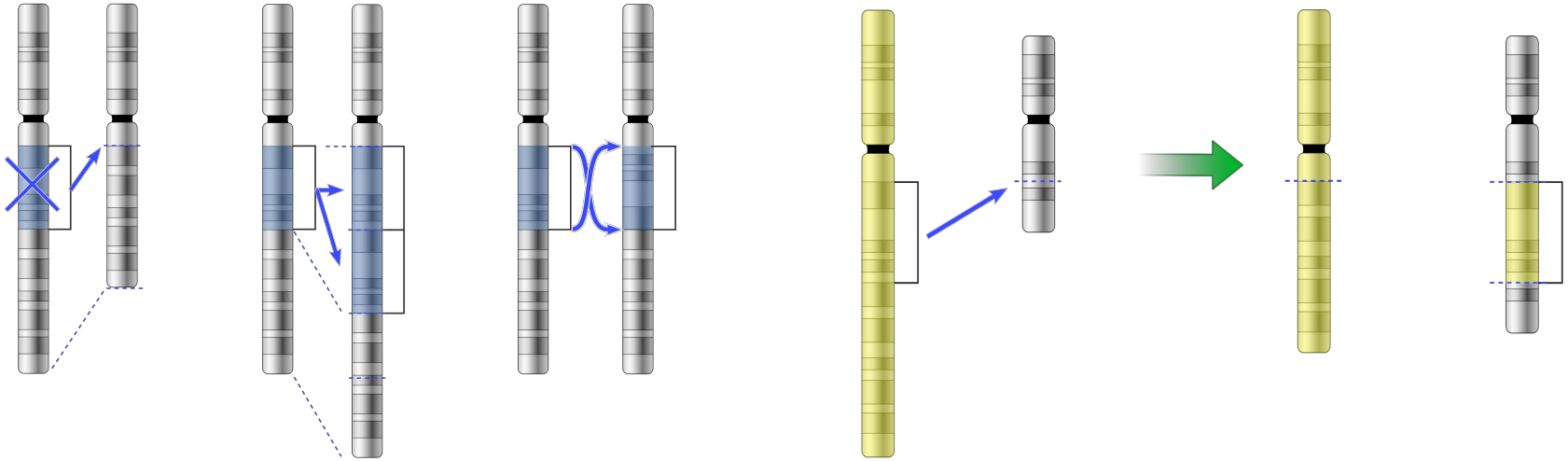
- Variation in the number of copy of small repeat units (classically of 4 bp)

```
ATGCTATATATATA----------GCATG
ATGCTATATATATATATATATAGCATG
```

- Can have many alleles and thus is more informative for example for individual identification or familial relationships reconstruction.

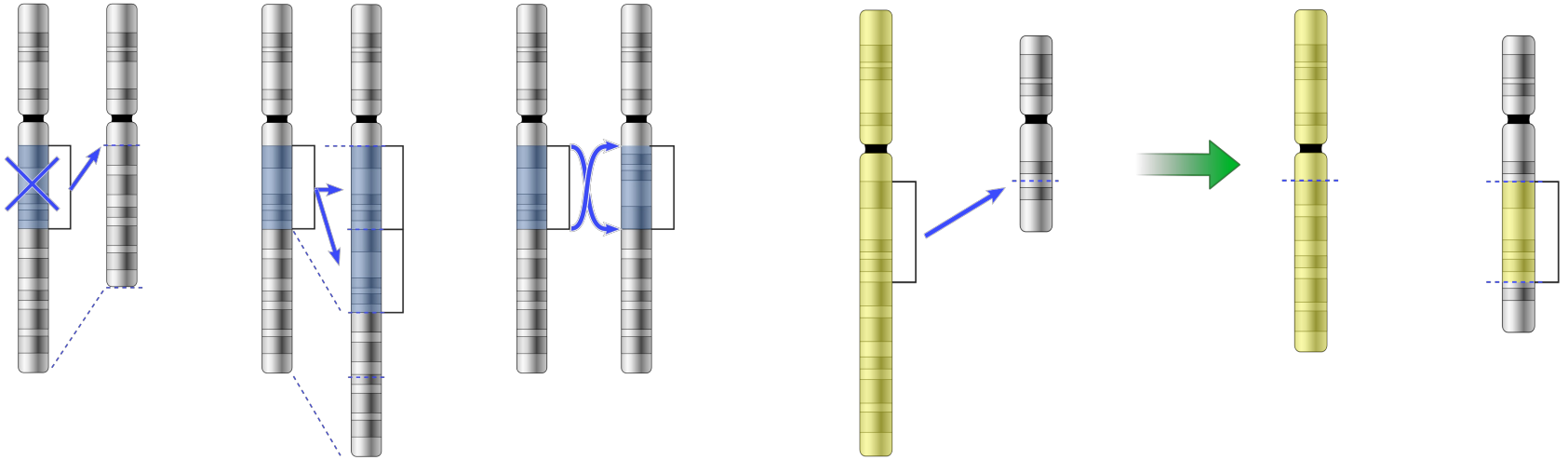
Segmental mutations

- Involving a change of ≥ 50 bp (chromosomal rearrangements)
  - Large insertions and deletions
  - Duplications
  - Inversions
  - Translocations



Segmental mutations

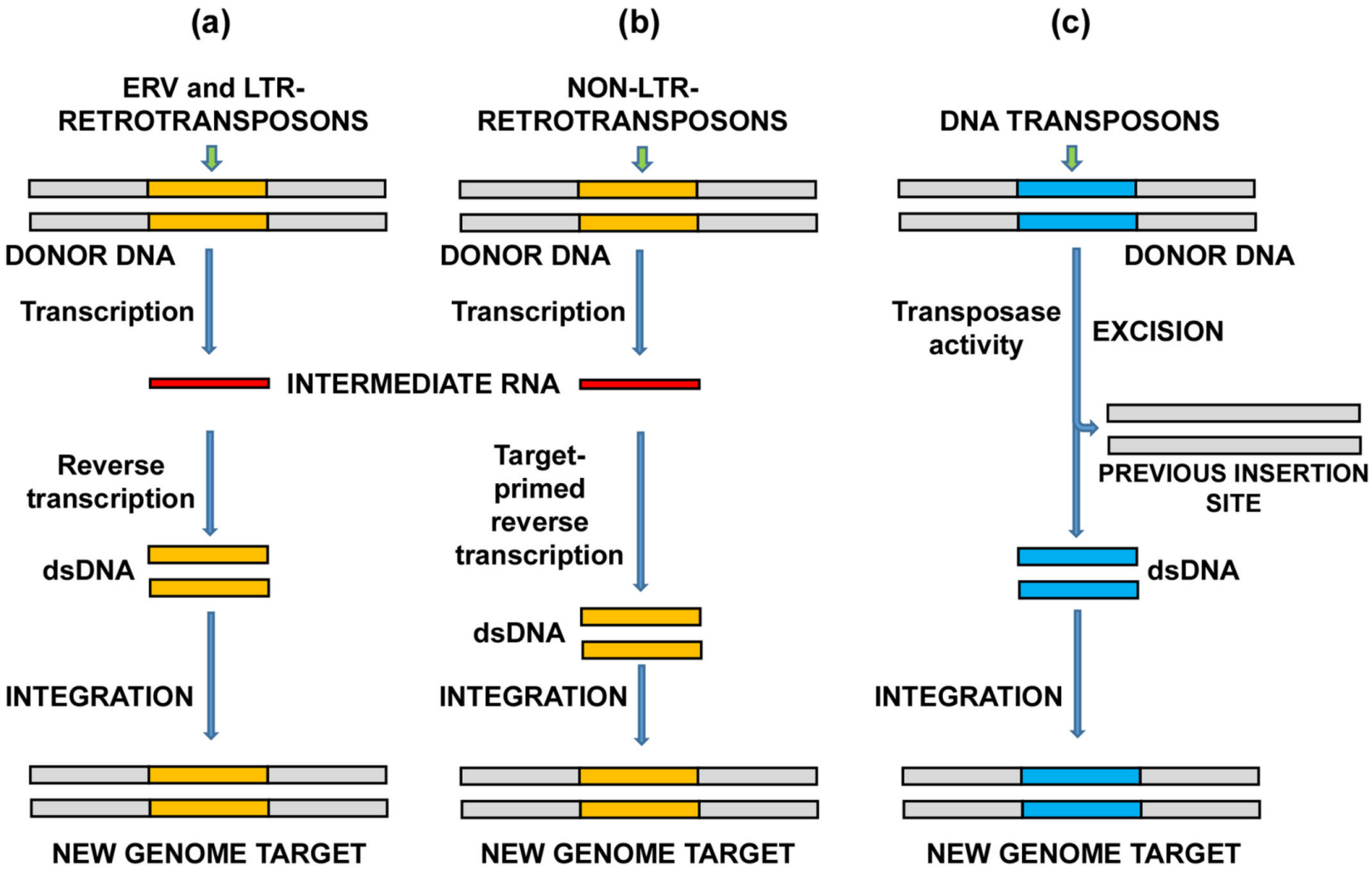
- Involving a change of ≥ 50 bp (chromosomal rearrangements)
  - Large insertions and deletions
  - Duplications
  - Inversions
  - Translocations
- Segmental mutations can lead to copy number variation (CNV) among individual genomes.



Transposable elements

Fambrini et al. 2020 *Genesis*

There are two main classes of transposable elements

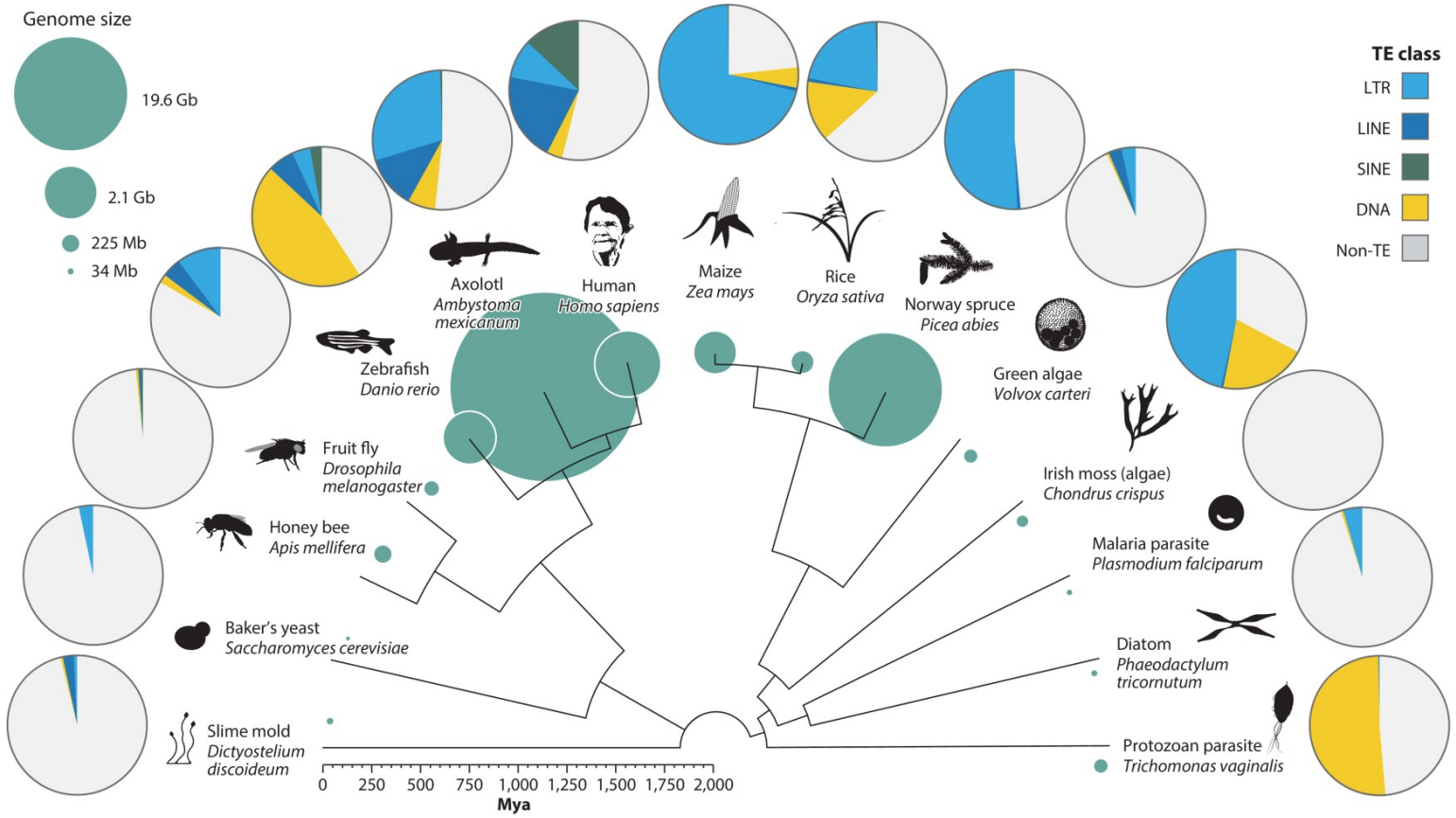


Copy and Paste

Cut and Paste

Transposable elements

- They occur in variable proportions in different genomes



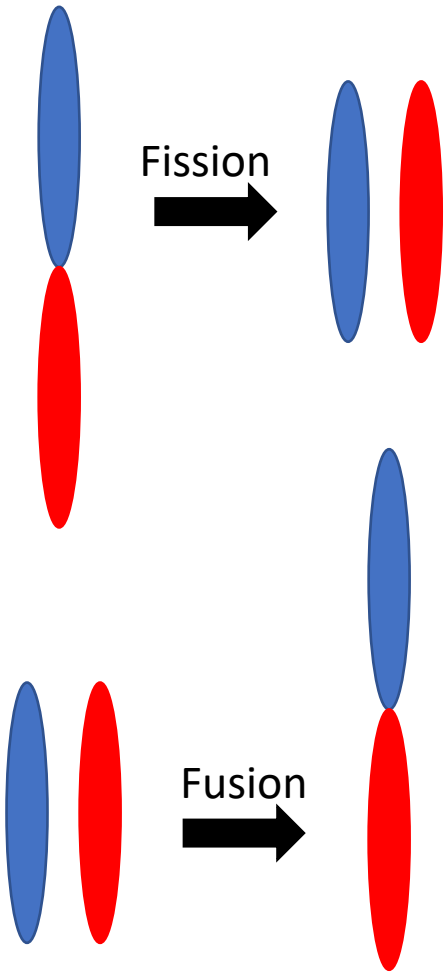
Wells JN, Feschotte C. 2020
Annu. Rev. Genet. 54:539–61

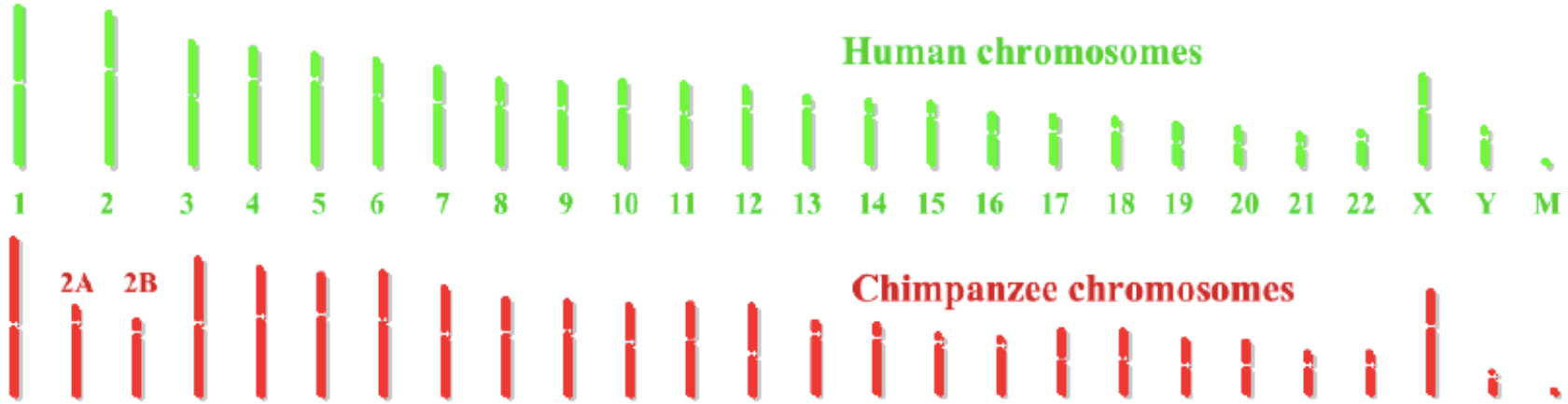
Wells & Feschotte. 2020 *Annual Rev. Genet.*

Transposons and repetitive DNA sequences in the genome

- DNA transposons create mutations by inserting themselves (presence/absence)
- Retrotransposons create copy-number variants
- They are one type of repetitive DNA (stretches of DNA sequences that occur in multiple copies in the genome)
- Repetitive DNA also corresponds to specific genome structures as centromeres & telomeres…

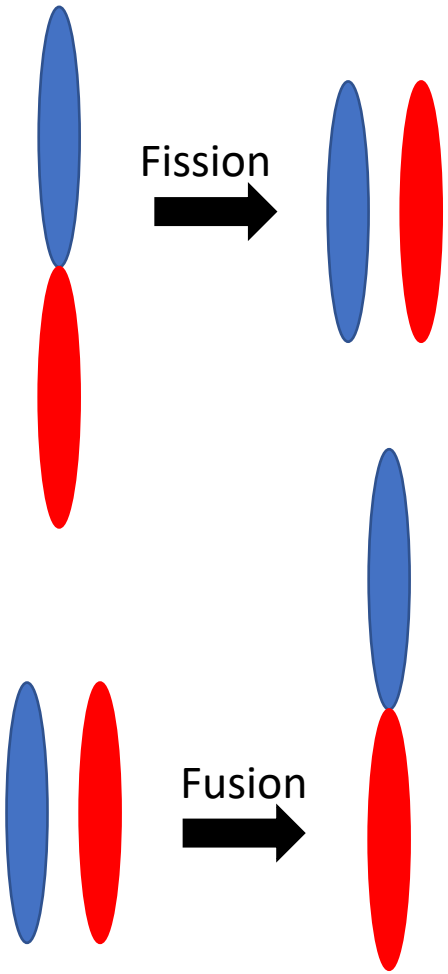
Chromosome fission/fusion

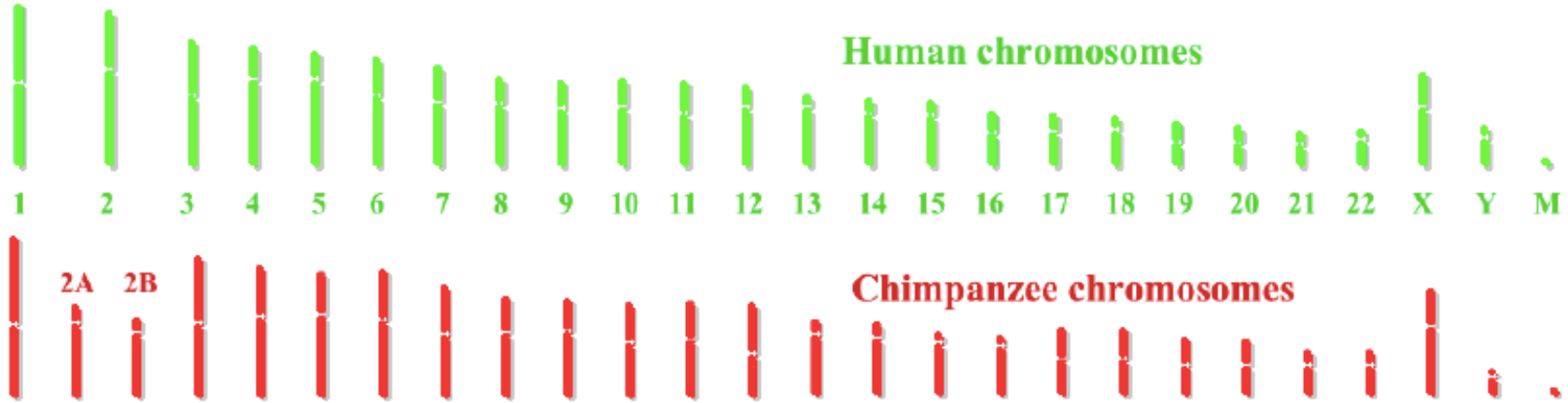




1 « extra » chromosome in chimpanzee

Chromosome fission/fusion





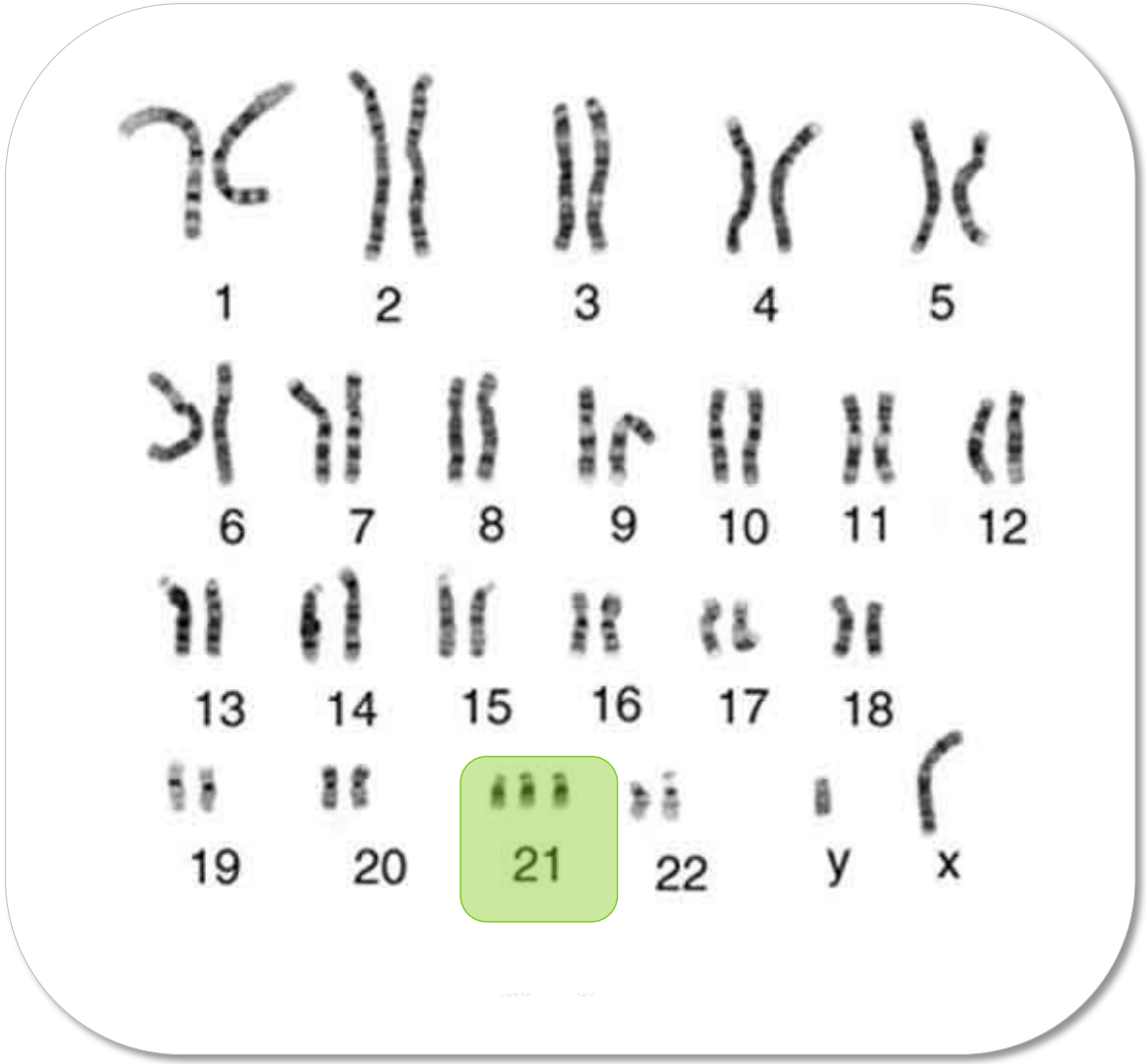
Bonobo/Chimpanzé	Homme	Gorille	Orang	Gibbons
24	23	24	24	19 à 26

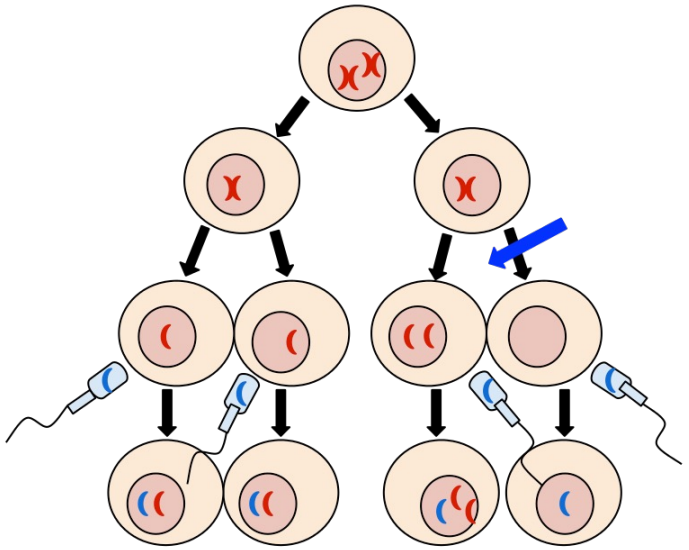
1 « less » chromosome in humans
Robertsonian fusion

Acceleration of chromosomal rearrangments

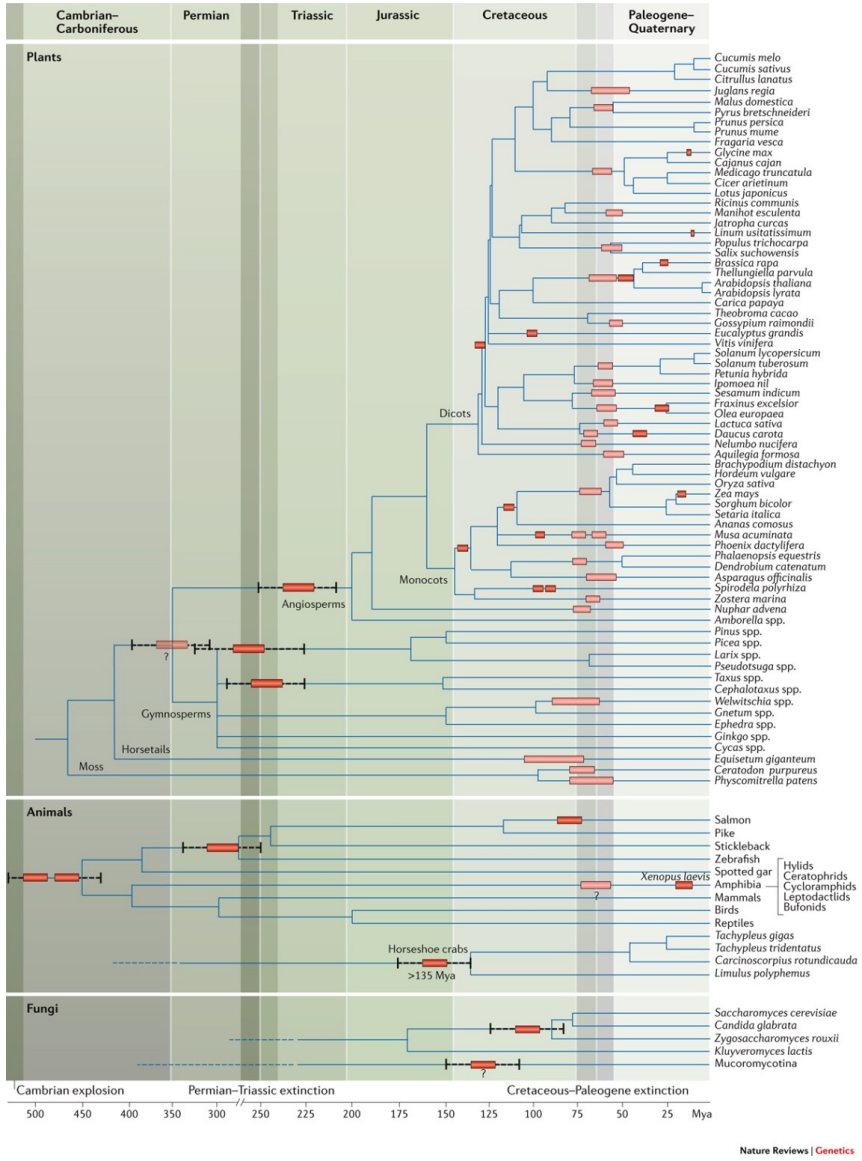
Aneuploidy

Chromosomal duplications or losses usually have severe effects on phenotype





Whole-genome duplication



- Polyploidy, or whole-genome duplication (WGD), is **usually an evolutionary dead-end**
- Although polyploidy is a frequent and recurrent phenomenon, the number of WGDs that have become established in the long term is low
- If established, WGD leads to a **sudden increase in genome size**
- WGD is most common in the plant kingdom

REF …

Different types of mutations that cause genetic variation

- point mutations (SNPs)
- small insertions/deletions (indels) – STRs (microsatellites)
- segmental mutations
- transpositions
- chromosome fission/fusion
- aneuploidy and whole genome duplications

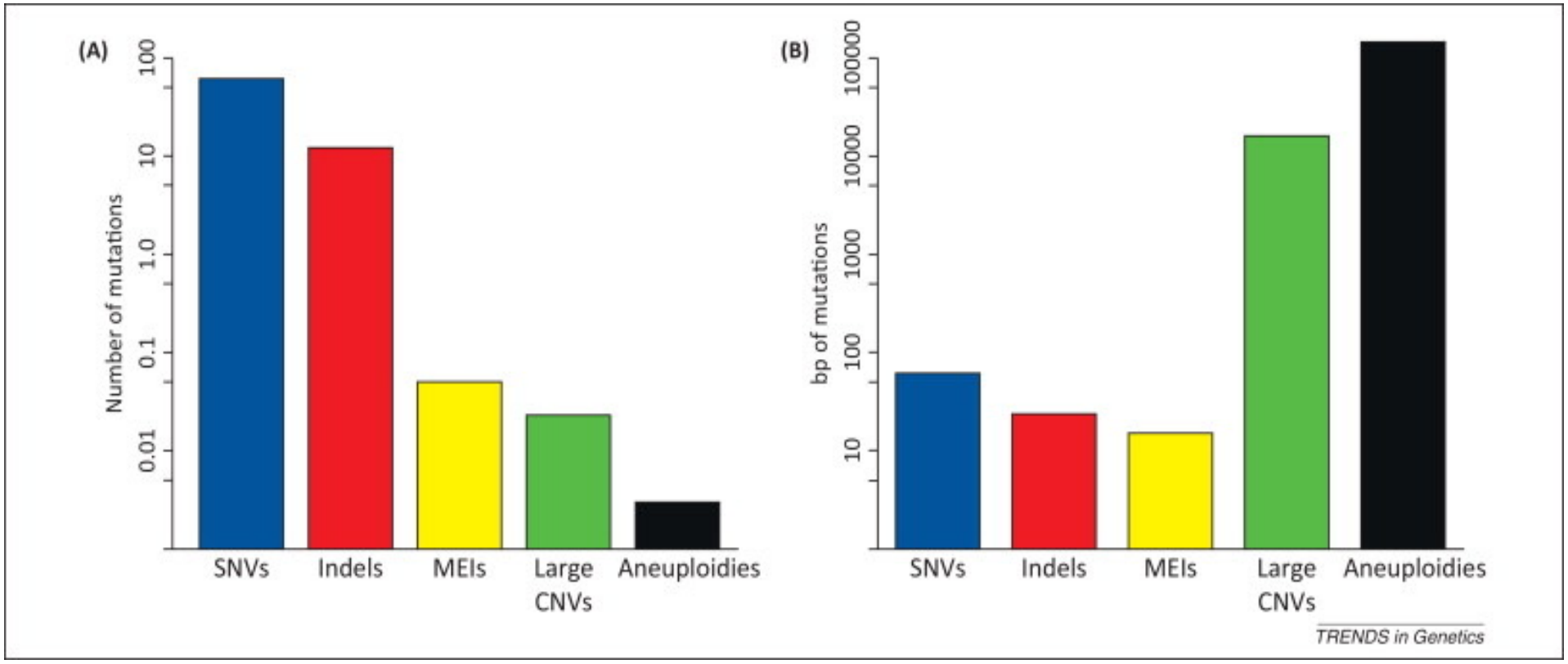
All these mutations have different underlying and repair mechanisms, and thus have different rates !

Describing genetic variation

Note log scale!

Number of events

Number of bp involved



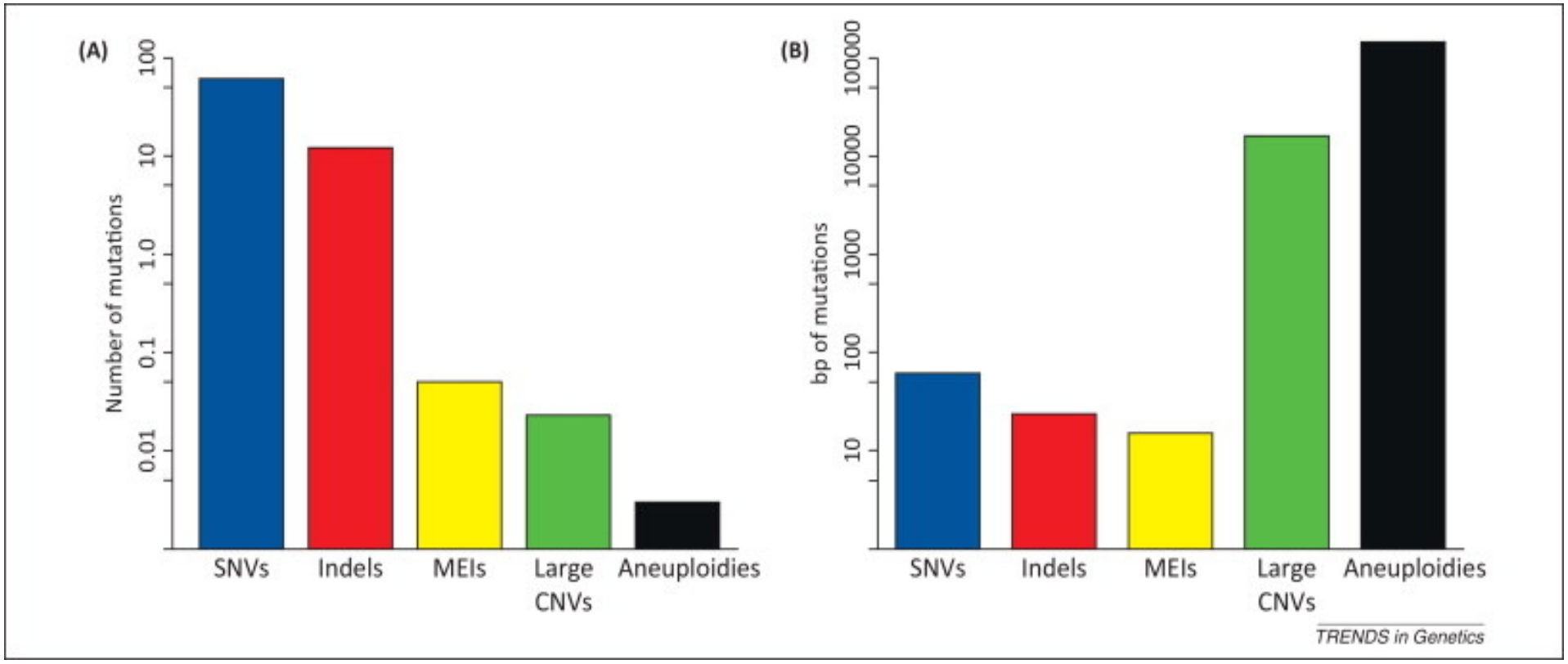
SNVs = single nucleotide variants
MEIs = transposabe (or mobile) element insertions
CNVs = copy number variants

Campbell & Eichler 2013 *Trends in Genet.*

Describing genetic variation

Number of events

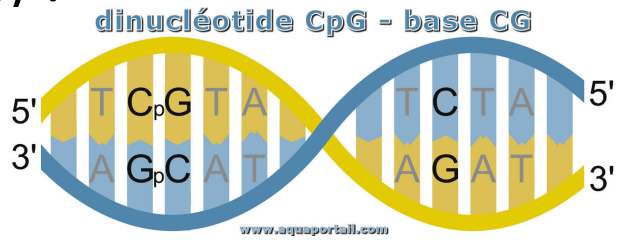
Number of bp involved



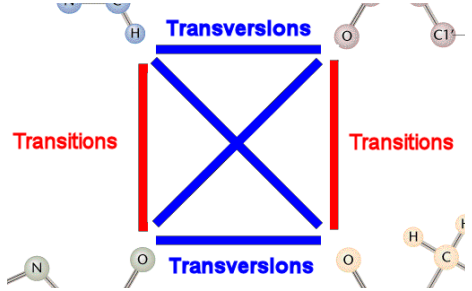
Much of the genetic variation we study consists of bi-allelic SNPs (mutation rate around 10^-8)

Describing genetic variation

But there is **also a lot of variability among point mutations** (SNPs) !



- CpG (13-fold more mutable than non-CpG because of s. pontaneous deamination of methylated Cs in mammals) : CpG = 2% of the human genome, 19% of de novo mutations…
- AT-rich regions more easily repaired than CG-rich ones because double-bonded
- Transitions mutation rate 2-fold higher than Transversions mutation rate (G-T and A-C mispairing being most common)
- Male versus female mutations (number of replications)…



Describing genetic variation

(here, haploid)

Ind1	A	C	T	T	A	G	A	T
Ind2	G	C	T	C	A	G	T	C
Ind3	G	C	G	C	A	G	T	C
Ind4	A	C	T	T	-	G	T	A
Ind5	G	C	T	C	A	G	T	C

For our analyses, we keep only bi-allelic SNPs, where we observe two alleles:

- the most common variant is called the **major allele** and the least common variant the minor allele

Describing genetic variation

(here, haploid)

Ind1	A	C	T	T	A	G	A	T
Ind2	G	C	T	C	A	G	T	C
Ind3	G	C	G	C	A	G	T	C
Ind4	A	C	T	T	-	G	T	A
Ind5	G	C	T	C	A	G	T	C
REF	G	C	T	T	-	G	T	A

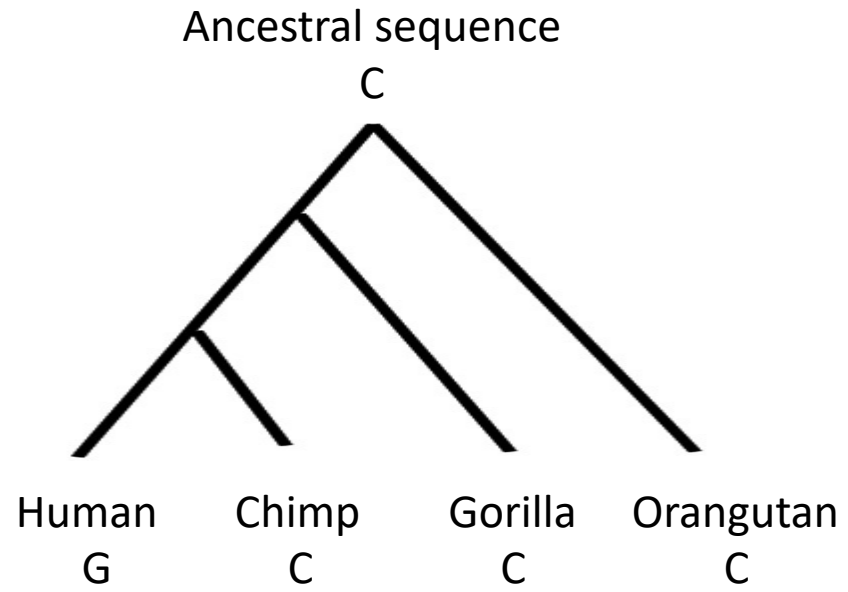
For our analyses, we keep only bi-allelic SNPs, where we observe two alleles:

- the most common variant is called the **major allele** and the least common variant the minor allele
- the variant similar to the reference genome is the **reference allele** and the other one the alternative allele

Describing genetic variation

(here, haploid)

Ind1	A	C	T	T	A	G	A	T
Ind2	G	C	T	C	A	G	T	C
Ind3	G	C	G	C	A	G	T	C
Ind4	A	C	T	T	-	G	T	A
Ind5	G	C	T	C	A	G	T	C
REF	G	C	T	T	-	G	T	A
ANC	G	C	T	T	A	C	T	C



For our analyses, we keep only bi-allelic SNPs, where we observe two alleles:

- the most common variant is called the **major allele** and the least common variant the minor allele
- the variant similar to the reference genome is the **reference allele** and the other one the alternative allele
- the variant similar to the ancestral genome is the **ancestral allele** and the other one the derived allele

Describing genetic variation

- Assessing the level of genetic variation

Sequence alignment of four DNA sequences consisting of 16 nucleotides

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
A	C	G	C	G	T	A	A	C	T	G	C	C	G	T	A
A	C	G	A	G	T	T	A	C	T	G	G	C	G	A	A
A	C	G	A	G	T	A	A	C	T	G	G	C	G	A	A
A	C	G	C	G	T	A	A	C	T	G	C	C	G	T	A

Describing genetic variation

- Assessing the level of genetic variation

Sequence alignment of four DNA sequences consisting of 16 nucleotides

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
A	C	G	C	G	T	A	A	C	T	G	C	C	G	T	A
A	C	G	A	G	T	T	A	C	T	G	G	C	G	A	A
A	C	G	A	G	T	A	A	C	T	G	G	C	G	A	A
A	C	G	C	G	T	A	A	C	T	G	C	C	G	T	A

- number of segregating sites *S*

Describing genetic variation

- Assessing the level of genetic variation

Sequence alignment of four DNA sequences consisting of 16 nucleotides

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
A	C	G	C	G	T	A	A	C	T	G	C	C	G	T	A
A	C	G	A	G	T	T	A	C	T	G	G	C	G	A	A
A	C	G	A	G	T	A	A	C	T	G	G	C	G	A	A
A	C	G	C	G	T	A	A	C	T	G	C	C	G	T	A

- number of segregating sites *S*
- $S = 4$
- or $4/16 = 0.25$ represents the density of segregating sites

Describing genetic variation

- Assessing the level of genetic variation

Sequence alignment of four DNA sequences consisting of 16 nucleotides

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
A	C	G	C	G	T	A	A	C	T	G	C	C	G	T	A
A	C	G	A	G	T	T	A	C	T	G	G	C	G	A	A
A	C	G	A	G	T	A	A	C	T	G	G	C	G	A	A
A	C	G	C	G	T	A	A	C	T	G	C	C	G	T	A

- heterozygosity

Describing genetic variation

- Assessing the level of genetic variation

Sequence alignment of four DNA sequences consisting of 16 nucleotides

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
A	C	G	C	G	T	A	A	C	T	G	C	C	G	T	A
A	C	G	A	G	T	T	A	C	T	G	G	C	G	A	A
A	C	G	A	G	T	A	A	C	T	G	G	C	G	A	A
A	C	G	C	G	T	A	A	C	T	G	C	C	G	T	A

- heterozygosity at a site *j* (site heterozygosity)

Describing genetic variation

- Assessing the level of genetic variation

Sequence alignment of four DNA sequences consisting of 16 nucleotides

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
A	C	G	C	G	T	A	A	C	T	G	C	C	G	T	A
A	C	G	A	G	T	T	A	C	T	G	G	C	G	A	A
A	C	G	A	G	T	A	A	C	T	G	G	C	G	A	A
A	C	G	C	G	T	A	A	C	T	G	C	C	G	T	A

- heterozygosity at a site j (site heterozygosity)
- $H_j = \frac{n}{n-1}\left(1 - \sum_i p_i^2\right)$, where n represents the number of sequences/samples and p_i the frequency of the ith allele at site j

Describing genetic variation

- Assessing the level of genetic variation

Sequence alignment of four DNA sequences consisting of 16 nucleotides

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
A	C	G	C	G	T	A	A	C	T	G	C	C	G	T	A
A	C	G	A	G	T	T	A	C	T	G	G	C	G	A	A
A	C	G	A	G	T	A	A	C	T	G	G	C	G	A	A
A	C	G	C	G	T	A	A	C	T	G	C	C	G	T	A

- heterozygosity at a site j (site heterozygosity)
- $H_j = \frac{n}{n-1}\left(1 - \sum_i p_i^2\right)$, where n represents the number of sequences/samples and p_i the frequency of the ith allele at site j; $H_7 = \frac{4}{3}\left(1 - \frac{1}{16} - \frac{9}{16}\right) = \frac{4}{3} \cdot \frac{6}{16} = \frac{2}{4} = 0.5$; $H_{12} = \frac{4}{3}\left(1 - \frac{1}{4} - \frac{1}{4}\right) = \frac{4}{3} \cdot \frac{2}{4} = \frac{2}{3} = 0.67$

Describing genetic variation

- Assessing the level of genetic variation

Sequence alignment of four DNA sequences consisting of 16 nucleotides

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
A	C	G	C	G	T	A	A	C	T	G	C	C	G	T	A
A	C	G	A	G	T	T	A	C	T	G	G	C	G	A	A
A	C	G	A	G	T	A	A	C	T	G	G	C	G	A	A
A	C	G	C	G	T	A	A	C	T	G	C	C	G	T	A

- nucleotide diversity

Describing genetic variation

- Assessing the level of genetic variation

Sequence alignment of four DNA sequences consisting of 16 nucleotides

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
A	C	G	C	G	T	A	A	C	T	G	C	C	G	T	A
A	C	G	A	G	T	T	A	C	T	G	G	C	G	A	A
A	C	G	A	G	T	A	A	C	T	G	G	C	G	A	A
A	C	G	C	G	T	A	A	C	T	G	C	C	G	T	A

- nucleotide diversity is the average site heterozygosity over all sites in the sequence

Describing genetic variation

- Assessing the level of genetic variation

Sequence alignment of four DNA sequences consisting of 16 nucleotides

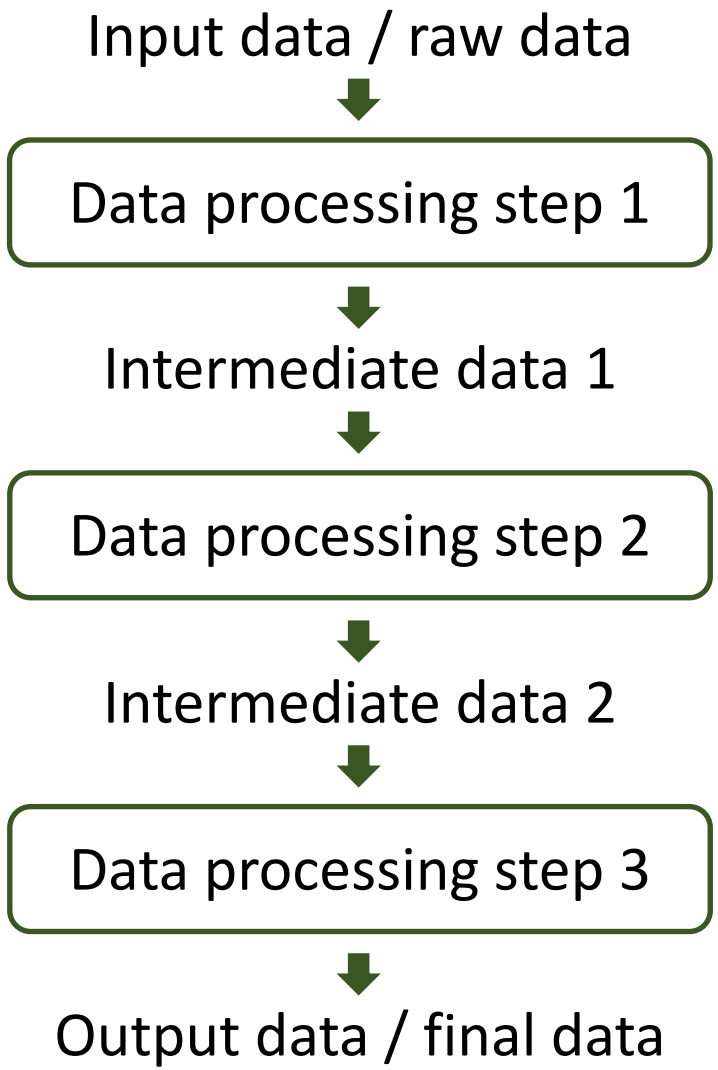
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
A	C	G	C	G	T	A	A	C	T	G	C	C	G	T	A
A	C	G	A	G	T	T	A	C	T	G	G	C	G	A	A
A	C	G	A	G	T	A	A	C	T	G	G	C	G	A	A
A	C	G	C	G	T	A	A	C	T	G	C	C	G	T	A

- nucleotide diversity is the average site heterozygosity over all sites in the sequence
- $\pi = \frac{1}{L}\sum_{j}^{L} H_j$, where L is the number of sites in the sequence, here $L = 16$

→ π = (0.667+0.500+0.667+0.667)/16 = 0.156

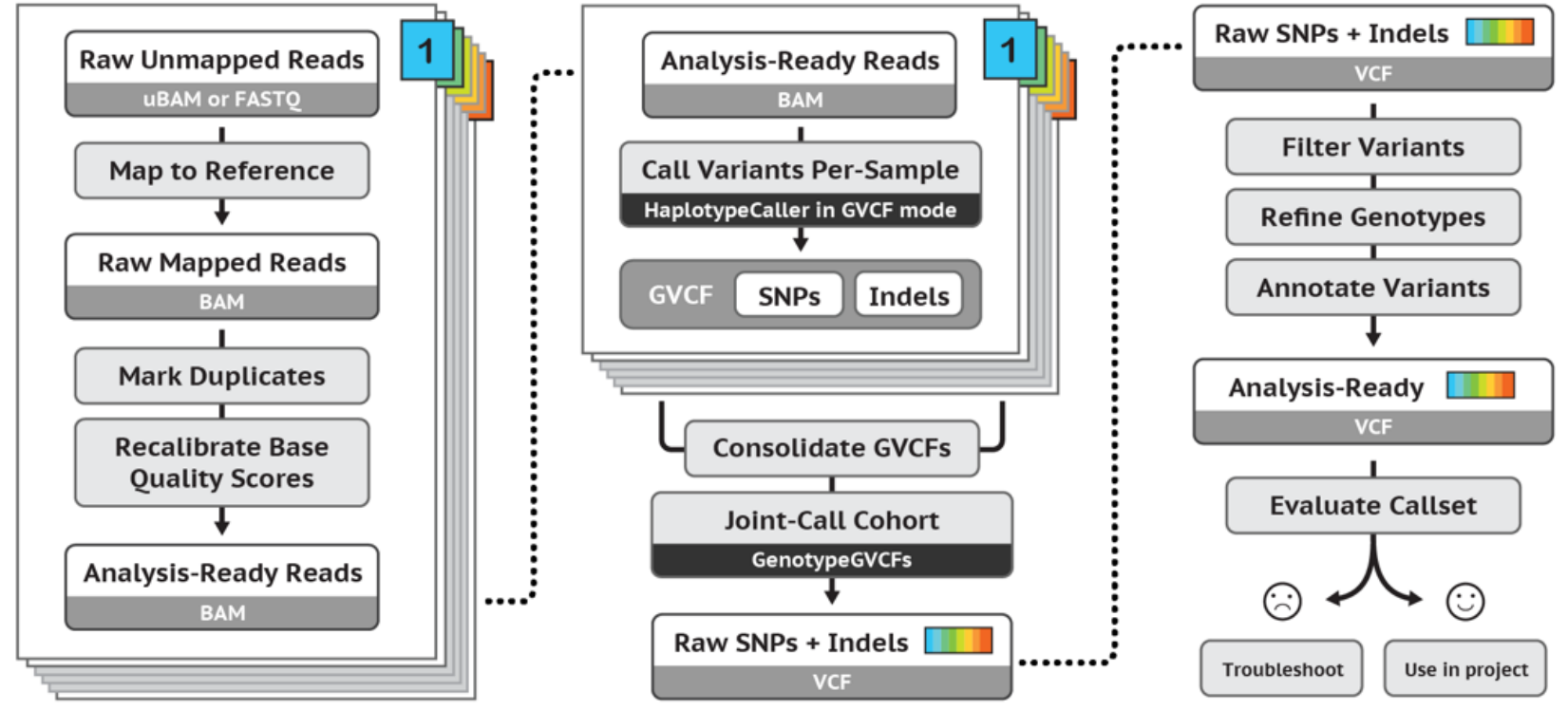
Table of contents

What is a workflow?



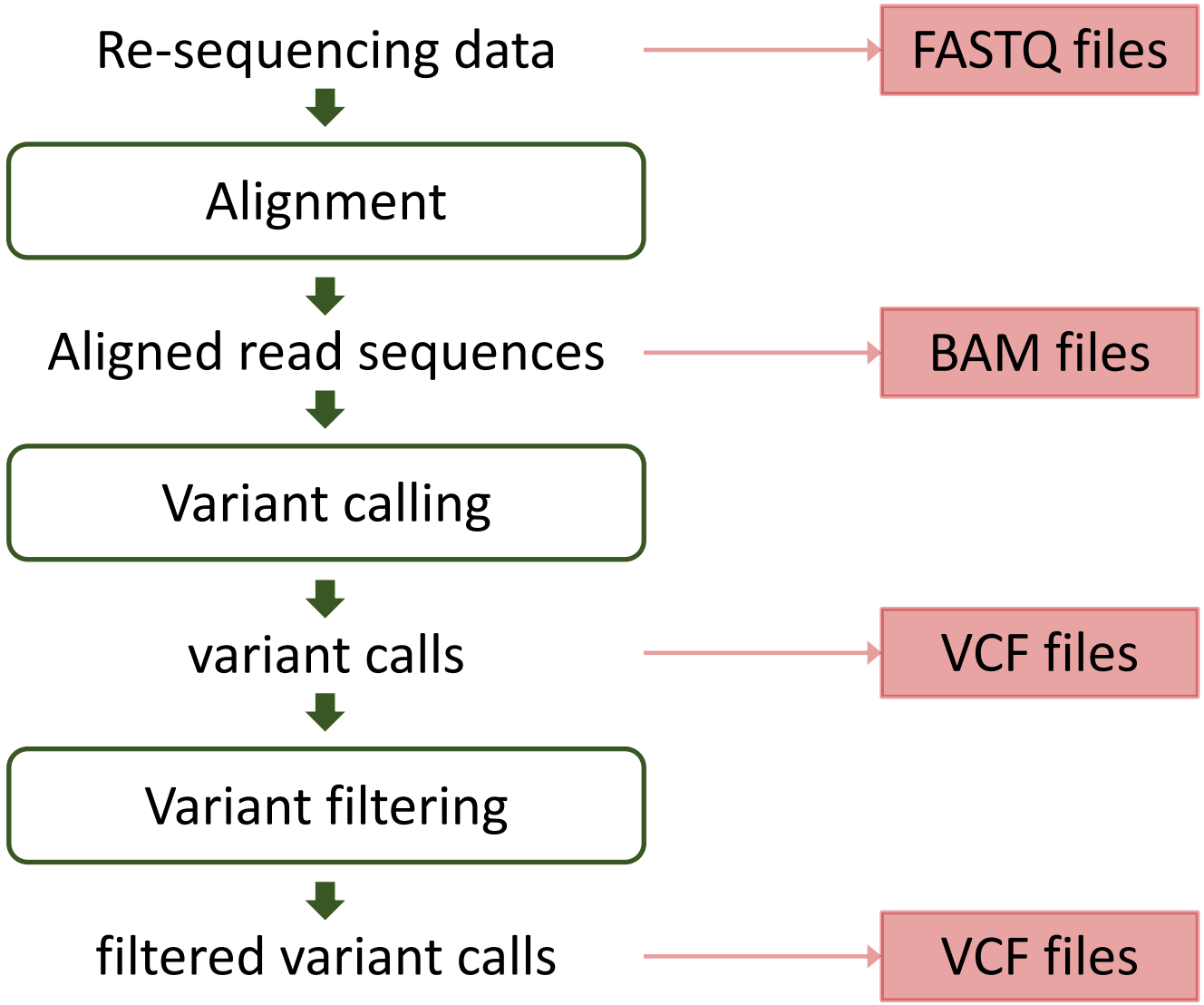
SNP calling workflow

https://gatk.broadinstitute.org



Best Practices for SNP and Indel discovery in germline DNA - leveraging groundbreaking methods for combined power and scalability.

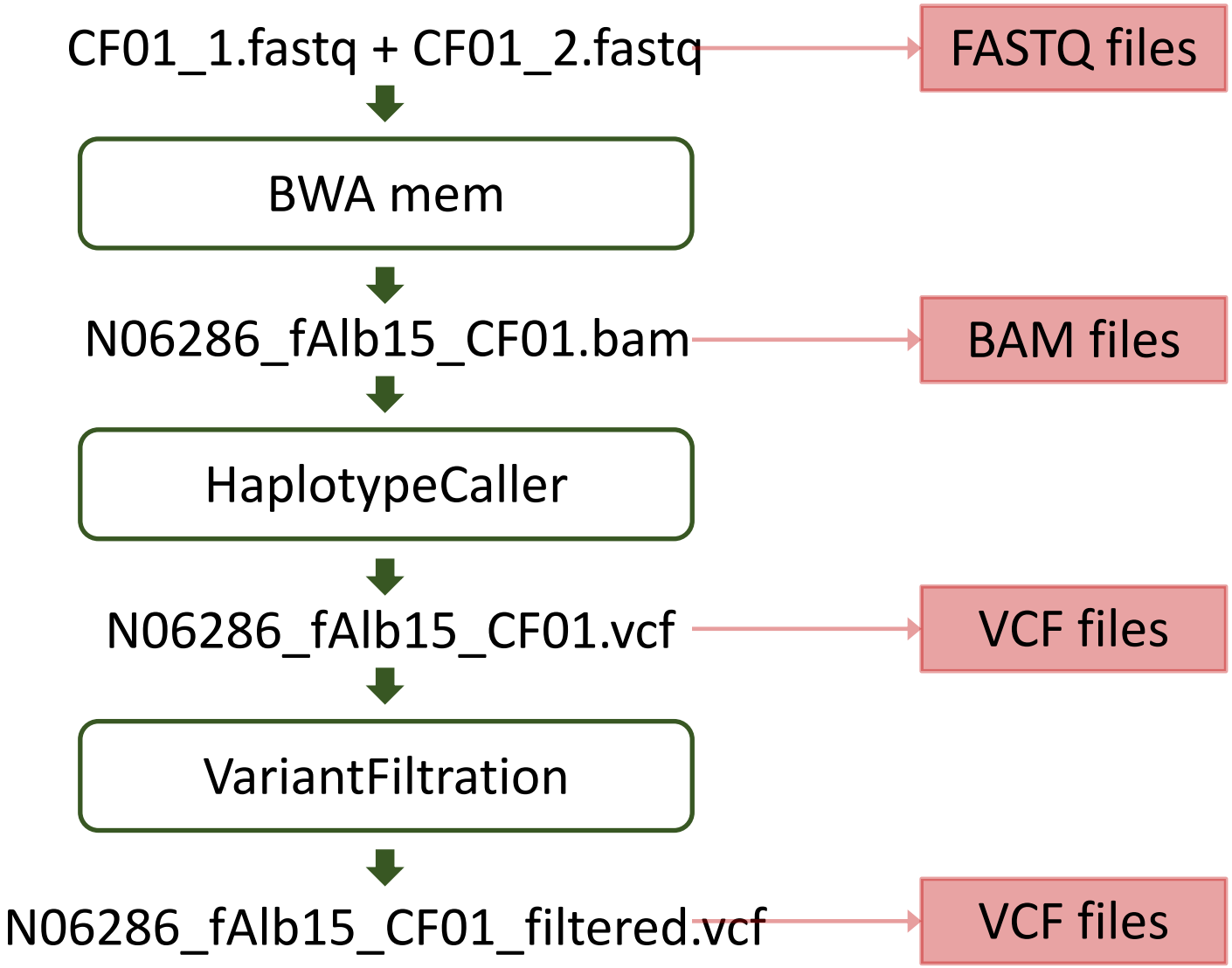
Basic workflow, one example



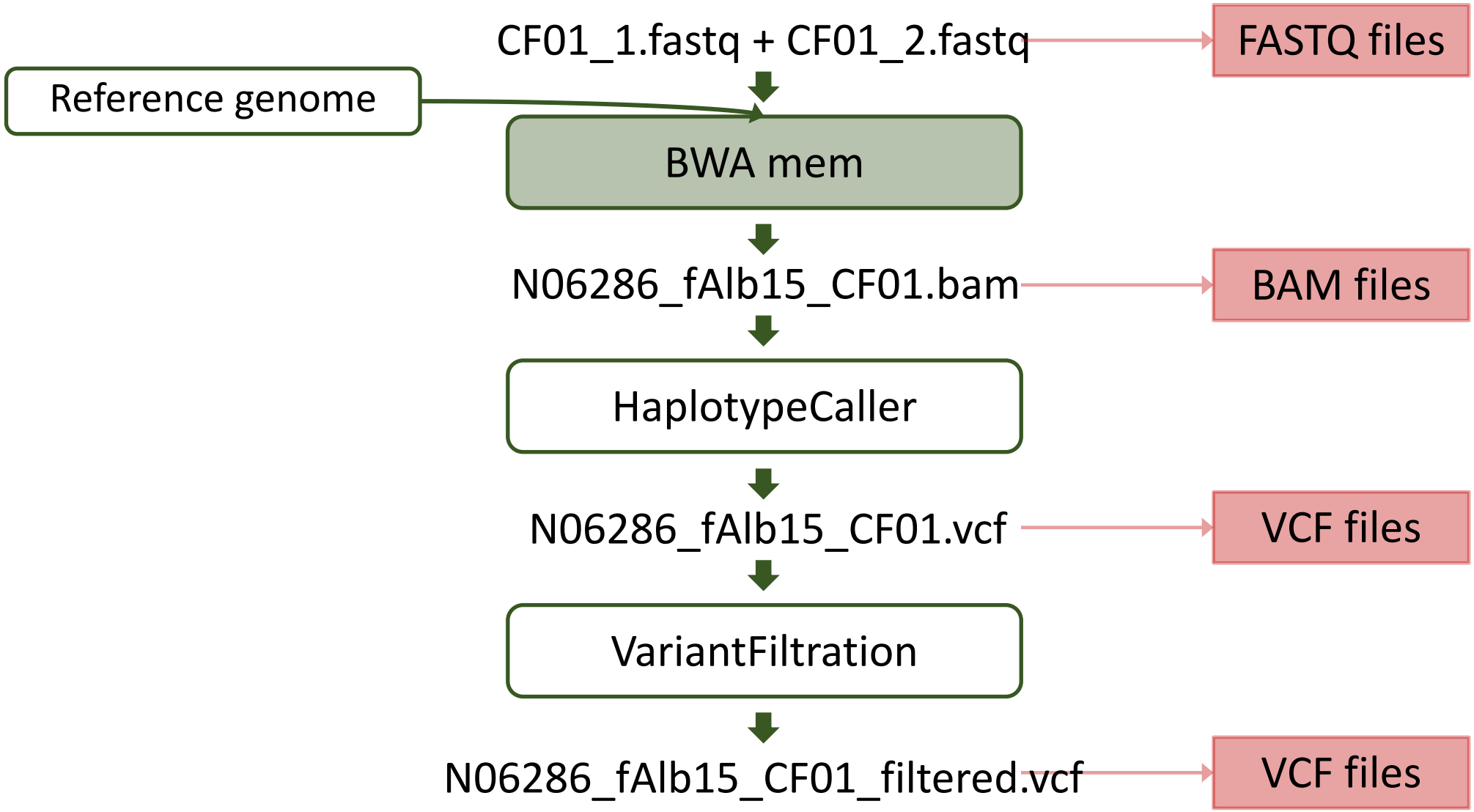
Workflow conventions

- Create a new output file in each processing step
  - Don’t overwrite the input file!
- Use informative file names
  - include information about the sample(s) and eventual other input data
  - include information about the processing step
  - Use the correct file extensions (.fastq, .bam, .vcf, …)
- Allocate appropriate computing resources

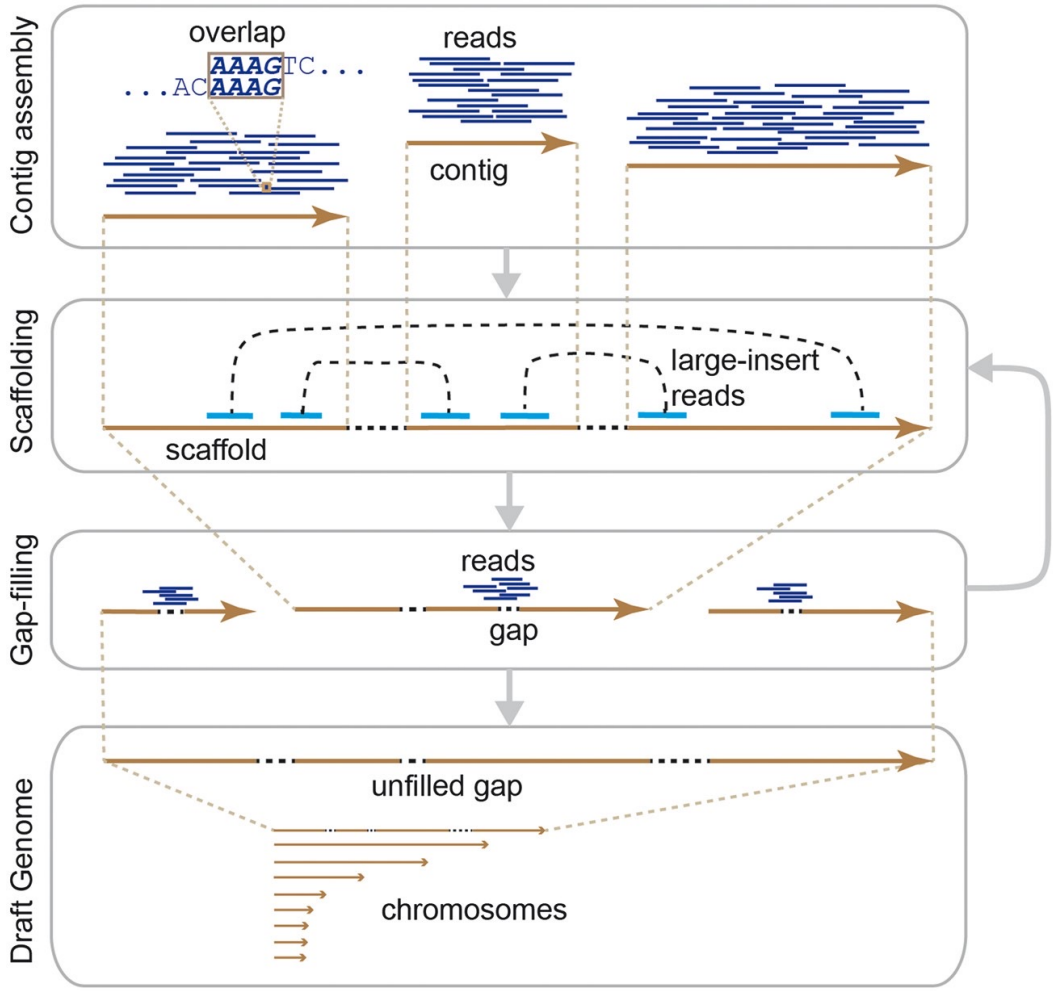
Basic variant calling workflow, one sample



Basic variant calling workflow, one sample



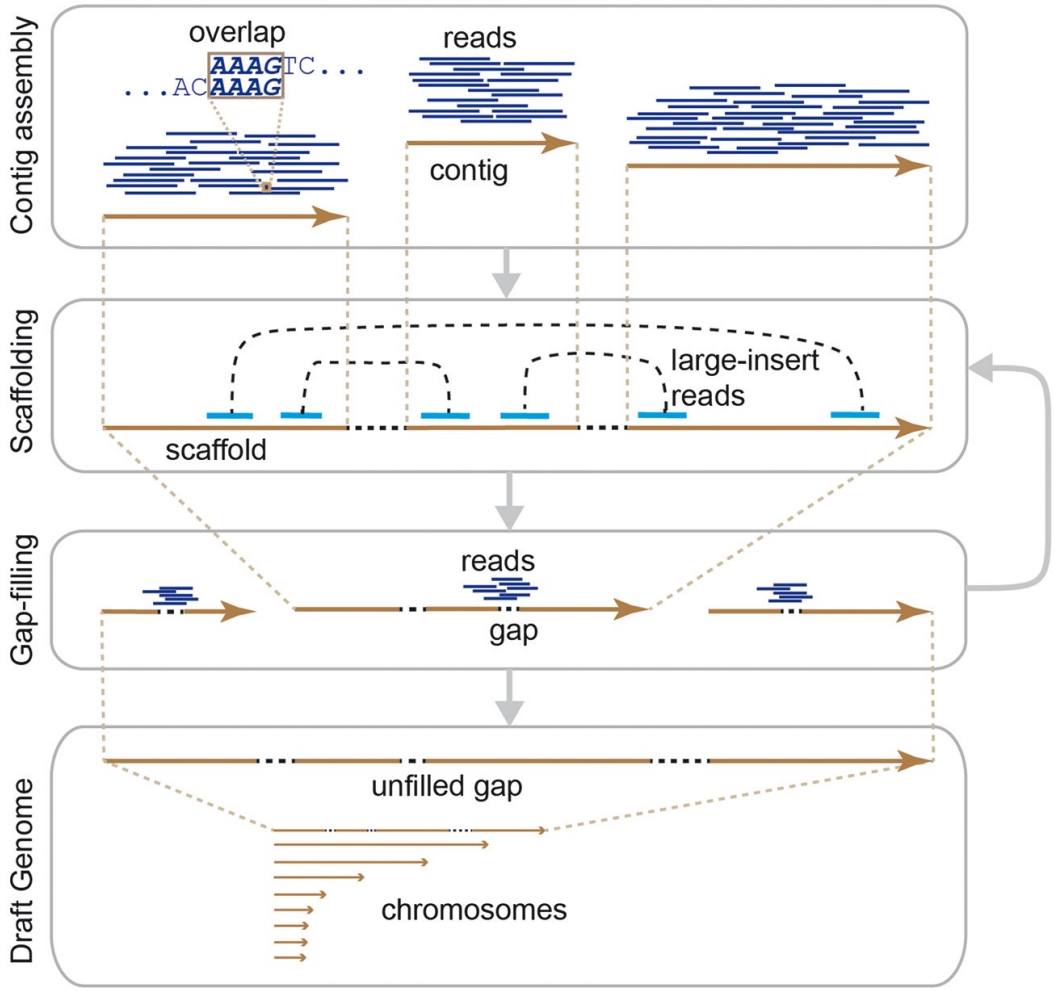
Reference genome



- The reference genome represents a **template genome sequence** of a species, typically the target species or a closely related species
- The reference genome covers those parts of the genome sequence that have been assembled and usually **includes several gaps** and may contain **misassembled regions**
- The reference genome can be assembled at the **scaffold-level** or at the **chromosome-level**

Sohn & Nam 2016 *Brief. Bioinform.*

Reference genome – alignment quality



- The **quality and contiguity of reference genome assemblies** influence the alignment quality
- Alignment of reads to a **divergent reference genome** influences the alignment quality
- The proportion of **repetitive DNA sequences** in the genome influences the alignment quality
- **Structural re-arrangements** among the genomes of sampled individuals and the reference genome influence the alignment quality

Sohn & Nam 2016 *Brief. Bioinform.*

Alignment

ACGTTTGCGTCCCGCCCGATNNNNNN----------CGTAGTCGGGGTATGTAGNNGATTCTCTCAGT

TCGGCGTATGTGGCGGATTCTCT

ATGTCTCG---TGTAGATCCG

Alignment

ACGTTTGCGTCCCGCCCGATNNNNNN----------CGTAGTCGGGGTATGTAGNNGATTCTCTCAGT

TCGGCGTATGTGGCGGATTCTCT

ATGTCTCG---TGTAGATCCG

Can we trust the alignment of the second read?

Alignment – Burrows-Wheeler Aligner (BWA)

- BWA is a software package for mapping low-divergent short-read sequences against a large reference genome
  - https://bio-bwa.sourceforge.net/
- BWA-MEM is the latest version and supports split alignment and is generally recommended for high-quality read sequences
- The output from read mapping is a SAM format
- The BAM file is a binary representation of the SAM file

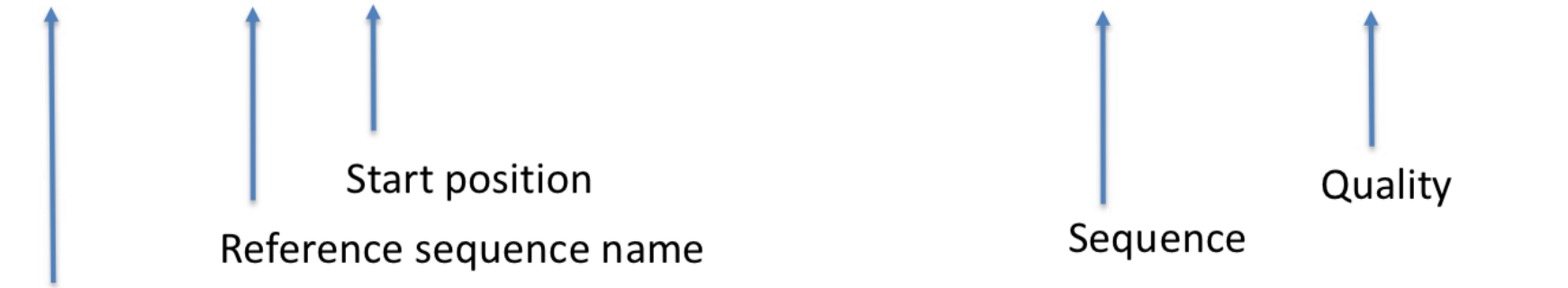
Sequence Alignment/Map (SAM) file

HEADER SECTION

```
@HD    VN:1.6SO:coordinate
@SQ    SN:2   LN:243199373
@PG    ID:bwaPN:bwaVN:0.7.17-r1188    CL:bwa mem -t 1 human_g1k_v37_chr2.fasta HG00097_1.fq HG00097_2.fq
@PG    ID:samtools PN:samtools PP:bwaVN:1.10      CL:samtools sort
@PG    ID:samtools.1      PN:samtools PP:samtools VN:1.10      CL:samtools view -H HG00097.bam
```

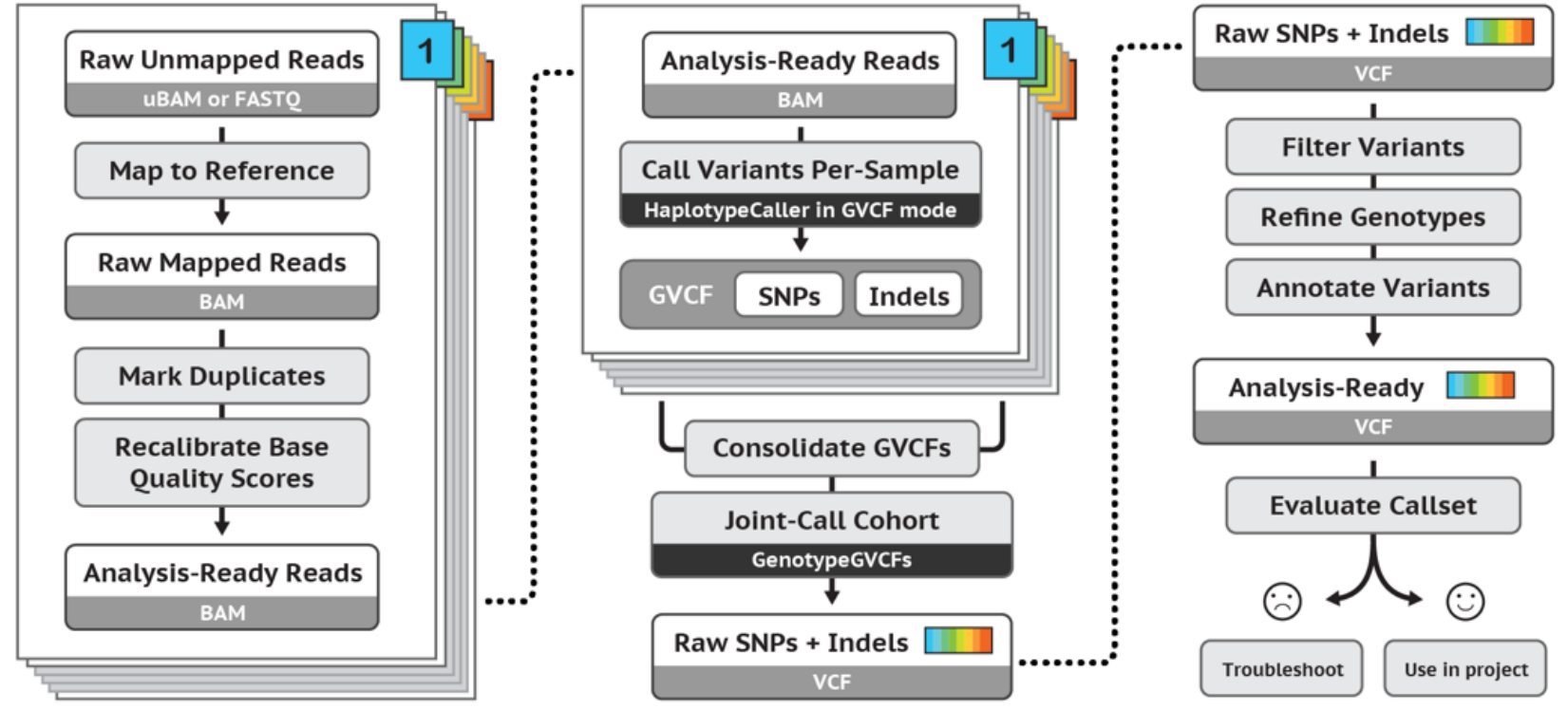
ALIGNMENT SECTION

```
Read_001      99      2      3843448      0      101M    =      3843625      278     TTTGGTTCCATATGAACTTT        0F<BFB<FFFBFBFFFBFBB
Read_001      147     2      3843625      0      101M    =      3843448      -278    TTATTTCATTGAGCAGTGGT        FBBI7IIFIB<BBBB<BBFF
Read_002      163     2      4210055      0      101M    =      4210377      423     TGGTACCAAAACAGAGATAT        0IIFBFFFIIIFFIFFFBBF
Read_003      99      2      4210066      0      101M    =      4210317      352     CAGAGATATAGATCAATGGA        0IIFFFIFFFIFIFIIIIIF
```



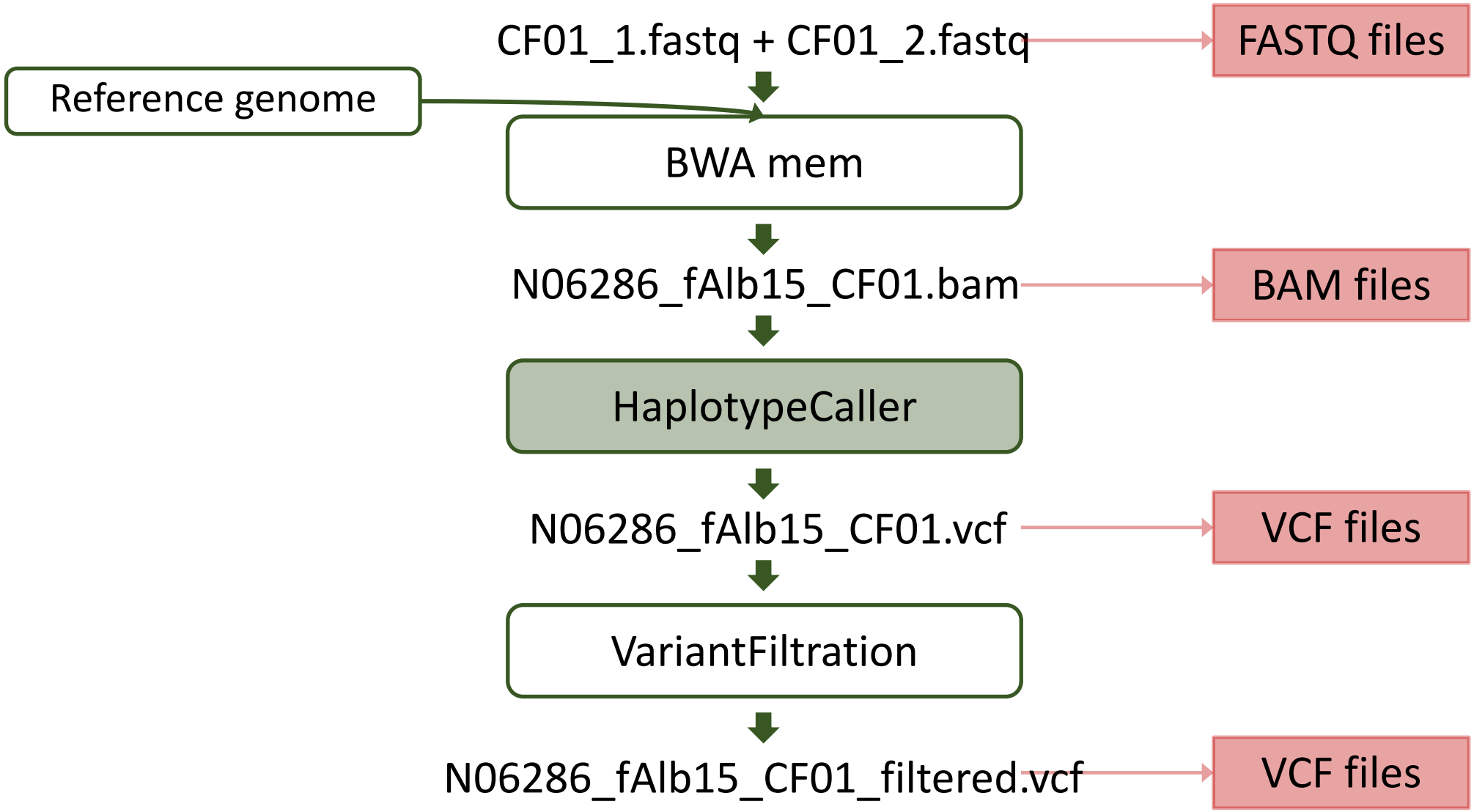
SNP calling workflow

https://gatk.broadinstitute.org



Best Practices for SNP and Indel discovery in germline DNA - leveraging groundbreaking methods for combined power and scalability.

Basic variant calling workflow, one sample



Detecting variants in reads

Reference: ACGTTTGCGTCCCGCCCGATNNNNNN----------CGTAGTCGGGGTATGTAGNNGATTCTCTCAGT

Samples:

…TCGGCGTATGTGGCGGATTCTCT…

…TCGGGGTATGTAGCGGATTCTCT …

…TCGGCGTATGTGGCGGATTCTCT…

…TCGGGGTATGTAGCGGATTCTCT …

…TCGGGGTATGTGGCGGATTCTCT …

…TCGGCGTATGTGGCGGATTCTCT…

…TCGGGGTATGTAGCGGATTCTCT …

…TCGGGGTATGTAGCGGATTCTCT …

GGGGTATGTGGCGGATTCTCT…

…TCGGGGTATGTGGCGGATTCTCT…

Reference and alternative alleles

Reference: ACGTTTGCGTCCCGCCCGATNNNNNN----------CGTAGTCGGGGTATGTAGNNGATTCTCTCAGT

Samples:

…TCGGCGTATGTGGCGGATTCTCT…

…TCGGGGTATGTAGCGGATTCTCT …

…TCGGCGTATGTGGCGGATTCTCT…

…TCGGGGTATGTAGCGGATTCTCT …

…TCGGGGTATGTGGCGGATTCTCT …

…TCGGCGTATGTGGCGGATTCTCT…

…TCGGGGTATGTAGCGGATTCTCT …

…TCGGGGTATGTAGCGGATTCTCT …

GGGGTATGTGGCGGATTCTCT…

…TCGGGGTATGTGGCGGATTCTCT…

Reference allele: the allele in the reference genome	G	A
Alternative allele: the allele NOT in the reference genome	C	G

Variant call format (VCF) file

- The variant call format (VCF) file consists of a header and a list of variant call records

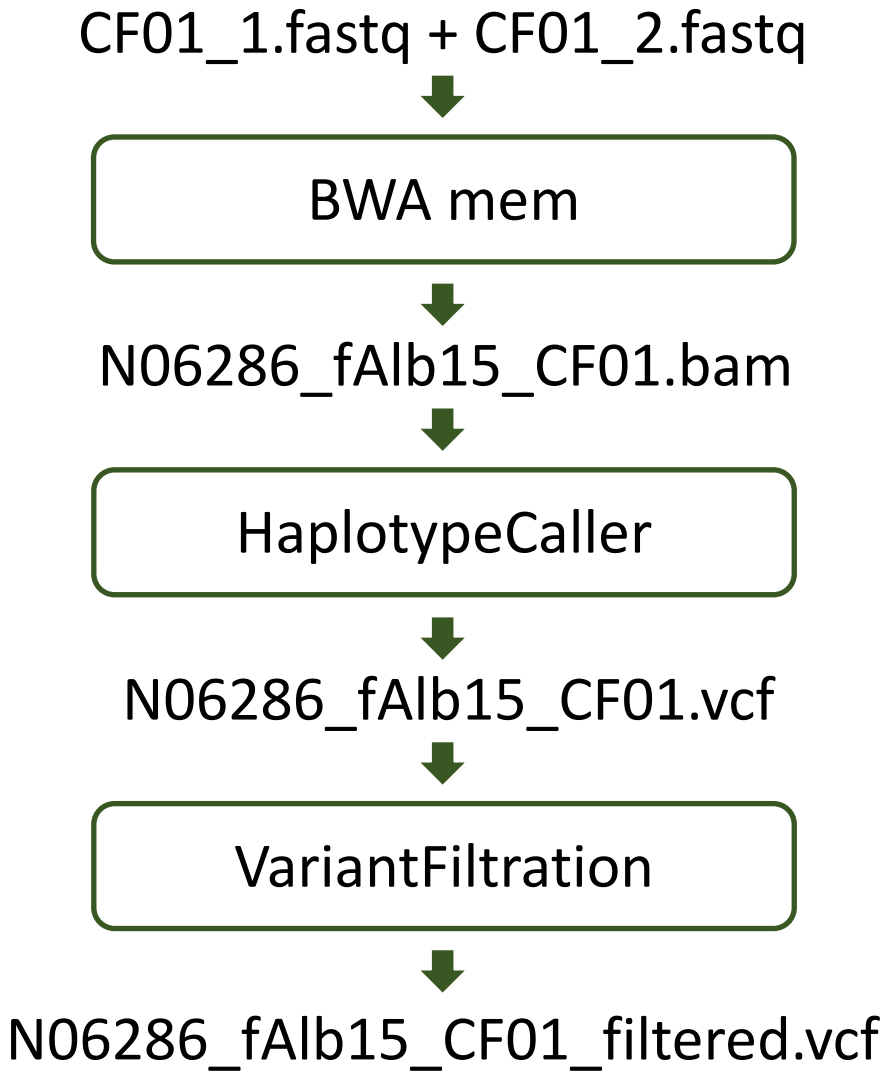
```
##fileformat=VCFv4.2
##ALT=<ID=NON_REF,Description="Represents any possible alternative allele not already represented at this location by REF and ALT">
##FILTER=<ID=LowQual,Description="Low quality">
##FILTER=<ID=PASS,Description="All filters passed">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic 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 filtered)">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=…
##GATKCommandLine= …
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the same order as listed">
##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 qualities">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth; some reads may have been filtered">
##INFO=<ID=…
##contig=<ID=N00001,length=26618703>
##source=HaplotypeCaller
#CHROM  POS    ID     REF    ALT    QUAL    FILTER  INFO    FORMAT  ATL_FSP08-001_M
N00001  14     .      G      A      2886.43 .       AC=30;AF=0.063;AN=478;BaseQRankSum=1.28;DP=1099;…      GT:AD:DP:GQ:PGT:PID:PL:PS       0/0:5,0:5:15:.:.:0,15,134
```

Variant call format (VCF) file

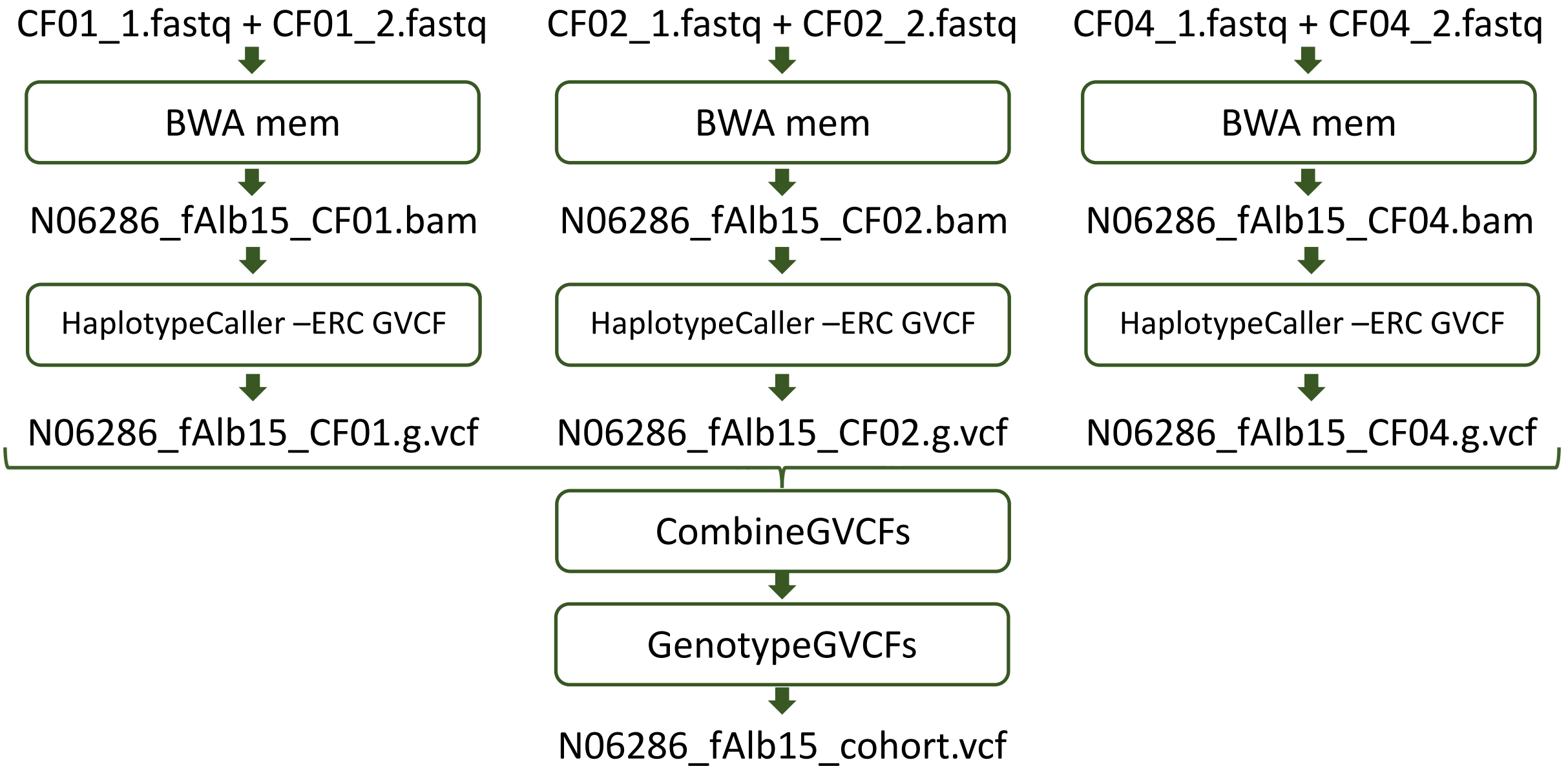
- The variant call format (VCF) file consists of a header and a list of variant call records

```
##fileformat=VCFv4.2
##ALT=<ID=NON_REF,Description="Represents any possible alternative allele not already represented at this location by REF and ALT">
##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 filtered)">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=…
##GATKCommandLine= …
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the same order as listed">
##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 qualities">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth; some reads may have been filtered">
##INFO=<ID=…
##contig=<ID=N00001,length=26618703>
##source=HaplotypeCaller
#CHROM  POS    ID     REF    ALT    QUAL    FILTER  INFO    FORMAT  CF01
N00001  14     .      G      A      2886.43 .       AC=30;AF=0.063;AN=478;BaseQRankSum=1.28;DP=1099;…      GT:AD:DP:GQ:PGT:PID:PL:PS       0/0:5,0:5:15:.:.:0,15,134
```

Basic variant calling workflow, one sample

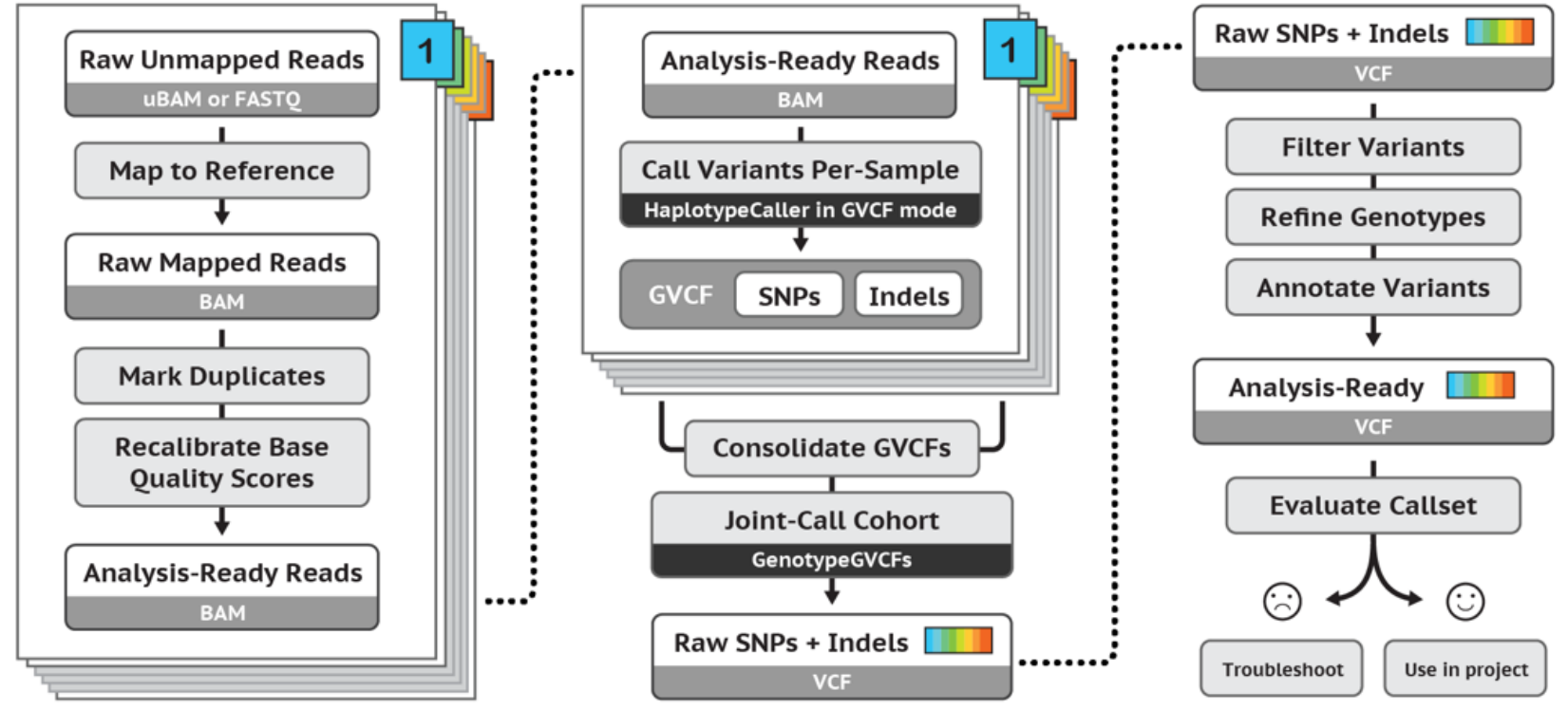


Basic variant calling workflow in cohort



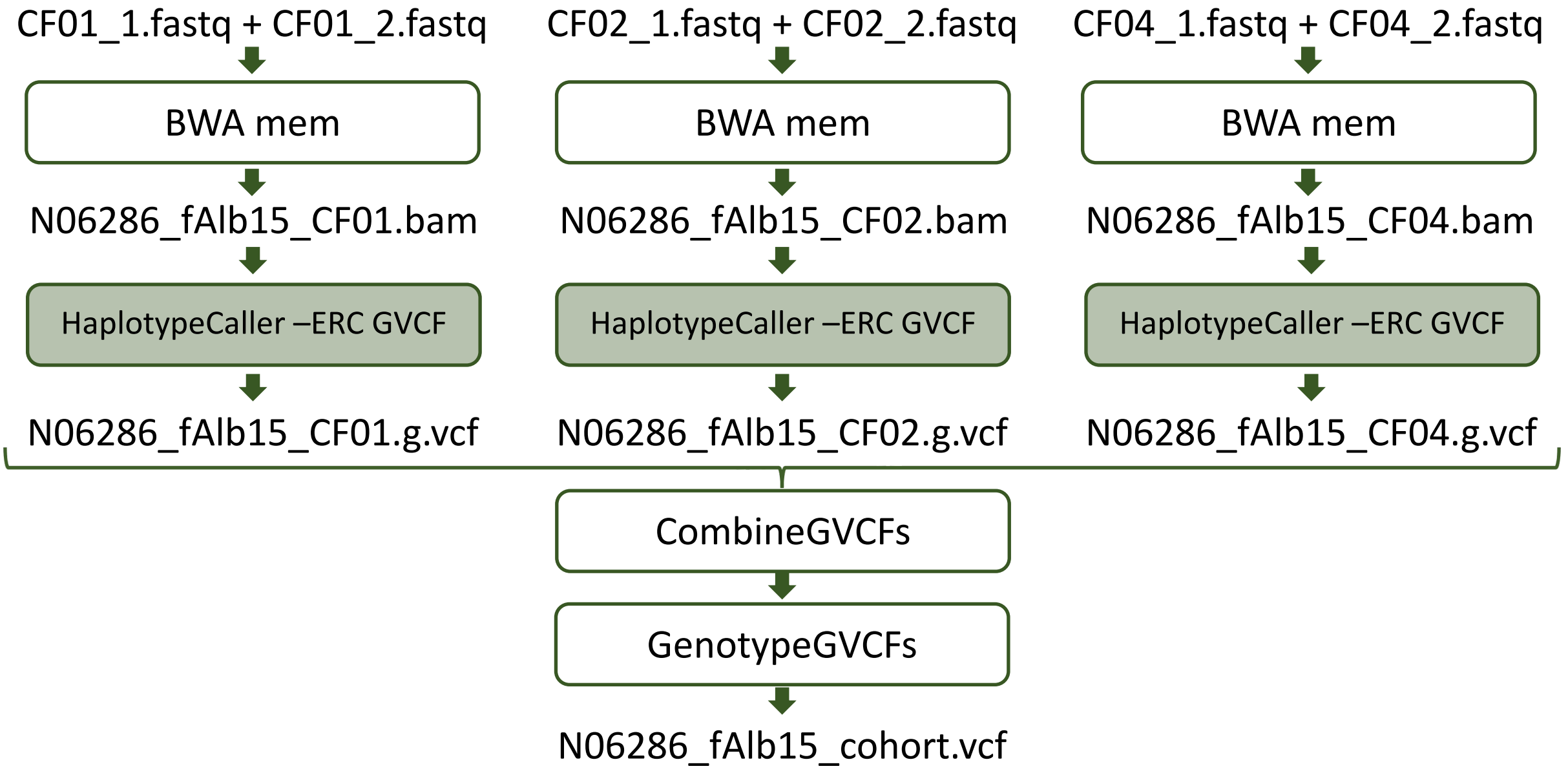
SNP calling workflow

https://gatk.broadinstitute.org



Best Practices for SNP and Indel discovery in germline DNA - leveraging groundbreaking methods for combined power and scalability.

Basic variant calling workflow in cohort



Difference between a GVCF and a VCF file

Regular VCF file

##fileformat
##ALT
##FILTER
##FORMAT
##GATKCommandLine
##INFO
##contig
##source

#record header
variant call records

GVCF file

##fileformat
##ALT
##FILTER
##FORMAT
##GATKCommandLine
##GVCFBlock
##INFO
##contig
##source

#record header
non-variant block records
variant call records

- A GVCF file has records for all sites, whether there is a variant call or not
- Adjacent non-variant sites merged into blocks

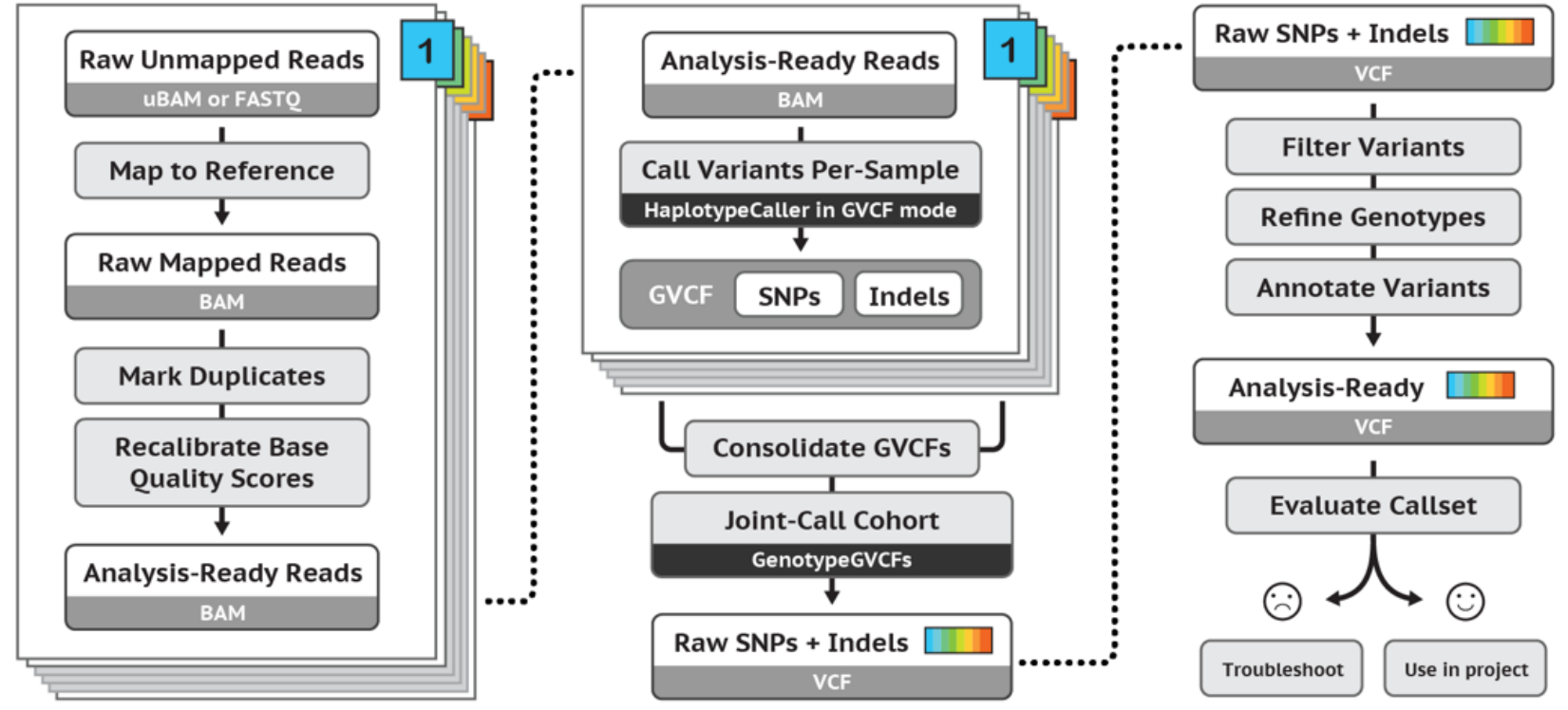
Variant call format (VCF) file for a cohort

- The variant call format (VCF) file consists of a header and a list of variant call records

```
##fileformat=VCFv4.2
##ALT=<ID=NON_REF,Description="Represents any possible alternative allele not already represented at this location by REF and ALT">
##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 filtered)">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=…
##GATKCommandLine= …
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the same order as listed">
##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 qualities">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth; some reads may have been filtered">
##INFO=<ID=…
##contig=<ID=N00001,length=26618703>
##source=GenomicsDBImport
##source=GenotypeGVCFs
##source=HaplotypeCaller
#CHROM  POS    ID     REF    ALT    QUAL   FILTER  INFO   FORMAT  CF01  CF02  CF04
```

SNP calling workflow

https://gatk.broadinstitute.org



Best Practices for SNP and Indel discovery in germline DNA - leveraging groundbreaking methods for combined power and scalability.

Variant filtering criteria

There are two recommended best practices for variant call filtering

- Variant quality score recalibration (VQSR)
  - VQSR is a machine learning algorithm than can be trained to recognize likely false variant calls
  - VQSR requires an input of likely true variant calls, it's application is thus limited to model organisms, but recommended if possible
- GATK hard filters
  - Filters based on information contained in the VCF

https://gatk.broadinstitute.org/hc/en-us/articles/360035531112--How-to-Filter-variants-either-with-VQSR-or-by-hard-filtering

GATK hard filters

- The variant call format (VCF) file consists of a header and a list of variant call records

```
##fileformat=VCFv4.2
##ALT=<ID=NON_REF,Description="Represents any possible alternative allele not already represented at this location by REF and ALT">
##FILTER=<ID=LowQual,Description="Low quality">
##FILTER=<ID=PASS,Description="All filters passed">
##FILTER=<ID=hard_filt,Description="QD < 2.0 || FS > 60.0 || MQ < 40.0 || MQRankSum < -12.5 || StrandOddsRatio > 3 || ReadPosRankSum < -8.0">
##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 filtered)">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=…
##GATKCommandLine= …
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the same order as listed">
##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 qualities">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth; some reads may have been filtered">
##INFO=<ID=…
##contig=<ID=N00001,length=26618703>
##source=GenomicsDBImport
##source=GenotypeGVCFs
##source=HaplotypeCaller
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  CF01  CF02  CF04
```

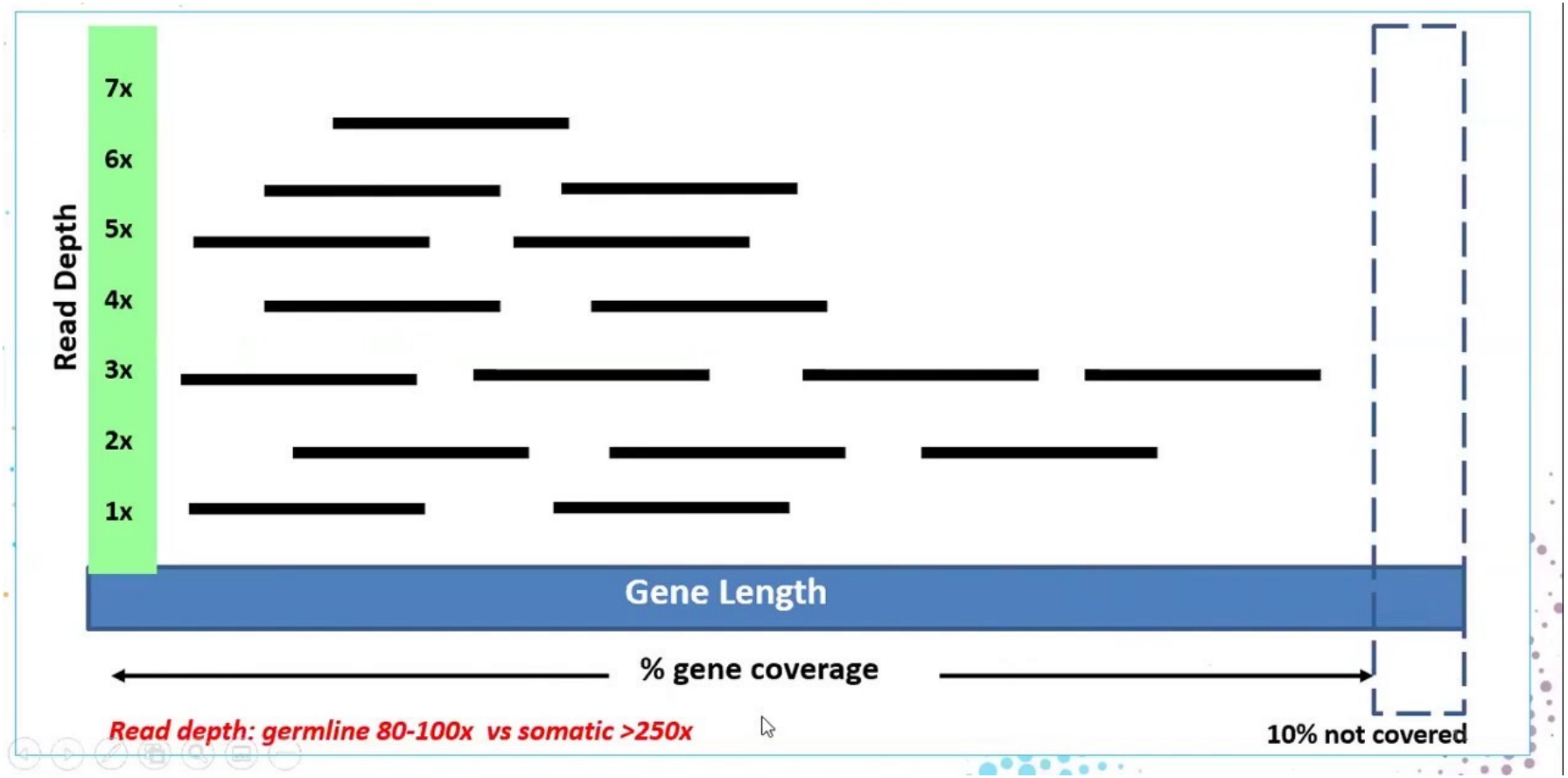
Additional variant filtering criteria

- In addition to the basic filtering steps, filtering adjusted to the study organism is recommended
- **Remember!**
- The quality and contiguity of reference genome assemblies influence the alignment and variant calling quality
- Alignment of reads to a divergent reference genome influences the alignment and variant calling quality
- The proportion of repetitive DNA sequences in the genome influences the alignment and variant calling quality
- Structural re-arrangements, such as CNVs, among the genomes of sampled individuals and the reference genome influence the alignment and variant calling quality

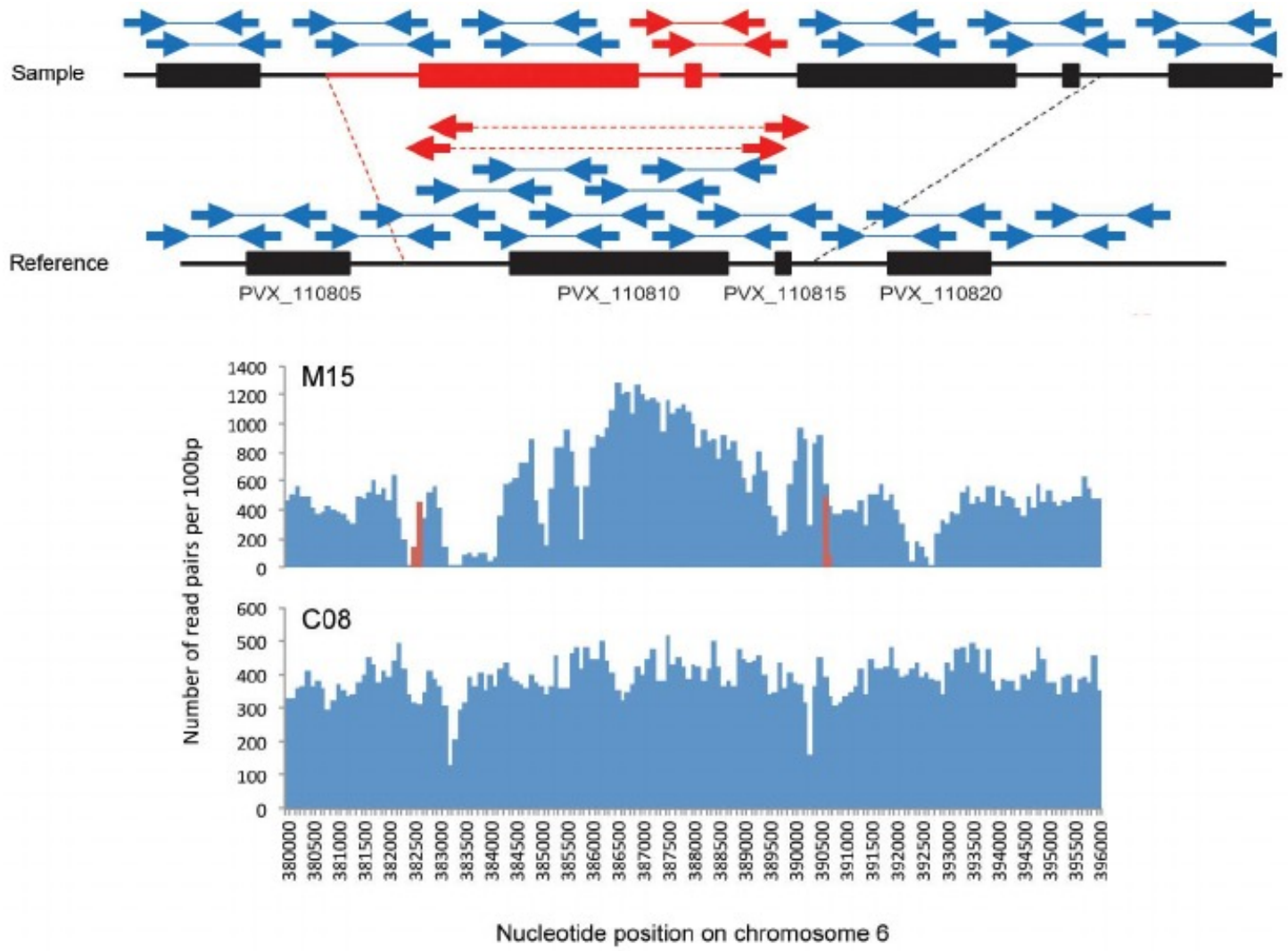
Additional variant filtering criteria

- Remove indels (GATK)
- Keep only mono-allelic and bi-allelic sites (GATK)
- Remove sites overlapping repetitive regions (VCFtools)
- Remove sites with extreme coverage values (VCFtools)
- Apply quality score filtering (VCFtools)
- Identify and remove sites overlapping with copy number variants
- …

Depth versus Coverage

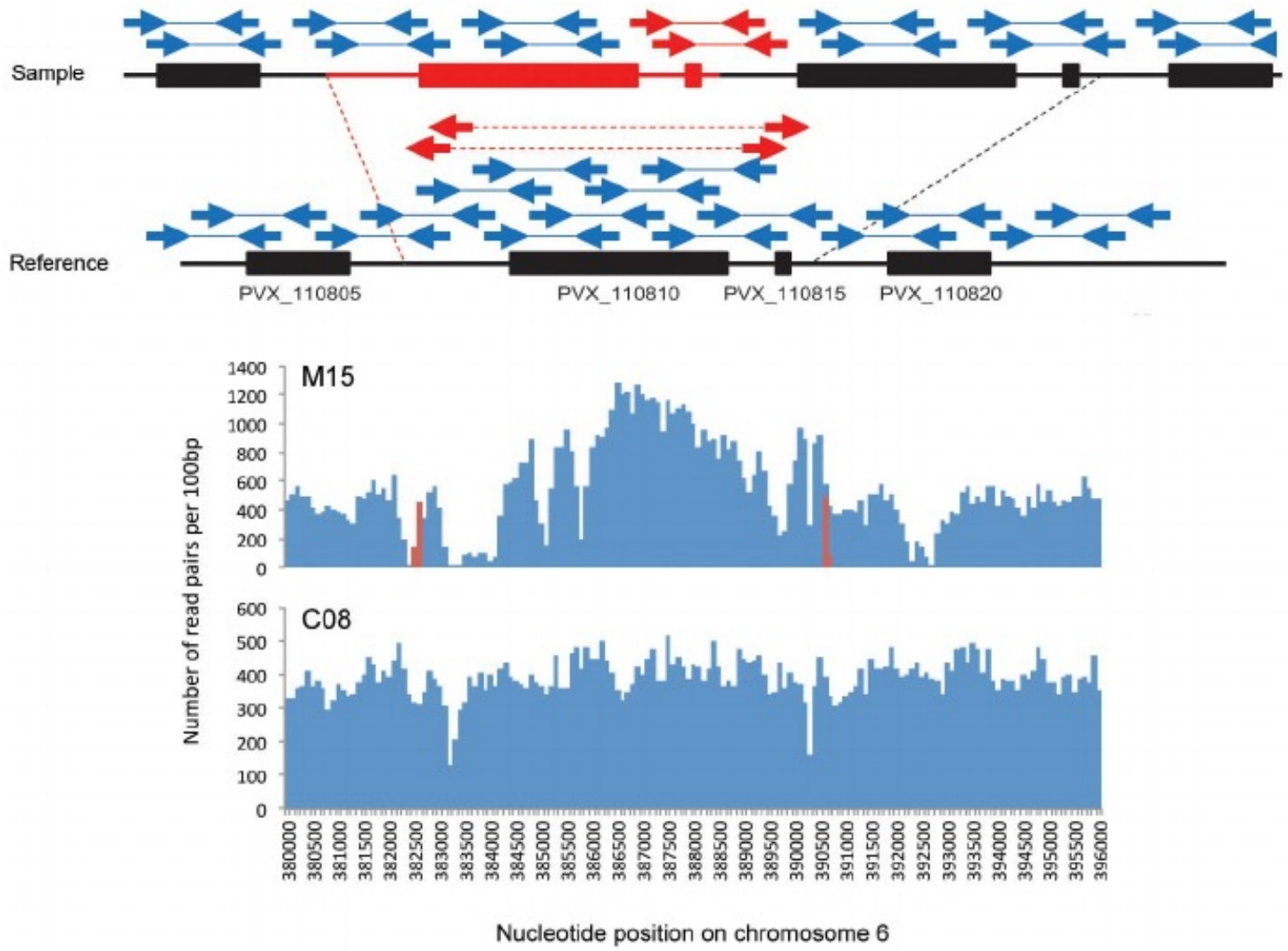


Mismapping



Due to

Mismapping



Due to errors in assembly or variation among individuals

Result in mismapping and collapsed regions

Result in variation of coverage

Table of contents

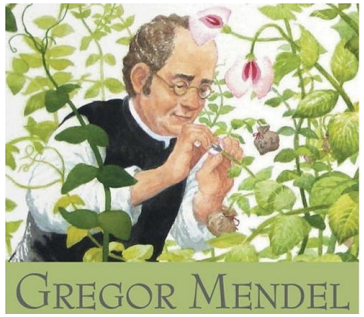
Evolution can be seen as simply a consequence of these conditions…

Variation

Individuals vary in traits that govern reproduction and survival…

…and resources are not endless such that there is competition and thus selection…

Selection



Heritability

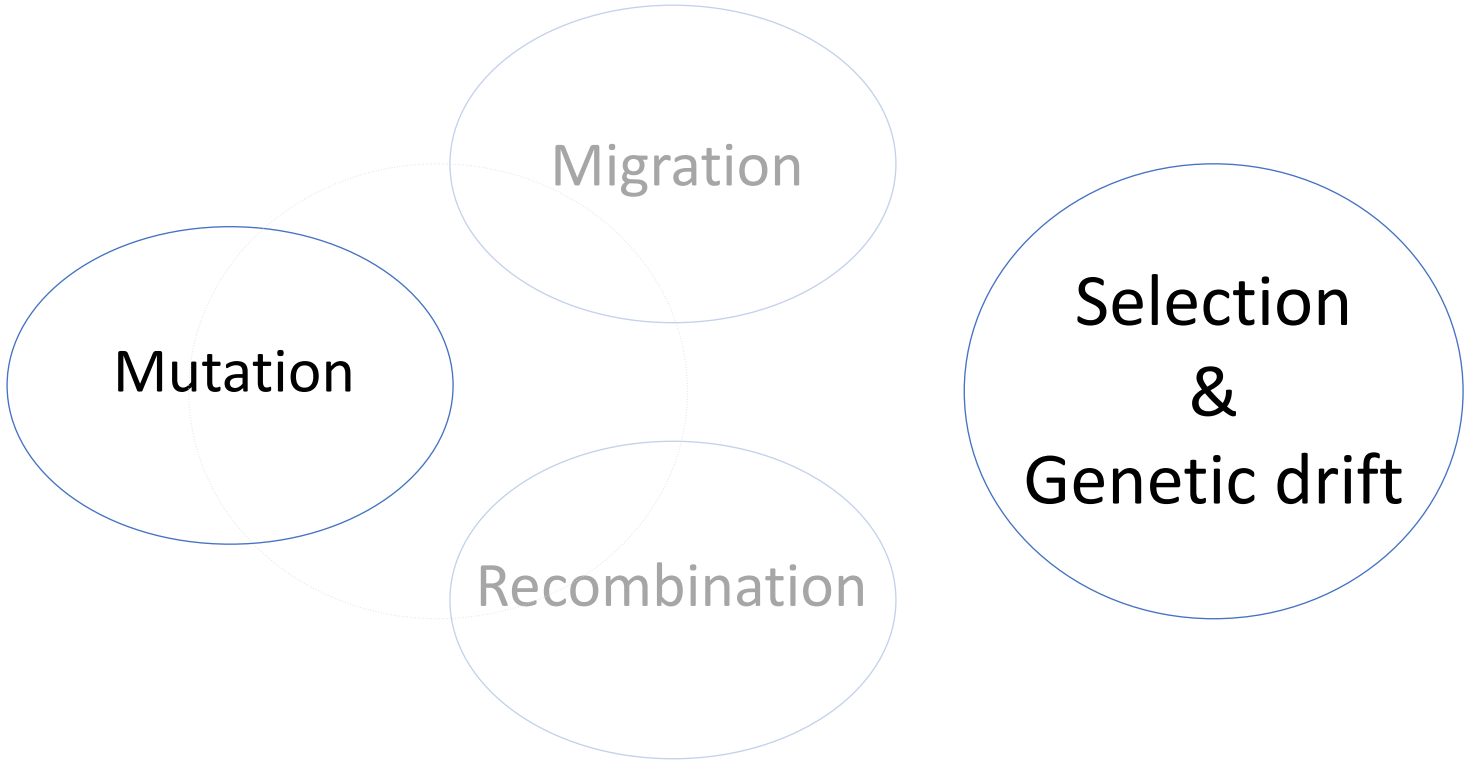
…and traits important to survival and reproduction are genetically controlled and inherited, then…

Evolution

Genetic variation

Selection & Genetic drift

Evolution


Migration
Mutation
Recombination
Selection
&
Genetic drift

Applications in ecology and evolution

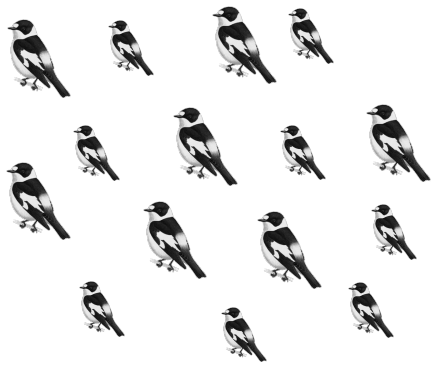
- Central questions in evolutionary genetics
  - How are changes in the genome generated?
  - Why is the genome changing over time?

Applications in ecology and evolution

- Central questions in evolutionary genetics
  - How are changes in the genome generated?
  - Why is the genome changing over time?

Evolution is a process influenced by

- mutation
- genetic drift
- natural selection
- demography
- recombination

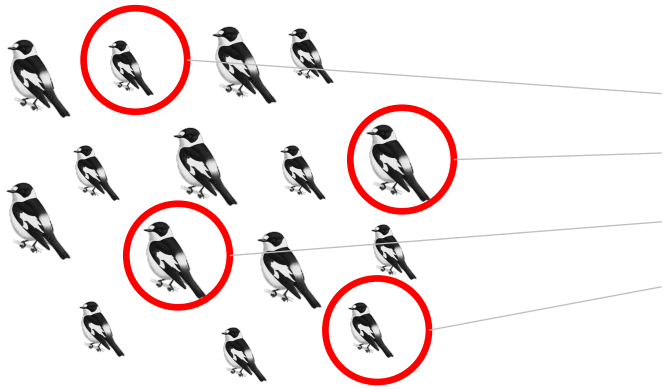


Applications in ecology and evolution

- Central questions in evolutionary genetics
  - How are changes in the genome generated?
  - Why is the genome changing over time?

Evolution is a process influenced by

- mutation
- genetic drift
- natural selection
- demography
- recombination



sequencing of a sample of individuals

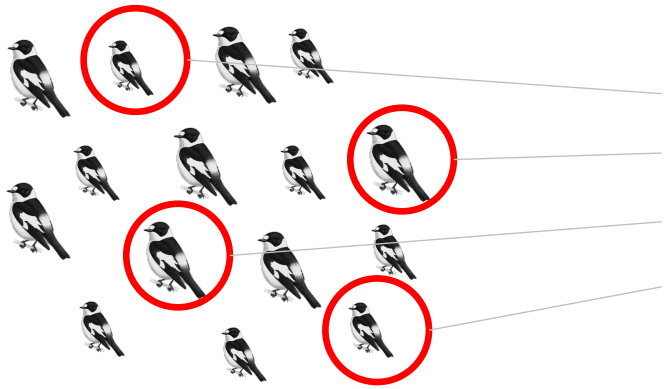
A	C	T	T	A	G	T	A
G	C	T	C	A	G	T	C
G	C	G	C	A	G	T	C
A	C	T	T	A	G	T	C

Applications in ecology and evolution

- Central questions in evolutionary genetics
  - How are changes in the genome generated?
  - Why is the genome changing over time?

Evolution is a process influenced by

- mutation
- genetic drift
- natural selection
- demography
- recombination



sequencing of a sample of individuals

A	C	T	T	A	G	T	A
G	C	T	C	A	G	T	C
G	C	G	C	A	G	T	C
A	C	T	T	A	G	T	C

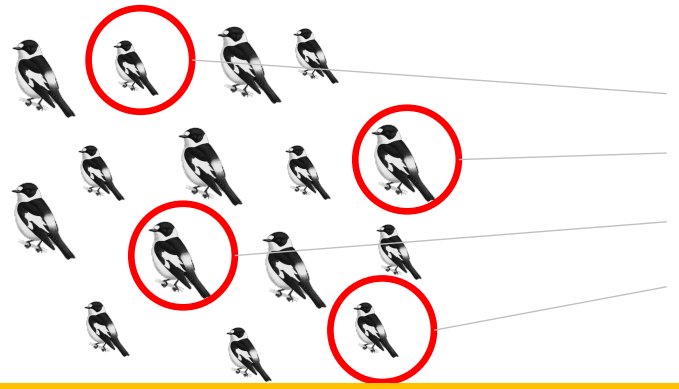
statistical inference

Applications in ecology and evolution

- Central questions in evolutionary genetics
  - How are changes in the genome generated?
  - Why is the genome changing over time?

Evolution is a process influenced by

- mutation
- genetic drift
- natural selection
- demography
- recombination



sequencing of a sample of individuals

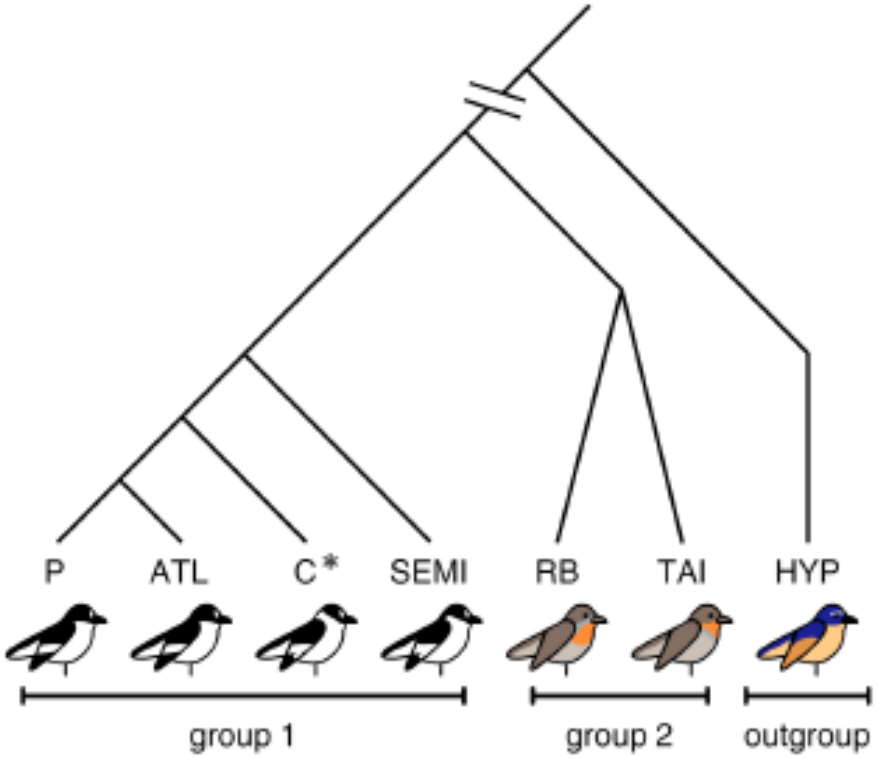
A	C	T	T	A	G	T	A
G	C	T	C	A	G	T	C
G	C	G	C	A	G	T	C
A	C	T	T	A	G	T	C

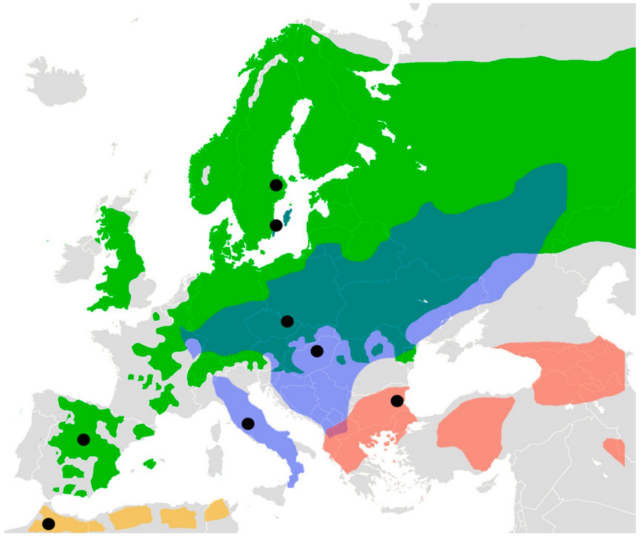
Information is contained in allele frequency data (amongst others)

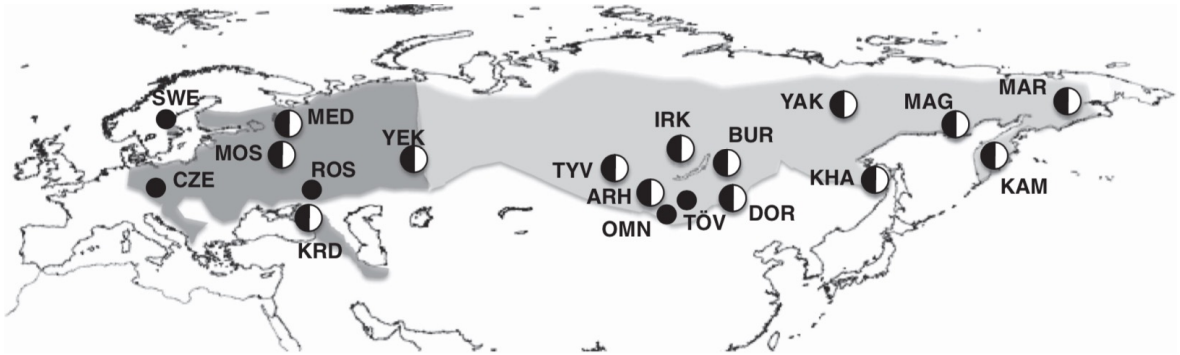
statistical inference

SNP calling practical - overview

- SNP calling and detection of balancing selection in *Ficedula* flycatchers







SNP calling practical - overview

- SNP calling and detection of balancing selection in *Ficedula* flycatchers
- Perform SNP calling in a subset of *Ficedula* flycatcher individuals
  - starting from recalibrated BAM files to a filtered VCF file
- Description of genetic variation and detection of balancing selection across two selected scaffolds
- Quality assessment and interpretation of signatures of balancing selection