SNV calling and the study of genetic variation in ecology and evolution

Carina MUGAL

carina.mugal@univ-lyon1.fr

CNRS Workshop NGS 2023

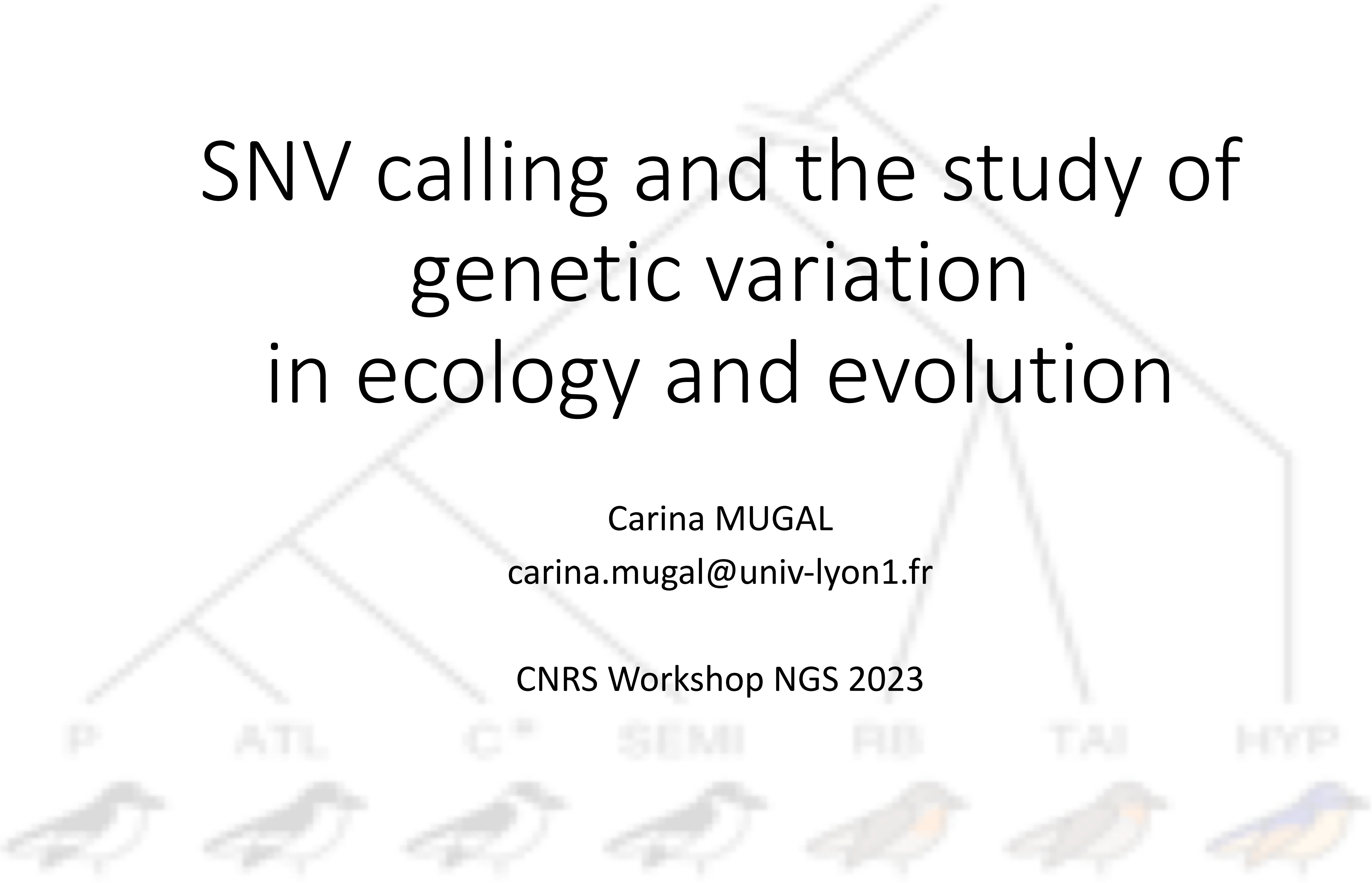


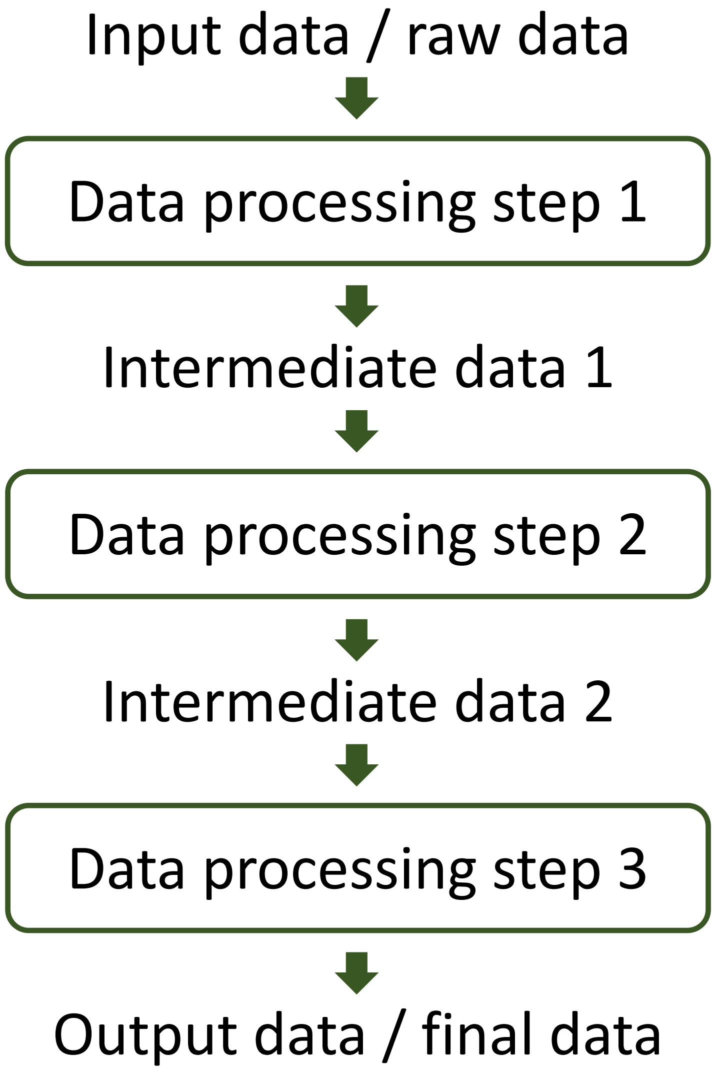
Table of contents

- SNV calling workflow
    - common software and file formats
    - reference genome
    - short-read alignment
    - SNV calling
    - filtering of variant calls
- Applications in ecology and evolution

Table of contents

- SNV calling workflow
  - common software and file formats
  - reference genome
  - short-read alignment
  - SNV calling
  - filtering of variant calls
- Applications in ecology and evolution

What is a workflow?



Input data / raw data

Data processing step 1

Intermediate data 1

Data processing step 2

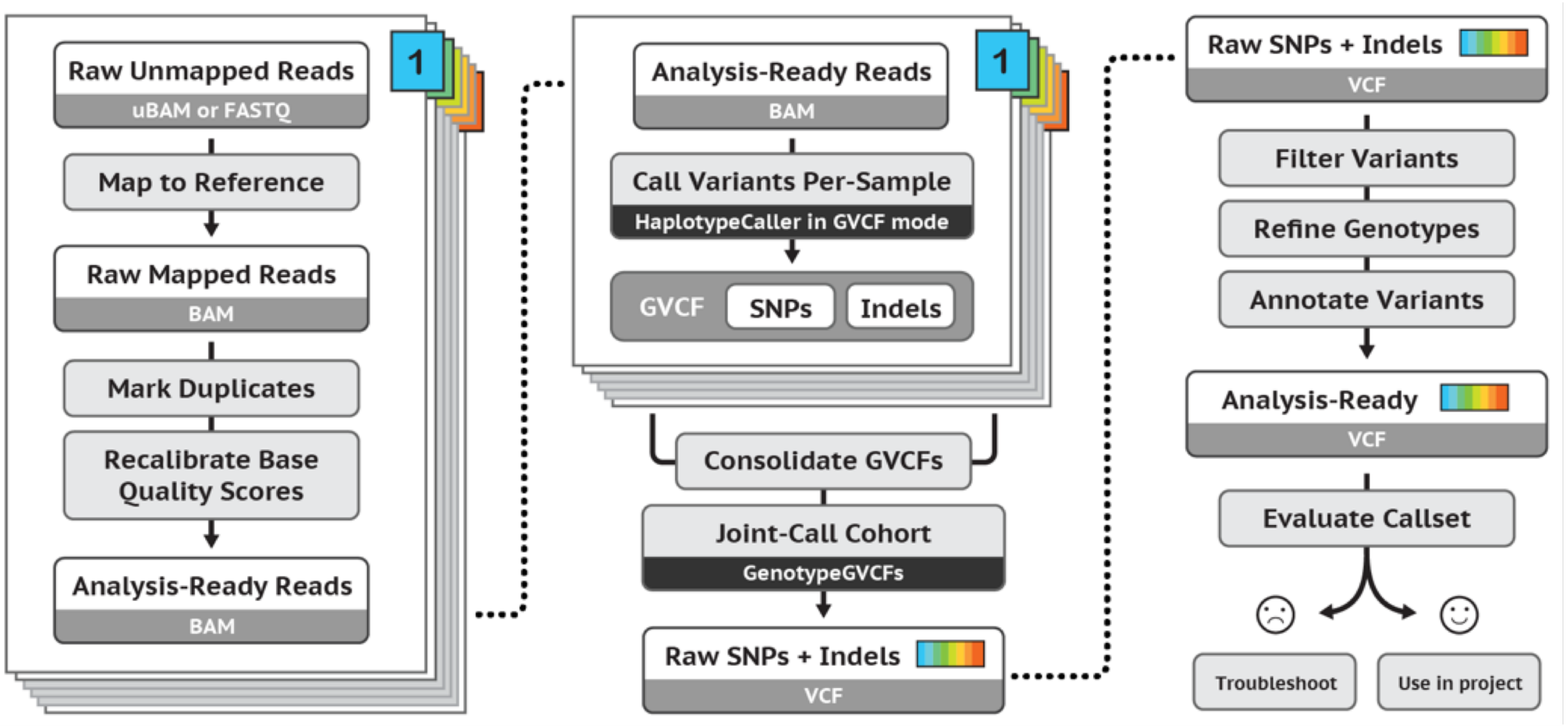
Intermediate data 2

Data processing step 3

Output data / final data

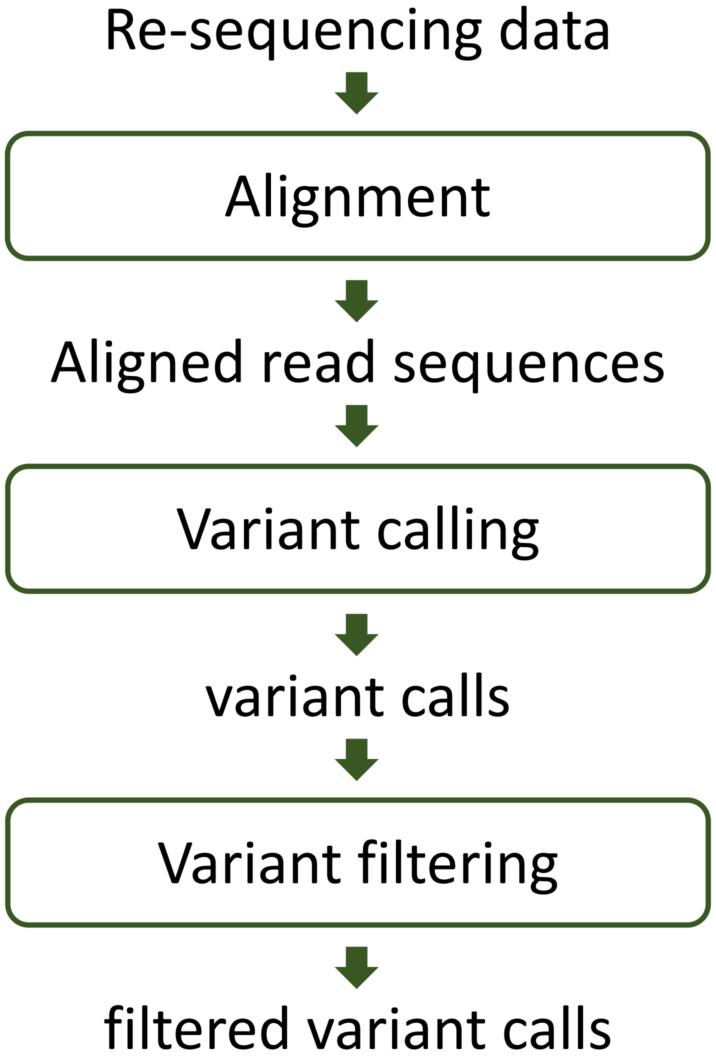
SNV calling workflow

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



Re-sequencing data

Alignment

Aligned read sequences

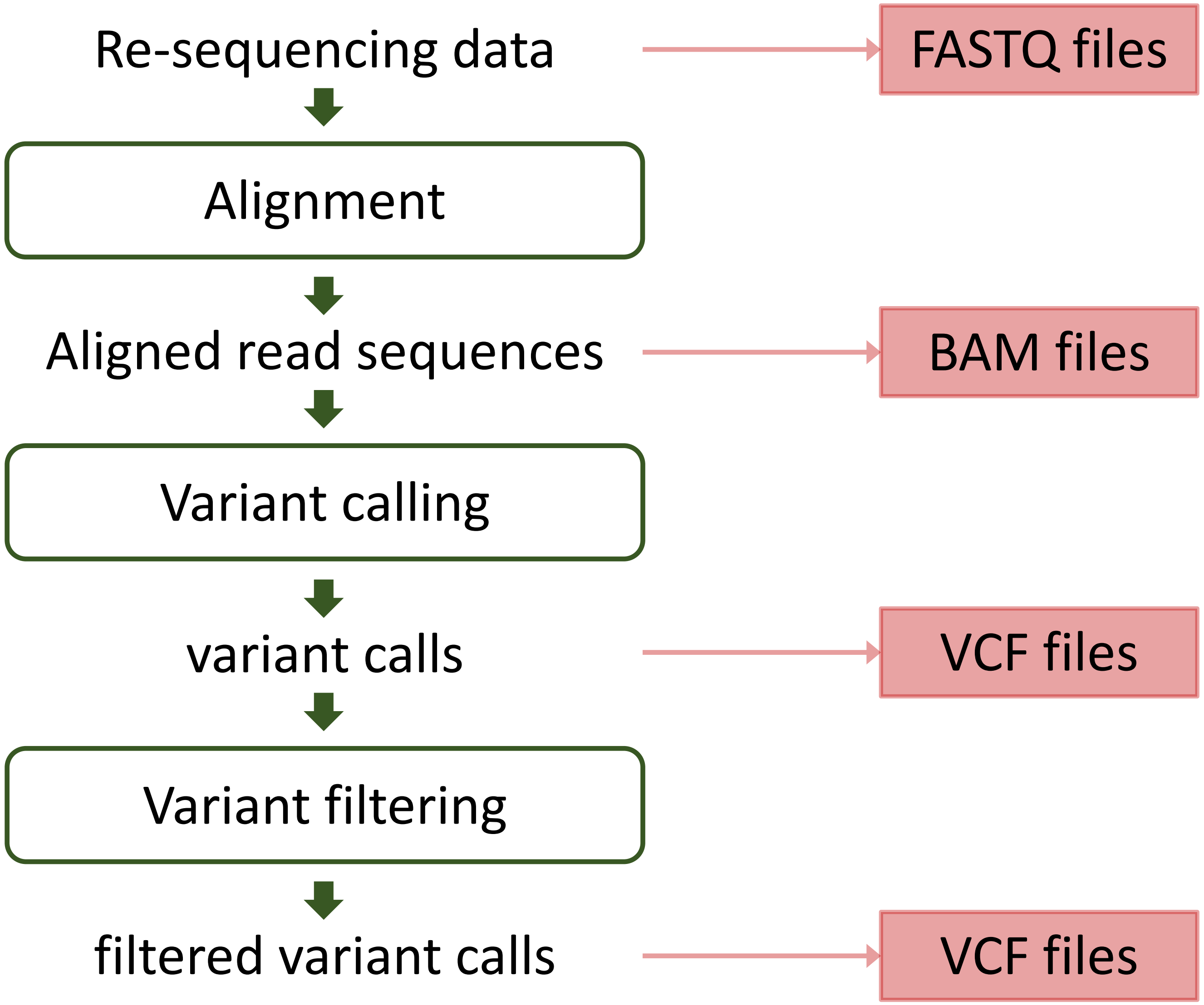
Variant calling

variant calls

Variant filtering

filtered variant calls

Basic workflow, one example



Re-sequencing data → FASTQ files

Alignment

Aligned read sequences → BAM files

Variant calling

variant calls → VCF files

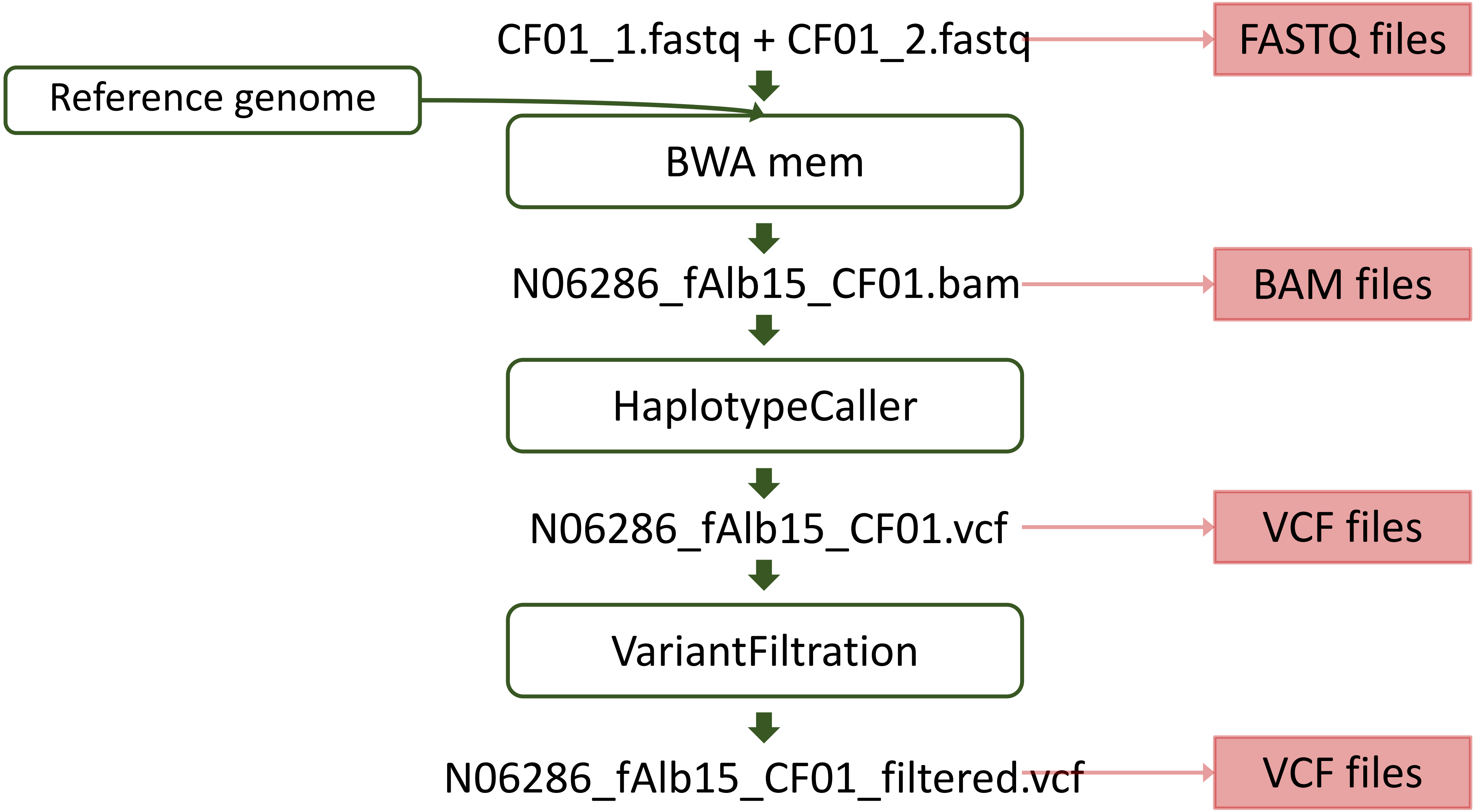
Variant filtering

filtered variant calls → VCF files

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



CF01_1.fastq + CF01_2.fastq

Reference genome

BWA mem

N06286_fAlb15_CF01.bam

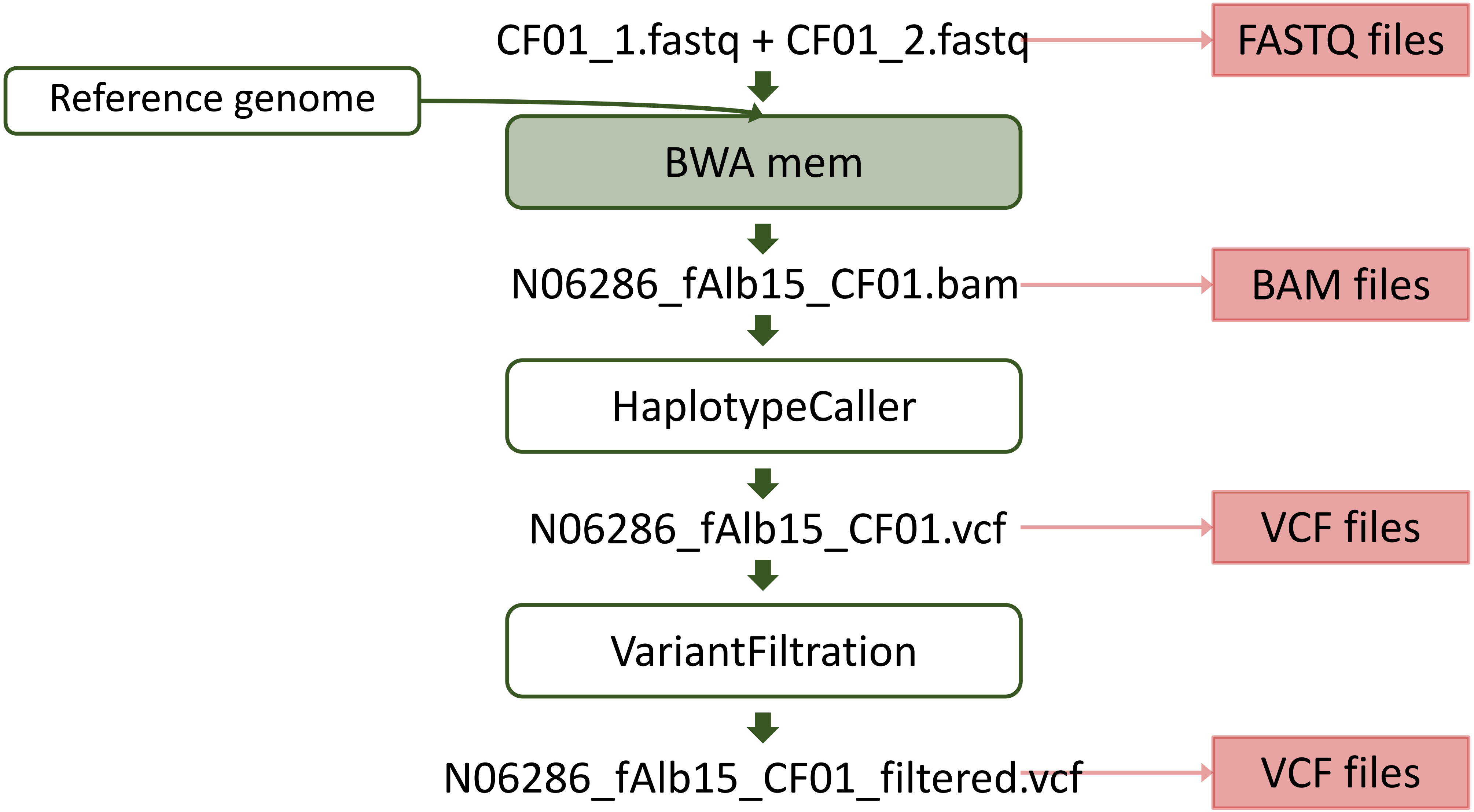
HaplotypeCaller

N06286_fAlb15_CF01.vcf

VariantFiltration

N06286_fAlb15_CF01_filtered.vcf

Basic variant calling workflow, one sample



CF01_1.fastq + CF01_2.fastq

FASTQ files

Reference genome

BWA mem

N06286_fAlb15_CF01.bam

BAM files

HaplotypeCaller

N06286_fAlb15_CF01.vcf

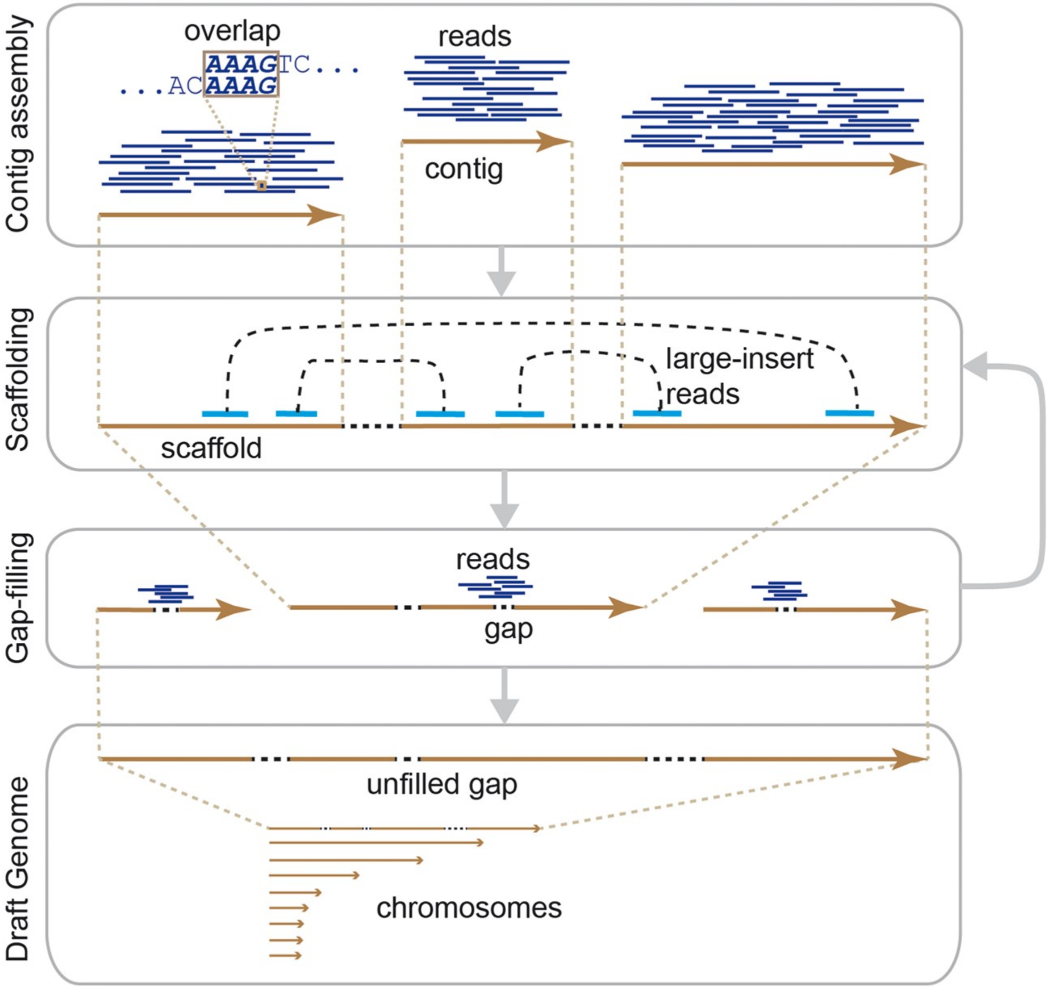
VCF files

VariantFiltration

N06286_fAlb15_CF01_filtered.vcf

VCF files

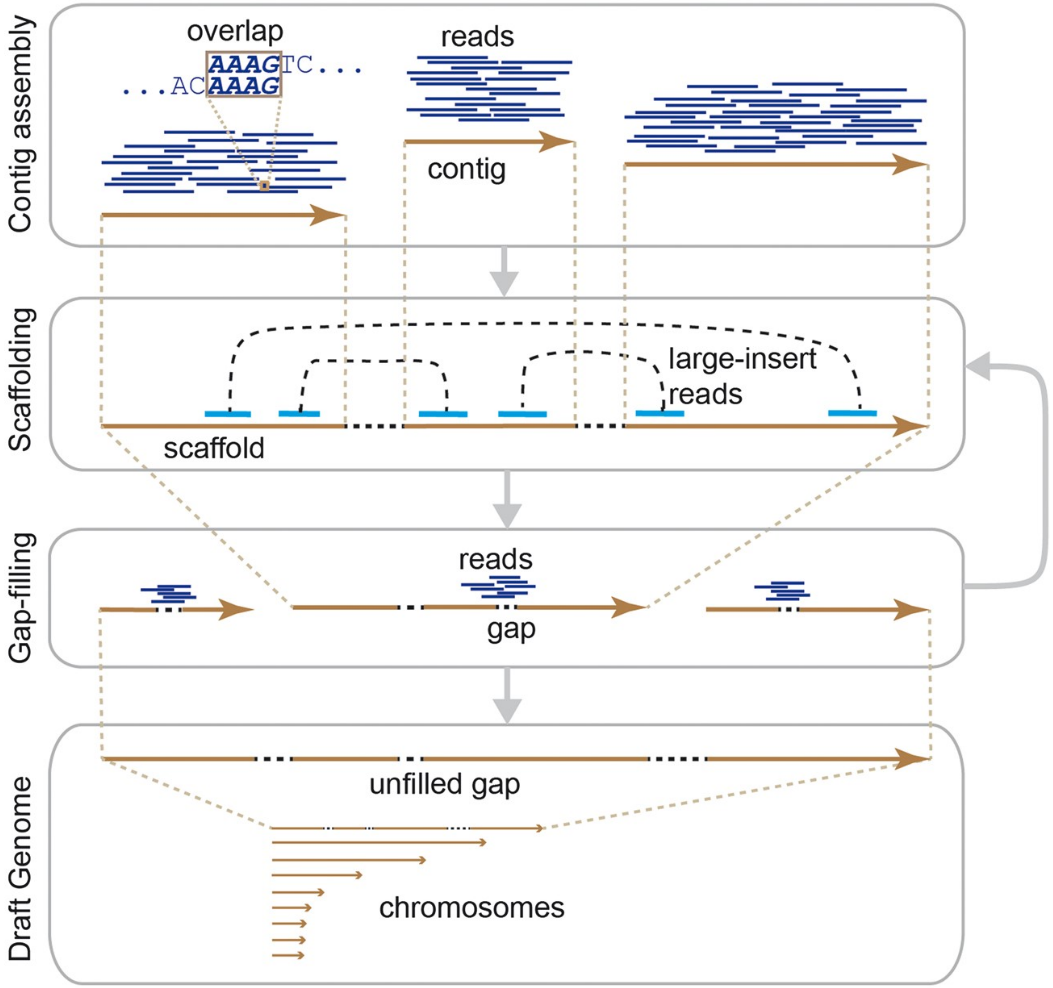
Reference genome



- 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
```

Alignment

```
ACGTTTGCGTCCCGCCCGATNNNNNN---------------CGTAGTCGGGGTATGTAGNNGATTCTCTCAGT
                                              TCGGCGTATGTGGCGGATTCTCT
                   ATGTCTCG---TGTAGATCCG
```

Alignment

```
ACGTTTGCGTCCCGCCCGATNNNNNNN---------------CGTAGTCGGGGTATGTAGNNGATTCTCTCAGT
                                               TCGGCGTATGTGGCGGATTCTCT
                    ATGTCTCG---TGTAGATCCG
```

Can we trust the alignment of the second read?

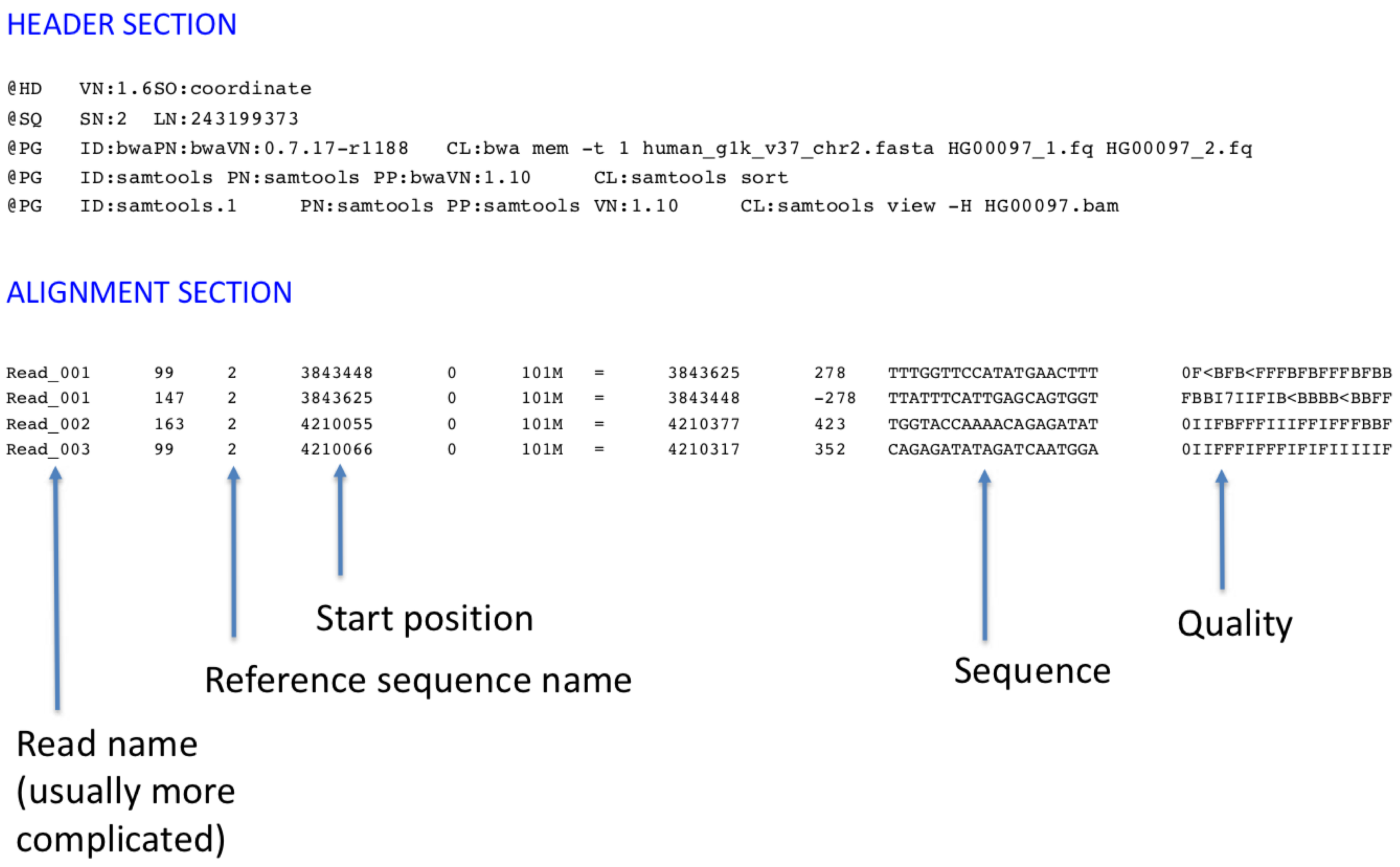
Alignment – Burrows-Wheeler Aligner (BWA)

- 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

Alignment – Burrows-Wheeler Aligner (BWA)

- bwa mem -t 4 -M {input.reference} {input_1.fastq} {input_2.fastq} > {output.sam}

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
```

Start position

Reference sequence name

Sequence

Quality

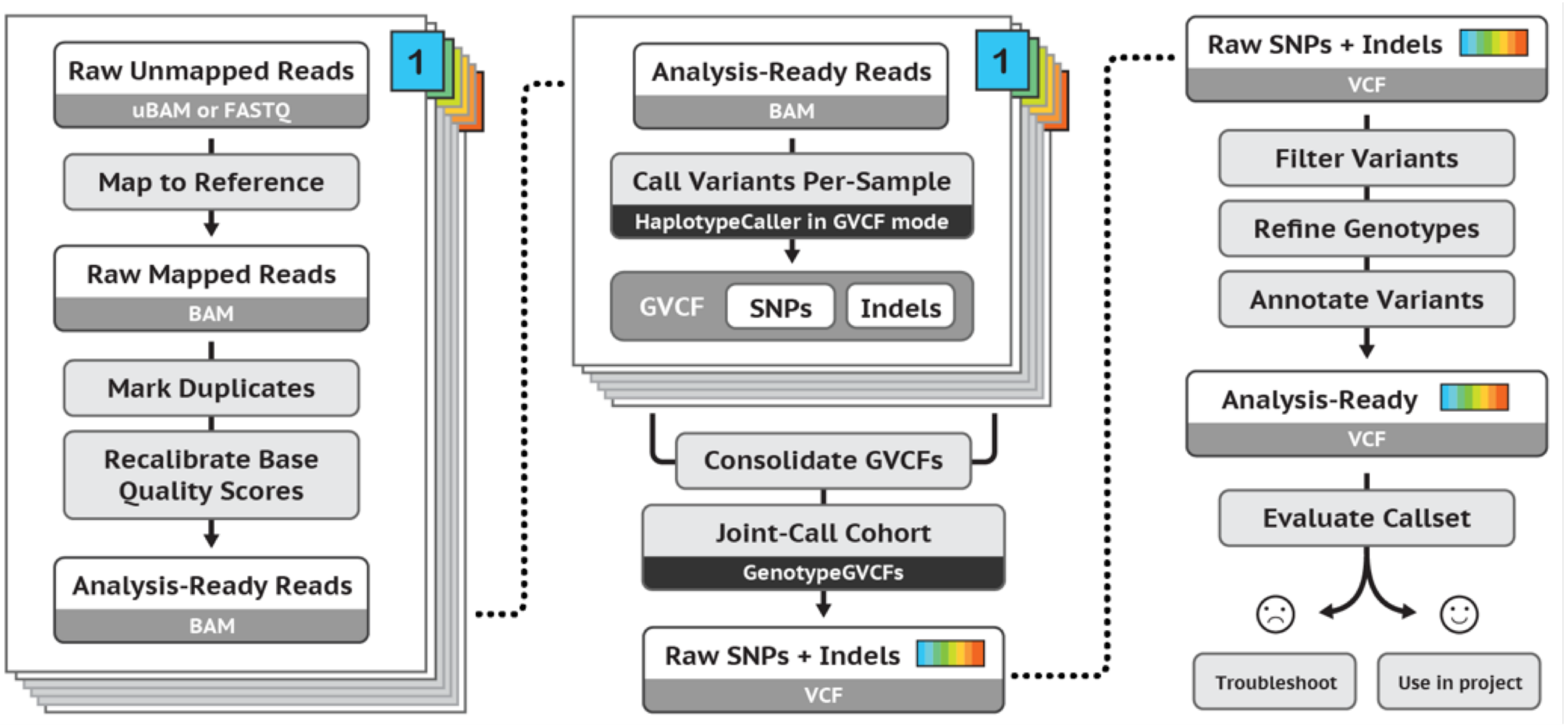
Read name (usually more complicated)

Alignment – Burrows-Wheeler Aligner (BWA)

- bwa mem -t 4 -M {input.reference} {input_1.fastq} {input_2.fastq} > {output.sam}
- samtools view -bhS {input.sam} -o {output.bam}
- samtools sort -o {output.sorted.bam} {input.bam}
- samtools index {input.sorted.bam}

SNV calling workflow

https://gatk.broadinstitute.org



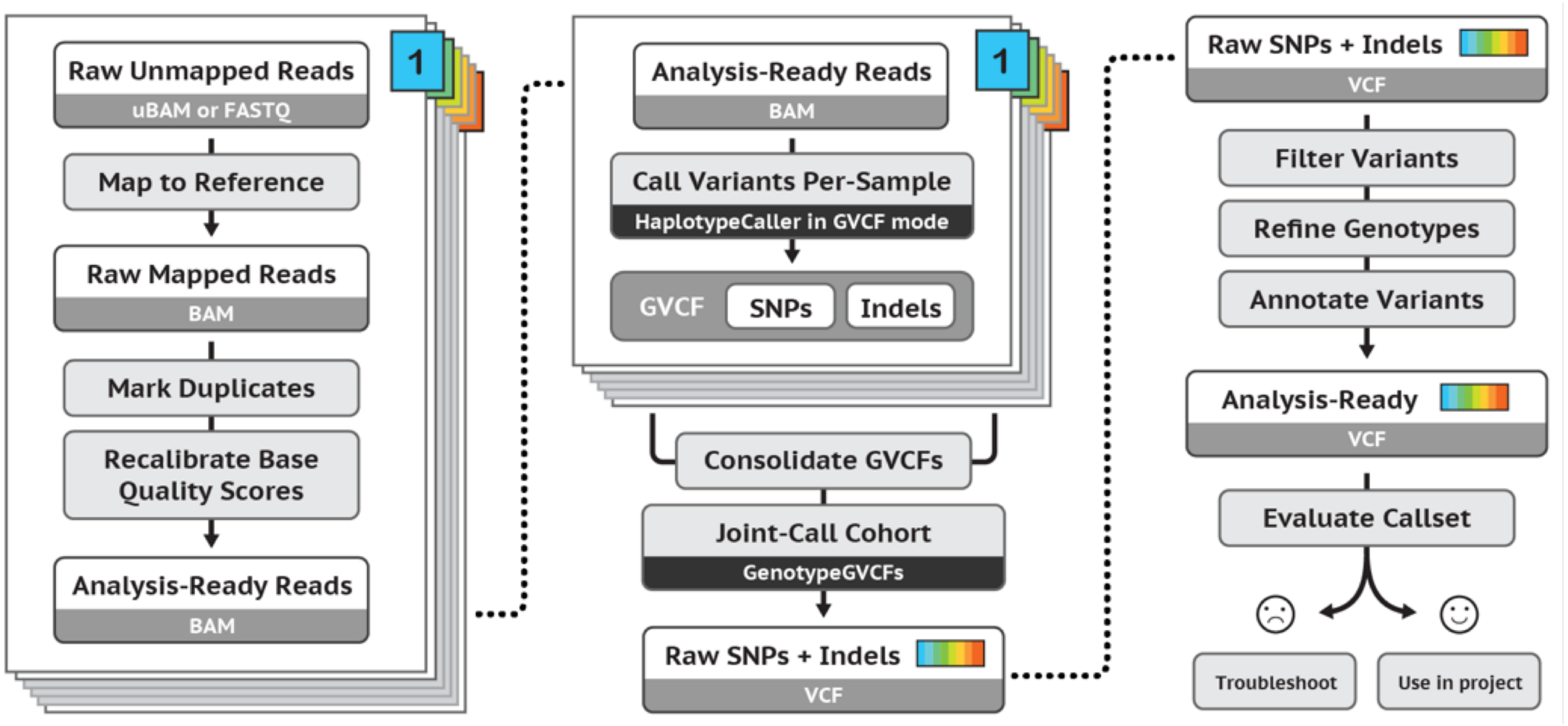
Best Practices for SNP and Indel discovery in germline DNA - leveraging groundbreaking methods for combined power and scalability.

Alignment – Burrows-Wheeler Aligner (BWA)

- bwa mem -t 4 -M {input.reference} {input_1.fastq} {input_2.fastq} > {output.sam}
- samtools view -bhS {input.sam} -o {output.bam}
- samtools sort -o {output.sorted.bam} {input.bam}
- samtools index {input.sorted.bam}
- java -jar $PICARD MarkDuplicates METRICS_FILE={metrics.txt} INPUT={input.sorted.bam} OUTPUT={output.sorted.markedDup.bam}
- samtools view -h -f 0x2 -F 0x4 -F0x8 -F 0x100 {input.sorted.markedDup.bam} > {output.filtered.sam}

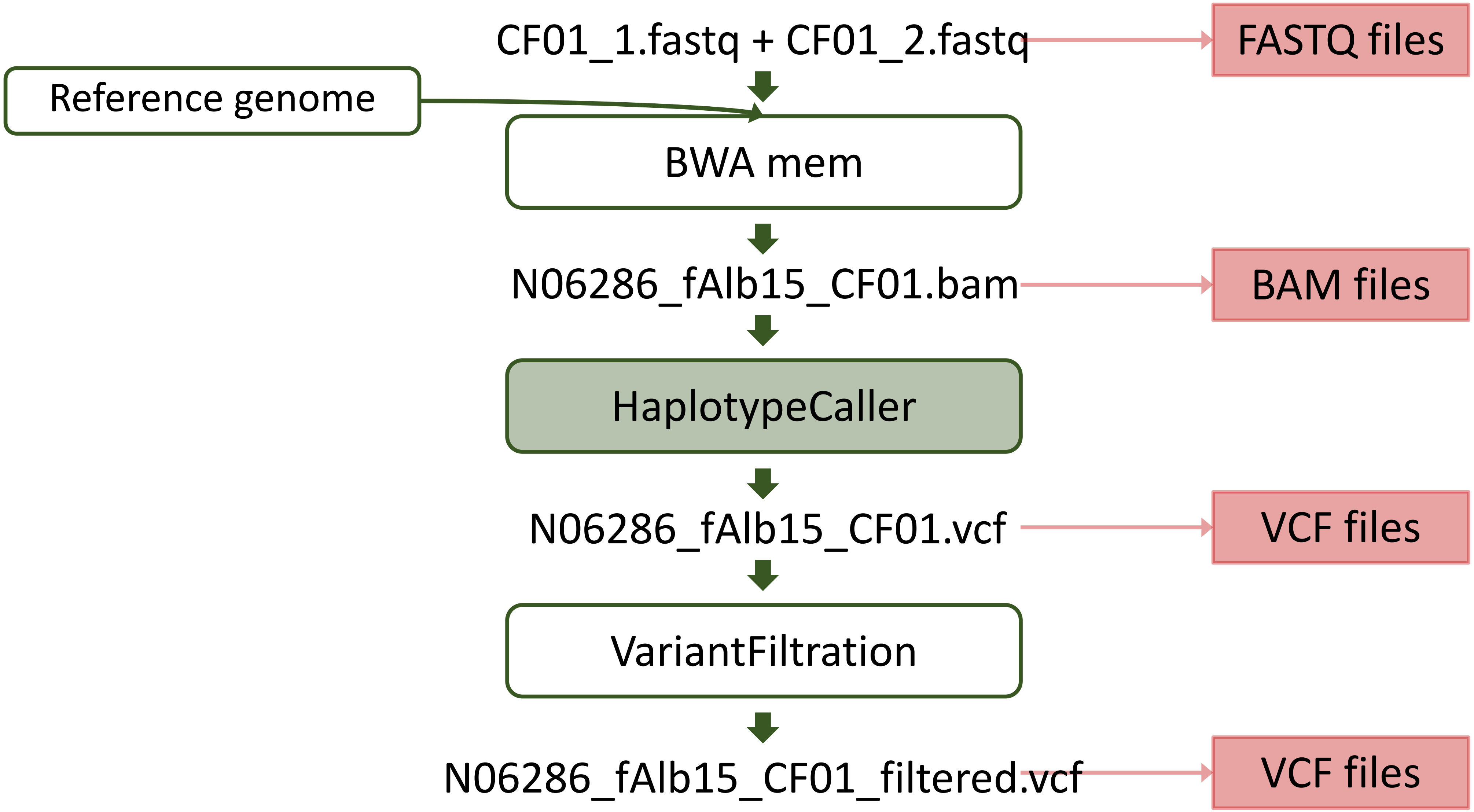
SNV calling workflow

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



CF01_1.fastq + CF01_2.fastq

FASTQ files

Reference genome

BWA mem

N06286_fAlb15_CF01.bam

BAM files

HaplotypeCaller

N06286_fAlb15_CF01.vcf

VCF files

VariantFiltration

N06286_fAlb15_CF01_filtered.vcf

VCF files

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

Alternative allele: the allele NOT in the reference genome C

Reference and alternative alleles

```
Reference:   ACGTTTGCGTCCCGCCCGATNNNNNNN---------------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

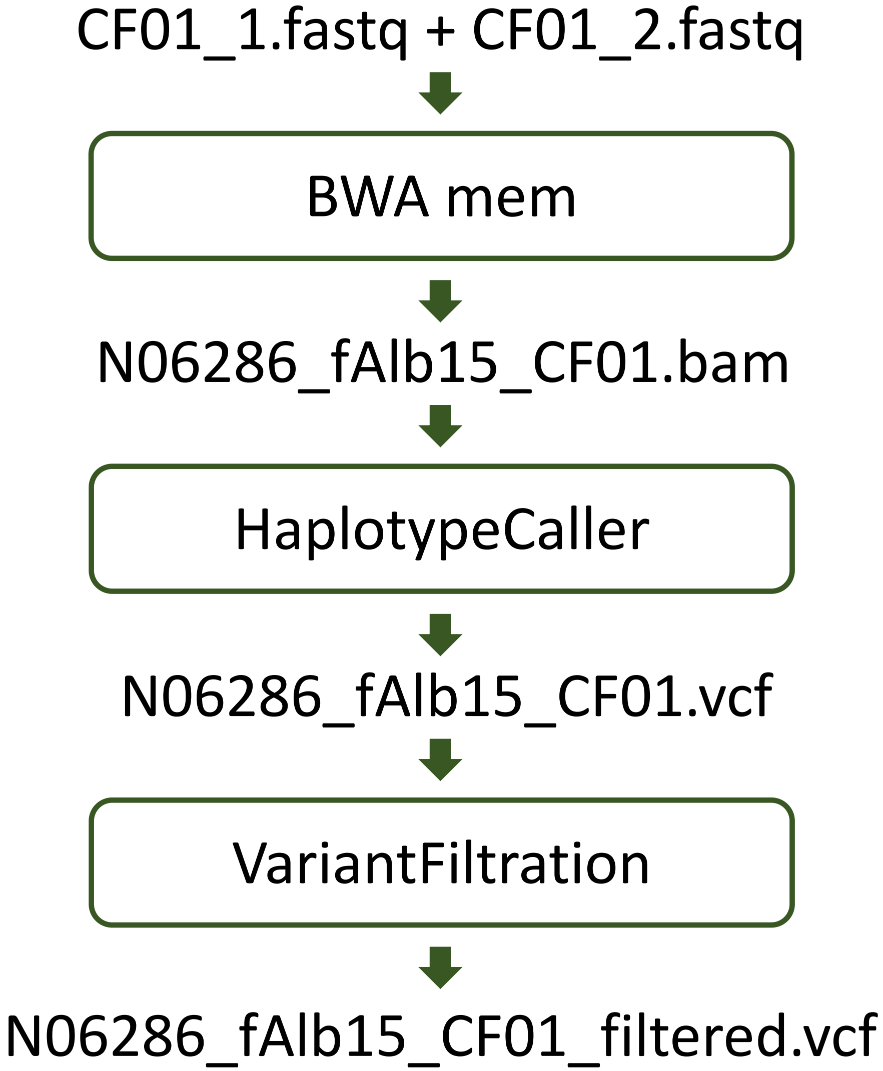
```
##fileformat=VCFv4.2
##ALT=<ID=NON_REF,Description="Represents any possible alternative allele not already represented at this location by REF and ALT">
##FILTER=<ID=LowQual,Description="Low quality">
##FILTER=<ID=PASS,Description="All filters passed">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic 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



CF01_1.fastq + CF01_2.fastq

BWA mem

N06286_fAlb15_CF01.bam

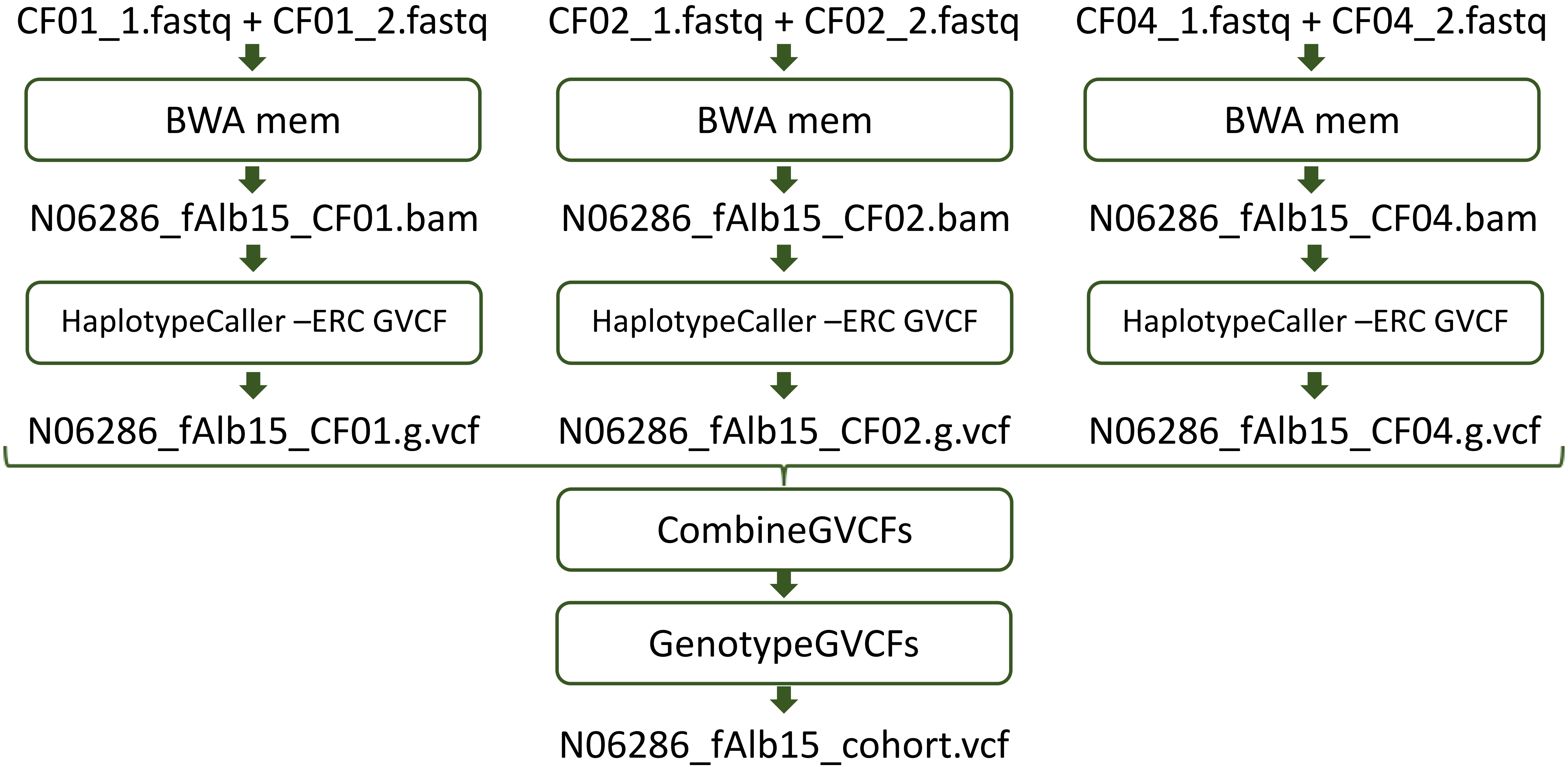
HaplotypeCaller

N06286_fAlb15_CF01.vcf

VariantFiltration

N06286_fAlb15_CF01_filtered.vcf

Basic variant calling workflow in cohort



CF01_1.fastq + CF01_2.fastq

BWA mem

N06286_fAlb15_CF01.bam

HaplotypeCaller –ERC GVCF

N06286_fAlb15_CF01.g.vcf

CF02_1.fastq + CF02_2.fastq

BWA mem

N06286_fAlb15_CF02.bam

HaplotypeCaller –ERC GVCF

N06286_fAlb15_CF02.g.vcf

CF04_1.fastq + CF04_2.fastq

BWA mem

N06286_fAlb15_CF04.bam

HaplotypeCaller –ERC GVCF

N06286_fAlb15_CF04.g.vcf

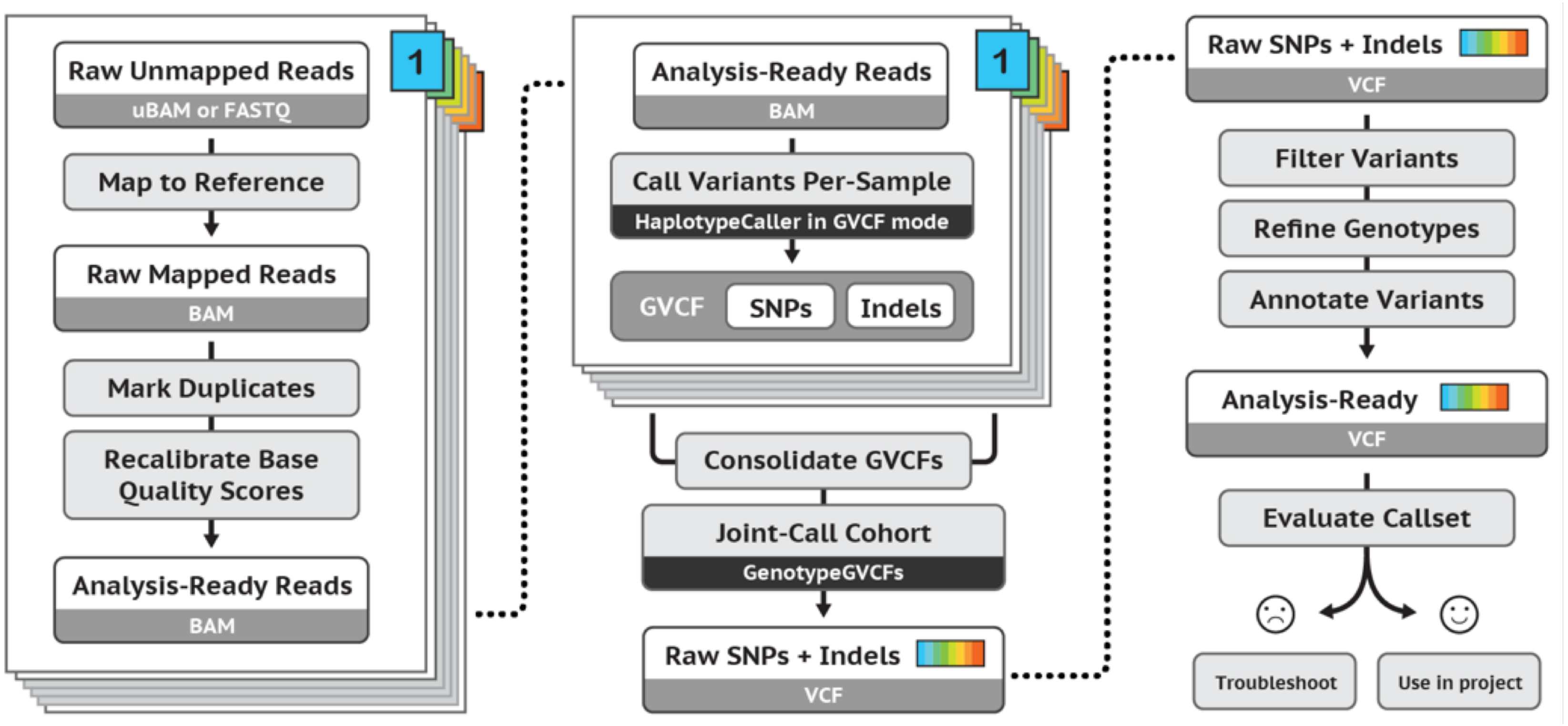
CombineGVCFs

GenotypeGVCFs

N06286_fAlb15_cohort.vcf

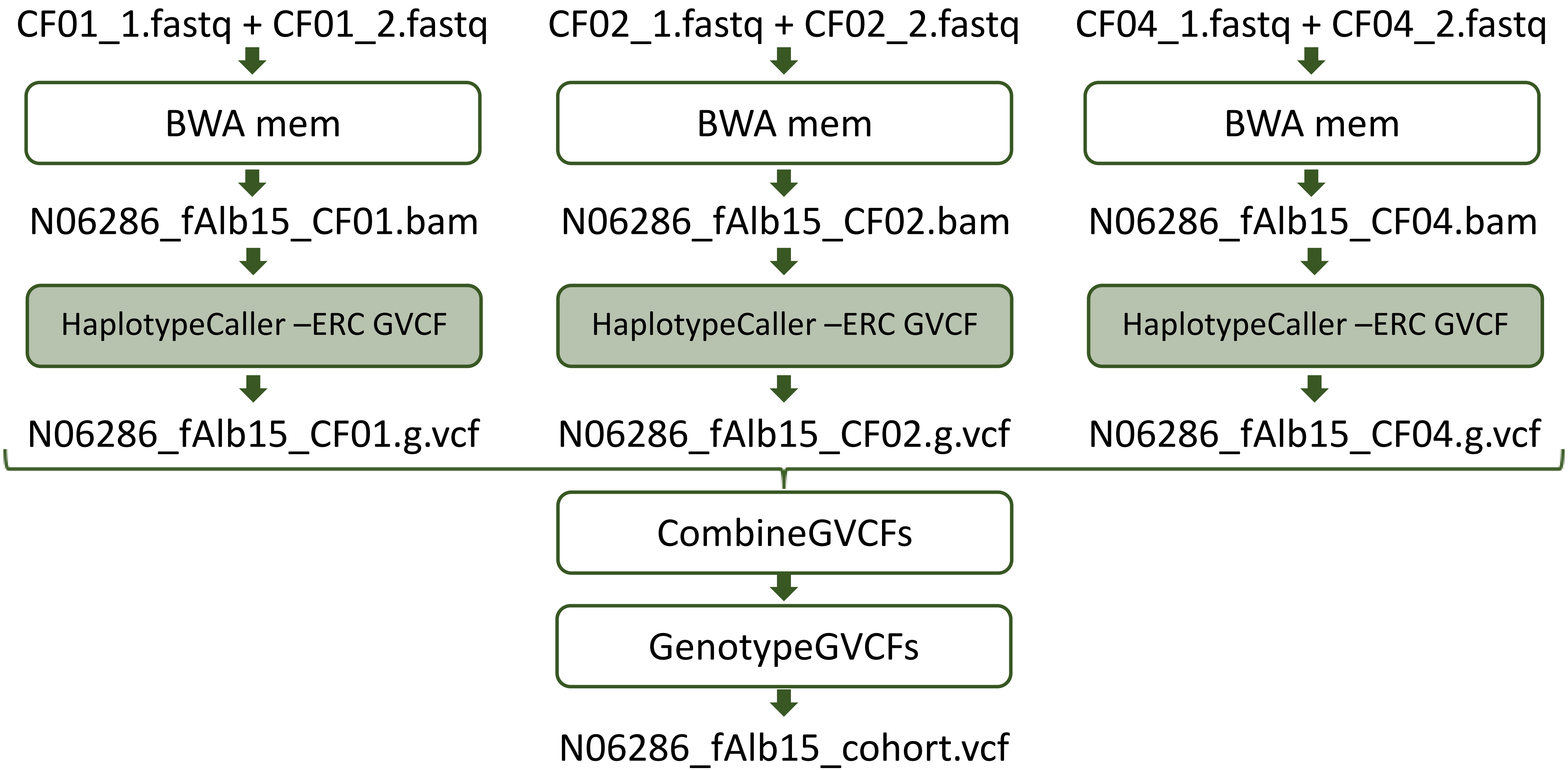
SNV calling workflow

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



CF01_1.fastq + CF01_2.fastq

BWA mem

N06286_fAlb15_CF01.bam

HaplotypeCaller –ERC GVCF

N06286_fAlb15_CF01.g.vcf

CF02_1.fastq + CF02_2.fastq

BWA mem

N06286_fAlb15_CF02.bam

HaplotypeCaller –ERC GVCF

N06286_fAlb15_CF02.g.vcf

CF04_1.fastq + CF04_2.fastq

BWA mem

N06286_fAlb15_CF04.bam

HaplotypeCaller –ERC GVCF

N06286_fAlb15_CF04.g.vcf

CombineGVCFs

GenotypeGVCFs

N06286_fAlb15_cohort.vcf

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

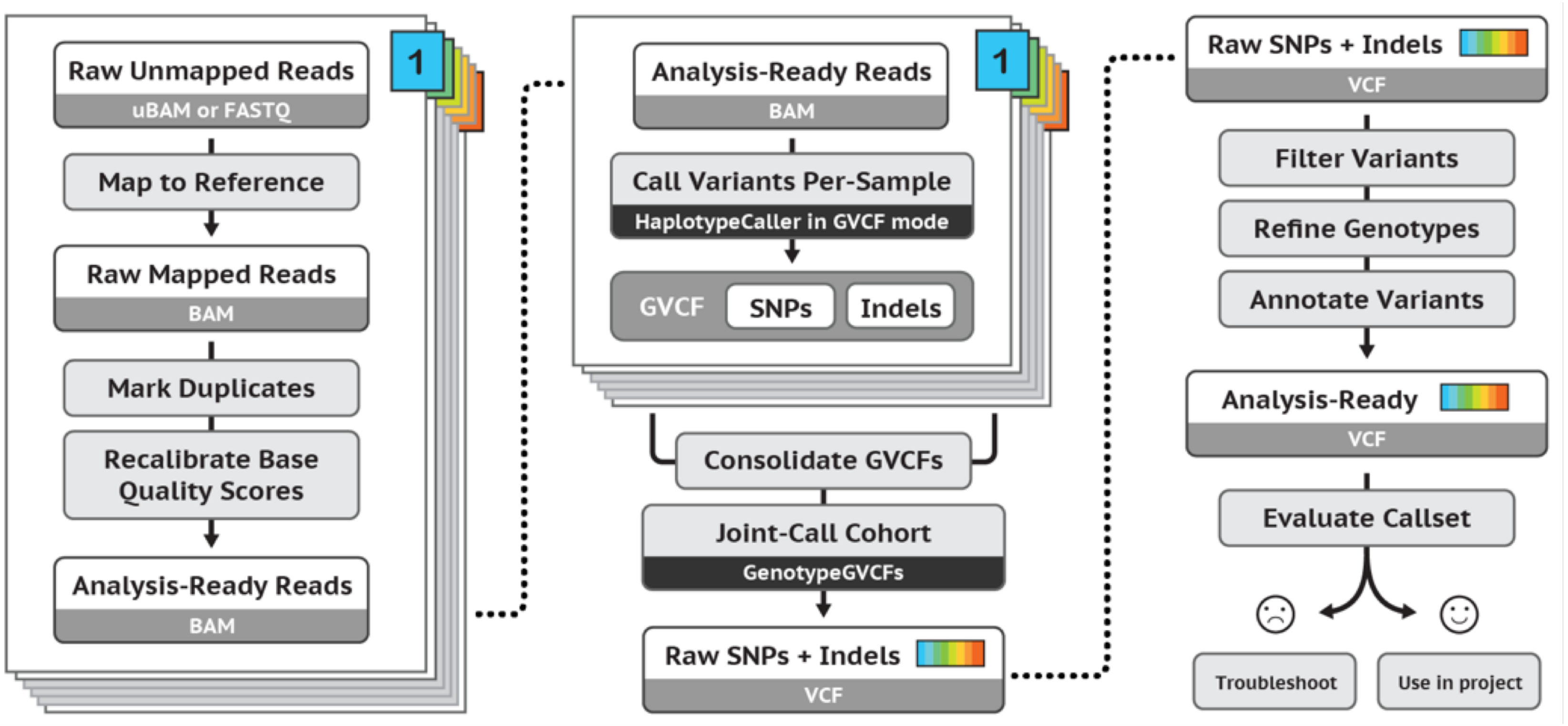
Variant call format (VCF) file for a cohort

- 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
```

SNV 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 <u>header</u> and a list of variant call <u>records</u>

```
##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
- …

Table of contents

- SNV calling workflow
  - common software and file formats
  - reference genome
  - short-read alignment
  - SNV calling
  - filtering of variant calls
- Applications in ecology and evolution

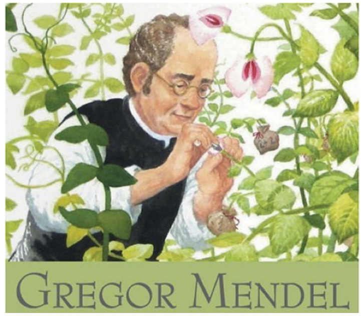
Evolution can be seen as simply a consequence of these conditions…

Variation

Individuals vary in traits that govern reproduction and survival…

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

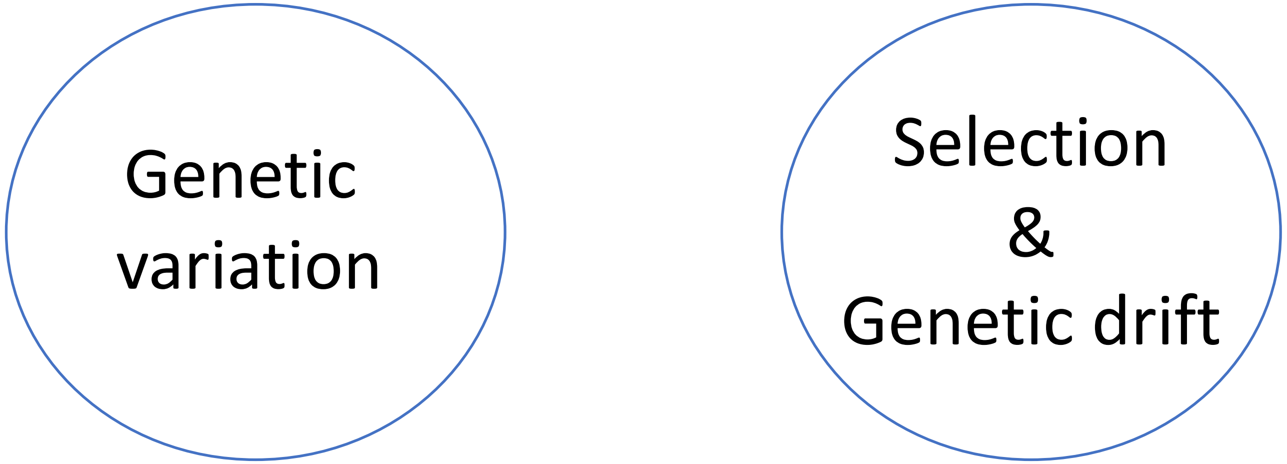
Selection



Heritability

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

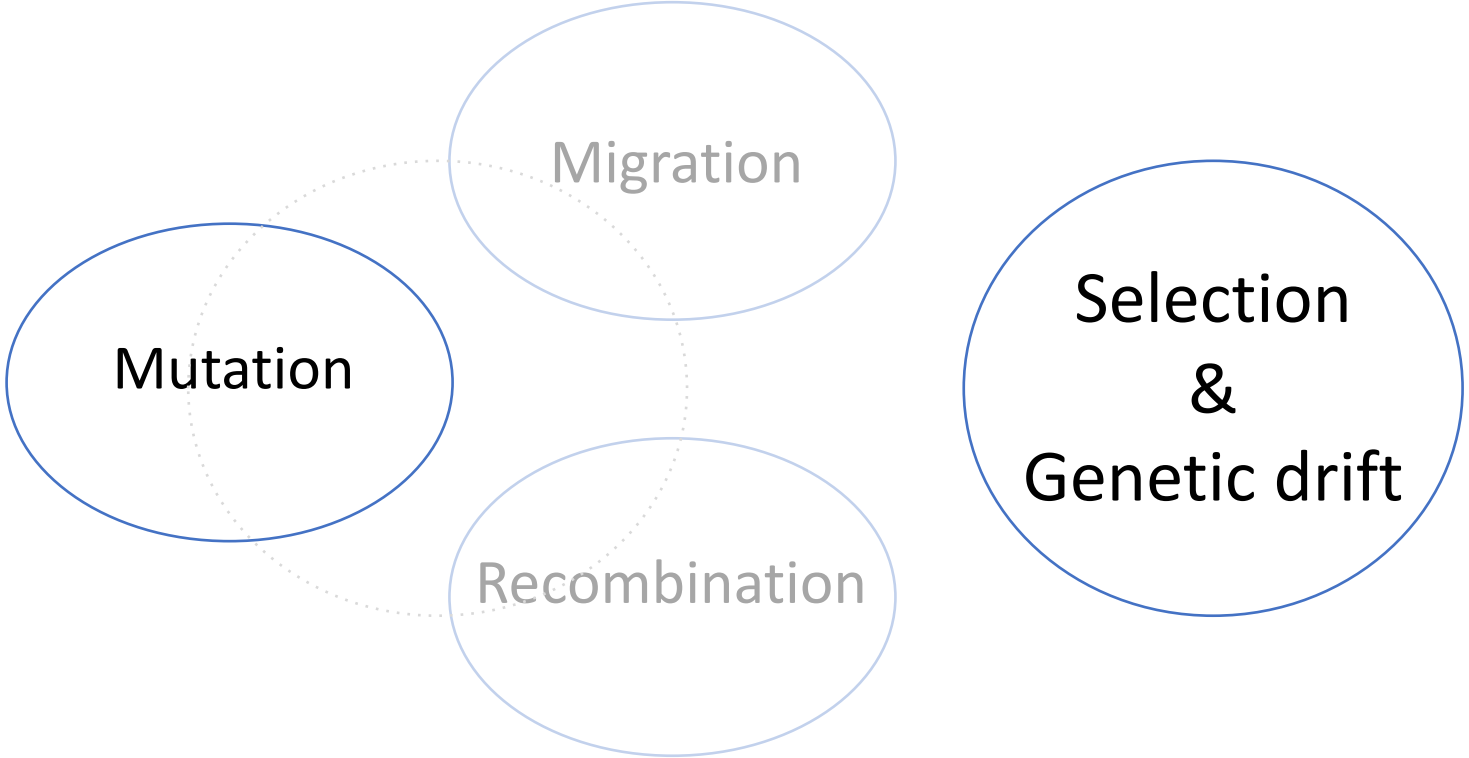
Evolution



Genetic variation

Selection & Genetic drift

Evolution


Migration
Mutation
Selection
&
Genetic drift
Recombination

Applications in ecology and evolution

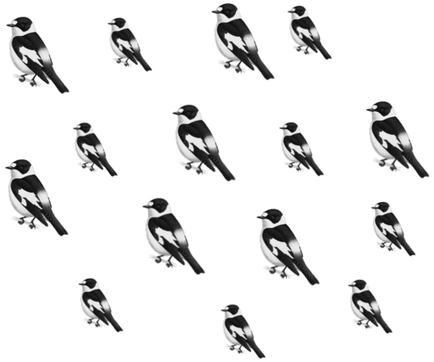
- Central questions in evolutionary genetics
  - How are changes in the genome generated?
  - Why is the genome changing over time?

Applications in ecology and evolution

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

Evolution is a process influenced by

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

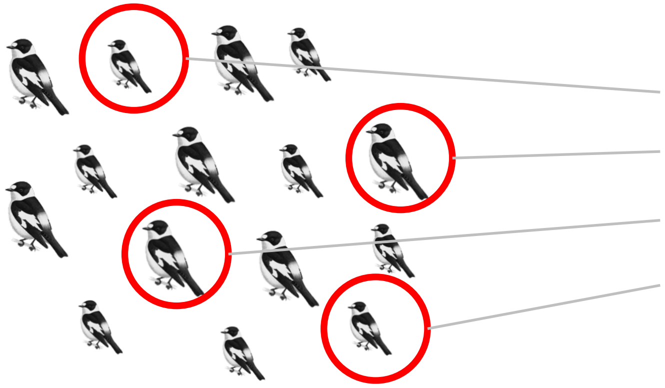


Applications in ecology and evolution

- 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

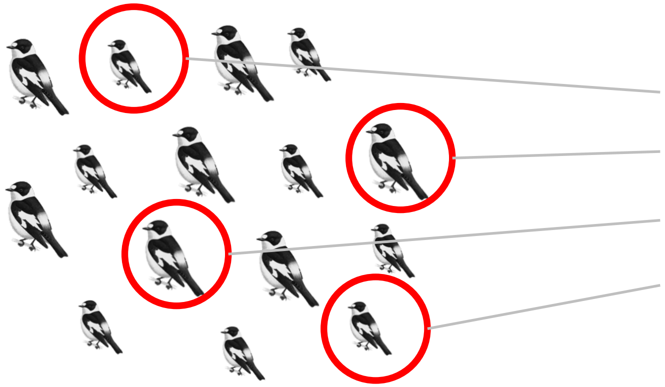
A	C	T	T	A	G	T	A
G	C	T	C	A	G	T	C
G	C	G	C	A	G	T	C
A	C	T	T	A	G	T	C

Applications in ecology and evolution

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

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

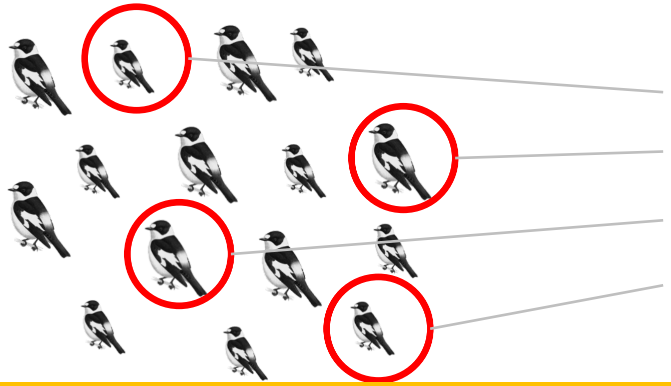
statistical inference

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

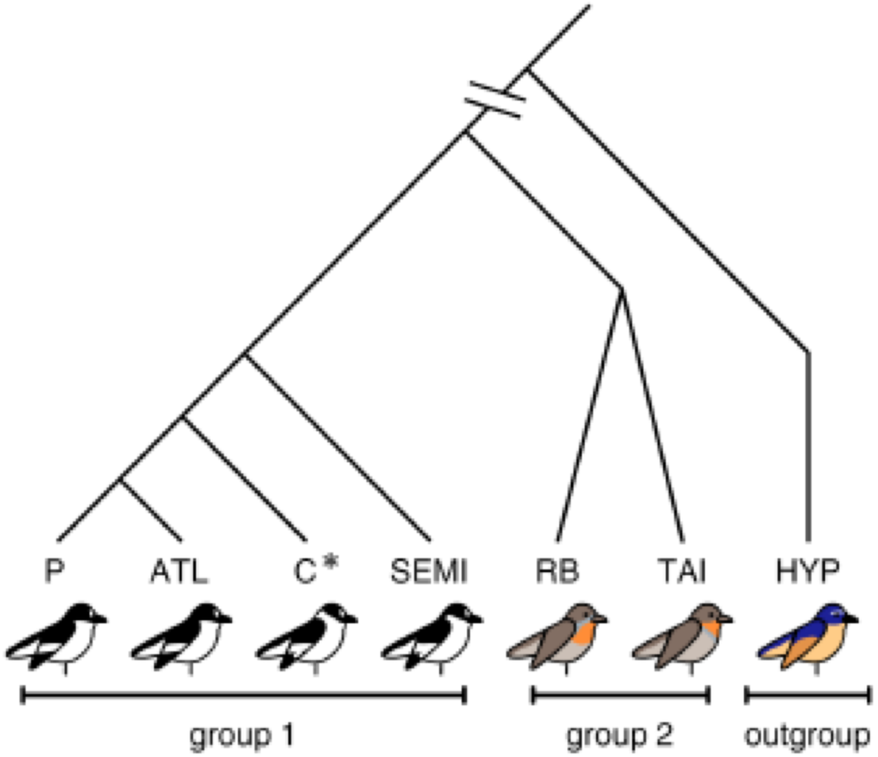
A	C	T	T	A	G	T	A
G	C	T	C	A	G	T	C
G	C	G	C	A	G	T	C
A	C	T	T	A	G	T	C

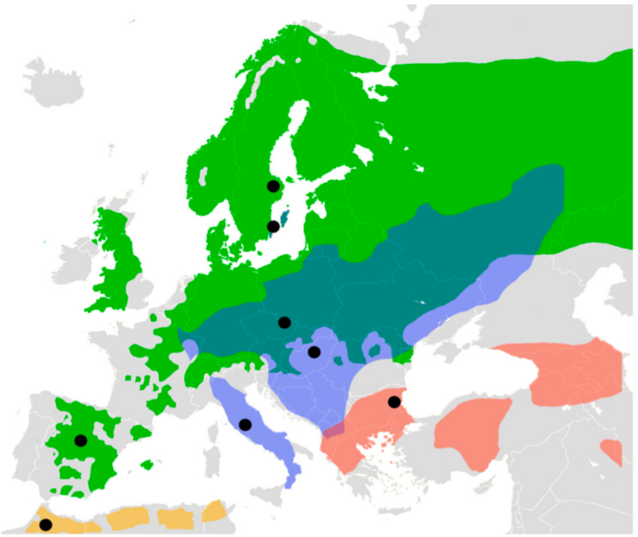
Information is contained in allele frequency data (amongst others)

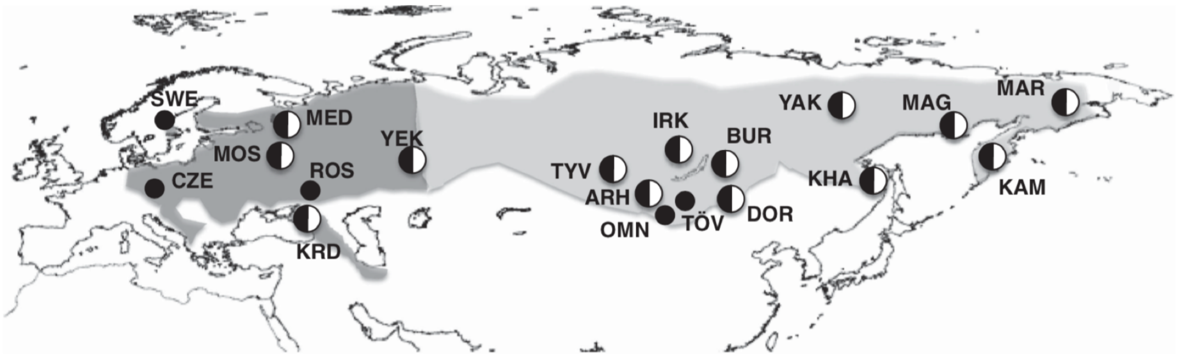
statistical inference

SNV calling practical - overview

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

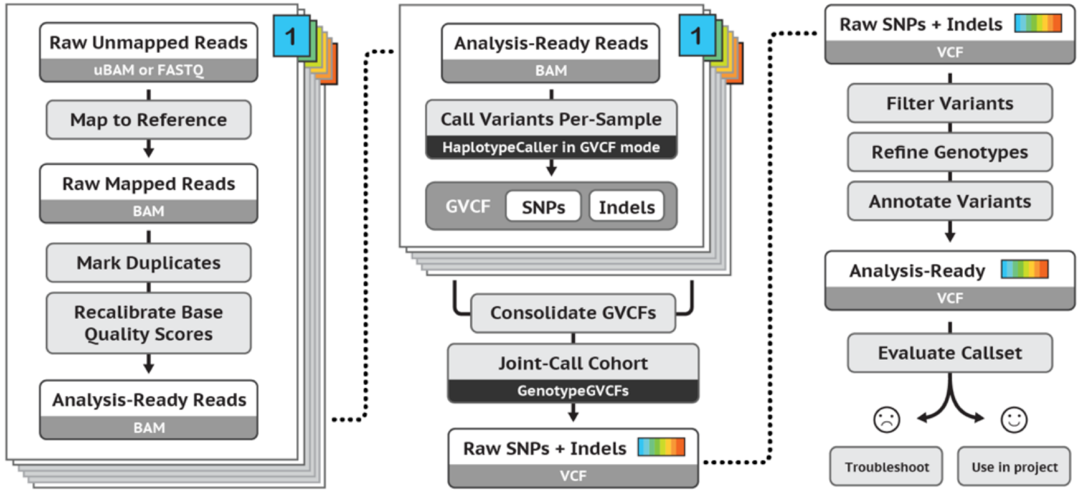






SNV calling practical - overview

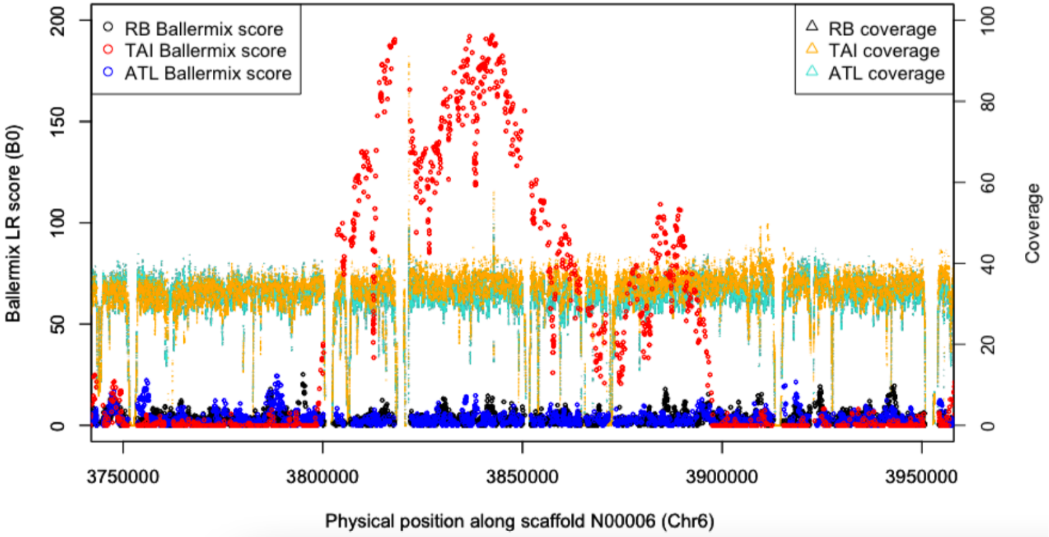
- SNV calling and detection of balancing selection in *Ficedula* flycatchers
- Perform SNV calling in a subset of *Ficedula* flycatcher individuals
  - starting from recalibrated BAM files to a filtered VCF file
- Description of genetic variation and detection of balancing selection across two selected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



Raw Unmapped Reads
uBAM or FASTQ
Map to Reference
Raw Mapped Reads
BAM
Mark Duplicates
Recalibrate Base Quality Scores
Analysis-Ready Reads
BAM
1

Analysis-Ready Reads
BAM
Call Variants Per-Sample
HaplotypeCaller in GVCF mode
GVCF SNPs Indels
1
Consolidate GVCFs
Joint-Call Cohort
GenotypeGVCFs
Raw SNPs + Indels
VCF

Raw SNPs + Indels
VCF
Filter Variants
Refine Genotypes
Annotate Variants
Analysis-Ready
VCF
Evaluate Callset
☹ ☺
Troubleshoot
Use in project


○ RB Ballermix score
○ TAI Ballermix score
○ ATL Ballermix score
△ RB coverage
△ TAI coverage
△ ATL coverage
Ballermix LR score (B0)
Coverage
Physical position along scaffold N00006 (Chr6)