Exercises in Marine Ecological Genetics

05. Variant calling and SNPs

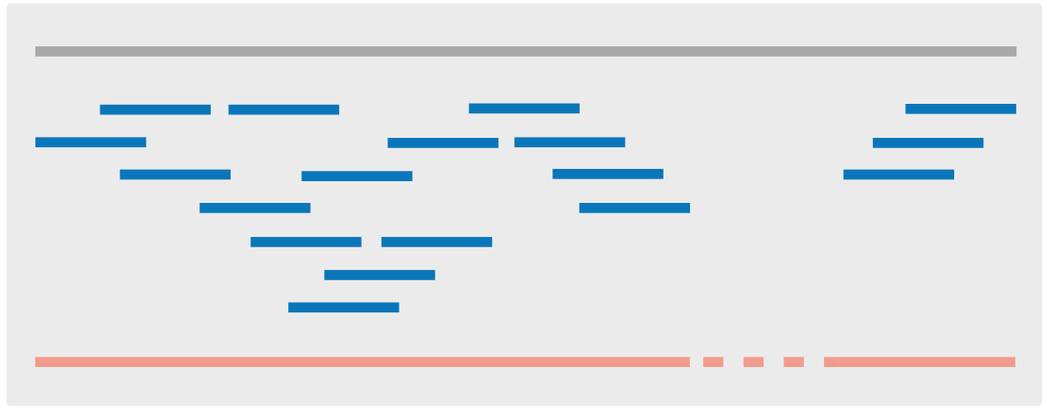
- Clean up short reads
- Map reads to reference genome
- Call variants
- Became familiar with VCF files

Martin Helmkampf

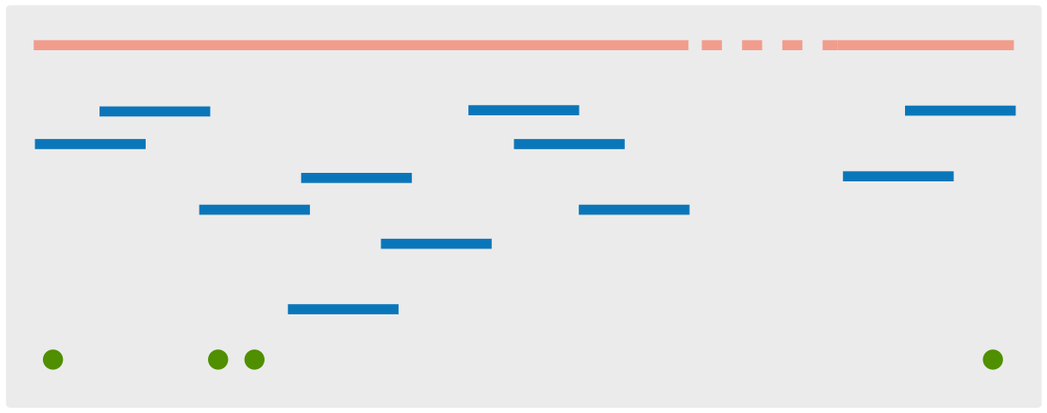
https://github.com/mhelmkampf/meg25

Genome sequencing strategies

De novo

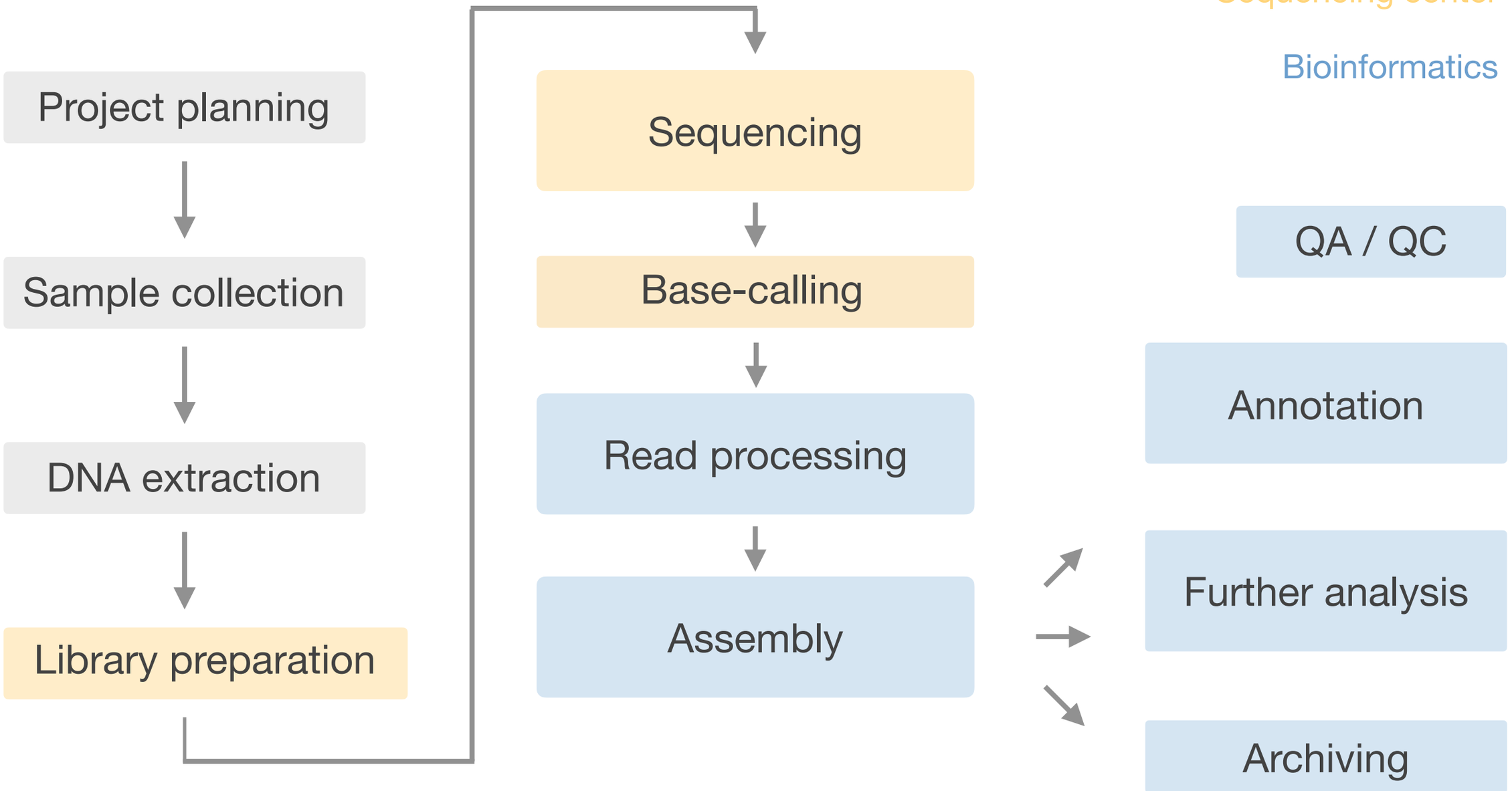


Re-sequencing

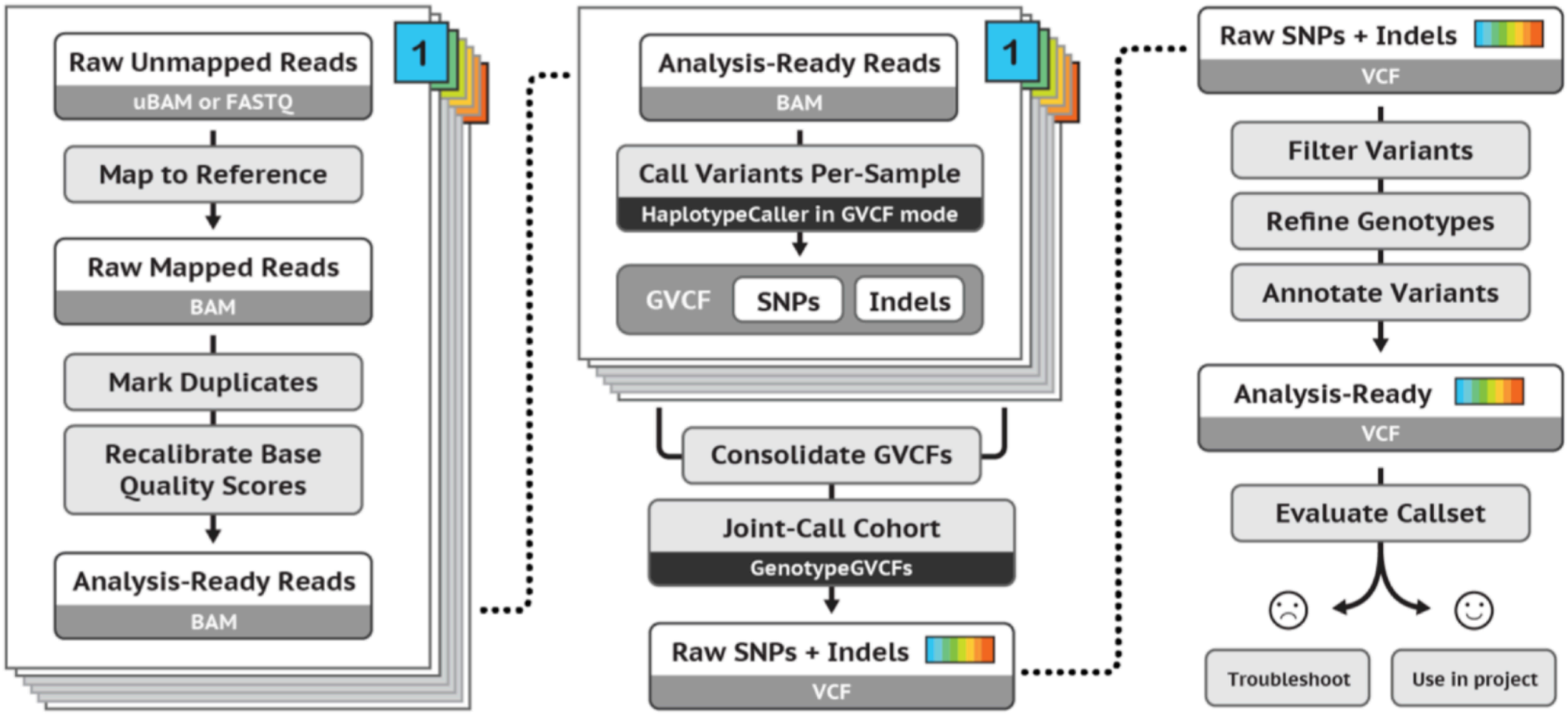


~ Reduced representation sequencing, e.g. RADseq

De novo genome sequencing workflow



Whole-genome genotyping workflow with GATK



gatk.broadinstitute.org

Sequence alignment

```
G A T G T T C G A A
G A T C G A A
G A C C T C G T
        ↓
G A T G T T C G A A
G A T C - - - G A A
G A C C - T C G - T
```

Arranges nucleotide or amino acid sequences

so that the number of mismatches and gaps are minimized

- Multiple sequence alignments can be constructed progressively from pairwise alignments
- Computationally complex, often requires heuristic solutions
- Key to identify evolutionary relationships between sequences (e.g. homology)

Sequence Alignment Map (SAM) format

```
samtools view -F 4 indbel-mtg_unsorted.sam | head -n 1  # print 1st mapped read
```

```
E00489:149:H3H77CCXY:7:1101:27783:2364          99          LG_M
11013       60          150M        =           11247       360
GATAAAAAGACTAATTGTTTCGATGACAATCAGGACAGGAATTAGAGGGCCGGGGGTTCCTTCTGGAAGAAGATGGCCTA
ACGCGTGAGTTGGCTGATTACGCATTCCAATTAGGACGGTTGCTAGTCATAGGGGGGTTGCAATTCCAAG
AAAFFJJJJJJJJJAJJJJ-7FJAFJFJJJJAJJJJJJJJJJJJJJJFJJJJJJJJ7A7FJFJ<FJJJJJJJJJ7FFFJ
JJJAJ7AFF-7FFJJFJJJJ77J7-A-FJF<FJ7F7AJJ--FA-F---FF<AAAJJA))AFFA--<F7J7
NM:i:2      MD:Z:60C82G6 MC:Z:126M    AS:i:140     XS:i:0
```

Key steps in genotyping

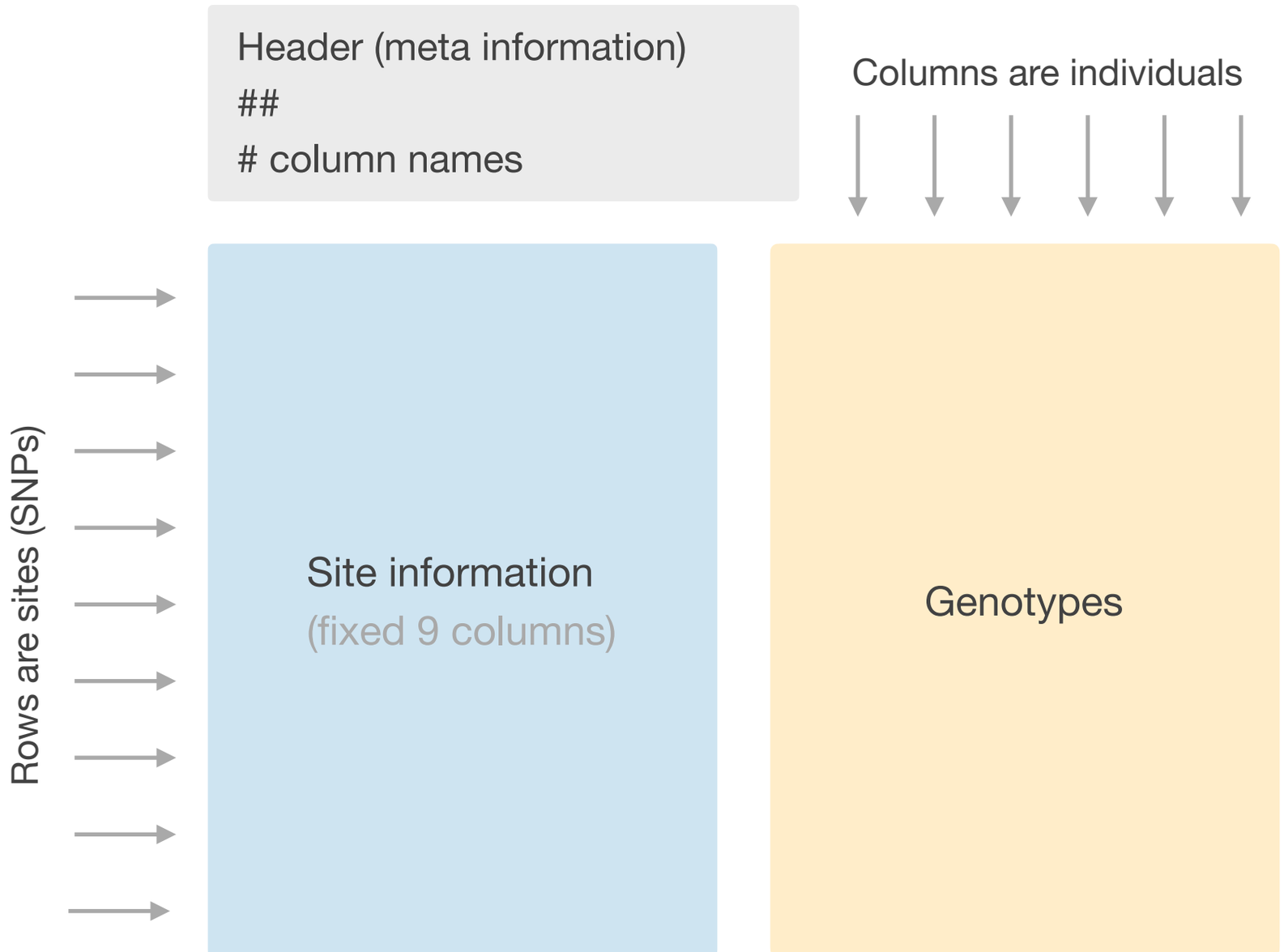
1. **Genotype likelihood calculation**
    - Base calls (A, C, G, T)
    - Base quality scores
    - Mapping quality
    - Read depth

    ➡ Probabilities for each genotype (e.g. AA, AC, CC)

2. **Variant calling**
    - Genotype likelihoods
    - Other probabilities (e.g. mutation rate, population-level information)

    ➡ VCF file with genotype calls and confidence scores

Variant call format (VCF)



Efficiently stores

genetic variation data

Optional data e.g:

- allele frequencies
- no. samples with data
- read depth
- genotype likelihood

Read depth and mapping quality

Phred quality score: $Q = -10 \log_{10} P$

Quality score	*P* incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%

Read depth:

```
Genome: CGTAATGGCATATCGCCTAGATTCGAAACG
Read 1:   TAATGGCATATCGCCTAGAT
Read 2:         CATATCGCCTAGATTCGAAA
Read 3:           TATCGCCTAGATTCGAAACG
Depth:  001111112233333333333222222211
```

Variant calling and SNPs

Recap

```
cutadapt [options] [-o output.fastq] input.fastq    # trim reads by quality etc.

bwa mem [options] reference input.fastq                # map reads to reference

bcftools mpileup [options] input.bam           # calculate genotype likelihoods

bcftools call [options] input.vcf              # call variants and generate VCF
```

- Accurate variant calling relies on high-quality read alignment, which requires read processing steps such as adapter trimming and removal of low-quality bases
- Filtering and thresholds (e.g. depth, base and mapping quality) are essential to distinguish true genetic variants from sequencing or alignment errors
- Genotype data form the basis of downstream analyses, including estimates of genetic diversity, population structure and signals of selection