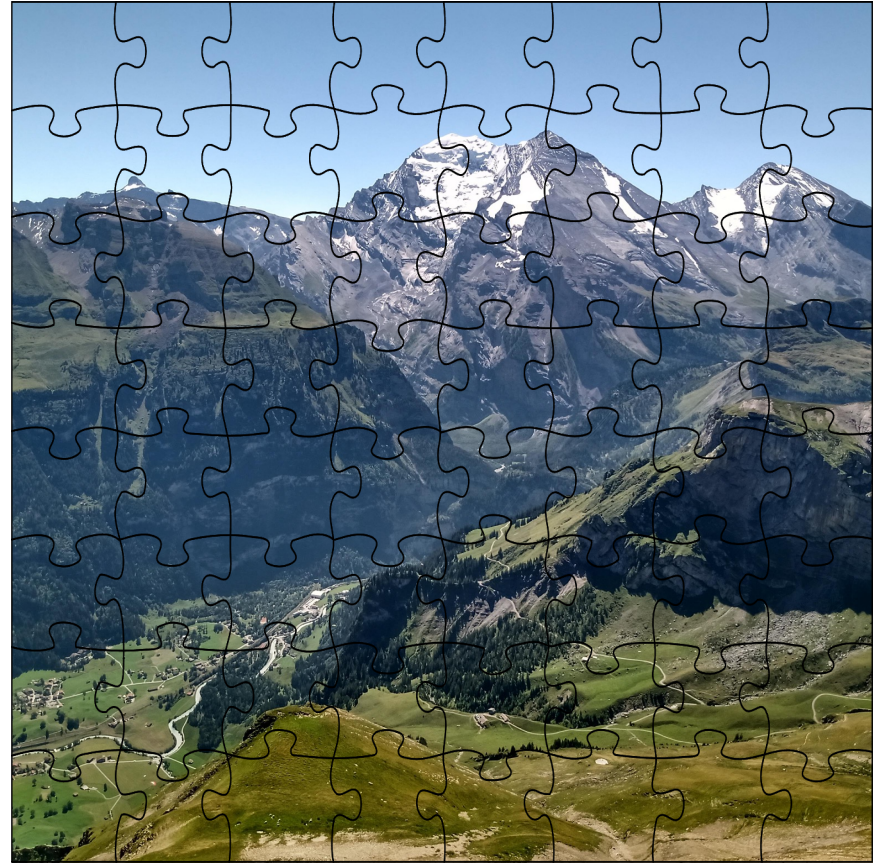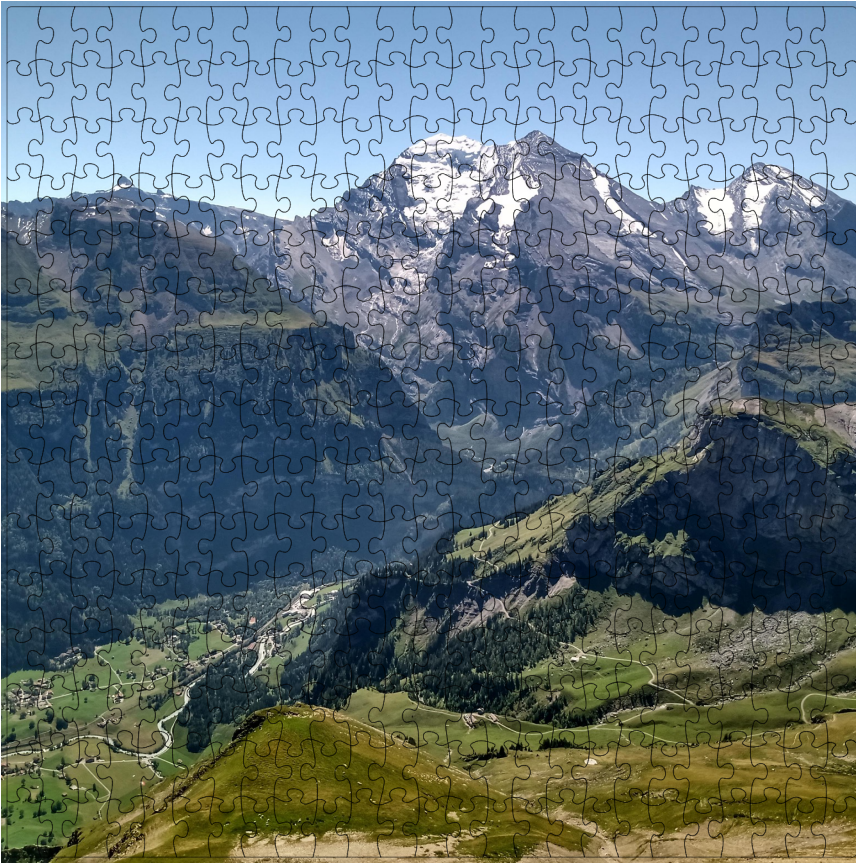# Long-read sequence analysis

Applications

Why long reads?

Applications

- (Genome) assembly
- Variant analysis
- Transcriptome analysis
- Epigenetics
- Metagenomics

Assembly

- Reconstructing a DNA sequence out of fragmented sequences:
  - Better understand variations between and within species
  - Reduces resources and increases accuracy for many applications
- Longer reads:
  - Less assembly errors
  - Higher contiguity (longer contigs)
  - Lower computational resources required

LR assembly software

- Microbial:
  - Flye
  - Miniasm
  - Unicycler
  - Trycycler

- Large(r) genomes:
  - Shasta (ONT only)
  - Falcon (PacBio only)
  - Canu
  - Flye
  - Hifiasm
  - IPA
  - Peregrine
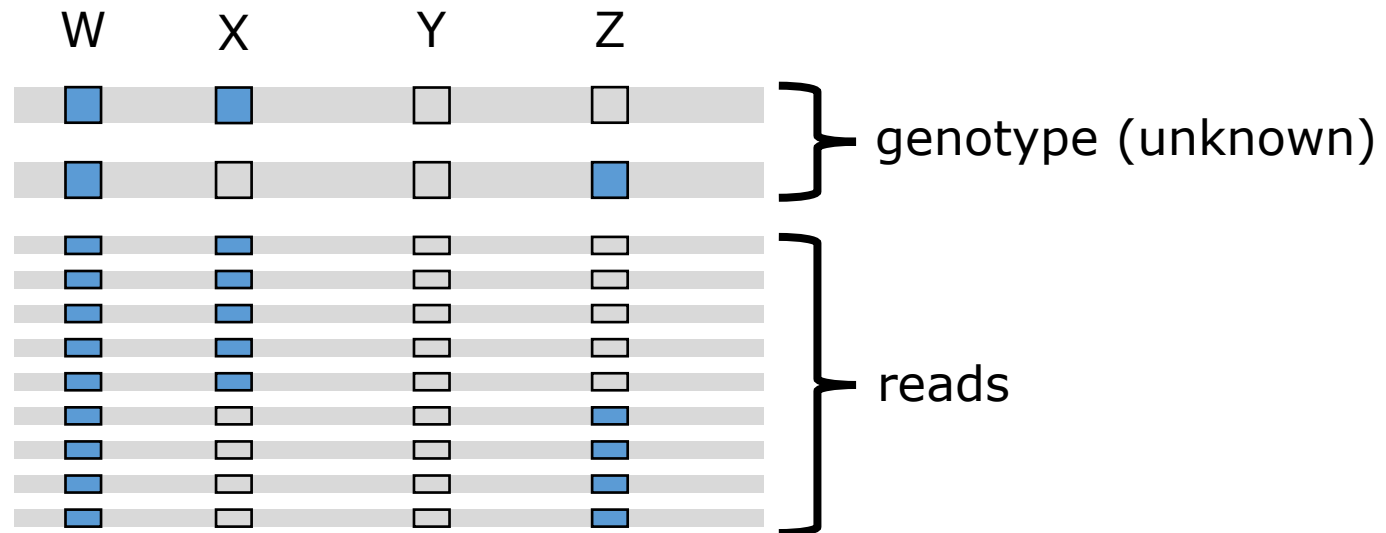 - Verkko (integration PacBio and ONT)

Variant analysis

- Low accuracy long reads: not particularly good at variant analysis

- HiFi: good for large and short variation

- Major applications
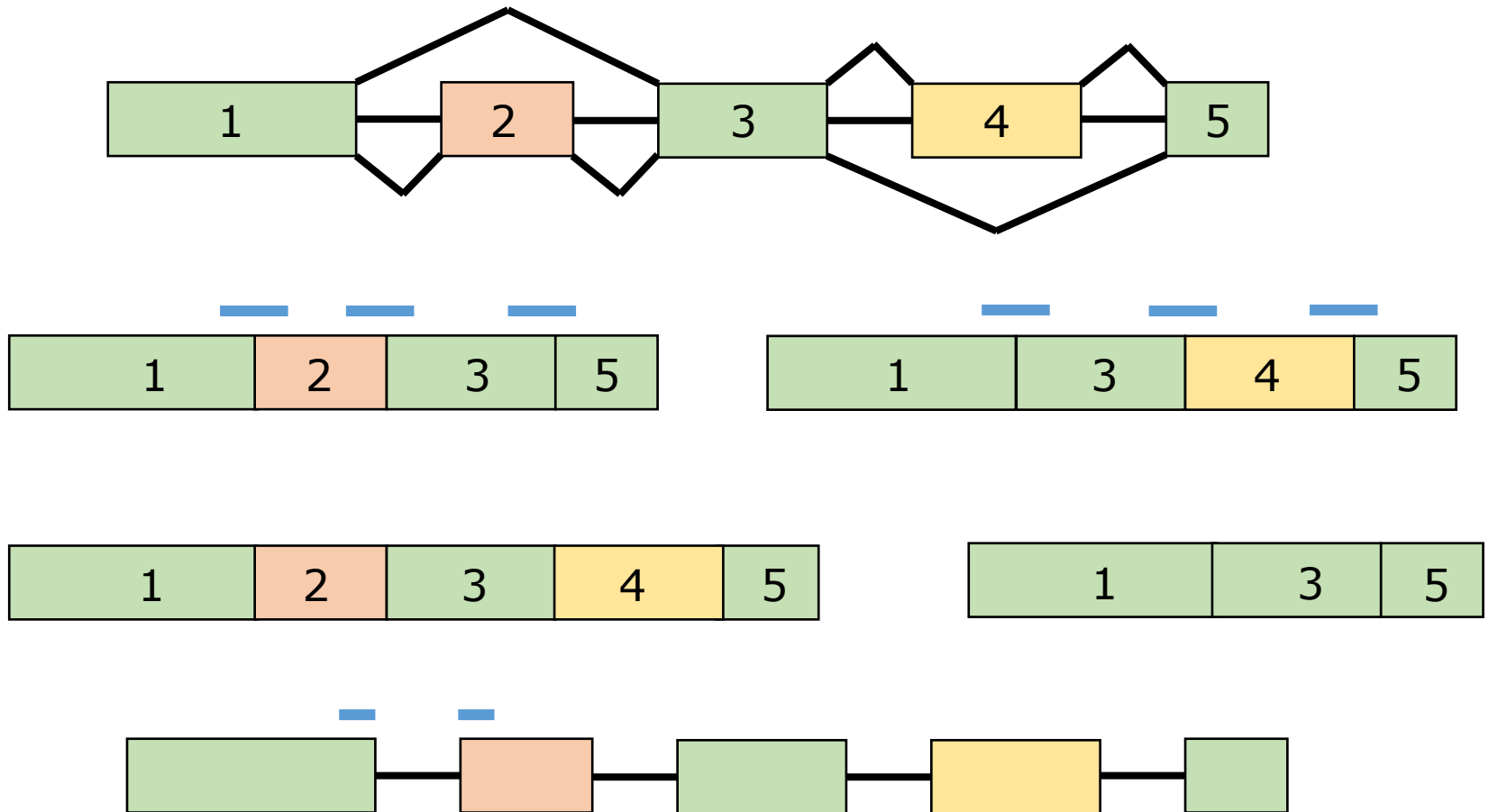 - Structural variation/large repeats
  - Phasing of variants
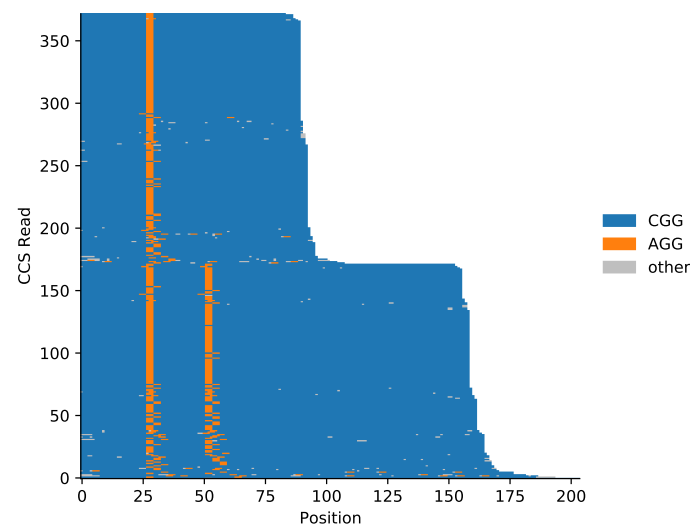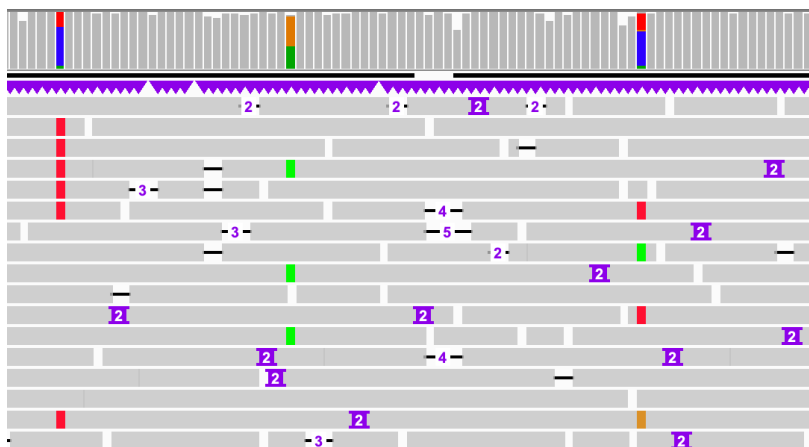
Phasing

- Variant calling:
  - DeepVariant (Pacbio + ONT)
  - Clair3 (ONT)
- Phasing: Whatshap

# Repeat expansion

Transcriptome analysis

Transcriptome analysis

- Full transcripts: less ambiguous identification + quantification of transcript isoforms
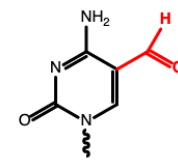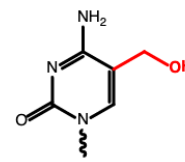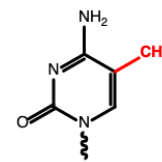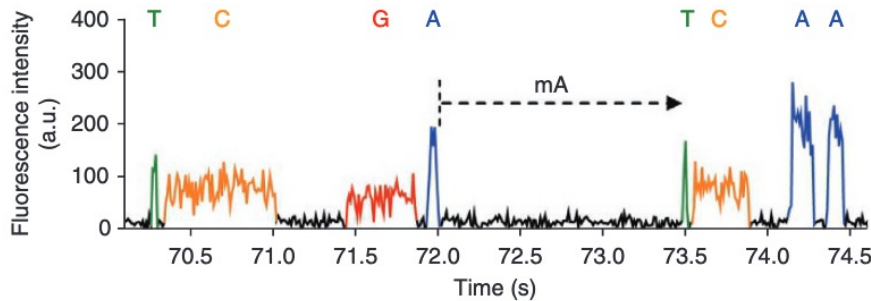
- Better gene model prediction (e.g. for genome annotation)

- ONT direct RNA-seq:
  - No PCR amplification
  - No reverse transcription
  - Base modification identification

Epigenetics

- Base modification in non-amplified libraries (Pacbio and ONT)
- No multiplexing
- Also possible in direct RNA-seq (ONT)

https://www.nature.com/articles/nmeth.1459.pdf
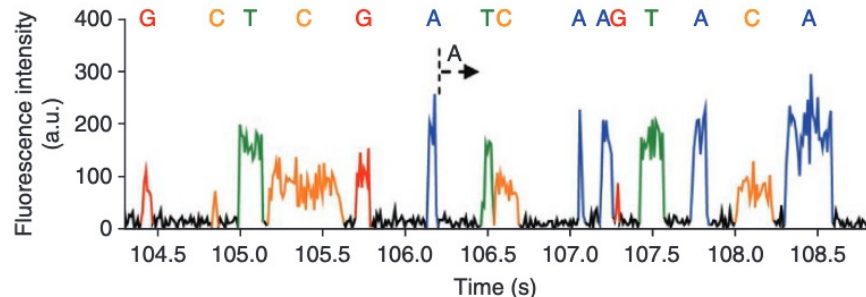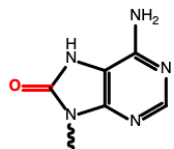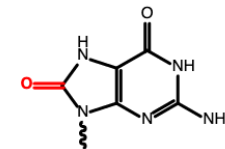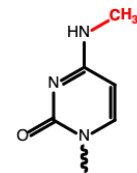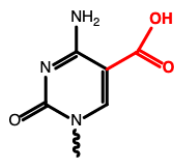
Metagenomics

- Full length 16S genes: better classification (PacBio CCS)

- Metagenome Assembled Genomes (MAGs) -> `flye --meta`

Downstream analysis

- ONT:
  - Workflows on EPI2ME
- PacBio:
  - pbbioconda
- nf-core
  - https://nf-co.re/bacass
  - https://nf-co.re/viralrecon
  - https://nf-co.re/isoseq
  - https://nf-co.re/mag
  - https://nf-co.re/nanoseq
  - https://nf-co.re/ampliseq