
DTU

22126: Next Generation Sequencing Analysis
DTU - January 2026
Mick Westbury

Mick Westbury
Associate Professor
Section of Bioinformatics
Technical University of Denmark
micwe@dtu.dk

VARIANT FILTERING

NGS Analysis workflow

Question

Raw data

Pre-processing

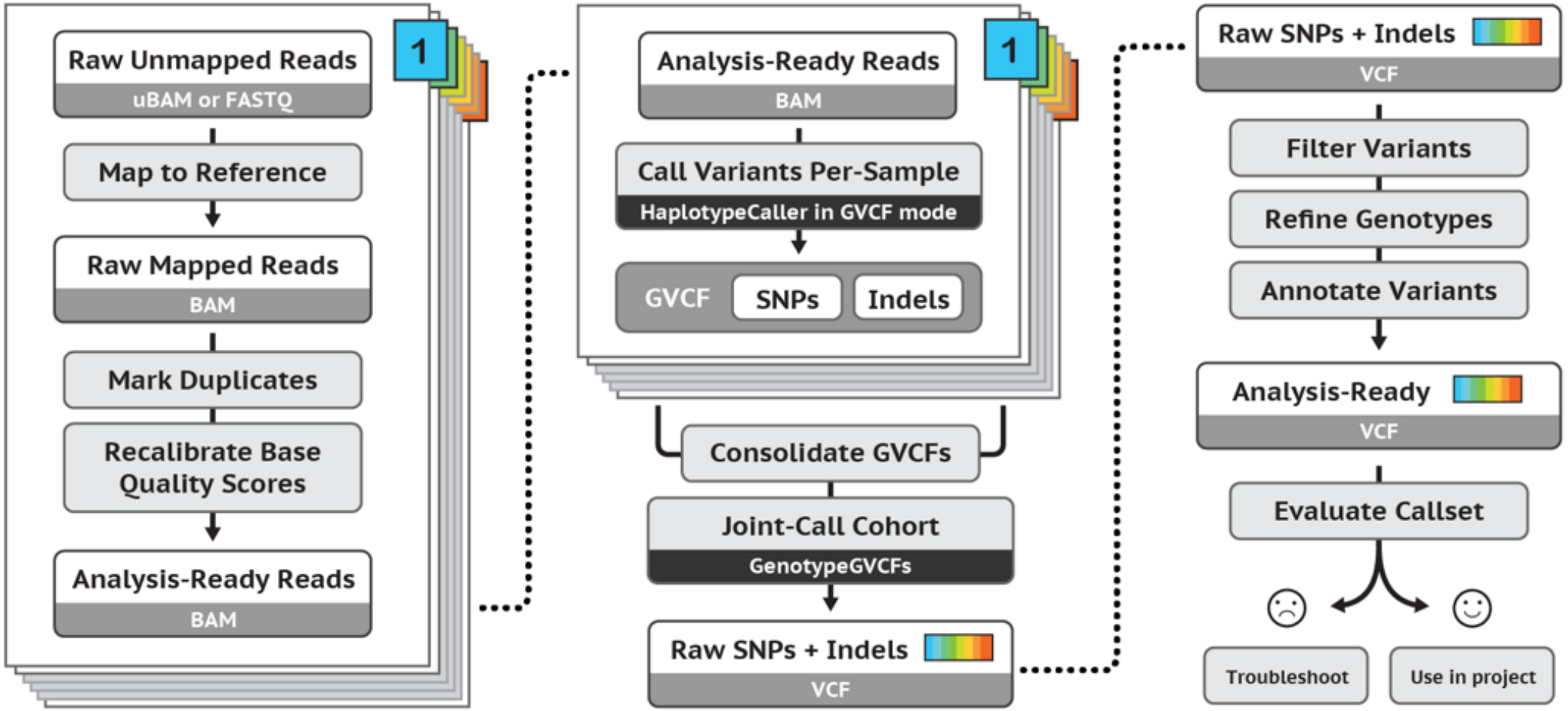
Assembly – mapping or de novo

Variant calling

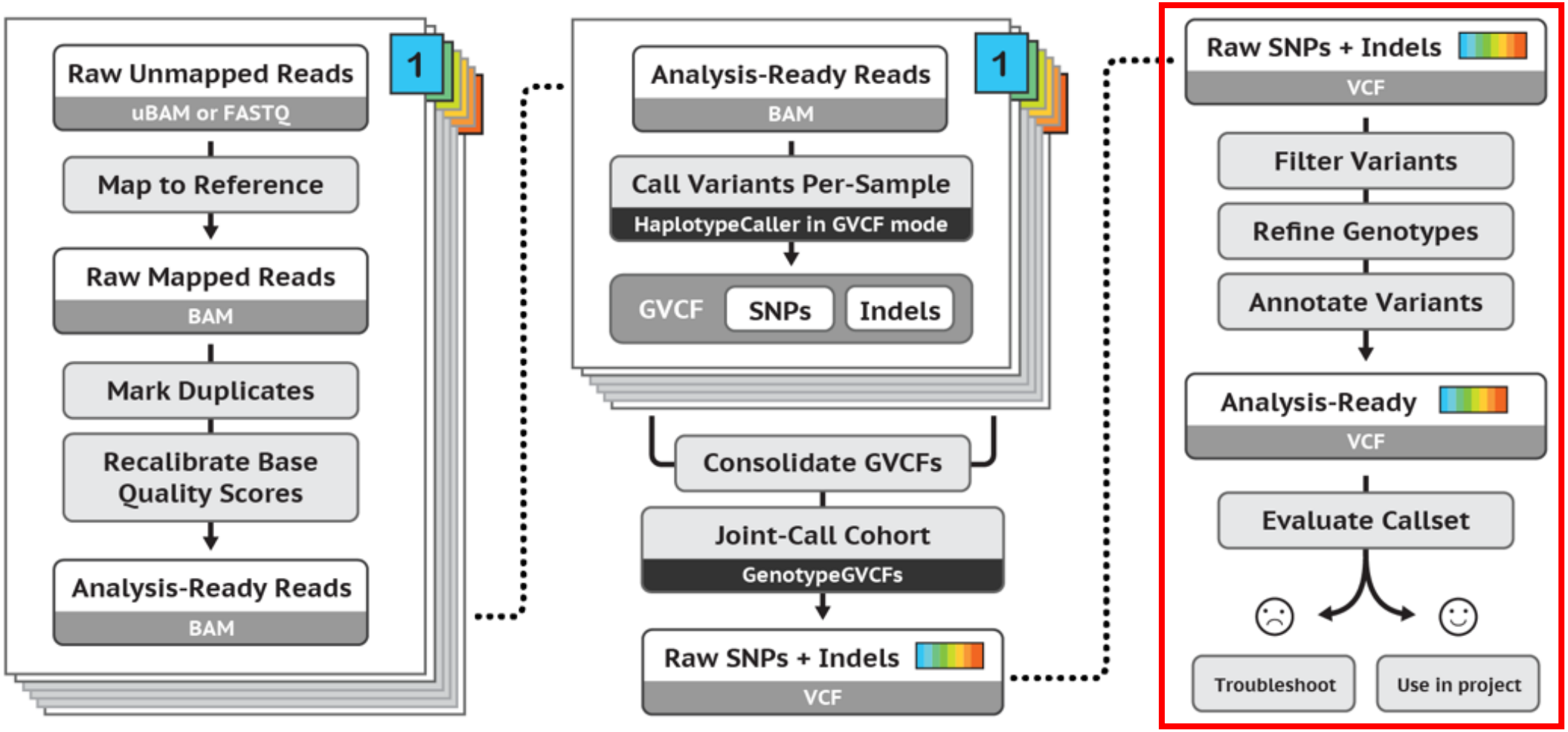
Post-processing

Comparison

Answer



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



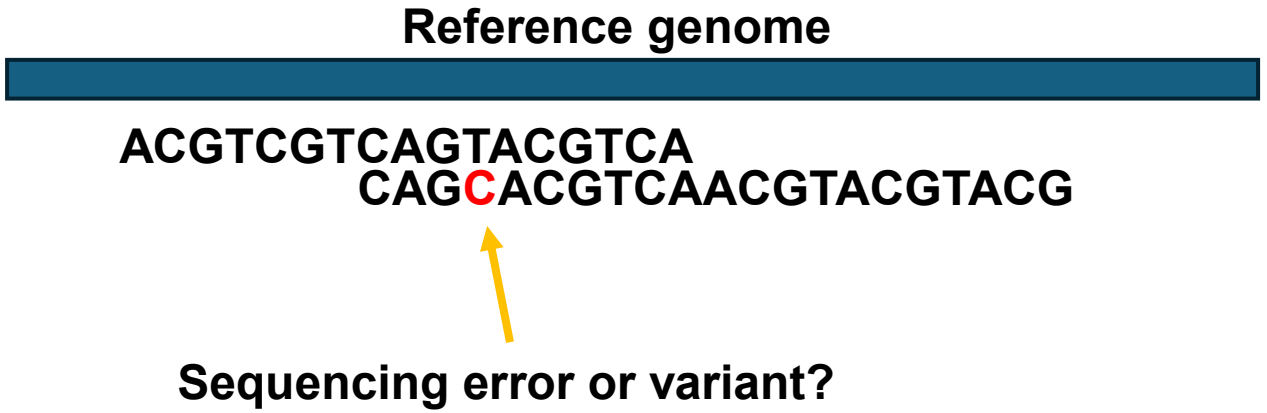
Best Practices for SNP and Indel discovery in germline DNA - leveraging groundbreaking methods for combined power and scalability.

Why Post-process?

- Raw VCFs contain many false positives
- Causes of false positives:
  - Low depth
  - Extremely high depth
  - Alignment artifacts
  - Repetitive regions

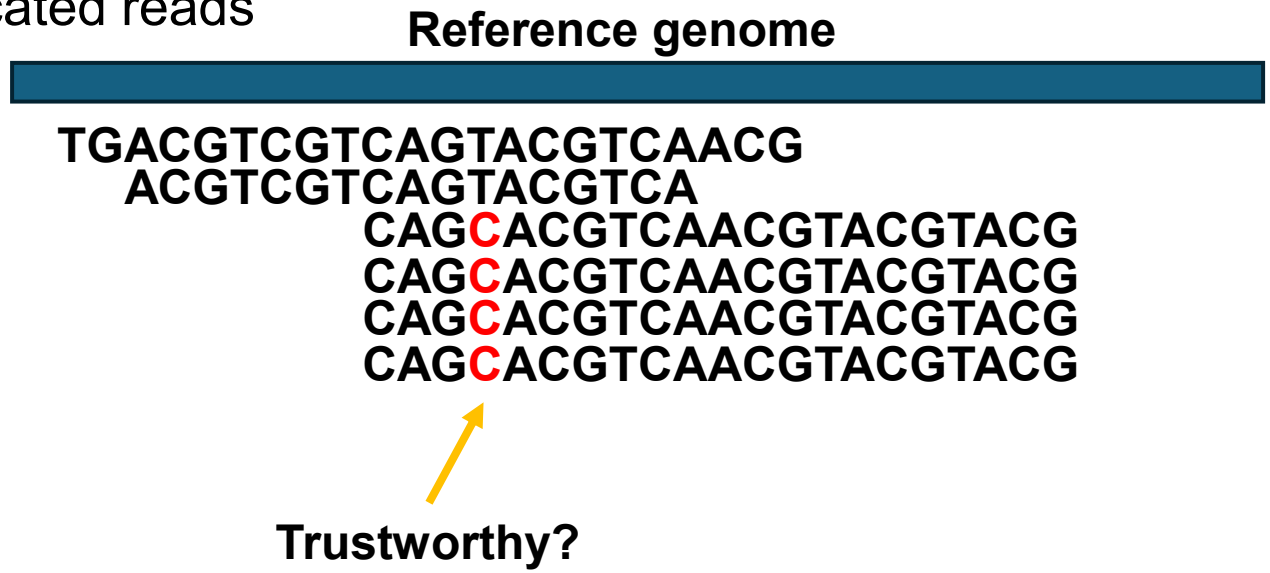
Why Post-process?

- Raw VCFs contain many false positives
- Causes of false positives:
  - **Low depth**
    - Insufficient evidence → random sequencing errors called as variants
  - Extremely high depth
  - Alignment artifacts
  - Repetitive regions



Why Post-process?

- Raw VCFs contain many false positives
- Causes of false positives:
  - Low depth
  - **Extremely high depth**
    - Often indicates mis-mapping or duplicated reads
  - Alignment artifacts
  - Repetitive regions



Why Post-process?

- Raw VCFs contain many false positives
- Causes of false positives:
  - Low depth
  - Extremely high depth
  - **Alignment artifacts**
    - Indels, soft-clipping, local misalignment near variants
  - Repetitive regions

Why Post-process?

- Raw VCFs contain many false positives
- Causes of false positives:
  - Low depth
  - Extremely high depth
  - Alignment artifacts
  - **Repetitive regions**
    - Reads map ambiguously → inflated depth and spurious variants

Why Post-process?

- Raw VCFs contain many false positives
- Causes of false positives:
  - Low depth
  - Extremely high depth
  - Alignment artifacts
  - Repetitive regions

Reflection prompt (1 min):
What's worse: missing a real variant (false negative) or believing a false variant (false positive)? Why?

Goals of Variant Post-processing

- Remove technical artifacts
  - Filter variants caused by sequencing, PCR, or alignment errors

Goals of Variant Post-processing

- Remove technical artifacts
  - Filter variants caused by sequencing, PCR, or alignment errors
- Retain biologically plausible variants
  - Consistent with expected allele balance, depth
  - Respect organism biology (ploidy, heterozygosity, mutation rate)

Goals of Variant Post-processing

- Remove technical artifacts
  - Filter variants caused by sequencing, PCR, or alignment errors
- Retain biologically plausible variants
  - Consistent with expected allele balance, depth
  - Respect organism biology (ploidy, heterozygosity, mutation rate)
- Increase precision with minimal sensitivity loss
  - Reduce false positives while keeping true variants
  - Accept that some true variants may be filtered

Variant Post-processing Approaches

- Hard filtering
- Variant Quality Score Recalibration (VQSR)

Hard Filtering

- Apply fixed, user-defined thresholds to variant-level metrics
- Variants failing any threshold are removed
- Key properties
  - Simple and transparent
  - Easy to reproduce
  - No model training required

Hard Filtering Metrics

- Variant quality (QUAL, QUAL/DP)
  - Confidence of the variant call relative to depth
- Depth (DP)
  - Remove low-support and abnormally high-coverage variants
- Mapping quality (MQ) / mappability
  - Poor or ambiguous read placement
- Read / base quality metrics
  - Low-quality evidence for the alternate allele

Mappability

- Measures uniqueness of read placement in the genome
- Low mappability → repetitive regions → ambiguous mapping
- Causes inflated depth and false-positive variants
- Often handled using genome masks or mappability tracks

Hard Filtering Limitations

- Thresholds depend on the data and project
  - Coverage, library prep, organism, ploidy
  - Assumes one set of cutoffs fits all variants
- Introduces systematic bias
  - Against indels
  - Against low-frequency variants
  - Against variants in difficult regions
- Can reduce sensitivity if applied too aggressively

Variant Quality Score Recalibration

- Model-based variant filtering (GATK)
- Learns characteristics of true variants from known datasets
- Assigns a probability-based quality score to each variant
- Improves precision–recall balance compared to hard filtering
- Requires large datasets and high-quality truth sets

Variant Quality Score Recalibration

- Model-based variant filtering (GATK)
- Learns characteristics of true variants from known datasets
- Assigns a probability-based quality score to each variant
- Improves precision–recall balance compared to hard filtering
- Requires large datasets and high-quality truth sets

Reflection prompt:
What kind of datasets is this NOT suitable for? Why?

Genotype Refinement

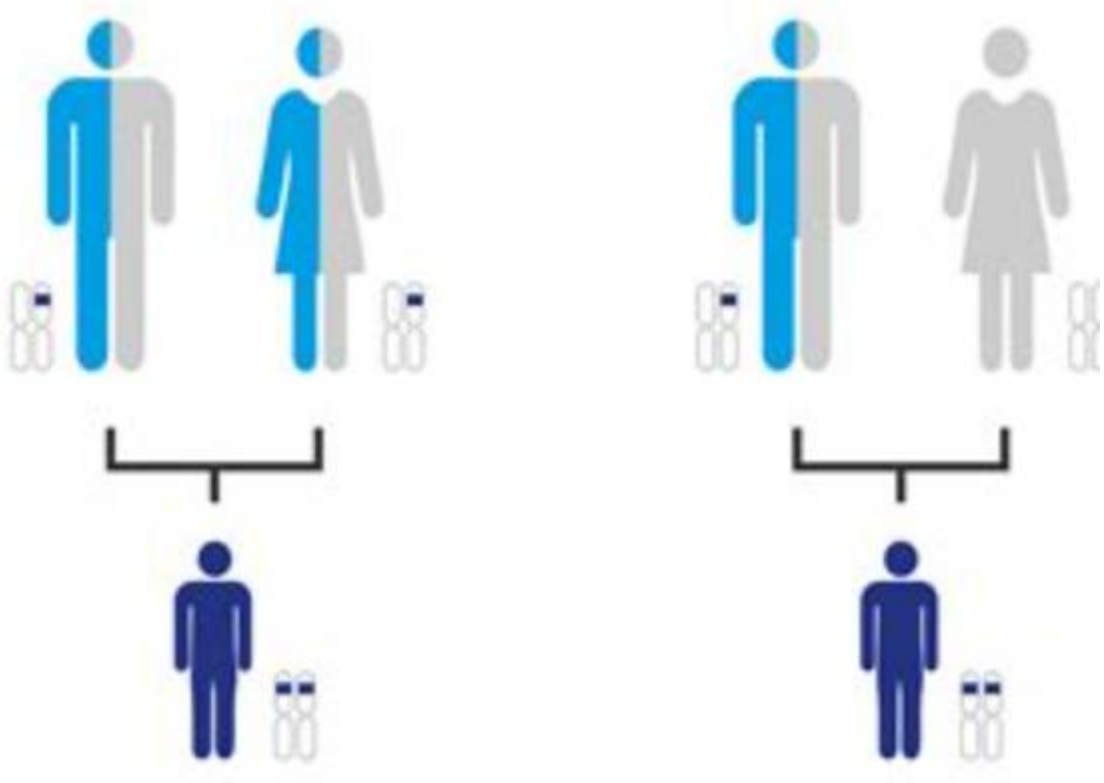
- Improve genotype accuracy after variant calling
- Resolve uncertain or low-confidence genotype assignments
- Key idea
  - Combine sequencing evidence with biological constraints and prior information

Genotype Refinement

- Family-based refinement
  - Enforces Mendelian consistency in trios and pedigrees
  - Corrects genotypes inconsistent with inheritance
- Population-based refinement
  - Uses allele frequency priors
  - Penalizes unlikely genotypes given population context
- Recalculation of genotype likelihoods
  - Updates genotype probabilities using:
    - Depth, base quality, allele balance

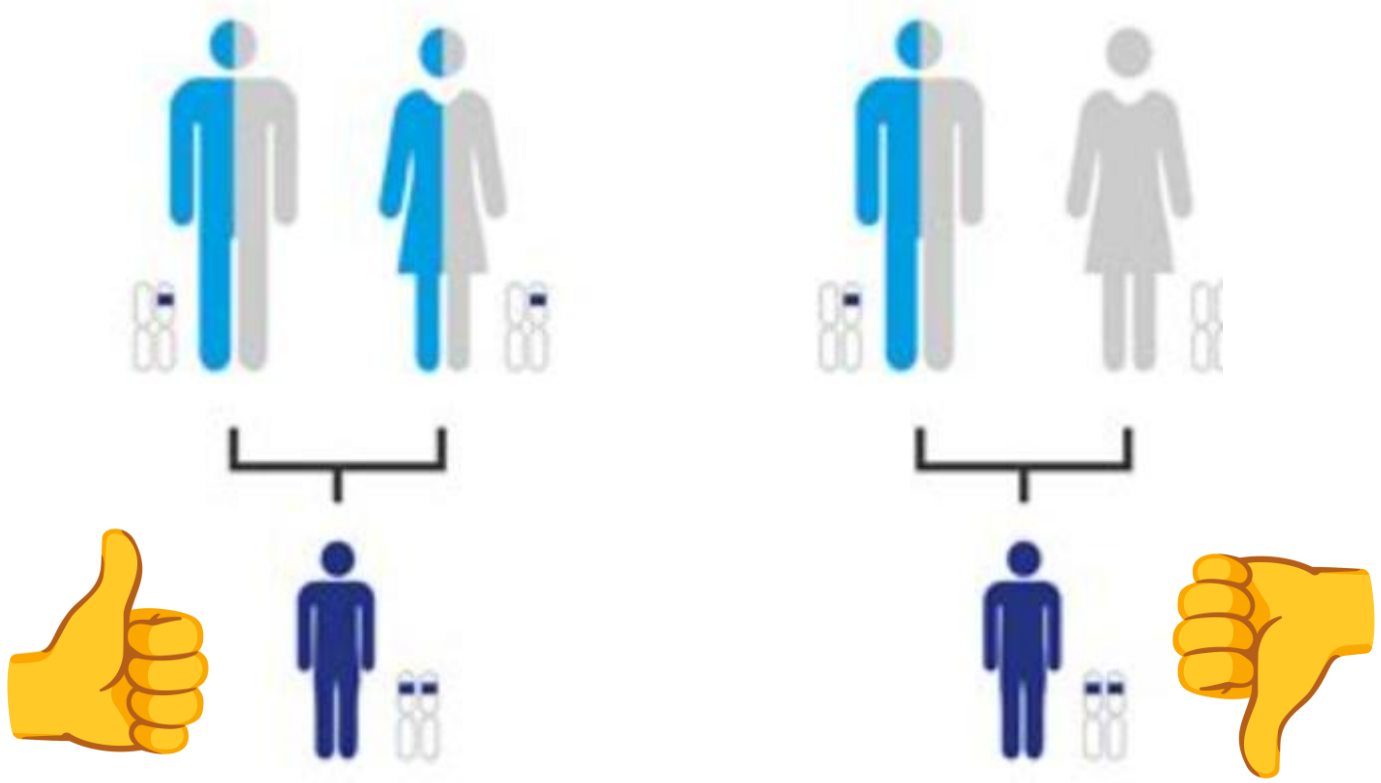
Genotype Refinement

- Family-based refinement



Genotype Refinement

- Family-based refinement



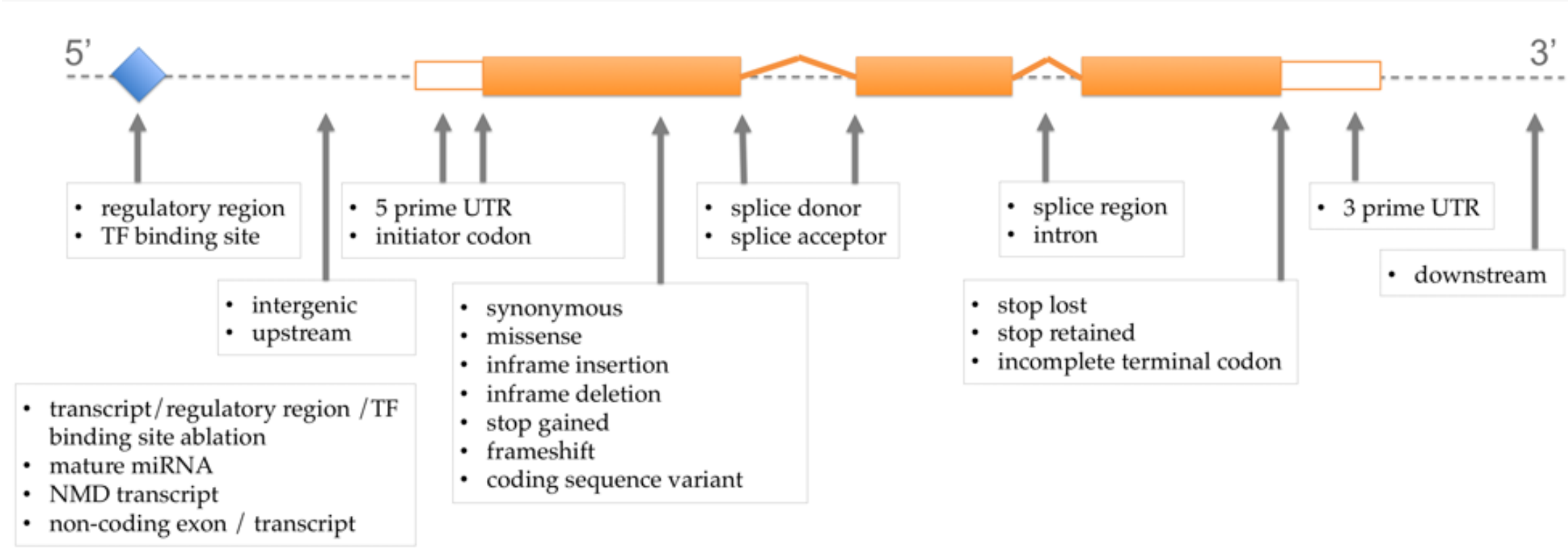
Genotype Refinement

- Requires accurate pedigree or population information
- Population priors can bias:
  - Rare variants
  - Population-specific alleles
- Limited benefit for:
  - Single samples
  - Small datasets
  - Non-model organisms

Variant Annotation

- Understand biological relevance
  - Clinical Significance (pathogenic or not)
  - Effects on protein
  - Drug-response interpretation
  - Enable filtering & prioritisation of variants in analysis pipelines

Variant Annotation



Variant Annotation

- Protein functional categories
  - –Synonymous (same amino acid)
  - –Missense (difference amino acid)
  - –Nonsense (loss of function)
- Impact predictions
- Splice-site and regulatory variants

Variant Annotation

- Protein functional categories
  - –Synonymous (same amino acid)
  - –Missense (difference amino acid)
  - –Nonsense (loss of function)
- Impact predictions
- Splice-site and regulatory variants

Reflection prompt:
A variant is classified as ‘likely pathogenic’—what additional information would convince you it really causes disease?

ETHICS

Ethics in Variant Analysis

- **Privacy and re-identification**
  - Genomic data is inherently identifiable
  - Even “anonymised” variant data can enable re-identification
  - Data sharing requires strict access control and consent
- **Population ancestry inference**
  - Variants are often correlated with ancestry
  - Can lead to misinterpretation or misuse
  - Risk of reinforcing biological determinism or social bias

Interpretation/reporting responsibility

- **Actionable vs non-actionable variants**
  - *Actionable*: established clinical relevance and available intervention
  - *Non-actionable*: uncertain significance or no available treatment
  - Many variants fall into variants of uncertain significance (VUS)
- **Reporting guidelines**
  - Not all detected variants should be reported
  - Clinical reporting follows strict standards
  - Over-reporting increases anxiety and misinterpretation

Reflection prompt

- If you discovered a potentially pathogenic variant in yourself, would you want to know? Why or why not?
- Points to consider
  - Psychological impact
  - Medical usefulness
  - Implications for family members
  - Right to know vs right *not* to know

Not only medical related…

A prosecutor reveals new details about the capture of one of America's most notorious serial killers

UPDATED NOV 20, 2025

By Faith Karimi

Not only medical related…

- Golden state killer
  - Series of violent crimes in California (1970s–1980s)
  - Genetic approach
    - DNA recovered from historical crime scene evidence
    - Profile uploaded to a public genealogy database
    - No direct match to the suspect
  - Familial matching
    - Partial matches to distant relatives
    - Construction of extended family trees

Not only medical related…

- Golden state killer
  - Key ethical issues
    - Relatives' genetic data used without their consent
    - Identification possible even if the individual never shared DNA
  - Original intent of data use:
    - Genealogy and recreation
    - Not law enforcement
  - Broader implications
    - Demonstrates re-identification risk
    - Genomic data affects families, not just individuals
    - Blurred boundaries between:
      - Consumer genomics, Research, Forensic use