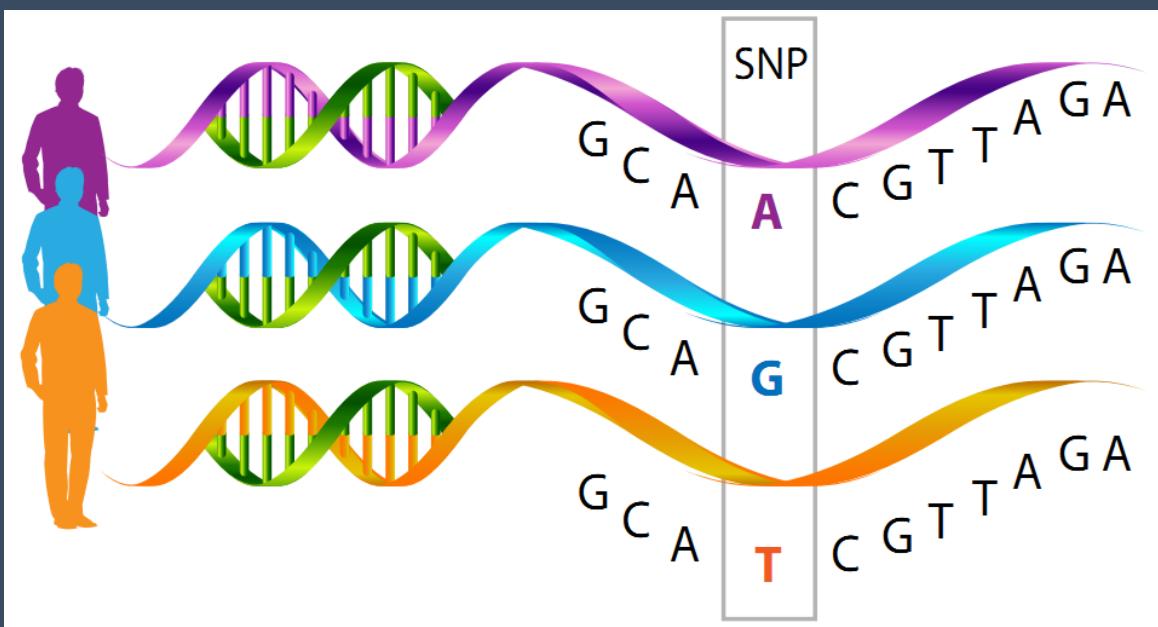




**HBC**  
Harvard Chan Bioinformatics Core

# Concepts and considerations for variant analysis in clinical genomics



Elizabeth Partan, PhD  
Harvard Chan Bioinformatics Core  
May 20, 2025

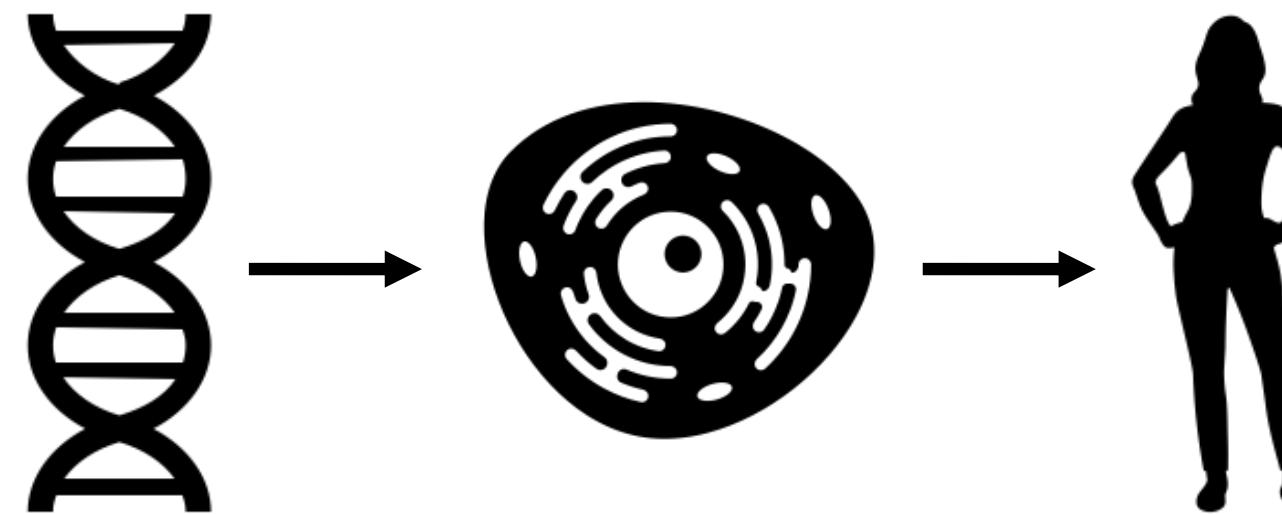
# Outline

- Background
- DNA sequencing methods
- Reference genomes
- Variant analysis workflow overview
- Cancer-specific considerations
- Variant annotations and prioritization
- Example research application
- Implications and societal considerations

# Outline

- **Background**
- DNA sequencing methods
- Reference genomes
- Variant analysis workflow overview
- Cancer-specific considerations
- Variant annotations and prioritization
- Example research application
- Implications and societal considerations

# DNA sequencing allows us to understand and treat disease

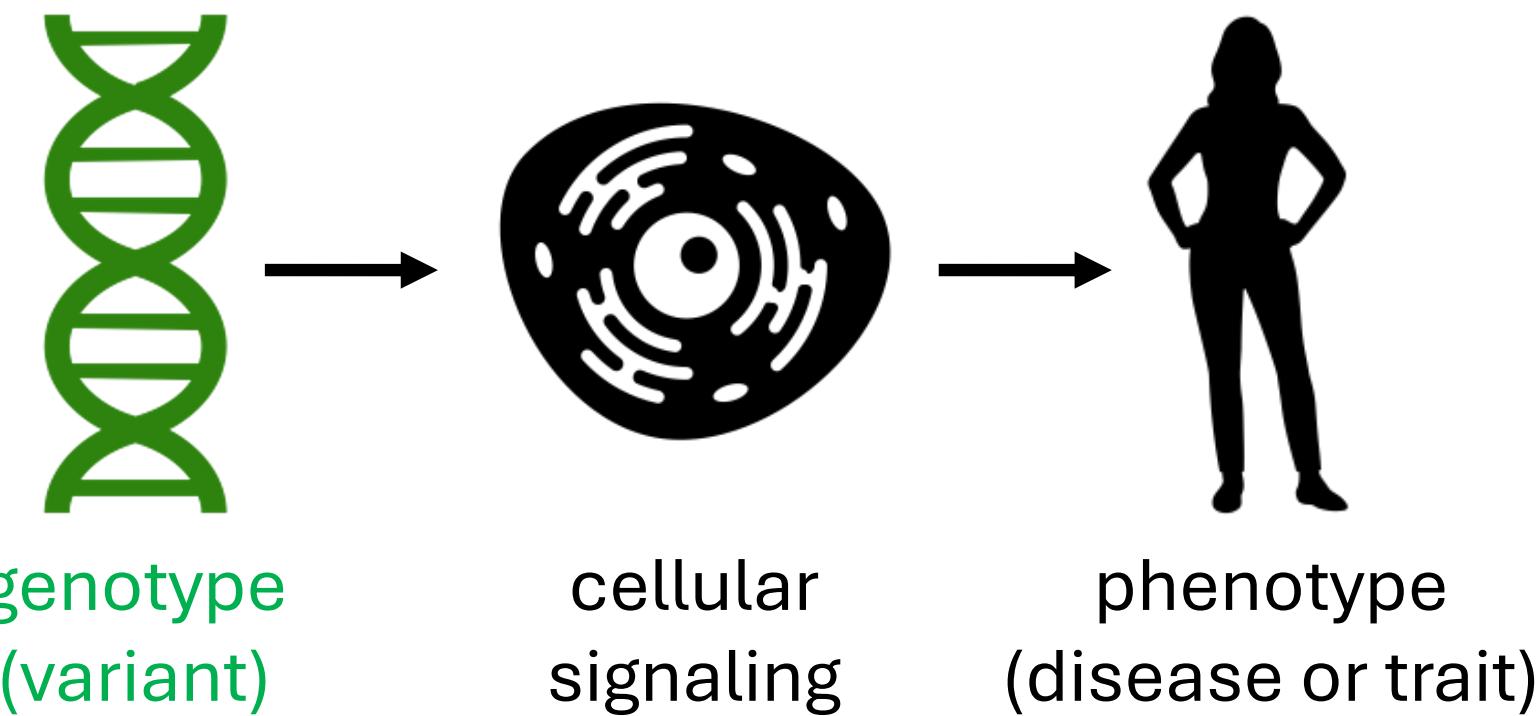


genotype  
(variant)

cellular  
signaling

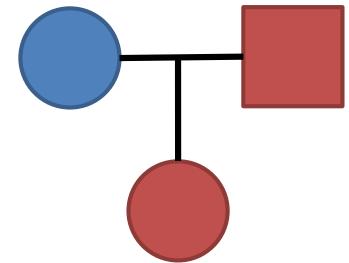
phenotype  
(disease or trait)

# DNA sequencing allows us to understand and treat disease



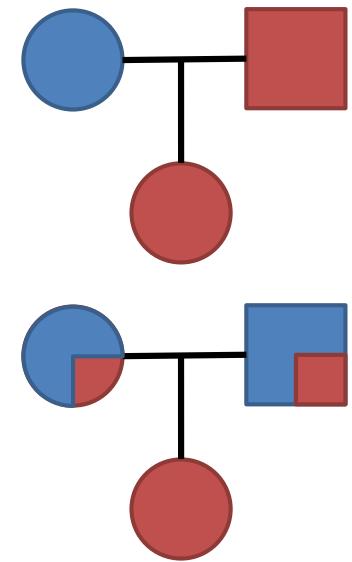
# Variant and disease inheritance

- **Germline:** Variant occurs throughout the body
  - Dominant: Inherited from an affected parent, only need one variant to cause disease



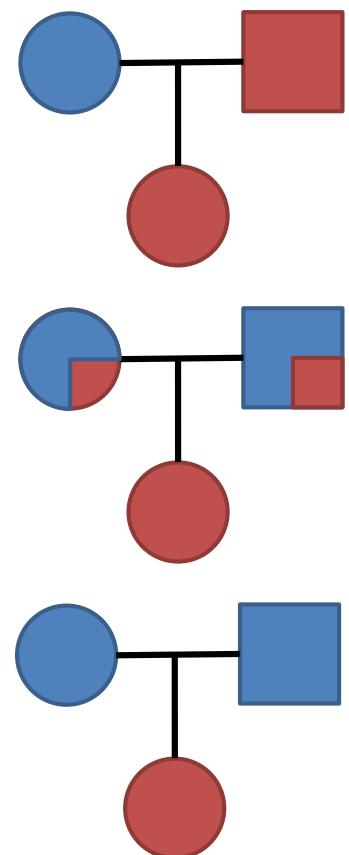
# Variant and disease inheritance

- **Germline:** Variant occurs throughout the body
  - Dominant: Inherited from an affected parent, only need one variant to cause disease
  - Recessive: Inherited from two unaffected parents (carriers), need two variants



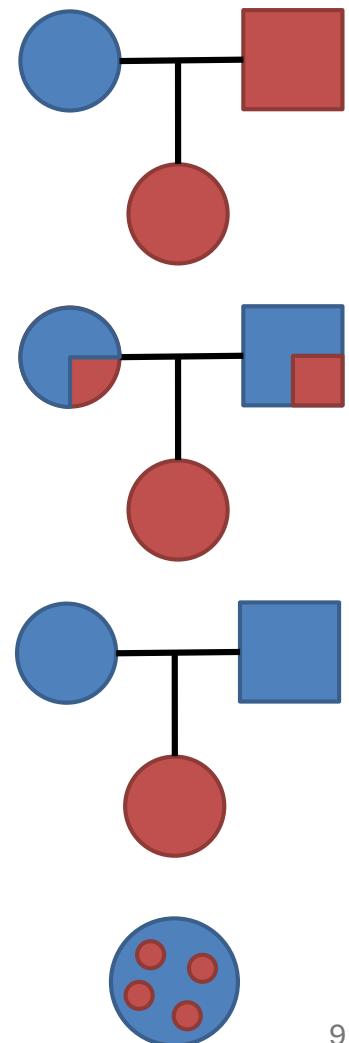
# Variant and disease inheritance

- **Germline:** Variant occurs throughout the body
  - Dominant: Inherited from an affected parent, only need one variant to cause disease
  - Recessive: Inherited from two unaffected parents (carriers), need two variants
  - *De novo:* Occurs for the first time in the child
    - Mutation during egg/sperm formation or in early embryogenesis
    - Can then be inherited in either pattern

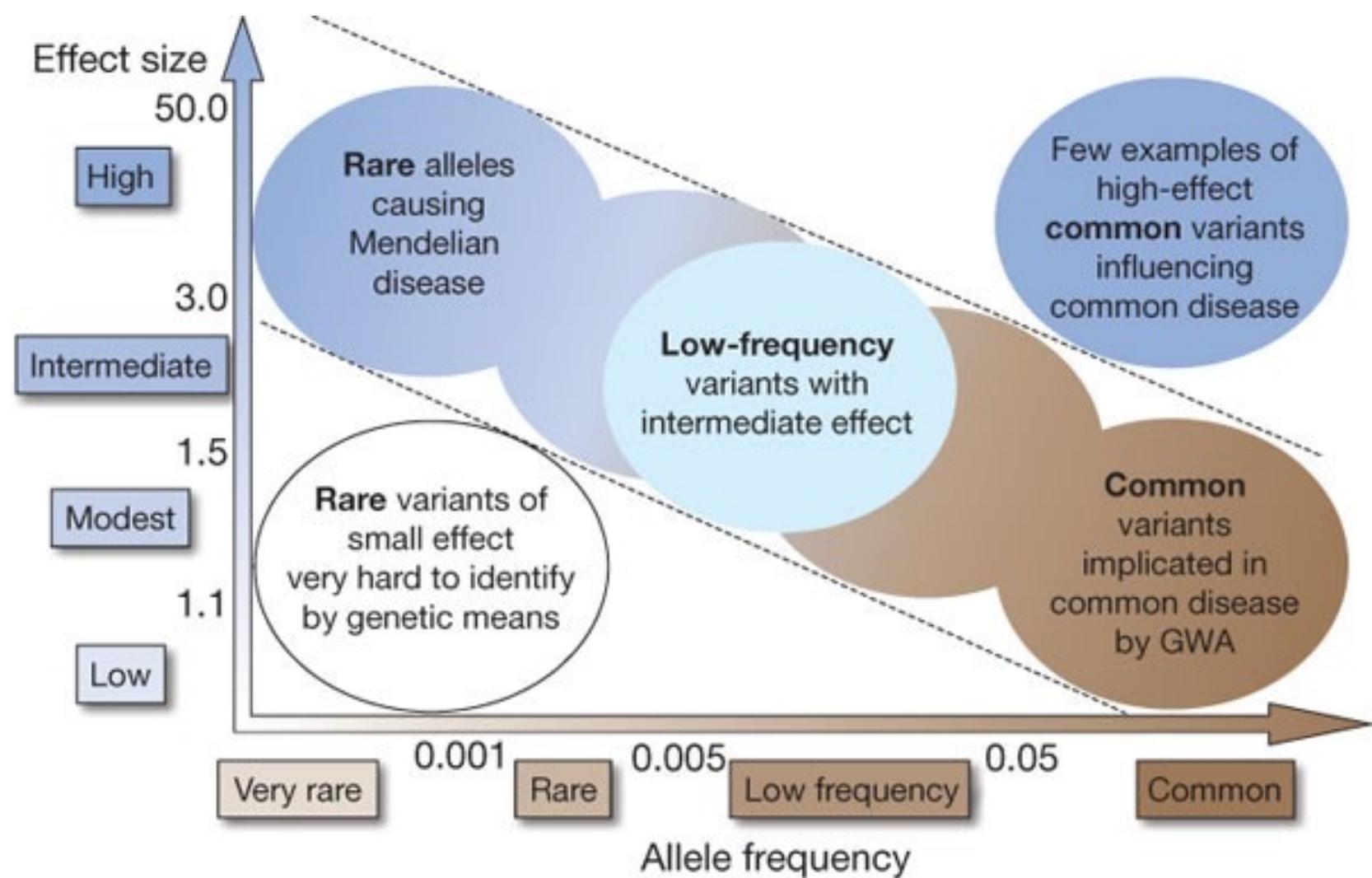


# Variant and disease inheritance

- **Germline:** Variant occurs throughout the body
  - Dominant: Inherited from an affected parent, only need one variant to cause disease
  - Recessive: Inherited from two unaffected parents (carriers), need two variants
  - *De novo:* Occurs for the first time in the child
    - Mutation during egg/sperm formation or in early embryogenesis
    - Can then be inherited in either pattern
- **Somatic (mosaic):** Variant only occurs in some cells
  - Mutation occurs later, not generally inheritable



# Types of genetic disease: Mendelian ↔ complex

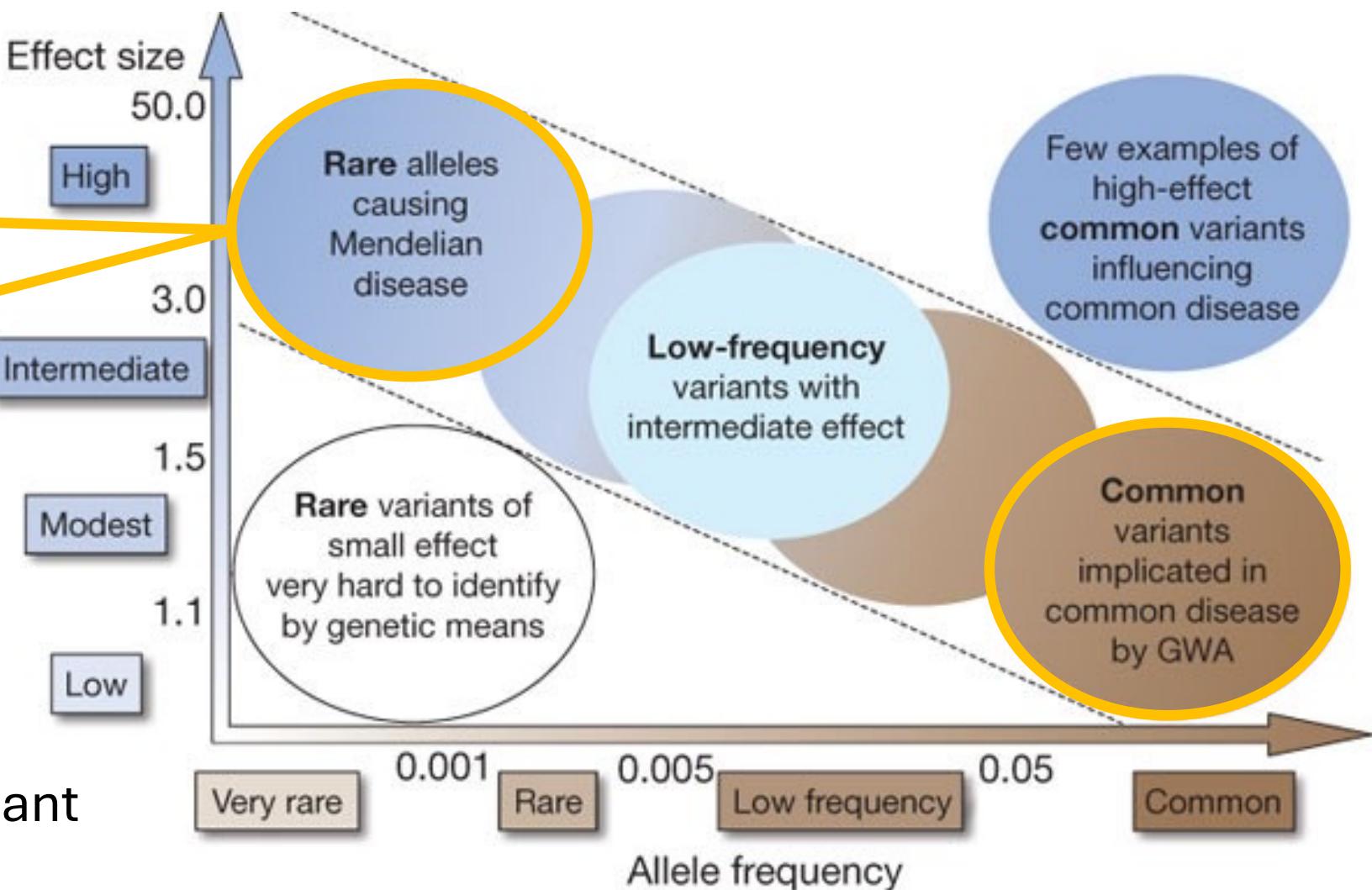


# Types of genetic disease: Mendelian ↔ complex

Huntington's

Phenylketonuria

Mendelian disease:  
single gene / variant



Diabetes  
Heart disease  
Cancer

Complex disease:  
many variants + environment

# Outline

- Background
- **DNA sequencing methods**
- Reference genomes
- Variant analysis workflow overview
- Cancer-specific considerations
- Variant annotations and prioritization
- Example research application
- Implications and societal considerations

# How much of the DNA can we look at?

- Gene panels: Selected genes or variants
  - Usually disease-focused
- Exome: Exons, which encode proteins
  - 30-50 million base pairs (1-2% of the genome)
- Genome: Everything \*
  - 3.2 billion base pairs



panel



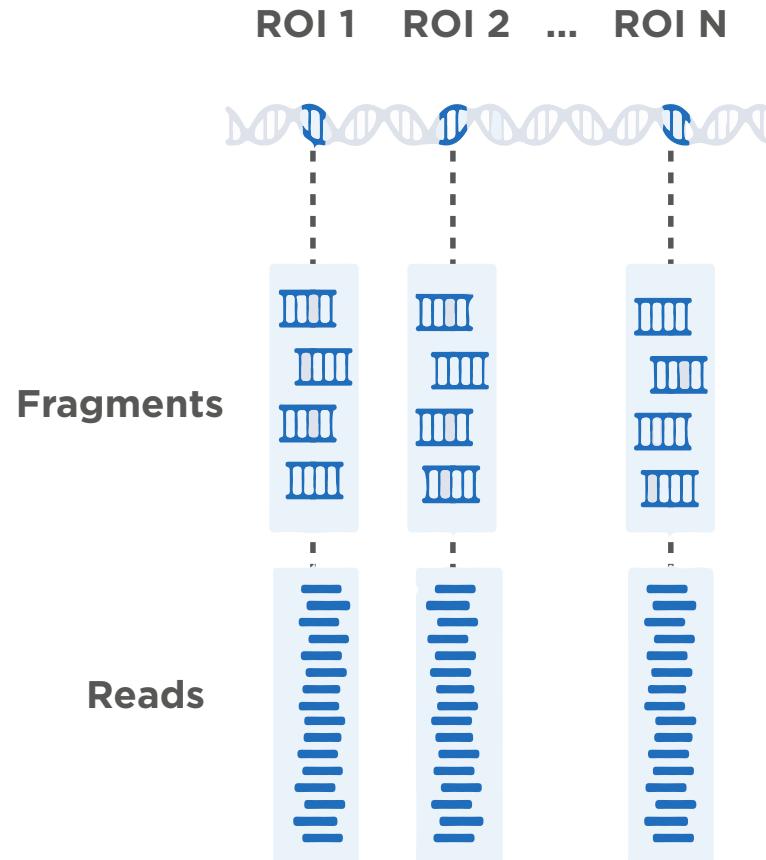
whole exome (WES)



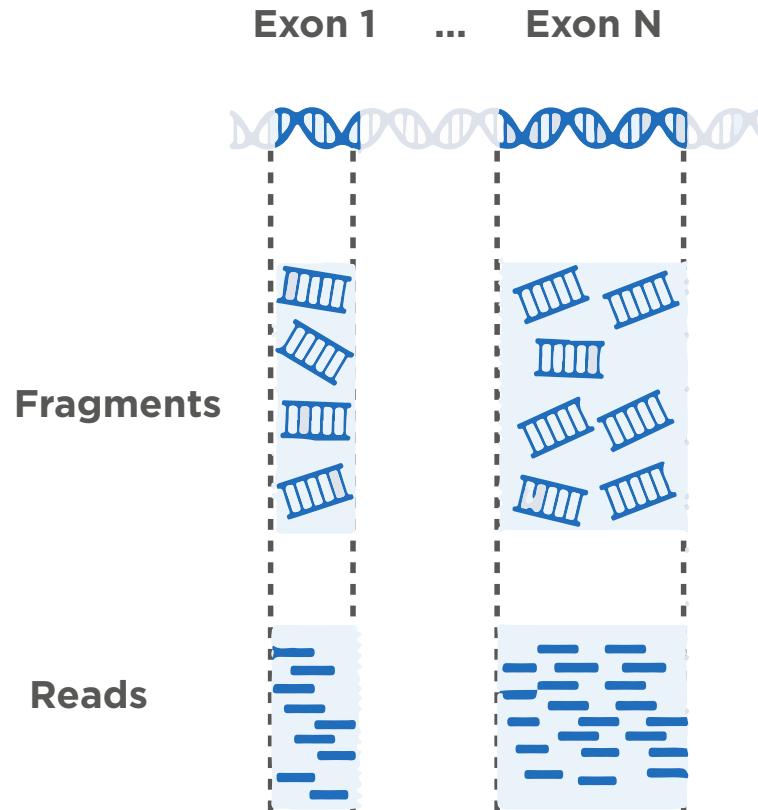
whole genome (WGS)

# How does DNA sequencing work?

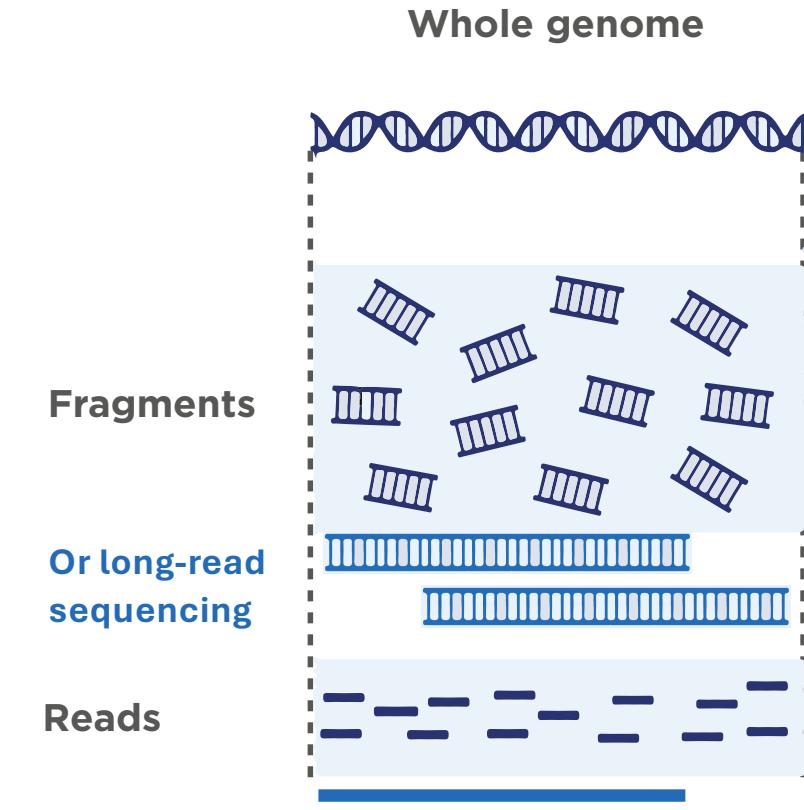
Targeted gene panel



Exome sequencing



Genome sequencing



# Which sequencing method is appropriate?

- Gene panels
  - Most common for clinical diagnostic analysis: cheapest, fastest, fewest variants of uncertain significance (VUS)
  - Need to know disease-associated variants in advance

# Which sequencing method is appropriate?

- Gene panels
  - Most common for clinical diagnostic analysis: cheapest, fastest, fewest variants of uncertain significance (VUS)
  - Need to know disease-associated variants in advance
- Exome
  - Multisystemic diseases not covered by a single panel
  - Will miss everything not captured (just exons, 3' and 5' UTR)

# Which sequencing method is appropriate?

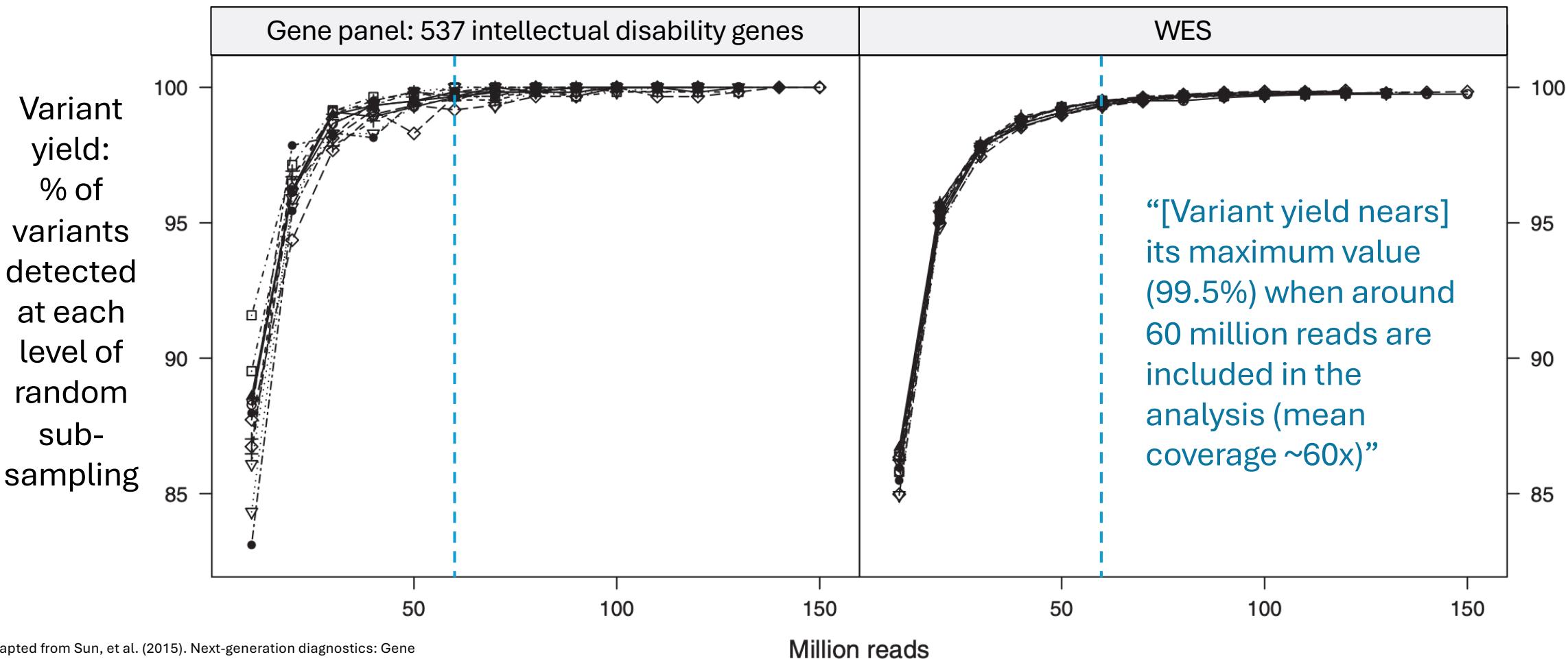
- Gene panels
  - Most common for clinical diagnostic analysis: cheapest, fastest, fewest variants of uncertain significance (VUS)
  - Need to know disease-associated variants in advance
- Exome
  - Multisystemic diseases not covered by a single panel
  - Will miss everything not captured (just exons, 3' and 5' UTR)
- Genome
  - Most data, but most expensive and most processing time
  - Many findings will be difficult to interpret: variants or genes of uncertain significance, including introns / intergenic regions

# Recommended sequencing depth



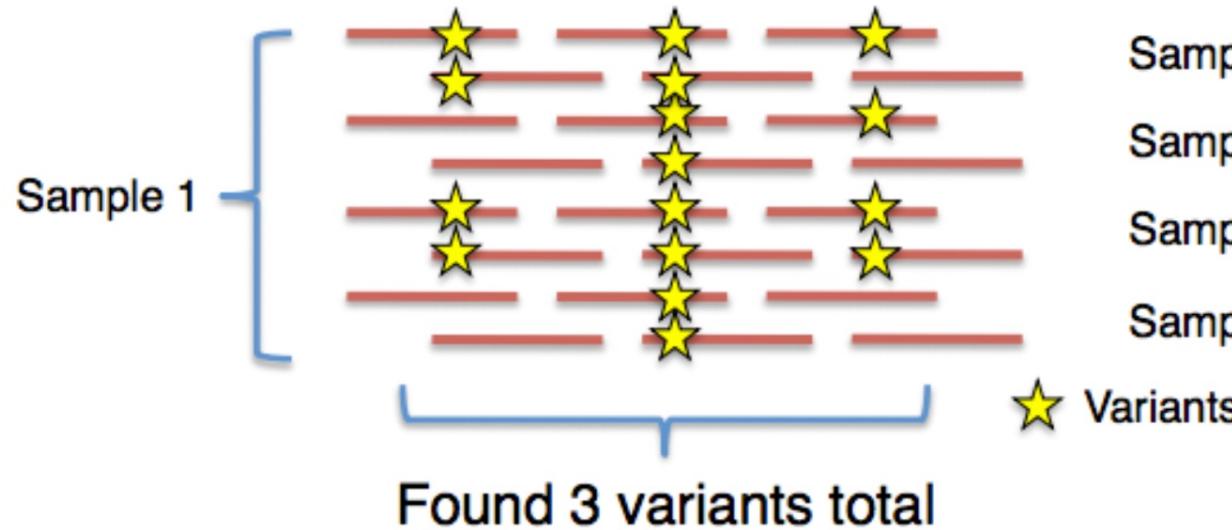
# Greater sequencing depth increases confidence but will not find new variants

Effect of additional sequencing on variant yield

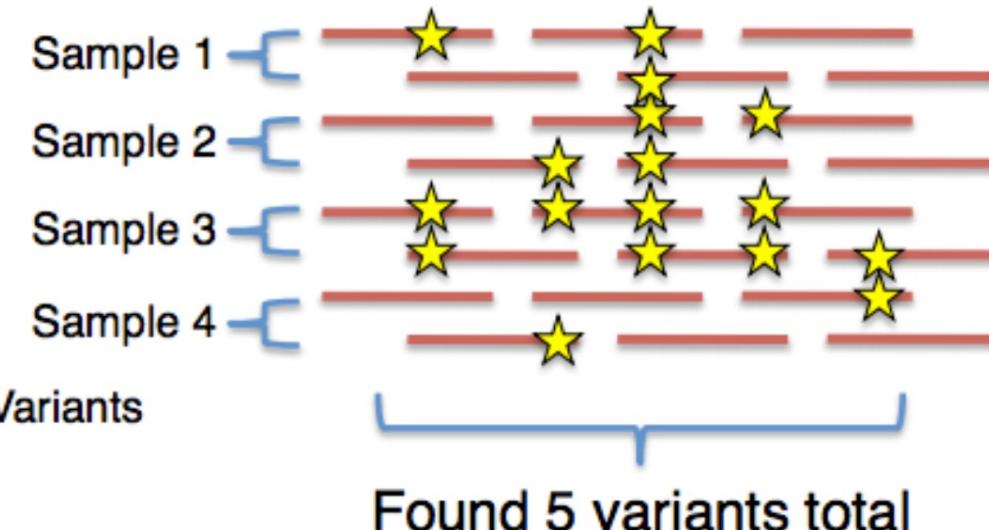


# Sequencing depth and cost: With a fixed sequencing budget, you can perform...

Deep single-sample sequencing



Shallower multi-sample sequencing

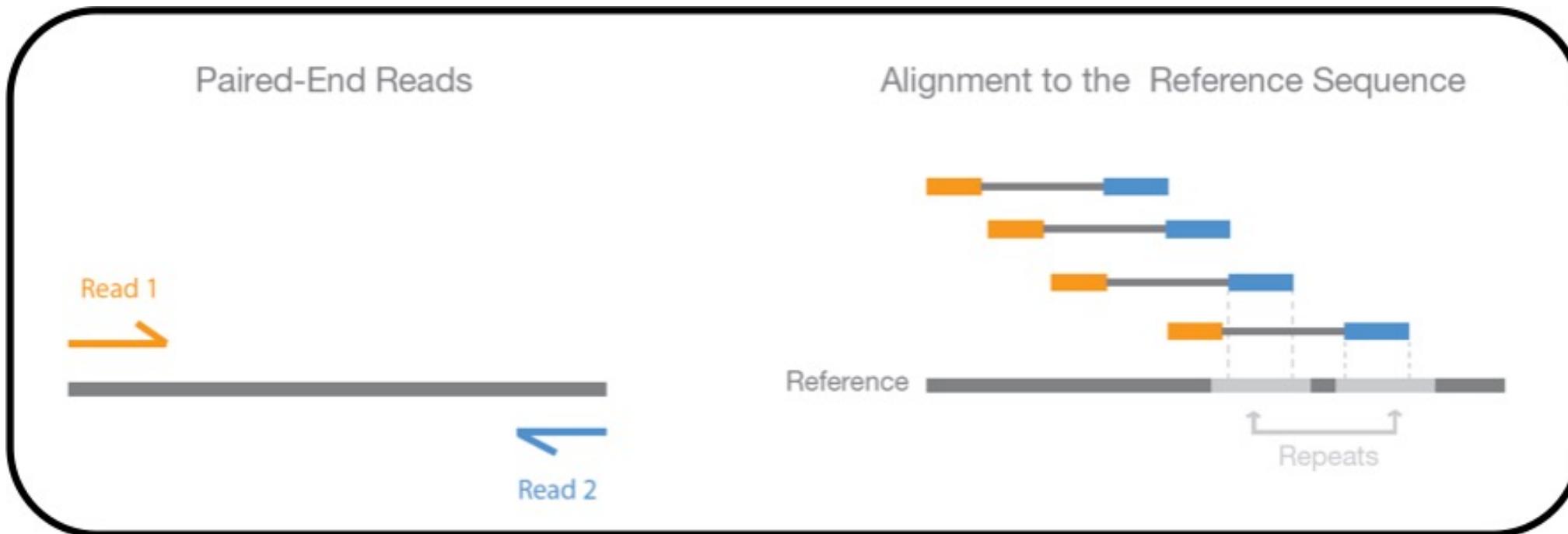


- Higher sensitivity for variants in the sample
- No information about other samples

- Sensitivity dependent on frequency of variation
- More total variants discovered

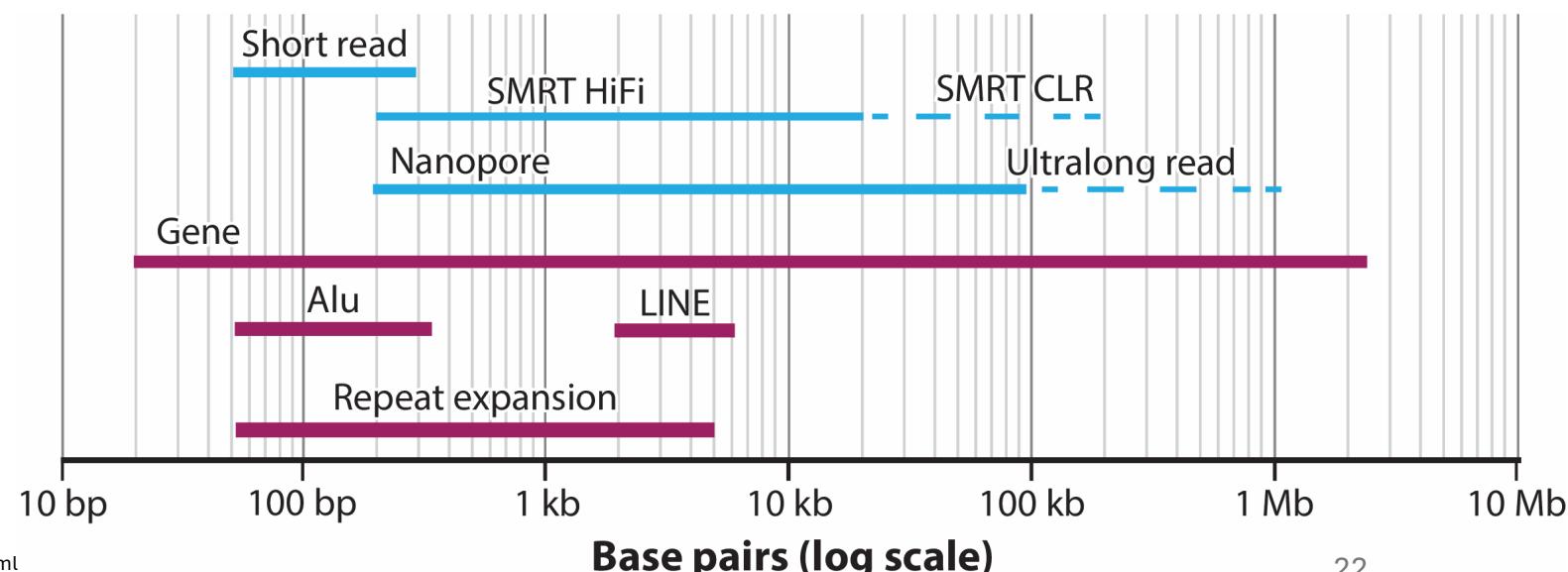
# Other experimental considerations: Single- vs paired-end sequencing

- Single-end is cheaper and is suitable for ChIPseq or small RNAseq
- Paired-end is better at resolving complex variants or alignments to repetitive sequences



# Other experimental considerations: Sequence lengths

- Short-read
  - Recommend 150bp for most DNA sequencing
  - Longer read lengths have lower quality, higher error rates (especially for reverse read)
- Long-read
  - Can resolve large, complex variants and repetitive regions
  - Accuracy may be lower than short-read



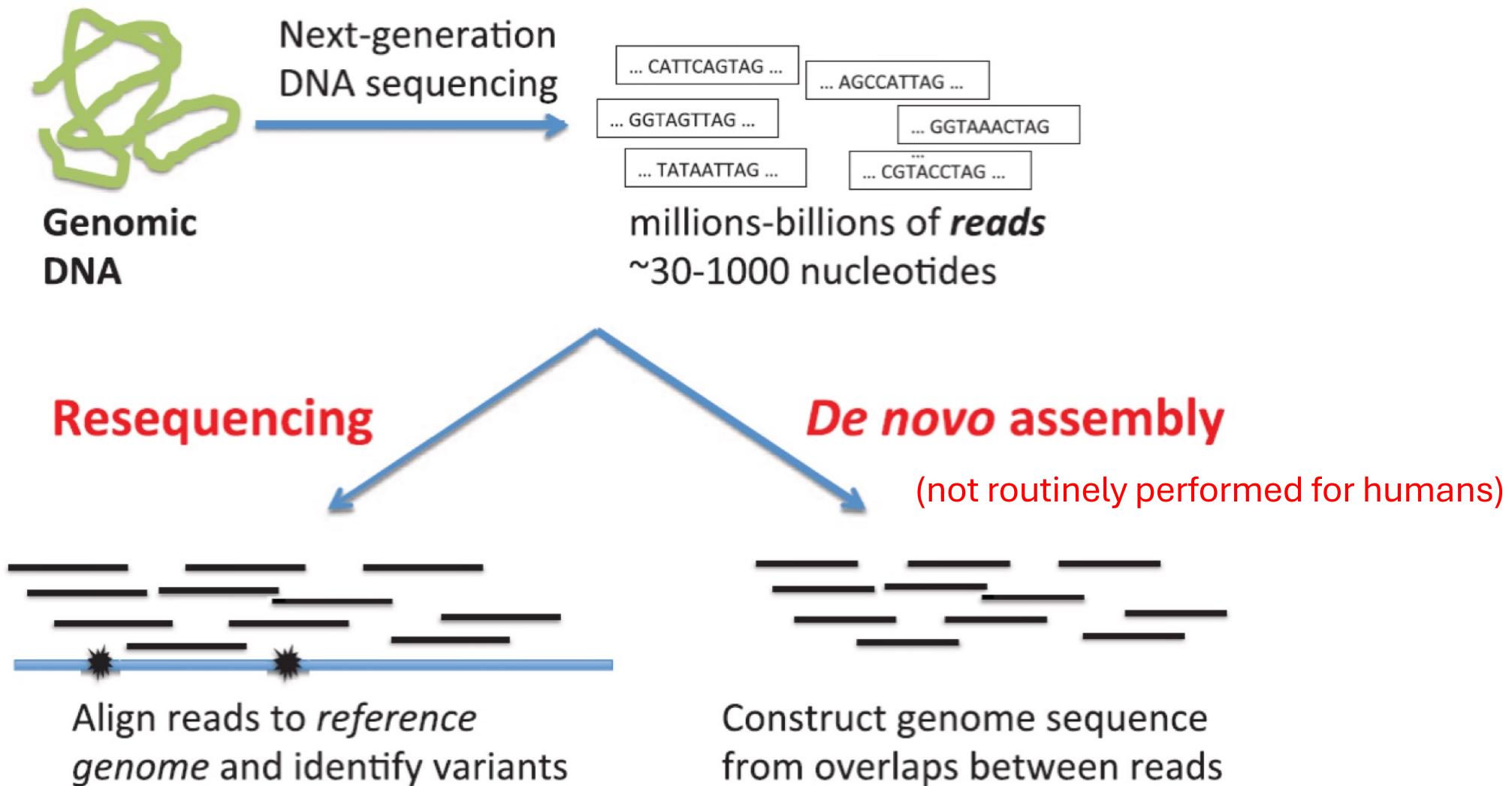
# Other experimental considerations: Long-read sequencing

- Requires long DNA stands
  - Be gentle when processing DNA
  - Avoid extraction methods that will mechanically fracture DNA
  - Use wide-bore tips when pipetting
- Requires high quality, high purity, high concentration DNA
  - Requires more input DNA than short-read sequencing
  - More sensitive to contaminants than short-read sequencing
  - Avoid DNA damage from, e.g., repeated freeze-thaw cycles, high temperatures, pH extremes

# Outline

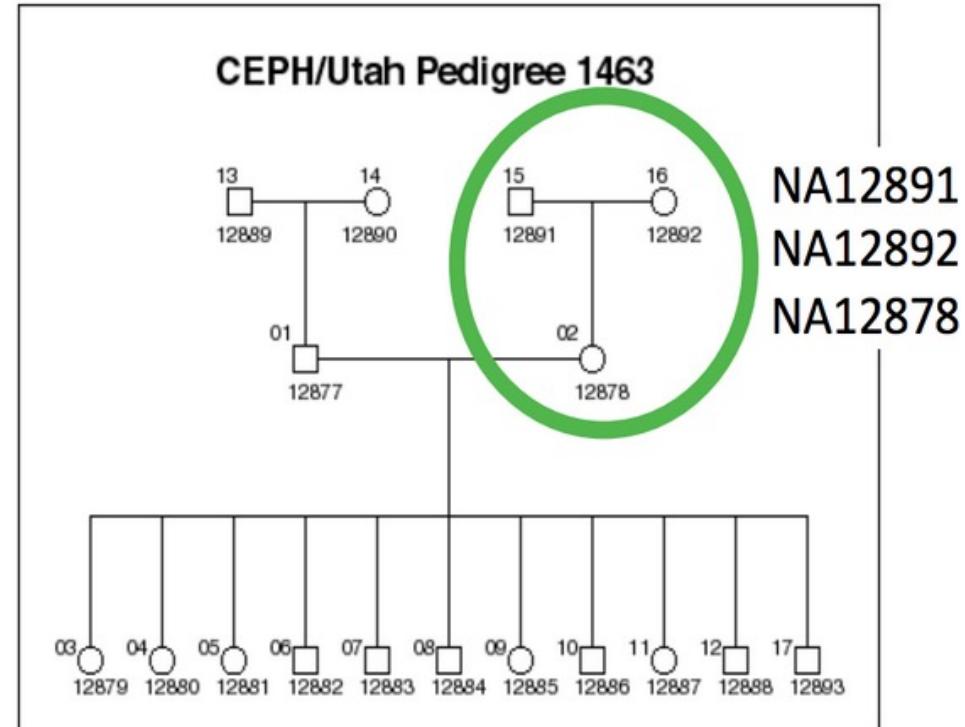
- Background
- DNA sequencing methods
- **Reference genomes**
- Variant analysis workflow overview
- Cancer-specific considerations
- Variant annotations and prioritization
- Example research application
- Implications and societal considerations

# Comparison to reference



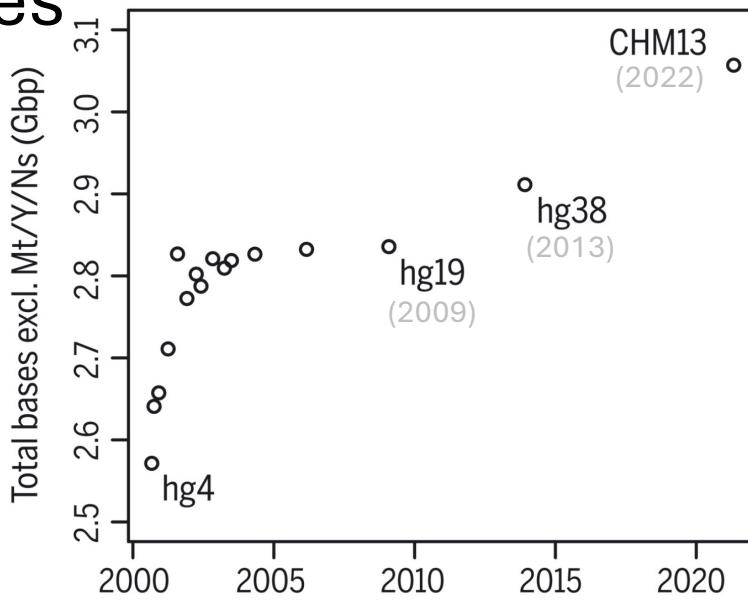
# Reference genomes: Individual humans

- Well-studied, gold standard
- Family (trio) data
- Used to benchmark sequencing platforms and analysis pipelines
- NA12891
- Genome in a Bottle (GIAB)
- HapMap
- 1000 Genomes

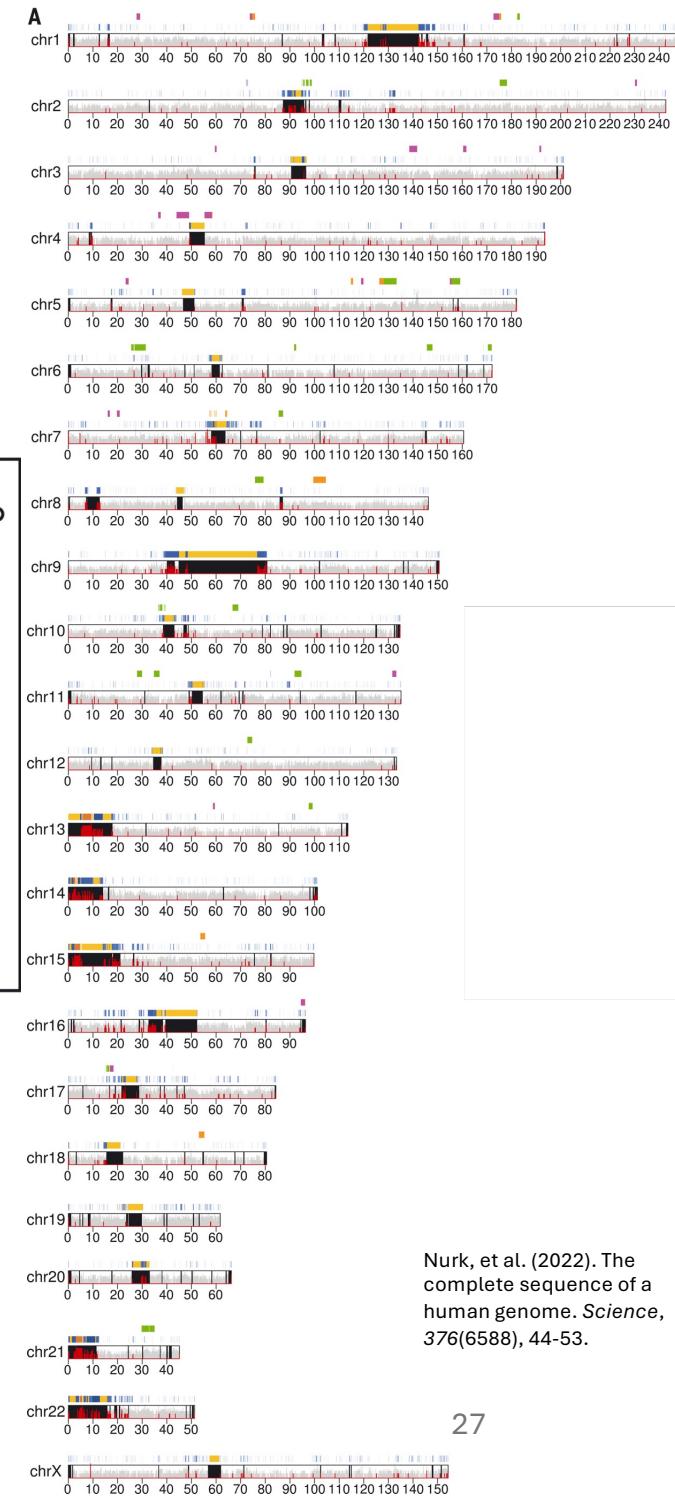


# Reference genomes: Many humans

- Built from individual references
- GRCh37 / hg19
- GRCh38 / hg38
- New: T2T / CHM13 / hs1
  - Telomere-to-telomere coverage using long reads
  - Improved repeat resolution
  - No Y chromosome
- Older genome builds still in use
  - Compare to older datasets
  - Tools not updated

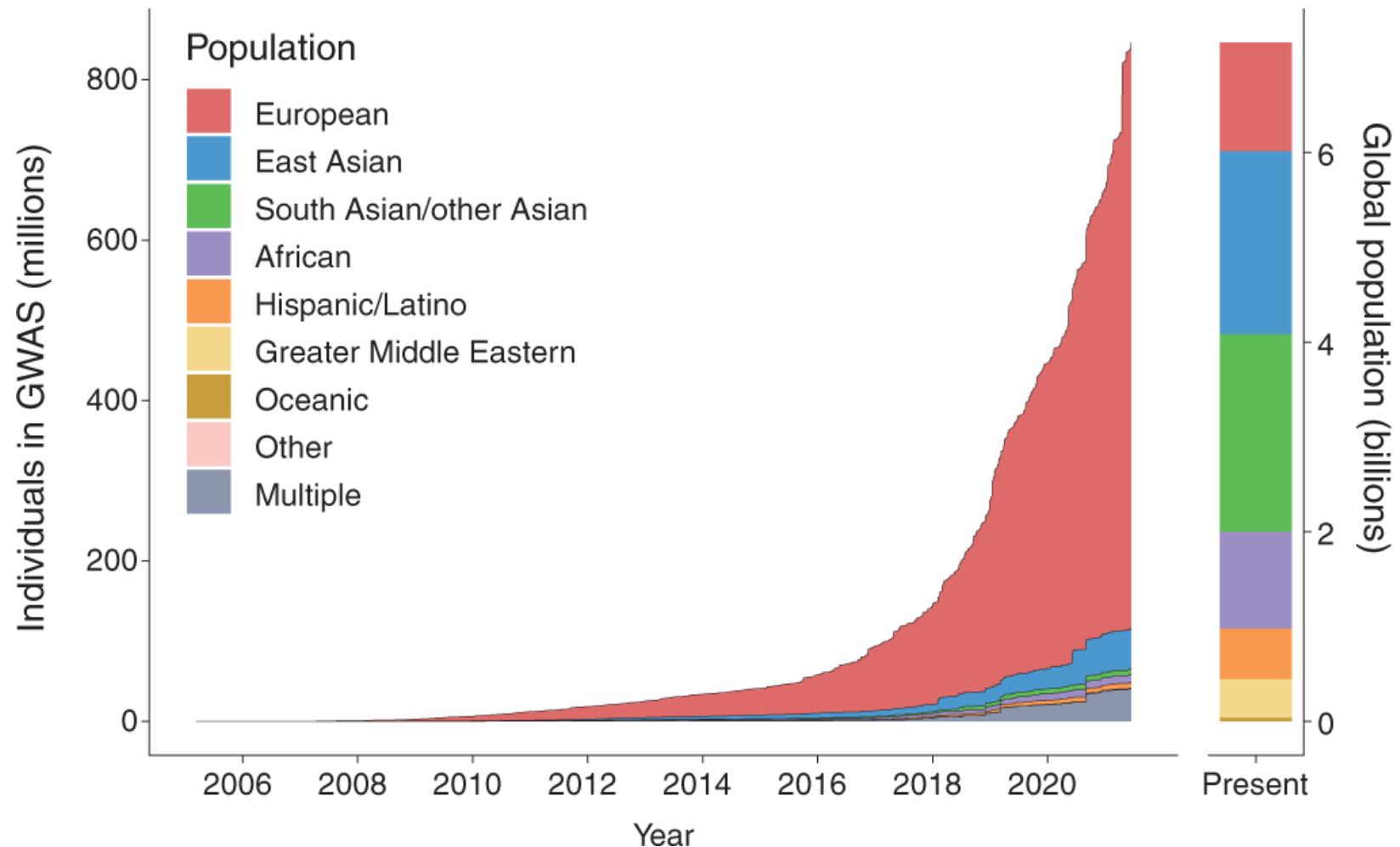


- Legend for Ancestry:
- EUR (White)
  - SAS (Green)
  - EAS (Purple)
  - AMR (Orange)
- Legend for genomic features:
- Centromeric satellites (Yellow)
  - Segmental duplications (Blue)
  - CHM13 exclusive gene density (Red)
  - GRCh38 gene density (Grey)
  - Fixed GRCh38 gaps and issues (Black)

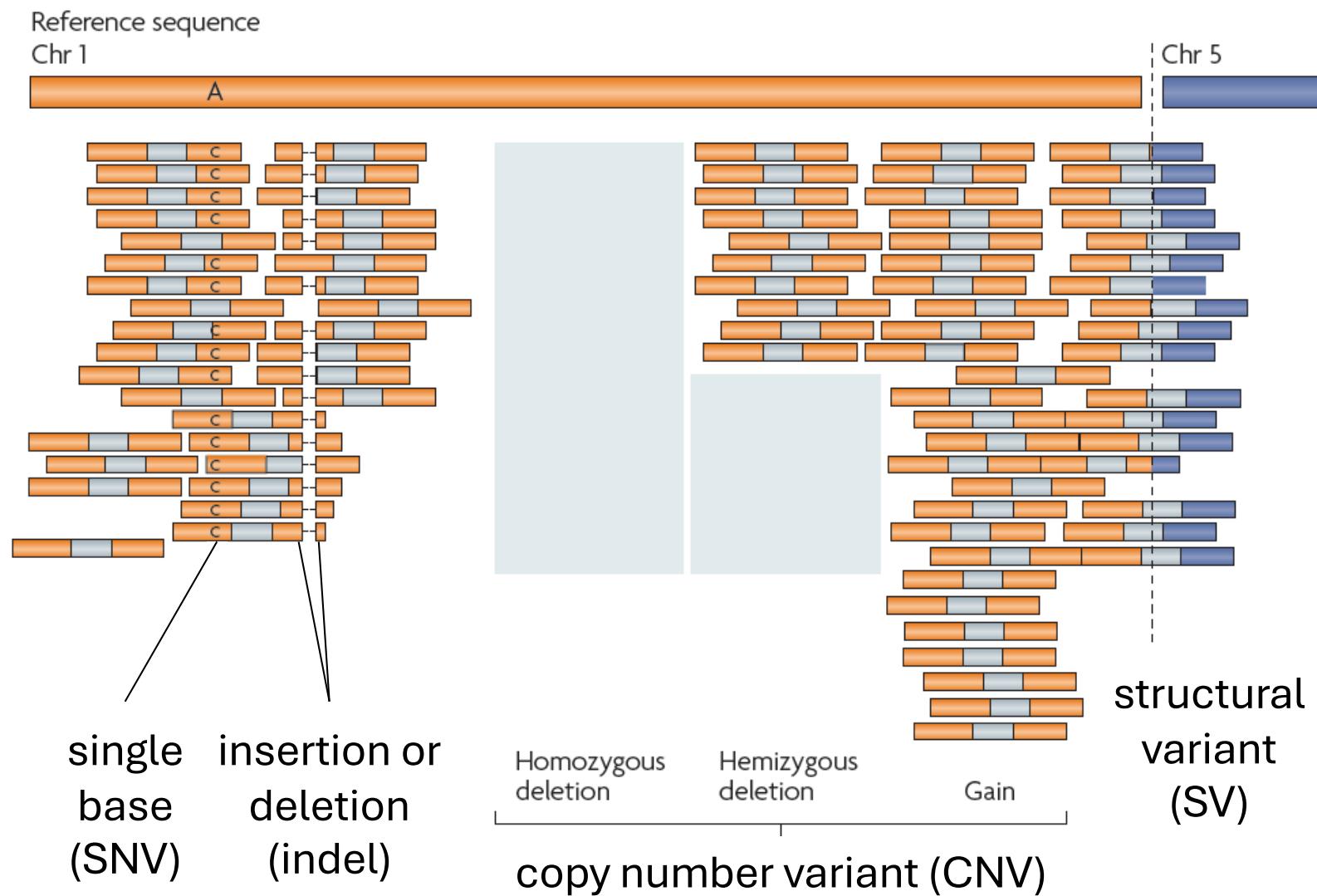


# Reference genomes: “Pan-genome”

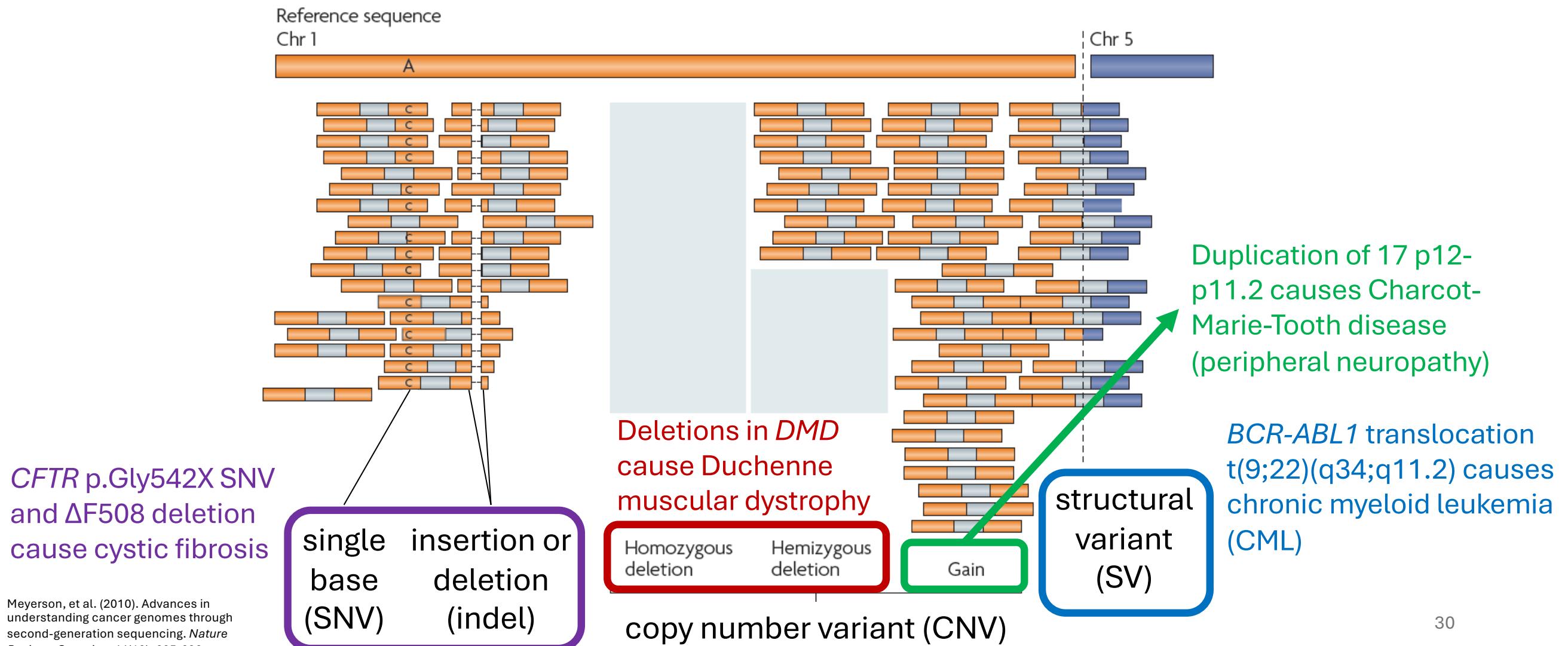
- Increase reference ancestral and allelic diversity
- Human Pan genome Project
- Human Pan genome Reference Project
- Global Alliance for Genomics & Health (GA4GH)



# What types of variants can DNA sequencing detect?



# What types of variants can DNA sequencing detect?

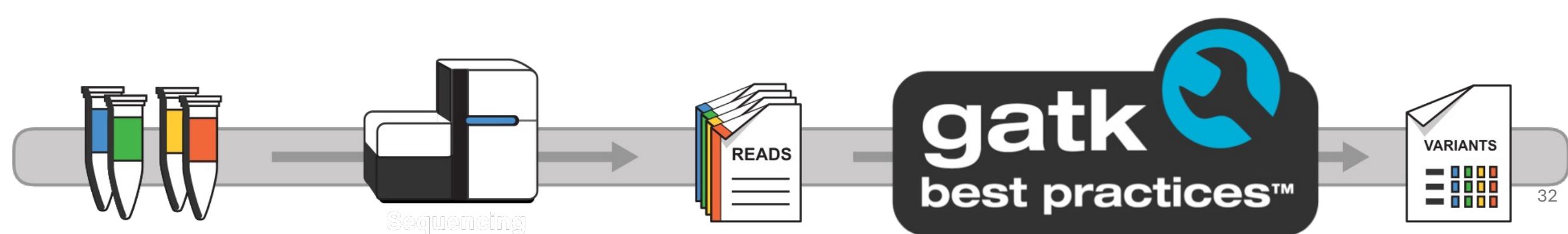


# Outline

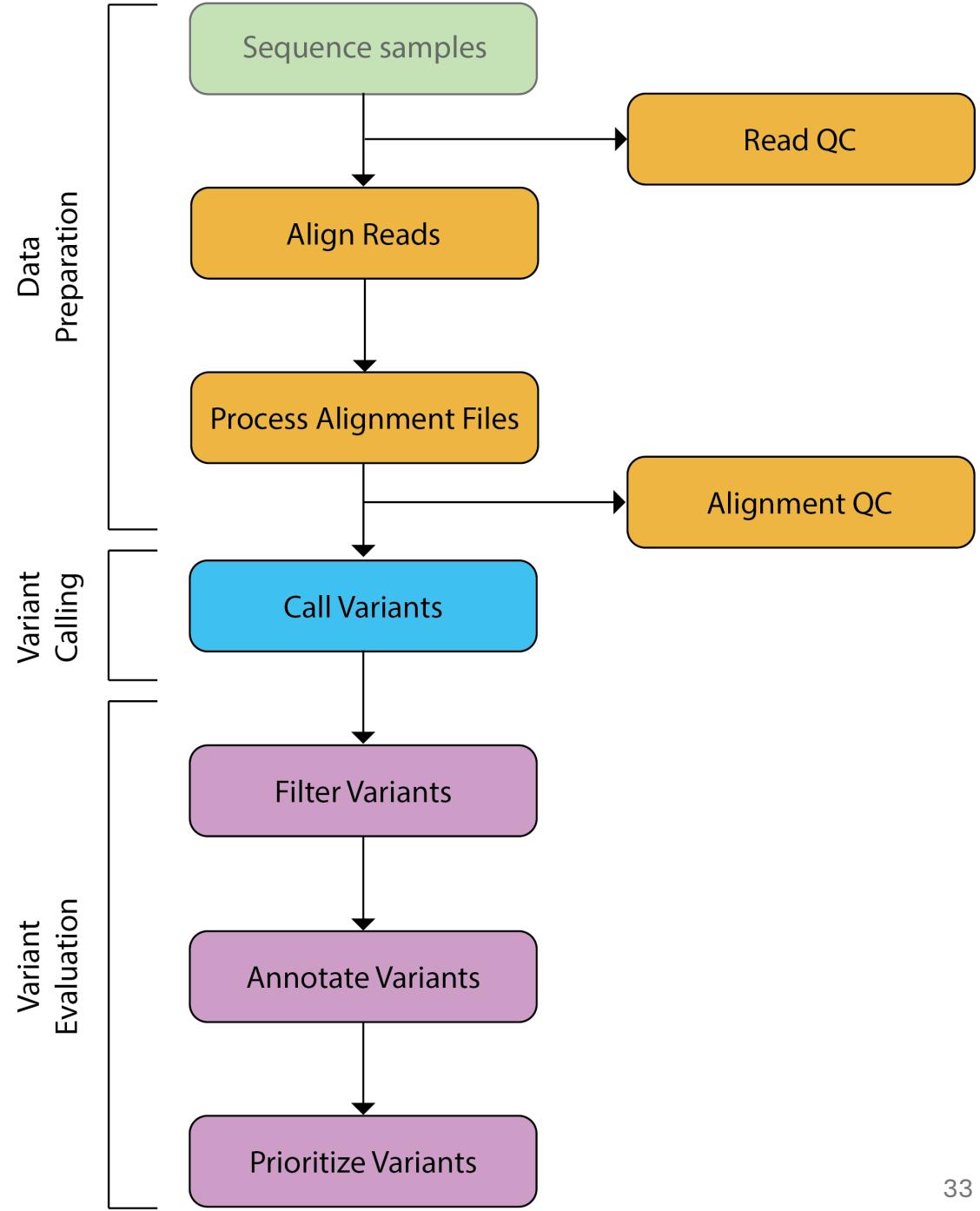
- Background
- DNA sequencing methods
- Reference genomes
- **Variant analysis workflow overview**
- Cancer-specific considerations
- Variant annotations and prioritization
- Example research application
- Implications and societal considerations

# Genome Analysis Tool Kit (GATK) best practices

- End-to-end workflows for multiple types of variant calling
- Designed for human genetic analysis but adaptable to other organisms
- Includes recommendations for experimental design, quality control, tools, implementation options
- Benchmarked against individual references e.g., NA12891, GIAB
- Evolving to reflect state of the field; currently v4



# Variant calling and analysis workflow



# Data preparation and QC

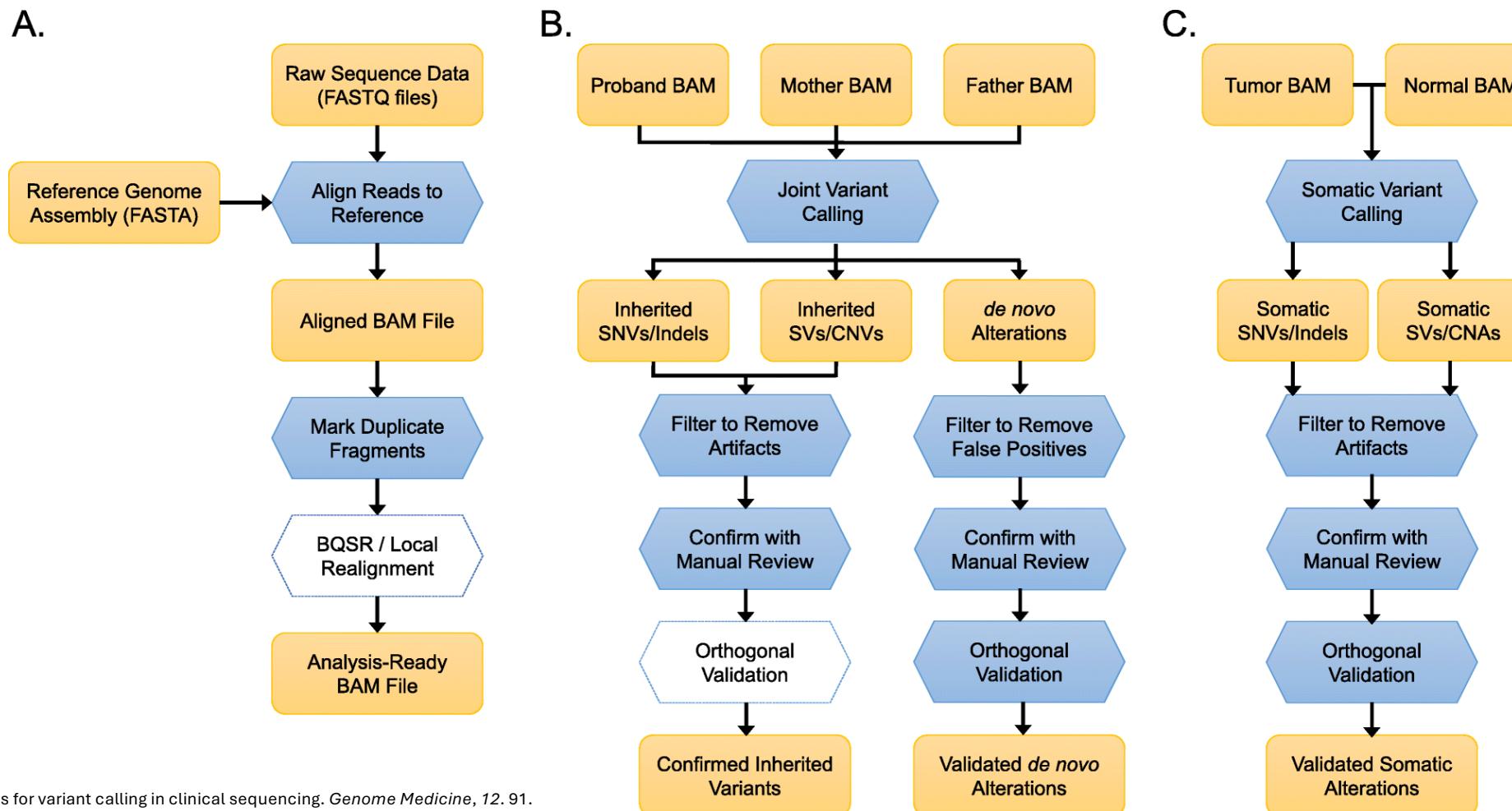
- **Goal: Reduce false positive variants due to technical artifacts**
- Remove low quality reads
- Remove low quality alignments
- Remove PCR duplicate reads
- Local realignment around indels
- Base quality score recalibration (BQSR)
  - \* May no longer be required due to technological improvements

# FFPE considerations

- Samples that are formalin-fixed and paraffin-embedded (FFPE) have DNA damage due to crosslinking
- Cancer specimens are also more likely to contain apoptotic and/or necrotic cells, again with DNA damage
- More stringent filtering is required
- Higher coverage may be required

# Variant calling

- Goal: Identify variants in sample(s) that are not seen in reference



# Variant calling: Individual

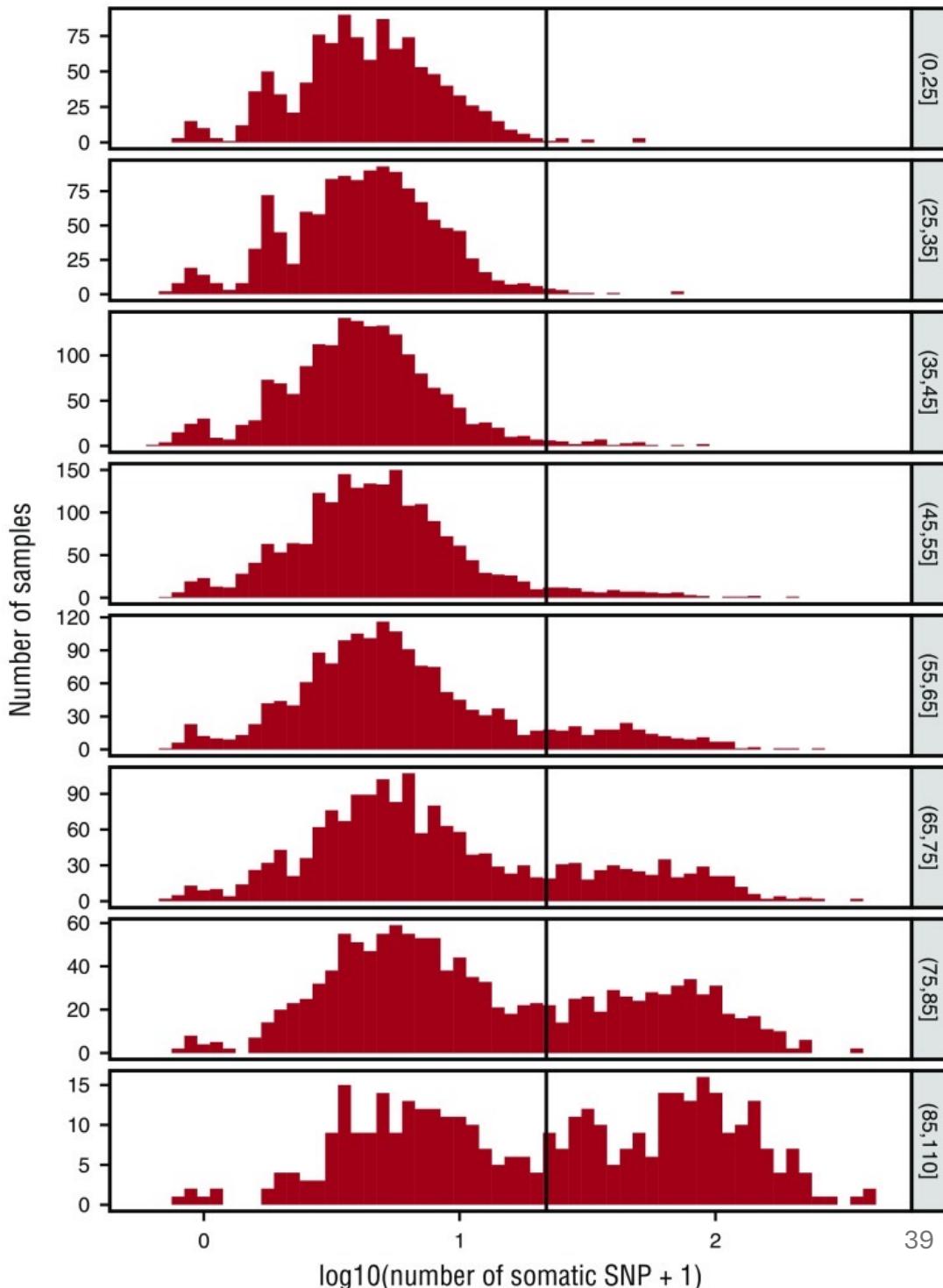
- **Single sample vs reference**
- Per person for **WGS** you expect:
  - ~ 3-5,000,000 SNVs (3-5mil nucleotides),
  - ~ 600,000 indels ( 2mil nucleotides),
  - ~ 160 CNVs
  - ~ 2,500 structural variants (>20mil nucleotides)
- Per person for **WES** you expect:
  - ~ 21,000 SNVs
  - ~ 1,000 indels

# Variant calling: Joint

- **Multiple samples vs reference**
- Contains genotypes for all variant positions, not just the variants detected in any one individual
- Family sequencing: Determine whether variants in child are in *cis*, in *trans*, or *de novo* based on parental sequences
  - Expect 70 *de novo* mutations per genome, or 1 per exome
- Can use information from one sample to infer the genotype in another

# Variant calling: Somatic

- **Affected (e.g., tumor) vs unaffected tissue within a single individual**
- Expect fewer variants than in germline, otherwise suspect DNA damage
- Higher frequency in sun-exposed tissue, older individuals

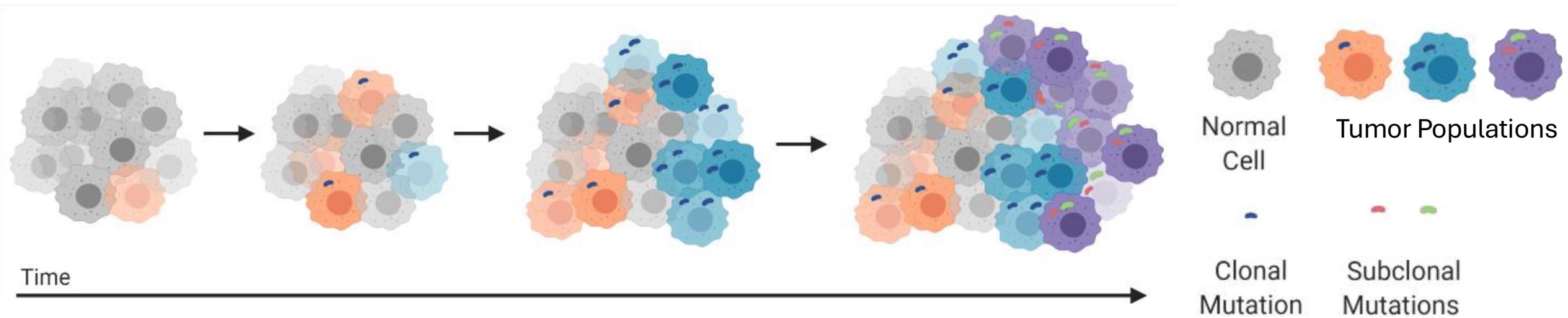


# Outline

- Background
- DNA sequencing methods
- Reference genomes
- Variant analysis workflow overview
- **Cancer-specific considerations**
- Variant annotations and prioritization
- Example research application
- Implications and societal considerations

# Cancer / somatic variant calling presents unique challenges

- Cancer specimens are usually a mix of tumor and non-tumor cells (low sample purity)
- Cancer heterogeneity: A single cancer is actually a diverse population of clones that may have different genomes
- May require higher sequencing depth to capture all variation



# Allele frequency: within a sample

- Variant allele frequency (VAF) =  
$$\frac{\text{[# reads supporting alternate allele]}}{\text{[# reads covering that genomic location]}}$$
- Germline: 0, 50, or 100% alternate allele
- Cancer: Influenced by proportion of tumor cells in sample (tumor purity), CNVs, cancer genetic heterogeneity

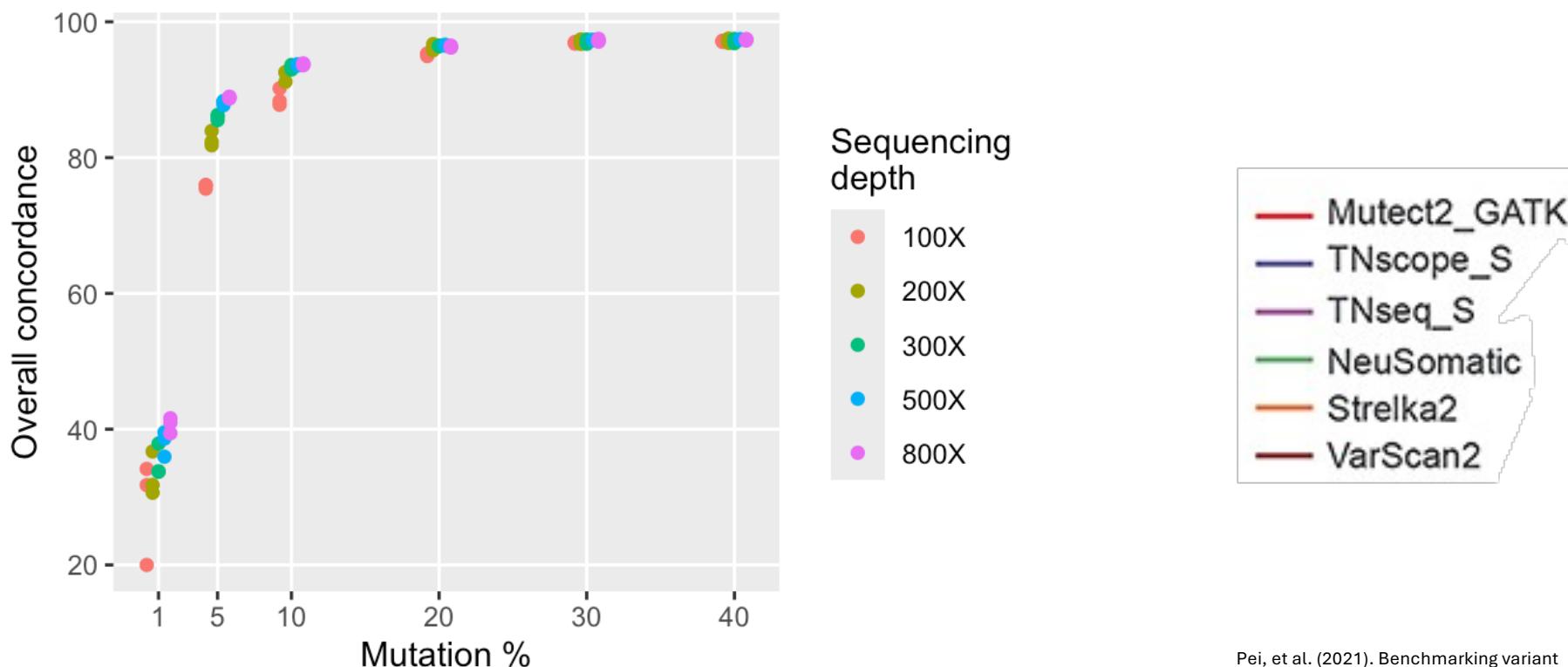
# Allele frequency: across a population

- **Allele frequency for a given population =**  
$$\frac{\text{[# chromosomes with allele]}}{\text{[size of population]}}$$
- Rare variant: < 1–5% minor allele frequency (MAF)

# Cancer / somatic variant calling presents unique challenges

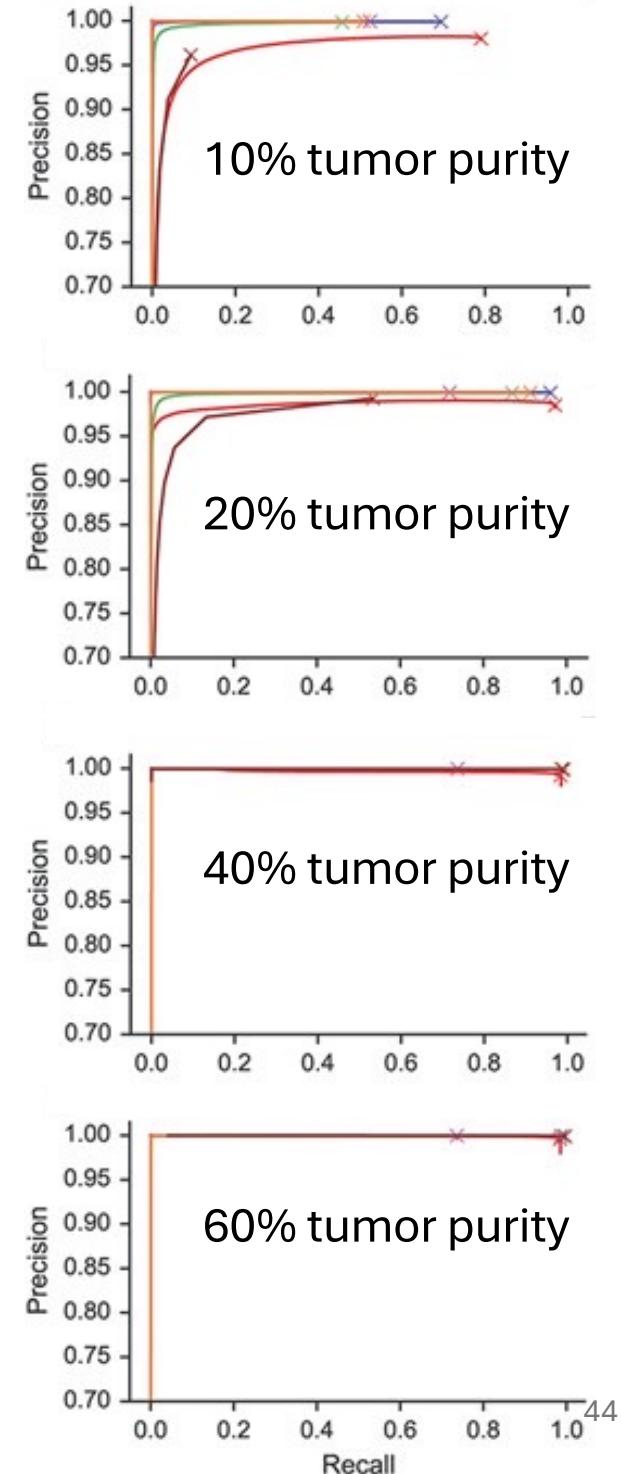
- Somatic variant callers frequently disagree

Comparison of MuTect2 (GATK) and Strelka somatic variant callers



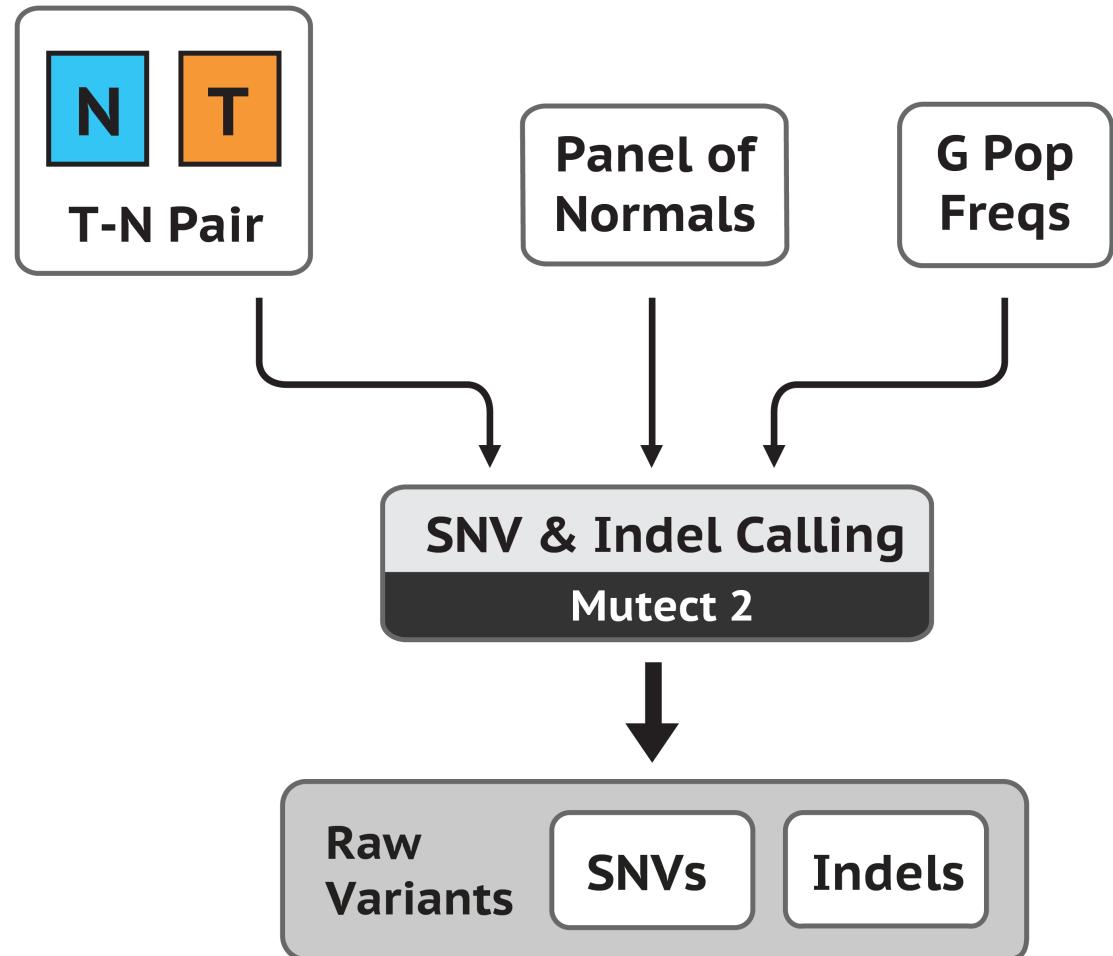
Adapted from Chen, et al. (2020). Systematic comparison of somatic variant calling performance among different sequencing depth and mutation frequency. *Scientific Reports*, 10(1), 3501.

Pei, et al. (2021). Benchmarking variant callers in next-generation and third-generation sequencing analysis. *Briefings in Bioinformatics*, 22(3), bbaa148.



# MuTect2

- Only calls short variants (SNVs, indels)
- Tumor-normal and tumor only modes
- Panel of Normals (PoN): remove sequencing noise



# Outline

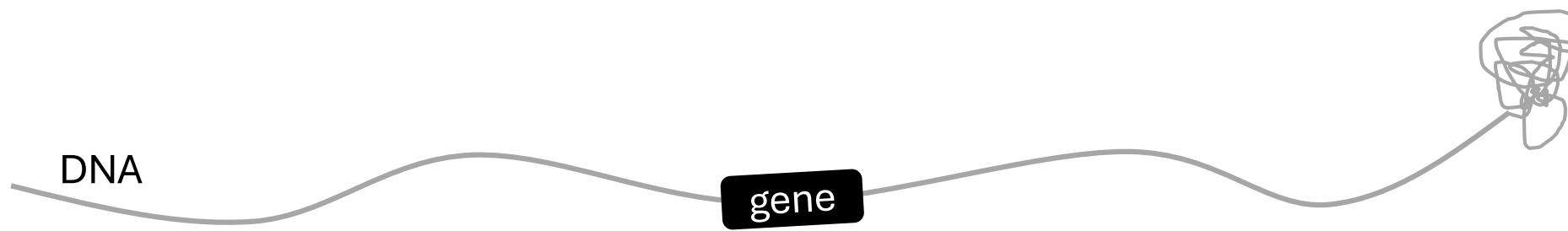
- Background
- DNA sequencing methods
- Reference genomes
- Variant analysis workflow overview
- Cancer-specific considerations
- **Variant annotations and prioritization**
- Example research application
- Implications and societal considerations

# Variant annotation

- **Goal: Identify (and prioritize) variants likely to have an effect**
- Most human genomic variants do not have any discernible phenotypic impact
- If they do, it can be...
  - **Positive:** confers a reproductive advantage
  - **Neutral:** no effect on fitness, but may affect traits such as height or hair color or be associated with ethnic origin
  - **Deleterious / damaging:** has a negative effect on protein structure, expression, and/or function
  - **Pathogenic:** causes disease

# Variant impacts

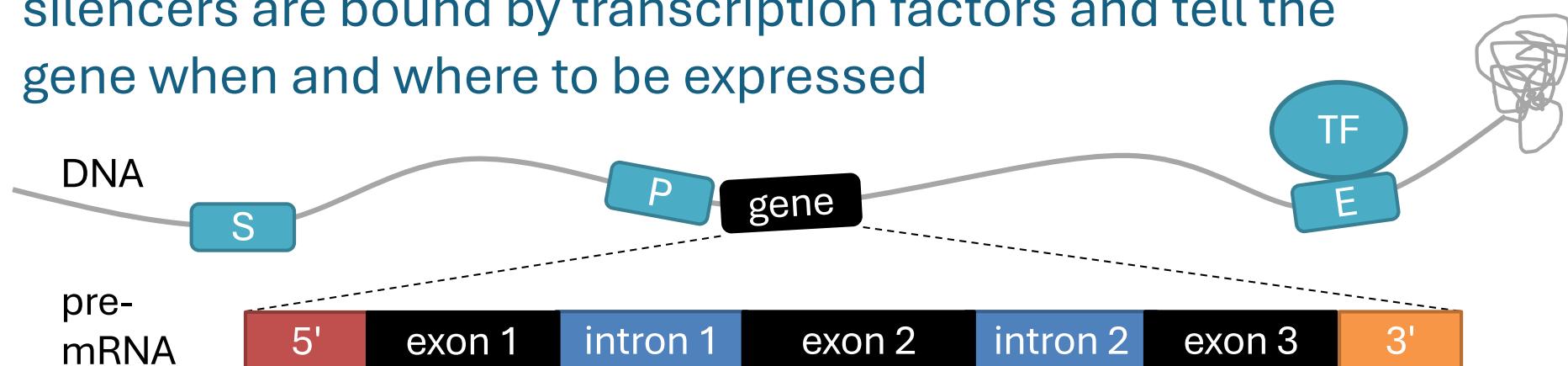
- Genes are only 1-2% of the genome
- Majority of called variants from WGS will be in noncoding regions



# Variant impacts

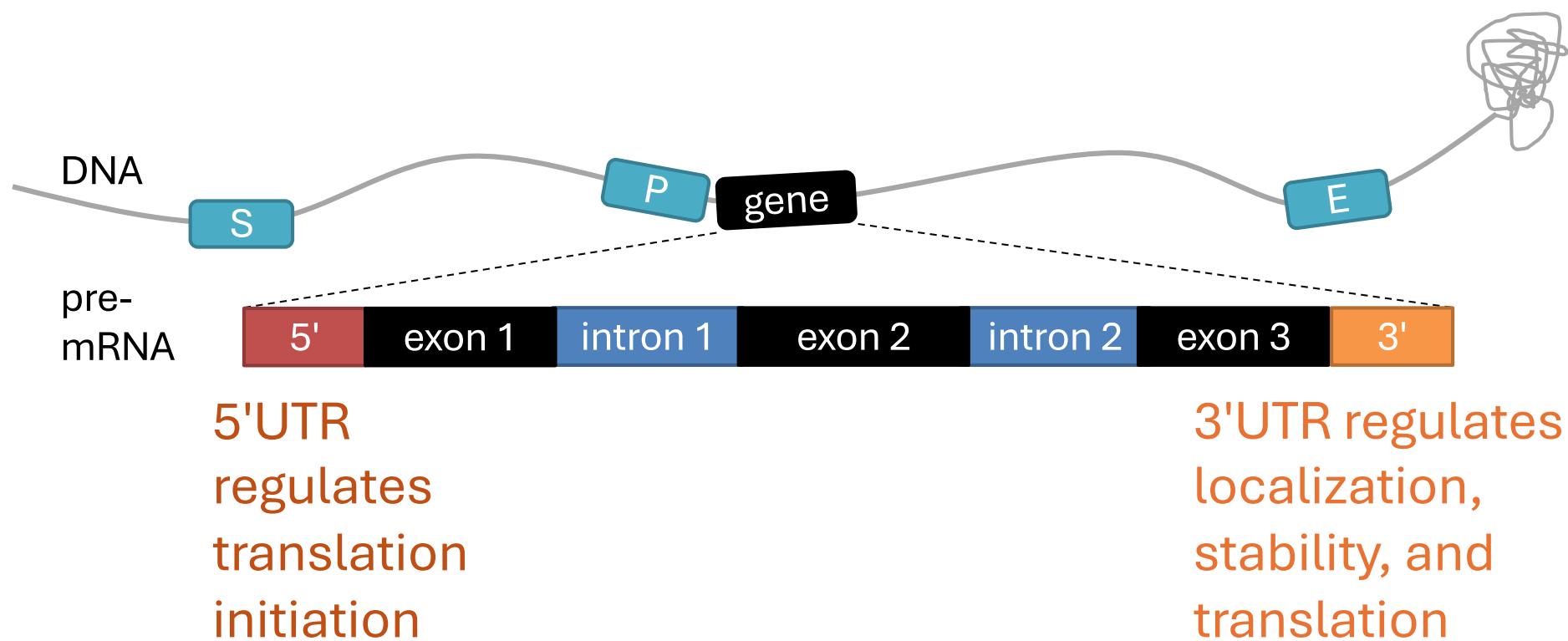
- What do noncoding regions do?

regulatory sequences like promoters, enhancers, and silencers are bound by transcription factors and tell the gene when and where to be expressed



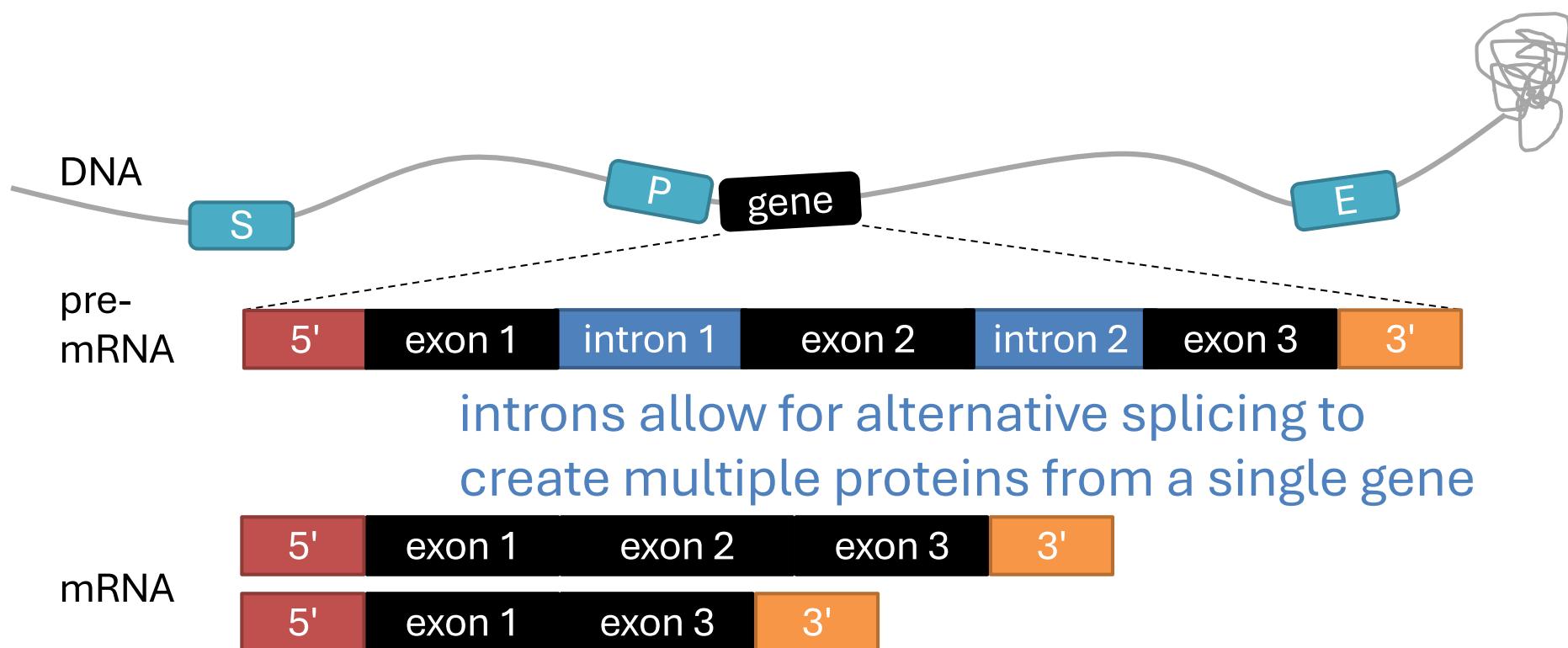
# Variant impacts

- What do noncoding regions do?



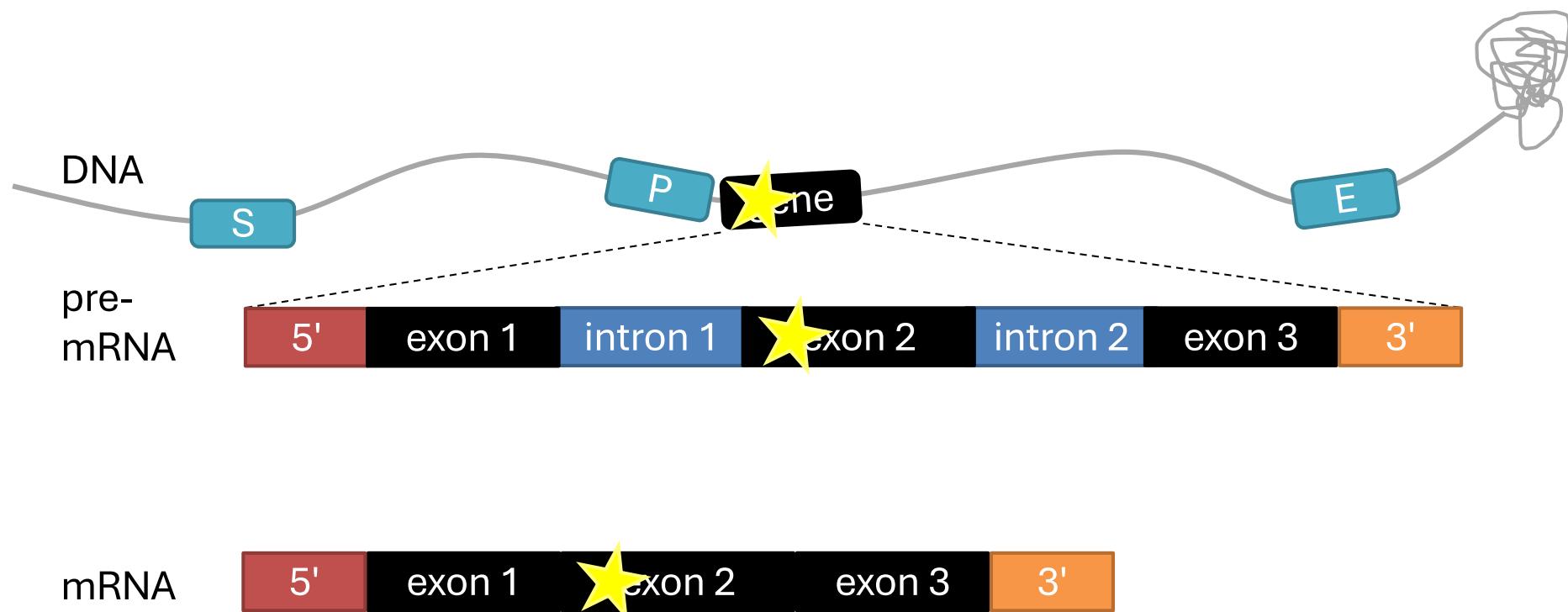
# Variant impacts

- What do noncoding regions do?



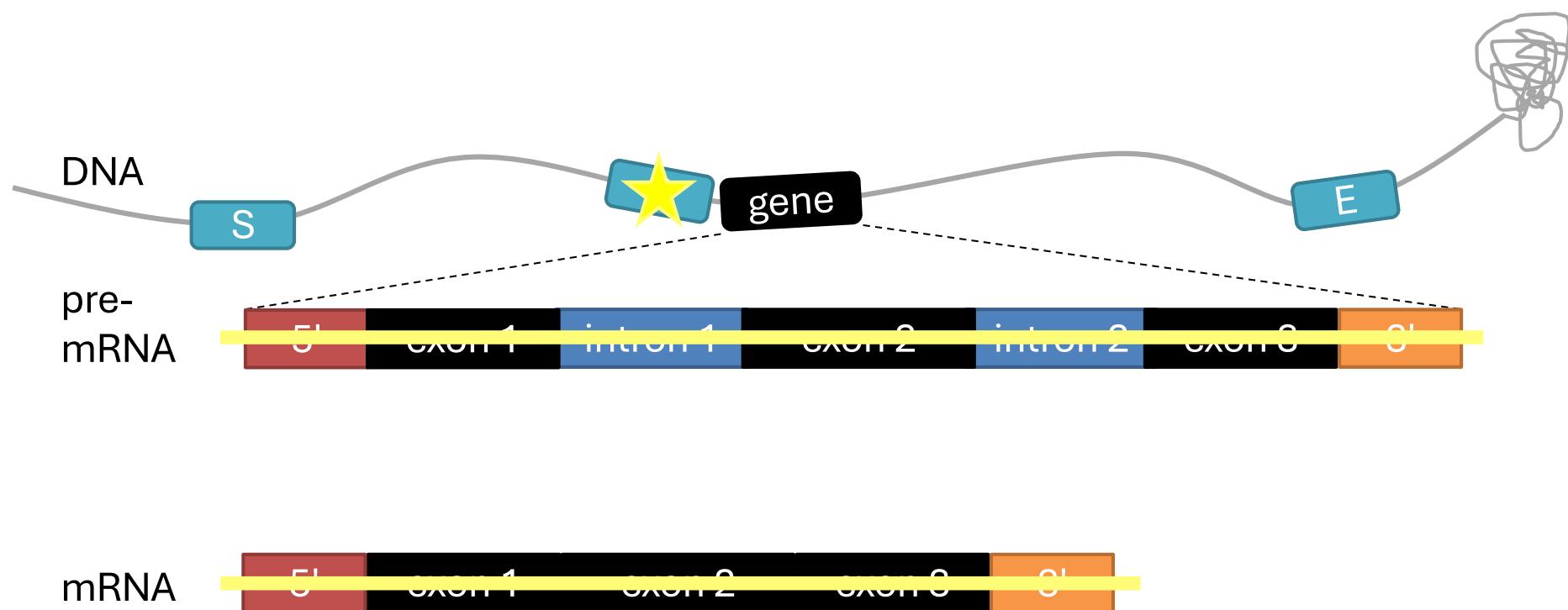
# How do you predict variant function?

- Disrupt coding sequences → change protein sequence



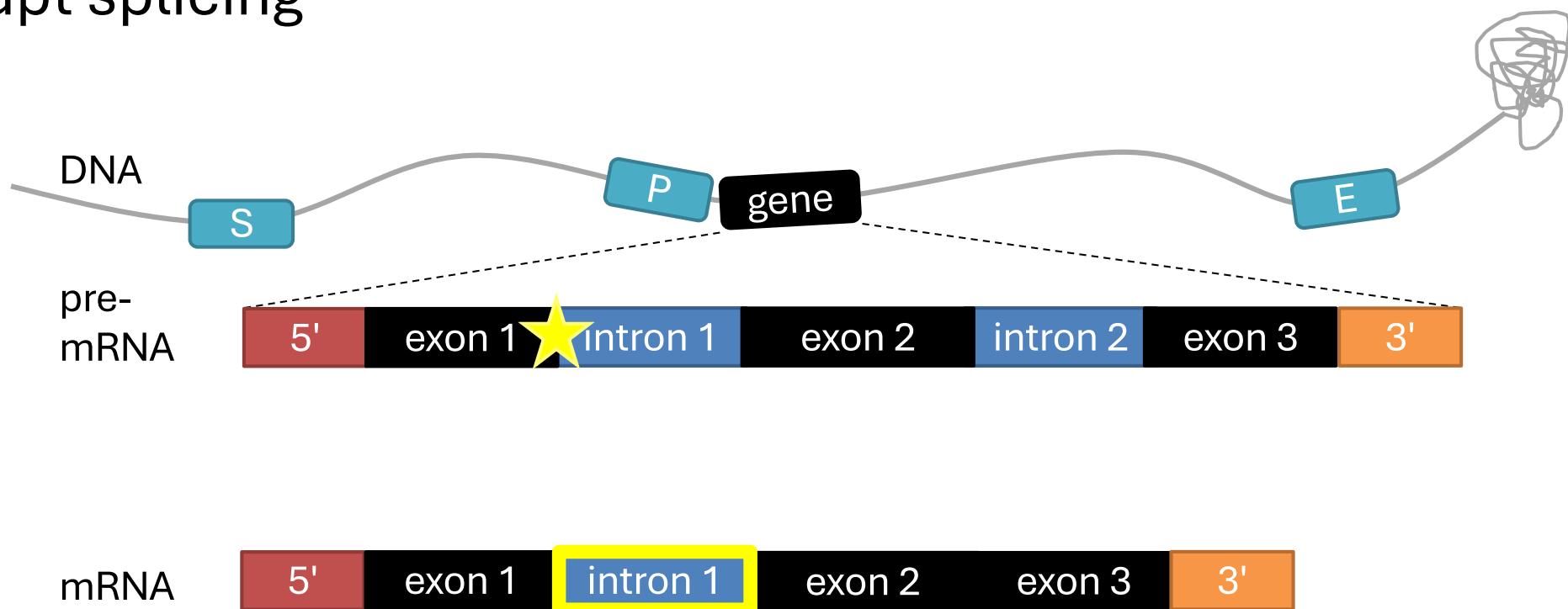
# How do you predict variant function?

- Disrupt coding sequences
- Disrupt regulatory regions



# How do you predict variant function?

- Disrupt coding sequences
- Disrupt regulatory regions
- Disrupt splicing



# How do you predict variant function?

- Disrupt coding sequences
- Disrupt regulatory regions
- Disrupt splicing
- Highly evolutionarily conserved

# Common annotations and databases

- **Variant context:** dbSNP
- **Variant population frequency:** gnomAD, ExAC
- **Variant association with human disease:** ClinVar, Human Gene Mutation Database (HGMD)
- **Gene association with human disease:** Online Mendelian Inheritance in Man (OMIM), genome-wide association study (GWAS)
- **Gene association with mouse phenotype:** Mouse Genome Informatics (MGI / Jax), Knockout Mouse Project (KOMP)
- **Cancer:** Catalogue Of Somatic Mutations In Cancer (COSMIC), cBioPortal, The Cancer Genome Atlas (TCGA)

# Functional consequence prediction

- **Effect on protein (exome)**
  - Pathogenicity prediction scores: MetaLR, MutationAssessor, MutationTaster, PolyPhen, PROVEAN, REVEL, SIFT, etc.
- **Disrupt regulatory regions**
  - Functional DNA prediction scores: CADD, DANN, EIGEN, FATHMM, GWAFA, etc.
  - Transcription factor binding sites (TFBS): ENCODE ChIPseq
  - Open chromatin: DNase hypersensitivity sites (DHS), Chromatin State Segmentation in relevant cell type
- **Disrupt splicing**
  - Splicing prediction scores: TraP, dbscSNV, regSNPintron, Human Splicing Finder (HSF)
- **Highly evolutionarily conserved**
  - Conservation scores: GERP, PhyloP, Residual Variation Intolerance Score (RVIS) 57

# Tools to add annotations

- Instead of looking up annotations one at a time, you can use...
- ANNOVAR
- **SnpEff / SnpSift**
- Variant Annotation, Analysis, and Search Tool (VAAST)
- Variant Effect Predictor (VEP)
- VarSome

# Variant prioritization: Germline

## QC filtering

- Alternative allele # > 5
- VAF > 20%
- Not in repeats

## Keep protein-altering variants in known genes

- Nonsense SNVs
- Missense SNVs
- Splice site SNVs
- Frameshift INDELs
- In-frame INDELs

## Remove common variants

- 1000 Genomes Project
- gnomAD
- ...

## Rank by annotation

- MAF (< 0.001 or not)
- Polyphen2
- CADD score
- REVEL score
- pLI score
- ...

## Find known pathogenic variants

- NCBI ClinVar
- InterVar
- IARC TP53
- BRCA Exchange
- ...

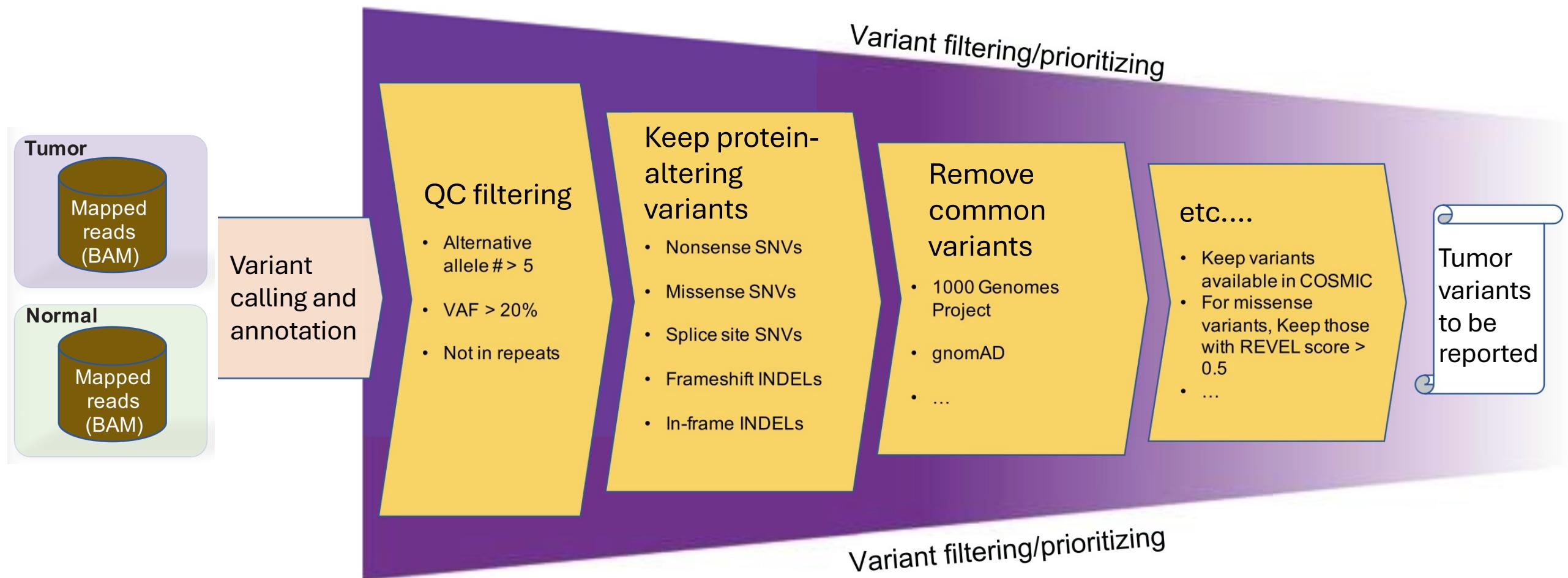
## 2<sup>nd</sup> hit for recessive disease

- Copy loss
- Promoter methylation
- ...

Germline variants to be reported

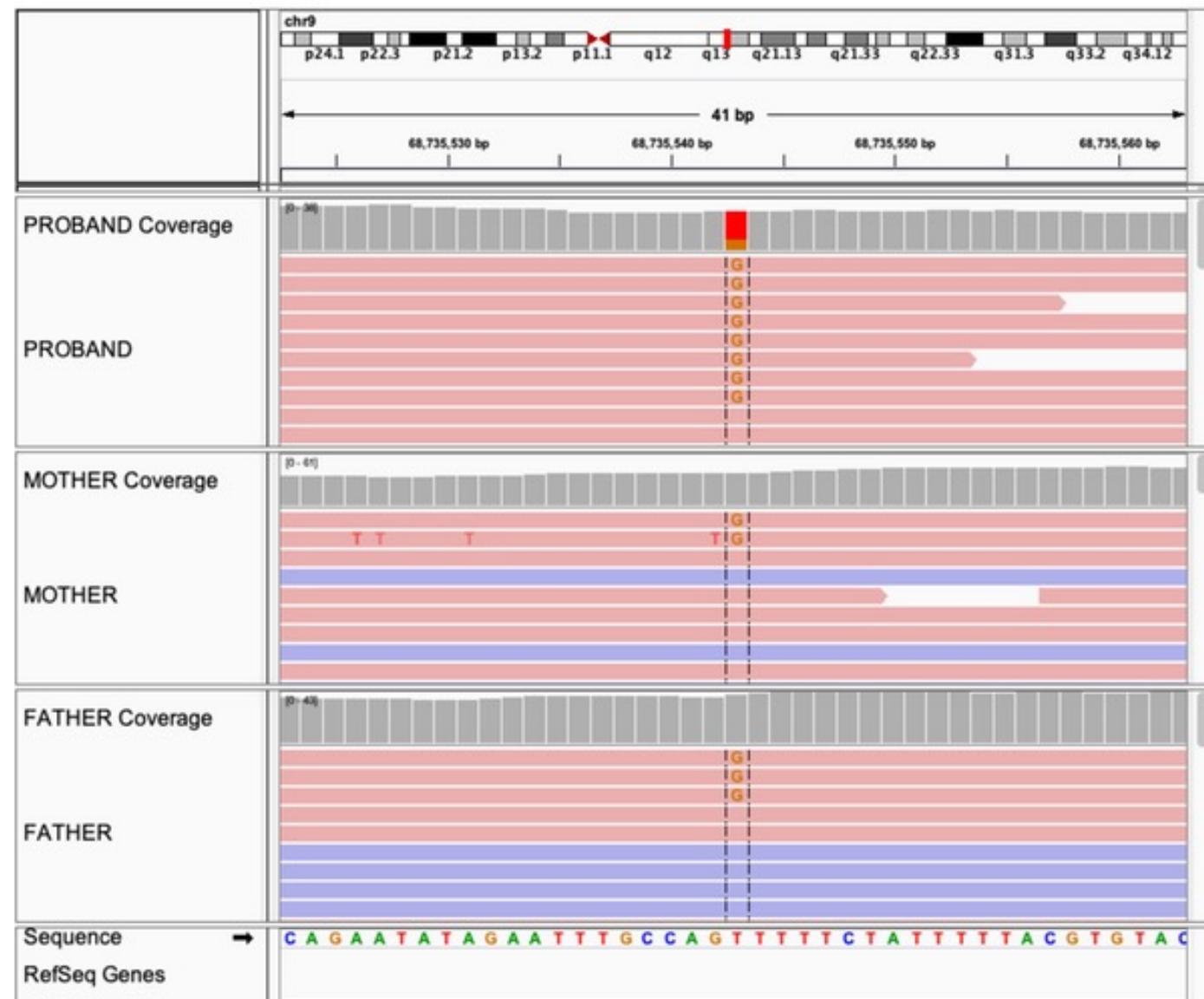
Variant filtering/prioritizing

# Variant prioritization: Tumor



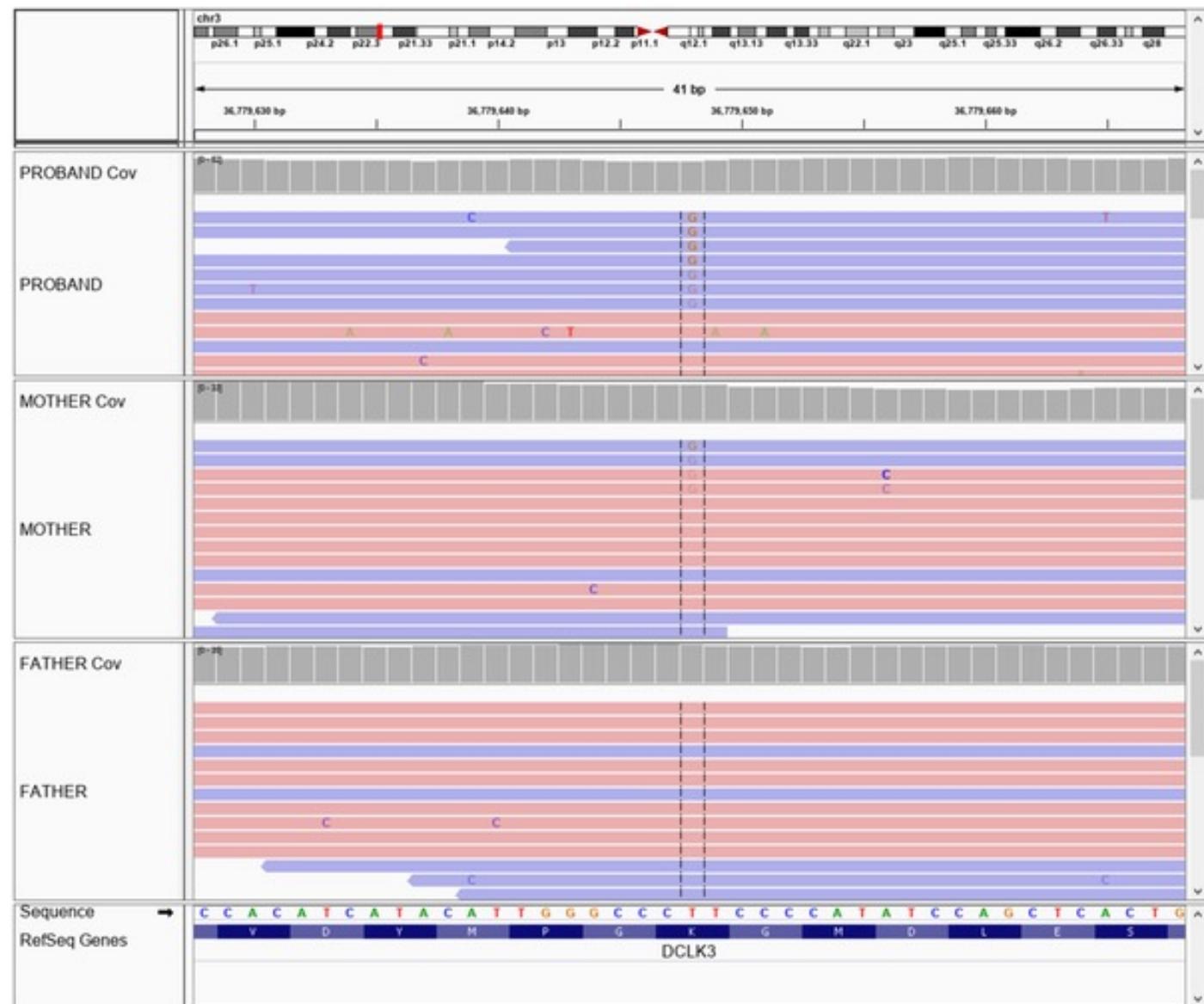
# Validation: Go look at the reads yourself in IGV

- **Strand bias**
  - Variants seen only on reverse strand (red)



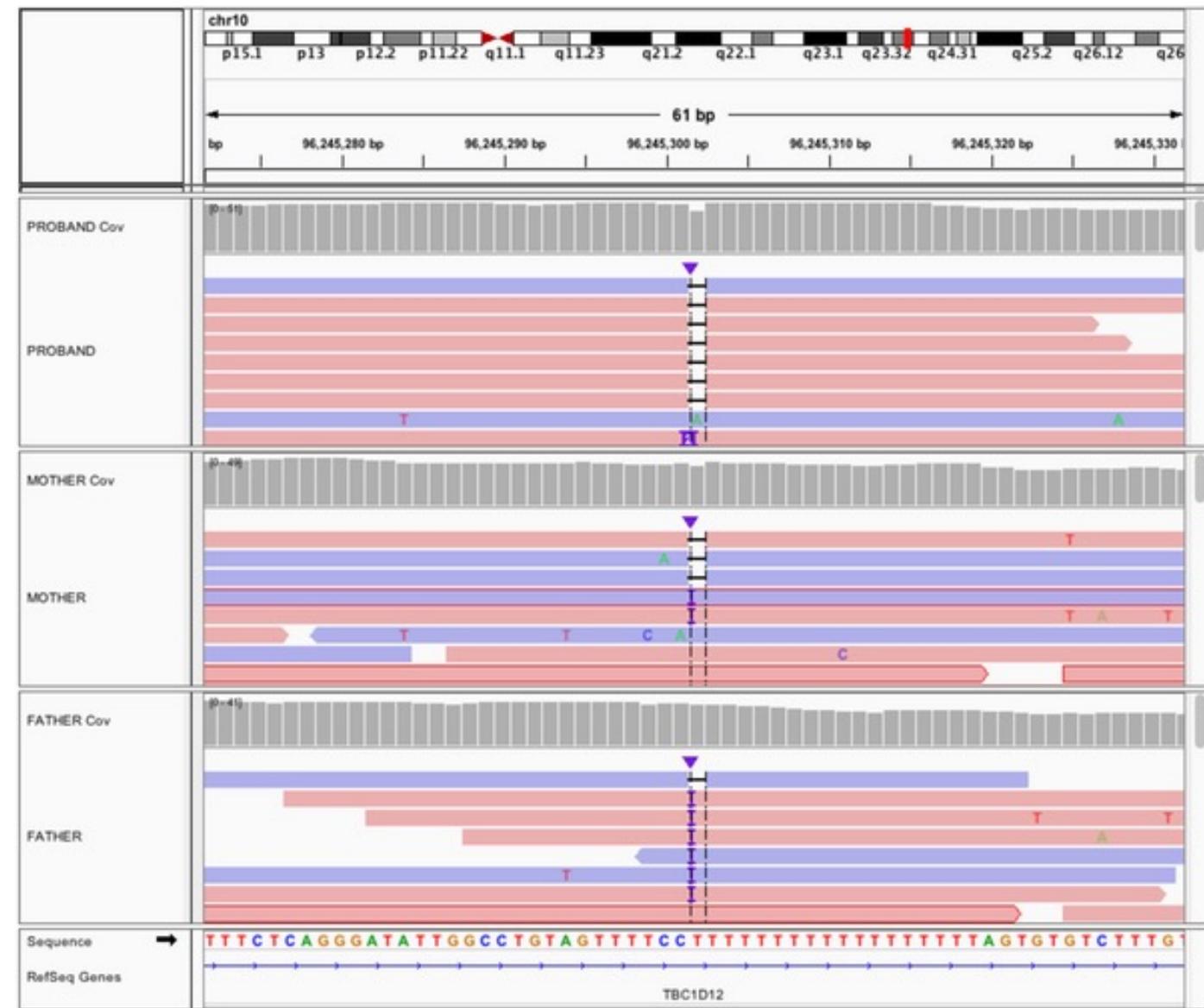
# Validation: Go look at the reads yourself in IGV

- Low quality
  - Variants are pale rather than dark



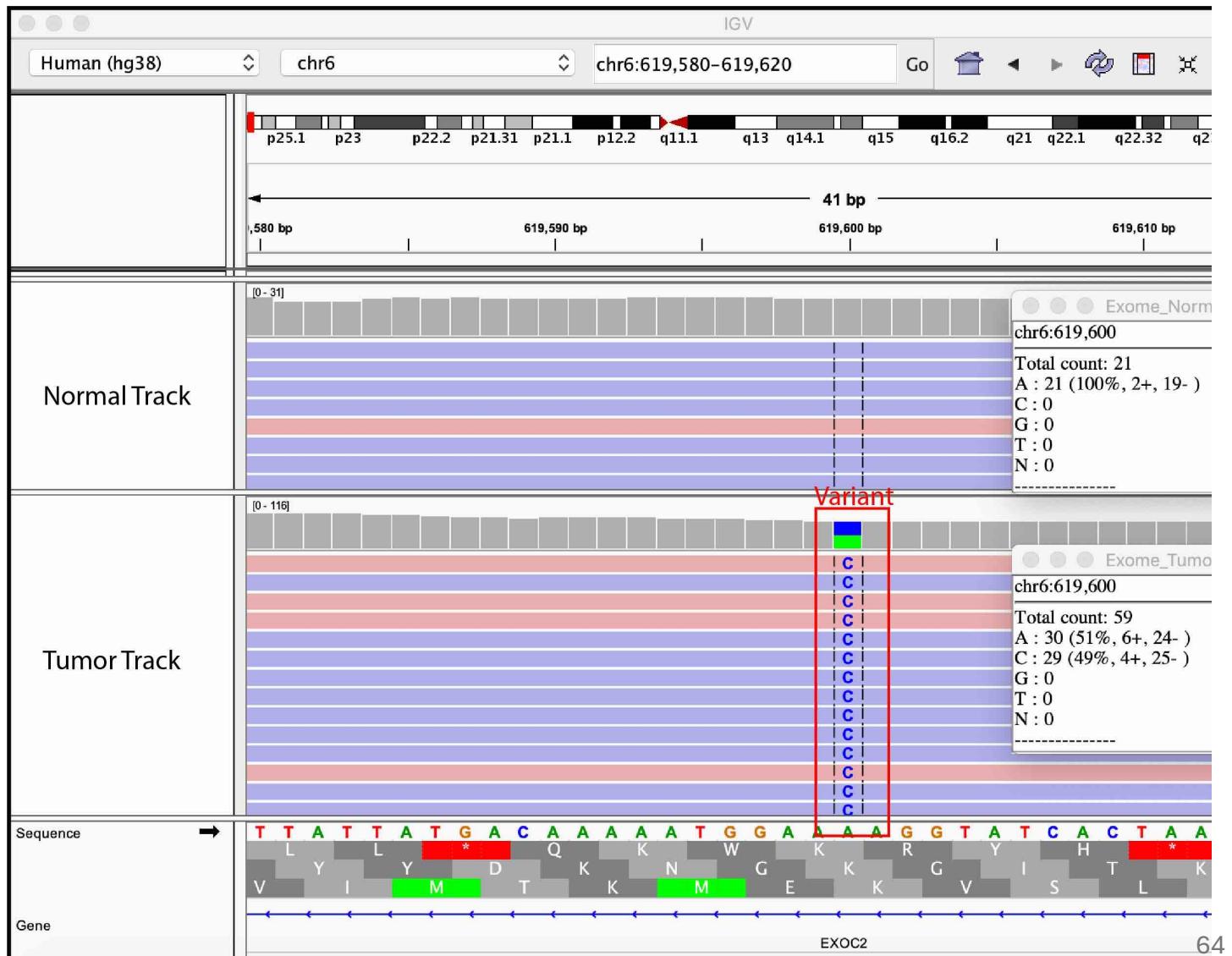
# Validation: Go look at the reads yourself in IGV

- Low complexity
  - Variants occur near homopolymer (TTTTTTTTTTTT)



# Validation: Go look at the reads yourself in IGV

- **Real variant!**
  - Adequate coverage
  - Seen in reads on both strands

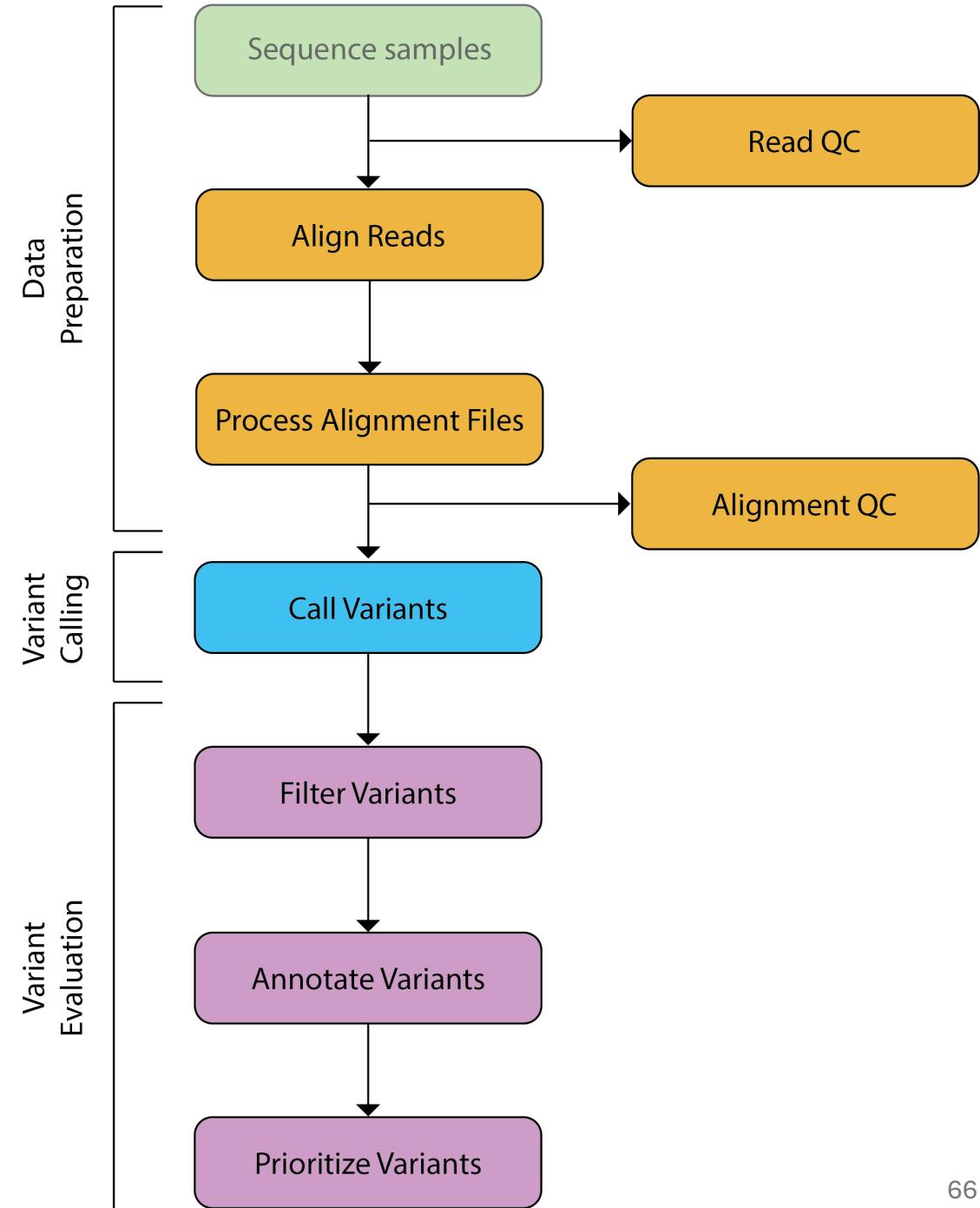


# Re-analysis to increase diagnostic yield

- Even with the same sequencing data, re-analysis after a period of time can yield new variants of interest
  - New reference version
  - New annotations
  - New disease associations
  - New recommendations for analysis parameters
  - etc.

# Variant calling and analysis workflow

- QC to reduce false positives
- Variant calling
- Validation
- Annotation, filtering, and prioritization



# Outline

- Background
- DNA sequencing methods
- Reference genomes
- Variant analysis workflow overview
- Cancer-specific considerations
- Variant annotations and prioritization
- **Example research application**
- Implications and societal considerations

# Example application: Rare disease analysis

## **Analysis of Whole Genome Sequencing in a Cohort of Individuals with PHACE Syndrome Suggests Dysregulation of RAS/PI3K Signaling**

✉ Elizabeth S. Partan, ✉ Francine Blei, ✉ Sarah L. Chamlin, ✉ Olivia M. T. Davies, ✉ Beth A. Drolet, ✉ Ilona J. Frieden, ✉ Ioannis Karakikes, ✉ Chien-Wei Lin, ✉ Anthony J. Mancini, ✉ Denise Metry, ✉ Anthony Oro, ✉ Nicole S. Stefanko, ✉ Laksshman Sundaram, ✉ Monika Tutaj, ✉ Alexander E. Urban, ✉ Kevin C. Wang, ✉ Xiaowei Zhu, ✉ Nara Sobreira, ✉ Dawn H. Siegel

**doi:** <https://doi.org/10.1101/2021.08.05.21261553>



# PHACE syndrome: Segmental hemangioma and 1+ other feature

- Posterior fossa brain malformations
- **Hemangiomas: benign vascular tumors**
- Arterial anomalies
- Cardiac anomalies
- Eye anomalies
- Overall incidence unknown; 300+ cases
- No known familial cases
- Inheritance thought to be *de novo*

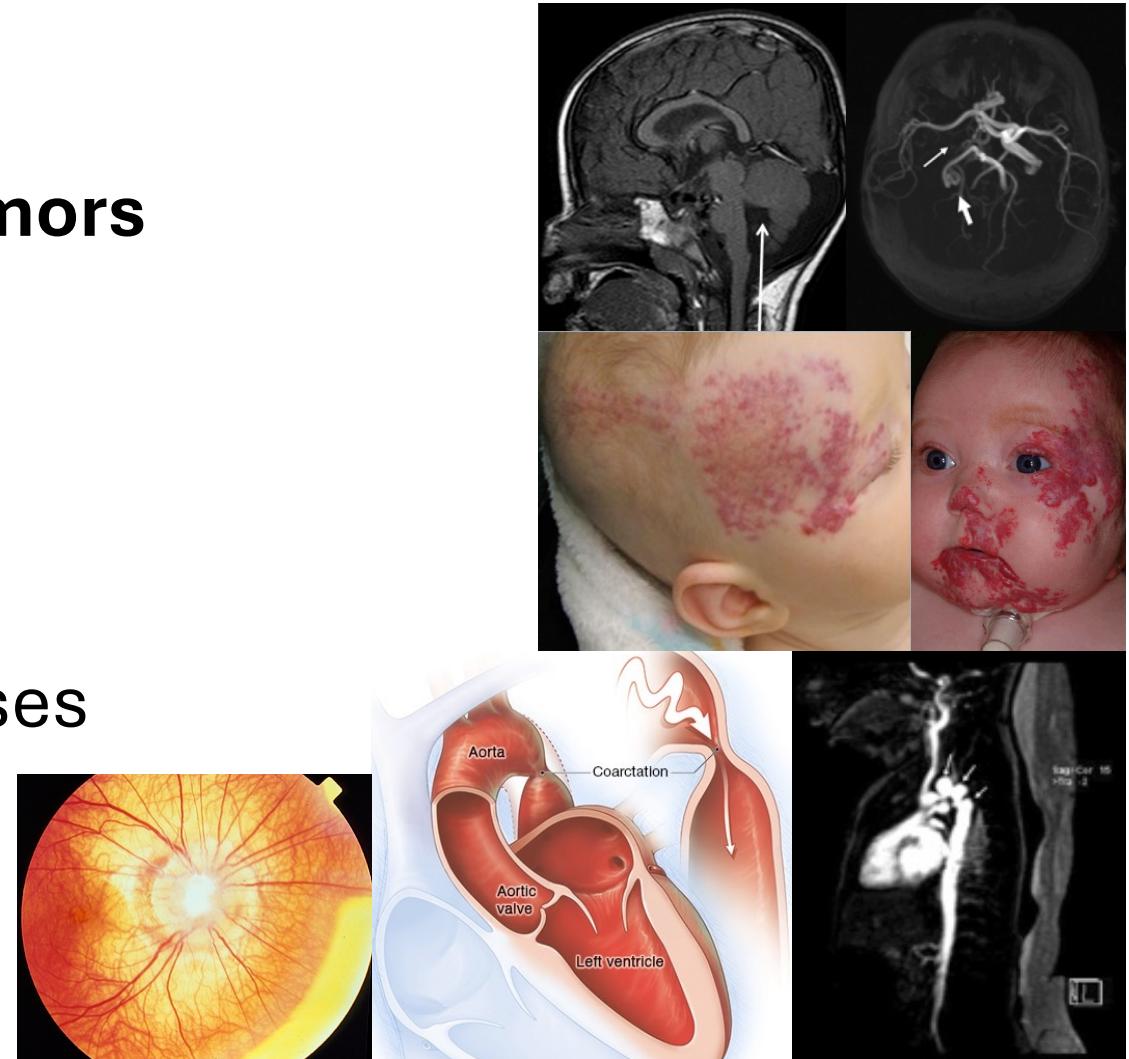


Image credits, left to right, by row: 1) Fernandes 2011, Metry 2009 2) Children's Hospital of Wisconsin, Metry 2006 3) Siegel 2007, Mayo Clinic, Metry 2009

# Variants called and annotated from 98 trios with germline WGS

- Gabriella Miller Kids First Pediatric Research Program
- 150 bp paired-end Illumina sequencing, aligned to hg38 using DRAGEN
- Identified *de novo* SNVs and indels using GATK HaplotypeCaller
- Added variant- and gene-level annotations using ANNOVAR

98 trios → patients *de novo* VCF

16,107 total SNVs (164 SNVs/patient)

112 coding

105 synonymous

15,880 noncoding

119 rare coding

100 rare synonymous

15,533 rare noncoding

# Variant prioritization

- Rare, *de novo* SNVs:
  - Known pathogenic variants — ClinVar
  - Gene with related human phenotype — OMIM, GWAS
  - Gene with related mouse model phenotype — MGI
  - Overlapping with CNVs previously identified in patients with PHACE syndrome and/or hemangioma
  - Genes with coding variants in multiple patients
- Noncoding variants were further investigated for:
  - Effects on functional DNA: prediction scores, transcription factor binding sites, open chromatin, and splicing

# Noncoding variant analysis

15,533 rare, *de novo*, noncoding variants

3,498 predicted  
to affect  
functional DNA

1,144 predicted  
to affect txn  
factor binding

5,452 predicted  
to affect open  
chromatin

4,770 predicted  
to affect  
splicing

3,894 predicted to be damaging by 2+ categories

795 by 3+ categories

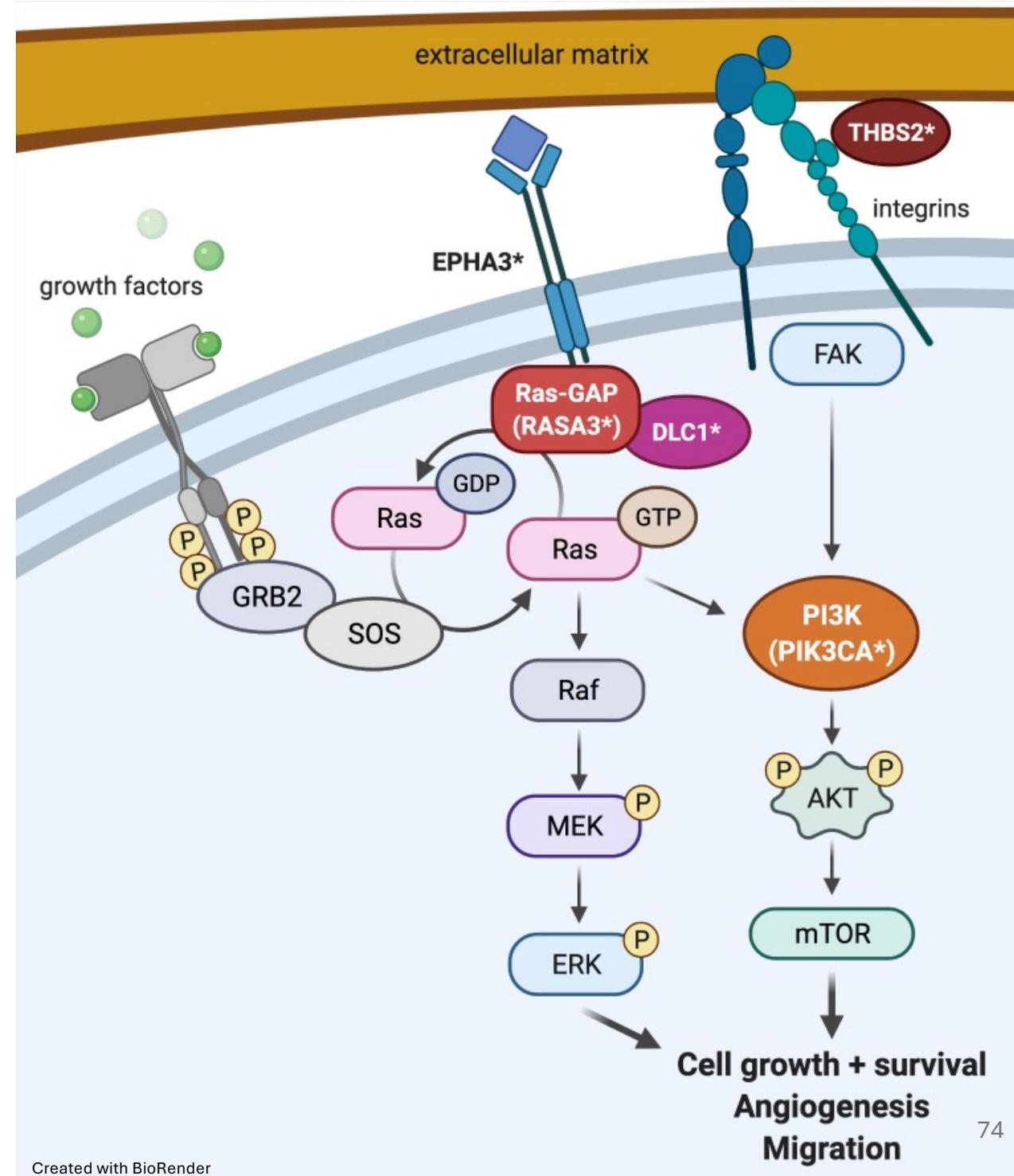
67 by 4

# Candidate causative variant summary

| Gene          | PHACE <i>de novo</i> variant(s)                                 | Variant-level evidence  | Gene-level evidence  |
|---------------|---|---|--|
| <i>THBS2</i>  | p.Asp859Asn;<br>1 intronic SNV and indel<br>in a single patient | <b>Coding:</b> residue required for protein folding<br>Predicted to affect TFBS | Vascular lethal KO mouse;<br>5 deletions (enriched)                    |
| <i>RASA3</i>  | p.Val85Met  | <b>Coding:</b> residue function unknown   | Vascular lethal KO mouse;<br>1 deletion                                |
| <i>PIK3CA</i> | 1 intronic SNV  | In open chromatin (transcribed)   | Human vascular disease;<br>Vascular lethal KO mouse;<br>2 duplications |
| <i>BCAS3</i>  | 2 intronic SNVs;<br>1 intronic indel                            | Predicted to affect TFBS, splicing  | Vascular lethal KO mouse;<br>1 deletion                                |
| <i>DLC1</i>   | 3 intronic SNVs   | In open chromatin (enhancer + transcribed)                                      | Vascular lethal KO mouse;<br>1 duplication                             |
| <i>EPHA3</i>  | 1 intronic SNV  | Predicted to affect TFBS, splicing  | Vascular lethal KO mouse;<br>1 duplication                             |
| <i>EXOC4</i>  | 4 intronic SNVs   | Predicted to affect TFBS, splicing  | Vascular lethal KO mouse;<br>1 deletion                                |

# Pathway analysis

- All genes with rare, *de novo* SNVs (4,320 genes) were analyzed using g:Profiler to understand broader patterns in gene regulation
- Top KEGG pathways all affect angiogenesis signaling



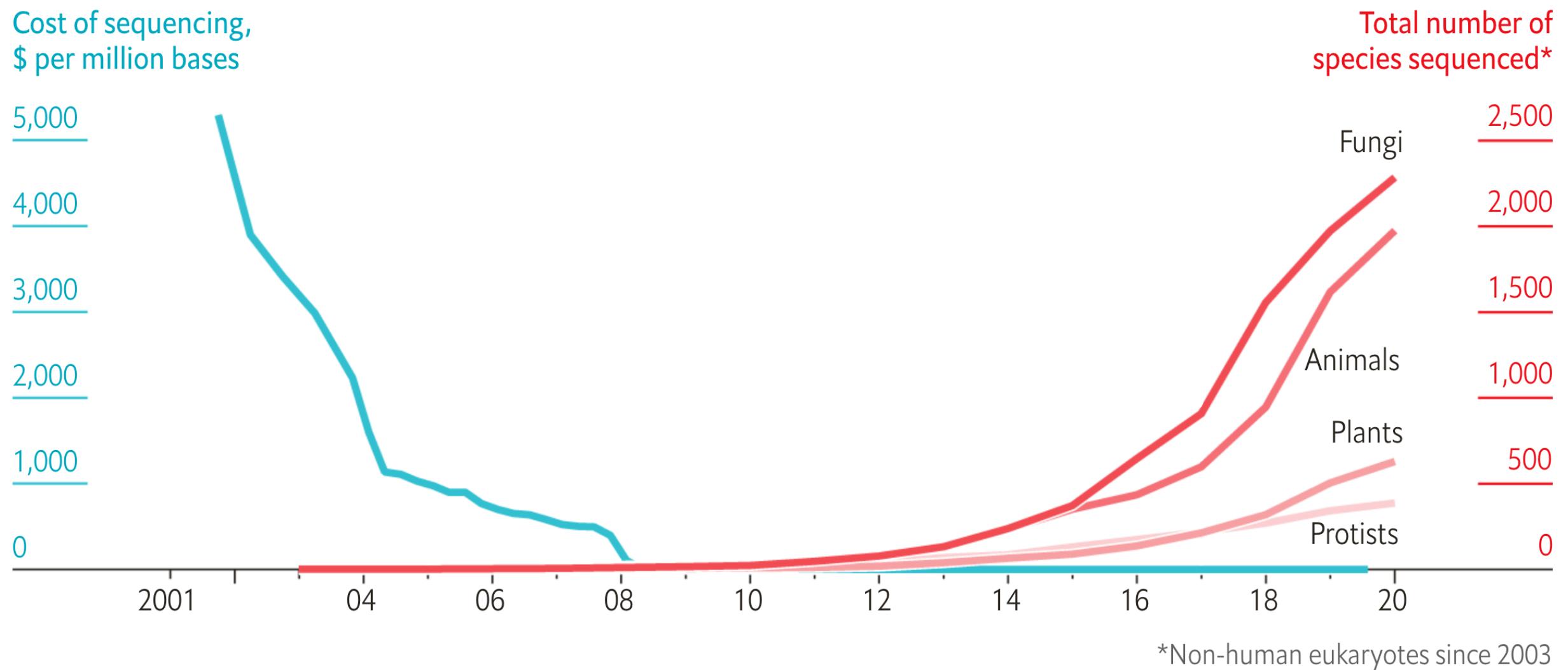
# Research summary: Matricellular signaling in PHACE syndrome

- PHACE syndrome is characterized by vascular anomalies and suspected to arise *de novo* based on inheritance pattern
- I found coding and noncoding *de novo* variants in 7 genes known to cause vascular malformations in a knockout mouse model
- *In silico* evidence suggests that these variants affect protein structure, splicing, and/or regulation
- Many of these gene products affect signaling through the Ras and PI3K pathways to regulate blood vessel growth
- Functional studies are underway to assess the pathogenicity of our candidate variants

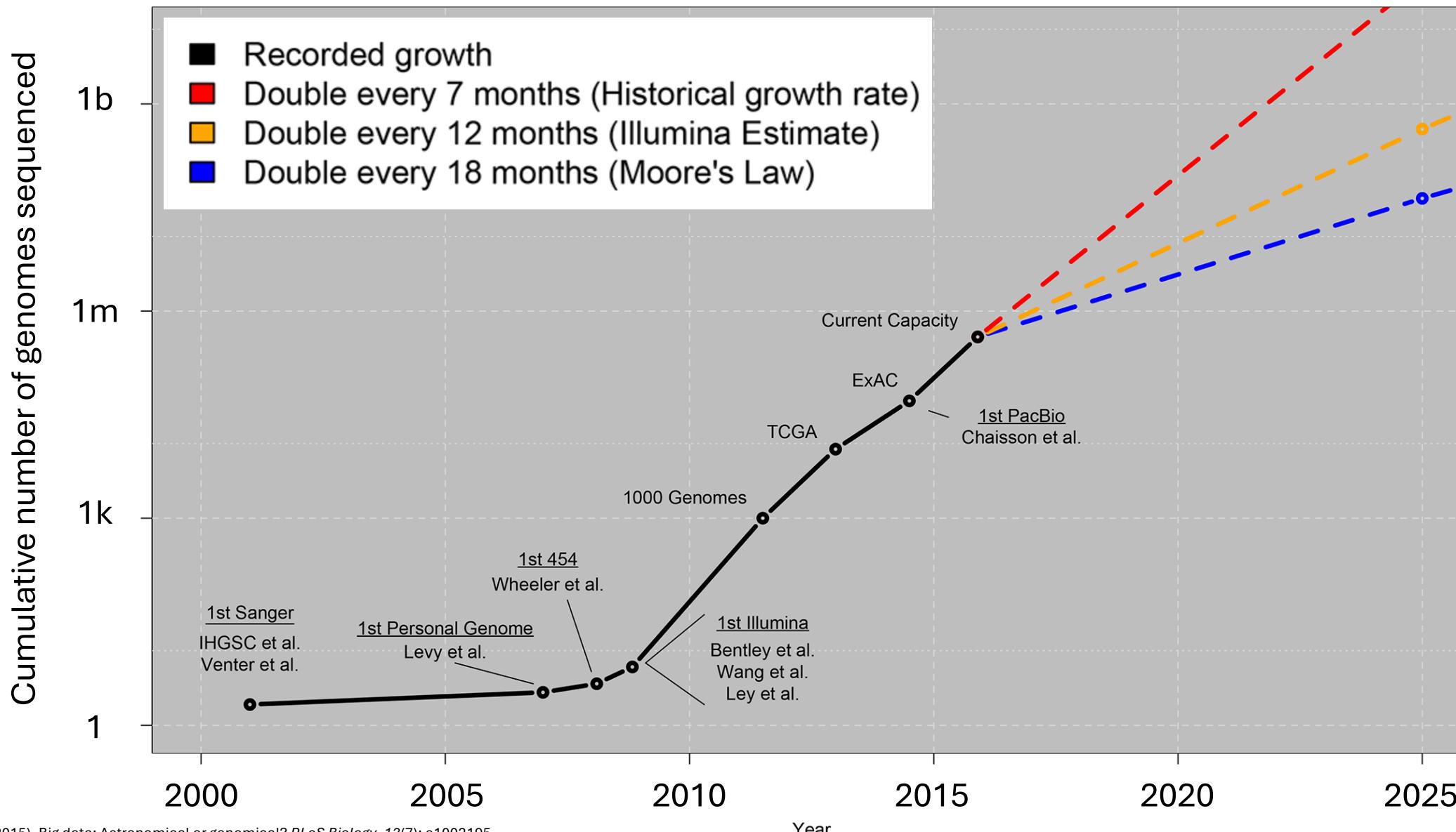
# Outline

- Background
- DNA sequencing methods
- Reference genomes
- Variant analysis workflow overview
- Cancer-specific considerations
- Variant annotations and prioritization
- Example research application
- **Implications and societal considerations**

# Growth of genome sequencing



# Growth of genome sequencing: Humans



# Growth of genome sequencing: Direct-to-consumer

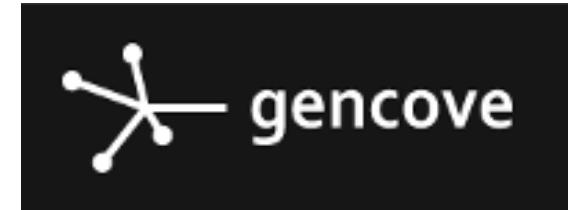
## Genotyping



## WES



## WGS



**Veritas**  
The Genome Company



... and more

# Growth of genome analysis



# Ethical, legal, and social implications (ELSI)

- Inequitable access to genomics
  - Financial
  - Ancestral
  - Geographic
  - Cultural / linguistic
- Genomic identifiability
- Potential harm to underrepresented communities
  - History of mistrust, lack of informed consent
  - Genealogy research can affect cultural identity, land rights
  - “Ownership” of biological materials

# Thank you!

*These materials have been developed by members of the teaching team at the [Harvard Chan Bioinformatics Core \(HBC\)](#). These are open access materials distributed under the terms of the [Creative Commons Attribution license \(CC BY 4.0\)](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.*

