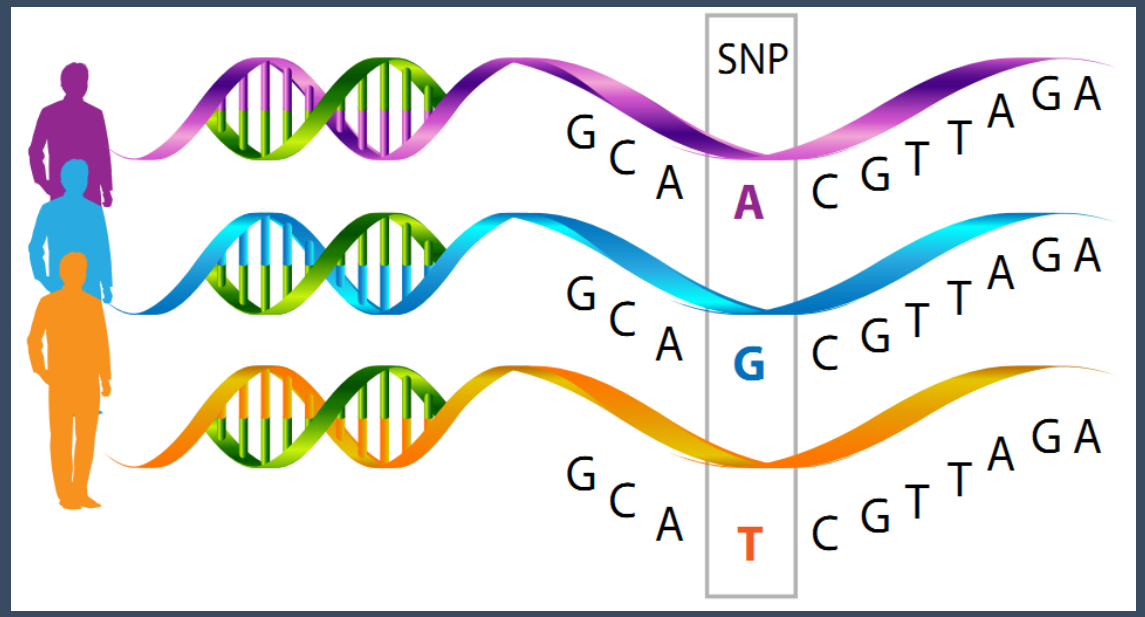
HBC
Harvard Chan Bioinformatics Core

Concepts and considerations for variant analysis in clinical genomics

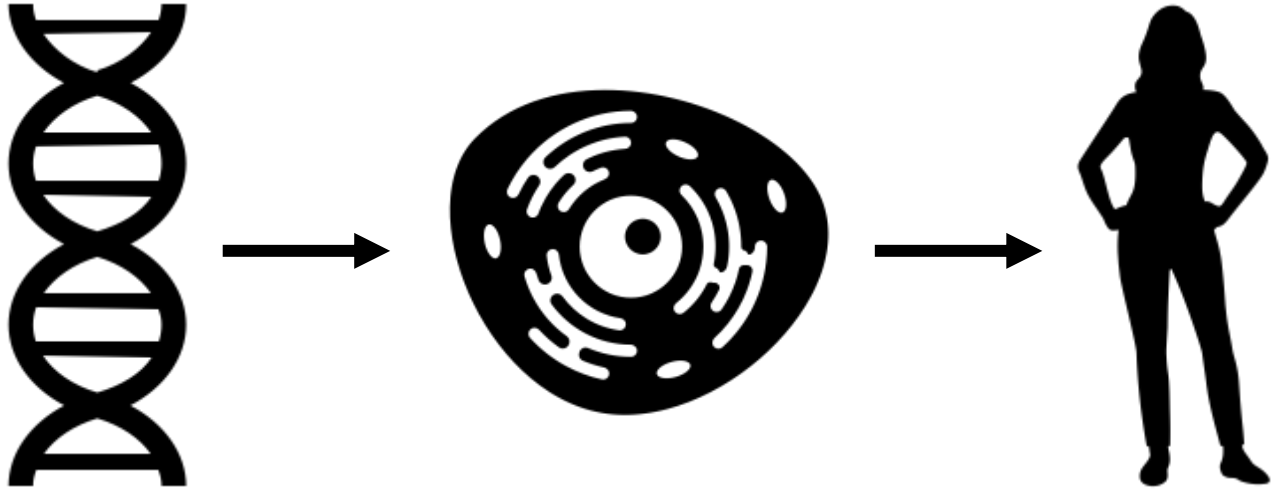


Elizabeth Partan, PhD
Harvard Chan Bioinformatics Core
May 28, 2024

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

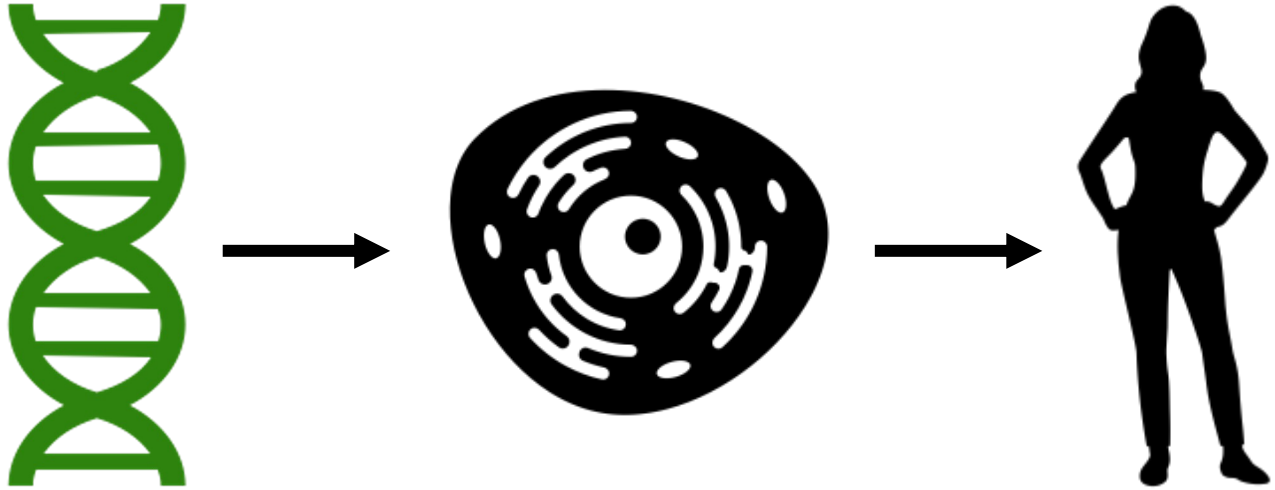


genotype (variant)

cellular signaling

phenotype (disease or trait)

DNA sequencing allows us to understand and treat disease



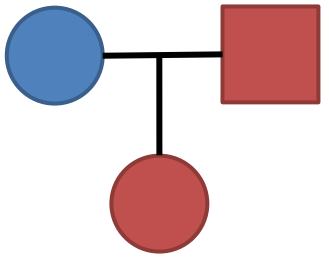
genotype (variant)

cellular signaling

phenotype (disease or trait)

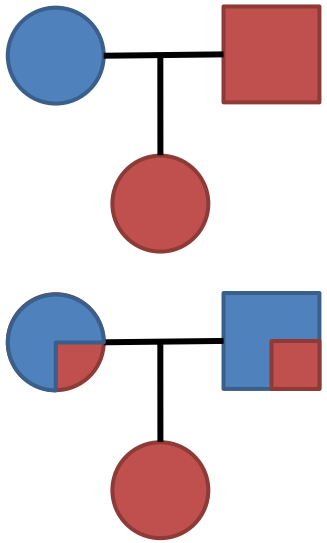
Variant and disease inheritance

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



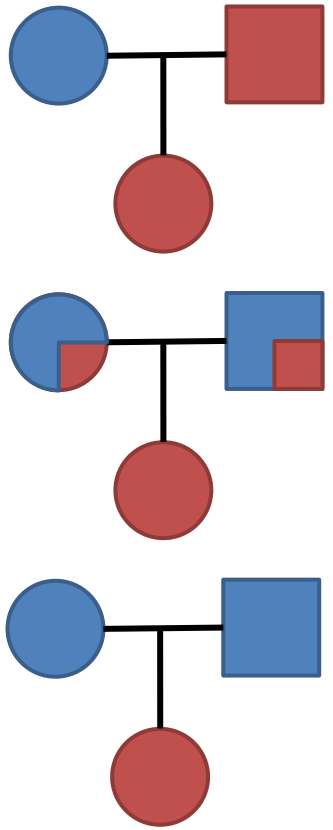
Variant and disease inheritance

- **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



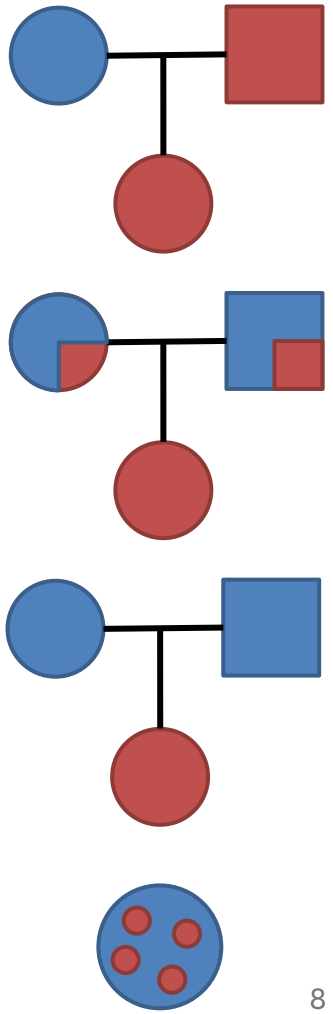
Variant and disease inheritance

- **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

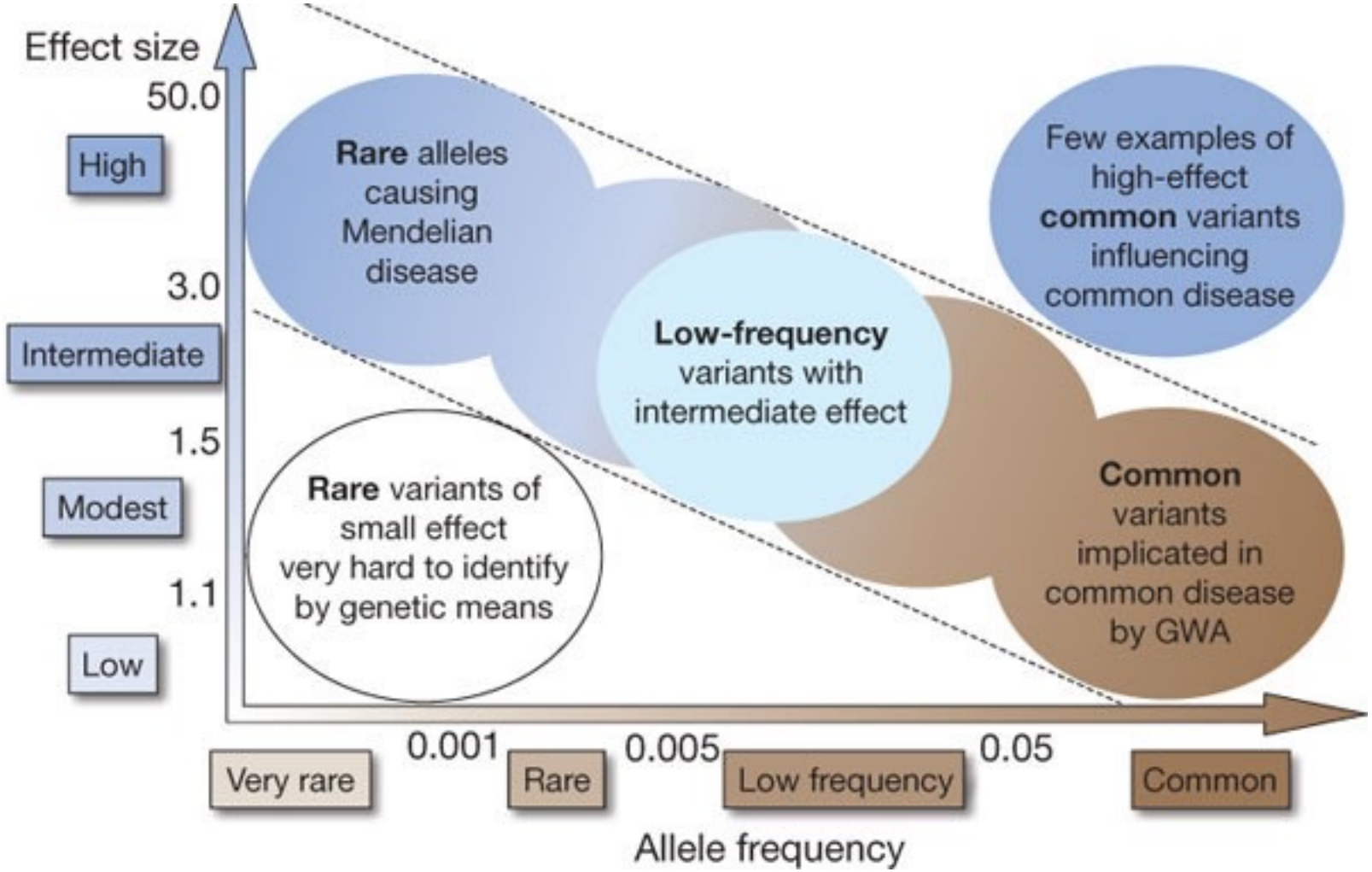


Variant and disease inheritance

- **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



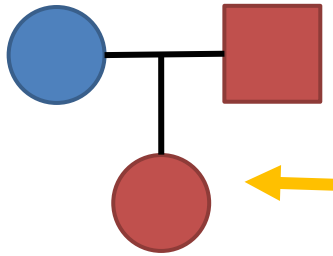
Types of genetic disease: Mendelian ↔ complex



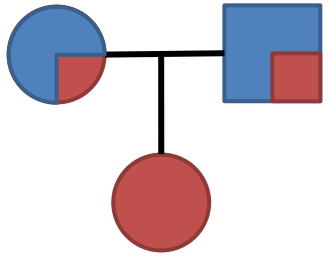
Manolio, et al. (2009). Finding the missing heritability of complex diseases. *Nature*, *461*, 747–753.

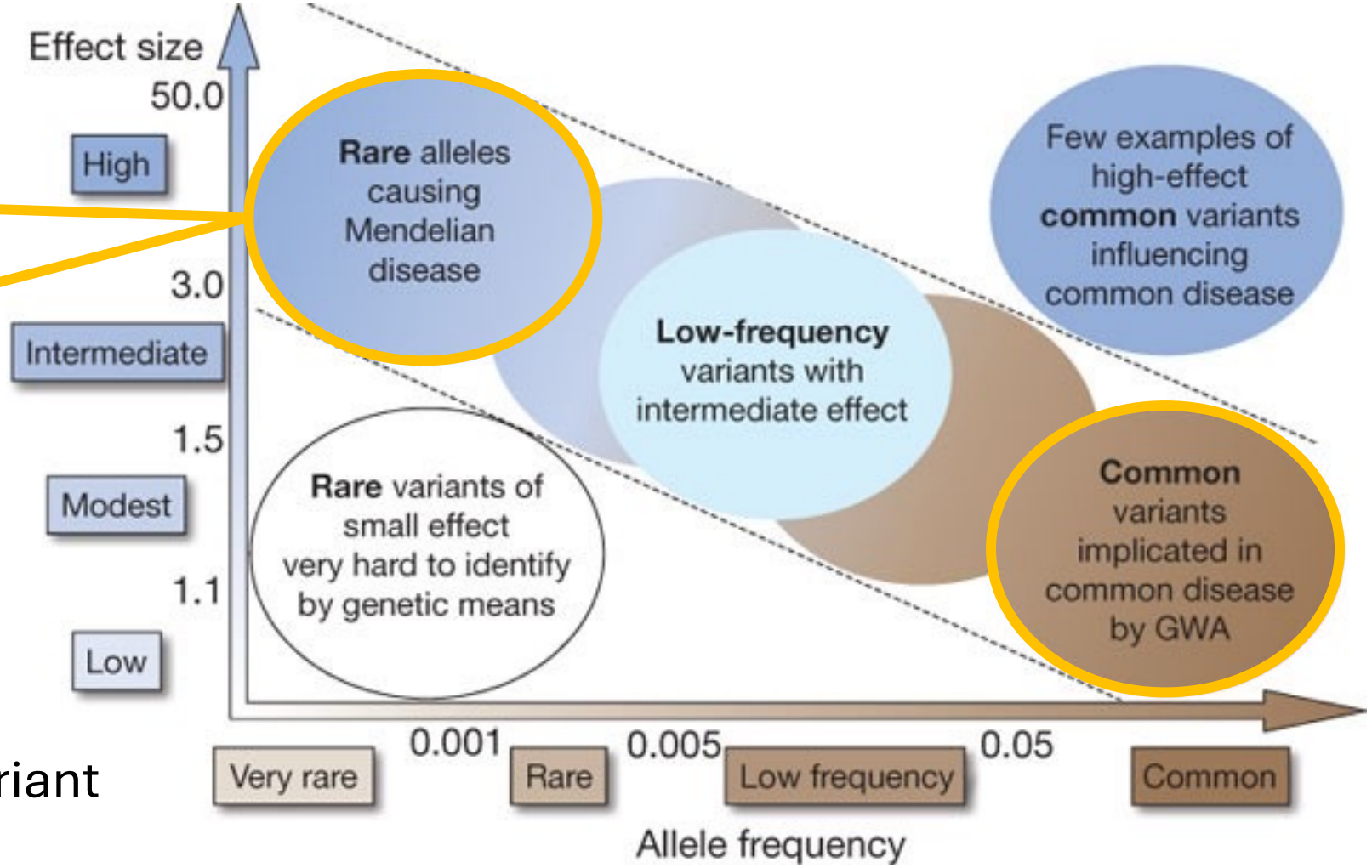
Types of genetic disease: Mendelian ↔ complex

Huntington's



Phenylketonuria





Mendelian disease:
single gene / variant

Diabetes
Heart disease
Cancer

Complex disease:
many variants + environment

Manolio, et al. (2009). Finding the missing heritability of complex diseases. *Nature*, *461*, 747–753.

How much of the DNA can we look at?

panel

- Gene panels: Selected genes or variants
  - Usually disease-focused

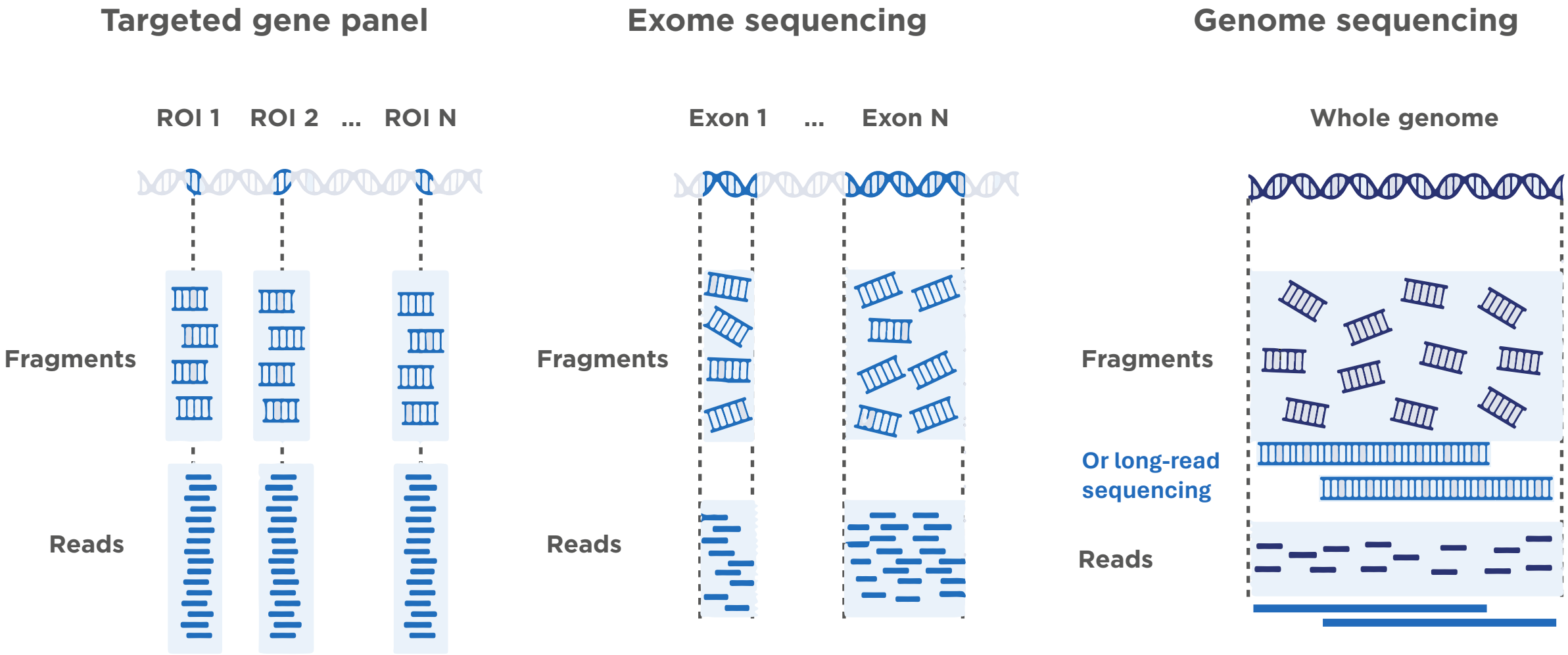
whole exome (WES)

- Exome: Exons, which encode proteins
  - 30-50 million base pairs (1-2% of the genome)

- Genome: Everything *
  - 3.2 billion base pairs

whole genome (WGS)

How does DNA sequencing work?



Adapted from Durães, et al. (2022). Demystifying the discussion of sequencing panel size in oncology genetic testing. *European Medical Journal*, *7*(2), 68-77.

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

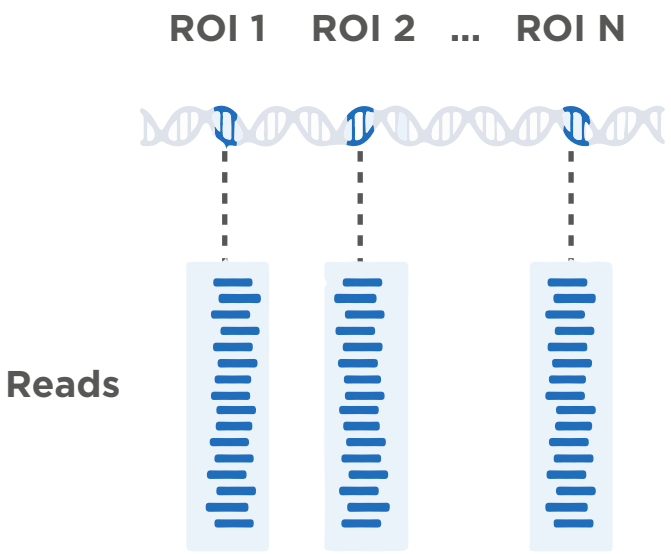
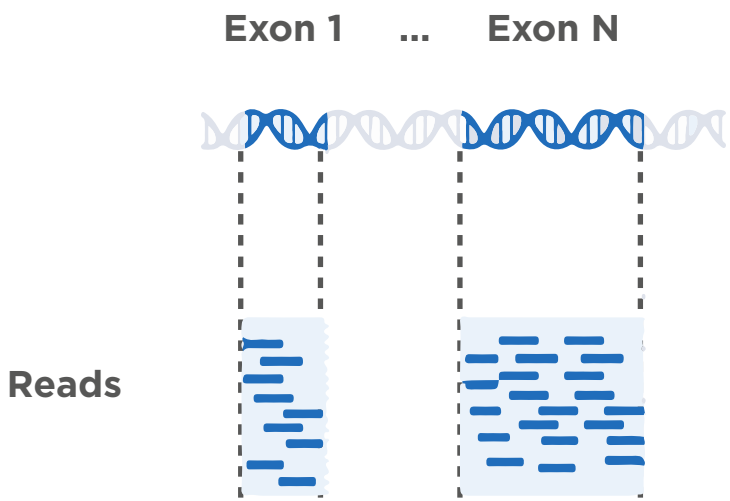
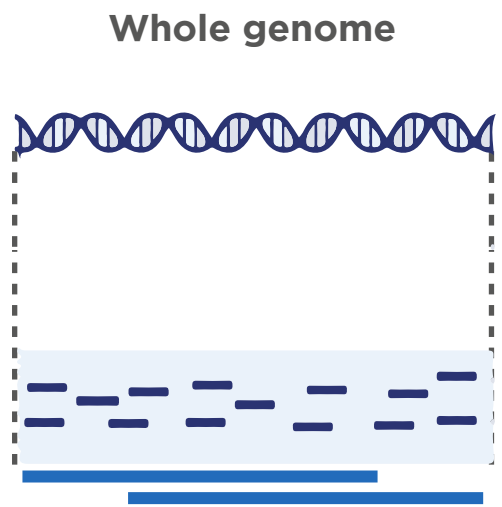
Which sequencing method is appropriate?

- 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?

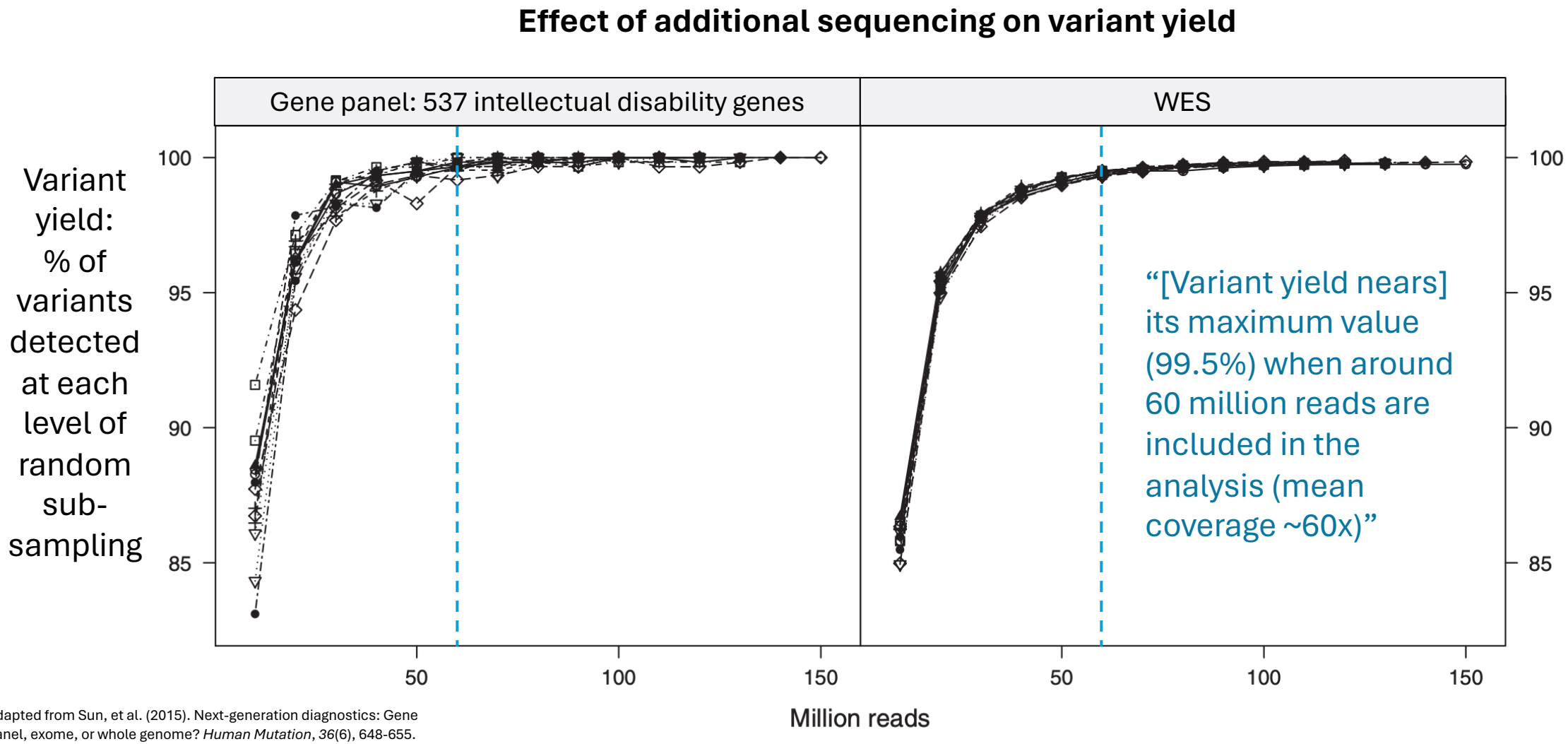
- Gene panels
  - Most common for clinical diagnostic analysis: cheapest, fastest, fewest variants of uncertain significance (VUS)
  - Need to know disease-associated variants in advance
- Exome
  - Multisystemic diseases not covered by a single panel
  - Will miss everything not captured (just exons,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

Targeted gene panel	Exome sequencing	Genome sequencing
		
Coverage 100–500X+	70–100X	Minimum 30X

Adapted from Durães, et al. (2022). Demystifying the discussion of sequencing panel size in oncology genetic testing. *European Medical Journal*, *7*(2), 68-77.

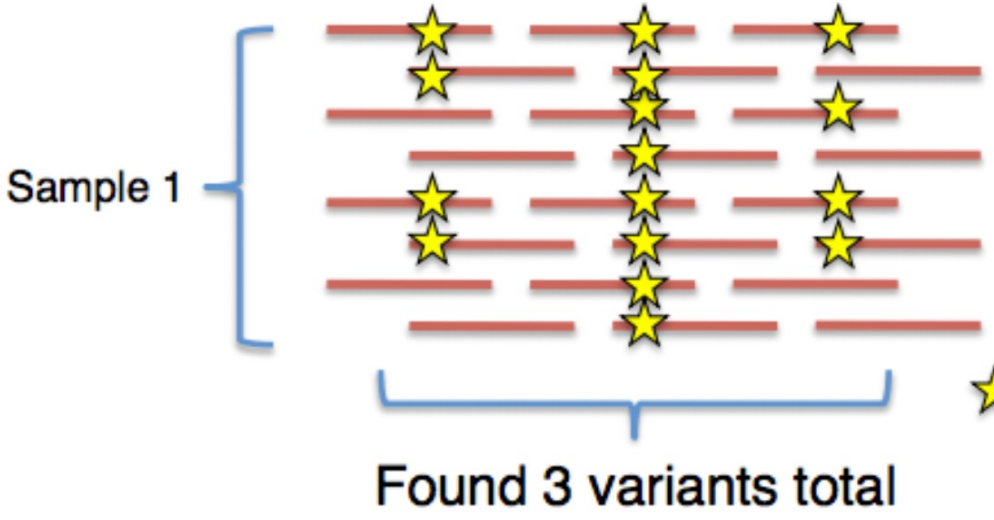
Greater sequencing depth increases confidence but will not find new variants



Adapted from Sun, et al. (2015). Next-generation diagnostics: Gene panel, exome, or whole genome? *Human Mutation*, *36*(6), 648-655.

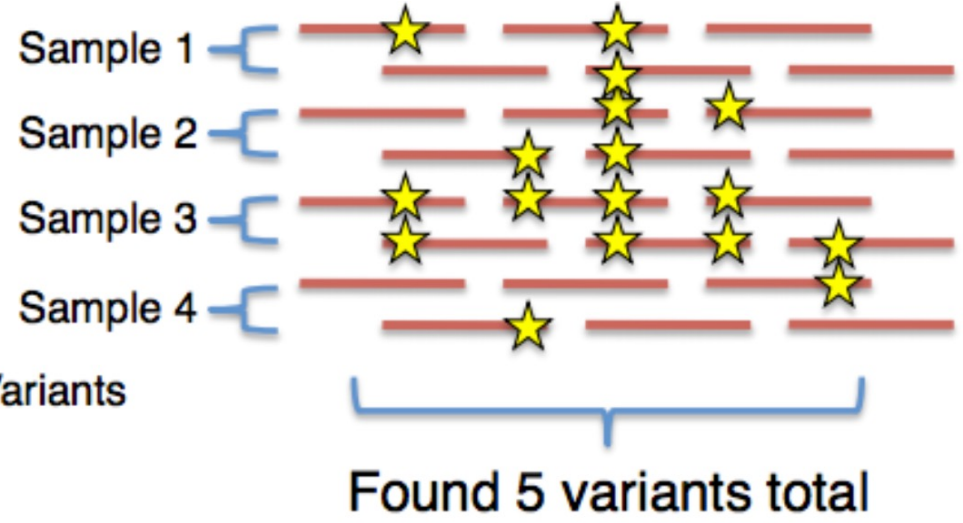
Sequencing depth and cost: With a fixed sequencing budget, you can perform…

Deep single-sample sequencing



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

Shallower multi-sample sequencing

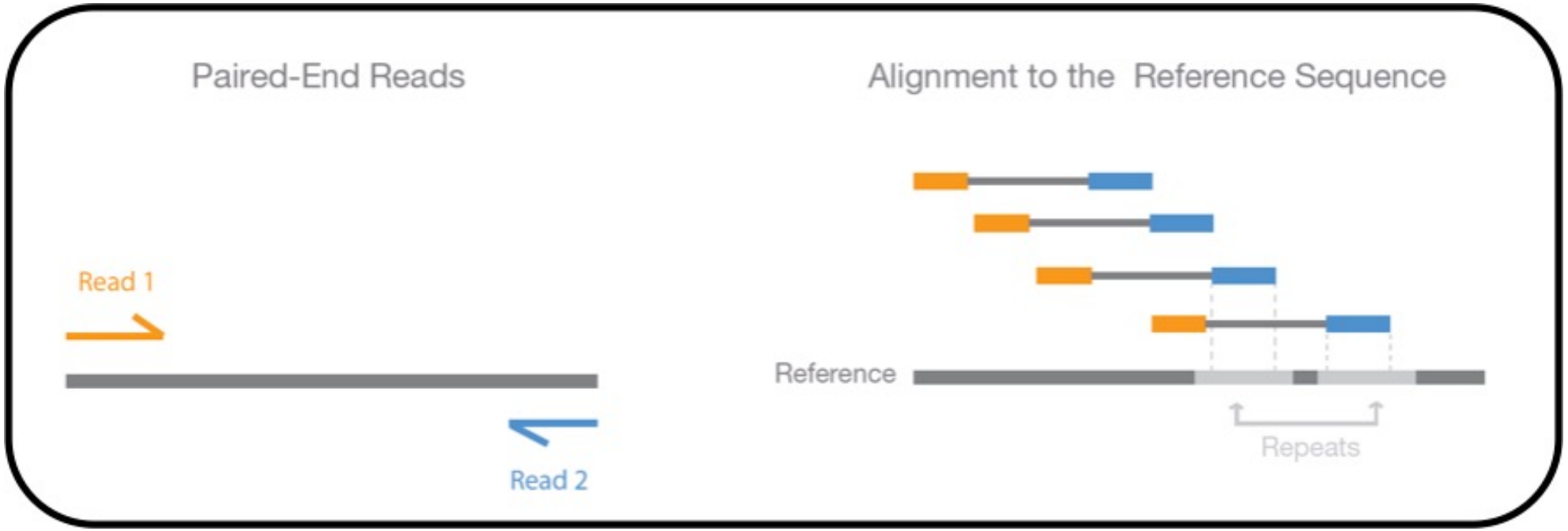


- Sensitivity dependent on frequency of variation
- More total variants discovered

Adapted from GATK best practices guidelines

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



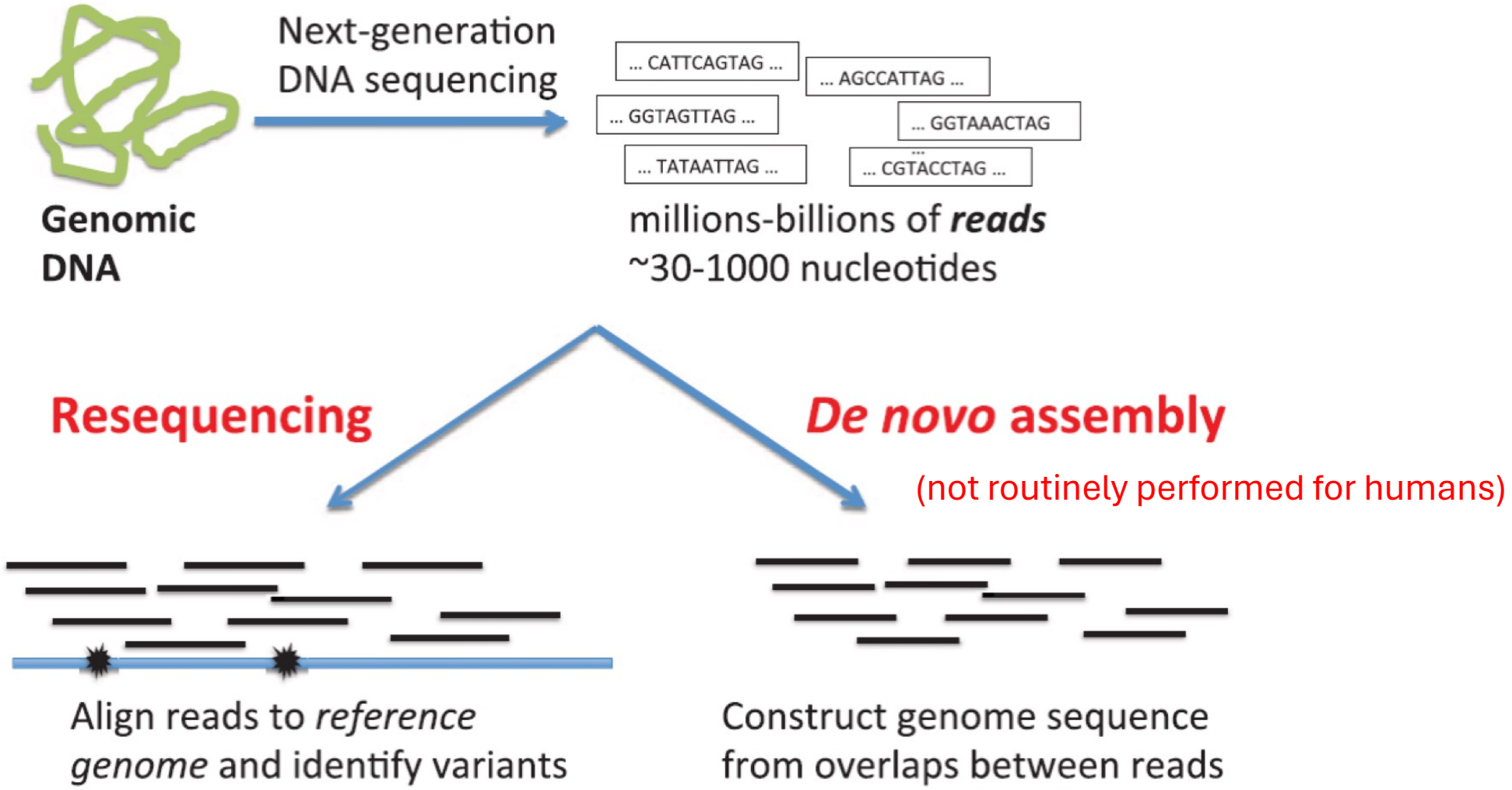
https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/paired-end-vs-single-read.html

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

https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/read-length.html
Tan, et al. (2019). Long fragments achieve lower base quality in Illumina paired-end sequencing. *Scientific Reports*, *9*(1), 2856.

Comparison to reference

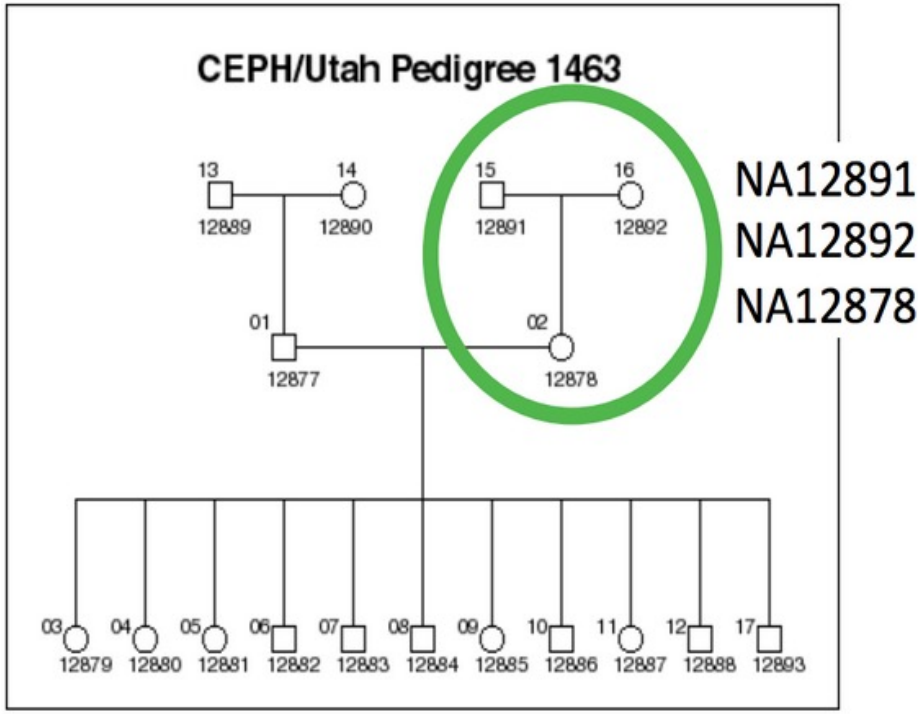


Raphael. (2012). Chapter 6: Structural Variation and Medical Genomics. *PLoS Computational Biology*, *8*(12). e1002821.

Reference genomes: Individual humans

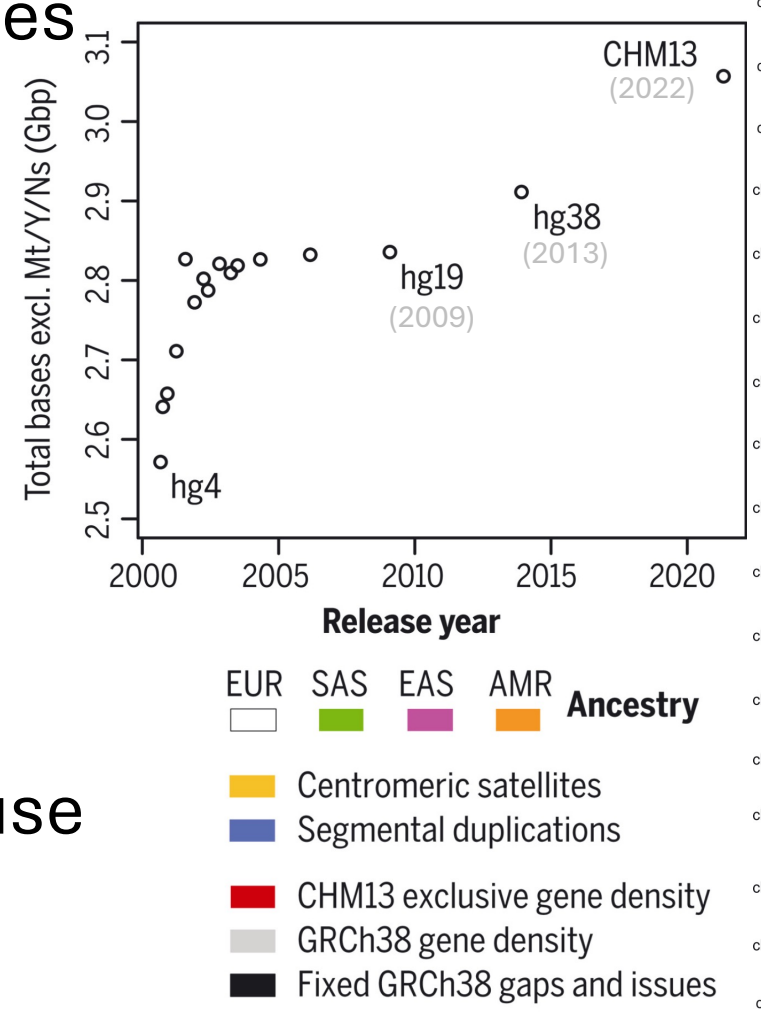
- Well studied, gold standard
- Family (trio) data
- Used to benchmark sequencing platforms and analysis pipelines

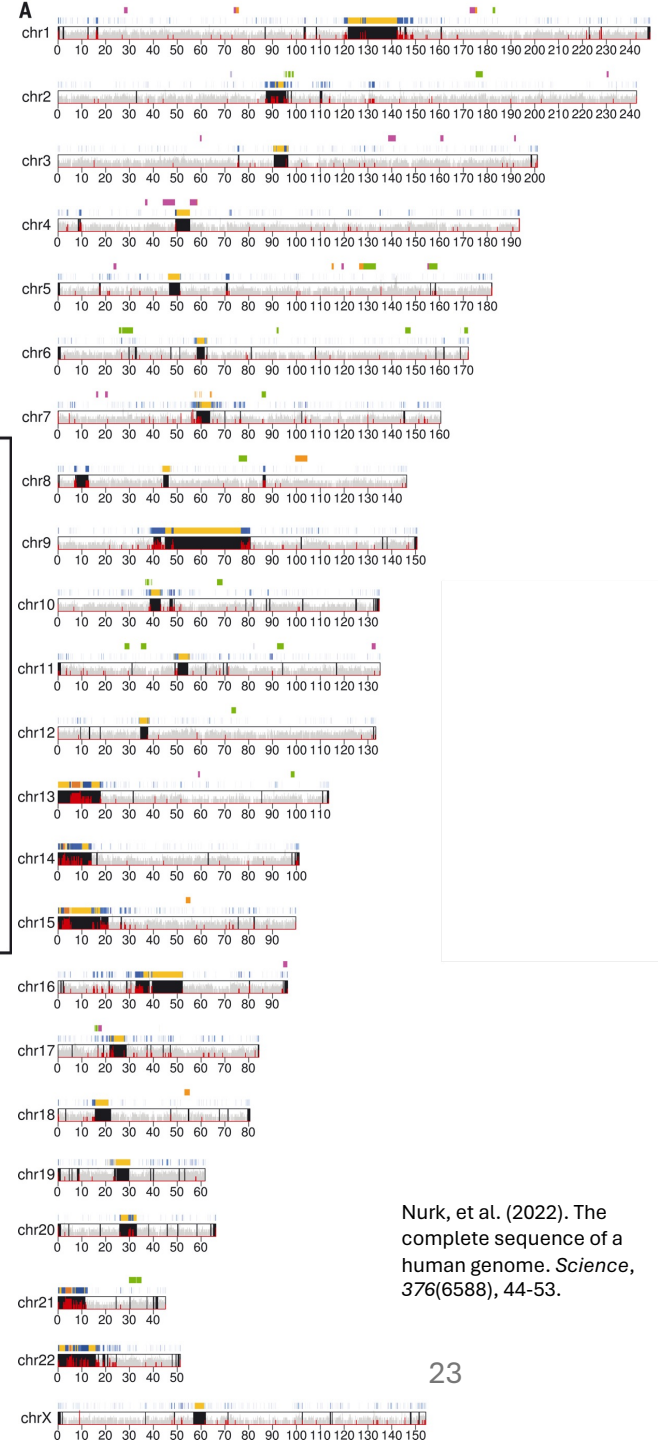
- NA12891
- Genome in a Bottle (GIAB)
- HapMap
- 1000 Genomes



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

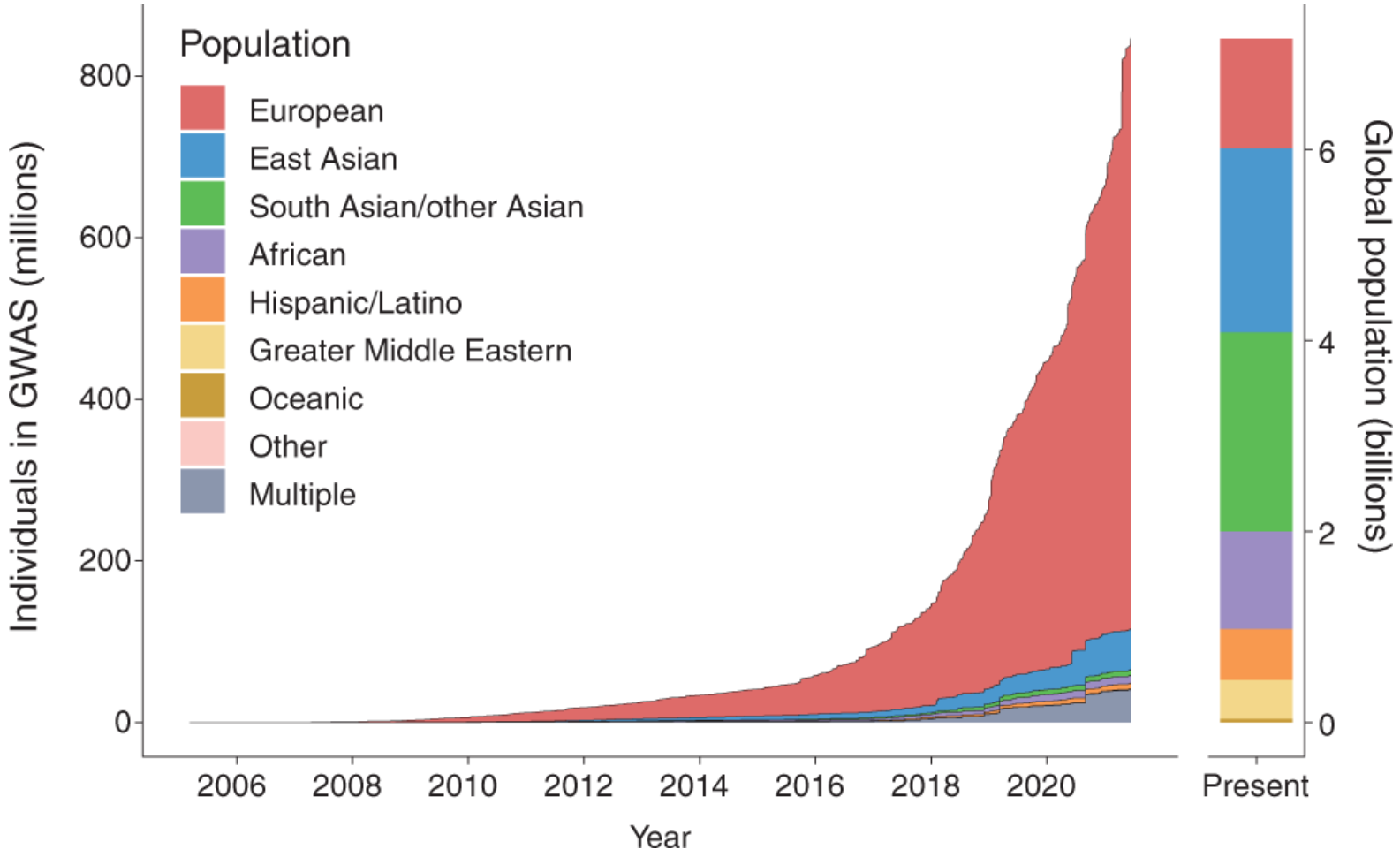




Nurk, et al. (2022). The complete sequence of a human genome. *Science*, *376*(6588), 44-53.

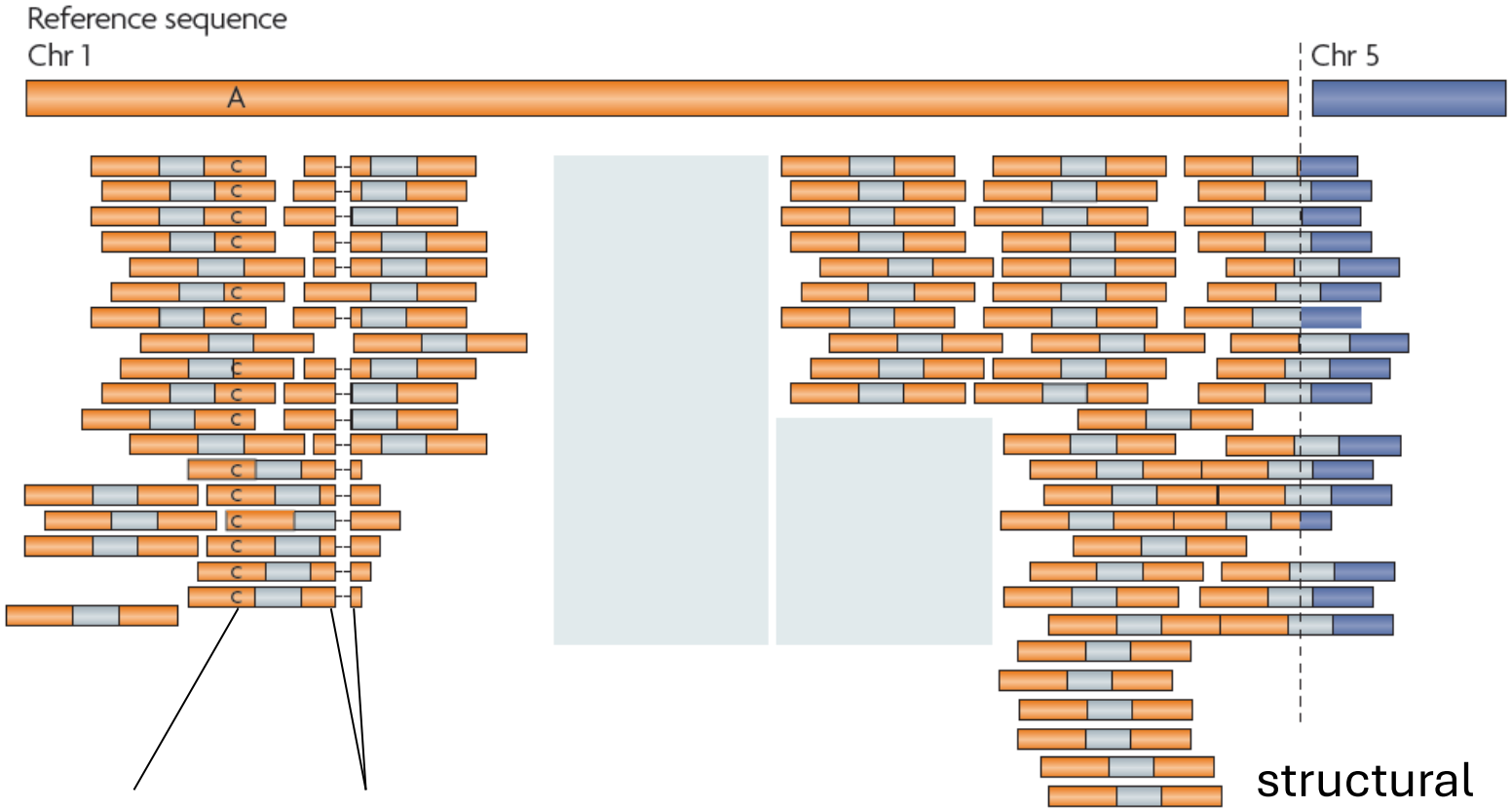
Reference genomes: “Pan-genome”

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



Fatumo, et al. (2022). A roadmap to increase diversity in genomic studies. *Nature Medicine*, *28*, 243–250.

What types of variants can DNA sequencing detect?



single base (SNV)

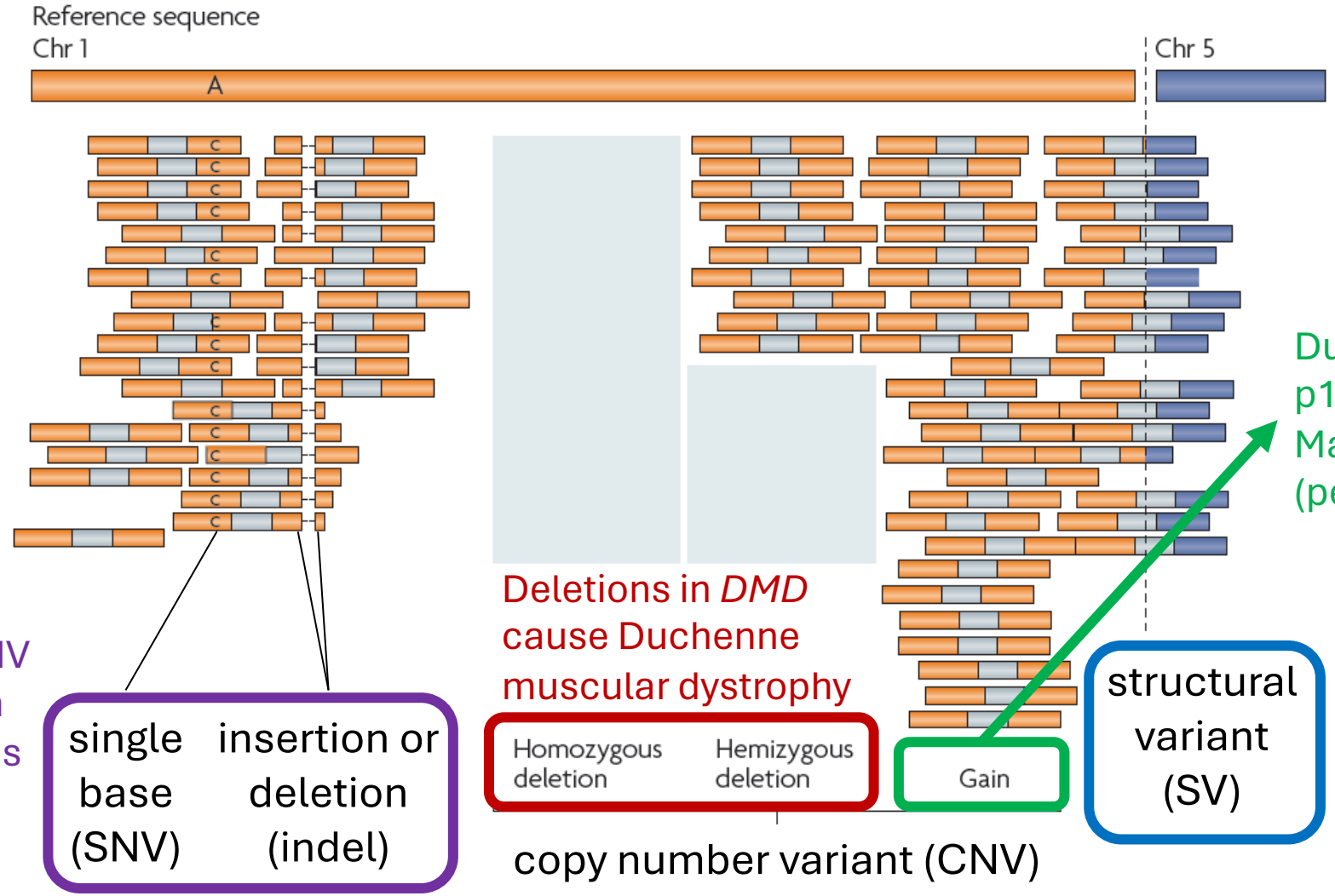
insertion or deletion (indel)

structural variant (SV)

copy number variant (CNV)

Meyerson, et al. (2010). Advances in understanding cancer genomes through second-generation sequencing. *Nature Reviews Genetics*, *11*(10), 685-696.

What types of variants can DNA sequencing detect?



CFTR p.Gly542X SNV and ΔF508 deletion cause cystic fibrosis

single base (SNV)

insertion or deletion (indel)

Deletions in *DMD* cause Duchenne muscular dystrophy

copy number variant (CNV)

Duplication of 17 p12-p11.2 causes Charcot-Marie-Tooth disease (peripheral neuropathy)

BCR-ABL1 translocation t(9;22)(q34;q11.2) causes chronic myeloid leukemia (CML)

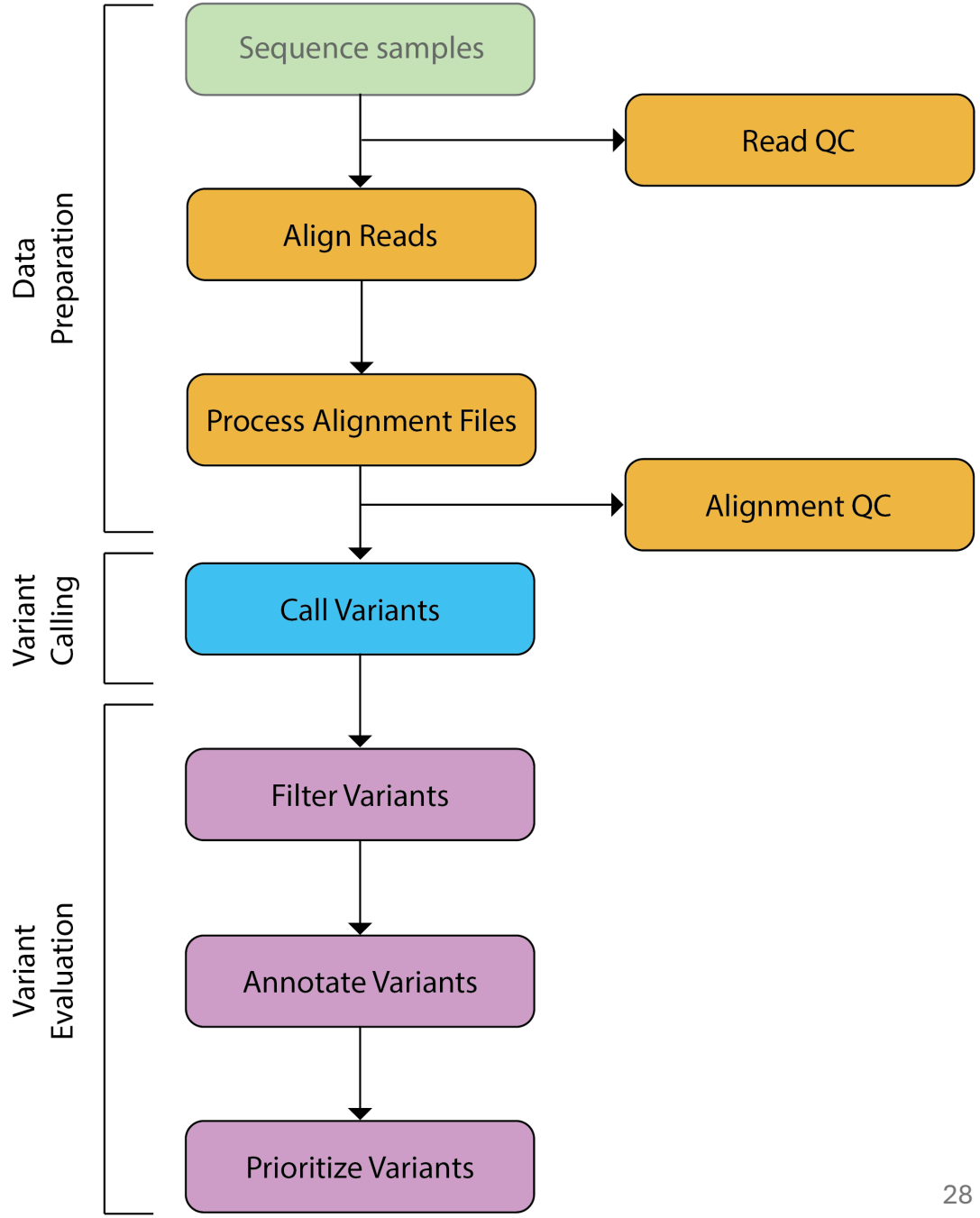
structural variant (SV)

Meyerson, et al. (2010). Advances in understanding cancer genomes through second-generation sequencing. *Nature Reviews Genetics*, *11*(10), 685-696.

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**

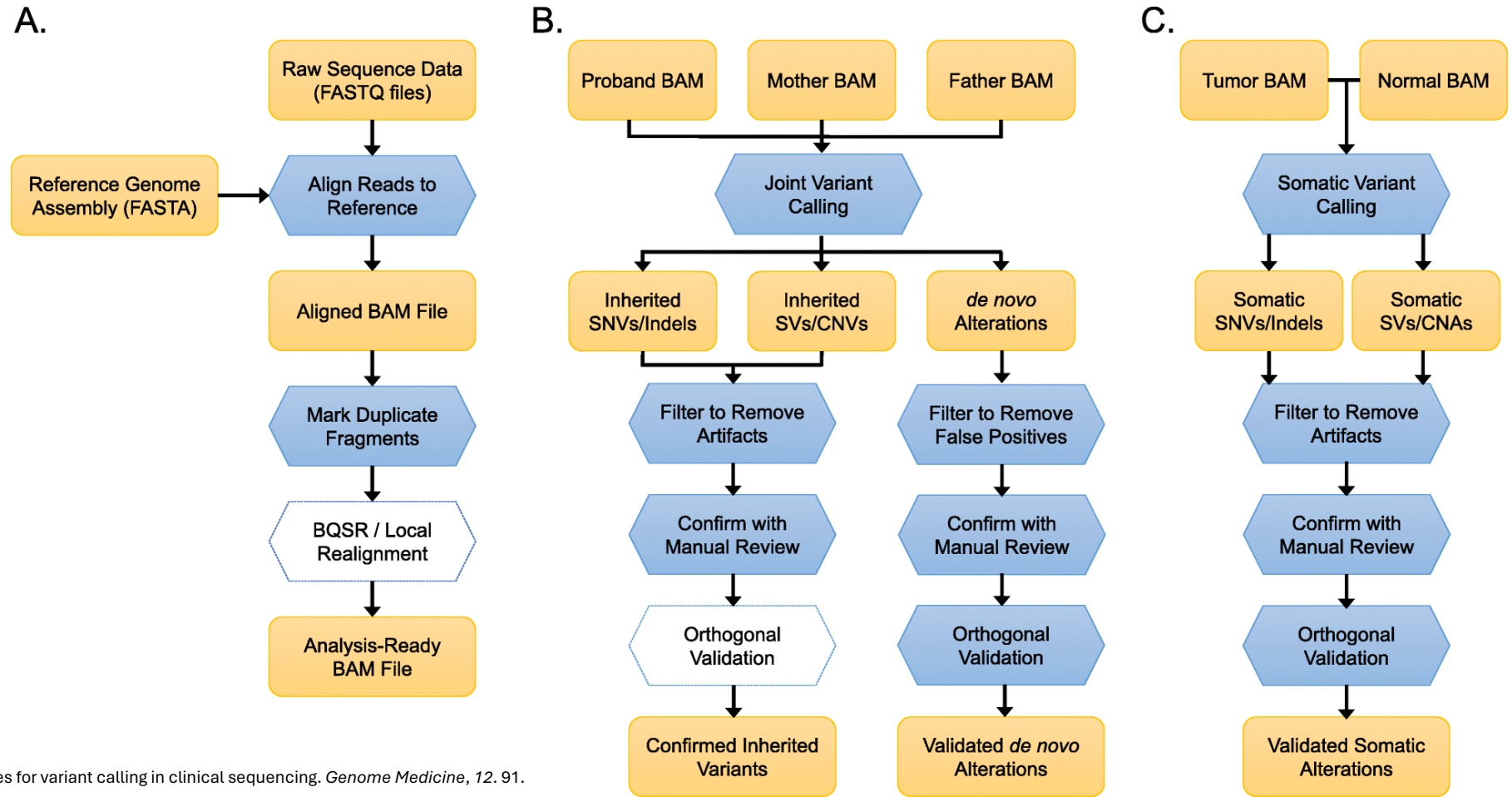
- Remove low quality reads
- Remove low quality alignments
- Remove PCR duplicate reads
- Local realignment around indels
- Base quality score recalibration (BQSR)
  * May no longer be required due to technological improvements

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**



Koboldt. (2020). Best practices for variant calling in clinical sequencing. *Genome Medicine*, *12*. 91.

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

WGS: The 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature*, *526*, 68-74.
WES: Dewey, et al. (2016). Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. *Science*, *354*(6319), aaf6814

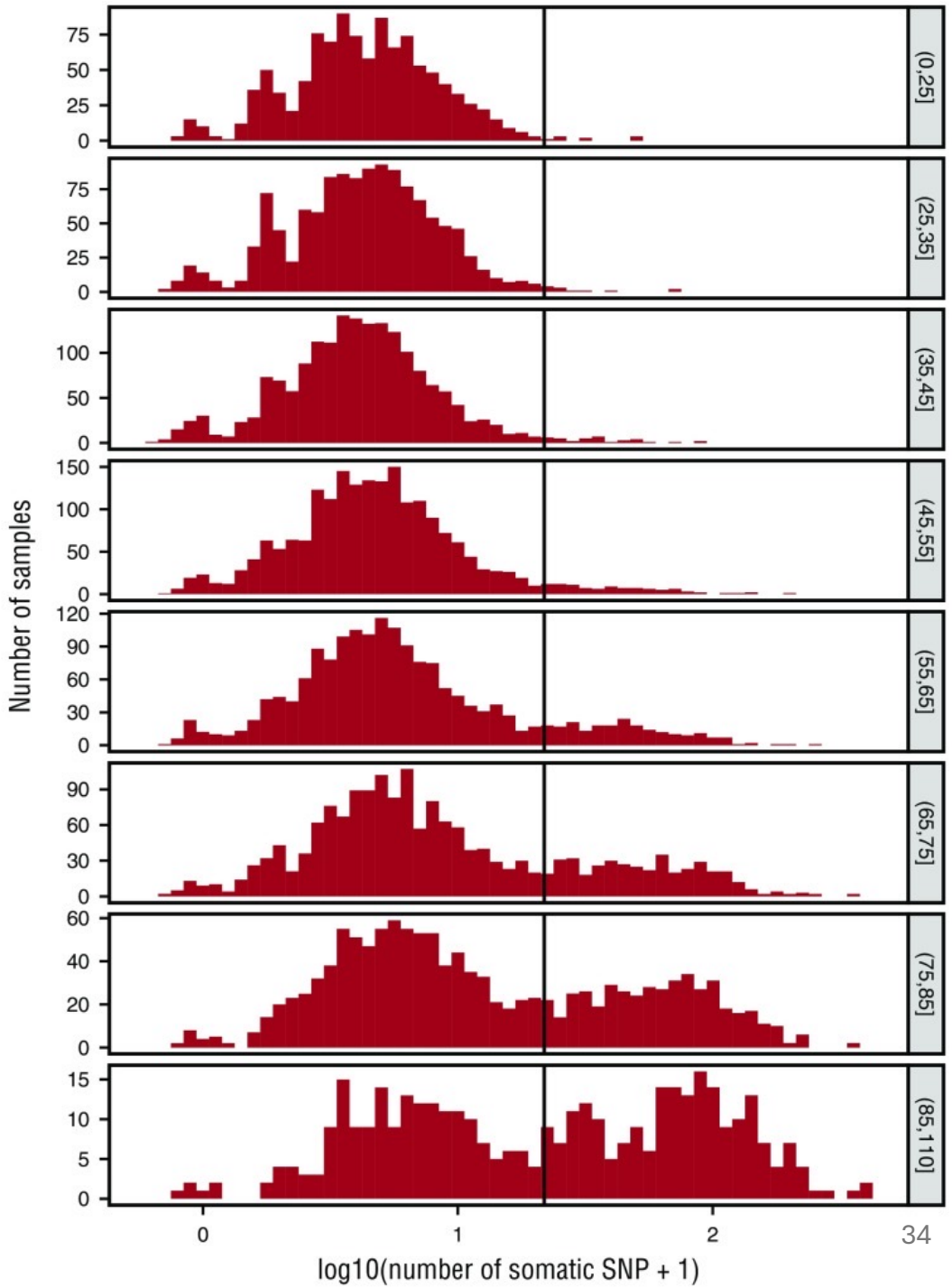
Variant calling: Joint

- **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

Koboldt. (2020). Best practices for variant calling in clinical sequencing. *Genome Medicine*, *12*. 91.

Variant calling: Somatic

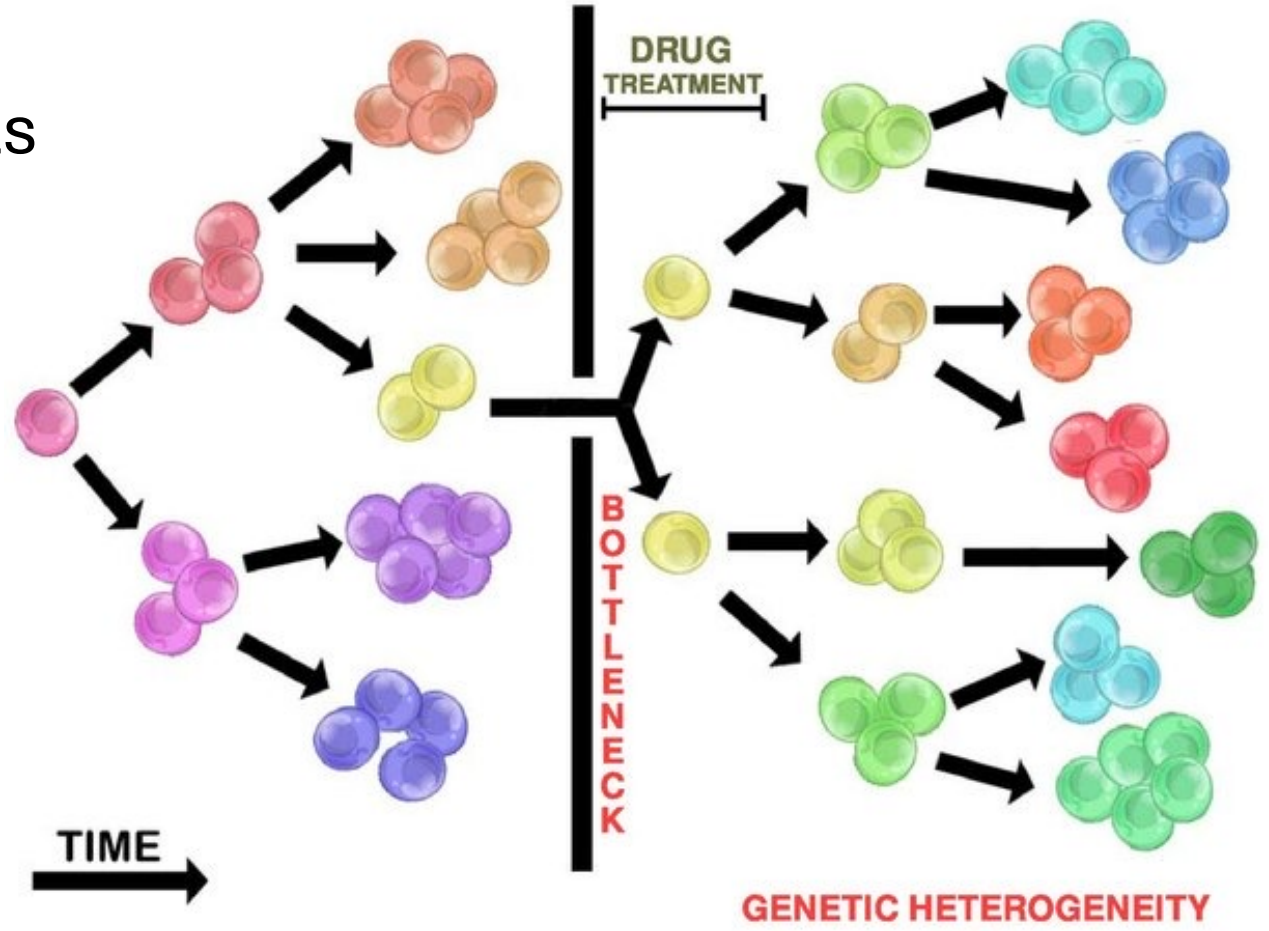
- **Affected (e.g., tumor) vs unaffected tissue within a single individual**
- Expect fewer variants than in germline, otherwise suspect DNA damage
- Higher frequency in sun-exposed tissue, older individuals



Zink, et al. (2017). Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly. *Blood*, *130*(6), 742–752. [Graph]
Koboldt. (2020). Best practices for variant calling in clinical sequencing. *Genome Medicine*, *12*. 91.

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



https://en.wikipedia.org/wiki/Tumour_heterogeneity
Meyerson, et al. (2010). Advances in understanding cancer genomes through second-generation sequencing. *Nature Reviews Genetics*, *11*(10), 685-696.

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

Chang, et al. (2022). Somatic and germline variant calling from next-generation sequencing data. In: Laganà (ed.), *Computational Methods for Precision Oncology*, *Advances in Experimental Medicine and Biology*, *1361*, 37-54.

Allele frequency: across a population

- **Allele frequency for a given population =**

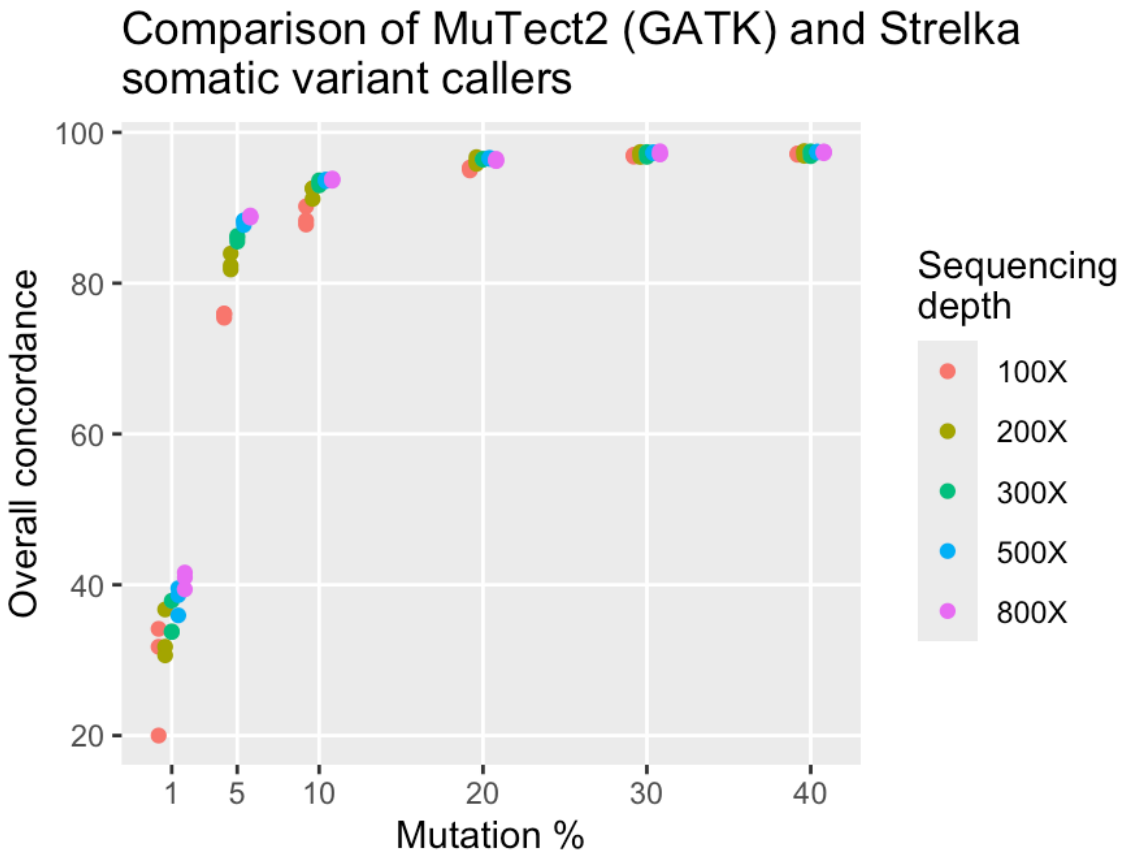
$$\frac{\textbf{[\# chromosomes with allele]}}{\textbf{[size of population]}}$$

- Rare variant: < 1–5% minor allele frequency (MAF)

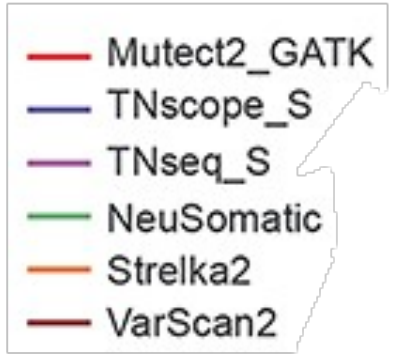
Chang, et al. (2022). Somatic and germline variant calling from next-generation sequencing data. In: Laganà (ed.), *Computational Methods for Precision Oncology*, *Advances in Experimental Medicine and Biology*, *1361*, 37-54.

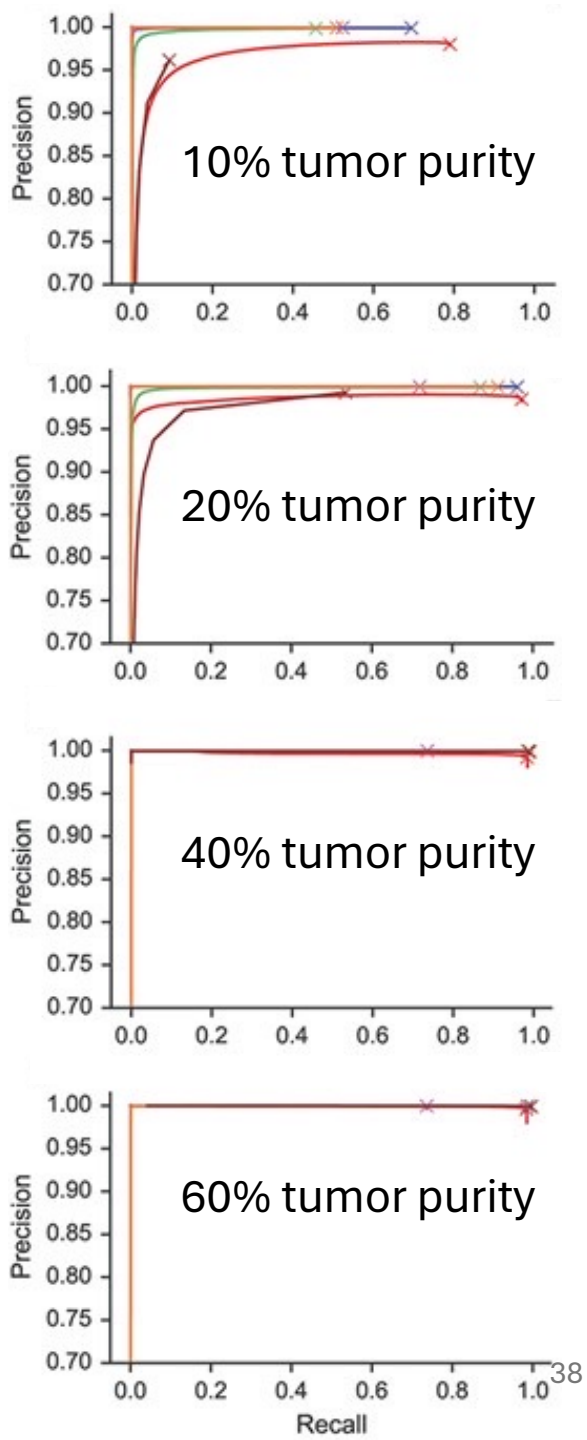
Cancer / somatic variant calling presents unique challenges

- Somatic variant callers frequently disagree



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.

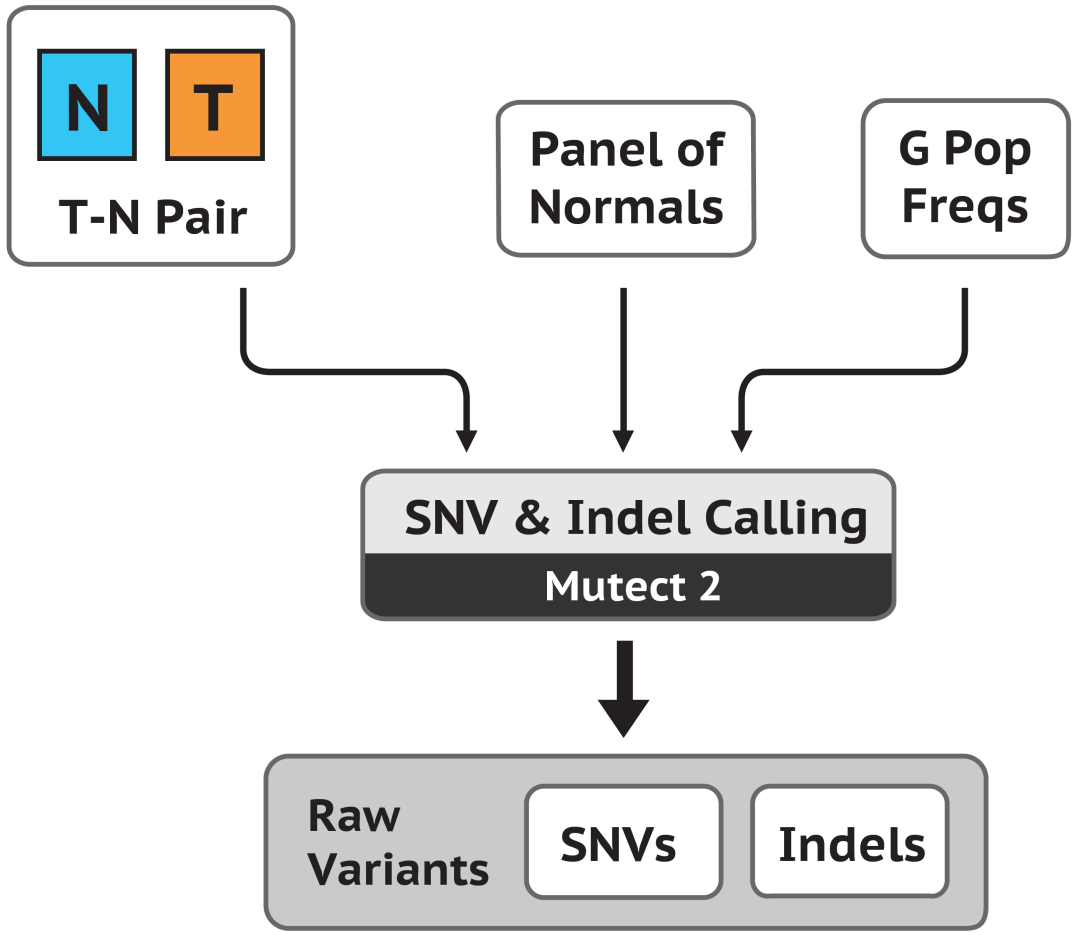




Pei, et al. (2021). Benchmarking variant callers in next-generation and third-generation sequencing analysis. *Briefings in Bioinformatics*, *22*(3), bbaa148.

MuTect2

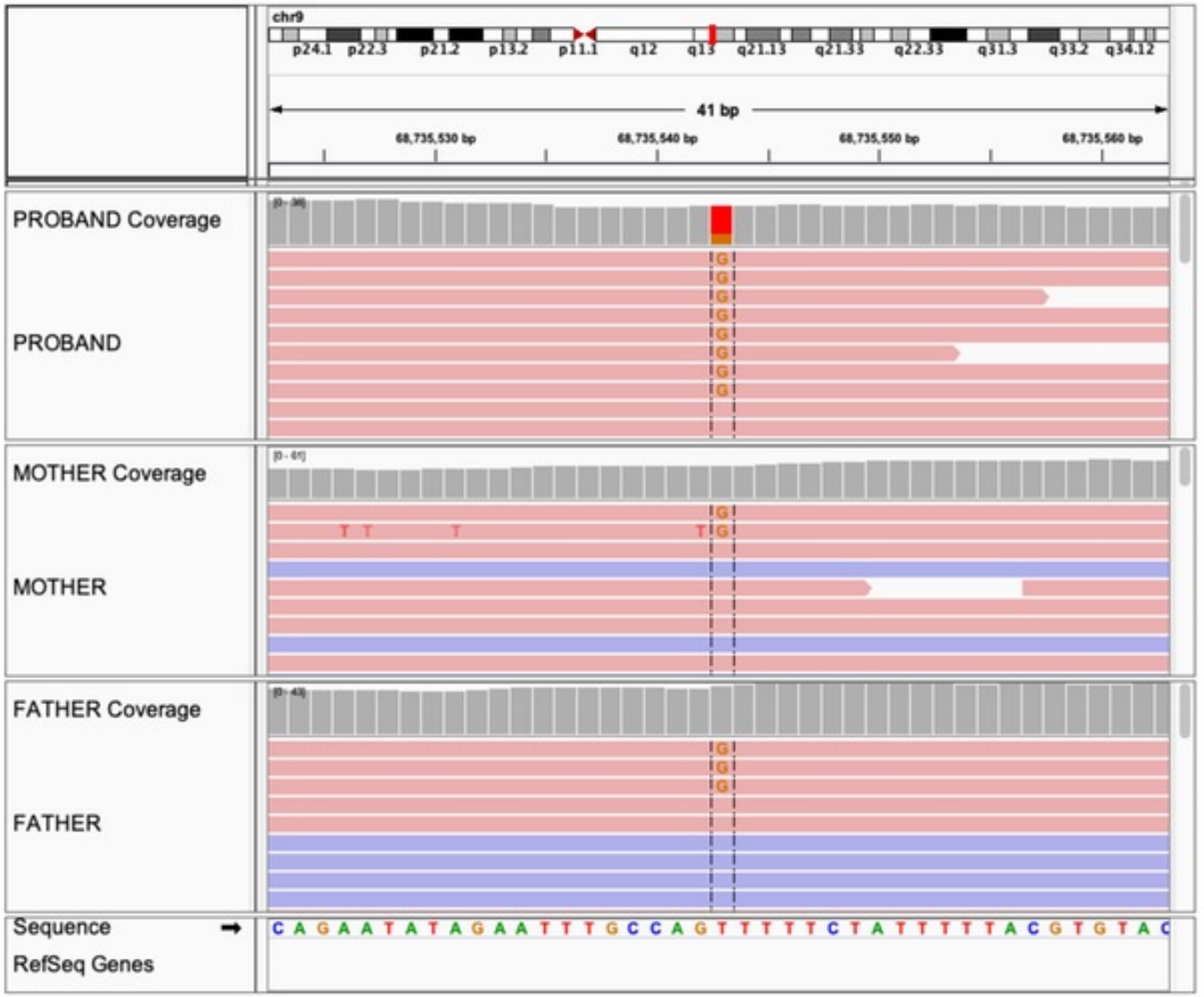
- Only calls short variants (SNVs, indels)
- Tumor-normal and tumor only modes
- Panel of Normals (PoN): remove sequencing noise



https://gatk.broadinstitute.org/hc/en-us/articles/360037593851-Mutect2
Image adapted from https://github.com/TRON-Bioinformatics/tronflow-mutect2

Validation: Go look at the reads yourself in IGV

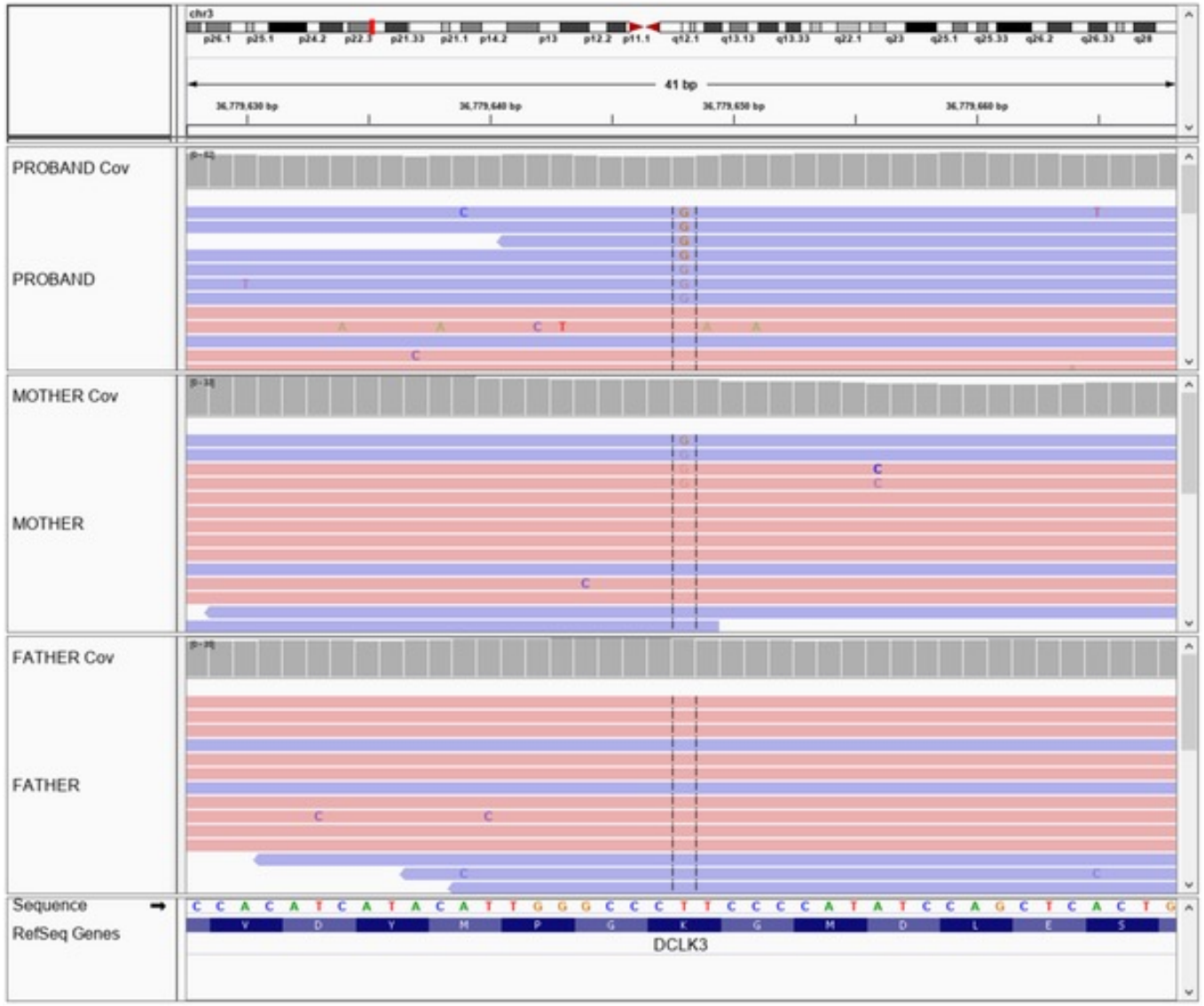
- **Strand bias**
  - Variants seen only on reverse strand (red)



Koboldt. (2020). Best practices for variant calling in clinical sequencing. *Genome Medicine*, *12*. 91.

Validation: Go look at the reads yourself in IGV

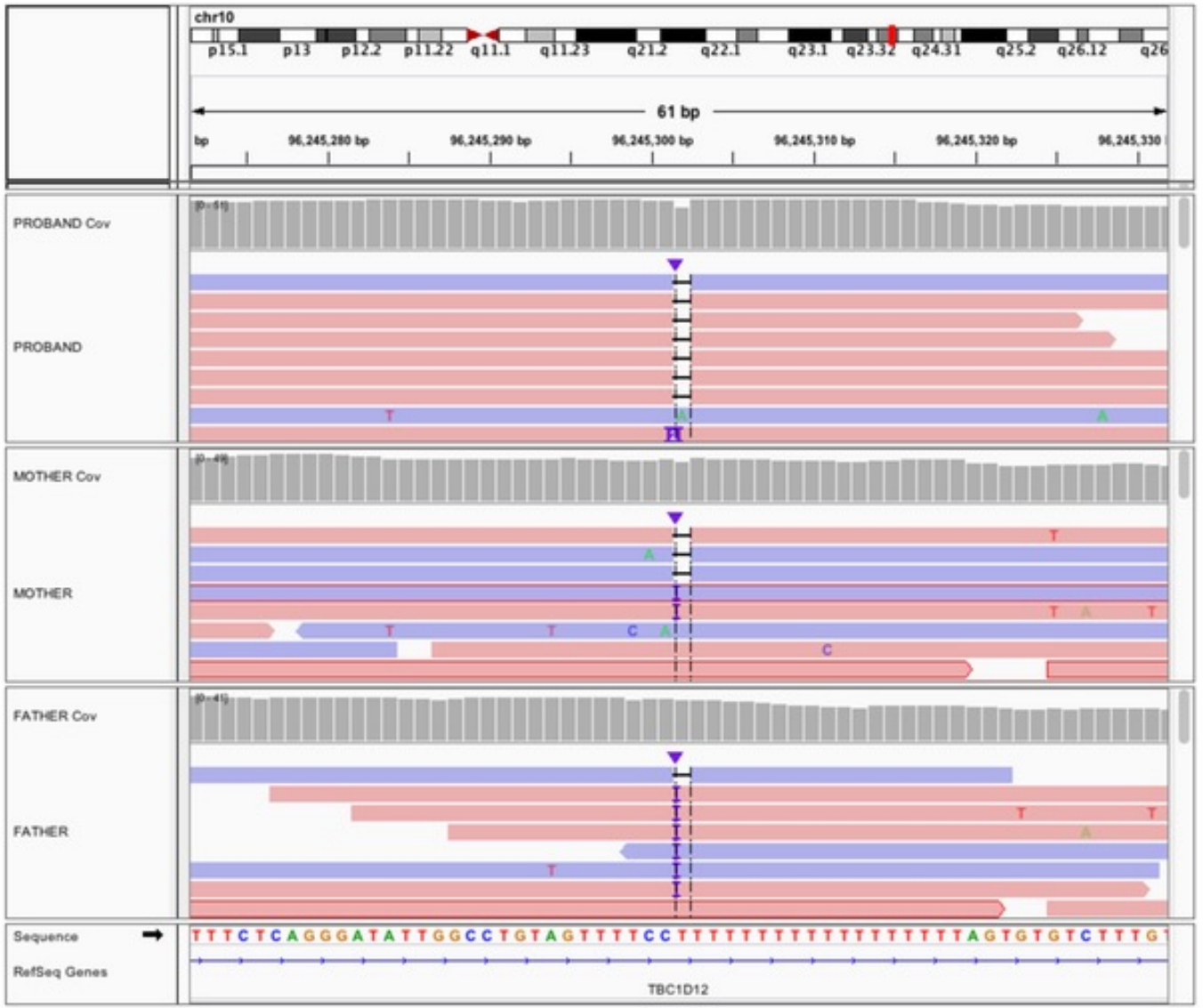
- **Low quality**
  - Variants are pale rather than dark



Koboldt. (2020). Best practices for variant calling in clinical sequencing. *Genome Medicine*, *12*. 91.

Validation: Go look at the reads yourself in IGV

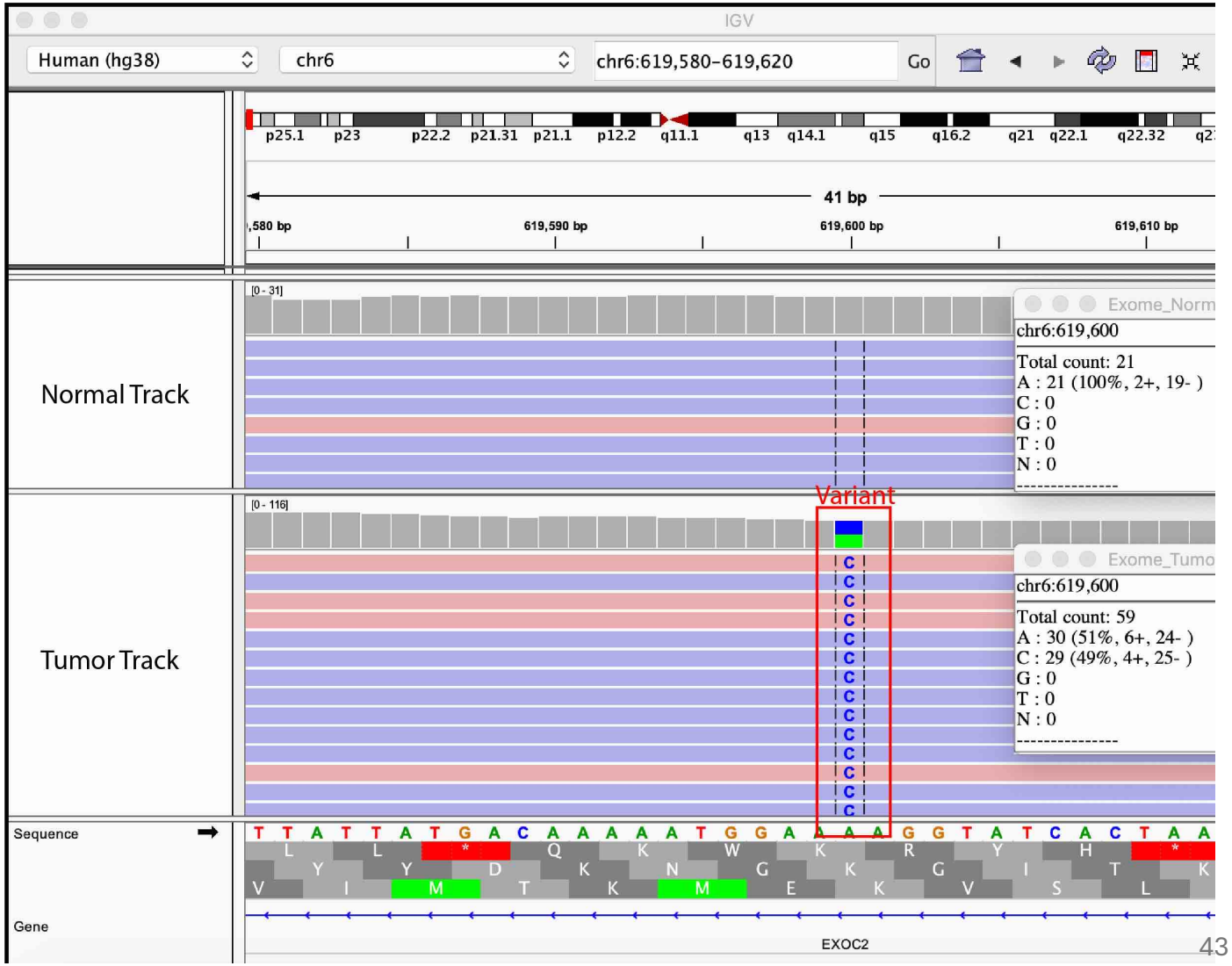
- **Low complexity**
  - Variants occur near homopolymer (TTTTTTTTTTTTT)



Koboldt. (2020). Best practices for variant calling in clinical sequencing. *Genome Medicine*, *12*. 91.

Validation: Go look at the reads yourself in IGV

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



Griffith lab, Washington University in Saint Louis: https://pmbio.org/module-05-somatic/0005/02/03/Somatic_SNV_and_Indel_Manual_Review/

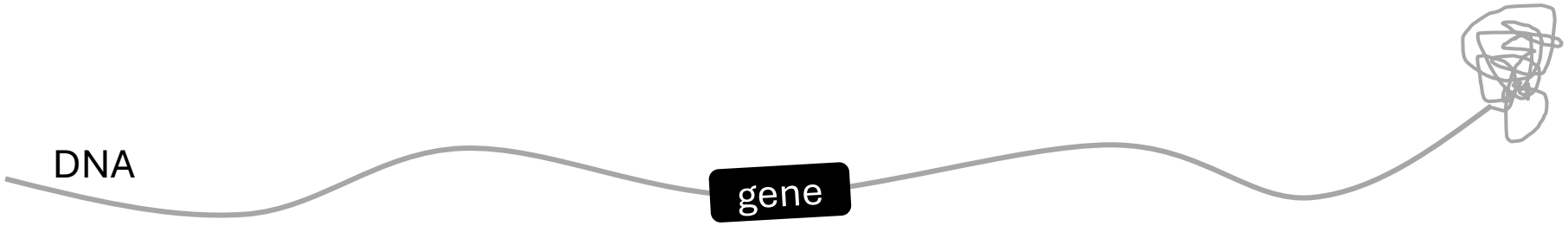
Variant annotation

- **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

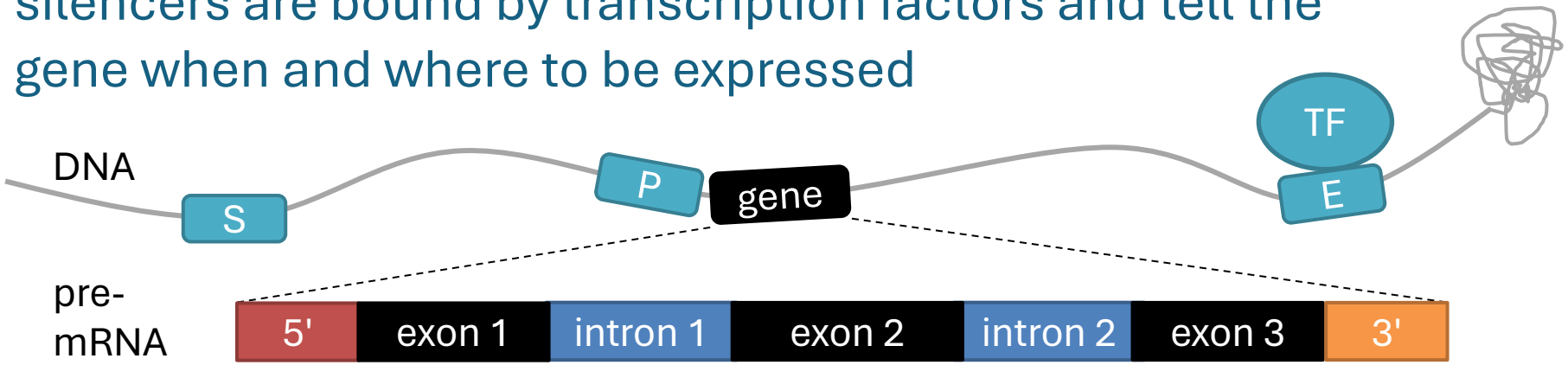
- Genes are only 1-2% of the genome
- Majority of called variants from WGS will be in noncoding regions



Variant impacts

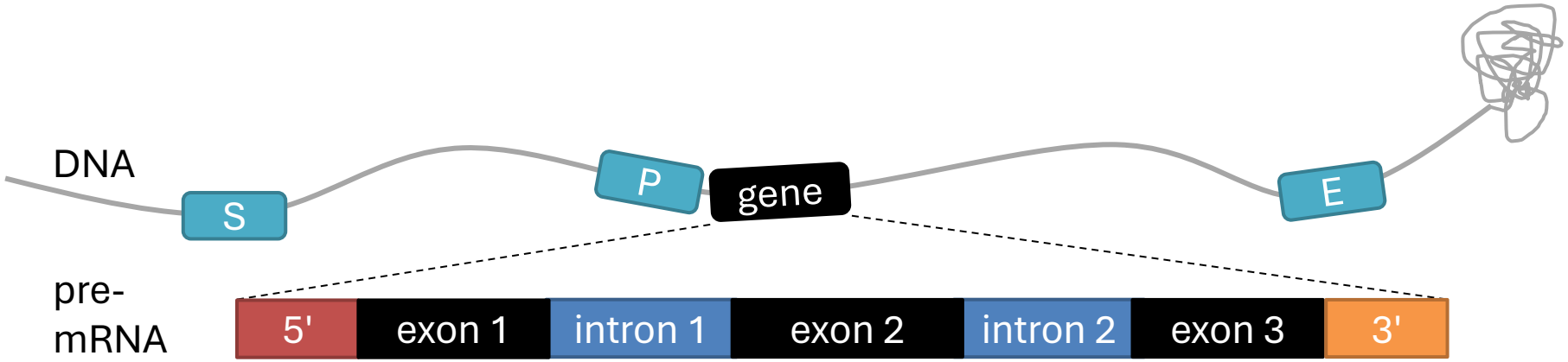
- What do noncoding regions do?

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?

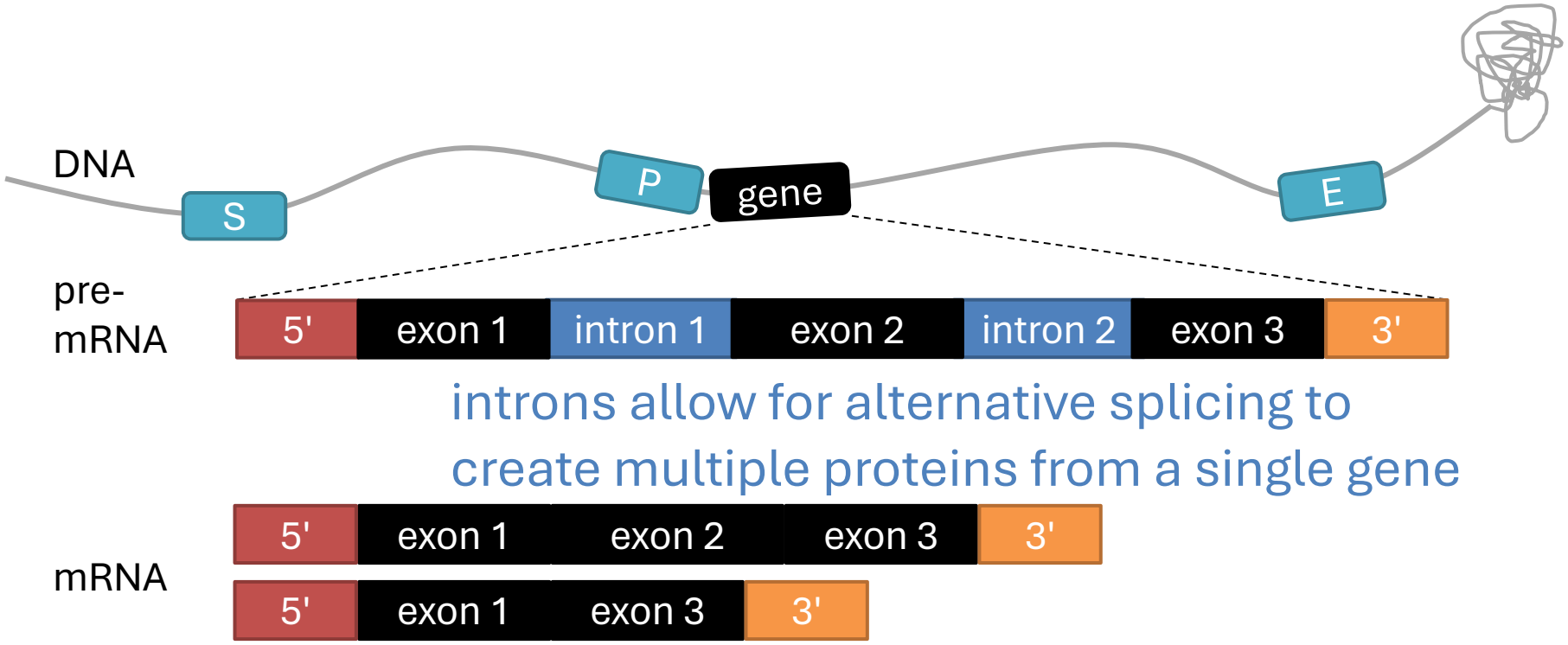


5'UTR regulates translation initiation

3'UTR regulates localization, stability, and translation

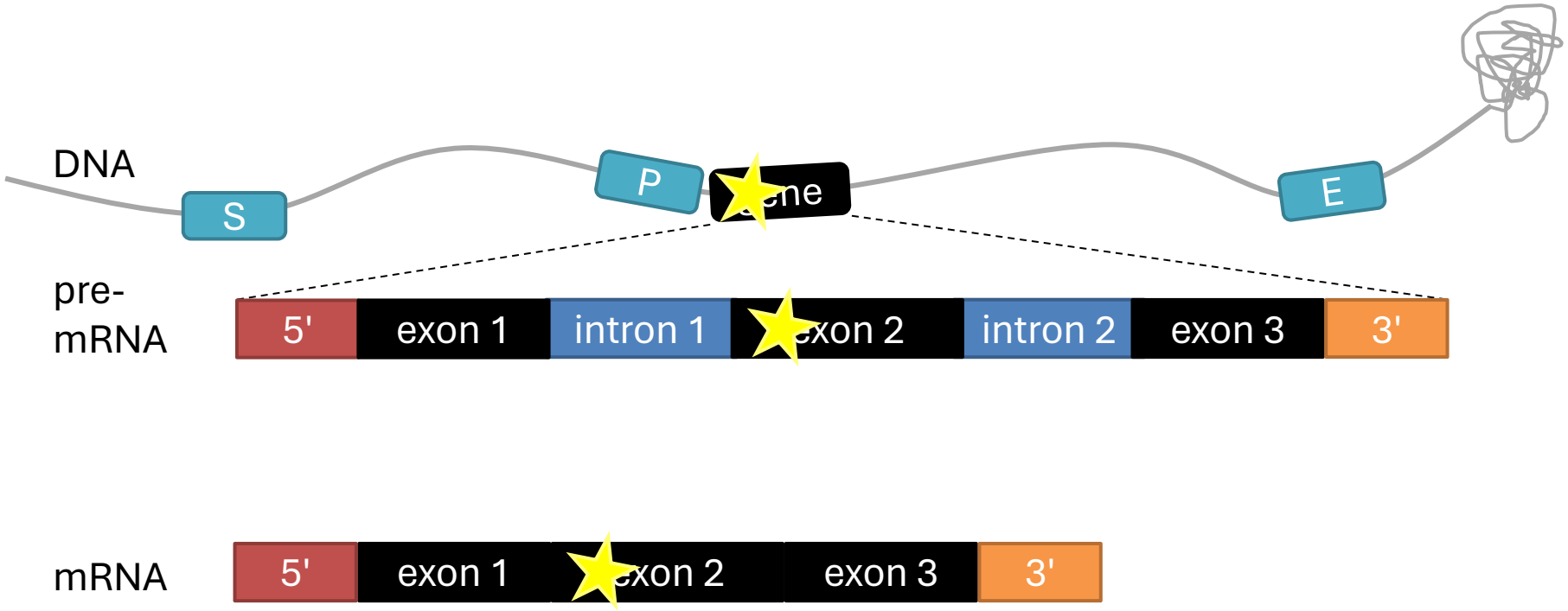
Variant impacts

- What do noncoding regions do?



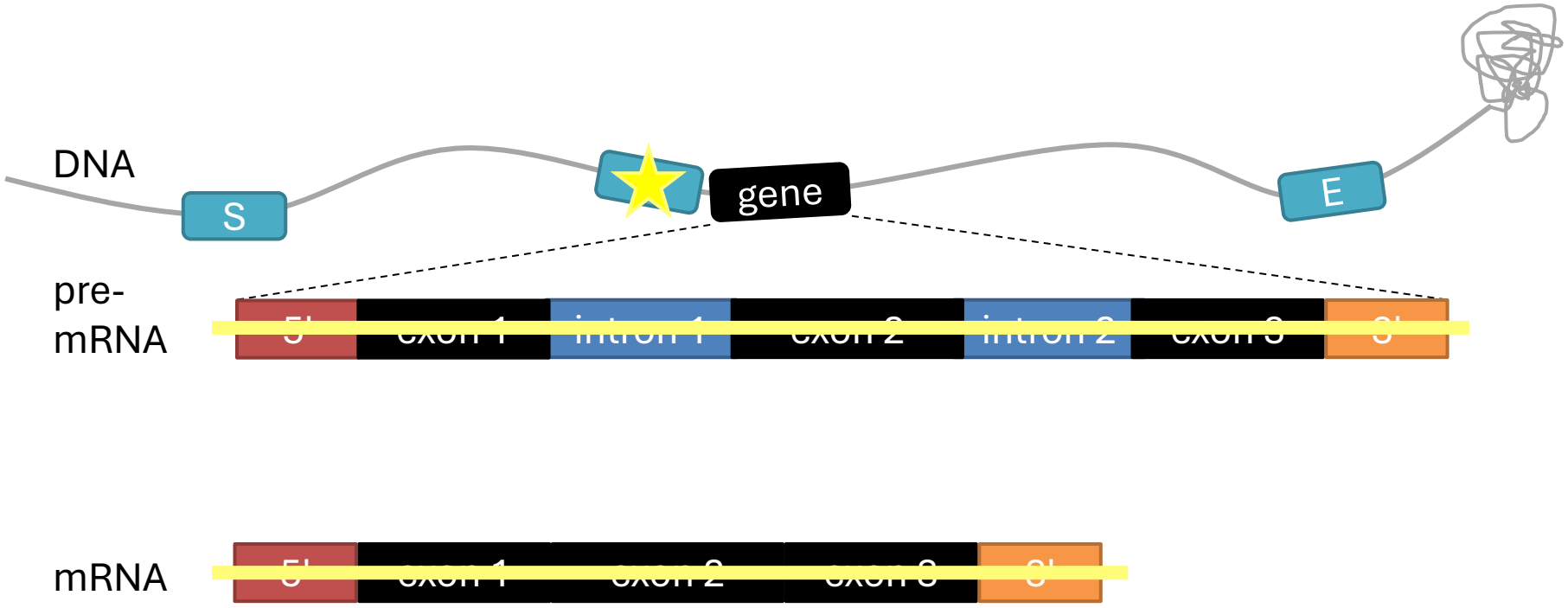
How do you predict variant function?

- Disrupt coding sequences → change protein sequence



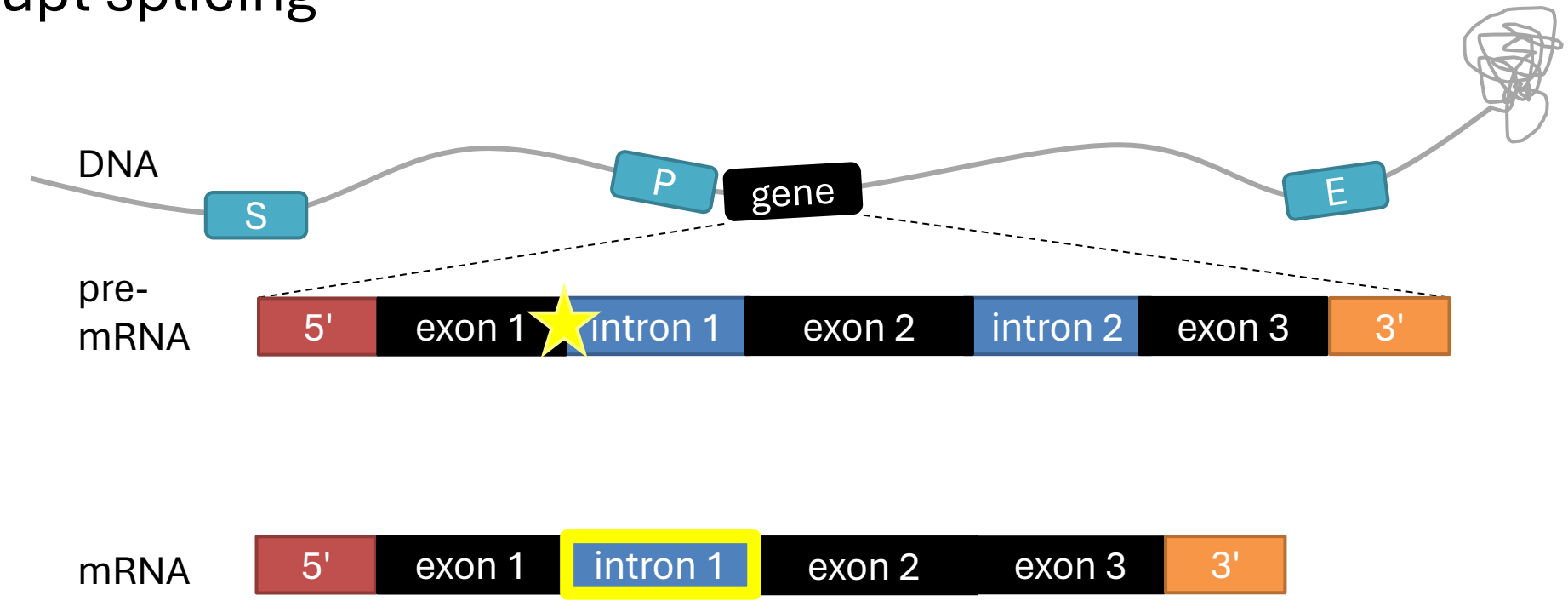
How do you predict variant function?

- 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

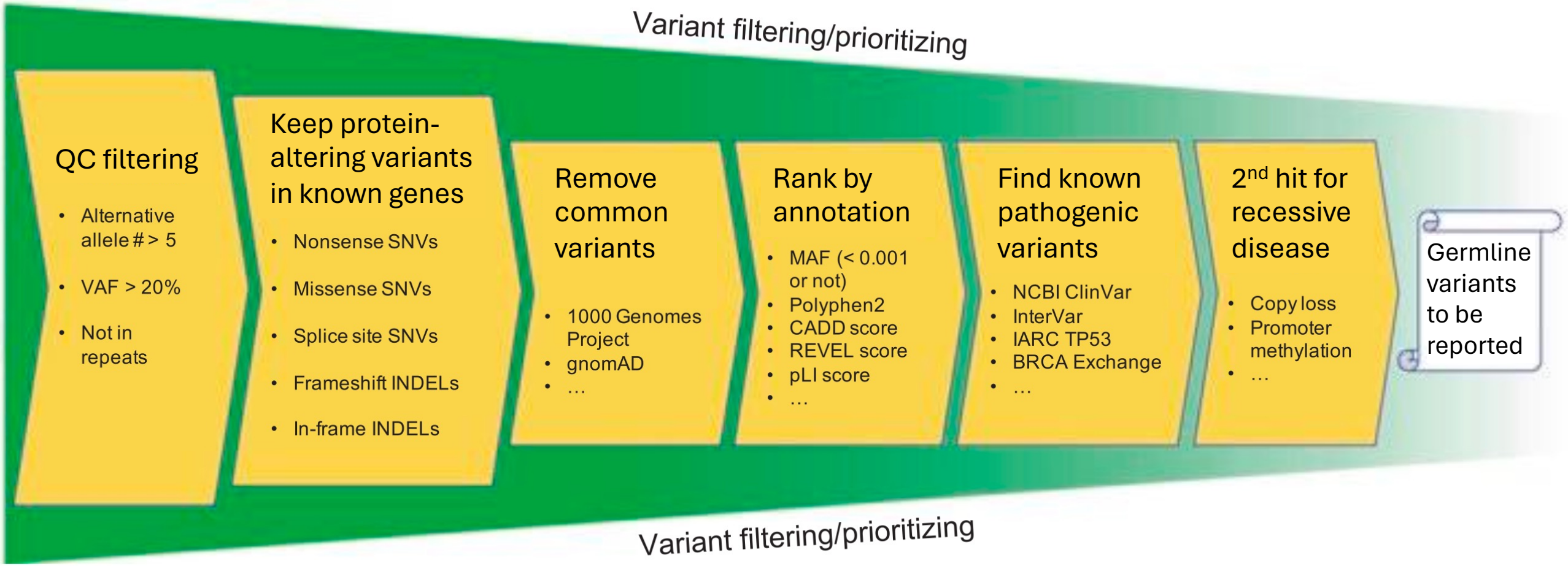
- **Effect on protein (exome)**
  - Pathogenicity prediction scores: MetaLR, MutationAssessor, MutationTaster, PolyPhen, PROVEAN, REVEL, SIFT, etc.
- **Disrupt regulatory regions**
  - Functional DNA prediction scores: CADD, DANN, EIGEN, FATHMM, GWAVA, etc.
  - Transcription factor binding sites (TFBS): 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)

Tools to add annotations

- Instead of looking up annotations one at a time, you can use…

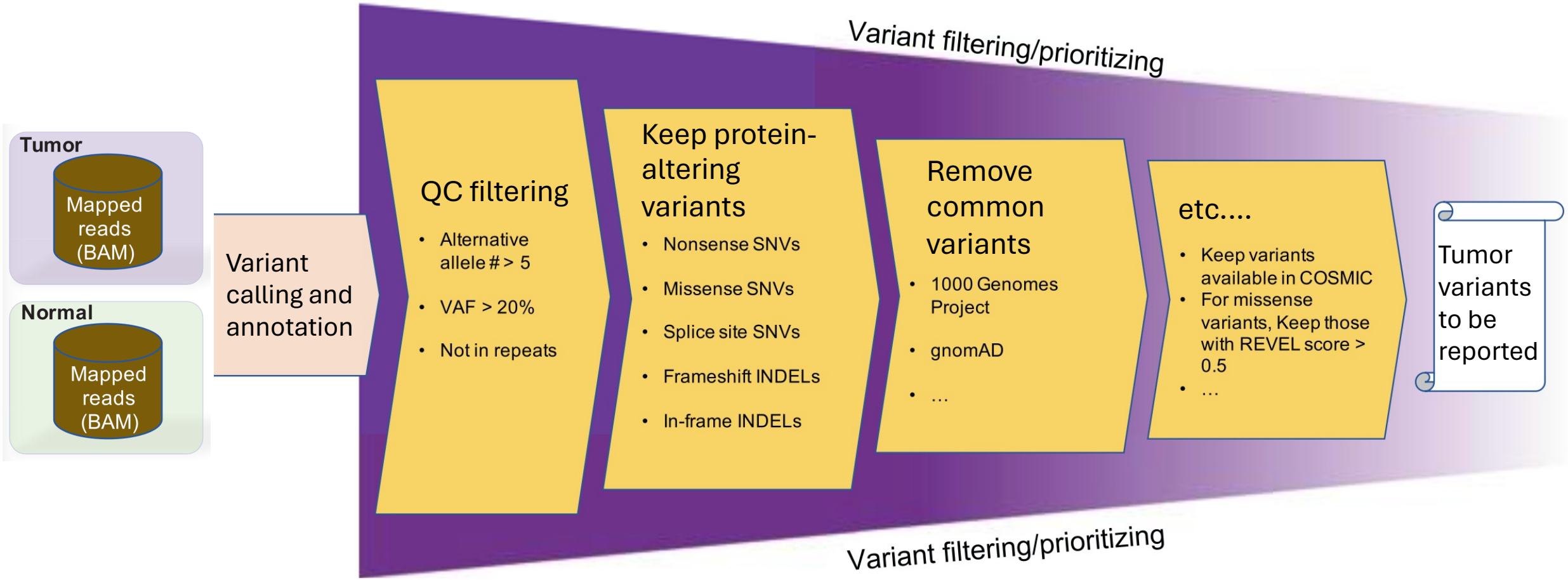
- ANNOVAR
- **SnpEff / SnpSift**
- Variant Annotation, Analysis, and Search Tool (VAAST)
- Variant Effect Predictor (VEP)
- VarSome

Variant prioritization: Germline



Adapted from Chang, et al. (2022). Somatic and germline variant calling from next-generation sequencing data. In: Laganà (ed.), *Computational Methods for Precision Oncology*, *Advances in Experimental Medicine and Biology*, *1361*, 37-54.

Variant prioritization: Tumor



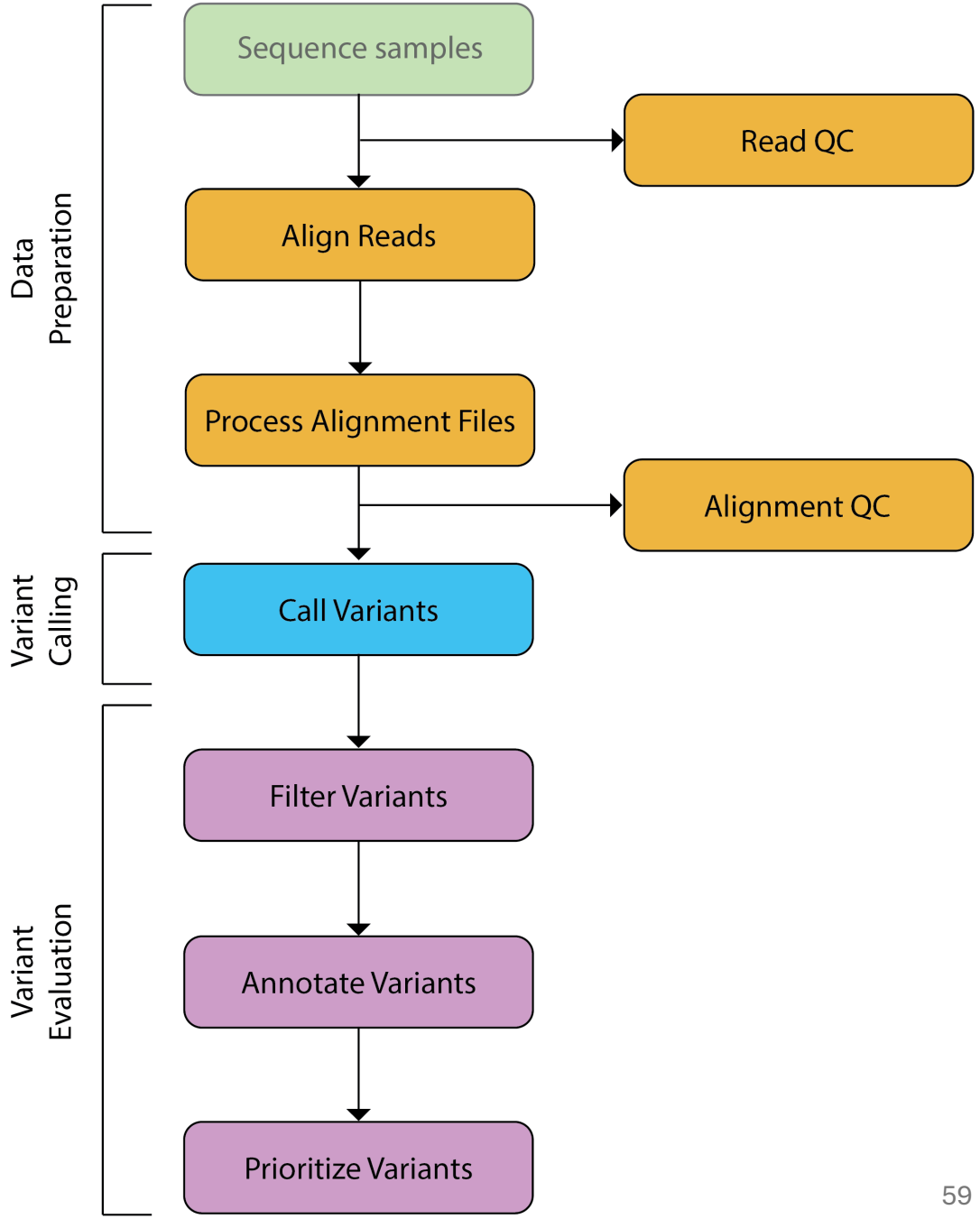
Adapted from Chang, et al. (2022). Somatic and germline variant calling from next-generation sequencing data. In: Laganà (ed.), *Computational Methods for Precision Oncology*, *Advances in Experimental Medicine and Biology*, *1361*, 37-54.

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



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

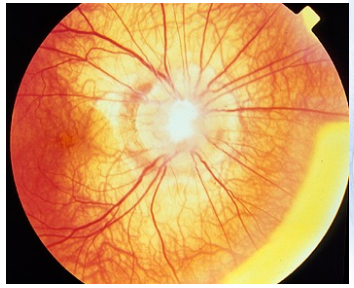
doi: https://doi.org/10.1101/2021.08.05.21261553

medRxiv
THE PREPRINT SERVER FOR HEALTH SCIENCES

PHACE syndrome: Segmental hemangioma and 1+ other feature

- **P**osterior fossa brain malformations
- **Hemangiomas: benign vascular tumors**
- **A**rterial anomalies
- **C**ardiac anomalies
- **E**ye anomalies

- Overall incidence unknown; 300+ cases
- No known familial cases
- Inheritance thought to be *de novo*



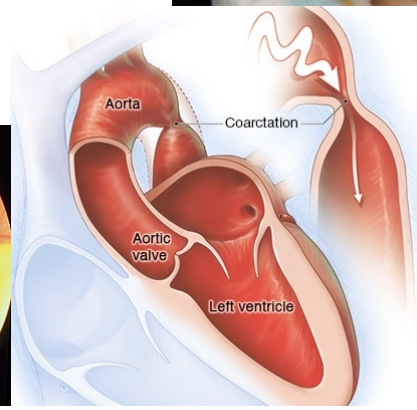


Image credits, left to right, by row: 1) Fernandes 2011, Metry 2009 2) Children's Hospital of Wisconsin, Metry 2006 3) Siegel 2007, Mayo Clinic, Metry 2009

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

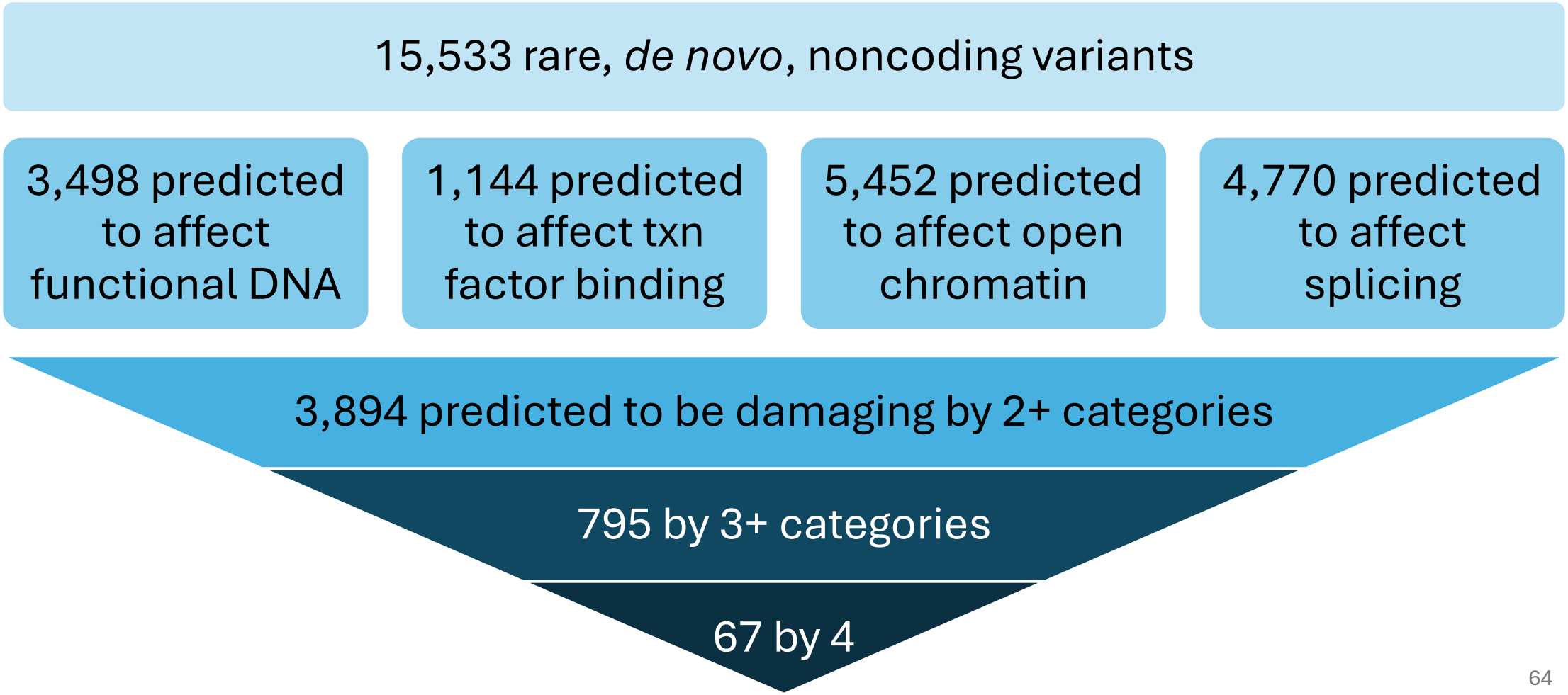
16,107 total SNVs (164 SNVs/patient)

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

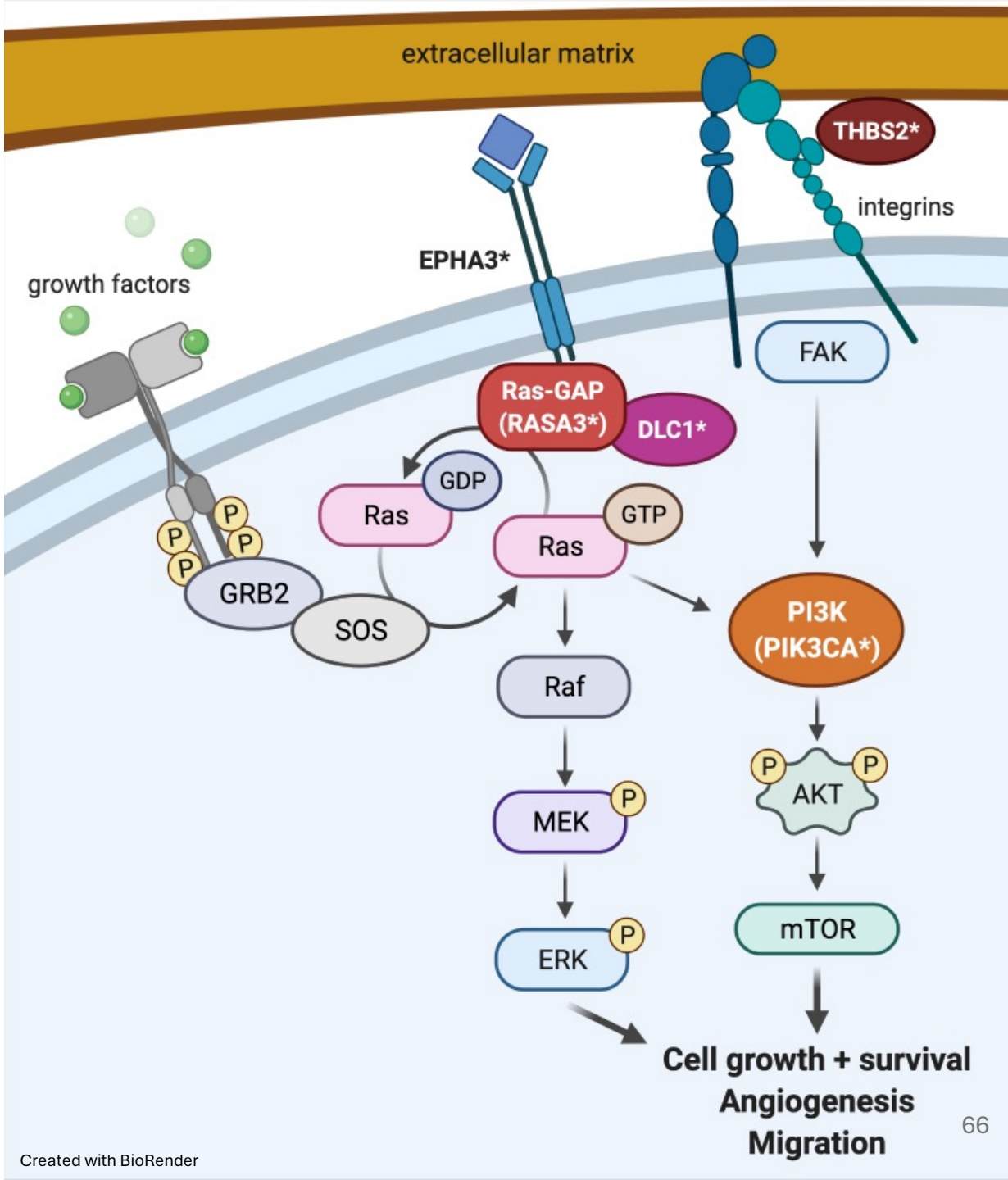


Candidate causative variant summary

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

Pathway analysis

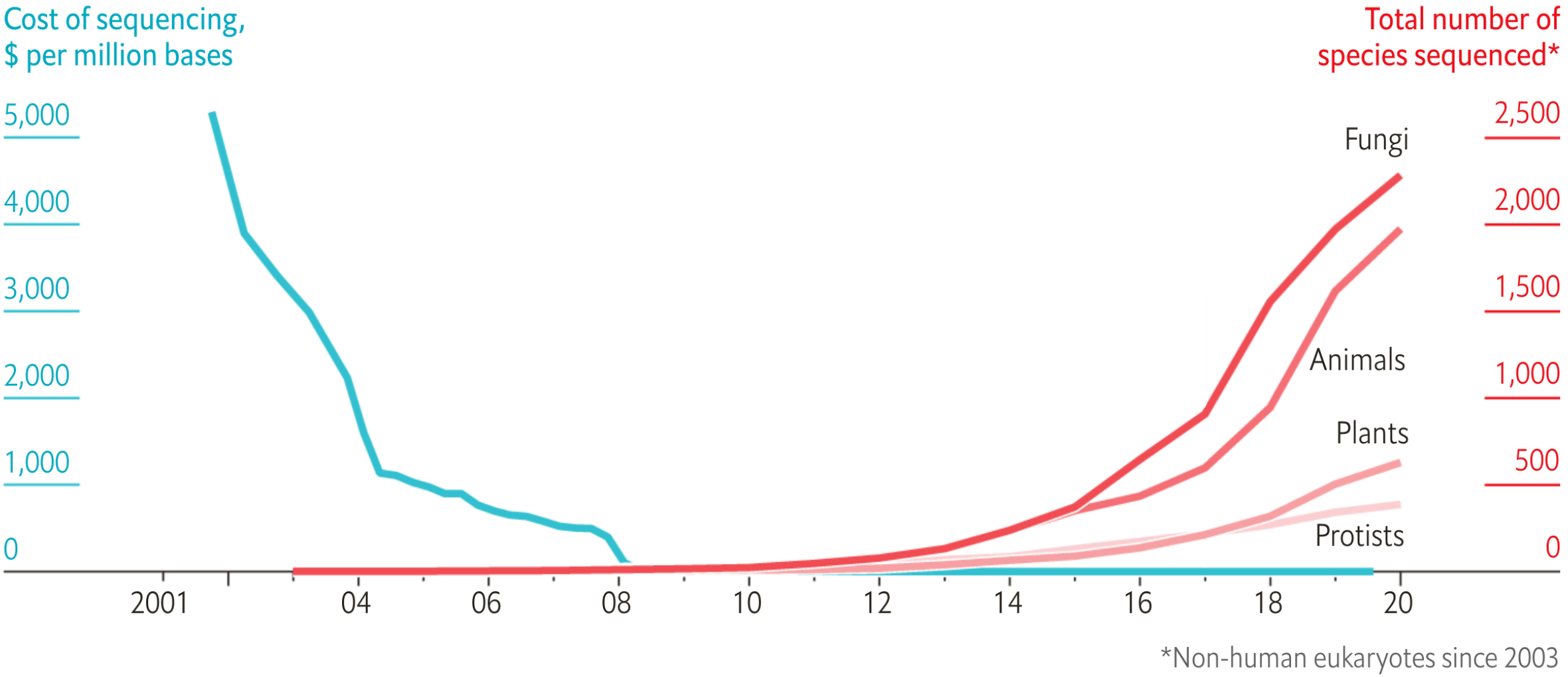
- All genes with rare, *de novo* SNVs (4,320 genes) were analyzed using g:Profiler to understand broader patterns in gene regulation
- Top KEGG pathways all affect angiogenesis signaling



Research summary: Matricellular signaling in PHACE syndrome

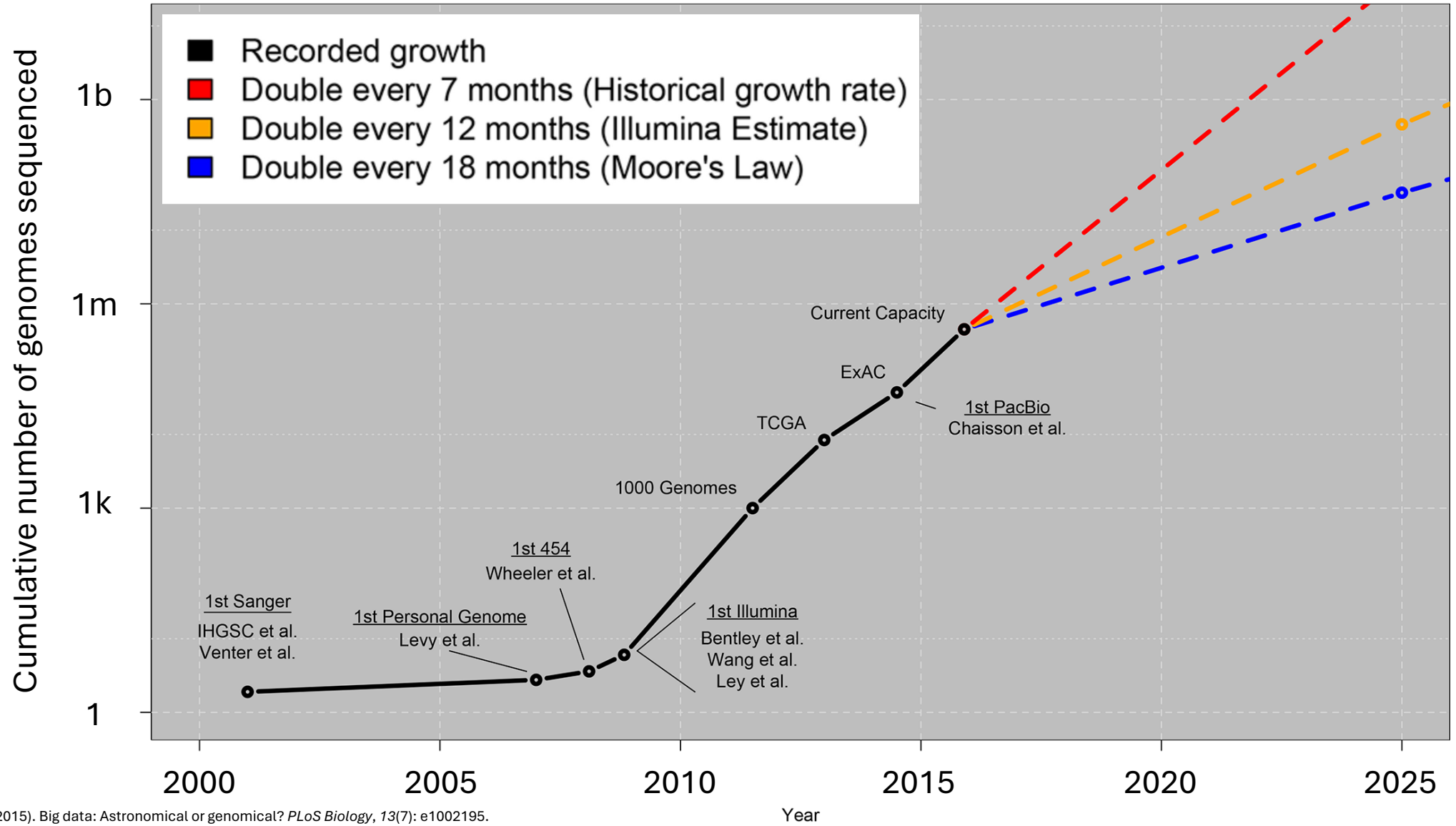
- PHACE syndrome is characterized by vascular anomalies
- I found coding and noncoding *de novo* variants in 7 genes known to cause vascular malformations in a knockout mouse model
- *In silico* evidence suggests that these variants affect protein structure, splicing, and/or regulation
- Many of these 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

Growth of genome sequencing



Adapted from https://www.economist.com/graphic-detail/2020/06/27/the-human-genome-project-transformed-biology-and-medicine

Growth of genome sequencing: Humans



Stephens, et al. (2015). Big data: Astronomical or genomical? *PLoS Biology*, *13*(7): e1002195.

Growth of genome sequencing: Direct-to-consumer

Genotyping

WES

WGS

... 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

American Society of Human Genetics (ASHG) 2023 annual meeting, “Addressing barriers to accessible genetic research and services” platform session
Shabani & Marelli. (2019). Re-identifiability of genomic data and the GDPR: Assessing the re-identifiability of genomic data in light of the EU General Data Protection Regulation. *EMBO Reports*, *20*(6), e48316.
Santos. (2008). Genetic research in native communities. *Progress in Community Health Partnerships*, *2*(4), 321-327.
Hudson, et al. (2020). Rights, interests and expectations: Indigenous perspectives on unrestricted access to genomic data. *Nature Reviews Genetics*, *21*, 377–384.

Thank you!