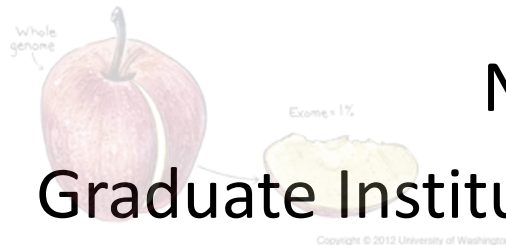
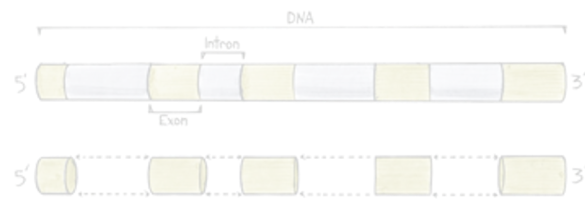


中研院生命科學圖書館【生物資訊系列】

NGS Workshop 4-遺傳變異偵測與註解



Jacob Shujui Hsu 許書睿

National Taiwan University

Graduate Institute of Medical Genomics and Proteomics

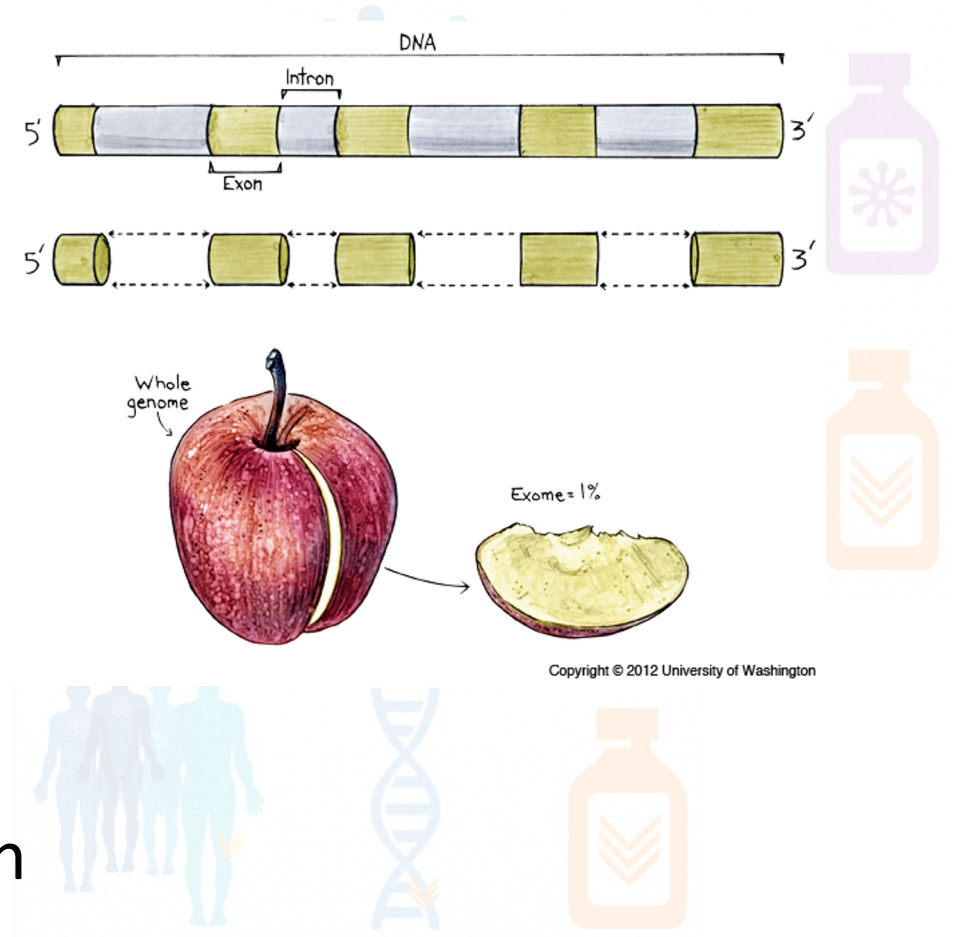
2025-11-05

Disclosures

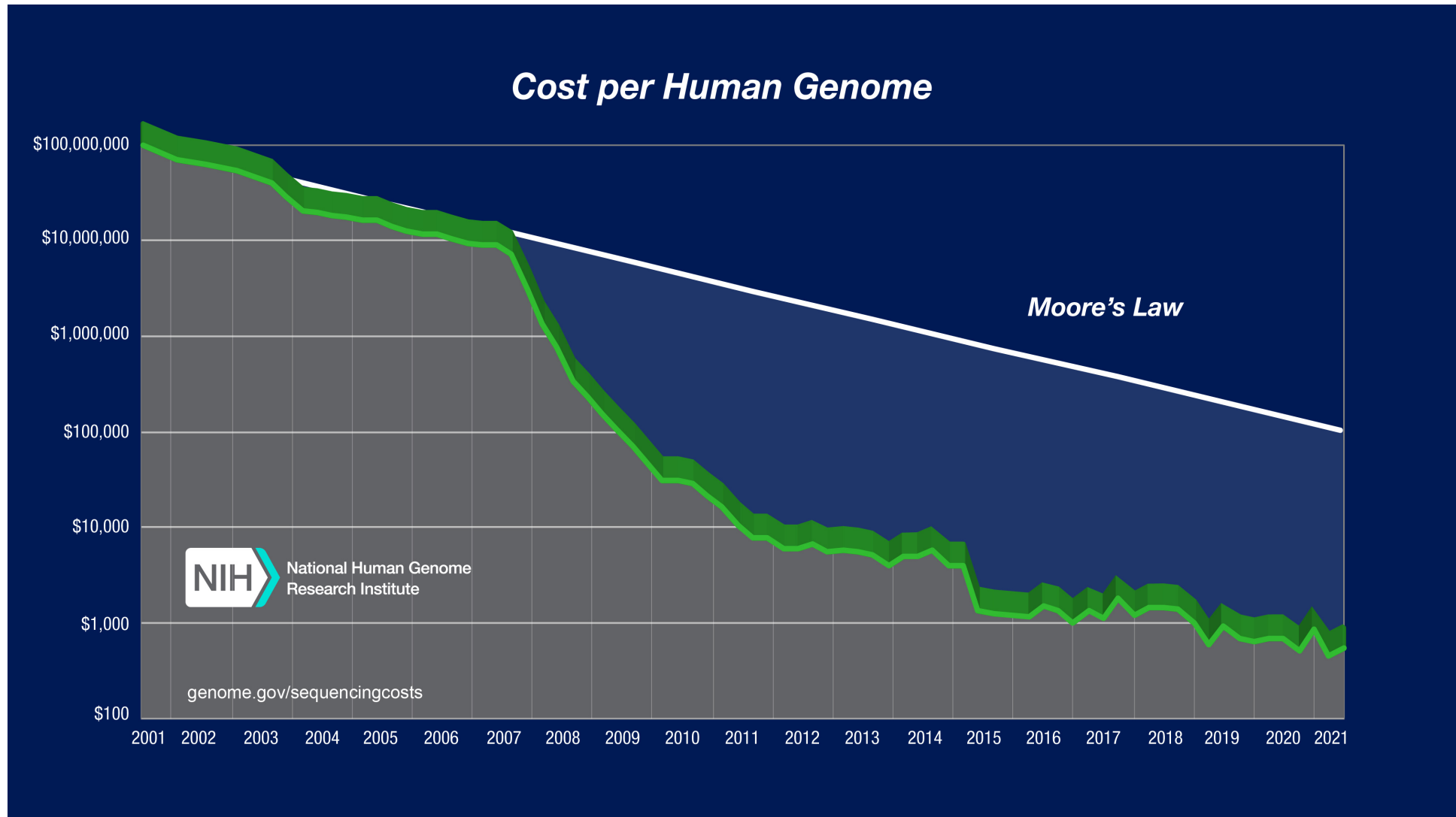
- I, Jacob Shujui Hsu, declare the following conflicts of interest in the context of the subject of this presentation:
- **Three courses**
 - NGS, bioinformatics and genomic medicine (Genom7009)
 - Principle of Statistical Genetics & Genome Data Interpretation (MOLMED7017)
 - Fantastic Genomic Biomarkers and Where to Find Them (Genom7019)
- **Industrial-academic Cooperative Project (NSTC)**
 - Establishment of an analysis pipeline for tumor tissue whole genome mutation profile (**Taiwan AI Labs**, 110-2622-B-002-012)
 - Utilizing FDA sequencing quality control materials to standardize tumor whole-genome mutation profile (**Taiwan AI Labs**, 111-2622-B-002-012, 112-2622-B-002 -012)
 - Benchmarking Long-Read Whole Genome Sequencing Analysis Workflow to Improve Molecular Diagnosis of Genetic Diseases (**Blossom Bio**, 114-2622-B-002 -009)

Learning Outcomes

- Human Genome Reference(s)
- Variant identification
 - Single Nucleotide Variant (SNV)
 - Small Insertion/Deletion (INDEL)
 - Structure Variant (SV)
 - Others
- Variant Annotation(s)
 - Functional Consequence
 - Predicted Effect
 - Pathogenicity & Clinical implications
- State of the art human genome research



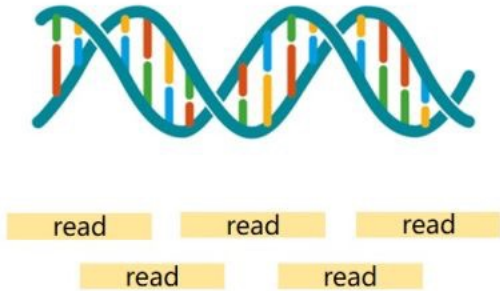
46,425,255 v.s. 30.73



Next Generation Sequencing (NGS)

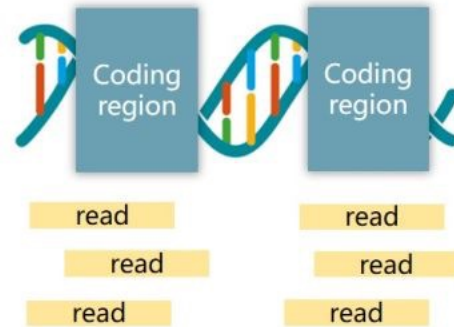
WGS

Whole genome sequencing



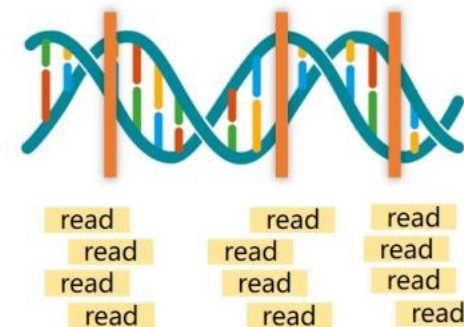
- **Sequencing region:** whole genome
- **Sequencing depth:** >30X
- **Sequencing feature:** covering everything, identify all kinds of variants including SNP, INDELs, CNV and SV.

Whole exome sequencing



- **Sequencing region:** whole exome
- **Sequencing depth:** >50X~100X
- **Sequencing feature:** identify all kinds of variants including SNP, INDELs and SV in coding region.

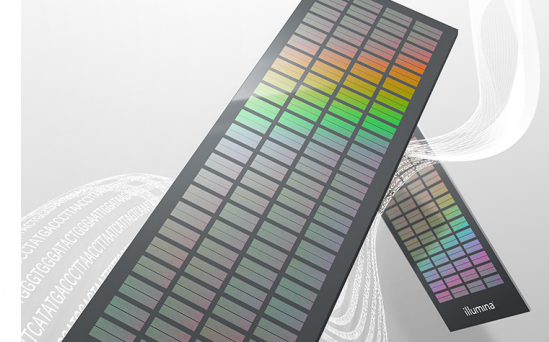
Targeted sequencing



- **Sequencing region:** customized regions
- **Sequencing depth:** >500X
- **Sequencing feature:** identify all kinds of variants including SNP, INDELs, CNV and SV in specific regions.

TPMI

Genotyping SNP array



- **Genotyping region:** A locus in a linkage block
- **Depth:** NA
- **Features:** phenotype-associated linkage blocks
- **Some known pathogenic variants**

WGS cost v.s. value



New

iPhone 17 Pro

Innovative design for ultimate performance and battery life.

From \$1099 or \$45.79/mo. for 24 mo.¹⁰

[Learn more](#)

[Buy >](#)



New

iPhone Air

The thinnest iPhone ever.
With the power of pro inside.

From \$999 or \$41.62/mo. for 24 mo.¹⁰

[Learn more](#)

[Buy >](#)



New

iPhone 17

Even more delightful.
Even more durable.

From \$799 or \$33.29/mo. for 24 mo.¹⁰

[Learn more](#)

[Buy >](#)



iPhone 16

Amazing performance.
Durable design.

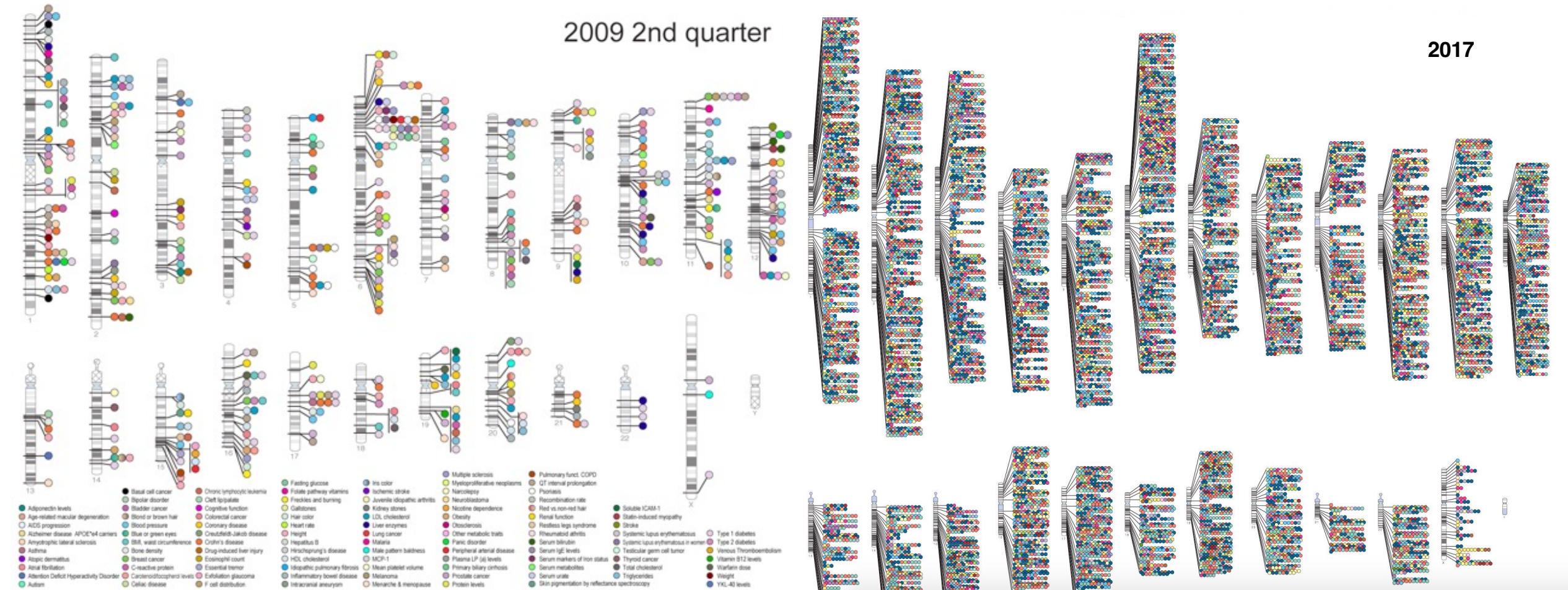
From \$699 or \$29.12/mo. for 24 mo.¹⁰

[Learn more](#)

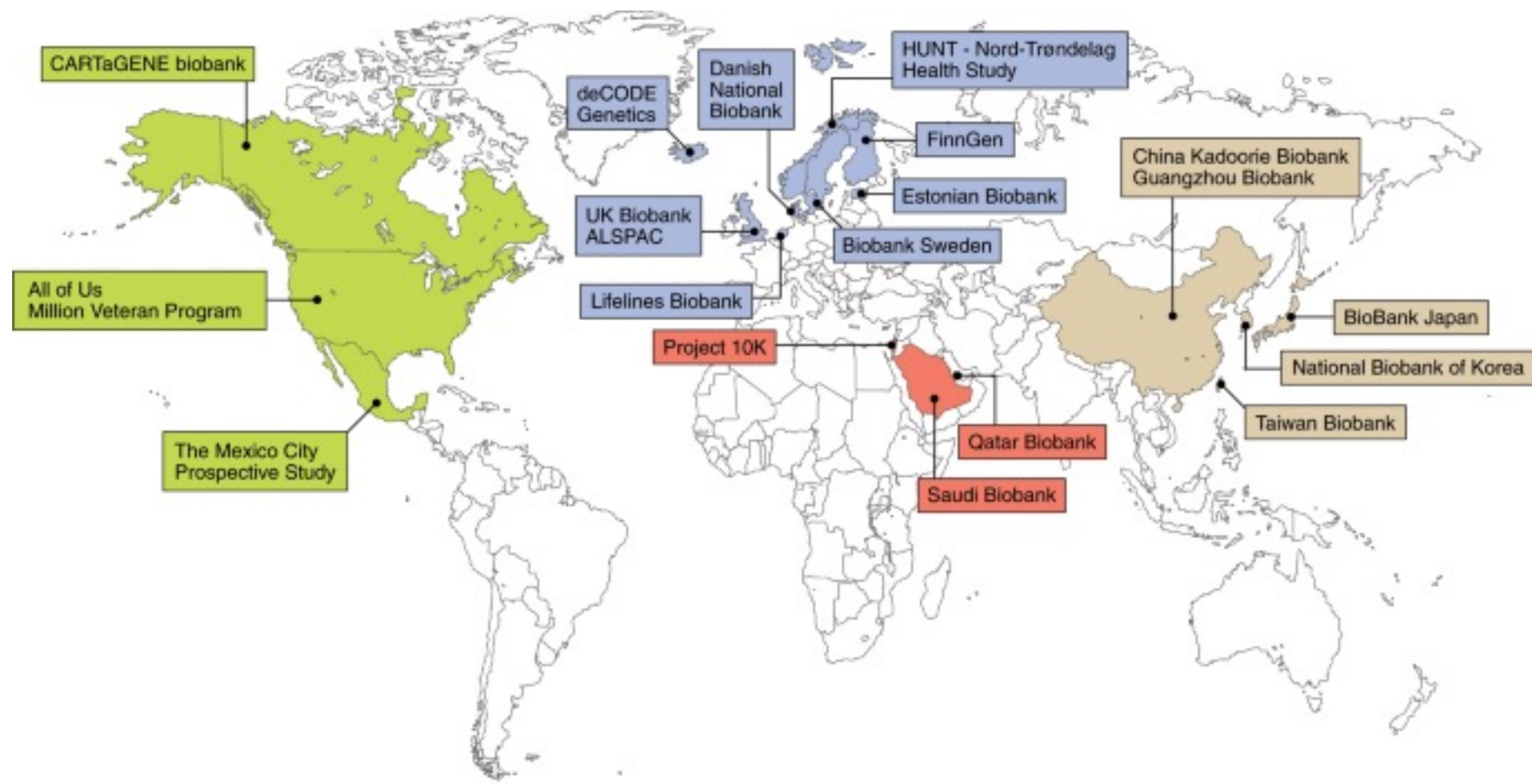
[Buy >](#)



Genome-wide association study (GWAS)



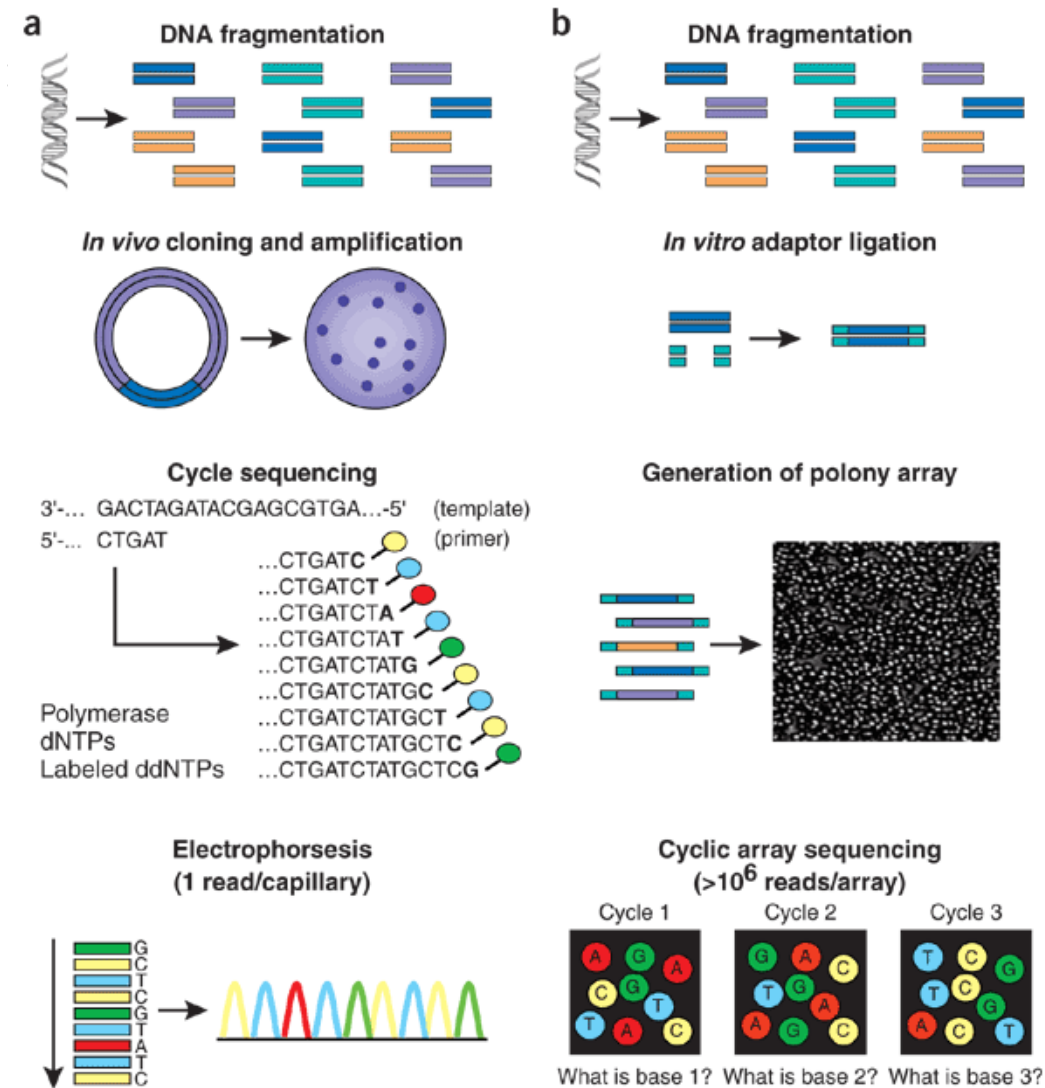
The world's most valuable resource is no longer oil, but data (The Economist, 2017)



Location	Biobank	N (goal)
Canada	CARTaGENE biobank ¹¹⁹	43,000
USA	All of Us ³³ Million Veteran Program ⁴⁹	1,000,000 > 600,000
Mexico	The Mexico City Prospective Study ⁵²	150,000
Iceland	deCODE Genetics	500,000
UK	UK Biobank ³⁸ Avon Longitudinal Study of Parents and Children (ALSPAC) ²⁰	500,000 > 15,000
Netherlands	Lifelines Biobank ¹²⁰	> 167,000
Denmark	Danish National Biobank ¹²¹	
Norway	HUNT - Nord-Trøndelag Health Study ¹²²	125,000
Sweden	Biobank Sweden	
Finland	FinnGen	500,000
Estonia	Estonian Biobank ¹²³	52,000
Israel	Project 10K	10,000
Saudi Arabia	Saudi Biobank	200,000
Qatar	Qatar Biobank ¹²⁴	60,000
China	China Kadoorie Biobank ⁵¹ Guangzhou Biobank ¹²⁵	> 500,000 30,000
Japan	BioBank Japan ¹²⁶	200,000
Korea	National Biobank of Korea ¹²⁷	500,000
Taiwan	Taiwan Biobank ¹²⁸	200,000

Next Generation Sequencing (NGS)

- **Sanger:** DNA is fragmented, then cloned to a plasmid vector
- High-resolution electrophoretic separation within one of **96 or 384** capillaries in one run of a sequencing instrument.
- **NGS**, adaptors are ligated to fragmented genomic DNA, which turns to an array of **millions** of spatially immobilized **PCR colonies**
- A single **microliter-scale** reagent volume can be applied **in parallel**.



a. PacBio SMRT sequencing

Template topology

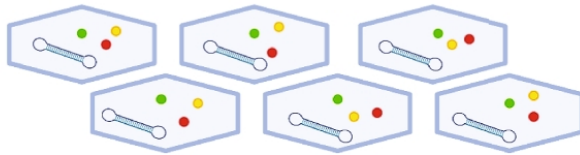
SMRT bell template



Polymerase

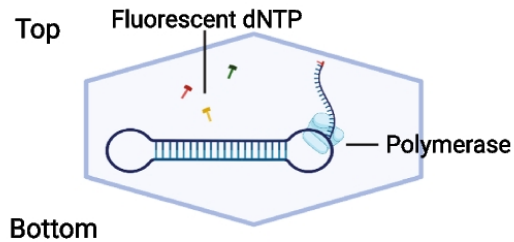
SMRTbell library preparation: Hairpin adapters are ligated to both of the ends of the template DNA which creates a single-stranded circular DNA

Flow cell top view



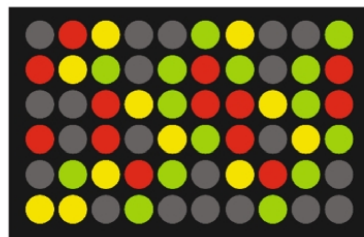
Several ZMW cells with immobilized DNA and polymerase at their base

Single Zero Mode Waveguide (ZMW)

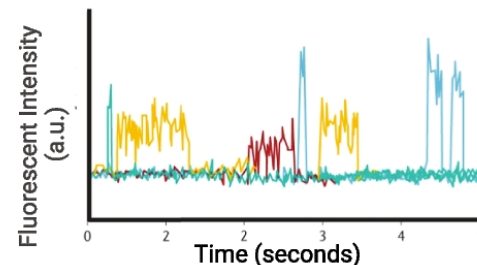


Sequencing involves the addition of differentially labelled dNTP. The fluorescence of each of the different ZMW is then recorded.

Readout

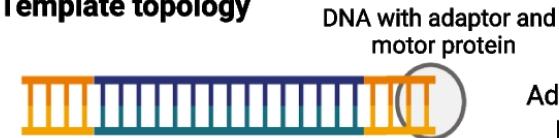


SMRT output is the fluorescence pattern in the ZMWs



b. ONT sequencing

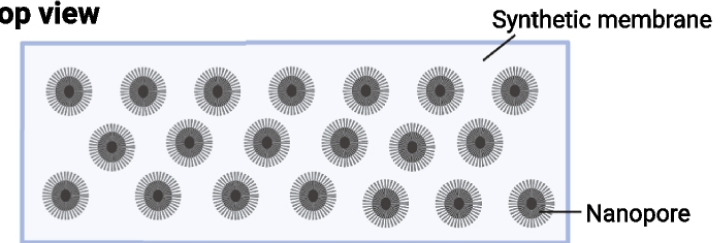
Template topology



DNA with adaptor and motor protein

DNA template: Adapters tagged with motor protein are ligated onto polynucleotides

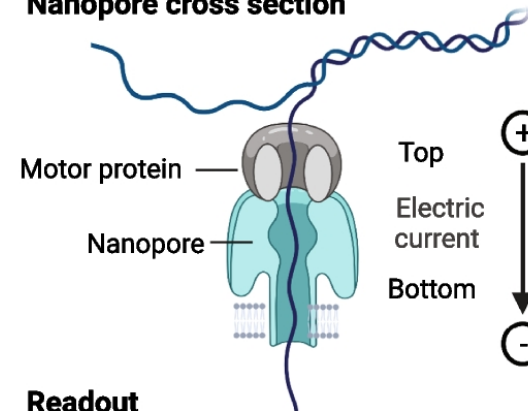
Flow cell top view



Synthetic membrane

Nanopore

Nanopore cross section



Motor protein

Nanopore

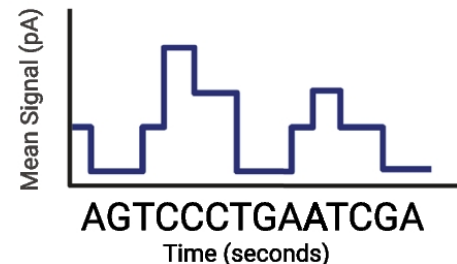
Top

Electric current

Bottom

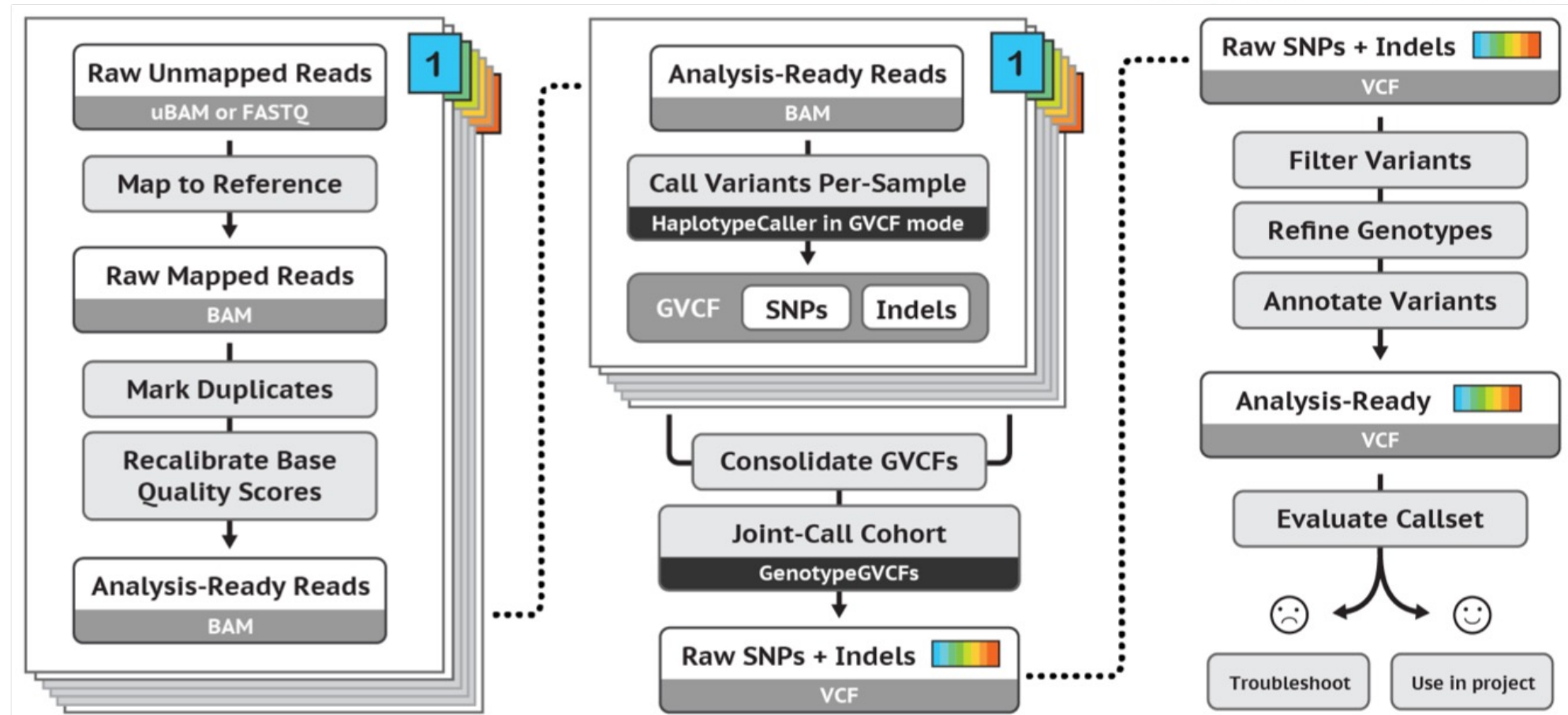
Sequencing is initiated by the adapter being guided to an available pore. The motor protein then attaches to the pore, where it enzymatically separates the strands into template and complementary strands.

Readout



ONT output is a change in the current across the synthetic membrane which corresponds to a particular k-mer that can be used to deduce the sequence

Germline short variant discovery – Which human genome reference should be used?



Whose DNA Makes Up the Reference Genome?

WANTED
20 Volunteers
to participate in the
Human Genome Project
a very large international scientific research effort.

The goal is to decode the human hereditary information (*human blueprints*) that determines all individual traits inherited from parents. The outcome of the project will have tremendous impact on future progress of medical science and lead to improved diagnosis and treatment of hereditary diseases.

Volunteers will receive information about the project from the Clinical Genetics Service at Roswell Park, and sign a consent form before participating.

No personal information will be maintained or transferred.

Volunteers will provide a one-time donation of a small blood specimen. A small monetary reimbursement will be provided to the participants for their time and effort.

Individuals must be at least 18 years of age.
Persons who have undergone chemotherapy are not eligible.

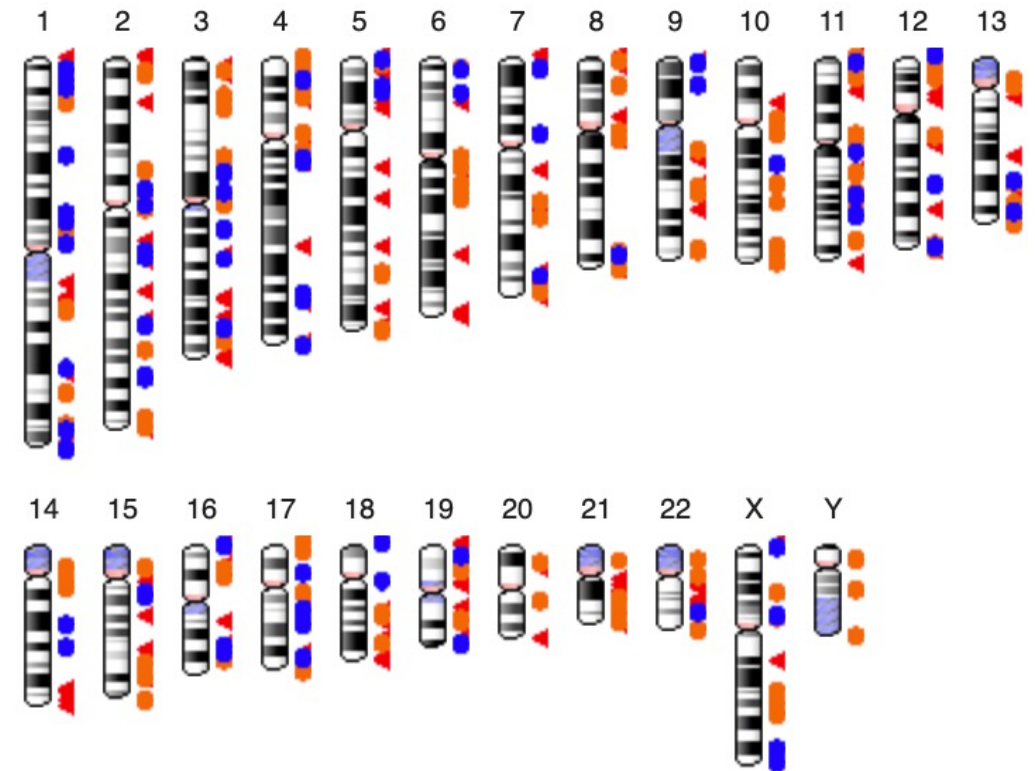
ROSWELL PARK
CANCER INSTITUTE

For more information please contact the
Clinical Genetics Service
845-5720 (9:00 am - 3:00 pm)
March 24 - 26, 1997

- Blood (female) and sperm (male) samples from many volunteers in Buffalo, NY
- >70% of the reference genome produced by the public HGP came from a **single anonymous male donor** from Buffalo, NY
- Human genome size ($\sim 3.23 \times 10^9$ bp)
- WGS data size (30X, 200Gb)
- Genetic Variations/Variants

Genome Reference Consortium

- Main Builds (**major**)
 - GRCh37 (Feb 27, 2009)
 - GRCh38 (May 07, 2014)
- Patches (**minor**)
 - GRCh37.p13 (June 28, 2013)
 - GRCh38.p13 (March 01, 2019)
 - GRCh38.p14 (**February 03, 2022**)
- Contigs
- GRch39? Telemore-to-Telomere?
- Genome graphs?
- Personal Genome Assembly

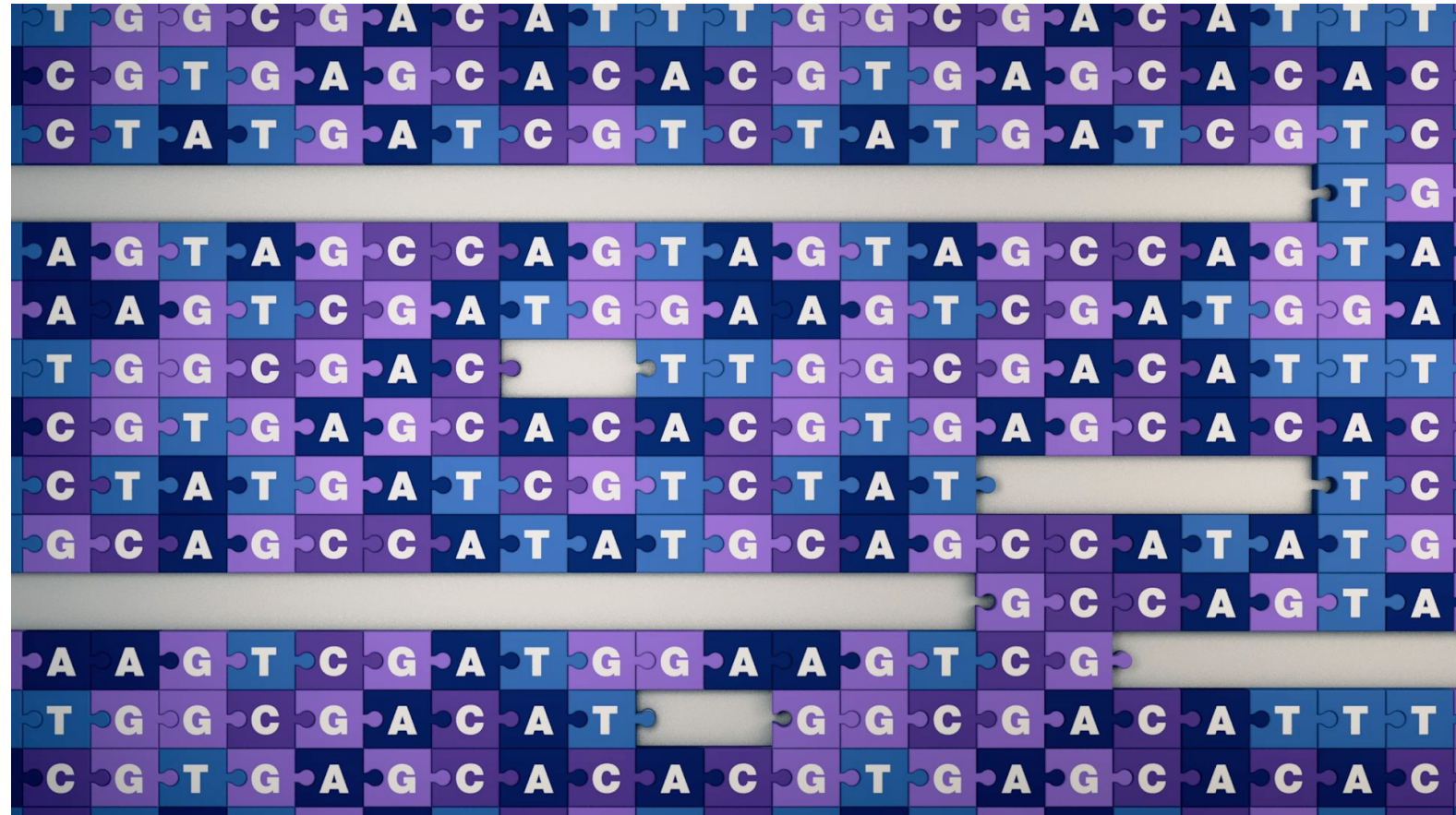


- ◀ Region containing alternate loci
- Region containing fix patches
- Region containing novel patches

Ideogram of the latest human assembly, GRCh38.p14

Complete human genome ???

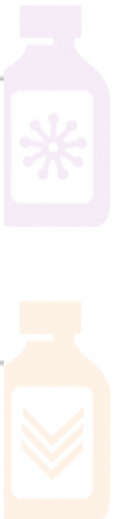
On **March 31, 2022**, the Telomere-to-Telomere (T2T) consortium announced that had filled in the remaining gaps and produced the first truly complete human genome sequence.



Flavor	Source	Name	Unplaced contigs	Unlocalized contigs	Alternate loci	 mitochondria	 Epstein-Barr virus	 decoy sequences	Remarks
GRCH		GRCh37	No canonical name	No canonical name	No canonical name	Maintained by Mitomap, distributed for convenience	✗	✗	
UCSC	GRCh37	hg19	chrUn_gl000212	chr1_gl000191_random	chr6_apd_hap1	NC_001807 (from build 36)	✗	✗	Chromosome names start by "chr" PAR regions on chrY are hard masked
Ensembl	GRCh37.p13	Ensembl API release 75 Homo_sapiens.GRCh37.75.dna.primary_assembly.fasta.gz	GL000211.1	GL000191.1	✗	NC_012920.1 Revised Cambridge Reference Sequence (rCRS)	✗	✗	Chromosome named "1" to "22", "X", "Y" and "MT"
1000 genomes project phase I & III	GRCh37.p2	hs37 g1k_v37 b37 human_g1k_v37.fasta.gz	GL000211.1	GL000191.1	✗	NC_012920.1 Revised Cambridge Reference Sequence (rCRS)	✗	✗	"1" to "22", "X", "Y" and "MT"



Flavor	Source	Name	Unplaced contigs	Unlocalized contigs	Alternate loci	 mitochondria	 Epstein-Barr virus	 decoy sequences	Remarks
1000 genomes project phase II	GRCh37. p4	hs37d5 b37+decoy +herpes hs37d5.fa.gz	GL000211.1	GL000191.1	✗	NC_012920.1 Revised Cambridge Reference Sequence (rCRS)	NC_00 7605	hs37d5 SS	pseudo-autosomal regions are hard-marked on Y chromosome
Illumina MiSeq Reporter + BSO	hg19	hg19	✗	✗	✗	NC_001807 (from build 36)	✗	✗	hg19 without unplaced/unlocaliz ed contigs nor alternate loci
Ion Torrent	hg19	hg19	✗	✗	✗	NC_012920.1 Revised Cambridge Reference Sequence (rCRS)	✗	✗	hg19 without unplaced/unlocaliz ed contigs nor alternate loci
GATK Bundle	GRCh37. p2	b37 + decoy	GL000211.1	GL000191.1	✗	NC_012920.1 Revised Cambridge Reference Sequence (rCRS)	✗	hs37d5 SS	"1" to "22", "X", "Y" and "MT"



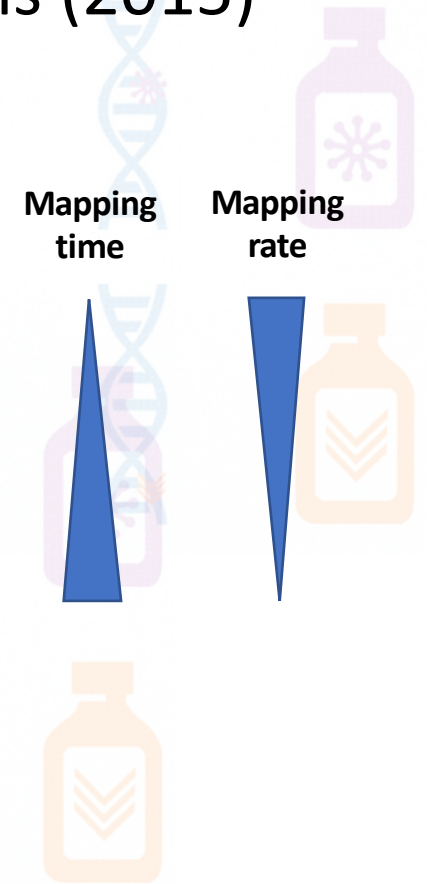
Human Genome Reference(s)

GATK reference type	B37	HG19	GRCh37(GRC)
(.fasta) file/MD5	human_g1k_v37 human_g1k_v37_ decoy	ucsc.hg19	
1b22b98cdeb4a9304cb5d48026a85128	1	chr1	chr1
a0d9851da00400dec1098a9255ac712e	2	chr2	chr2
23dccd106897542ad87d2765d28a19a1	4	chr4	chr4
1fa3474750af0948bdf97d5a0ee52e51	Y	----	----
6743bd63b3ff2b5b8985d8933c53290a	NC_007605	----	----
c68f52674c9fb33aef52dcf399755519	MT	----	chrM
fdfd811849cc2fadebc929bb925902e5	3	----	--
1e86411d73e6f00a10590f976be01623	----	chrY	chrY
641e4338fa8d52a5b781bd2a2c08d3c3	----	chr3	chr3
d2ed829b8a1628d16cbeee88e88e39eb	----	chrM	----

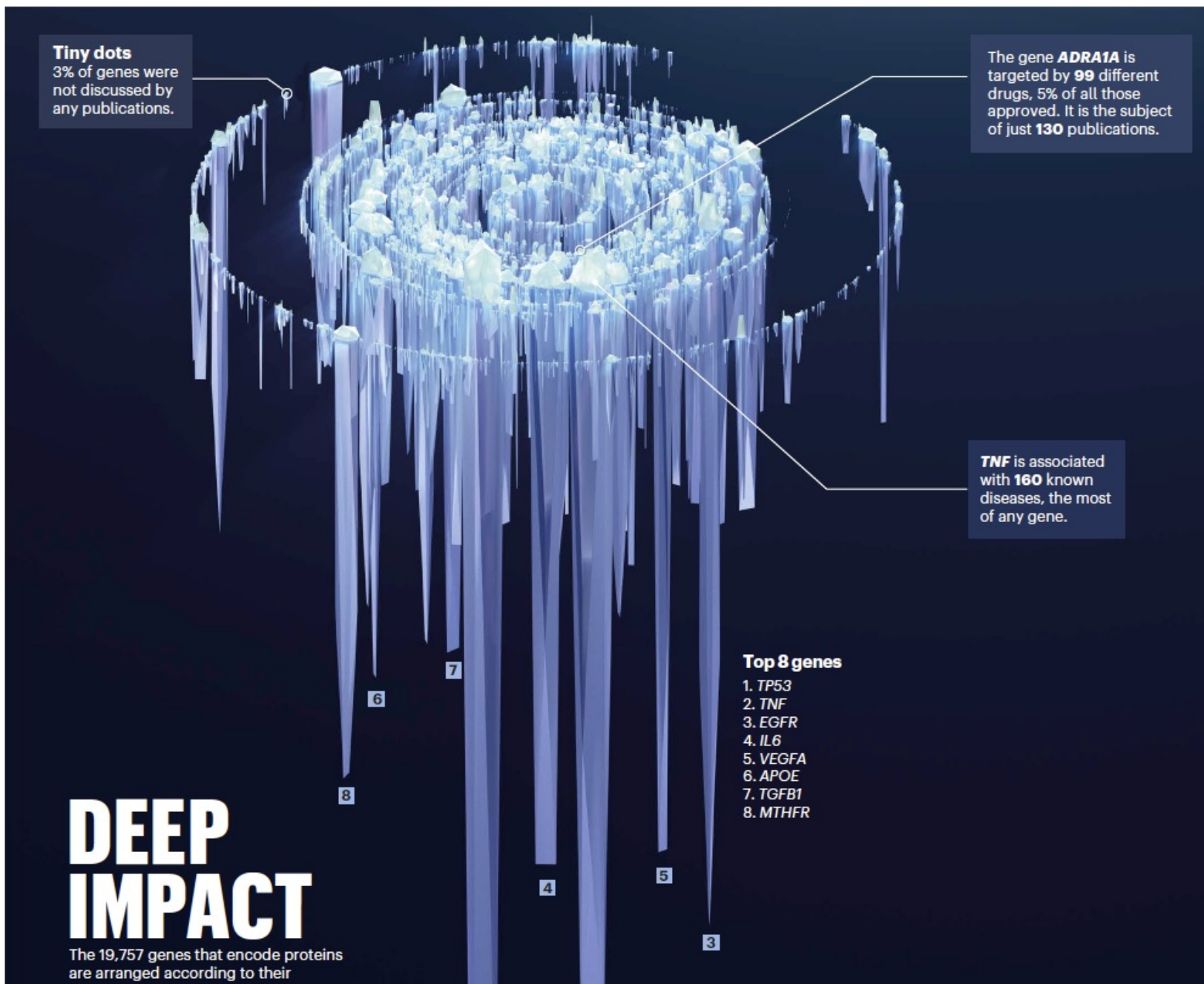
Mapping rate & Contigs

- A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing, Nature Communications (2015)

	Reads (control + tumor)			Minutes
	Total	Aligned	Uniquely aligned	CPU time
b37+d	2152793590 (100.0 %)	2112982704 (98.15 %)	2018366589 (93.76 %)	79348.8 (100.0 %)
b37	2152793590 (100.0 %)	2071267988 (96.21 %)	1996504352 (92.74 %)	92517.7 (116.6 %)
hg19r	2152793590 (100.0 %)	2058694644 (95.63 %)	1990780196 (92.47 %)	95894.3 (120.85 %)

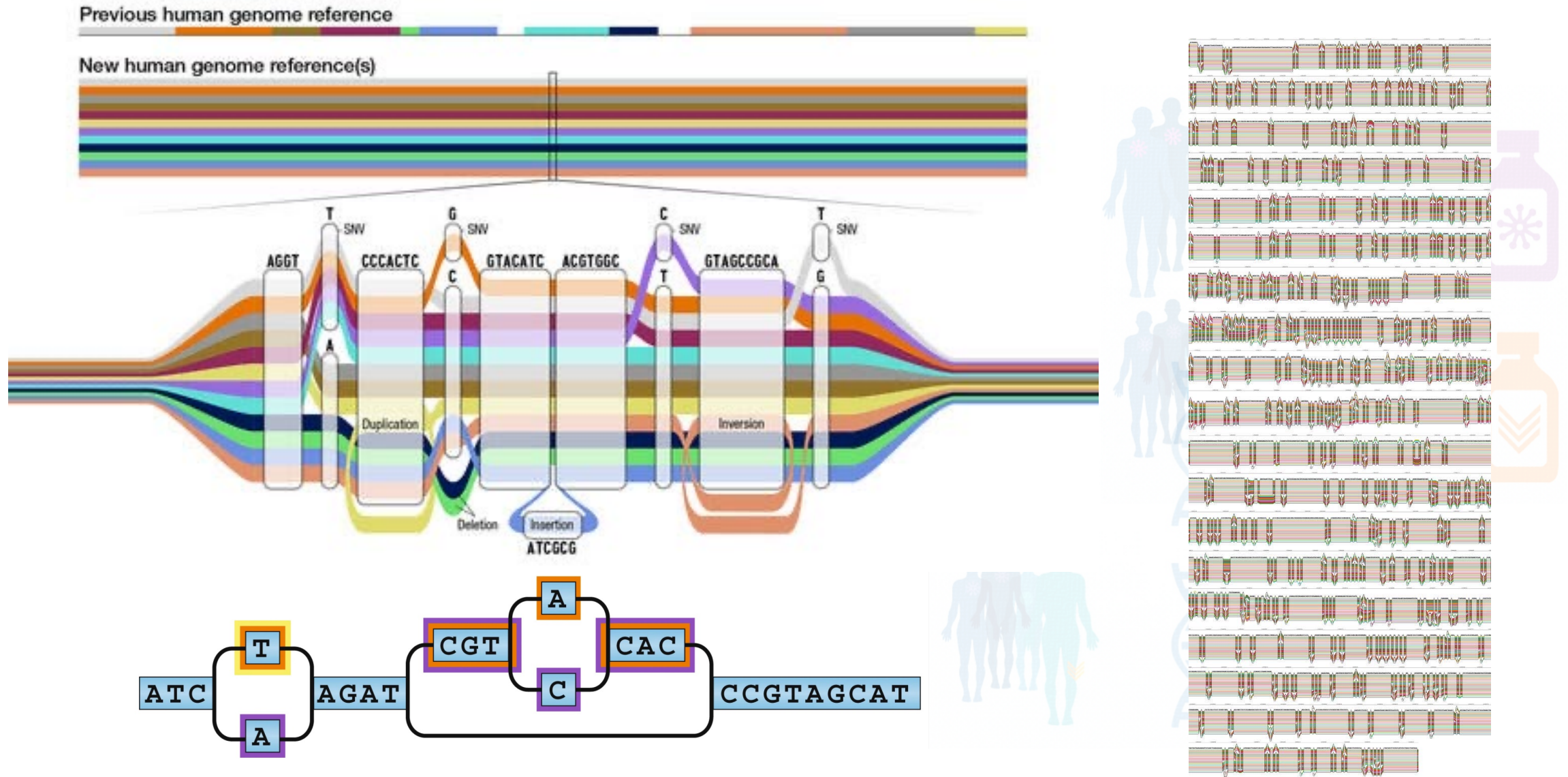


Supplementary Table 16. Comparisons of reference genome build versions on alignment rates and alignment times. (All mapping was performed with Novoalign2.) Using a larger reference genome build leads to higher mapping rates and shorter mapping times.



Superstar genes & Information bias

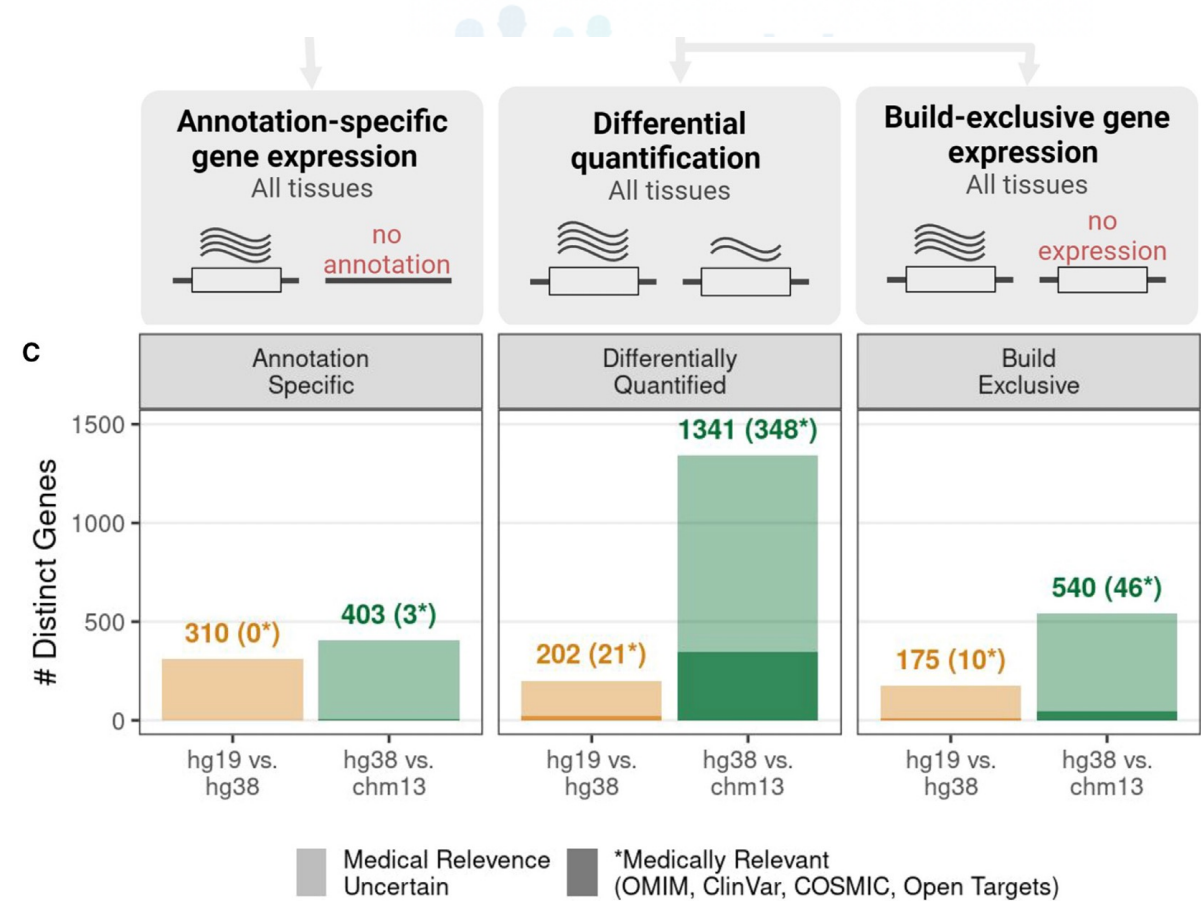
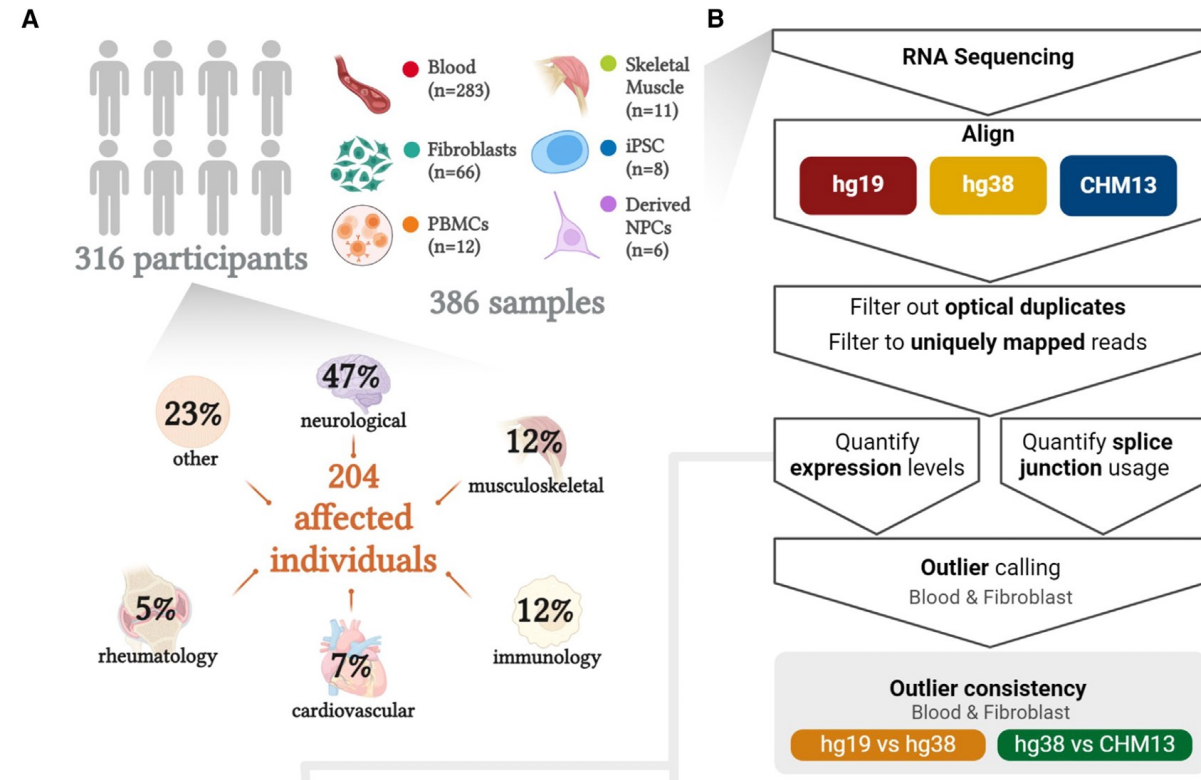
Human Pangenome Reference Consortium (HPRC)



The reference genome hash tables

Assembly	Type	Reference Genome Hash Table	File Name	Major DRAGEN Version	3.8	3.9	3.10	4.0	4.2	4.3	4.4
				Hash Table Version	8	8	8	8	9	10	11
hg38	Pangenome (Graph) reference	Homo sapiens [1000 Genomes] hg38 v5 Pangenome	hg38-alt_masked.cnv.graph.hla.methyl_cg.rna-11-r5.0-1	11							✓
		Homo sapiens [1000 Genomes] hg38 v4 Multigenome	hg38-alt_masked.cnv.graph.hla.rna-10-r4.0-1	10						✓	
		Homo sapiens [1000 Genomes] hg38 v3 Multigenome	hg38-alt_masked.cnv.graph.hla.rna-9-r3.0-1	9					✓		
		Homo sapiens [1000 Genomes] hg38 v2 Multigenome	hg38-alt_masked.cnv.graph.hla.rna-8-r2.0-1	8			✓	✓			
		Homo sapiens [1000 Genomes] hg38 alt-masked Multigenome	hg38_alt_masked+cnv+graph+rna-8-r1.0-1	8		✓					
		Homo sapiens [1000 Genomes] hg38 alt-aware Multigenome	hg38_alt_aware+cnv+graph+rna-8-r1.0-0	8	✓	✓	✓				
	Linear reference	Homo sapiens [1000 Genomes] hg38 v5	hg38-alt_masked.cnv.hla.methyl_cg.methylated_combined.rna-11-r5.0-1	11							✓
		Homo sapiens [1000 Genomes] hg38 v4	hg38-alt_masked.cnv.hla.methylated_combined.rna-10-r4.0-1	10						✓	
		Homo sapiens [1000 Genomes] hg38 v3	hg38-alt_masked.cnv.hla.rna-9-r3.0-1	9					✓		
		Homo sapiens [1000 Genomes] hg38 methylation v3	hg38-alt_masked.methylated_combined.methylation.seed_len27-9-r3.0-1	9					✓		
		Homo sapiens [1000 Genomes] hg38 v2	hg38-alt_masked.cnv.hla.rna-8-r2.0-1	8				✓			

Impact of genome build on RNA-seq interpretation and diagnostics



FastP report example

General

fastp version:	0.23.2 (https://github.com/OpenGene/fastp)
sequencing:	paired end (150 cycles + 150 cycles)
mean length before filtering:	150bp, 150bp
mean length after filtering:	143bp, 143bp
duplication rate:	34.004816%
Insert size peak:	151

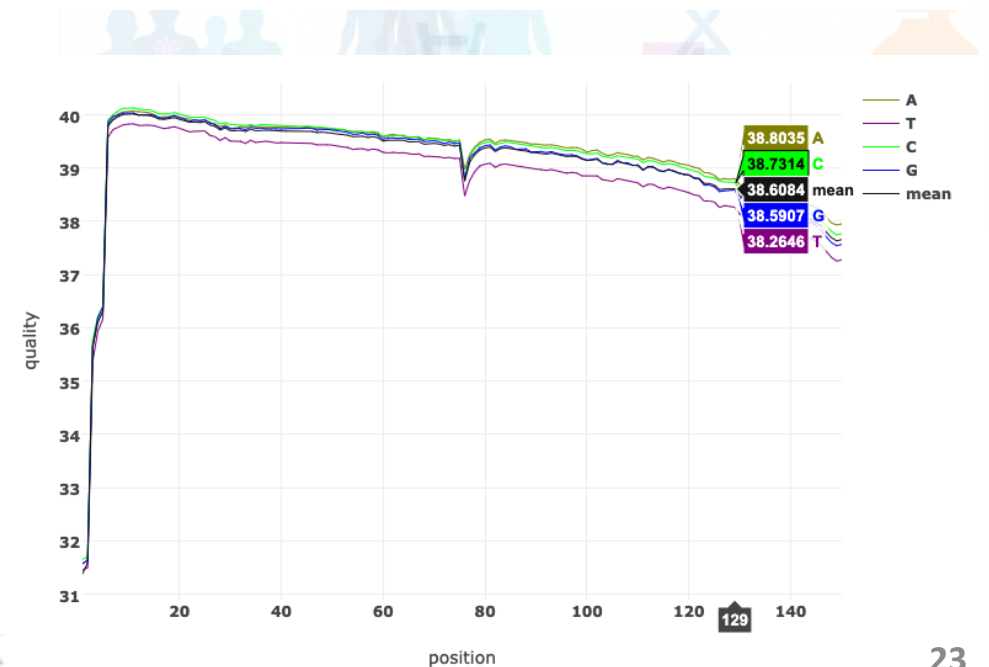
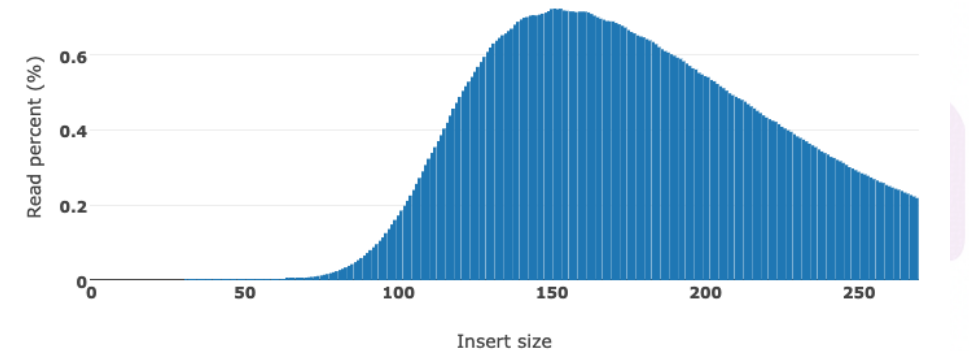
Before filtering

total reads:	843.667924 M
total bases:	126.550189 G
Q20 bases:	122.663198 G (96.928499%)
Q30 bases:	116.997262 G (92.451274%)
GC content:	51.848360%

After filtering

total reads:	827.131766 M
total bases:	118.758132 G
Q20 bases:	115.863008 G (97.562168%)
Q30 bases:	111.022069 G (93.485867%)
GC content:	51.843828%

Insert size distribution (14.149588% reads are with unknown length)



Read Mapping and Reference Genome

Sequencing lengths available

NGS Single End (50–300, Illumina)



NGS Paired End (2*75–300, total 150–600bp, Illumina)



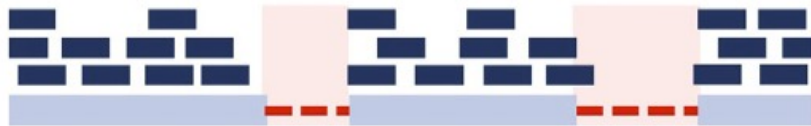
Long Read (>10 000, no fixed upper limit)



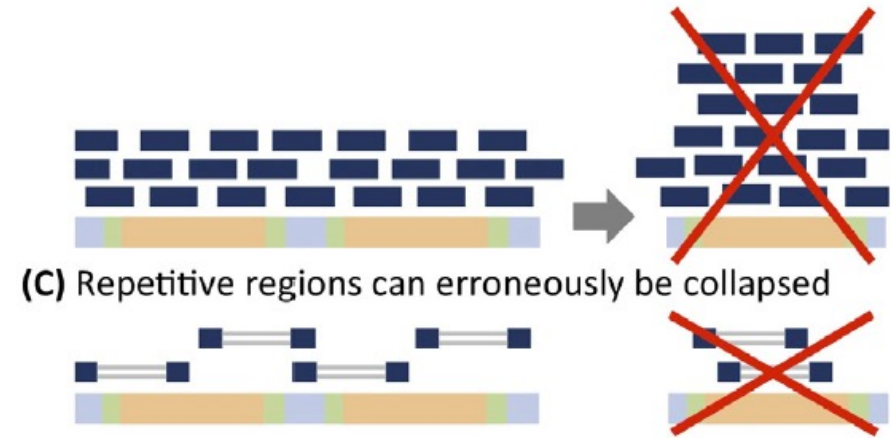
Assembling genome *de novo*



(A) Correct assembly with short reads



(B) Uneven coverage results in missing regions



(C) Repetitive regions can erroneously be collapsed

(D) Paired-end reads anchor repetitive sequences in unique flanking regions, providing the insert size is sufficient insert



(E) Long reads bridge regions, high error-rates



(F) Short reads can correct base calls of long reads

Read Mapping and Reference Genome

Sequencing lengths available

NGS Single End (50–300, Illumina)



NGS Paired End (2*75–300, total 150–600bp, Illumina)



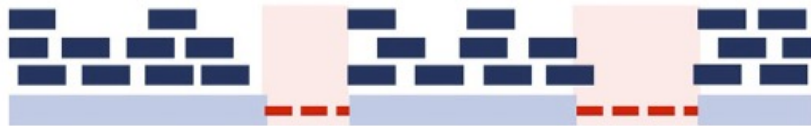
Long Read (>10 000, no fixed upper limit)



Assembling genome *de novo*

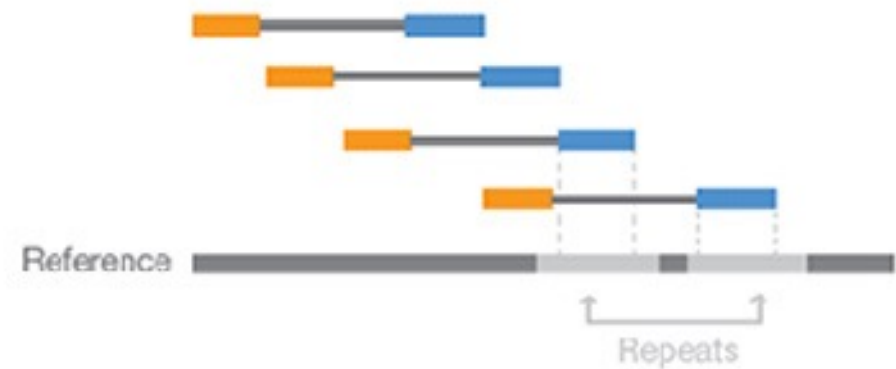


(A) Correct assembly with short reads



(B) Uneven coverage results in missing regions

Alignment to the Reference Sequence



(E) Long reads bridge regions, high error-rates



(F) Short reads can correct base calls of long reads

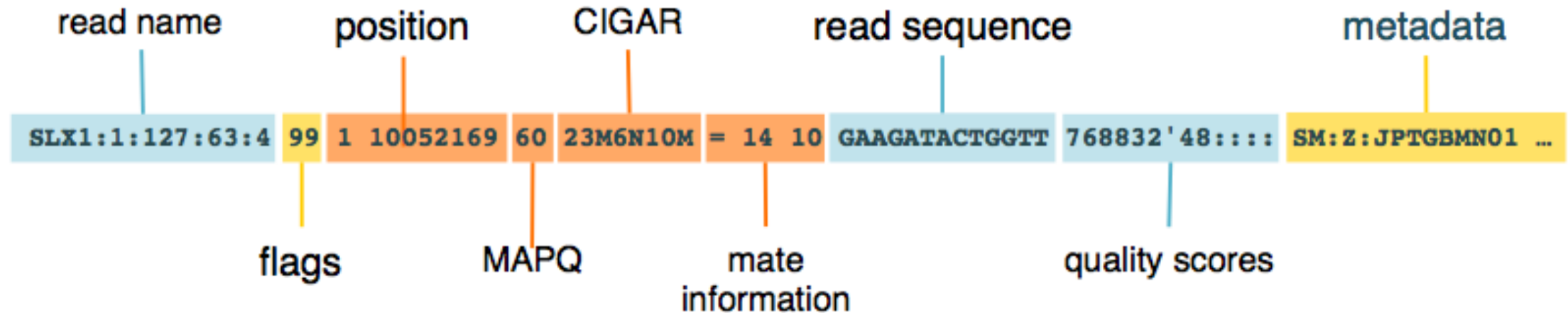
Alignment Map (mapped reads)

- Sequencing Alignment Map (**SAM**), [filename.sam](#)
 - Text format for human to read
 - Very Large
- Binary Alignment Map (**BAM**), [filename.bam](#)
 - Binary format for computer to read
 - View by Samtools (`samtools view Filename.bam`)
 - Large
- Compressed Reference Alignment Map (**CRAM**), [filename.cram](#)
 - 30%~60% smaller than BAM
 - For long-term storage (with reference)
- Cost vs. Informative
- **uBAM**: *unmapped* BAM, which holds unmapped read data (meta data)
- Aligned and un-aligned reads in SAMBAM/CRAM
- The Global Alliance for Genomics and Health (**GA4GH**)

Sequencing Alignment Map (SAM)

HEADER containing metadata (sequence dictionary, read group definitions etc)

RECORDS containing structured read information (1 line per read record)



The alignment map presents the information for **each read** against **reference genome**

- Are these fragments **aligned**?
- What is the **mapping quality**?
- Is there any **unmapped** fragment?
- Other information such as duplication, read-pair, etc.



SAM Format - CIGAR

```
Coord      12345678901234  5678901234567890123456789012345
ref        AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1    TTAGATAAAGGATA*CTG
+r002      aaaAGATAA*GGATA
+r003      gcctaAGCTAA
+r004      ATAGCT.....TCAGC
-r003      ttagctTAGGC
-r001/2    CAGCGGCAT
```

The corresponding SAM format is:¹

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001  99 ref  7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *
r002   0 ref  9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *
r003   0 ref  9 30 5S6M * 0 0 GCCTAAGCTAA * SA:Z:ref,29,-,6H5M,17,0;
r004   0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *
r003 2064 ref 29 17 6H5M * 0 0 TAGGC * SA:Z:ref,9,+,5S6M,30,1;
r001 147 ref 37 30 9M = 7 -39 CAGCGGCAT * NM:i:1
```


SAM Format - CIGAR

```
@HD VN:1.5 S0:coordinate
@SQ SN:ref LN:45
r001 99 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *
r003 0 ref 9 30 5S6M * 0 0 GCCTAAGCTAA * SA:Z:ref,29,-,6H5M,17,0;
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *
```

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

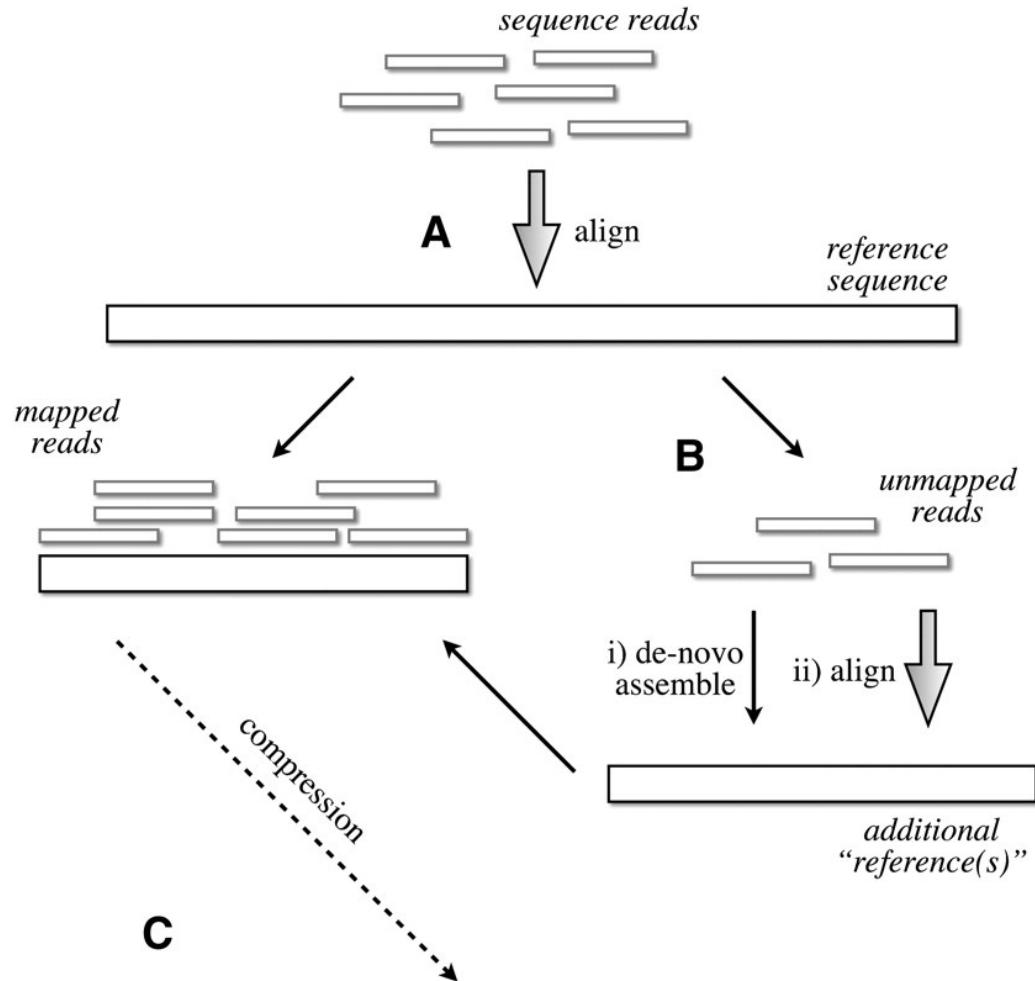
BAM

- BAM is the **compressed binary** version of the Sequence Alignment/Map (SAM) format
- A compact and **index-able** representation of nucleotide sequence alignments.
- Indexed BAM is that only **the portions of the files** needed to display a particular region are transferred to UCSC.
- View by Samtools (`samtools view Filename.bam`)

Alignment Map (mapped reads)

- Sequencing Alignment Map (**SAM**), filename.sam
 - Text format for human to read
 - Very Large
- Binary Alignment Map (**BAM**), filename.bam
 - Binary format for computer to read
 - View by Samtools (`samtools view Filename.bam`)
 - Large
- Compressed Reference Alignment Map (**CRAM**), filename.cram
 - 30%~60% smaller than BAM
 - For long-term storage (with reference)
- Cost vs. Informative
- **uBAM**: *unmapped* BAM, which holds unmapped read data (meta data)
- Aligned and un-aligned reads in SAMBAM/CRAM
- The Global Alliance for Genomics and Health (**GA4GH**)

CRAM Format

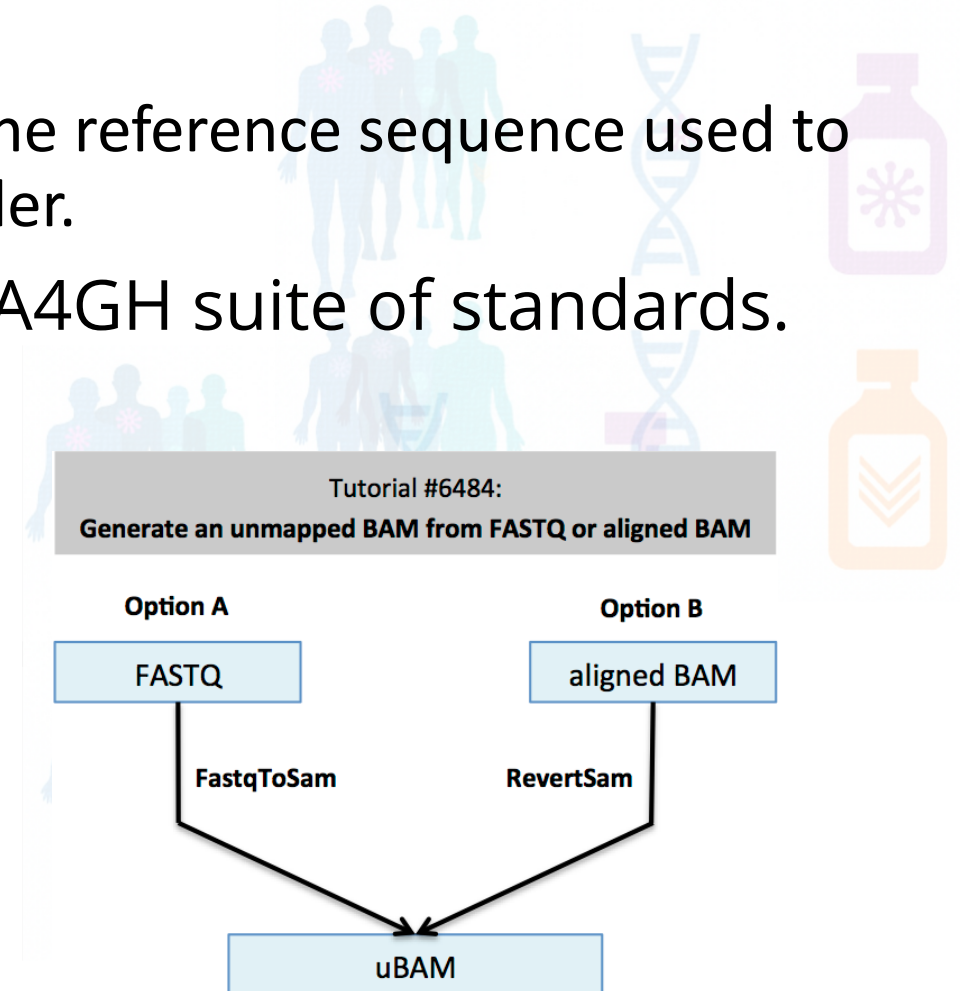


1 10 20 30
 | | | |
 ACGATCTTAATGCCTTACTTGTT--GG-CATTC *reference*
 ATC**G**TAAT---TTAC
 ATGCCTTACTTGT
 ACTTGTT**ATGGCC** *mapped reads*

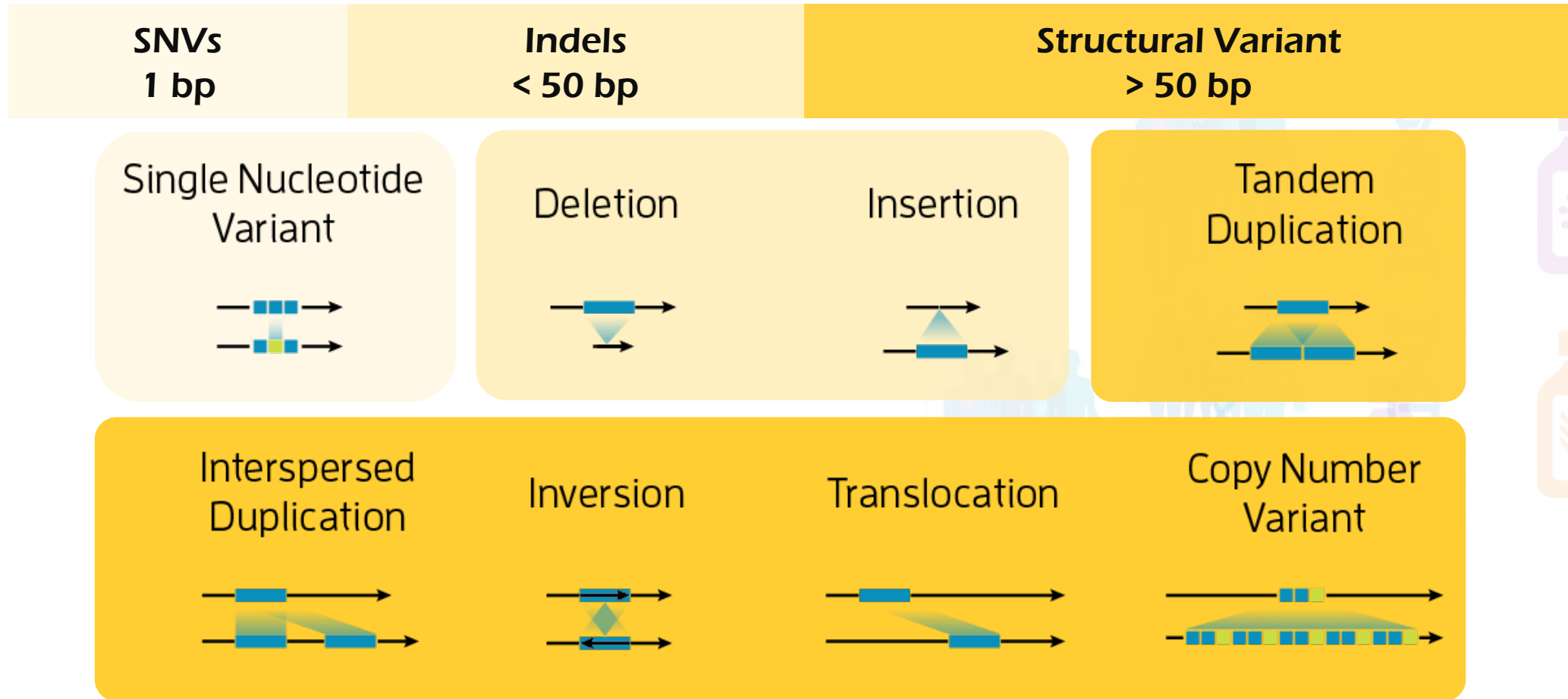
<u>Position</u>	<u>Strand</u>	<u>Substitutions</u>	<u>Insertions</u>	<u>Deletions</u>
4	+	4-G	<i>none</i>	5-3
6	+	<i>none</i>	<i>none</i>	<i>none</i>
7	+	<i>none</i>	8-AT 4-C	<i>none</i>

CRAM & uBAM

- CRAM files are smaller by taking advantage of an additional external "reference sequence" file.
- It is an expectation that the **checksum** of the reference sequence used to create the CRAM will be in the CRAM header.
- CRAM is a fundamental part of the GA4GH suite of standards.
(Ewan Birney, Director of EMBL-EBI)



Genetic variant calls



Types of Variant

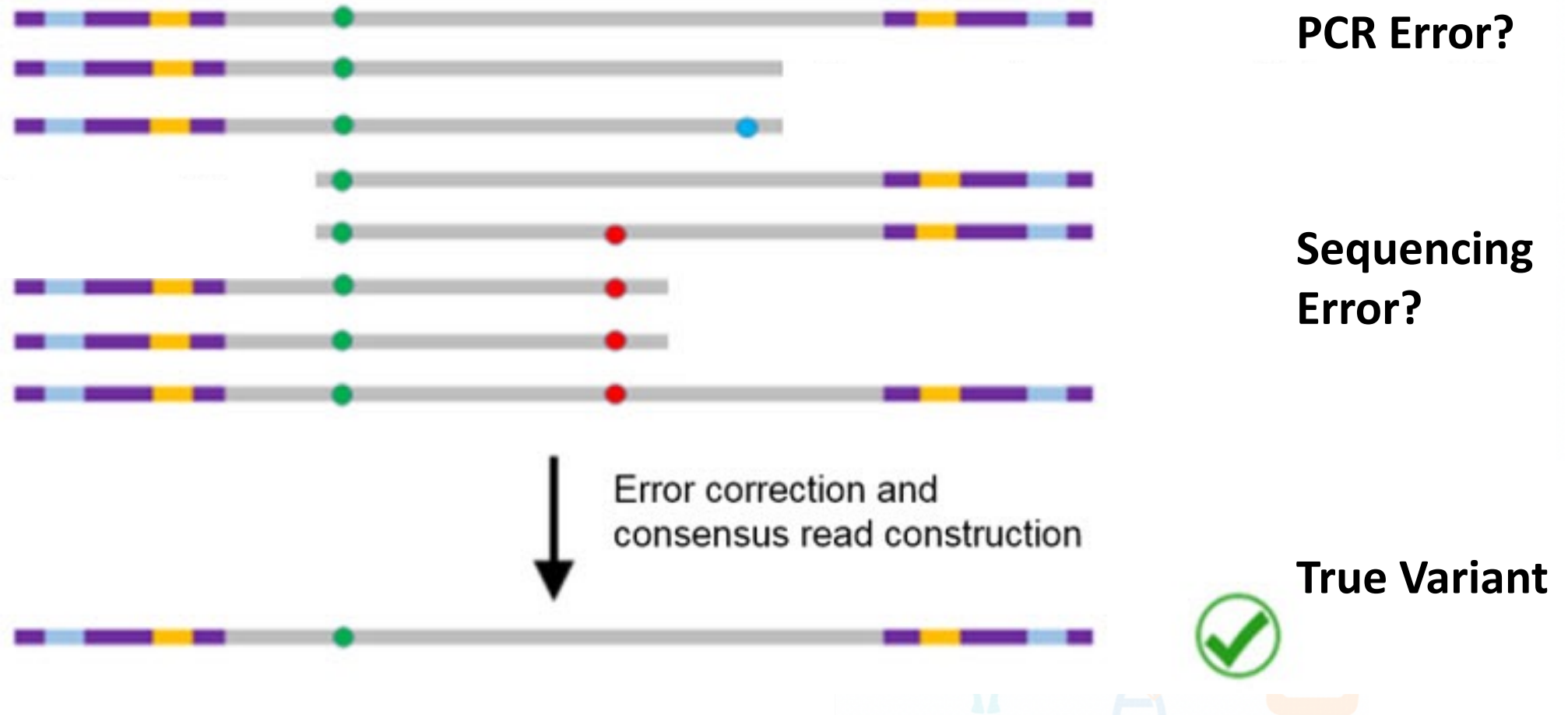
Learning Outcomes

- GATK Best Practices: **Variant Calling**
 - HaplotypeCaller
 - Mutect2
 - Variant Calling Format (VCF/GVCF)
- GATK Best Practices: **Quality Control**
 - Variant-level
 - Subject-level



Variant Discovery: Sequencing Error

PCR error/Sequencing error/True Variant?

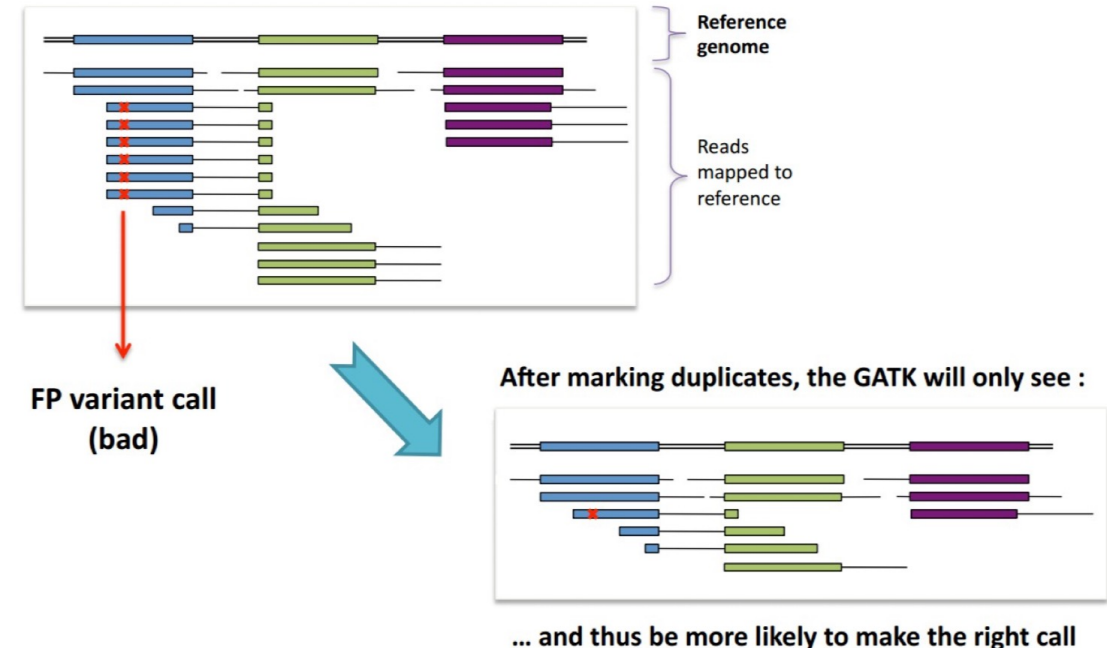


Variant Discovery

- To assign a genotype for each genetic locus (position)
- To determine/validate whether a genetic variant is true or not
- To determine TRUE genetic variant from artifacts
- It's difficult to distinguish by viewing bam files

- Issues: (experiment/analytic aspects)

- PCR artifacts
 - Mismatches in PCR cycles
 - PCR duplicates
- Sequencing Errors
 - erroneous calls (random?)
- Mapping Errors
 - often happens in low-complexity regions



- PCR error
- Sequencing error
- Homopolymers
- Mapping error
- Reference error

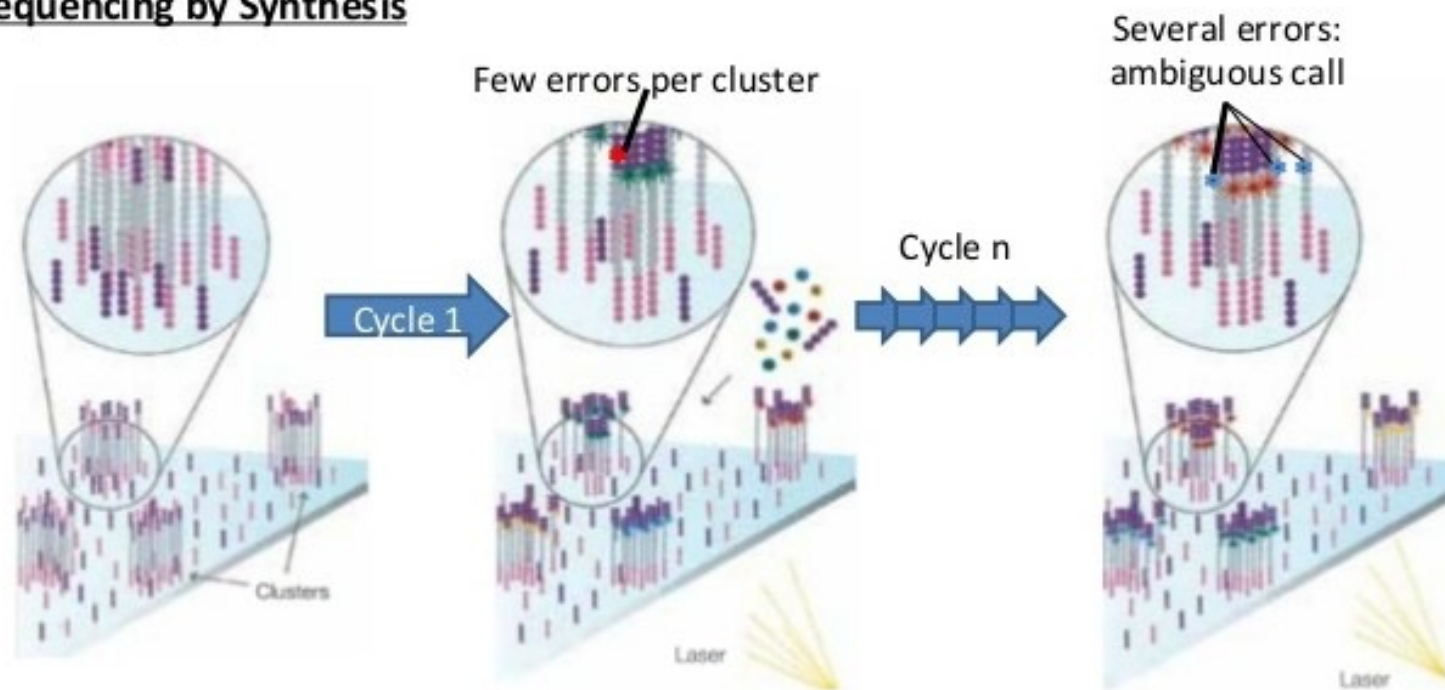
Sources of Error

- Reading error
- Single base error
- Insertion
- Deletion

ccatg -> cc**nn**g
 ccatg -> cc**tt**g
 ccatg -> ccat**cg**
 ccatg -> cc-tg

- Homopolymer errors
 aaaaatg -> aaaa-tg
 aaaaa**at**g
 aaaaa**at**g

Sequencing by Synthesis



Mapping Error & Local Realignment

Before

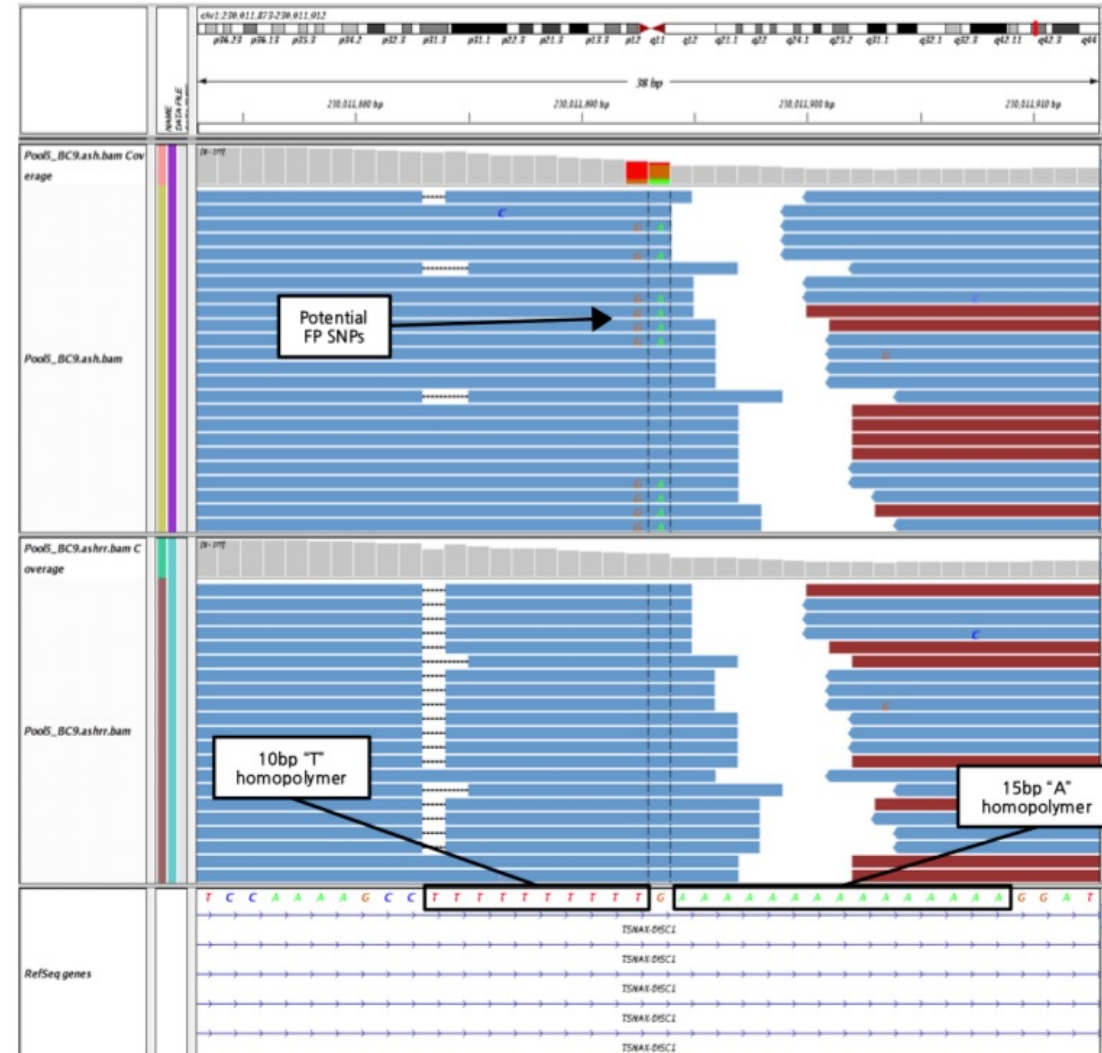
1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7

A A G C C T T T T T T T T G A

After

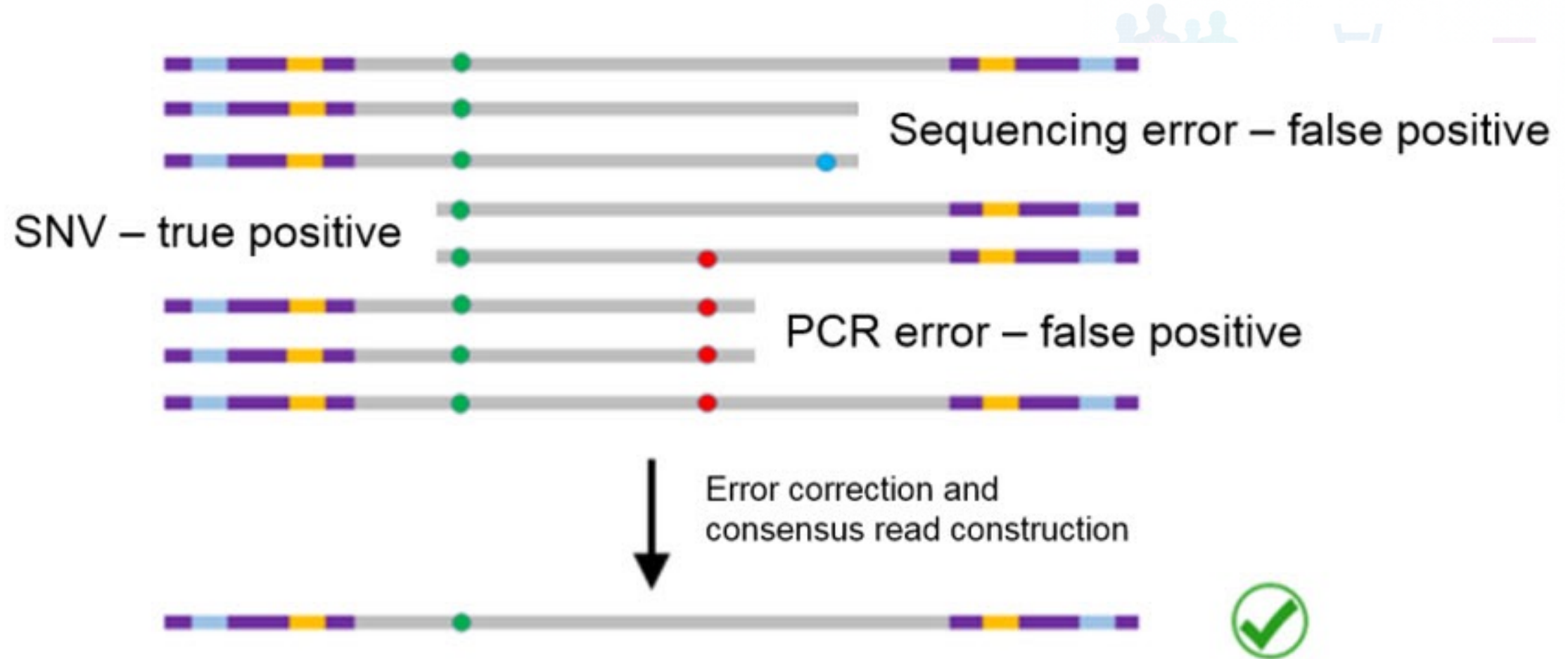
1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7

A A G C C _ T T T T T T T T T G A



Variant Discovery: Sequencing Error

PCR error/Sequencing error/True Variant?

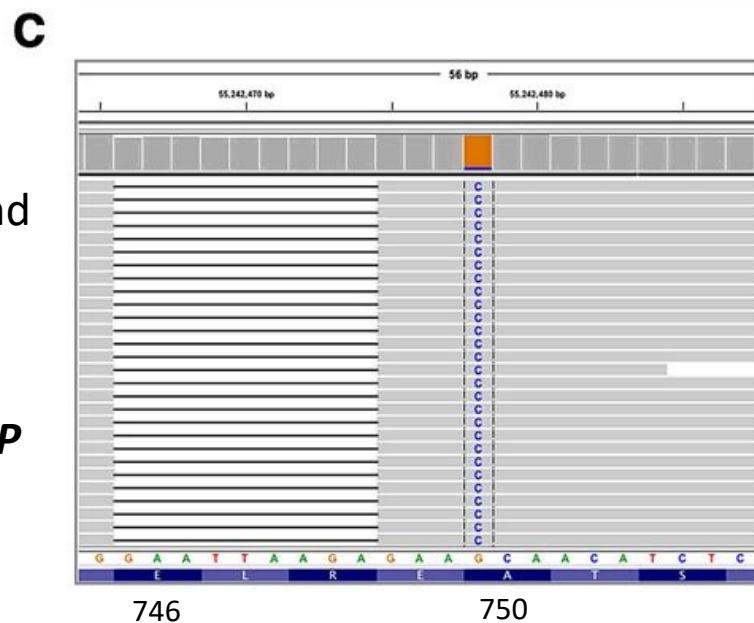
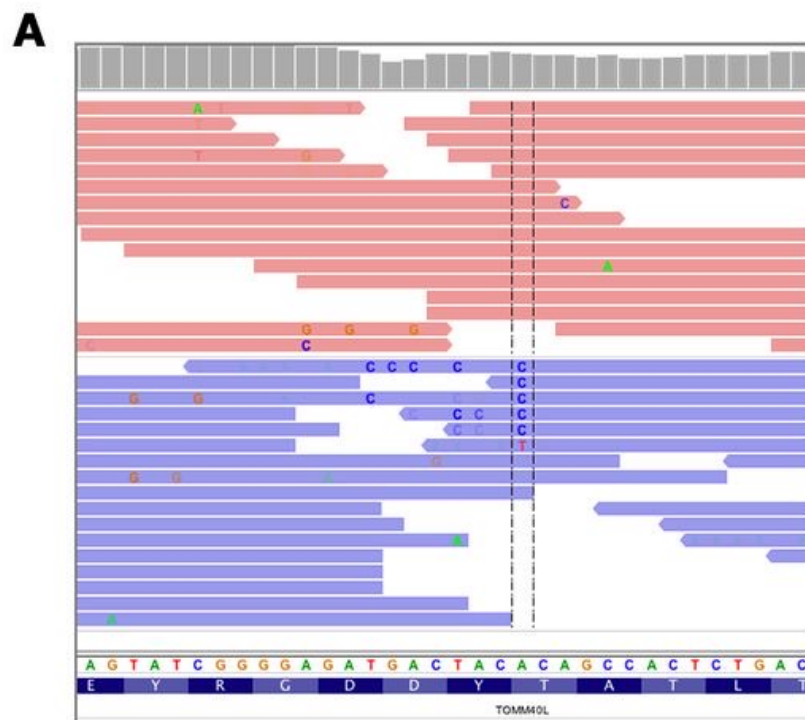




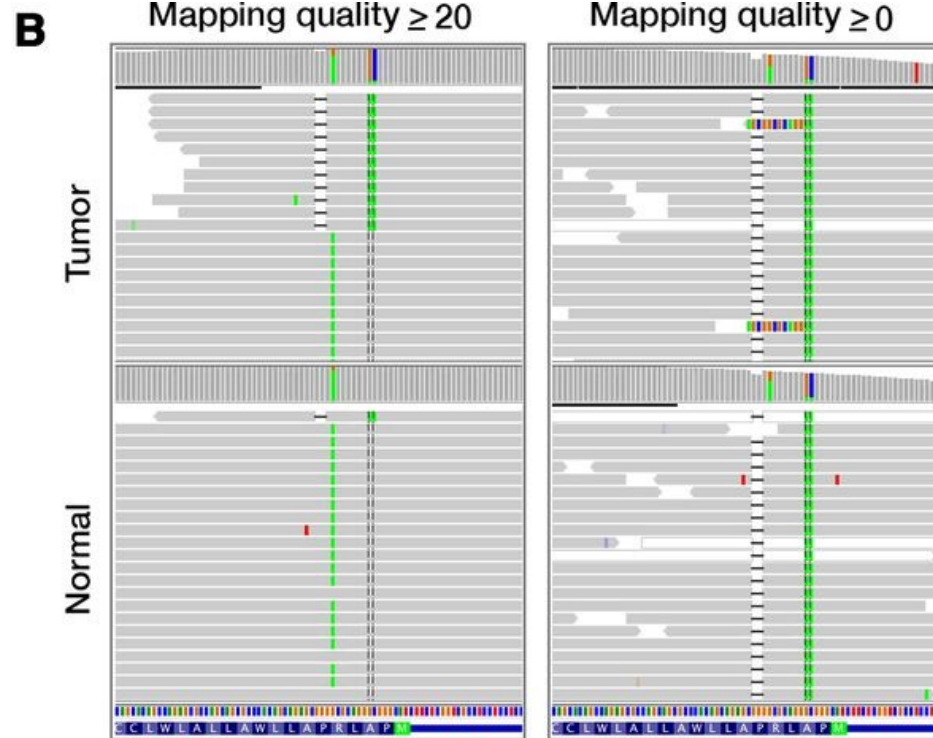
Integrative Genomics Viewer (IGV)



Color by
read **strand**



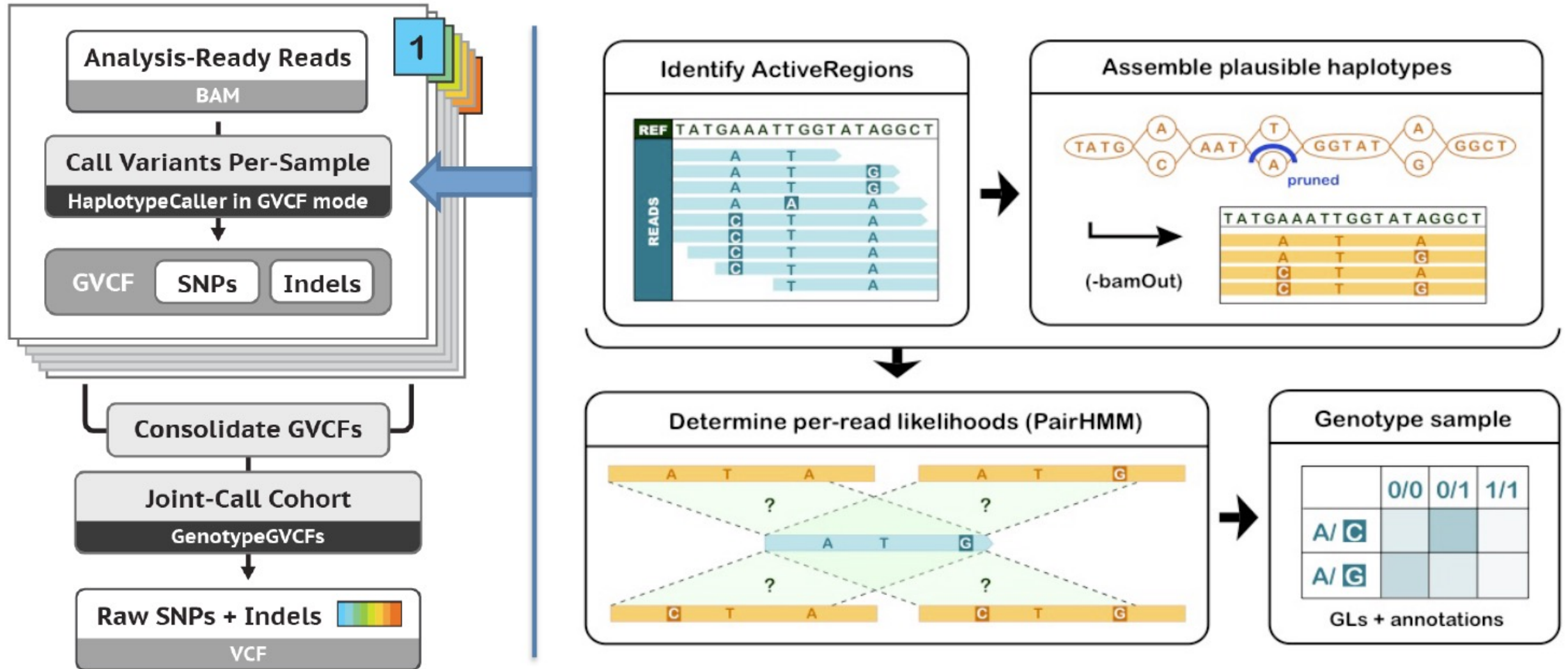
E746_R748del and
A750P
↓
EGFR–
L747_A750delinP



Mapping
quality
&
read with
del + SNV

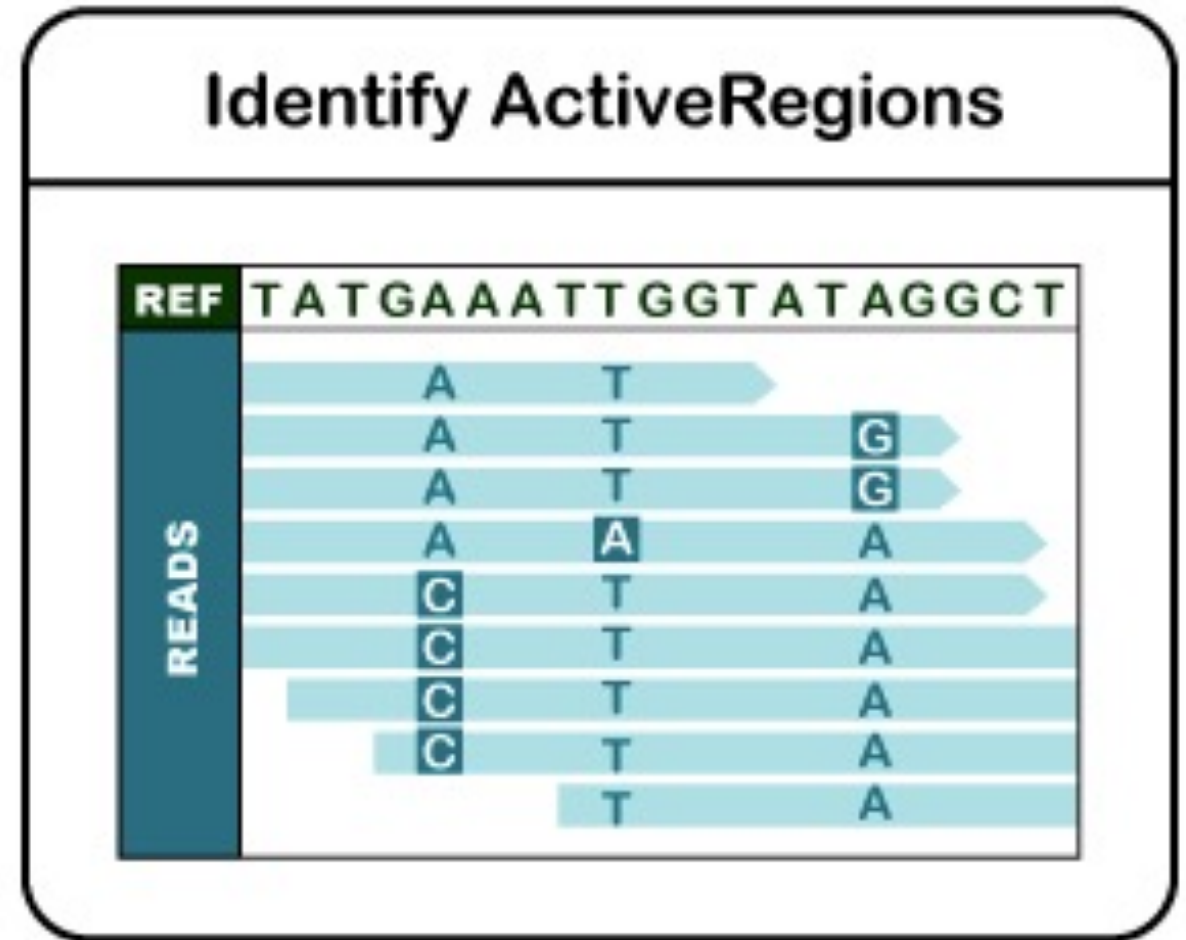
right right
&
left left

HaplotypeCaller (individual/cohort)



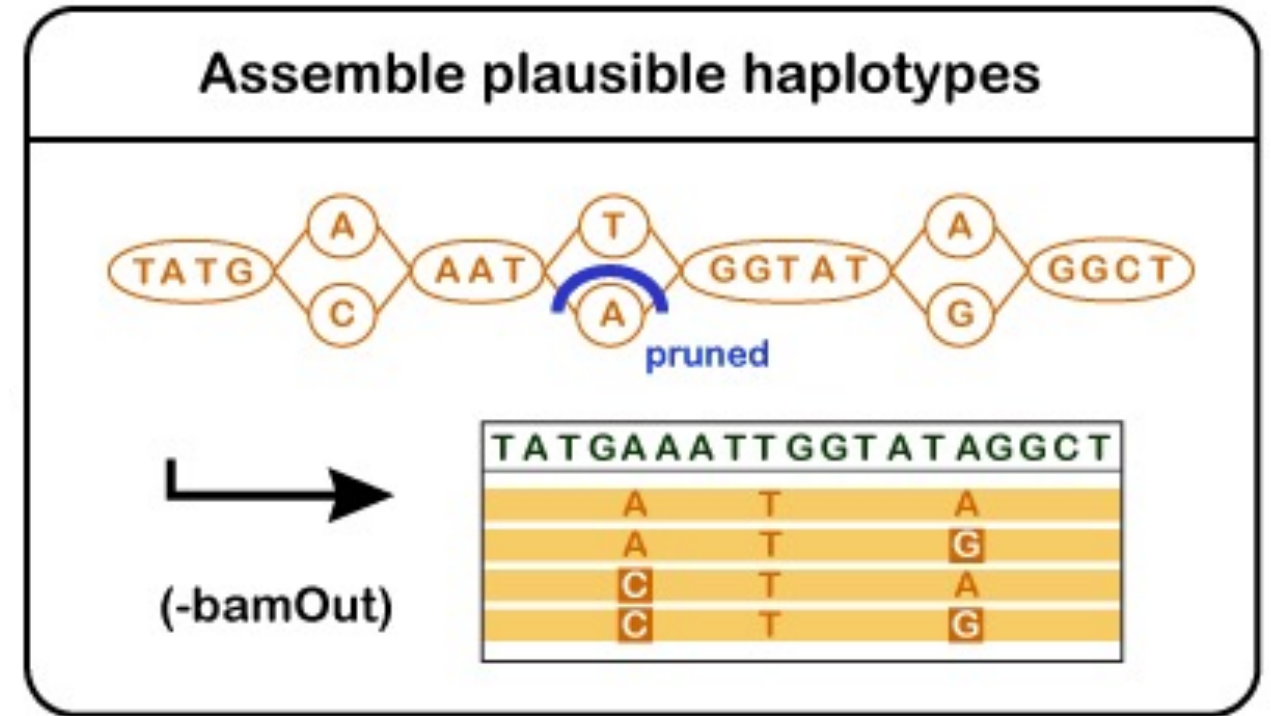
Identify Active Regions

- To identify **regions of the genomes** in which the samples being analyzed show substantial evidence of variation relative to the reference.
- Sliding window along the reference (50~300bp)
- Count mismatches, indels and soft-clips
- Trim and continue with Active Regions over threshold



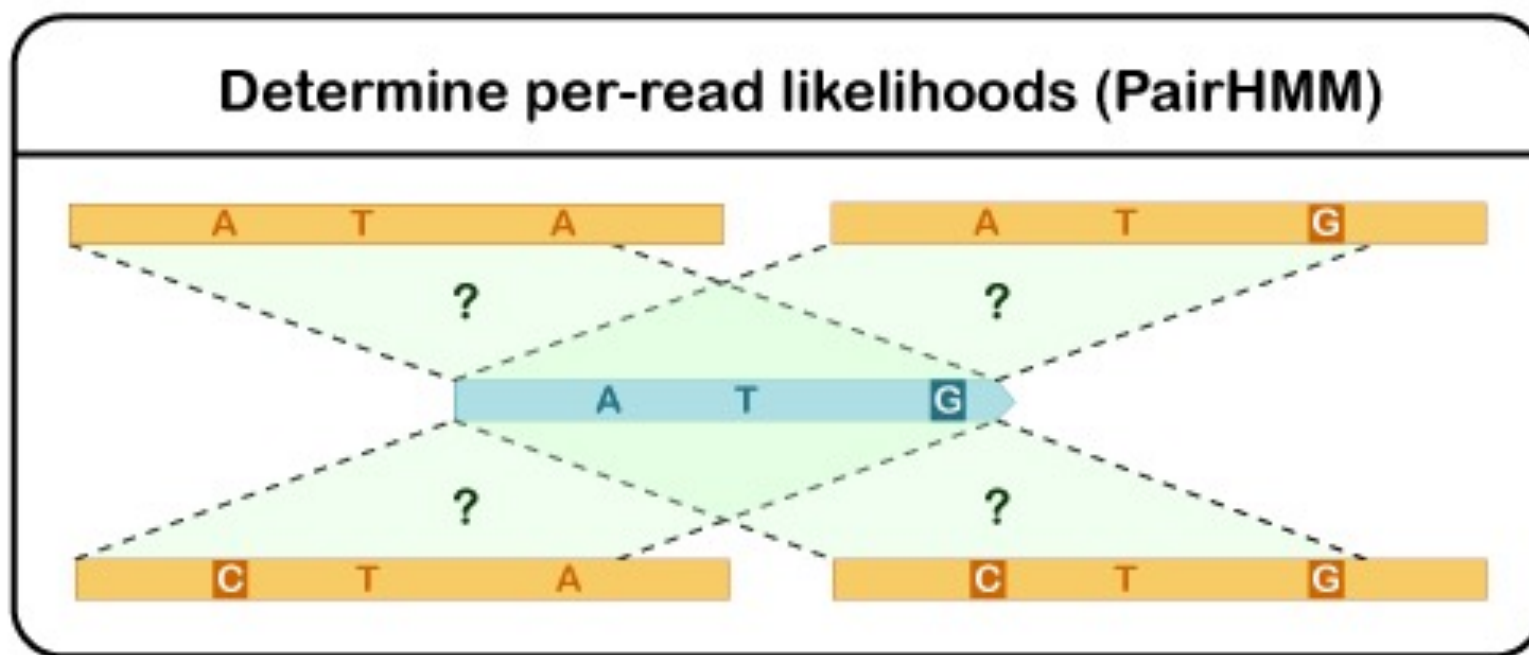
Re-assembly of the active region

- For each Active Region, the program builds a De Bruijn-like graph to reassemble the Active Region and identifies what are the **possible haplotypes** present in the data.
- The program then **realigns each haplotype** against the reference haplotype using the Smith-Waterman algorithm in order to identify potentially variant sites.



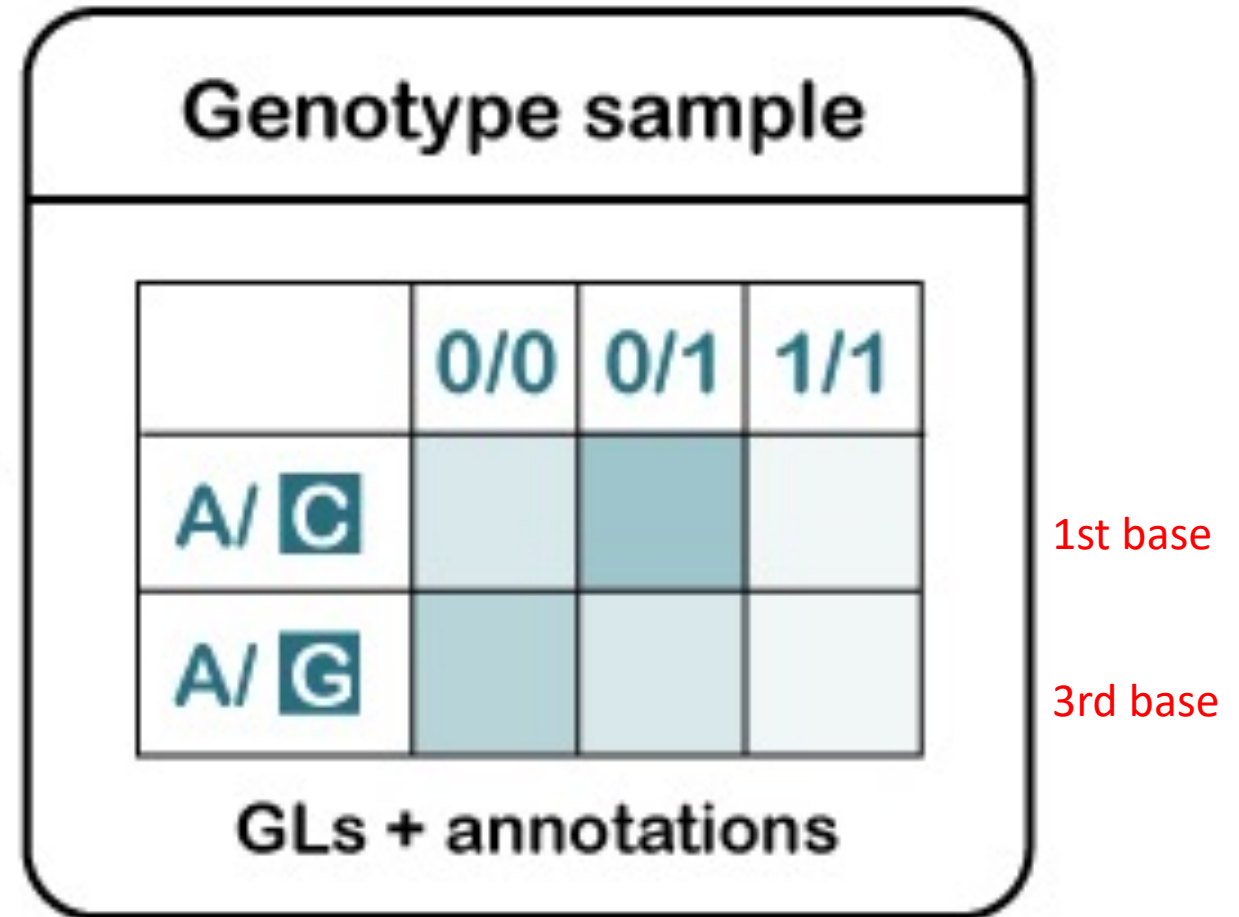
Likelihoods of the haplotypes

- For each ActiveRegion, the program performs a **pairwise** alignment of **each read** against **each haplotype** using the PairHMM algorithm.
- This outputs a score for each read-haplotype pairing, expressing **the likelihood of observing that read** given that haplotype.



Assign sample genotypes

- For each potentially variant site, the program applies Bayes' rule, using the likelihoods of alleles given **the read data** to calculate the posterior likelihoods of each genotype per sample **given the read data observed** for that sample.
- The most likely genotype is then assigned to the sample.



AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

[illegible]

AAAAAAAAAAAAAAAAAGGGGGGGGGGGGGGGGGGG

AAAAAAAAAAAAAAAAAAGGGGGGGGGGGGGCT

AAA GGG CCTT

$N=30, X=0 \rightarrow \mathbf{RR}$

$$N=30, \quad X=30 \rightarrow \mathbf{VV}$$
$$N=30, X=15 \rightarrow \mathbf{RV}$$

$N=30, X=12 \rightarrow \mathbf{RV}$

$N=10, X=3 \rightarrow \mathbf{RV?}$

- $c_1 = 10\%, c_2 = 30\%$

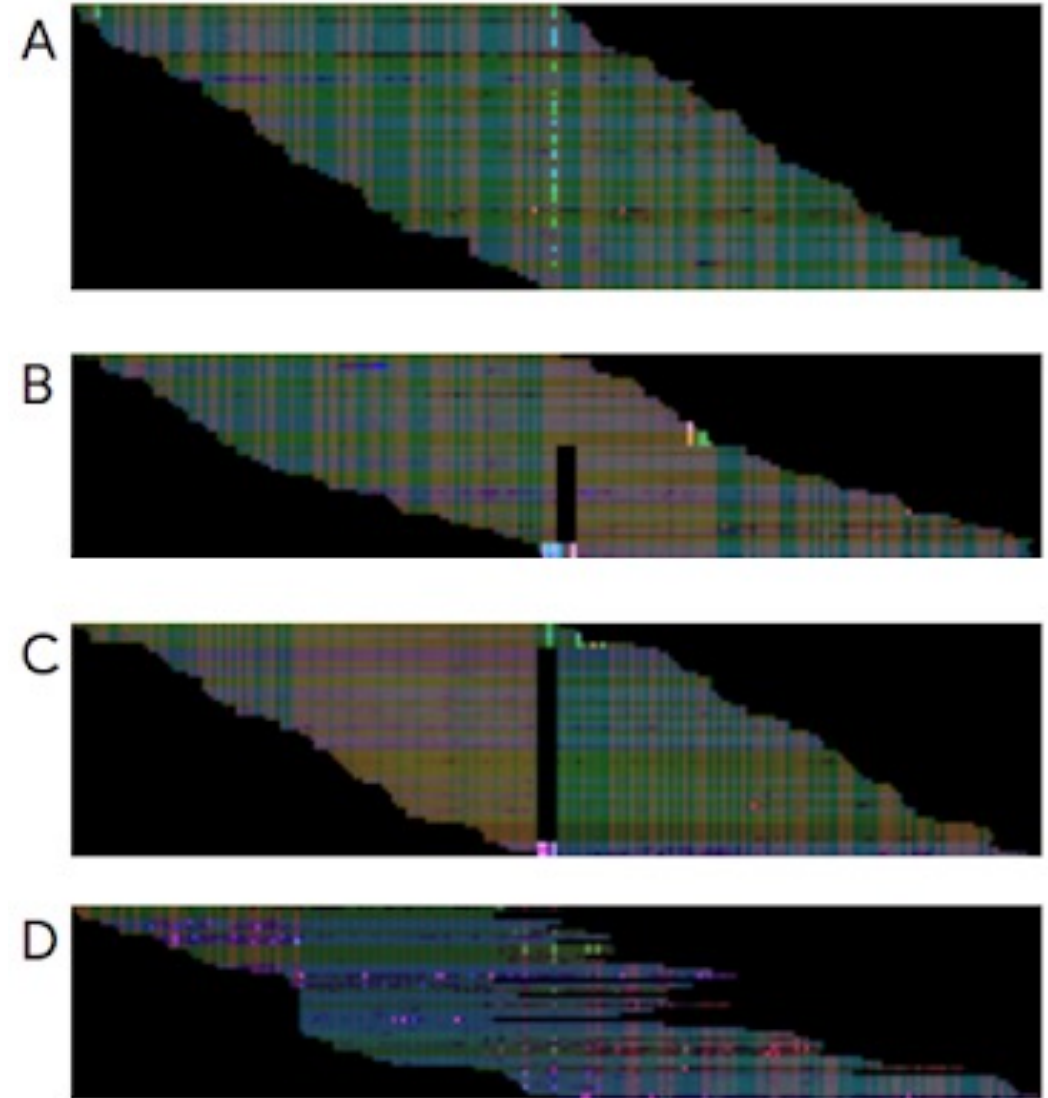
$X \leq c_1$	$\rightarrow$	RR
$c_1 < X < c_2$	$\rightarrow$	RV
$X \geq c_2$	$\rightarrow$	VV

Outcomes:

	RR	RV	VV
RR	0.00	0.00	0.00
RV	0.00	0.00	0.00
VV	0.00	0.00	0.00

Deep Variant (2017~)

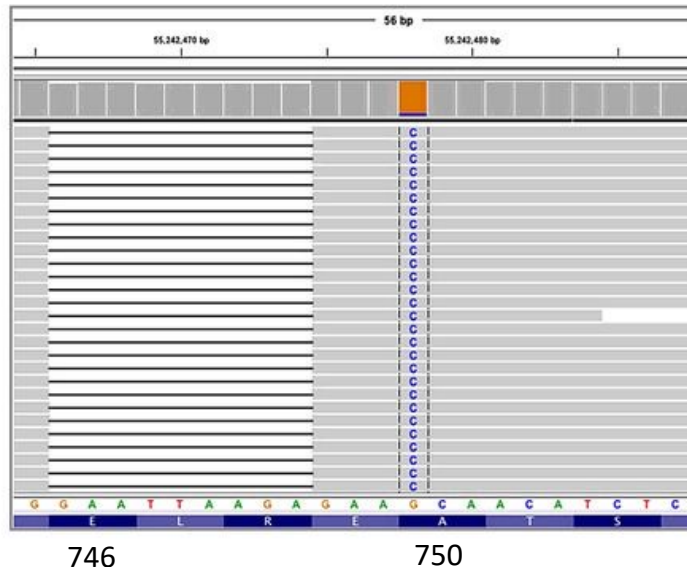
- DeepVariant transforms variant calling task into an image classification problem.
- Colors represent sequence/genotype annotations
- A: a true SNP
- B: a deletion on one chromosome,
- C: a deletion on both chromosomes,
- D: a false variant caused by errors.



Complex variant(s)

- Multi-Nucleotides Variant (MNV, delIns)
- Multi-allelic variant (left-aligned)
- MT variant
- XY chromosomes
- Repeat expansion

*E746_R748del and
A750P*
↓
*EGFR–
L747_A750delinP*



Reference and alternative
alleles of a CA short
tandem repeat (STR)

REF
ALT

GGGCACACACAGGG
GGGCACACAGGG

← CA deletion from the reference

Genome Reference

GGGCACACACAGGG

REF CA
ALT .
REF CAC
ALT C
REF GCACA
ALT GCA
REF GGCA
ALT GG
REF GCA
ALT G

Alleles represented against the human
genome reference. Allele pairs are
colored the same, all are representations
of the same variant.

Variant Call Format

POS	REF	ALT
8	CA	.
6	CAC	C
3	GCACA	GCA
2	GGCA	GG
3	GCA	G

Alleles represented in Variant Call Format,
all are representations of the same variant.

Not left aligned
and alternate
allele is empty
Not left aligned
but parsimonious
Not right trimmed
Not left trimmed
Normalized
(left aligned
& parsimonious)

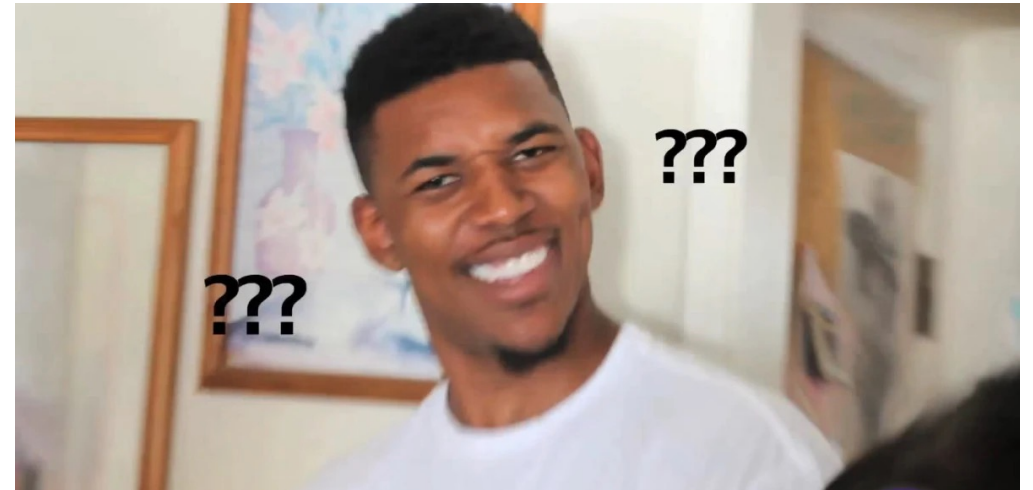
VCF (Variant Call Format)

#CHR	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA12778
M	282		G	A	[CLIPPED]			GT:AD:DP:GQ:PL	1/1:0,18:18:54:676,54,0

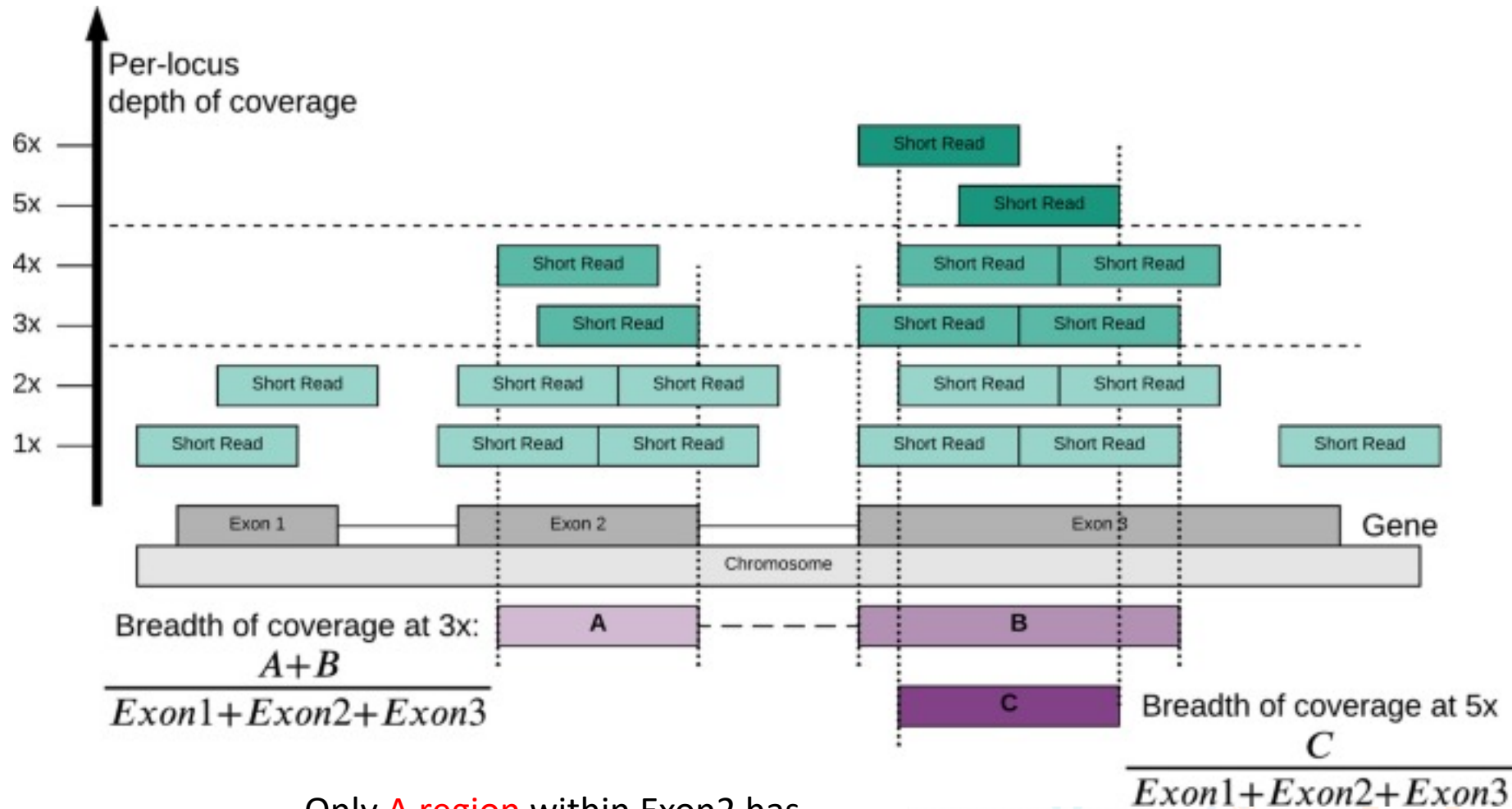
- GT : The genotype of this sample at this site (A/A, homozygous)
- AD : Allele depth (0,18)
- DP : Depth of coverage (18)
- GQ : Quality of the assigned genotype (54, $10^{(-5.4)} \sim 3.98 \times 10^{-6}$)
- PL : Normalized Phred-scaled likelihoods of the possible genotypes
 - (0/0,0/1,1/1), the probability that the genotype is **not** correct
 - PL(0/0) = 676
 - PL(0/1) = 54
 - PL(1/1) = 0

VCF records the variants

ALL other regions are reference homozygous variants ?



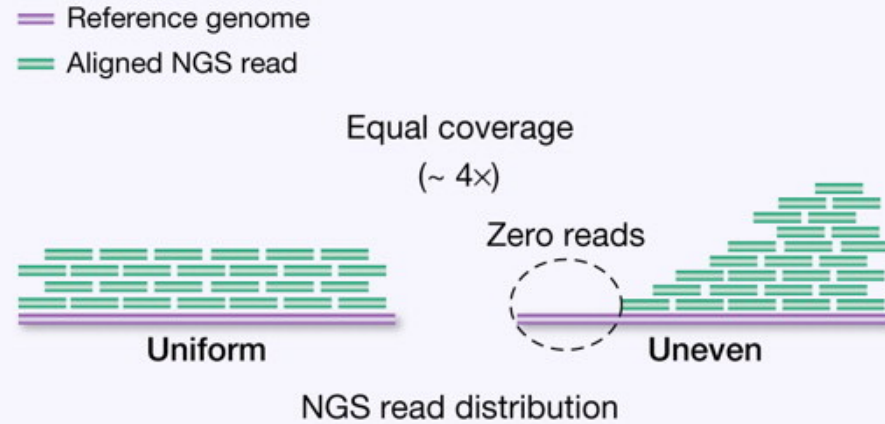
Breadth of coverage at 3x or 5x (bp)



Only **A region** within Exon2 has depth of coverage > 3X

C

Sequence coverage and error rates

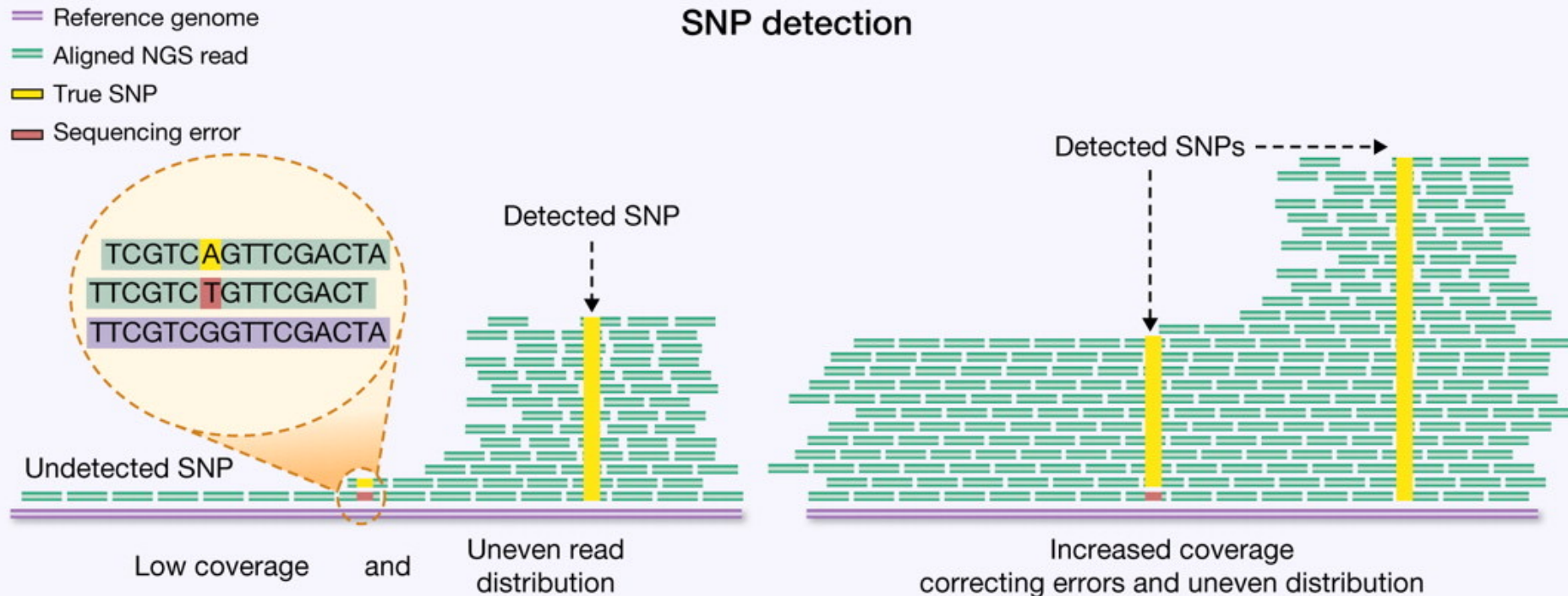


reference allele
or
missing genotype?



D

SNP detection



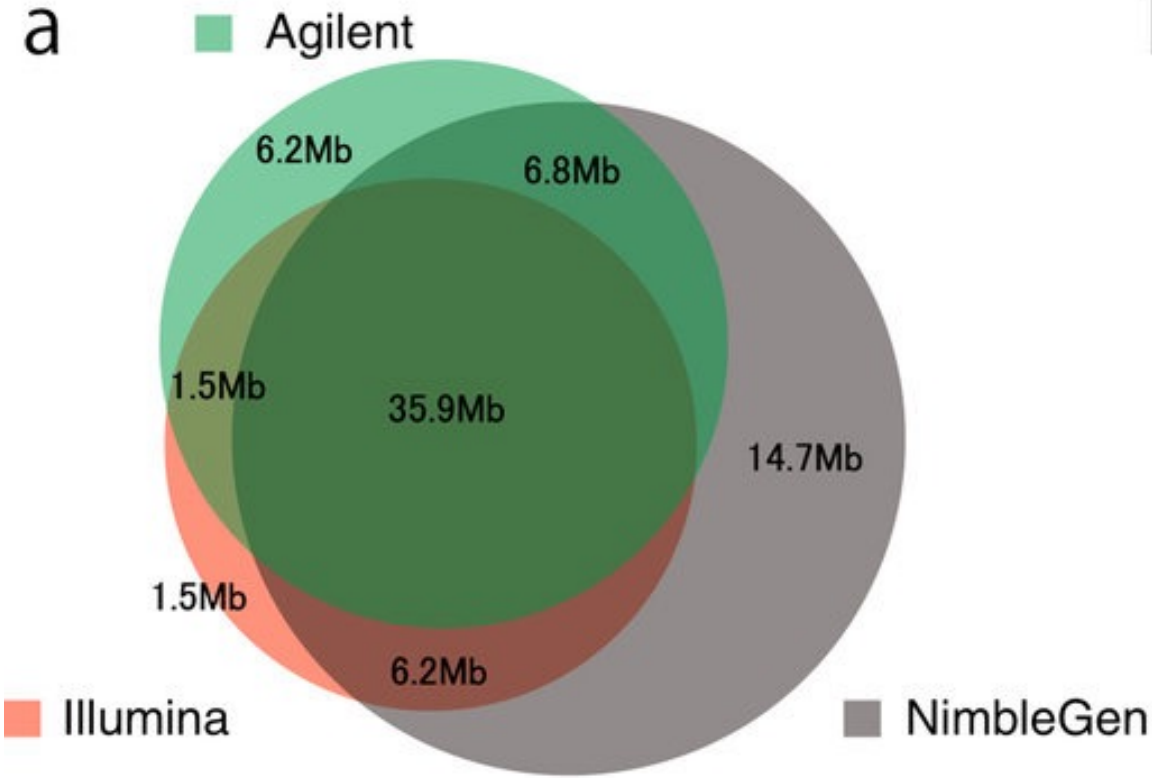
Comparison of Exome Capture Kits

Exome-Seq Kits	Targeted Region	Number of Probes	Probe Type	Genomic DNA input required	Adapter addition	Probe Length (mer)	Probe Design	Price per capture (negotiable)	Designed on build	Hybridization time (hours)
Agilent SureSelect XT2 V6 Exome	60 Mb	~758,086	biotinylated cRNA baits	100 ng	Ligation	120	Non-overlapping, paired-end reads used to fill gaps	\$270	GRCh37 (hg19)	16
Agilent SureSelect XT2 V5 Exome	51 Mb	~655,872	biotinylated cRNA baits	100 ng	Ligation	120	Non-overlapping, paired-end reads used to fill gaps	\$200	GRCh37 (hg19)	16
IDT xGEN Exome Panel	39 Mb	429,826	biotinylated DNA baits	500 ng	Ligation	not described	Non-overlapping	\$250	GRCh37 (hg19)	4
Illumina Nextera Rapid Capture Expanded Exome	62 Mb	>340,000	biotinylated DNA bait	50 ng	Transposase	95	Non-overlapping (adjacent to each other)	\$250	GRCh37 (hg19)	24-48
Roche Nimblegen SeqCap EZ Exome v3.0	64 Mb	>2,100,000	biotinylated DNA bait	1 ug	Ligation	60 - 90	Overlapping baits	\$600	GRCh37 (hg19)	72

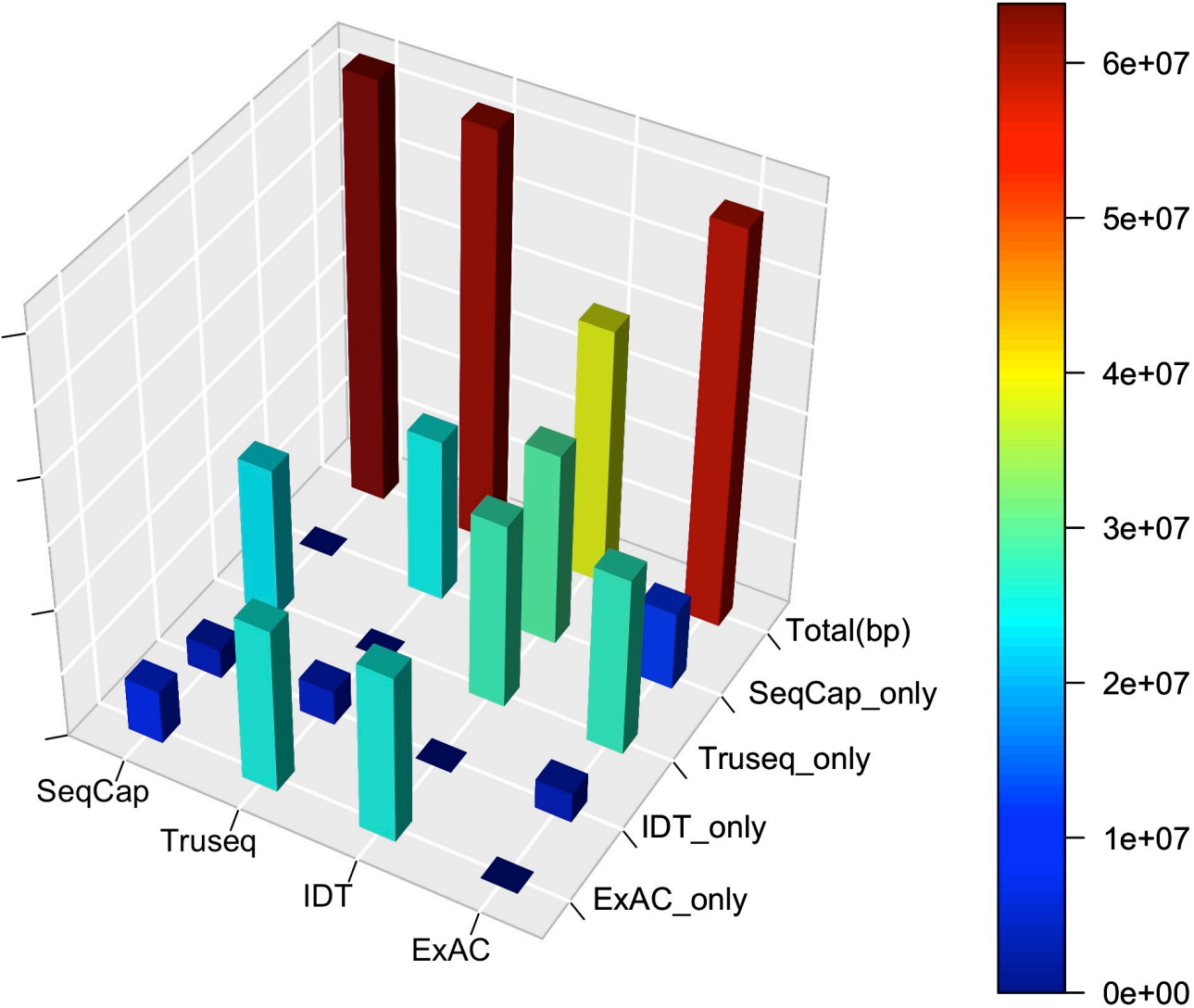


Human Exome Capture Regions

a



ExAC: 30.5~60 Mb (2016)
gnomAD: ~ 30.2 Mb (2020) (non-repetitive)
UKB: ~33.5 Mb (2021) (non-repetitive+flanking regions)

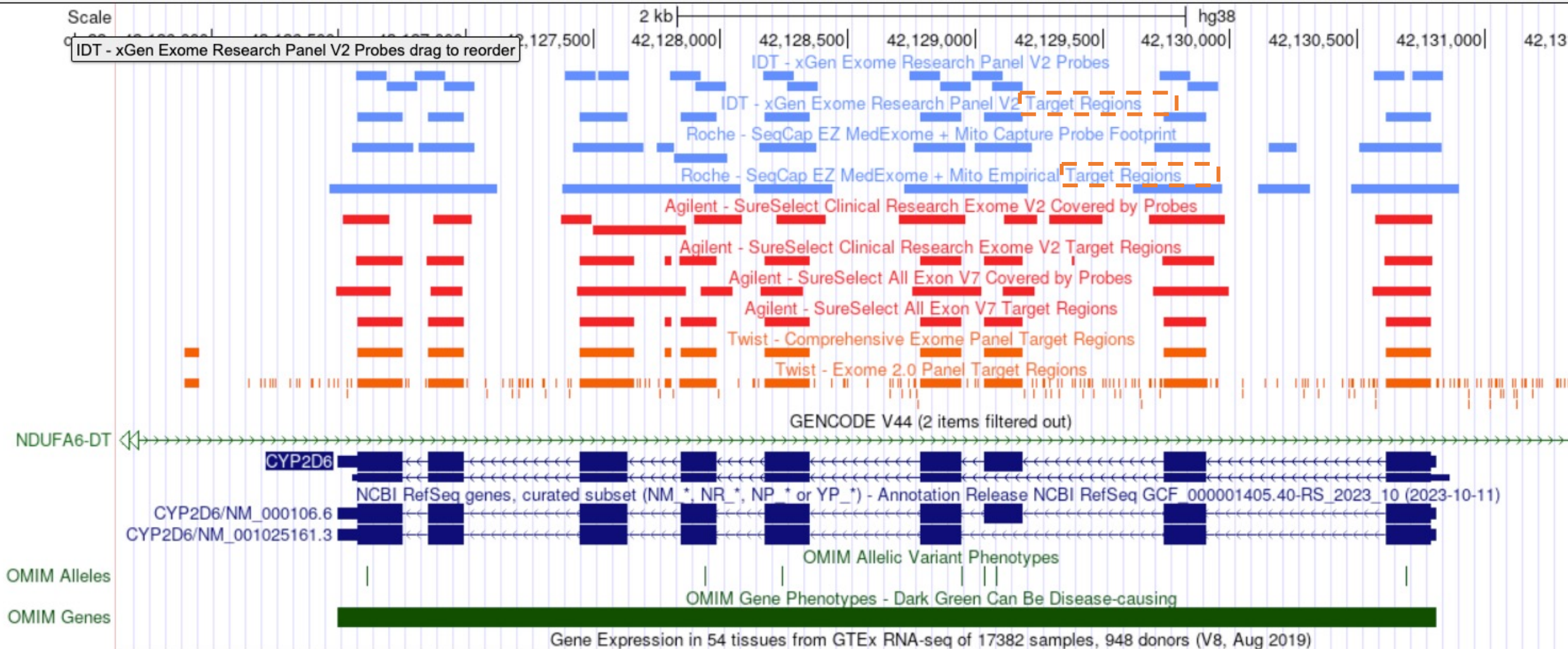


UCSC Genome Browser on Human (GRCh38/hg38)

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x

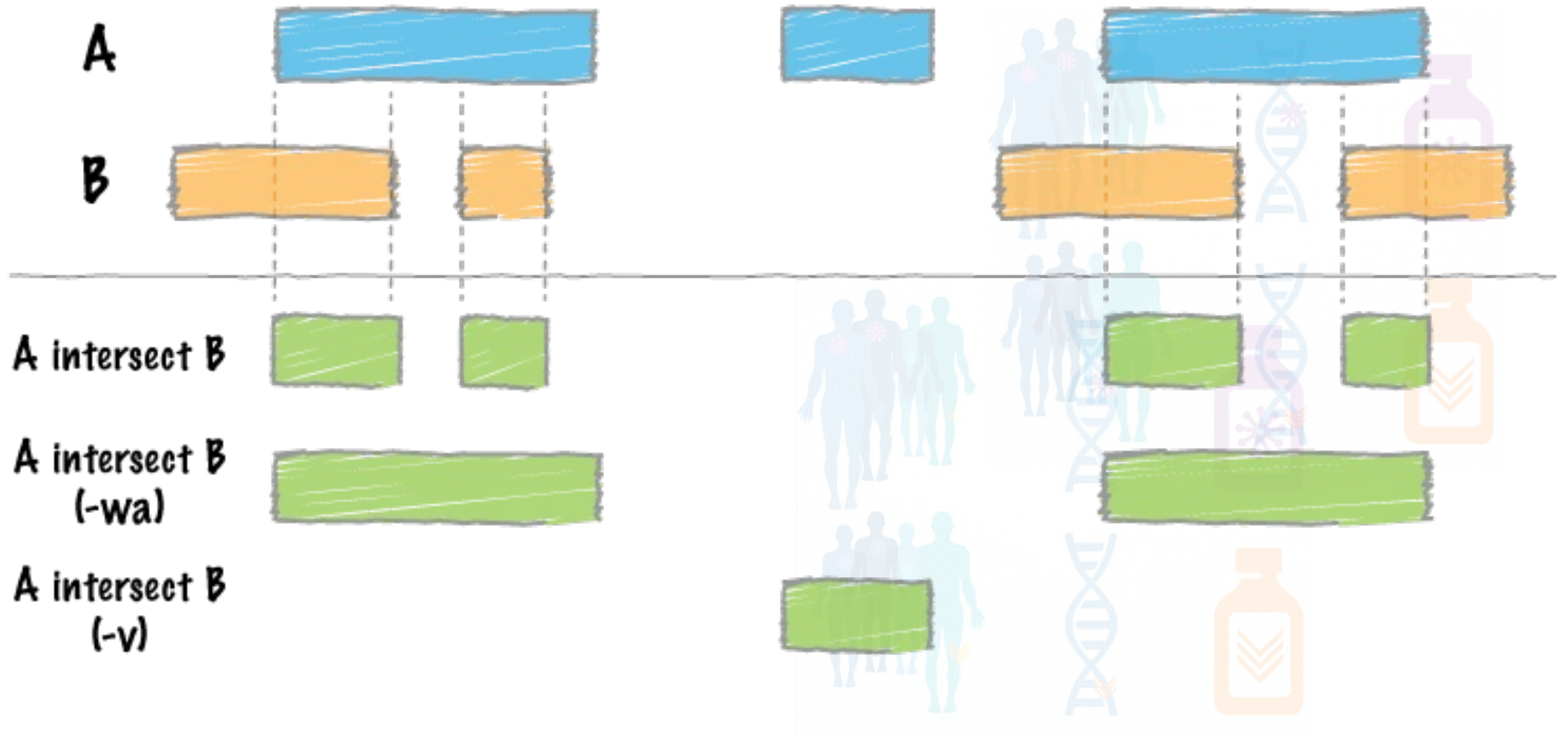
multi-region chr22:42,125,635-42,132,022 6,388 bp. gene, chromosome range, search terms, help pages, see € go [examples](#)

chr22 (q13.2) 22p13 22p12 22p11.2 22q11.21 q11.23 22q12.1 22q12.2 22q12.3 22q13.1 22q13.2 22q13.31 13.33

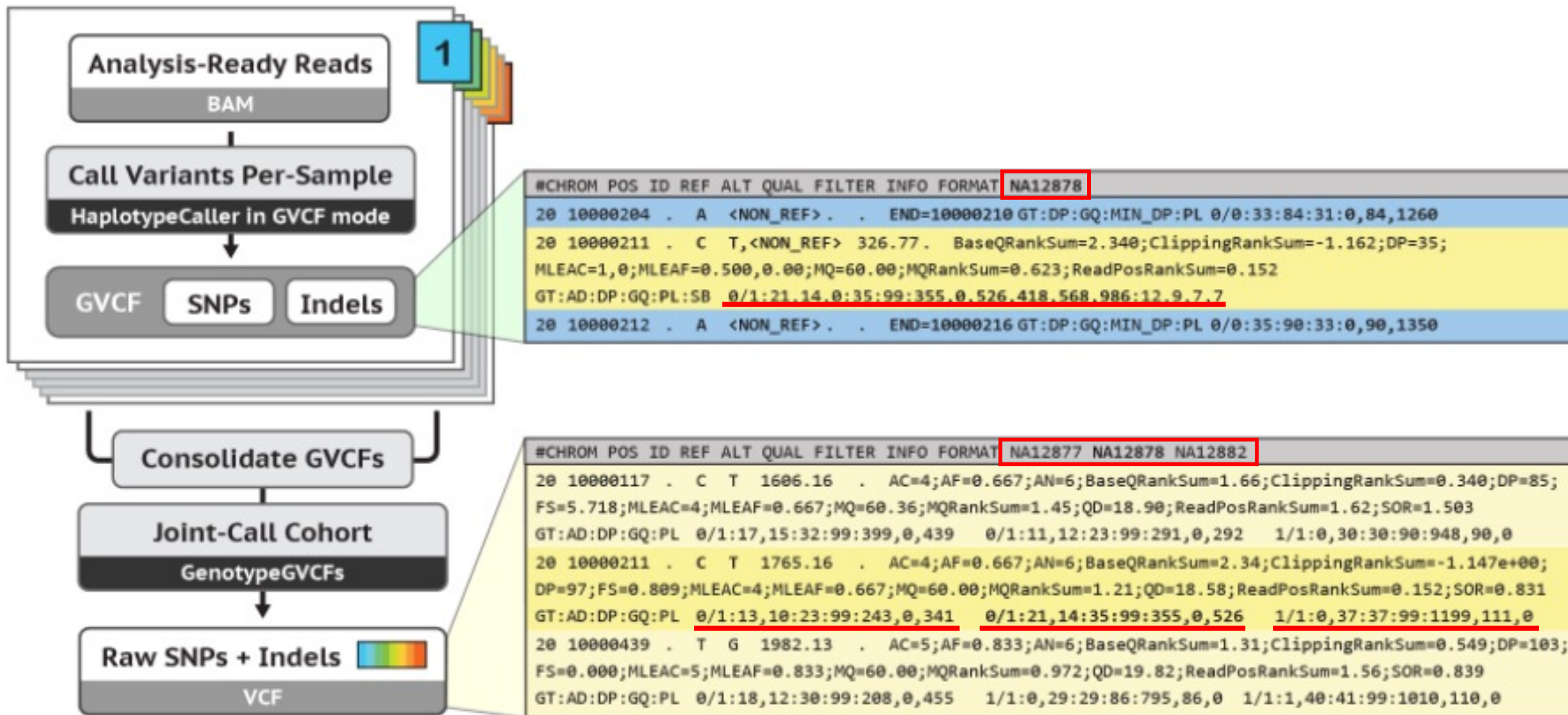


Bedtools *intersect*

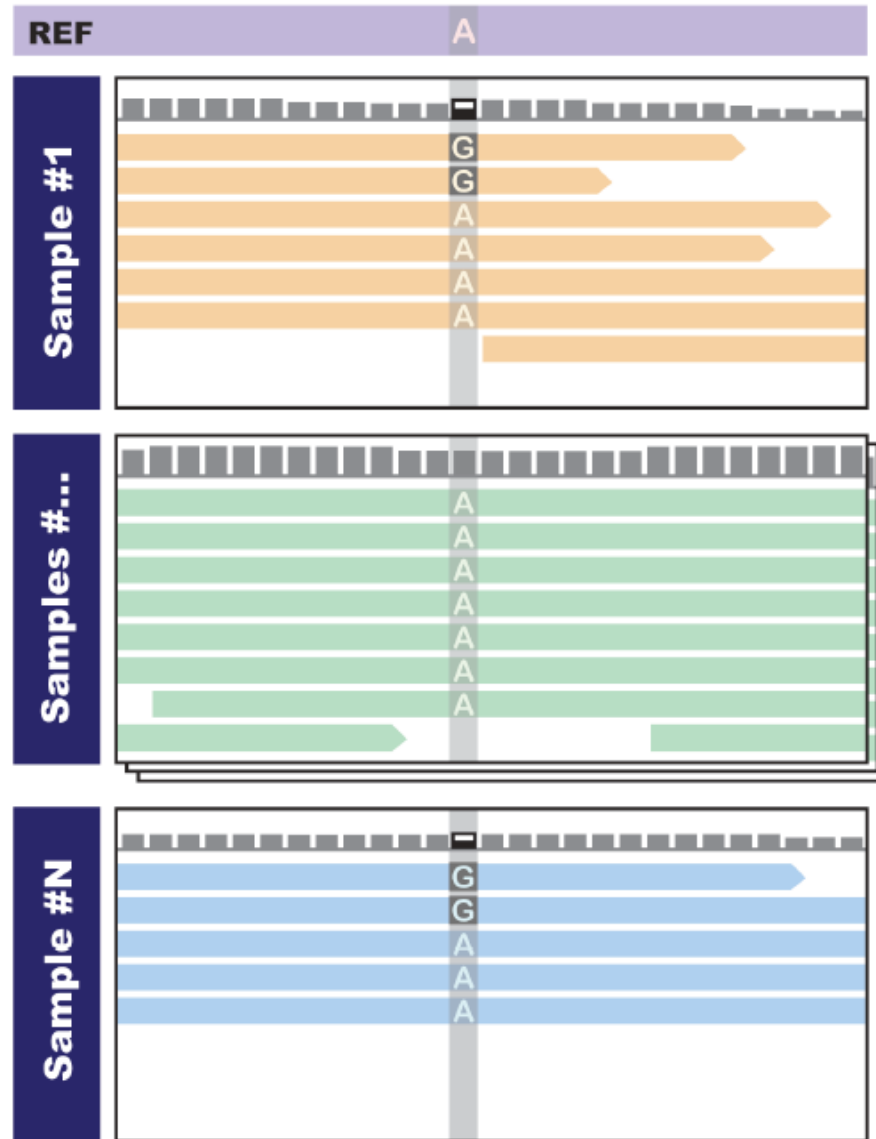
Intersect w/
1 database



G VCF and Joint-Call VCF



Joint Calling



- Sample #1 or Sample #N alone:
 - weak evidence for variant
 - may miss calling the variant
- Both samples seen together:
 - unlikely to be artifact
 - call the variant more confidently



VCF example (QUAL/GQ/HQ)

```
##fileformat=VCFv4.3
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
```

Header

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA00001	NA00002	NA00003
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0 0:48:1:51,51	1 0:48:8:51,51	1/1:43:5:.,.
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017	GT:GQ:DP:HQ	0 0:49:3:58,50	0 1:3:5:65,3	0/0:41:3
20	1110696	rs6040355	A	G,T	67	PASS	NS=2;DP=10;AF=0.333,0.667;AA=T;DB	GT:GQ:DP:HQ	1 2:21:6:23,27	2 1:2:0:18,2	2/2:35:4
20	1230237	.	T	.	47	PASS	NS=3;DP=13;AA=T	GT:GQ:DP:HQ	0 0:54:7:56,60	0 0:48:4:51,51	0/0:61:2
20	1234567	microsat1	GTC	G,GTCT	50	PASS	NS=3;DP=9;AA=G	GT:GQ:DP	0/1:35:4	0/2:17:2	1/1:40:3

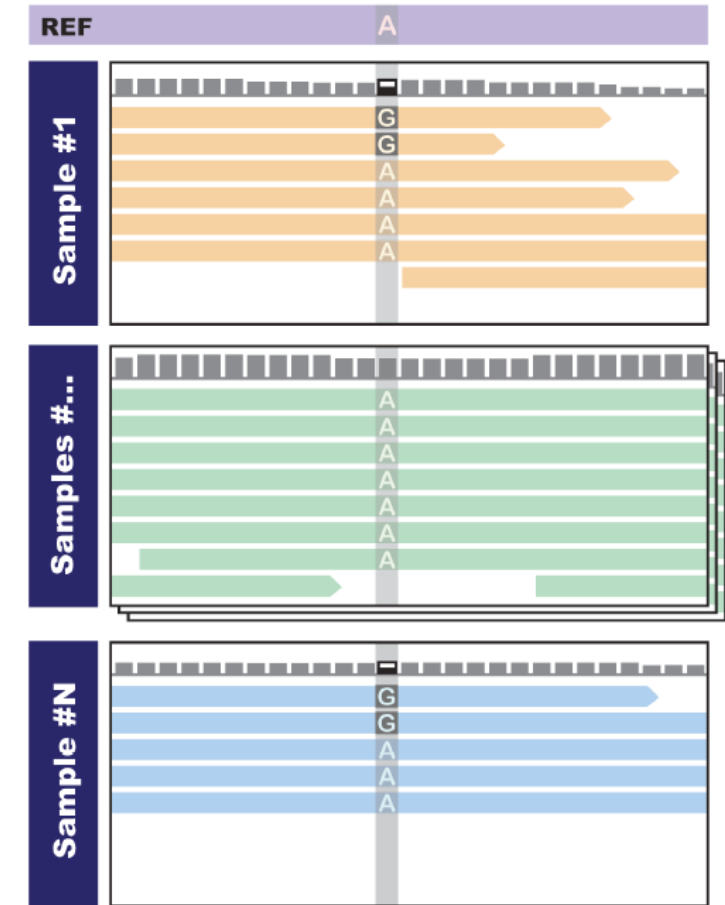
Variant
information

Genotype
information

VCF example (joint calling)

```
##fileformat=VCFv4.3
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sa
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT
20 14370 rs6054257 G A 29 PASS NS=3;DP=14;AF=0.5;DB;H2 GT:GQ:DP:HQ
20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017 GT:GQ:DP:HQ
20 1110696 rs6040355 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ
20 1230237 . T . 47 PASS NS=3;DP=13;AA=T GT:GQ:DP:HQ
20 1234567 microsat1 GTC G,GTCT 50 PASS NS=3;DP=9;AA=G GT:GQ:DP
```

Variant
information



Genotype
information

VCF example (Quality/RefHom)

```
##fileformat=VCFv4.3
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
```

Header

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NA00001	NA00002	NA00003
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2	GT:GQ:DP:HQ	0 0:48:1:51,51	1 0:48:8:51,51	1/1:43:5:.,.
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017	GT:GQ:DP:HQ	0 0:49:3:58,50	0 1:3:5:65,3	0/0:41:3
20	1110696	rs6040355	A	G,T	67	PASS	NS=2;DP=10;AF=0.333,0.667;AA=T;DB	GT:GQ:DP:HQ	1 2:21:6:23,27	2 1:2:0:18,2	2/2:35:4
20	1230237	.	T	.	47	PASS	NS=3;DP=13;AA=T	GT:GQ:DP:HQ	0 0:54:7:56,60	0 0:48:4:51,51	0/0:61:2
20	1234567	microsat1	GTC	G,GTCT	50	PASS	NS=3;DP=9;AA=G	GT:GQ:DP	0/1:35:4	0/2:17:2	1/1:40:3

Variant Quality

Genotype Quality

VCF for a family (FORMAT/INFO)

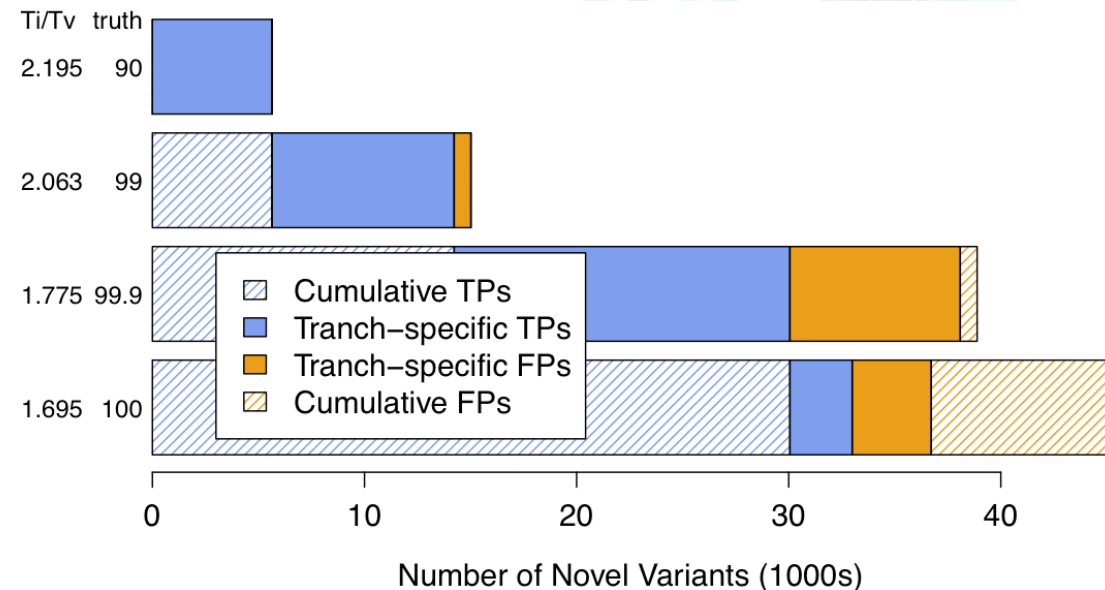
##fileformat=VCFv4.2												
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">												
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read depth">												
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype quality">												
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">												
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="The phred-scaled genotype likelihoods rounded to the closest integer">												
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the same order as listed">												
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele frequency, for each ALT allele, in the same order as listed">												
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">												
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership">												
##INFO=<ID=DP,Number=1,Type=Integer,Description="Combined depth across samples">												
##INFO=<ID=FS,Number=1,Type=Float,Description="Phred-scaled p-value using Fisher's exact test to detect strand bias">												
##INFO=<ID=MQ,Number=1,Type=Float,Description="RMS mapping quality">												
##contig=<ID=chrM,length=16571,assembly=hg19>												
##contig=<ID=chr1,length=249250621,assembly=hg19>												
##contig=<ID=chr20,length=63025520,assembly=hg19>												
##contig=<ID=chr21,length=48129895,assembly=hg19>												
##contig=<ID=chr22,length=51304566,assembly=hg19>												
##contig=<ID=chrX,length=155270560,assembly=hg19>												
##contig=<ID=chrY,length=59373566,assembly=hg19>												
##reference=file:///project/GP1/u3710062/AI_SHARE/reference/GATK_bundle/2.8/hg19/ucsc.hg19.fasta												
#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	Father	Mother	Son	Daughter
chr1	10178	.	G	.	149.69	.	AN=8;DP=89	GT:AD:DP	0/0:22:22	0/0:28:28	0/0:23:23	0/0:16:16
chr1	13178	.	G	A	7854.34	.	AC=1;AF=0.125;AN=8;	GT:AD:DP:GQ:PL	0/0:30,0:30:29:0,29,856	0/0:40,0:40:85:0,85,1221	0/1:0,50:50:60:60,0,1511	0/0:41,0:41:22:0,22,1151
chr20	115365	.	C	T	1089490	.	AC=2;AF=0.25;AN=8;D	GT:AD:DP:GQ:PL	0/0:102,0:102:99:0,120,1800	0/1:12,5:17:92:92,0,299	0/1:14,5:19:85:85,0,361	0/0:60,0:60:42:0,42,1687

Caller specific FORMAT and INFO

Filter variants by FORMAT & INFO columns

QC(GATK) – Call-set refinement (VQSR)

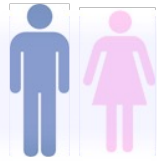
- Variant statistics
 - what the sequence context is like around the variant site
 - how many reads covered it
 - how many reads covered each allele
 - what proportion of reads were in forward vs reverse orientation
- Known variants
 - Hapmap
 - 1000G
 - Omni
 - dbSNP



Genotypes vs. Inheritance Modes

De novo

De novo



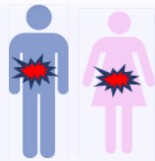
Gene A



Recessively inherited mutations

Unaffected parents

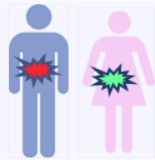
Cloaca patient



Gene A



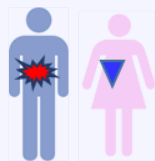
Homozygous



Gene A



Compound heterozygous

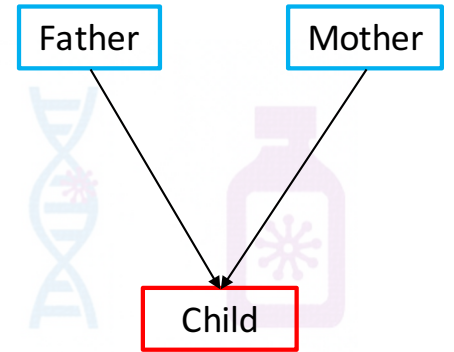


Gene A



Gene B

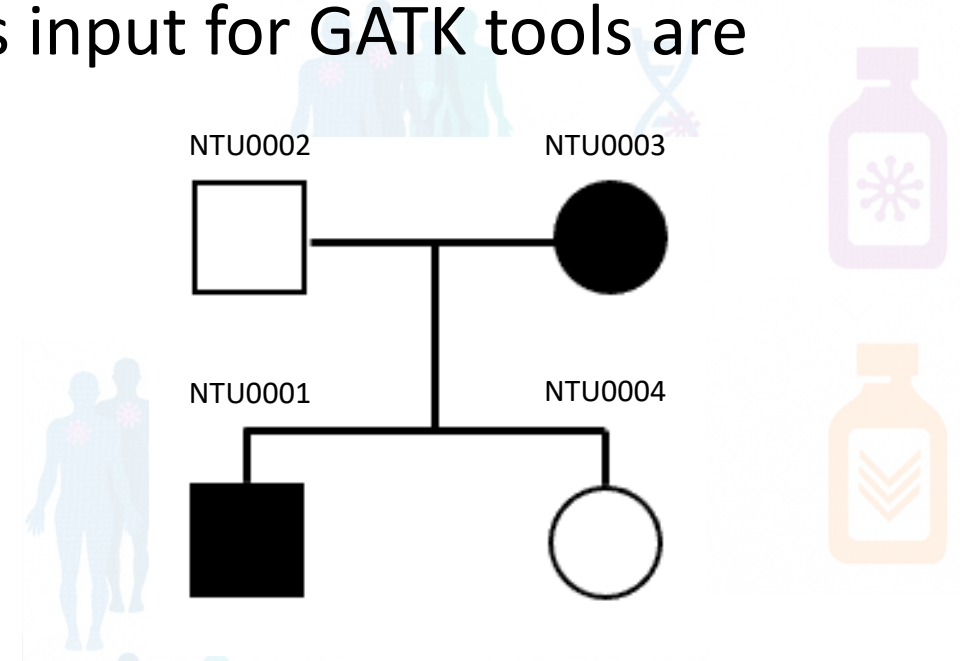
Di-genic



Pedigree format (PED)

- A pedigree is a structured description of the familial relationships between samples. The PED files used as input for GATK tools are based on PLINK pedigree files.

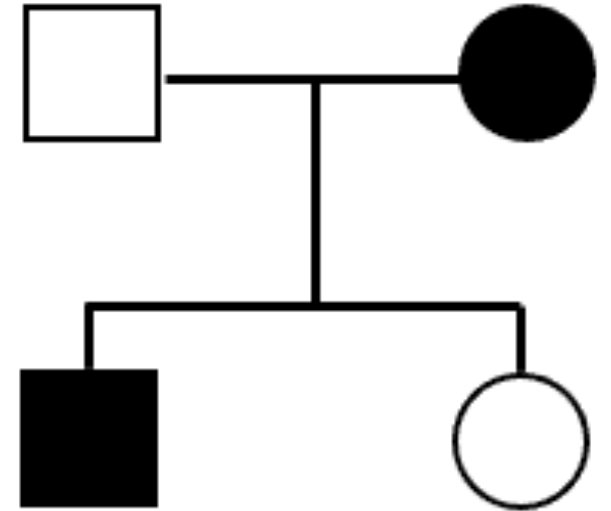
- Family ID
- Individual ID
- Paternal ID
- Maternal ID
- Sex (1=male; 2=female; other=unknown)
- Phenotype
 - 0 missing
 - 1 unaffected
 - 2 affected



FID	IID	PID	MID	Gender	Phe
F7009	NTU0001	NTU0002	NTU0003	1	2
F7009	NTU0002	.	.	1	1
F7009	NTU0003	.	.	2	1
F7009	NTU0004	NTU0002	NTU0003	2	1

Co-segregation (Inheritance mode)

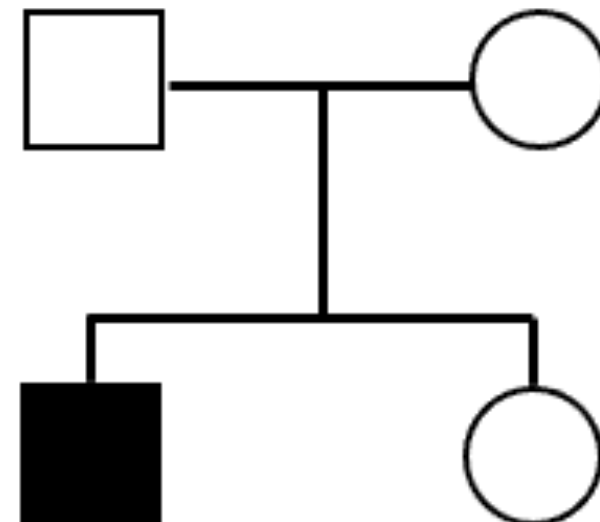
##fileformat=VCFv4.2												
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">												
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read depth">												
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype quality">												
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">												
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="The phred-scaled genotype likelihoods rounded to the closest integer">												
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the same order as listed">												
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele frequency, for each ALT allele, in the same order as listed">												
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">												
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership">												
##INFO=<ID=DP,Number=1,Type=Integer,Description="Combined depth across samples">												
##INFO=<ID=FS,Number=1,Type=Float,Description="Phred-scaled p-value using Fisher's exact test to detect strand bias">												
##INFO=<ID=MQ,Number=1,Type=Float,Description="RMS mapping quality">												
##contig=<ID=chrM,length=16571,assembly=hg19>												
##contig=<ID=chr1,length=249250621,assembly=hg19>												
##contig=<ID=chr20,length=63025520,assembly=hg19>												
##contig=<ID=chr21,length=48129895,assembly=hg19>												
##contig=<ID=chr22,length=51304566,assembly=hg19>												
##contig=<ID=chrX,length=155270560,assembly=hg19>												
##contig=<ID=chrY,length=59373566,assembly=hg19>												
##reference=file:///project/GP1/u3710062/AI_SHARE/reference/GATK_bundle/2.8/hg19/ucsc.hg19.fasta												
#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	Father	Mother	Son	Daughter
chr1	10178	.	G	.	149.69	.	AN=8;DP=89	GT:AD:DP	0/0:22:22	0/0:28:28	0/0:23:23	0/0:16:16
chr1	13178	.	G	A	7854.34	.	AC=1;AF=0.125;AN=8;	GT:AD:DP:GQ:PL	0/0:30,0:30:29:0,29,856	0/0:40,0:40:85:0,85,1221	0/1:0,50:50:60:60,0,1511	0/0:41,0:41:22:0,22,1151
chr20	115365	.	C	T	1089490	.	AC=2;AF=0.25;AN=8;D	GT:AD:DP:GQ:PL	0/0:102,0:102:99:0,120,1800	0/1:12,5:17:92:92,0,299	0/1:14,5:19:85:85,0,361	0/0:60,0:60:42:0,42,1687



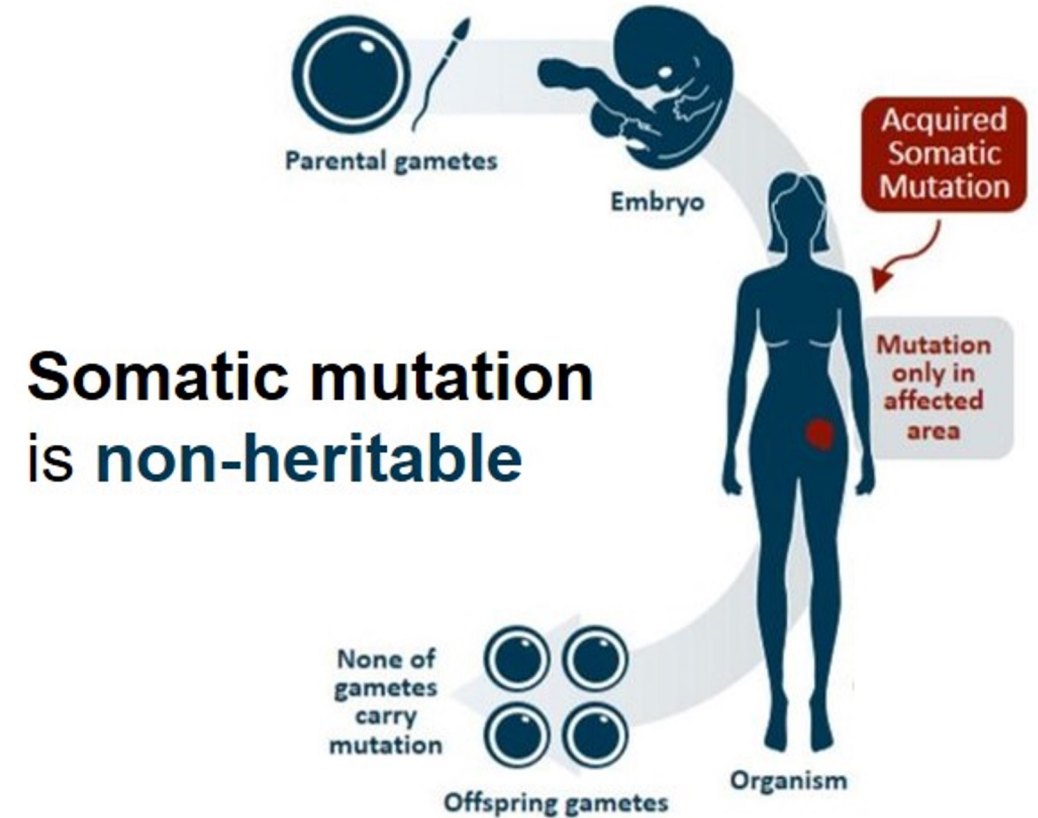
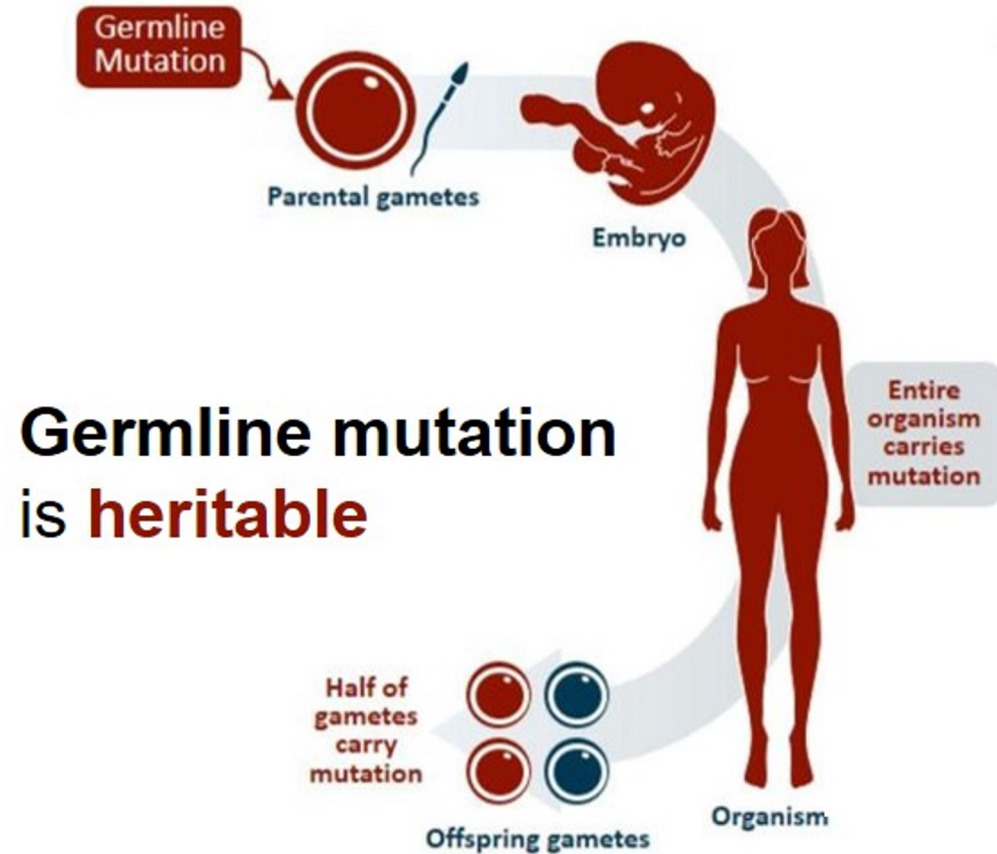
De novo mutation

```
##fileformat=VCFv4.2
##FORMAT=<ID=AD,Number=.,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read depth">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="The phred-scaled genotype likelihoods rounded to the closest integer">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes, for each ALT allele, in the same order as listed">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele frequency, for each ALT allele, in the same order as listed">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Combined depth across samples">
##INFO=<ID=FS,Number=1,Type=Float,Description="Phred-scaled p-value using Fisher's exact test to detect strand bias">
##INFO=<ID=MQ,Number=1,Type=Float,Description="RMS mapping quality">
##contig=<ID=chrM,length=16571,assembly=hg19>
##contig=<ID=chr1,length=249250621,assembly=hg19>
##contig=<ID=chr20,length=63025520,assembly=hg19>
##contig=<ID=chr21,length=48129895,assembly=hg19>
##contig=<ID=chr22,length=51304566,assembly=hg19>
##contig=<ID=chrX,length=155270560,assembly=hg19>
##contig=<ID=chrY,length=59373566,assembly=hg19>
##reference=file:///project/GP1/u3710062/AI_SHARE/reference/GATK_bundle/2.8/hg19/ucsc.hg19.fasta
```

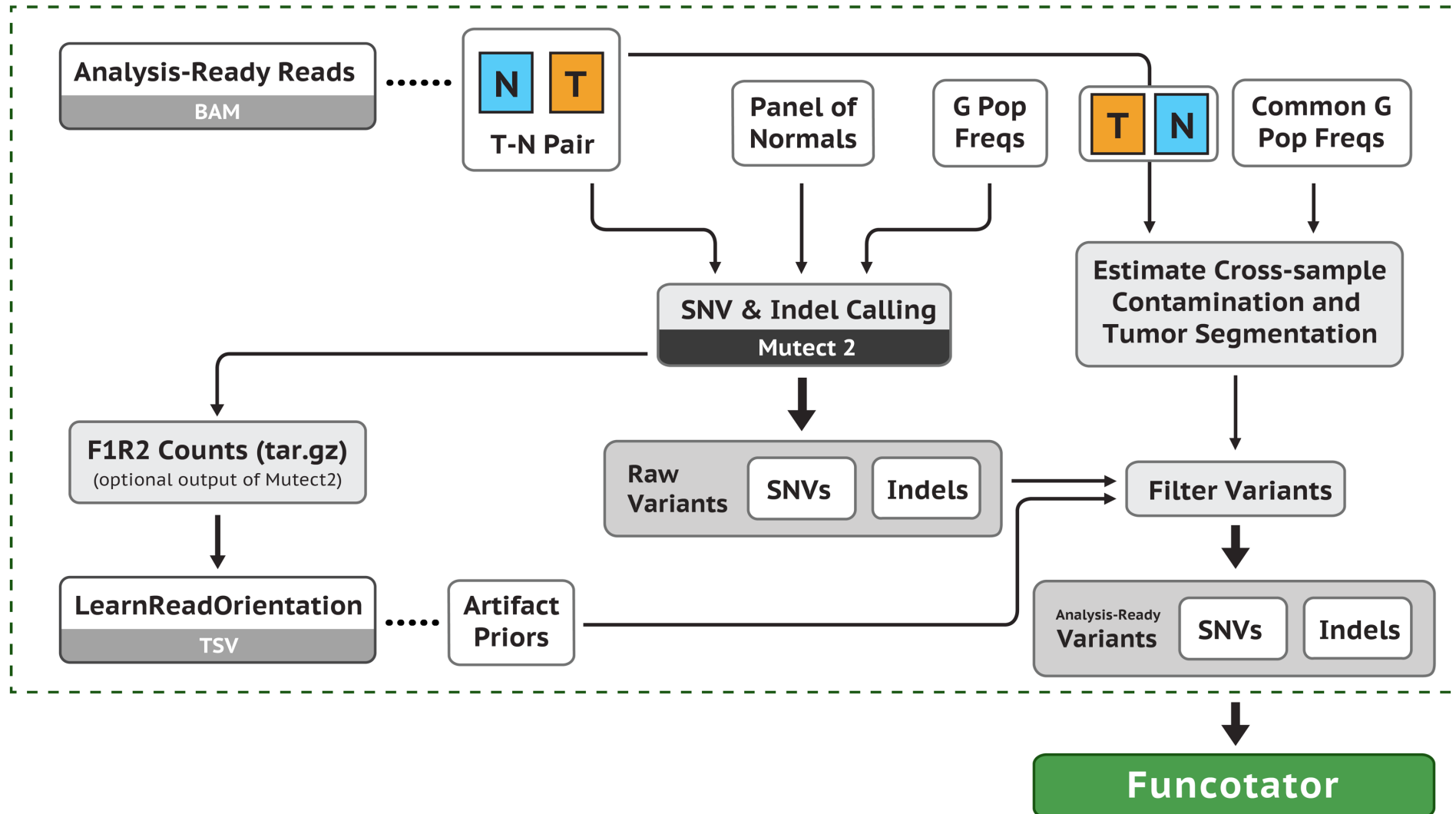
#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	Father	Mother	Son	Daughter
chr1	10178	.	G	.	149.69	.	AN=8;DP=89	GT:AD:DP	0/0:22:22	0/0:28:28	0/0:23:23	0/0:16:16
chr1	13178	.	G	A	7854.34	.	AC=1;AF=0.125;AN=8;	GT:AD:DP:GQ:PL	0/0:30,0:30:29:0,29,856	0/0:40,0:40:85:0,85,1221	<u>0/1:0,50:50:60:60,0,1511</u>	0/0:41,0:41:22:0,22,1151
chr20	115365	.	C	T	1089490	.	AC=2;AF=0.25;AN=8;D	GT:AD:DP:GQ:PL	0/0:102,0:102:99:0,120,1800	0/1:12,5:17:92:92,0,299	0/1:14,5:19:85:85,0,361	0/0:60,0:60:42:0,42,1687



Germline and Somatic DNA Variant

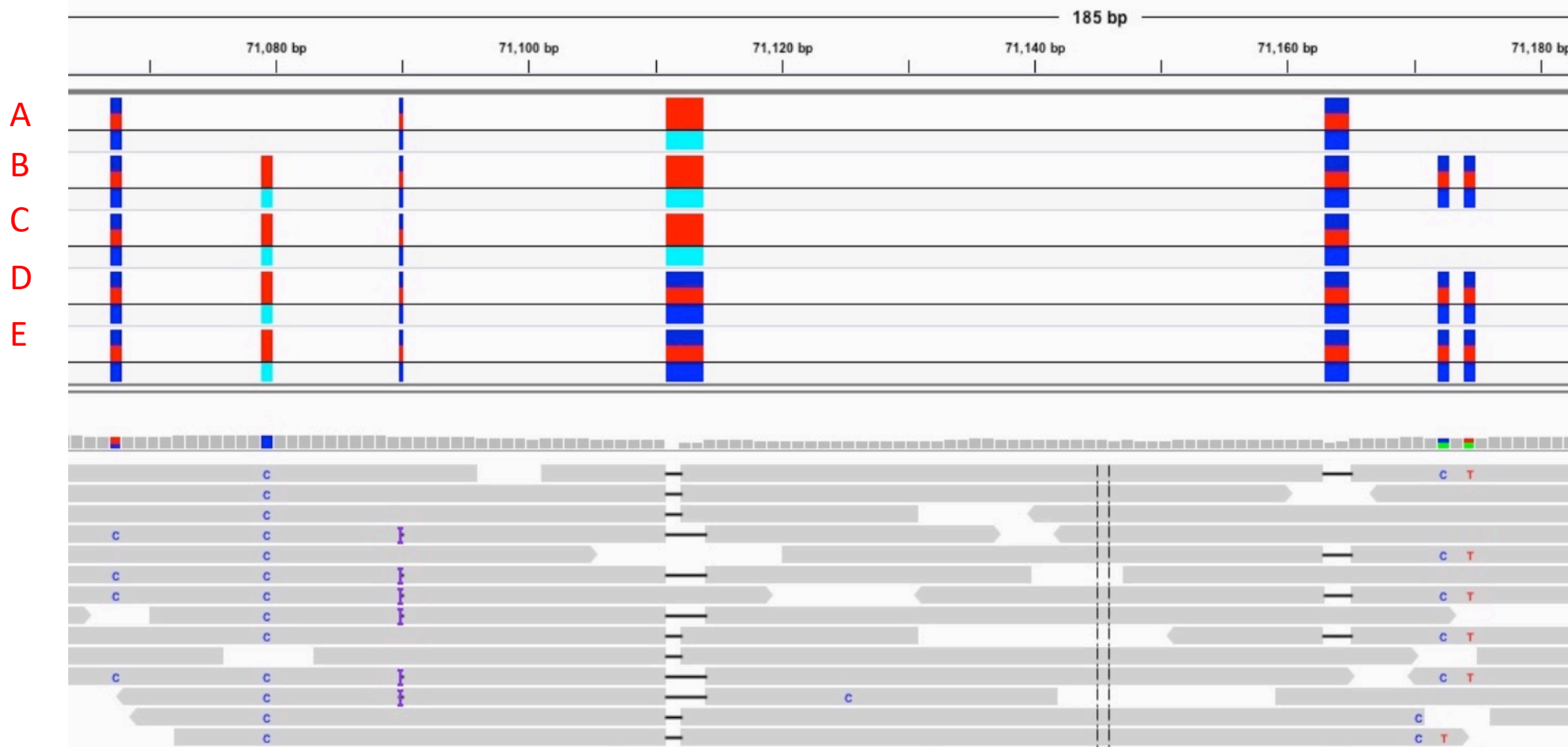


Mutect2 (variable ploidy assumption)



SNV/INDEL Benchmarking & Other variant types

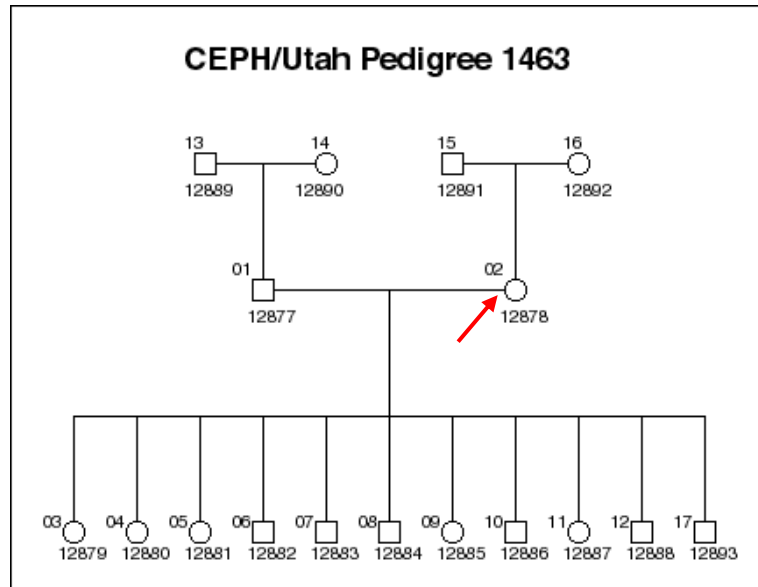
Variants Discovery (GATK)



Samtools_O
Samtools_B
Samtools_E
Unifygenotyper
Haplotyper

Genome in a Bottle (GIAB)

- National Institute of Standards and Technology
- Benchmark human genomes (n=7, **hs37d5**)
 - NA12878/HG001
 - A trio of Ashkenazi Jewish ancestry
 - A trio of chinese ancestry



Variant calling accuracy (confusion matrix)

• NA12878 (NIST v3.3.2)

- DRAGEN
 - Sentieon
 - Haplotypemapper
 - CLC
 - Mutect2
 - Varscan
- (parameters)

		Actual	
		Positive	Negative
Predicted	Positive	TP (True Positive)	FP (False Positive)
	Negative	FN (False Negative)	TN (True Negative)

Precision

$$\frac{TP}{(TP + FP)}$$

Recall

$$\frac{TP}{(TP + FN)}$$

(Sensitivity)

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{(\text{Precision} + \text{Recall})}$$

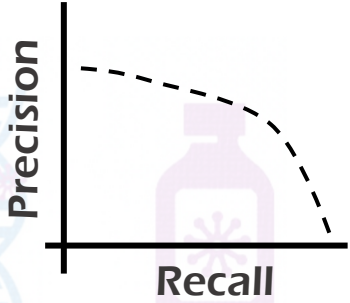
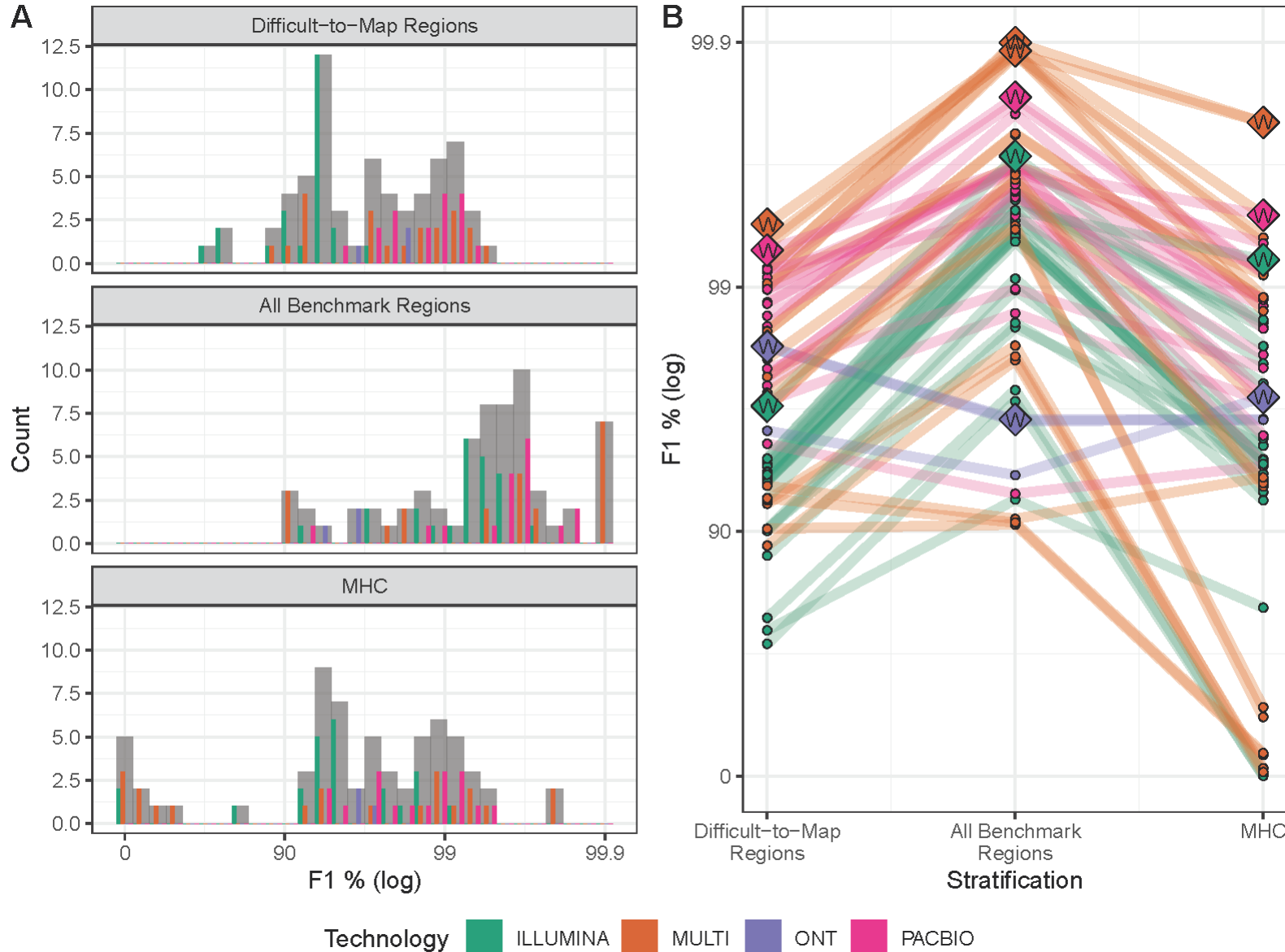


Table 2 | Examples of several combinations of truth and query SNV genotypes and how they are counted as TP, FP, FN, FP.GT, and FP.AL

REF	Truth	Query	Counted as
A	C/C	C/C	1 TP
A	A/A	C/C	1 FP
A	C/C	A/A	1 FN

PrecisionFDA V2 (GRCh38)



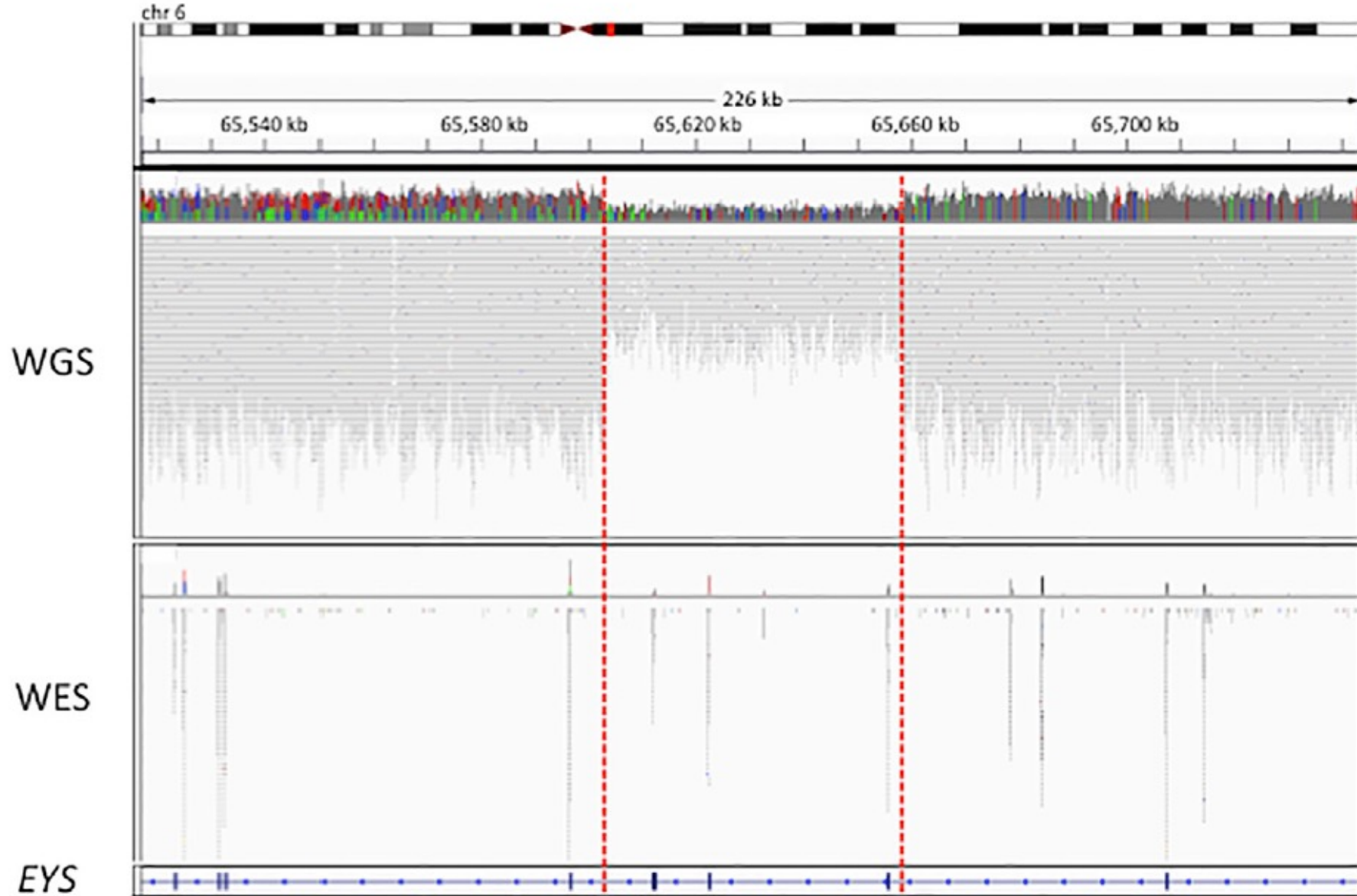
- Benchmark -> MHC -> Difficult
- MULTI -> PACBIO -> Illumina > ONT
- Algorithms matter !!!

PrecisionFDA V2 (GRCh38)

BEST PERFORMANCE Difficult-to-Map Regions – PACBIO in the precisionFDA Truth Challenge V2  AWARDED TO Sentieon PacBio HIFI only submission	BEST PERFORMANCE All Benchmark Regions – PACBIO in the precisionFDA Truth Challenge V2  AWARDED TO The Genomics Team in Google Health DeepVariant PacBio	BEST PERFORMANCE MHC – PACBIO in the precisionFDA Truth Challenge V2  AWARDED TO Sentieon PacBio HIFI only submission
BEST PERFORMANCE Difficult-to-Map Regions – ONT in the precisionFDA Truth Challenge V2  AWARDED TO UCSC CGL and Google Health PEPPER-DeepVariant (ONT_9to6)	BEST PERFORMANCE All Benchmark Regions – ONT in the precisionFDA Truth Challenge V2  AWARDED TO UCSC CGL and Google Health PEPPER-DeepVariant (ONT_9to6)	BEST PERFORMANCE MHC – ONT in the precisionFDA Truth Challenge V2  AWARDED TO Wang Genomics Lab ONT NanoCaller Clair Medaka Ensemble
BEST PERFORMANCE Difficult-to-Map Regions – ILLUMINA in the precisionFDA Truth Challenge V2  AWARDED TO DRAGEN team at Illumina DRAGEN Experimental Extension into Difficult Regions - Illumina reads	BEST PERFORMANCE All Benchmark Regions – ILLUMINA in the precisionFDA Truth Challenge V2  AWARDED TO DRAGEN team at Illumina DRAGEN Experimental Extension into Difficult Regions - Illumina reads	BEST PERFORMANCE MHC – ILLUMINA in the precisionFDA Truth Challenge V2  AWARDED TO Seven Bridges Genomics Seven Bridges GRAF - Illumina

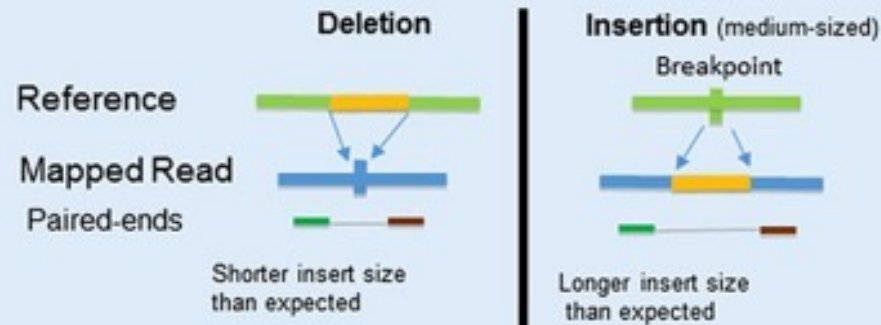
BEST PERFORMANCE Difficult-to-Map Regions – MULTI in the precisionFDA Truth Challenge V2  AWARDED TO Roche Sequencing Solutions RN-Illumina-PacBio-ONT	BEST PERFORMANCE All Benchmark Regions – MULTI in the precisionFDA Truth Challenge V2  AWARDED TO Sentieon Combination of Illumina, PacBio HIFI, and Oxford Nanopore submission	BEST PERFORMANCE MHC – MULTI in the precisionFDA Truth Challenge V2  AWARDED TO Sentieon Combination of Illumina, PacBio HIFI, and Oxford Nanopore submission
TIED		
BEST PERFORMANCE All Benchmark Regions – MULTI in the precisionFDA Truth Challenge V2  AWARDED TO Roche Sequencing Solutions RN-Illumina-PacBio-ONT		
TIED		
BEST PERFORMANCE All Benchmark Regions – MULTI in the precisionFDA Truth Challenge V2  AWARDED TO The Genomics Team in Google Health DeepVariant Hybrid PacBio-Illumina		

Deletion can be determine by coverage

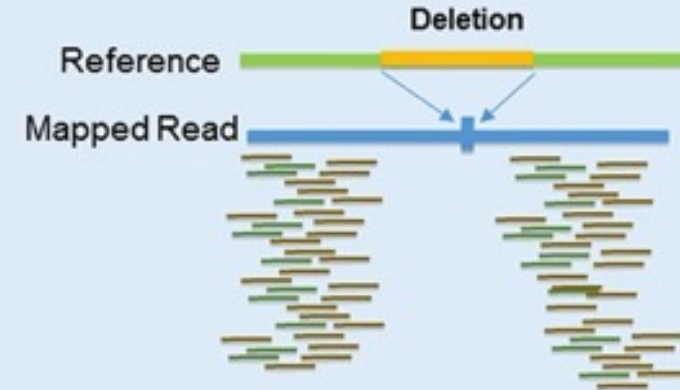


Algorithms to detect SV

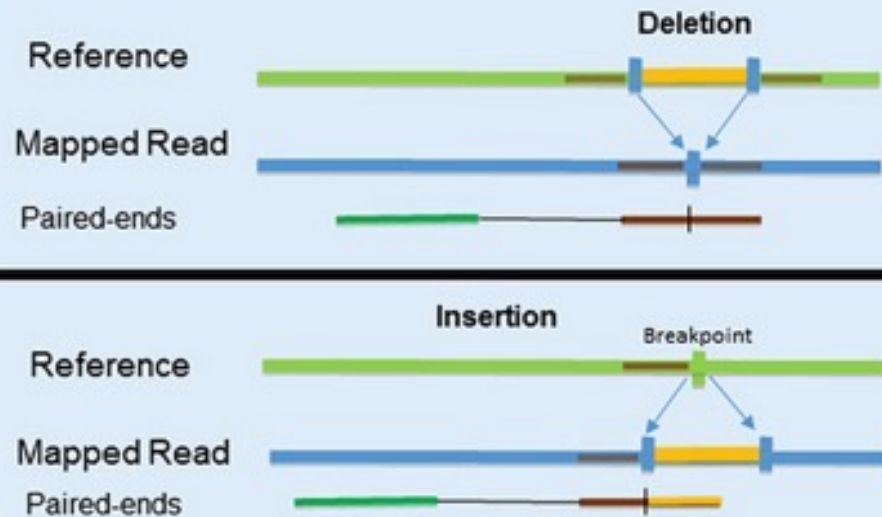
1. Read-Pair (RP)



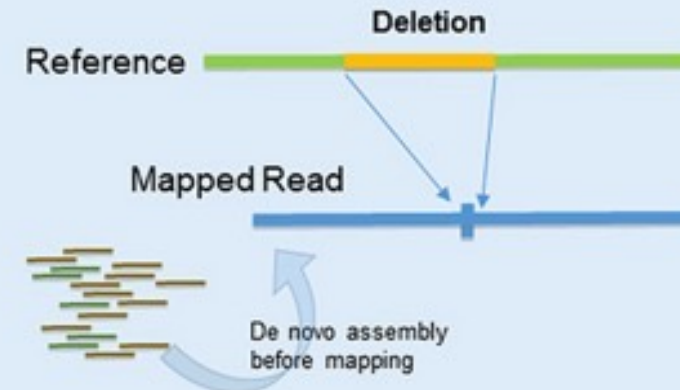
3. Read-depth (RD)



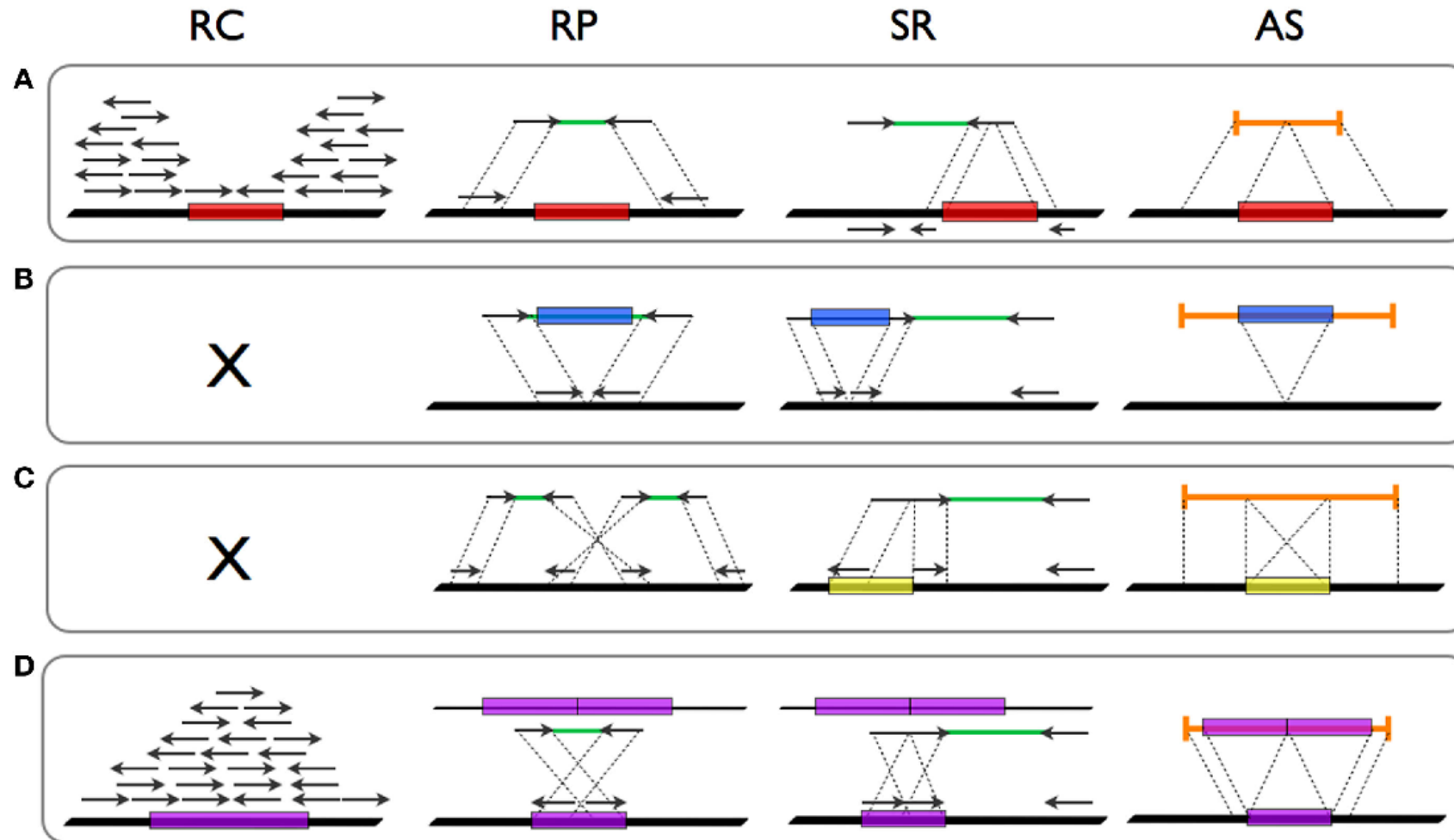
2. Split-read (SR)



4. Assembly (AS)



Algorithms to detect SV (cont'd)



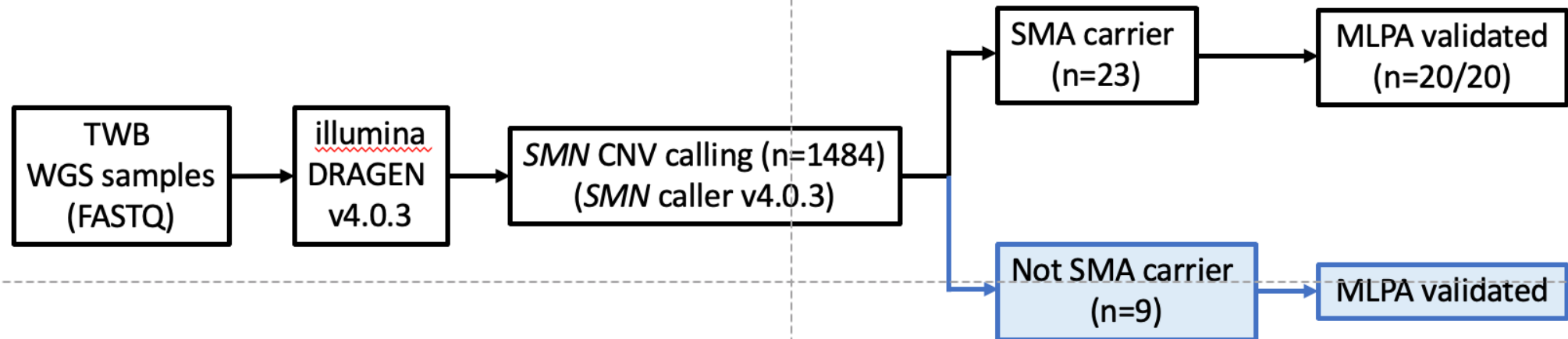
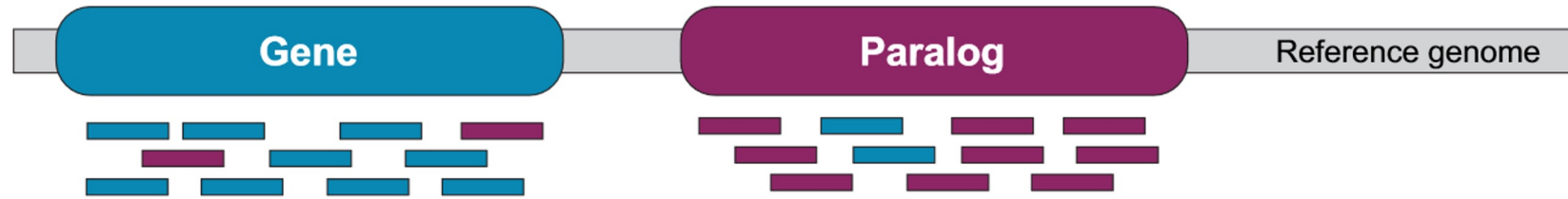
DRAGEN targeted callers and performance summary

Targeted caller	Performance summary	Source
SMN	•99.8% of <i>SMN1</i> and 99.7% of <i>SMN2</i> copy number calls agreed with orthogonal methods (MLPA/droplet digital PCR) •Recall of 100% for spinal muscular atrophy and 97.8% for carriers and a precision of 100% for both spinal muscular atrophy and carriers	Chen et al. ²⁷
CYP2D6	•99.3% concordance of star allele calls against orthogonal calls in a cohort of samples characterized by Get-RM	Chen et al. ²⁶
CYP2B6	•70/76 (92.1%) concordant star allele calls against orthogonal characterization results in the Get-RM cohort •123/125 (98.4%) concordant star allele calls against manually curated star allele calls from long-read technology (PacBio)	Mukherjee et al. ⁷⁶
GBA	•100% concordance with targeted long-read-based method (ONT) across 42 samples with diverse variant types •Significant enrichment of pathogenic <i>GBA</i> calls detected in Parkinson's disease relative to the control cohort	Toffoli et al. ²⁸
CYP21A2	•100% recall of pathogenic variants detected by the combination of MLPA and Sanger sequencing across the cohort of 16 clinical samples and four cell lines •98.5% (201/204) concordance in total <i>RCCX</i> locus copy numbers against calls made by Bionano optical mapping	Belyeu et al. ⁷⁷

HBA	•100% (222/222) concordance with manually curated calls from orthogonal long-read technology (PacBio) across all major copy number genotypes •1.6% (4/226) no call rate across the validation cohort	Han et al. ⁷⁸
LPA	•98.2% correlation between DRAGEN-estimated KIV-2 total copy number and KIV-2 copy number extracted from Bionano optical mapping across 154 samples •99.7% correlation between allele-specific KIV-2 copy number and Bionano-based KIV-2 allele-specific copy number estimates	Behera et al. ²⁹
HLA	•97.91% concordance in <i>HLA</i> alleles detected across 1,102 samples compared to results generated by Sanger sequencing •Comparable two-field accuracy results to OptiType⁷⁵ against their own truth set for 164 samples	Mukherjee et al. ⁷⁶
RHD/RHCE	•The <i>RHD/RHCE</i> exon 2 gene conversion caller was evaluated by comparing the calls of 252 1kGP samples to long-read assembly dipcall calls, resulting in a 96.82% concordance rate	NA

Spinal Muscular dystrophy *SMN1/SMN2*

Challenge: Misalignment of reads derived from genes with highly similar copies



Carrier frequency of SMA in TWB cohort
= 23/1480
= 1.6% (1/64)

Estimation of α -thalassemia carrier rate in Taiwan



- 5.1 % of α -thalassemia carrier rate

GENE	Disease	Alpha-Thalassemia type	CHROM:POS	LENGTH	IMPACT	# of HET	# of HOM	AF	Carrier frequency	
									TWB	Westemeyer et al.
HBA1/HBA2	Alpha-Thalassemia	--THAI	chr16:149861	-33449	HIGH	2	0	0.00068	0.04939	0.083333
		--FIL	chr16:151659	-30617	HIGH	4	0	0.00135		
		--SEA	chr16:165393	-19307	HIGH	60	0	0.02027		
		- $\alpha^{4.2}$	chr16:170204	-4412	HIGH	1	0	0.00034		
		- $\alpha^{3.7}$	chr16:173557	-3799	HIGH	8	1	0.00338		

Prevalence of NOTCH2NLC and FMR1 Repeat Expansions in Atypical Parkinsonism Compared to Asymptomatic Elderly Individuals

TABLE 1 Demographics and prevalence of NOTCH2NLC and FMR1 repeat expansions in patients with atypical parkinsonism and asymptomatic individuals over 60 years old

	Community recruited asymptomatic elderly (n = 341)	Patients with atypical parkinsonism (n = 252)			P value
		MSA (n = 165)	PSP (n = 58)	CBS (n = 29)	
Sex (male, %)	192 (56.30)	92 (55.76)	36 (61.02)	10 (34.48)	0.10
Age at enrollment (y)	64.12 ± 2.92	68.55 ± 8.19	73.56 ± 6.72	73.72 ± 7.34	<0.01 ^{★★}
Onset age (cases)	N.A.	60.49 ± 7.65	66.03 ± 6.88	66.86 ± 8.52	–
Family history, n (%)	N.A.	21 (12.73)	7 (12.07)	1 (3.45)	–
<i>NOTCH2NLC</i> repeat number					
Normal (7–39)	337 (98.83%)	160 (96.97%)	56 (94.92%)	29 (100%)	0.29
Intermediate (41–59)	2 (0.59%)	1 (0.61%)	0	0	1
Pathogenic (≥ 60)	2 (0.59%)	4 (2.42%) ^a	2 (3.39%) ^b	0	0.07
<i>FMR1</i> repeat number					
Normal (6–44)	339 (99.41%)	163 (98.78%)	57 (98.28%)	29 (100%)	0.60
Intermediate (45–54)	1 (0.29%)	2 (1.21%)	1 (1.79%)	0	0.29
Premutation (55–200)	1 (0.29%)	0	0	0	1

Note: Family history, defined as having family members with parkinsonism or dementia within three generations.

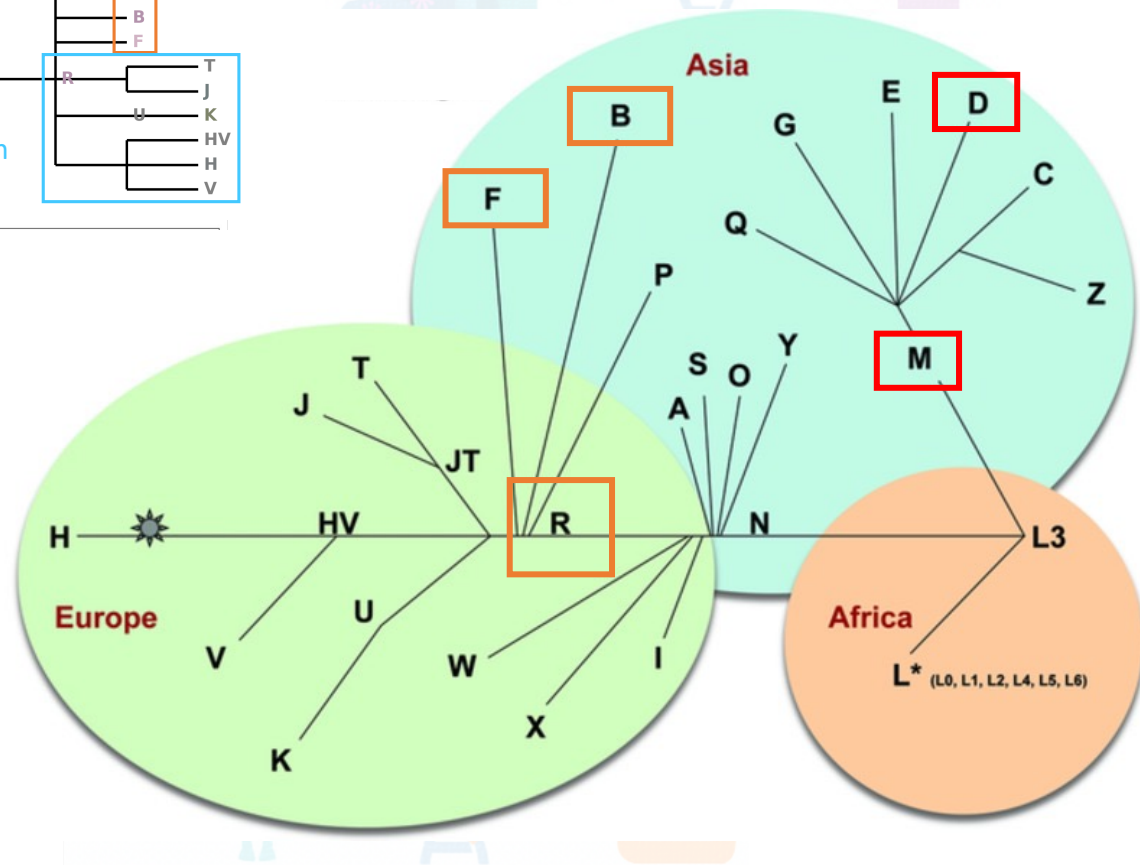
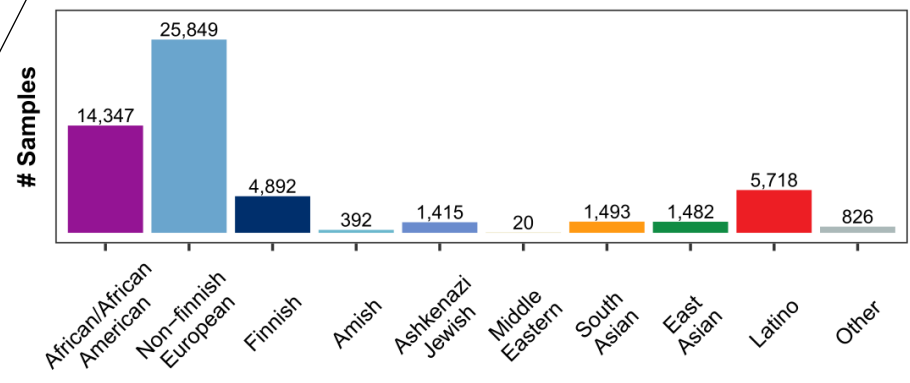
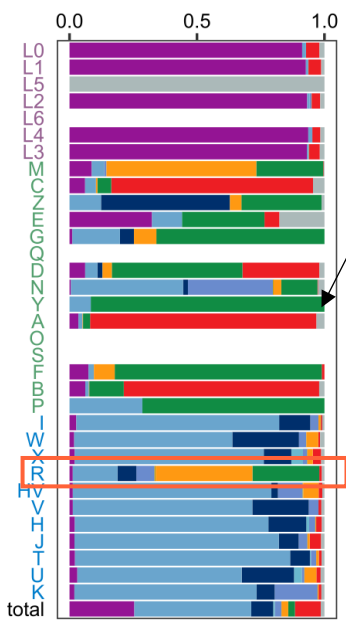
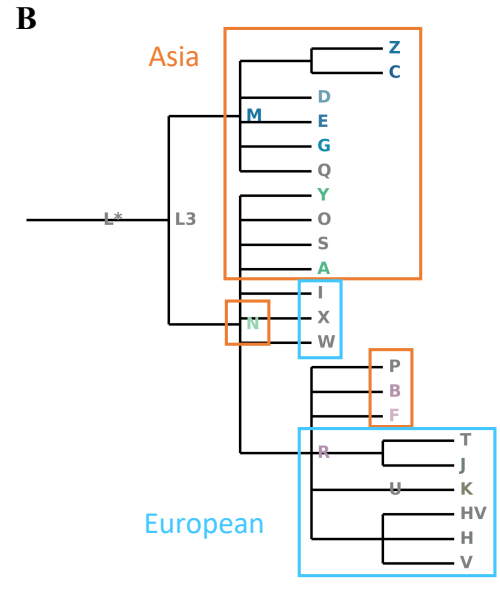
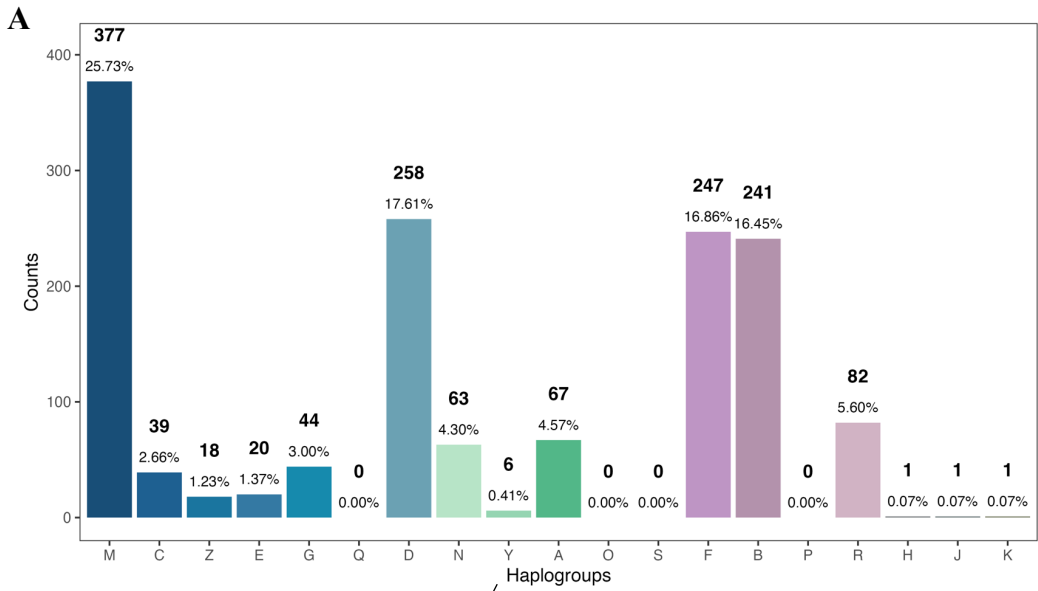
Abbreviations: MSA, multiple system atrophy; PSP, progressive supranuclear palsy; CBS, corticobasal syndrome; N.A., not available.

^aAsymptomatic individuals versus MSA: *P* = 0.07.

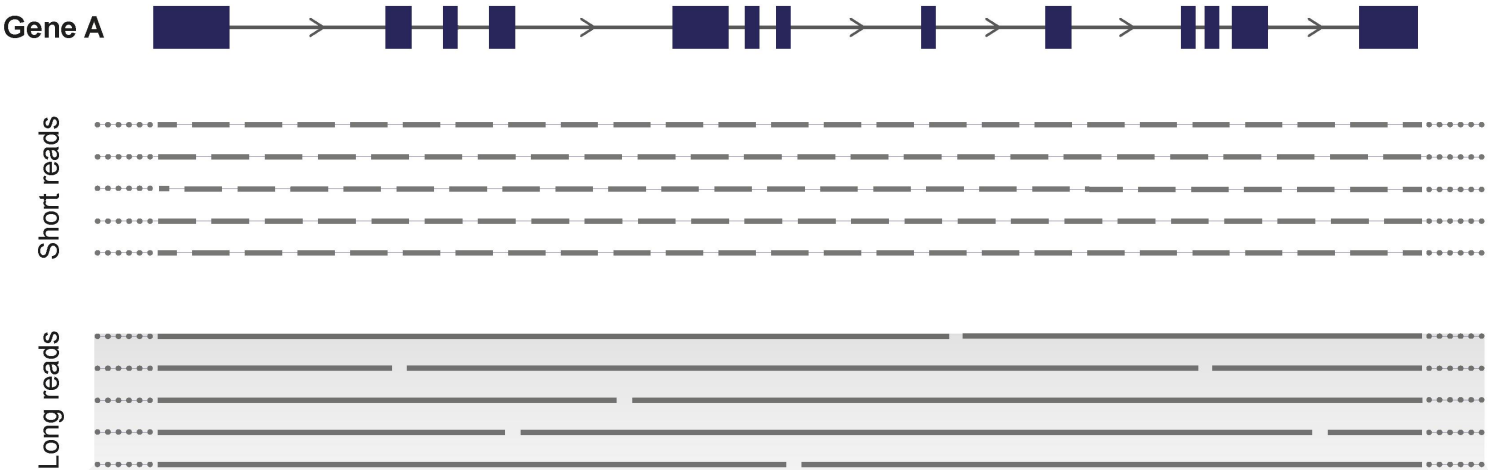
^bAsymptomatic individuals versus PSP: *P* = 0.09.

★★*P* < 0.01





Advantages of Long-Read Sequencing



1. Structural variation
e.g. *PRKAR1A*, *G6PC*, *BBS9*, *ARGHEF9*, *TAF1*

Ref. Gene A

A 919D

Ref. Gene A

Gene A

2. Repeat expansion
e.g. *FMR1*, *DMPK*, *ATXN10*, *HTT*

Ref. CGG CGG n~50

CGG CGG CGG CGG n>200

3. Phasing
e.g. Compound heterozygosity, Parental origin of de novo mutations, Mosaicism

mat. Gene A *

pat. Gene A *

/

mat. Gene A

pat. Gene A *

4. Pseudogenes
e.g. *PMS2*, *CYP2D6*, *CHEK2*, *SMN1*, *PKD1*

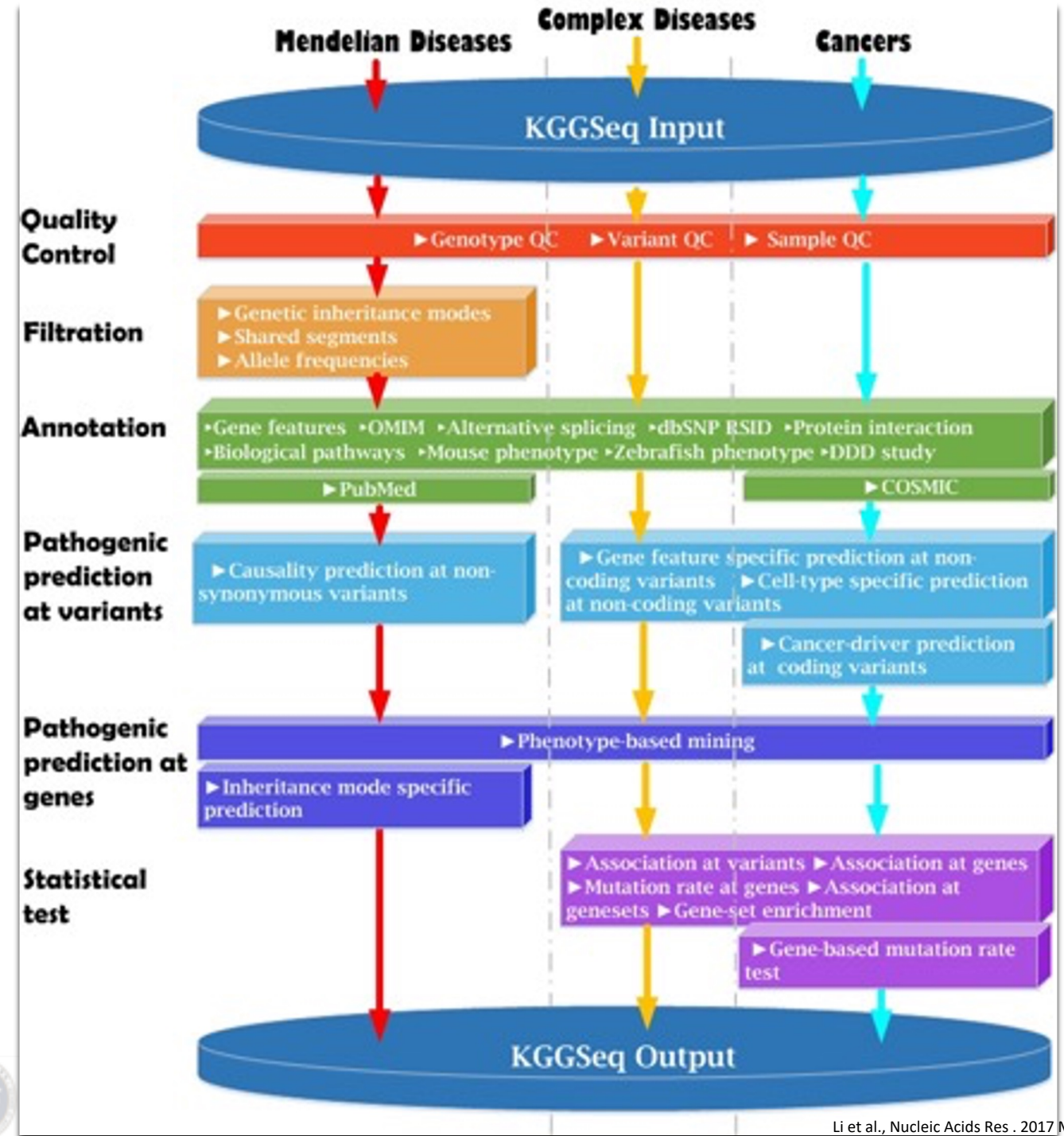
Ref. Gene A

* Gene A'

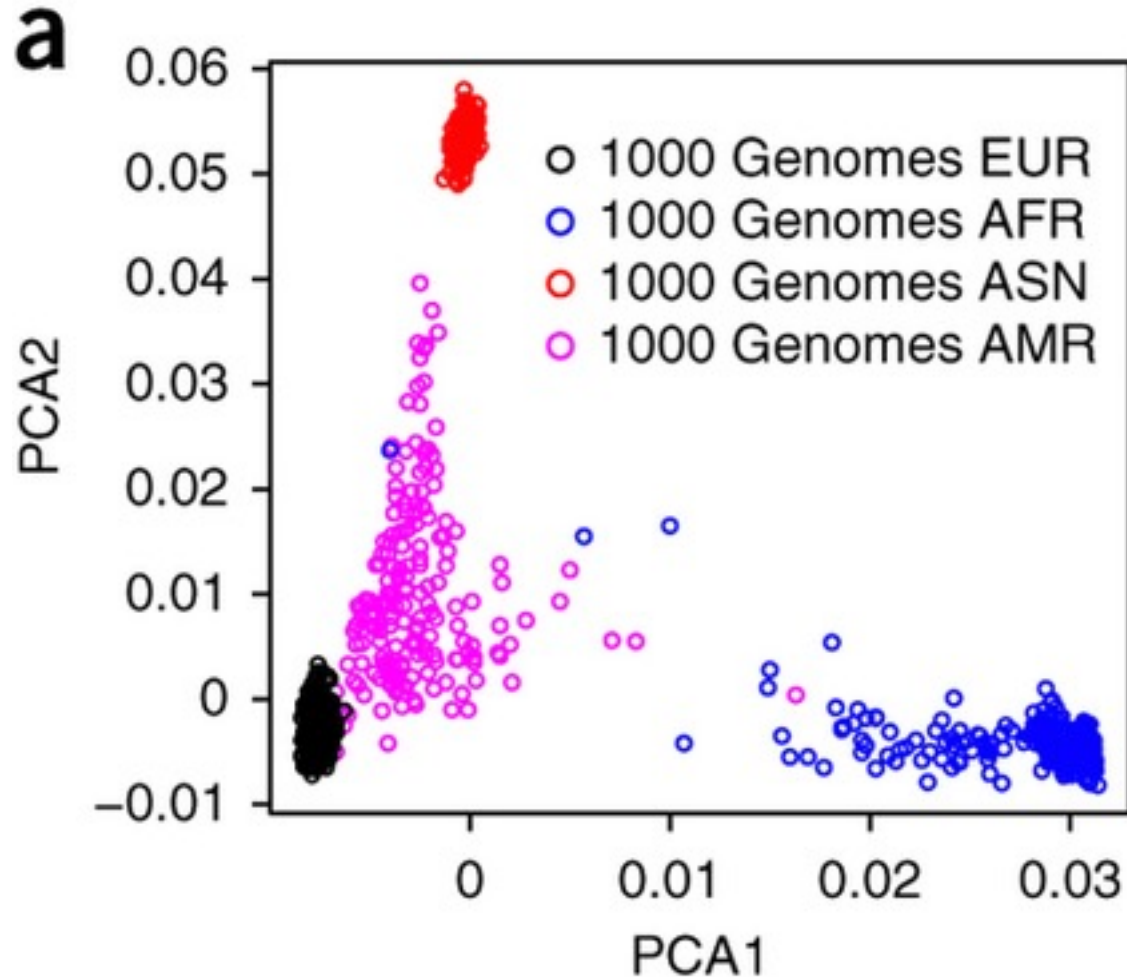


Variant Filtering

- Genotype/Variant QC
 - Overall coverage
 - On-target coverage
 - Sequence & Mapping Quality
- Sample QC
 - Gender check
 - Relatedness check
 - Population stratification



QC(plink) – Gender/Relatedness/Population check

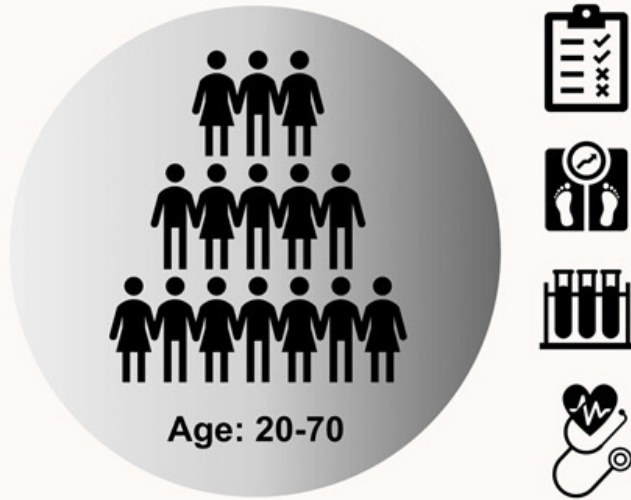


Number of Subjects		Caseness	
		Controls	Cases
Country	S	500	50
	I	500	950

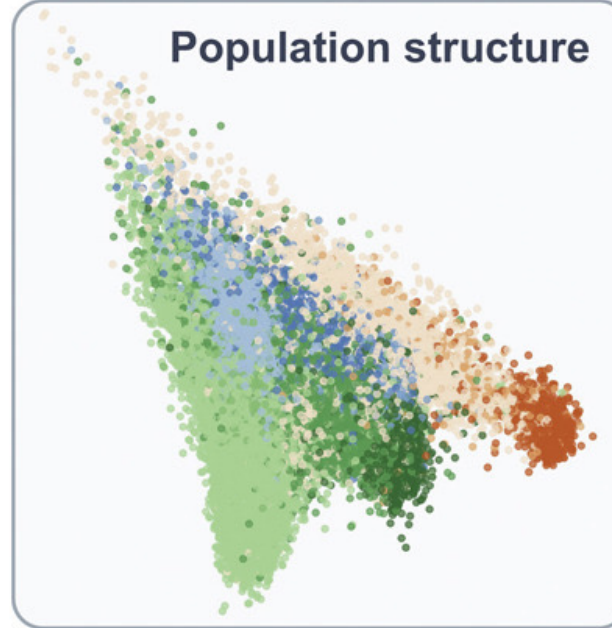
- Ethnicity predicts cases
- So does any genetic difference between country S and I
- Population Stratification is the confounding (association) of disease with ethnicity
- Any markers which correlates with ethnicity will also correlate with disease status

Taiwan Biobank (TWB)

Community-based phenotyping



Population structure



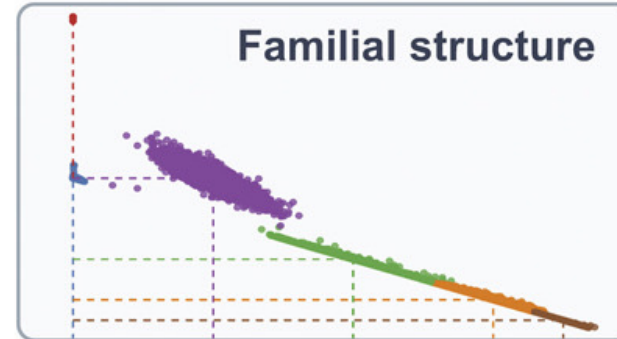
Registry data linkage



Omics data generation



Familial structure



2012

Recruitment began

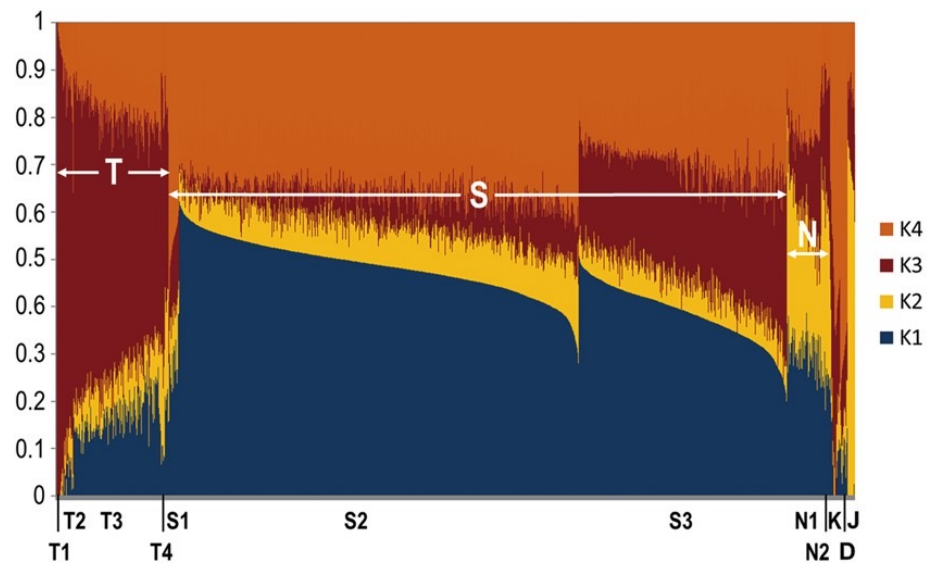
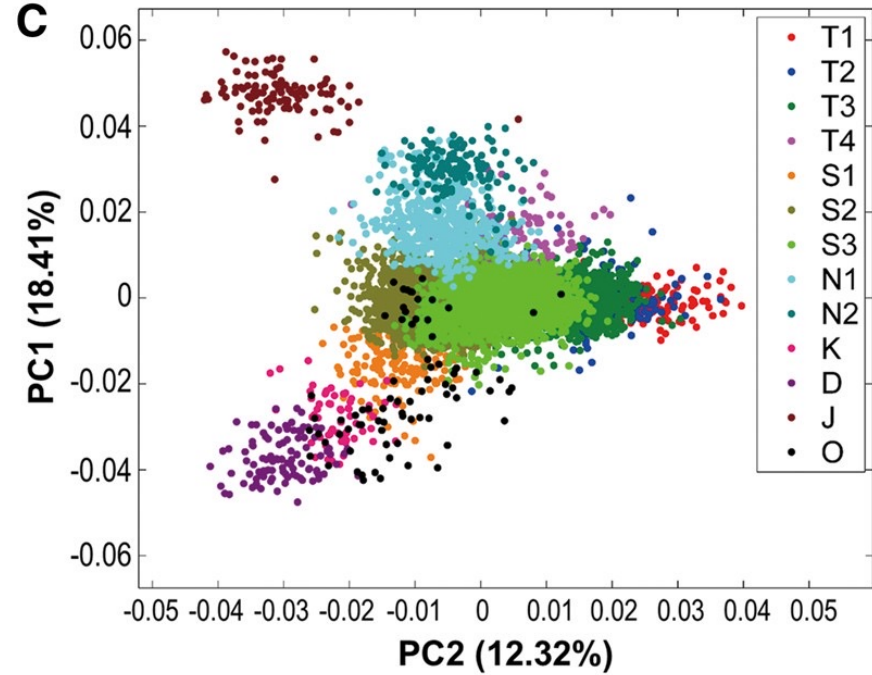
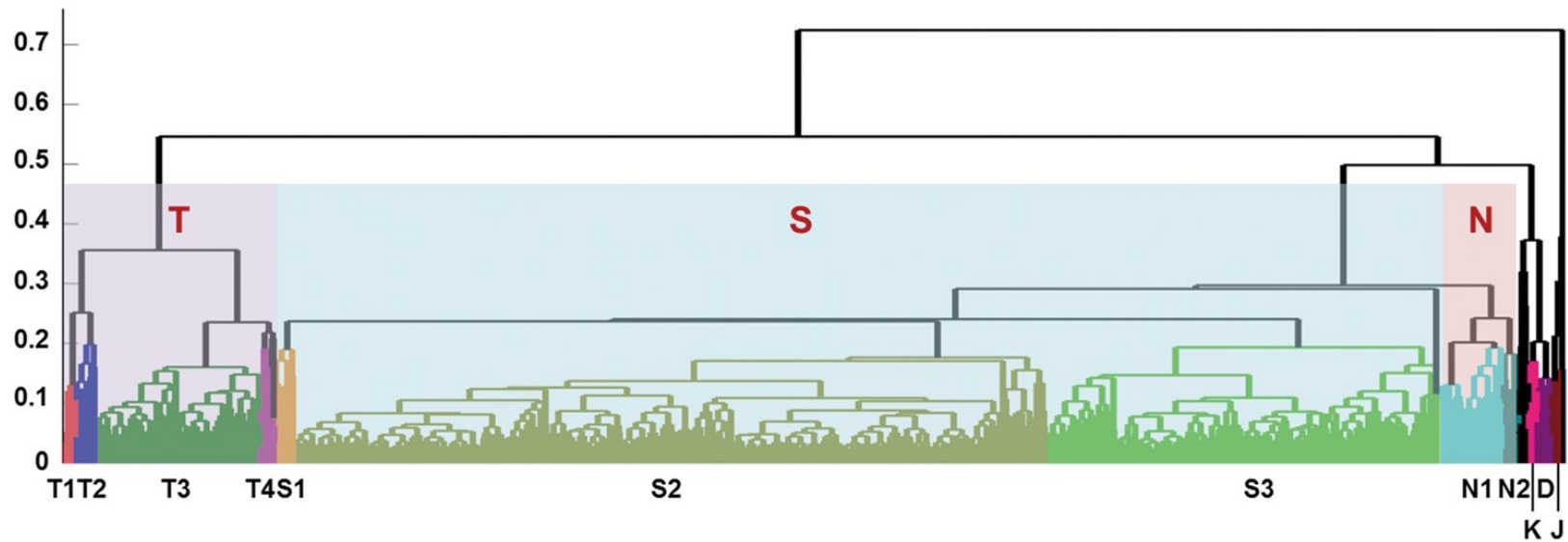
Aug. 2021

N ≈ 150,000

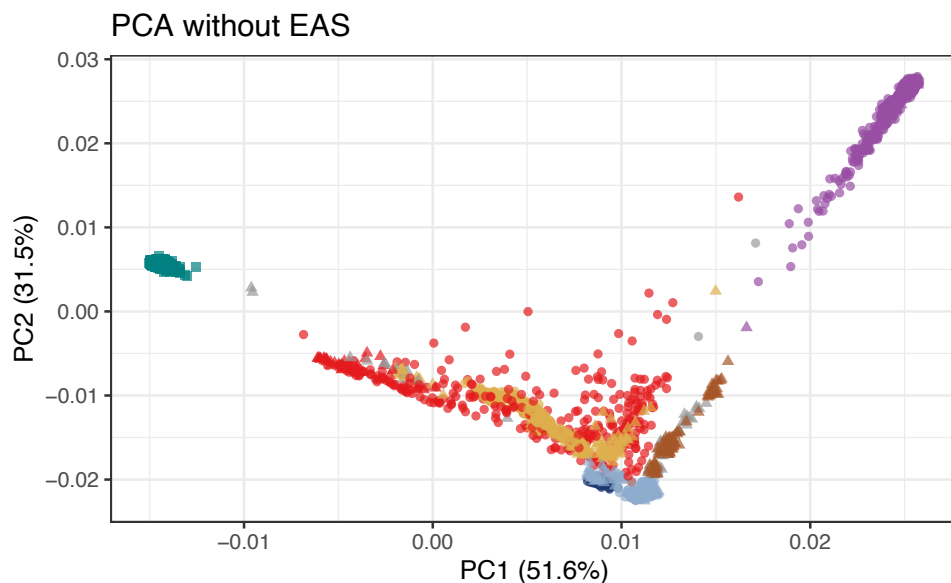
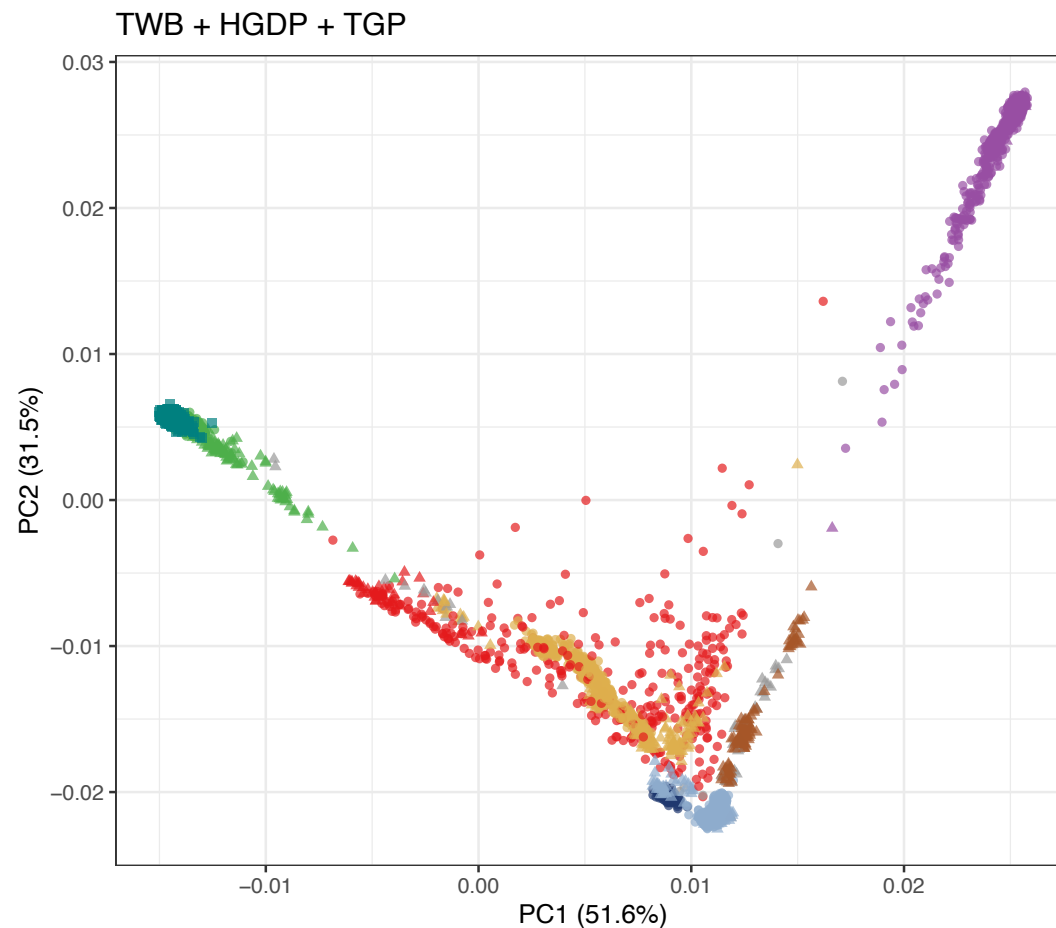
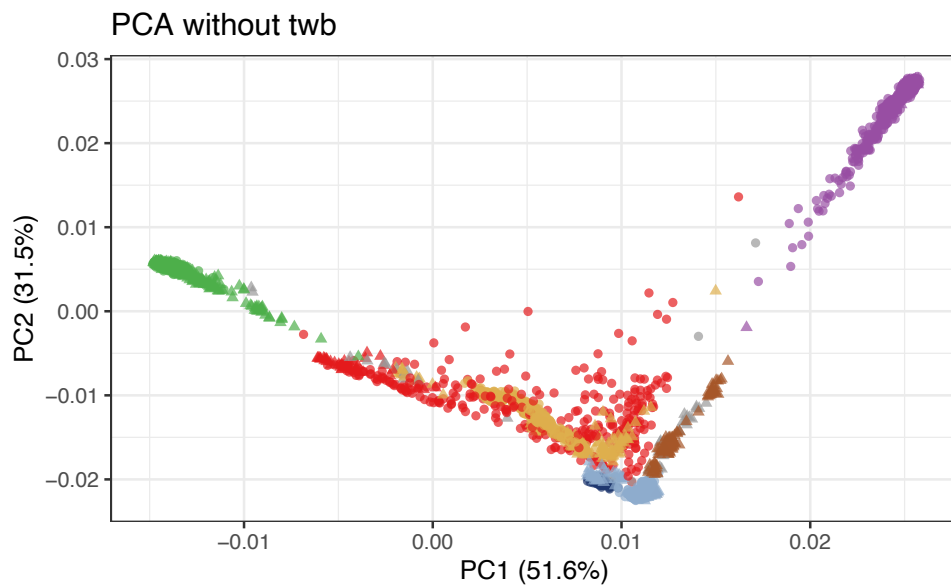
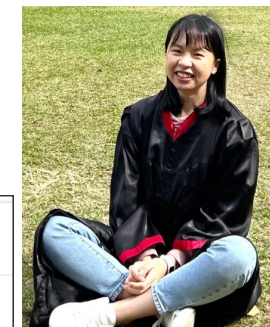
Target size

N ≈ 200,000

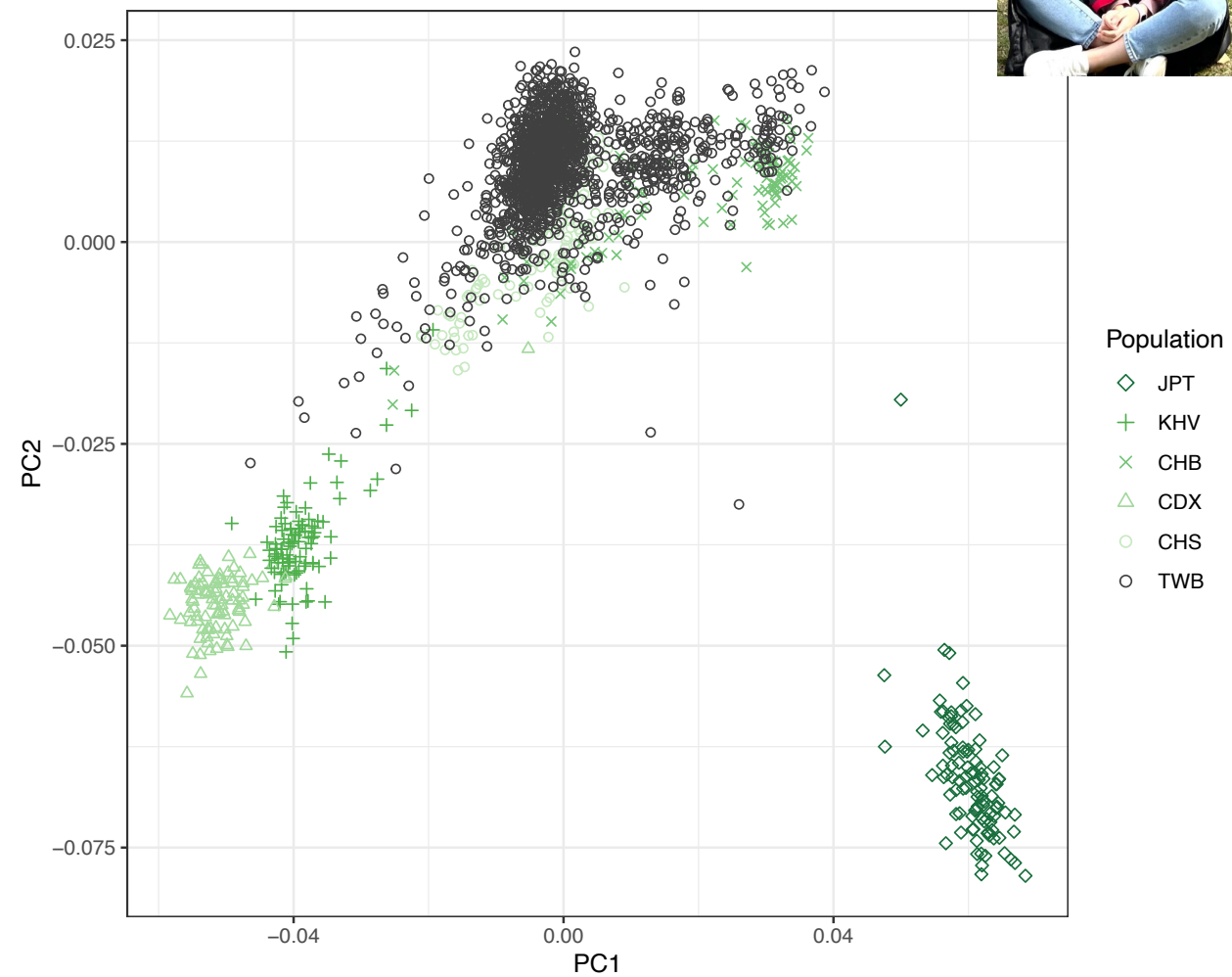
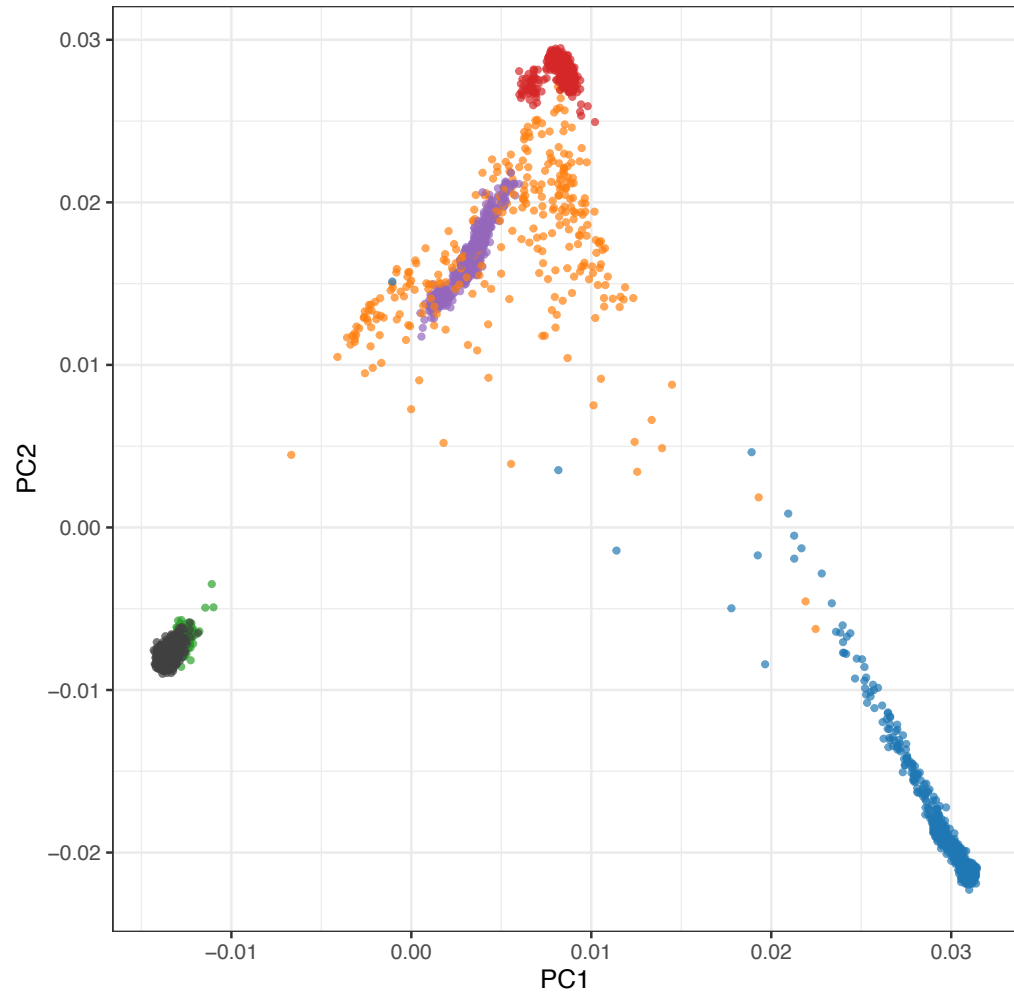


A**C****B**

TWB + 1kGP PCA



TWB + 1kGP PCA

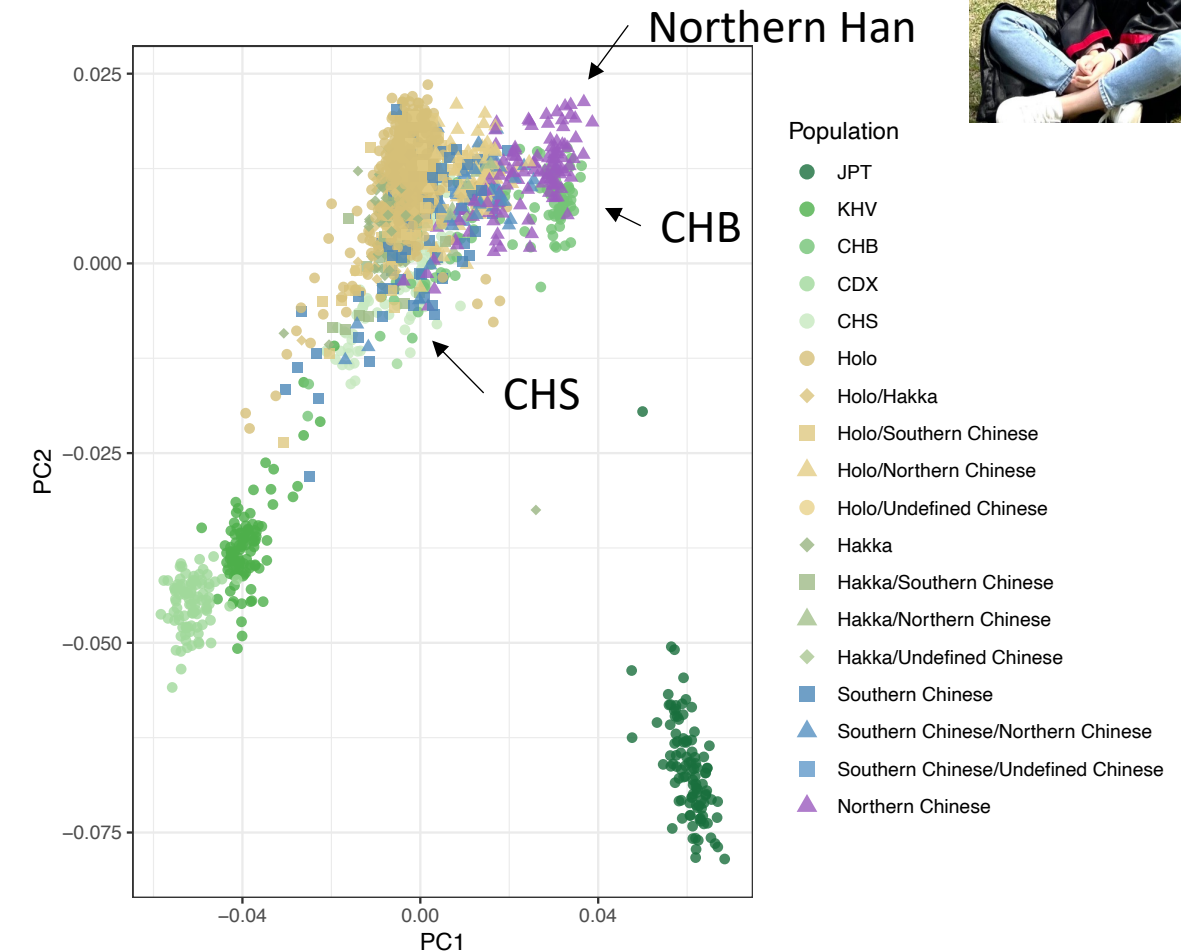
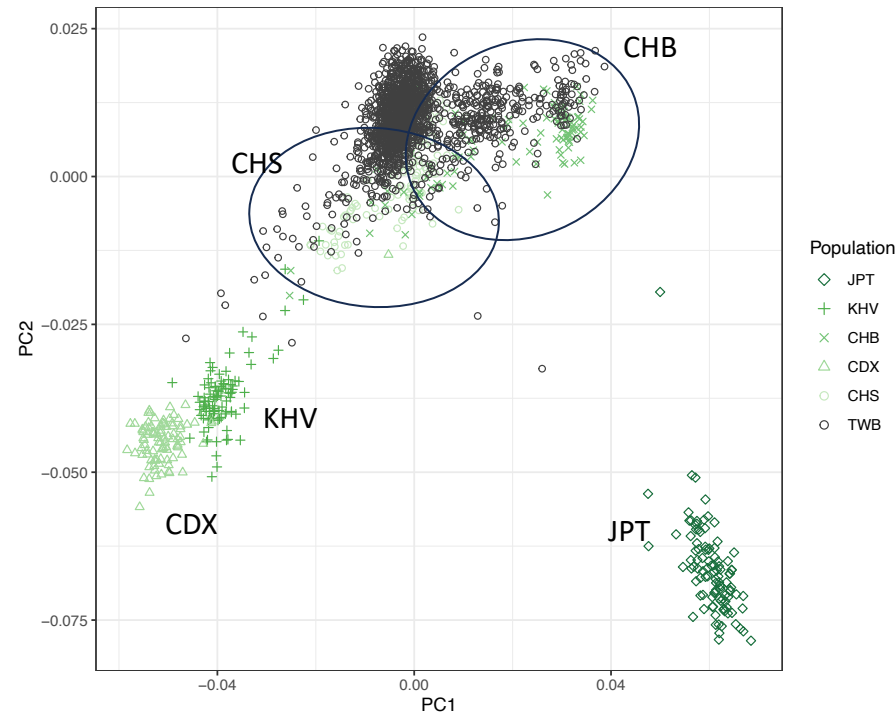


SNV: 76,651 variants; individual: TWB1492+1kGP (1492+2504=3996)

TWB + 1kGP_EAS PCA



Population Description	Population Code	Super Population
Chinese Dai in Xishuangbanna, China	CDX	EAS
Han Chinese in Beijing, China	CHB (106)	EAS
Japanese in Tokyo, Japan	JPT	EAS
Kinh in Ho Chi Minh City, Vietnam	KHV	EAS
Southern Han Chinese, China	CHS (105)	EAS



SNV: 76,651 variants; individual:
TWB1492+1kGP_eas (1492+504=1996)

Minor Allele Frequency (population)

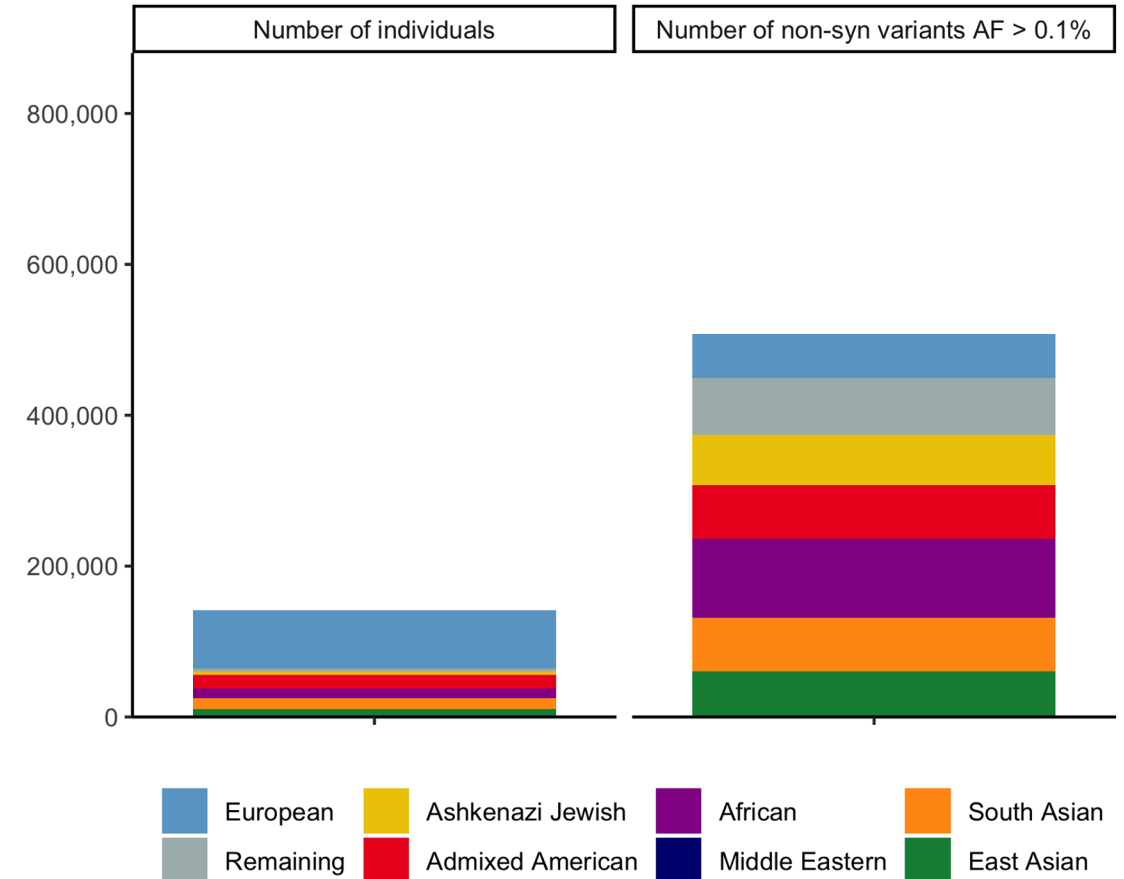
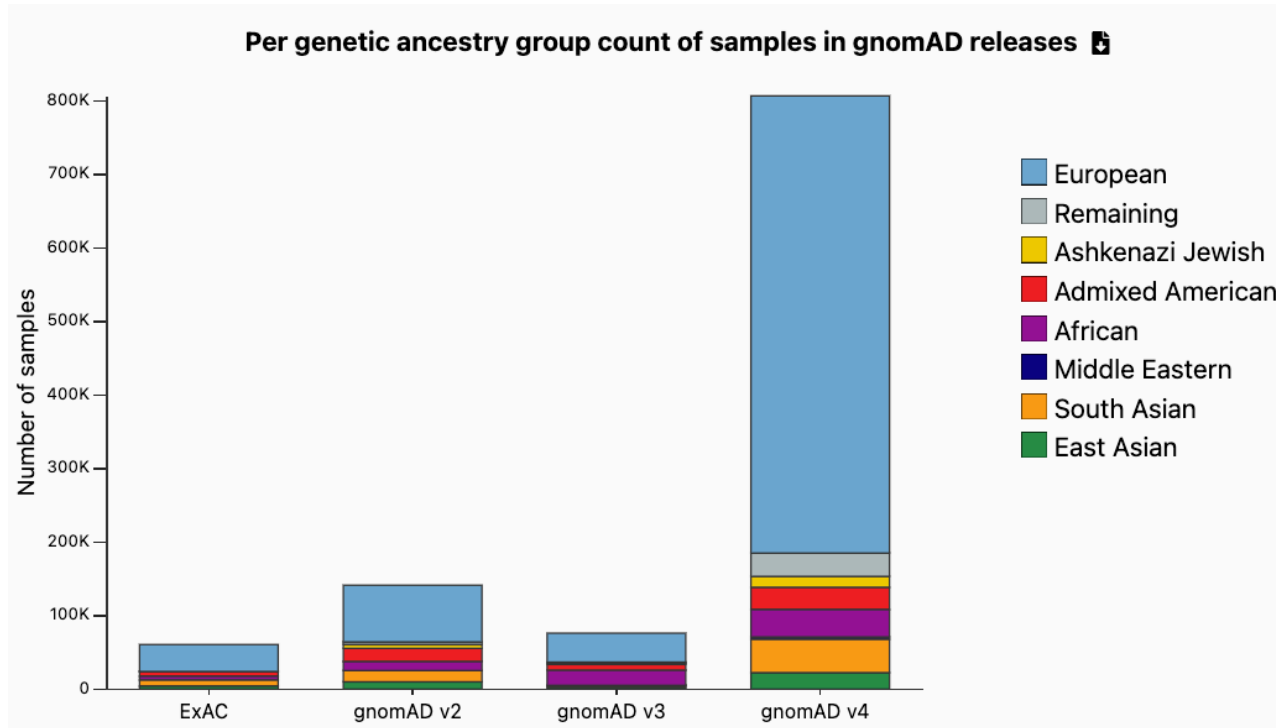
- Major allele frequency
- Minor allele frequency (MAF)
 - Disease prevalence/incidence rate
 - Evolution
 - Selection
 - Migration



Minor Allele Frequency (population)

- Minor Allele Frequency (MAF) is the frequency at which **the second most** common allele in a given population. If there are **two alleles**, with frequencies of 0.51, 0.49
 - MAF will be reported as 0.49.
- If there are **three** alleles, with frequencies of 0.51, 0.48, 0.01
 - MAF will be reported as 0.48 and 0.01 (each)
- For example, References SNP page for rs222 reports:
 - “MAF/MinorAlleleCount:G=0.249/542”.
 - This means that for rs222, minor allele is ‘G’ and has a frequency of 24.9% in the 1000Genome phase 1 population and that 'G' is observed **542** times in the sample population of **1088** people (or **2176** chromosomes).
- MAF is limited by the sample size

Sample sizes & Variant counts





Genome Aggregation Database

NM_004004.6(GJB2):c.109G>A (p.Val37Ile)

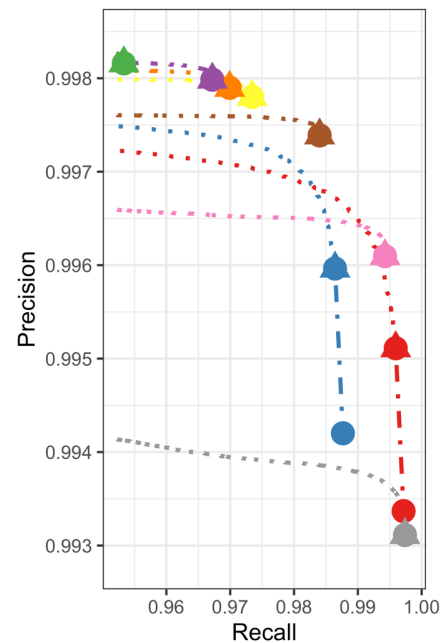
Genetic Ancestry Group	Allele Count	Allele Number	Number of Homozygotes	Allele Frequency
▸ East Asian	2013	44852	90	0.04488
▸ Ashkenazi Jewish	228	29584	0	0.007707
▸ Remaining	433	62494	6	0.006929
▸ Admixed American	171	60032	1	0.002848
▸ Middle Eastern	11	6058	0	0.001816
▸ European (Finnish)	106	64030	0	0.001655
▸ European (non-Finnish)	1040	1179624	2	0.0008816
▸ African/African American	55	75030	1	0.0007330
▸ South Asian	14	91070	1	0.0001537
▸ Amish	0	912	0	0.000
XX	2131	812302	59	0.002623
XY	1940	801384	42	0.002421
Total	4071	1613686	101	0.002523



Why Taiwan Biobank?

Genetic Ancestry		v4 Exomes			v4 Genomes			Combined		
Genetic Ancestry	Genetic Ancestry	Sample Count	XX	XY	Sample Count	XX	XY	Sample Count	XX	XY
Admixed American	amr	22,362	12,845	9,517	7,657	3,399	4,258			
African/African American	afr	16,740	9,663	7,077	20,805	11,094	9,711			
Ashkenazi Jewish	asj	13,068	6,318	6,750	1,736	934	802			
Amish	ami	-	-	-	456	235	221			
East Asian	eas	19,850	10,356	9,494	2,598	1,136	1,462			
European (Finnish)	fin	26,710	13,824	12,886	5,316	1,287	4,029			
European (non-Finnish)	nfe	556,006	286,144	269,862	34,025	19,683	14,342			
Middle Eastern	mid	2,884	1,253	1,631	147	72	75	3,031	1,325	1,706
Remaining	remaining	30,198	15,900	14,298	1,058	525	533	31,256	16,425	14,831
South Asian	sas	43,129	11,020	32,109	2,417	577	1,840	45,546	11,597	33,949
Total		730,947	367,323	363,624	76,215	38,942	37,273	807,162	406,265	400,897

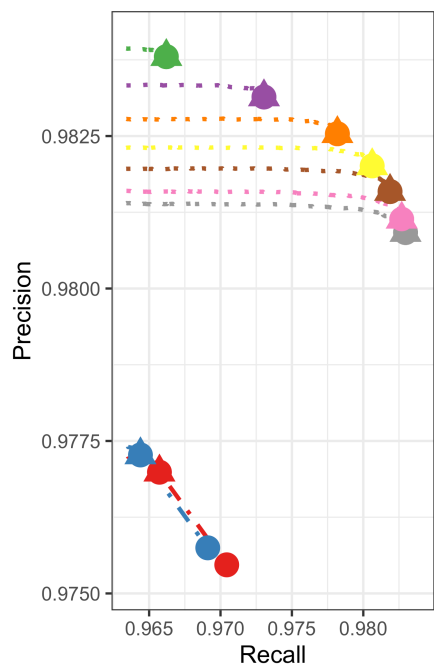
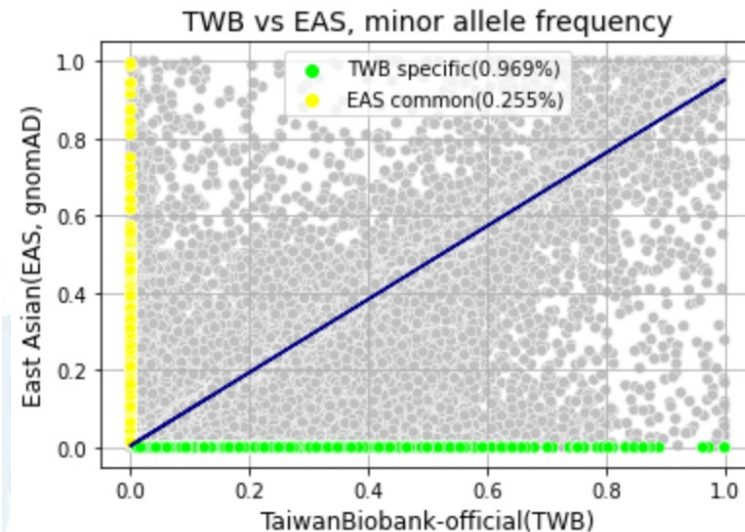
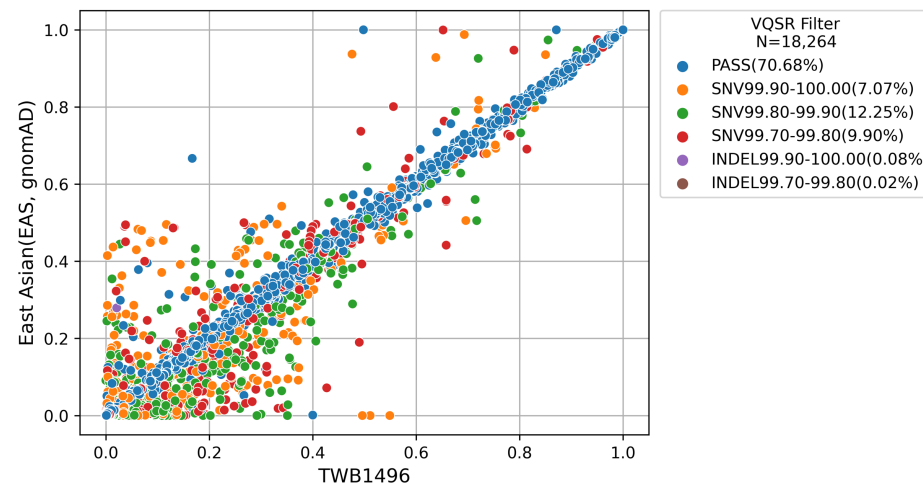
Inferred sex counts of gnomAD v4 samples per genetic ancestry group



SNV

- NA12878_30x/0.995
- NA12878_VQSR99.9/0.991
- VQSR99.0/0.975
- VQSR99.5/0.982
- VQSR99.6/0.984
- VQSR99.7/0.985
- VQSR99.8/0.991
- VQSR99.9/0.995
- VQSR_100/0.995

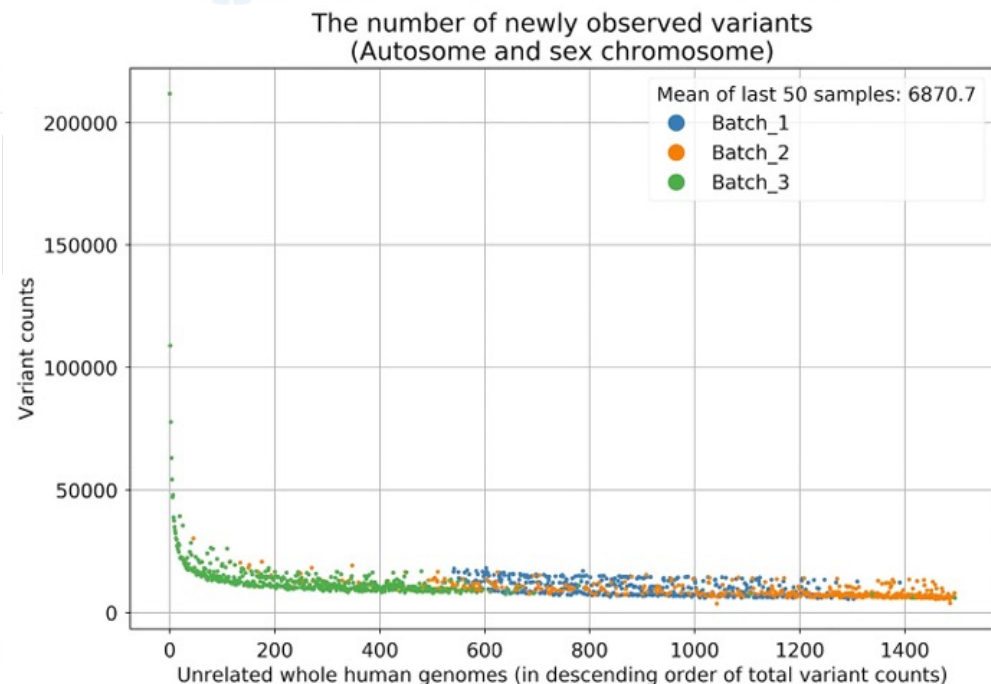
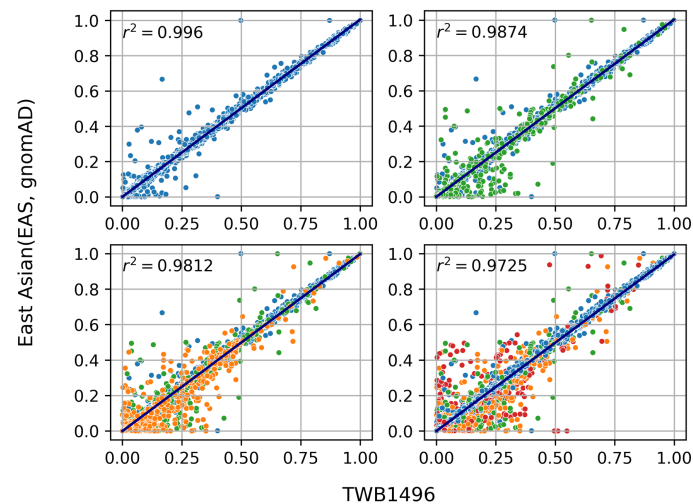
Filter
● ALL
▲ PASS

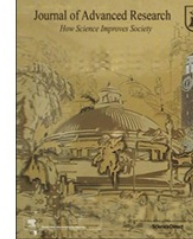


Indel

Filter
● ALL
▲ PASS

- NA12878_30x/0.973
- NA12878_VQSR99.9/0.972
- VQSR99.0/0.975
- VQSR99.5/0.978
- VQSR99.6/0.980
- VQSR99.7/0.981
- VQSR99.8/0.982
- VQSR99.9/0.982
- VQSR_100/0.982





Complete genomic profiles of 1496 Taiwanese reveal curated medical insights

Jacob Shujui Hsu^{a,b,1}, Dung-Chi Wu^{c,1}, Shang-Hung Shih^c, Jen-Feng Liu^b, Ya-Chen Tsai^d, Tung-Lin Lee^e, Wei-An Chen^e, Yi-Hsuan Tseng^a, Yi-Chung Lo^f, Hong-Ye Lin^d, Yi-Chieh Chen^a, Jing-Yi Chen^f, Ting-Hsuan Chou^a, Darby Tien-Hao Chang^{f,g}, Ming Wei Su^h, Wei-Hong Guo^d, Hsin-Hsiang Mao^d, Chien-Yu Chen^{c,d,*}, Pei-Lung Chen^{a,b,c,e,i,*}

^a Graduate Institute of Medical Genomics and Proteomics, National Taiwan University College of Medicine, Taipei 100025, Taiwan

^b Institute of Molecular Medicine, National Taiwan University College of Medicine, Taipei 100233, Taiwan

^c Genome and Systems Biology Degree Program, National Taiwan University and Academia Sinica, Taipei 10617, Taiwan

^d Department of Biomechatronics Engineering, National Taiwan University, Taipei 10617, Taiwan

^e Department of Medical Genetics, National Taiwan University Hospital, Taipei 100226, Taiwan

^f Department of Electrical Engineering, National Cheng-Kung University, Tainan 701401, Taiwan

^g Digital Technology Division, SinoPac Holdings, Taiwan

^h Institute of Biomedical Sciences, Academia Sinica, Taipei 115201, Taiwan

ⁱ Graduate Institute of Clinical Medicine, National Taiwan University College of Medicine, Taipei 100233, Taiwan



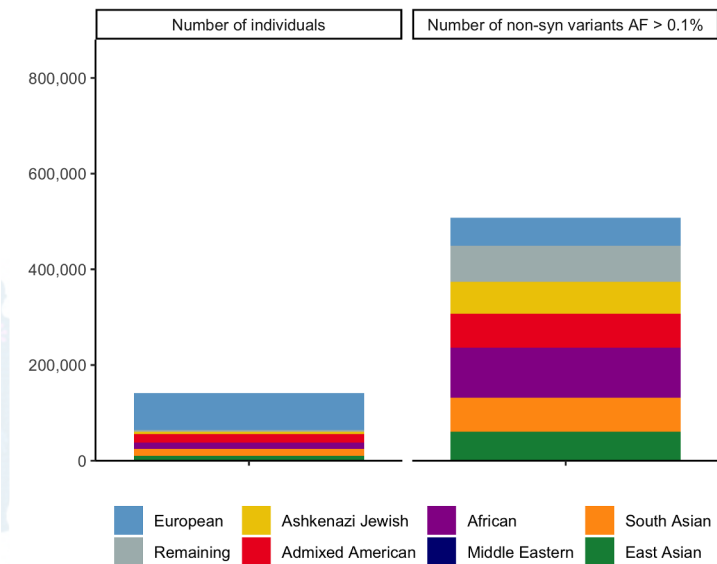
Pei-Lung Chen
2025-11-05
#3192268



Chien-Yu Chen



Tien-Hao Chang



TaiwanGenomes

Search variants by genomic position, rsID, gene name, clinical signature or their combinations. One may leave the search field empty to [browse all variants](#).

Q rs72474224

Search

Search samples:

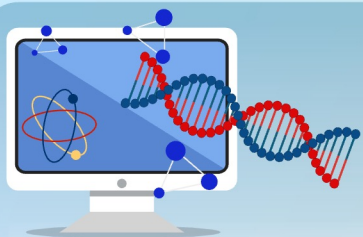
chr1:10000-20000
(gene position)

rs201752861
(rsID)

BRCA1 BRCA2
(gene name)

pathogenic likely_pathogenic
(clinical signature)

Genomic Profile of Taiwan Population WGS



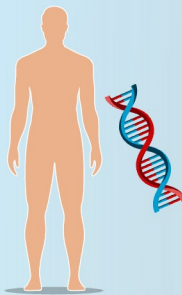
1,496 whole-genome sequencing datasets from the Taiwan Biobank



- Single nucleotide variants
- Small insertions/deletions
- Pharmacogenomic drug-associated alleles
- Copy number variants

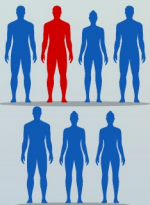


Reviewed by multiple practising clinicians



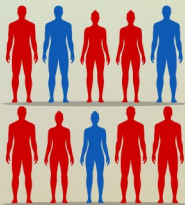
Average of 6,870.7 globally novel variants for each individual

54 Cytochrome P450 variant-drug pairs with high-level evidence



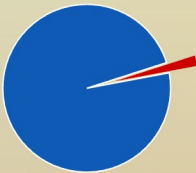
Population frequency: 13.2%

At least one human leukocyte antigen risk allele



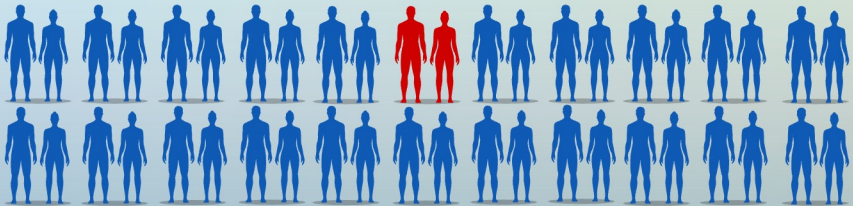
Population frequency: 19.1% (1A) 75.3% (1A/2A/2B)

At least one medical actionable variant



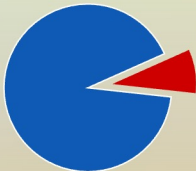
Population frequency: 1.67%

Risk of having offspring with at least one pathogenic variant



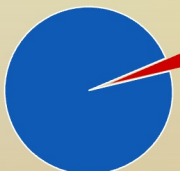
4.52%

Carrier frequency for alpha thalassemia (N=1484)



5.12%

Carrier frequency for Spinal Muscular Atrophy (N=1480)



1.6%



The reference genome hash tables

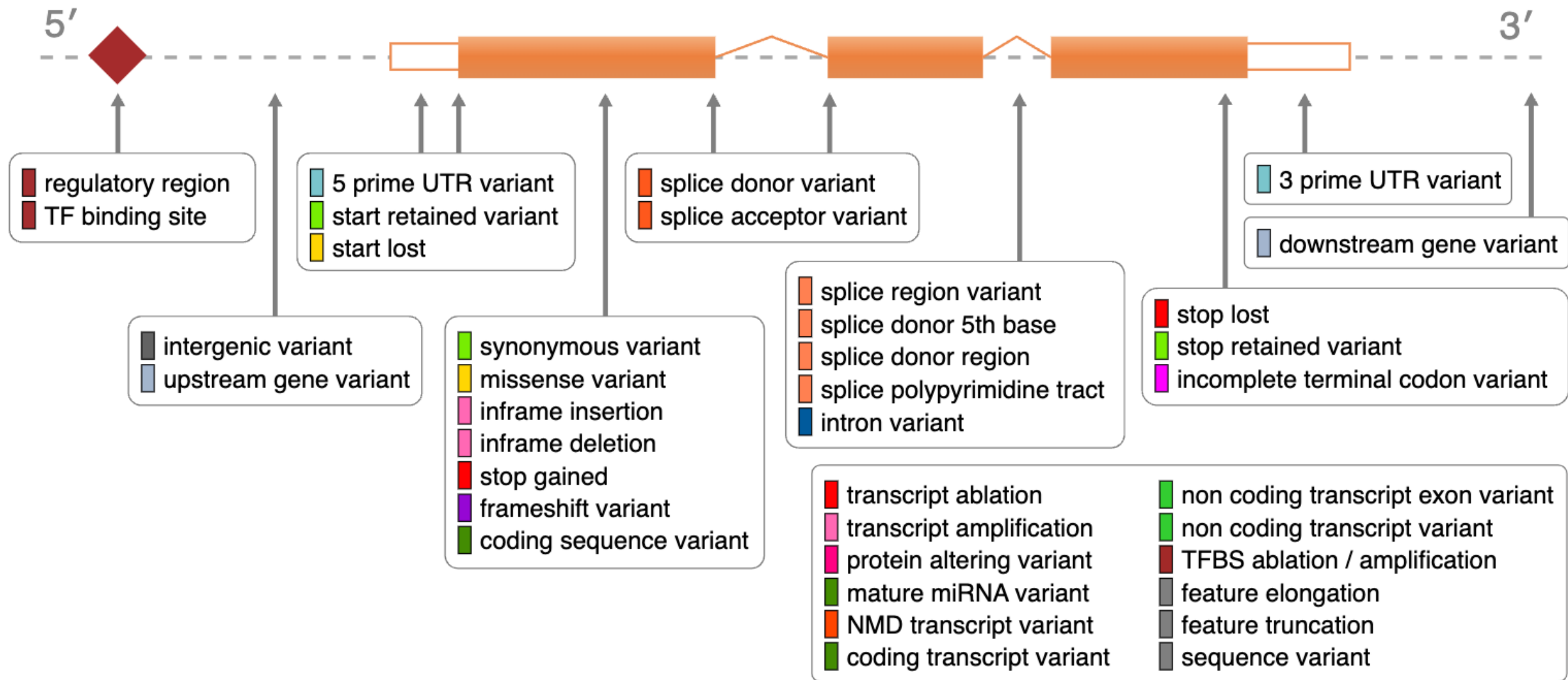
Assembly	Type	Reference Genome Hash Table	File Name	Major DRAGEN Version	3.8	3.9	3.10	4.0	4.2	4.3	4.4
				Hash Table Version	8	8	8	8	9	10	11
hg38	Pangenome (Graph) reference	Homo sapiens [1000 Genomes] hg38 v5 Pangenome	hg38-alt_masked.cnv.graph.hla.methyl_cg.rna-11-r5.0-1	11							✓
		Homo sapiens [1000 Genomes] hg38 v4 Multigenome	hg38-alt_masked.cnv.graph.hla.rna-10-r4.0-1	10						✓	
		Homo sapiens [1000 Genomes] hg38 v3 Multigenome	hg38-alt_masked.cnv.graph.hla.rna-9-r3.0-1	9					✓		
		Homo sapiens [1000 Genomes] hg38 v2 Multigenome	hg38-alt_masked.cnv.graph.hla.rna-8-r2.0-1	8			✓	✓			
		Homo sapiens [1000 Genomes] hg38 alt-masked Multigenome	hg38_alt_masked+cnv+graph+rna-8-r1.0-1	8		✓					
		Homo sapiens [1000 Genomes] hg38 alt-aware Multigenome	hg38_alt_aware+cnv+graph+rna-8-r1.0-0	8	✓	✓	✓				
	Linear reference	Homo sapiens [1000 Genomes] hg38 v5	hg38-alt_masked.cnv.hla.methyl_cg.methylated_combined.rna-11-r5.0-1	11							✓
		Homo sapiens [1000 Genomes] hg38 v4	hg38-alt_masked.cnv.hla.methylated_combined.rna-10-r4.0-1	10						✓	
		Homo sapiens [1000 Genomes] hg38 v3	hg38-alt_masked.cnv.hla.rna-9-r3.0-1	9					✓		
		Homo sapiens [1000 Genomes] hg38 methylation v3	hg38-alt_masked.methylated_combined.methylation.seed_len27-9-r3.0-1	9					✓		
Homo sapiens [1000 Genomes] hg38 v2		hg38-alt_masked.cnv.hla.rna-8-r2.0-1	8				✓				



Annotation

- Oxford Dictionary:
 - a note or notes added to a book or text giving **explanations** or **comments**; the act of adding these notes
- Variant Refinement/Filtering (genotype annotation)
 - Genotype level (###**FORMAT**)
 - Variant level (###**INFO**)
 - Subject level (Co-segregation/Inheritance mode)
- Variant Annotation
 - Genome information (refGene/knownGene/ensGene)
 - Minor allele frequency (ExAC/gnomAD/TWB/others)
 - Functional annotation (Knowledge-based/Data-driven/Both)
 - Variant effect prediction (SpliceAI/PrimateAI/others)

Ensembl Variant Effect Predictor (VEP)



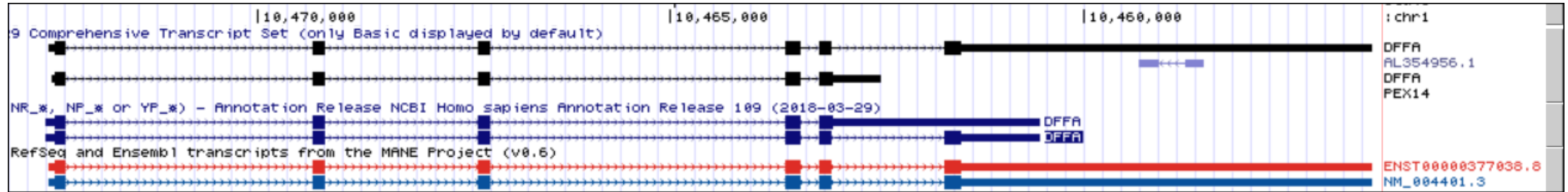
* SO term	SO description	SO accession	Display term	IMPACT
transcript_ablation	A feature ablation whereby the deleted region includes a transcript feature	SO:0001893	Transcript ablation	HIGH
splice_acceptor_variant	A splice variant that changes the 2 base region at the 3' end of an intron	SO:0001574	Splice acceptor variant	HIGH
splice_donor_variant	A splice variant that changes the 2 base region at the 5' end of an intron	SO:0001575	Splice donor variant	HIGH
stop_gained	A sequence variant whereby at least one base of a codon is changed, resulting in a premature stop codon, leading to a shortened transcript	SO:0001587	Stop gained	HIGH
frameshift_variant	A sequence variant which causes a disruption of the translational reading frame, because the number of nucleotides inserted or deleted is not a multiple of three	SO:0001589	Frameshift variant	HIGH
stop_lost	A sequence variant where at least one base of the terminator codon (stop) is changed, resulting in an elongated transcript	SO:0001578	Stop lost	HIGH
start_lost	A codon variant that changes at least one base of the canonical start codon	SO:0002012	Start lost	HIGH
transcript_amplification	A feature amplification of a region containing a transcript	SO:0001889	Transcript amplification	HIGH
feature_elongation	A sequence variant that causes the extension of a genomic feature, with regard to the reference sequence	SO:0001907	Feature elongation	HIGH
feature_truncation	A sequence variant that causes the reduction of a genomic feature, with regard to the reference sequence	SO:0001906	Feature truncation	HIGH
inframe_insertion	An inframe non synonymous variant that inserts bases into in the coding sequence	SO:0001821	Inframe insertion	MODERATE
inframe_deletion	An inframe non synonymous variant that deletes bases from the coding sequence	SO:0001822	Inframe deletion	MODERATE
missense_variant	A sequence variant, that changes one or more bases, resulting in a different amino acid sequence but where the length is preserved	SO:0001583	Missense variant	MODERATE
protein_altering_variant	A sequence_variant which is predicted to change the protein encoded in the coding sequence	SO:0001818	Protein altering variant	MODERATE
splice_donor_5th_base_variant	A sequence variant that causes a change at the 5th base pair after the start of the intron in the orientation of the transcript	SO:0001787	Splice donor 5th base variant	LOW
splice_region_variant	A sequence variant in which a change has occurred within the region of the splice site, either within 1-3 bases of the exon or 3-8 bases of the intron	SO:0001630	Splice region variant	LOW
splice_donor_region_variant	A sequence variant that falls in the region between the 3rd and 6th base after splice junction (5' end of intron)	SO:0002170	Splice donor region variant	LOW
splice_polypyrimidine_tract_variant	A sequence variant that falls in the polypyrimidine tract at 3' end of intron between 17 and 3 bases from the end (acceptor -3 to acceptor -17)	SO:0002169	Splice polypyrimidine tract variant	LOW
incomplete_terminal_codon_variant	A sequence variant where at least one base of the final codon of an incompletely annotated transcript is changed	SO:0001626	Incomplete terminal codon variant	LOW
start_retained_variant	A sequence variant where at least one base in the start codon is changed, but the start remains	SO:0002019	Start retained variant	LOW
stop_retained_variant	A sequence variant where at least one base in the terminator codon is changed, but the terminator remains	SO:0001567	Stop retained variant	LOW
synonymous_variant	A sequence variant where there is no resulting change to the encoded amino acid	SO:0001819	Synonymous variant	LOW
coding_sequence_variant	A sequence variant that changes the coding sequence	SO:0001580	Coding sequence variant	MODIFIER
mature_miRNA_variant	A transcript variant located with the sequence of the mature miRNA	SO:0001620	Mature miRNA variant	MODIFIER
5_prime_UTR_variant	A UTR variant of the 5' UTR	SO:0001623	5 prime UTR variant	MODIFIER
3_prime_UTR_variant	A UTR variant of the 3' UTR	SO:0001624	3 prime UTR variant	MODIFIER
non_coding_transcript_exon_variant	A sequence variant that changes non-coding exon sequence in a non-coding transcript	SO:0001792	Non coding transcript exon variant	MODIFIER
intron_variant	A transcript variant occurring within an intron	SO:0001627	Intron variant	MODIFIER
NMD_transcript_variant	A variant in a transcript that is the target of NMD	SO:0001621	NMD transcript variant	MODIFIER
non_coding_transcript_variant	A transcript variant of a non coding RNA gene	SO:0001619	Non coding transcript variant	MODIFIER
coding_transcript_variant	A transcript variant of a protein coding gene	SO:0001968	Coding transcript variant	MODIFIER
upstream_gene_variant	A sequence variant located 5' of a gene	SO:0001631	Upstream gene variant	MODIFIER
downstream_gene_variant	A sequence variant located 3' of a gene	SO:0001632	Downstream gene variant	MODIFIER
TFBS_ablation	A feature ablation whereby the deleted region includes a transcription factor binding site	SO:0001895	TFBS ablation	MODIFIER
TFBS_amplification	A feature amplification of a region containing a transcription factor binding site	SO:0001892	TFBS amplification	MODIFIER
TF_binding_site_variant	A sequence variant located within a transcription factor binding site	SO:0001782	TF binding site variant	MODIFIER
regulatory_region_ablation	A feature ablation whereby the deleted region includes a regulatory region	SO:0001894	Regulatory region ablation	MODIFIER
regulatory_region_amplification	A feature amplification of a region containing a regulatory region	SO:0001891	Regulatory region amplification	MODIFIER
regulatory_region_variant	A sequence variant located within a regulatory region	SO:0001566	Regulatory region variant	MODIFIER
intergenic_variant	A sequence variant located in the intergenic region, between genes	SO:0001628	Intergenic variant	MODIFIER
sequence_variant	A sequence_variant is a non exact copy of a sequence_feature or genome exhibiting one or more sequence_alteration	SO:0001060	Sequence variant	MODIFIER



Gene Feature (Annotation)

Feature	Code	Explanation	
Frameshift	0	Short insertion or deletion result in a completely different translation from the original.	Non-synonymous
Nonframeshift	1	Short insertion or deletion results in loss of amino acids in the translated proteins.	
Startloss	2	Indels or nucleotide substitution result in the loss of start codon(ATG) (mutated into a non-start codon).	
Stoploss	3	Indels or nucleotide substitution result in the loss of stop codons (TAG, TAA, TGA)	
Stopgain	4	Indels or nucleotide substitution result in the new stop codons (TAG, TAA, TGA), which may truncate the protein.	
Splicing	5	variant is within 2-bp of a splicing junction (use <code>--splicing x</code> to change this, the unit of x is base-pair)	
Missense	6	Variants result in a codon coding for a different amino acid (missense)	
Synonymous	7	Nucleotide substitution does not change amino acid.	
Exonic	8	Due to loss of sequences in reference database, this variant can only be mapped into exonic region without more precise annotation.	
UTR5	9	Within a 5' untranslated region	
UTR3	10	Within a 3' untranslated region	
Intronic	11	Within an intron	
Upstream	12	Within 1-kb region upstream of transcription start site (use <code>--neargene x</code> to change this, the unit of x is base-pair)	
Downstream	13	Within 1-kb region downstream of transcription end site (use <code>--neargene x</code> to change this, the unit of x is base-pair)	
ncRNA	14	Within a transcript without protein-coding annotation in the gene definition (see Notes below for more explanation)	
Intergenic	15	variant is in intergenic region	
Monomorphic	16	It is not a sequence variation actually, which may be resulted from bugs in reference genome in variant calling.	
Unknown	17	Variants has no annotation.	

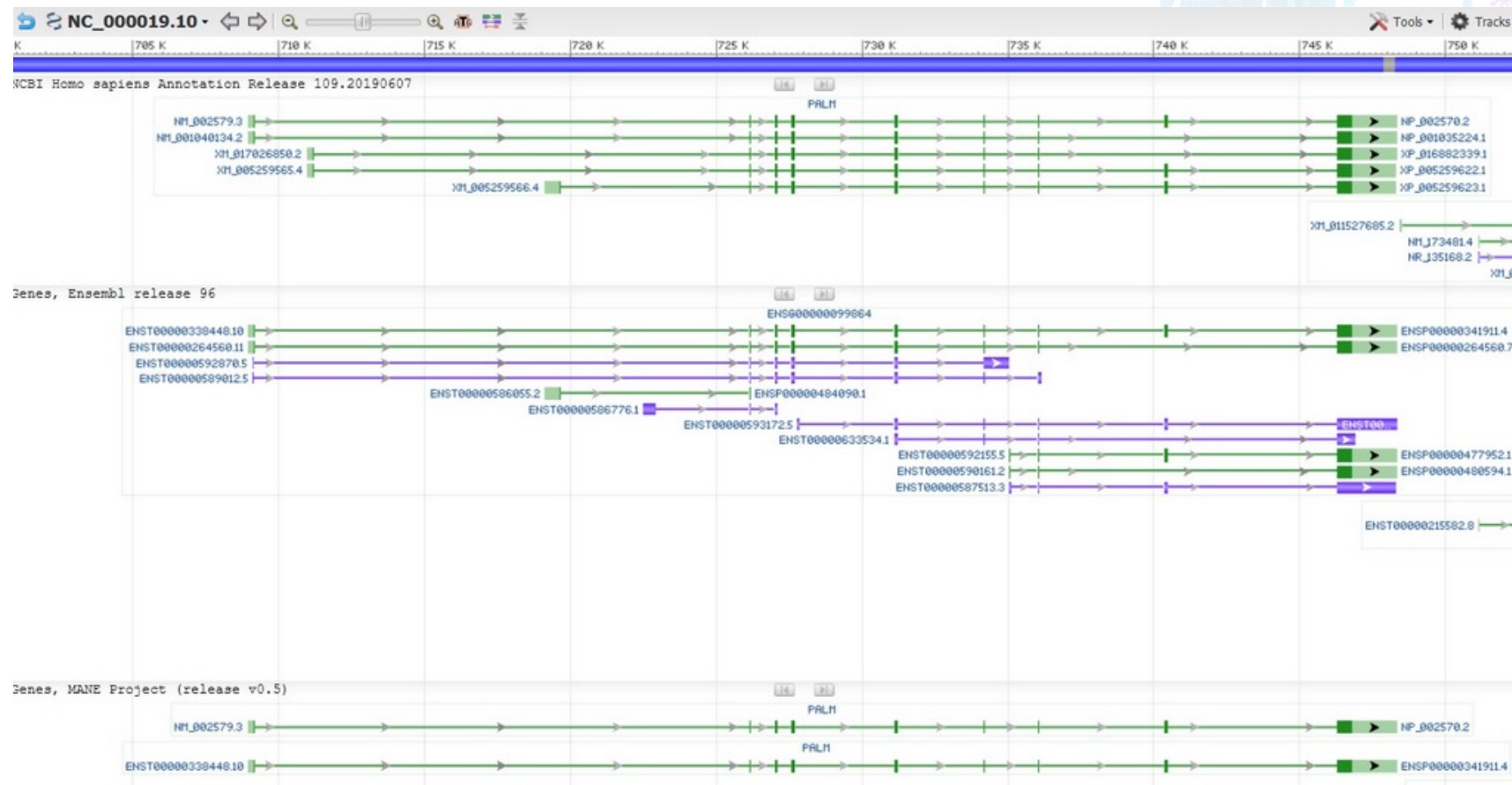
RNA Transcripts in two databases



NCBI

Ensembl

MANE



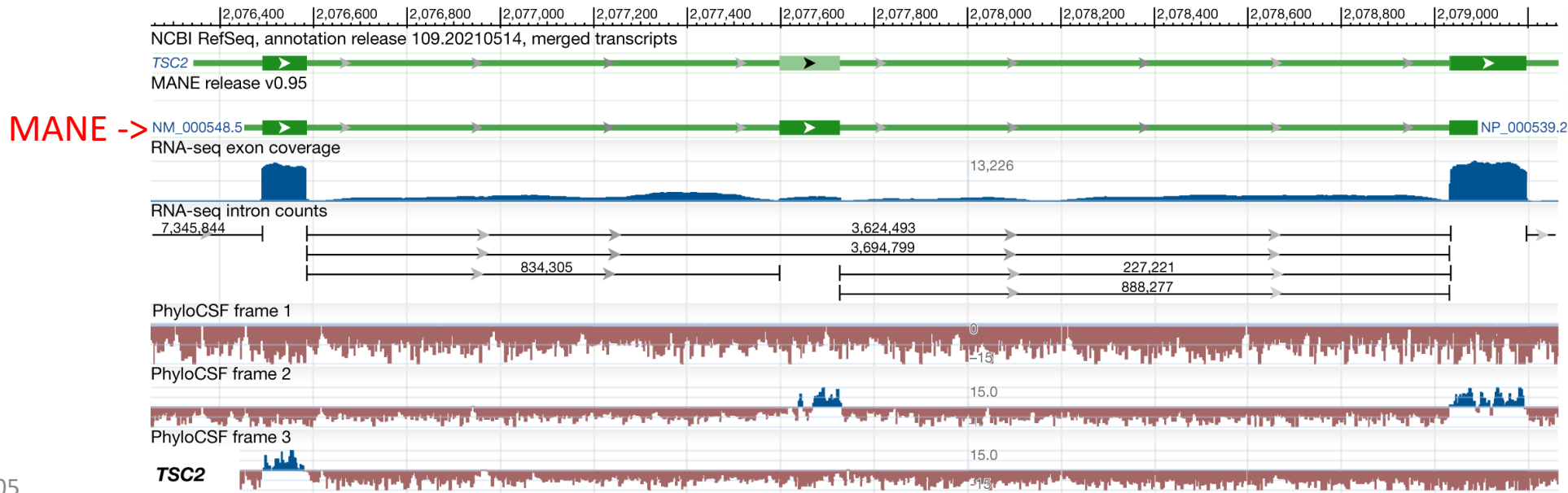
基因體暨蛋白質醫學研究所
Graduate Institute of Medical Genomics and Proteomics



Matched Annotation from NCBI and EMBL-EBI (MANE)



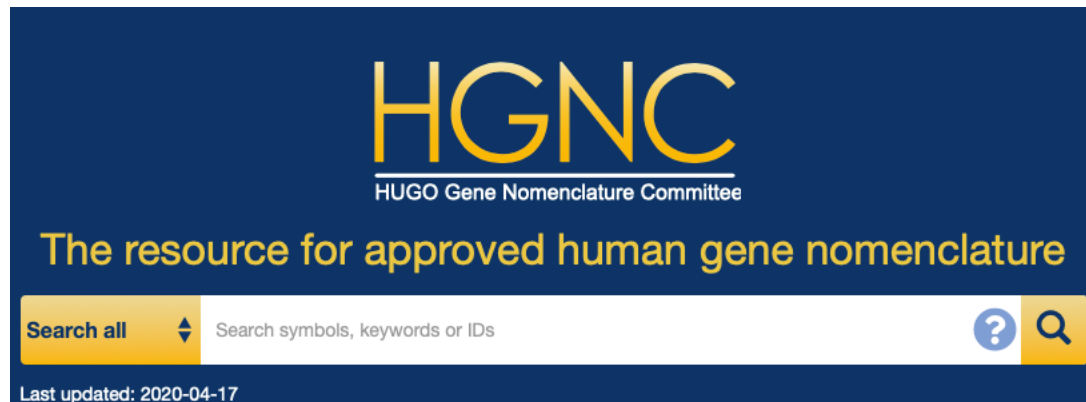
Low expression
No conservation



Low expression
conservation

HUGO Gene Nomenclature Committee (HGNC)

- Human Genome Organization (HUGO) that sets the standards for human gene nomenclature. The HGNC approves a **unique** and **meaningful** name for every known human gene
- The work of the HGNC is supported by National Human Genome Research Institute (NHGRI)
- Approved Gene Symbol
- *BRCA1* (*RNF53, BRCC1, PPP1R53, FANCS*)



Sequence Variant Nomenclature (HGVS)

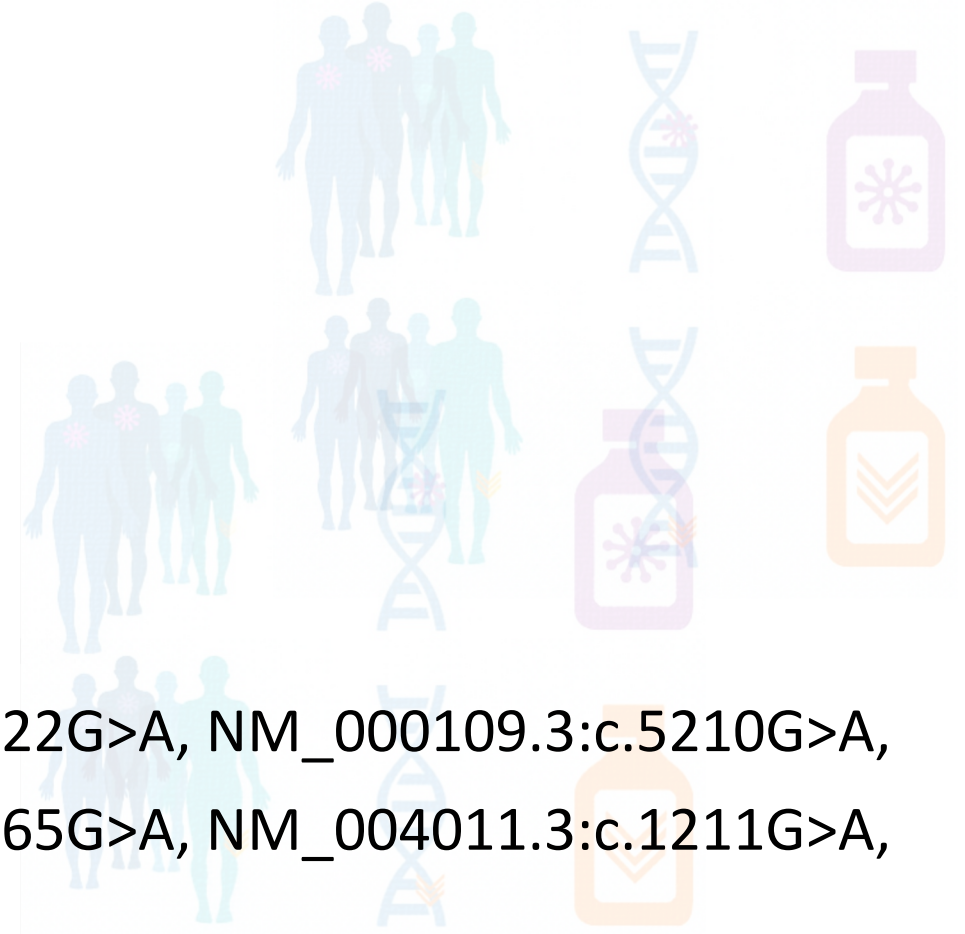
- Human Genome Variation Society (HGVS):
 - NC_ a genomic reference sequence based on a chromosome
 - NC_000023.10:g.32407761G>A (Feb.2009: h19, GRCh37)
 - NC_000023.11:g.32389644G>A (Dec.2013: hg38, GRCh38)
 - NG_ a genomic reference sequence based on a Gene or Genomic region
 - NG_012232.1:g.954966C>T
 - LRG_ a genomic reference sequence, based on a Gene or Genomic region (clinical)
 - LRG_199:g.954966C>T
- transcript (RNA, nucleotide)
 - NM_ a reference sequence based on a protein coding RNA (mRNA)
 - NM_004006.2:c.4375C>T
 - NR_ a reference sequence based on a non-protein coding RNA
 - NR_002196.1:c.601G>T
- protein (amino acid)
 - NP_ a reference sequence based on a protein (amino acid) sequence
 - NP_003997.1:p.Arg1459* (p.Arg1459Ter)
 - () indicate uncertainties and predicted consequences

Sequence Variant Nomenclature (HGVS)

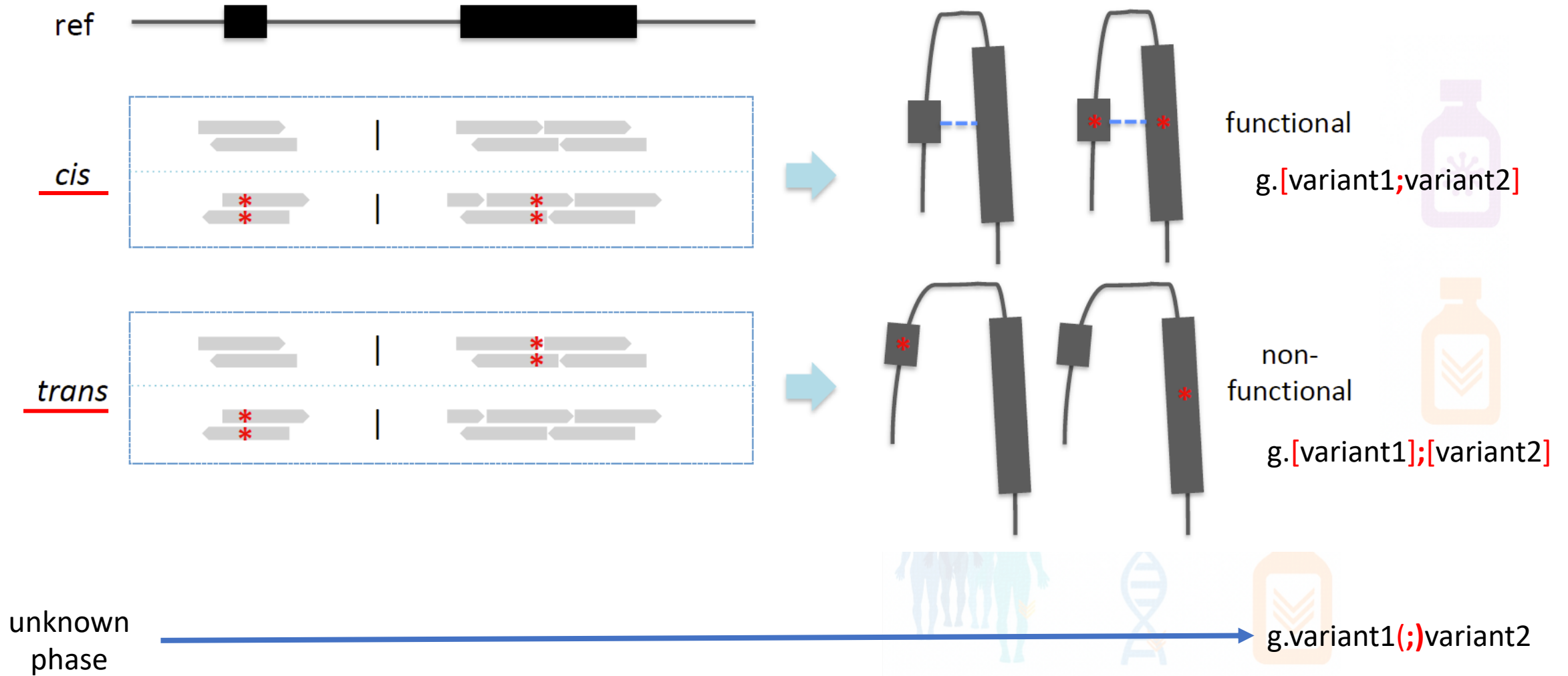
- Type of Variants
- Substitution: NM_00XXXX.X: c.4375C>T
- Deletion: NM_00XXXX.X: c.4375_4379del
 - NM_00XXXX.X: c.4375_4379delCGATT.
- Duplication: NM_00XXXX.X: c.4375_4385dup
 - c.4375_4385dupCGATTATTCCA
- Insertion: NM_00XXXX.X: c.4375_4376insACCT
- Intronic: NG_012232.1(NM_004006.2):c.93+1G>T
- Multi-nucleotide: NM_004006.2:c.145_147delinsTGG
 - two substitutions replacing codon CGC (position c.145 to c.147) by TGG

Sequence Variant Nomenclature (Aliases)

- NC_000023.9:g.32290917C>T,
- NC_000023.10:g.32380996C>T,
- NC_000023.11:g.32362879C>T,
- NG_012232.1:g.981731G>A,
- LRG_199:g.981731G>A
- LRG_199t1:c.5234G>A
- NM_004006.2:c.5234G>A, NM_004009.3:c.5222G>A, NM_000109.3:c.5210G>A,
NM_004007.2:c.4865G>A, NM_004010.3:c.4865G>A, NM_004011.3:c.1211G>A,
NM_004012.3:c.1202G>A

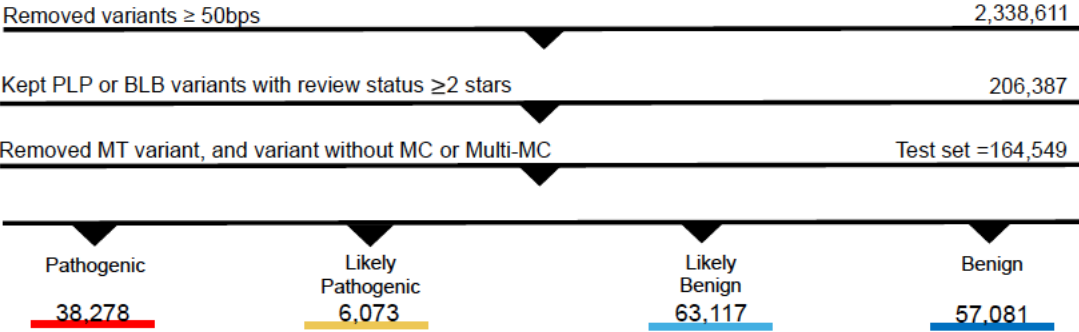


Haplotype & Phasing



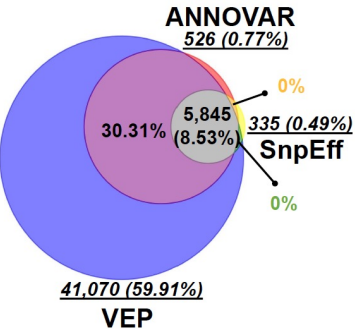
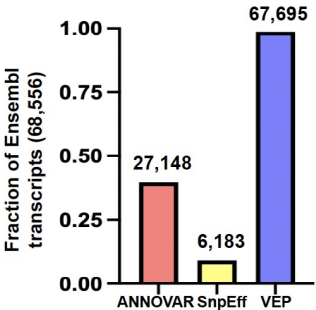
Toward Streamlined Variant Interpretation: Discordant Genetic Interpretation Across Variant Annotations for ClinVar Pathogenic Variants

ClinVar 20240107
GRCh38 VCF ClinVar = 2,349,000

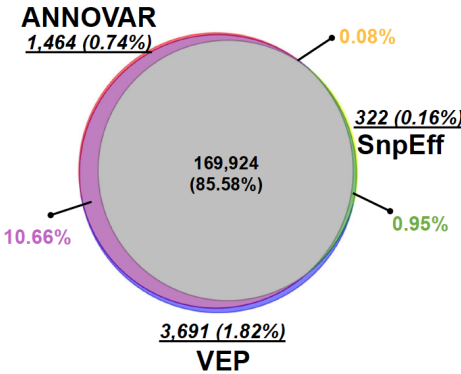


B

Ensembl

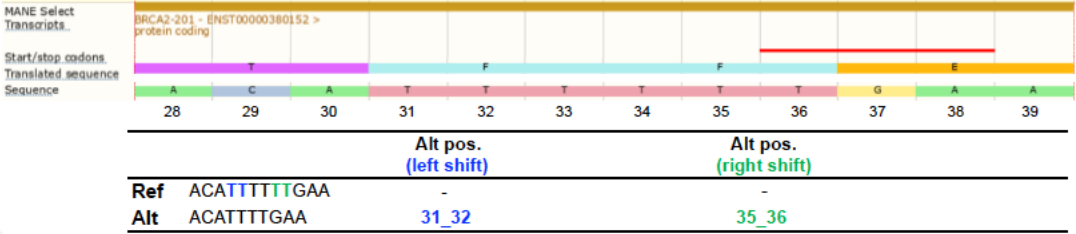


E Coding impact (n=198,545)



Left or right align

chr13: 32316491-32316492 ClinVar Variation ID: 1696818
NM_000059.4:c.35_36del



Annotation discordance

chr13: 32316491-32316492 ClinVar Variation ID: 1696818
NM_000059.4:c.35_36del

Annotator	Transcript set	Coding impact	HGVSc	HGVSp
ANNOVAR	Ensembl	stopgain	ENST00000380152.8:c.31_32del	ENSP00000369497.3:p.F12*
	RefSeq	stopgain	NM_000059.4:c.31_32del	NP_000050.3:p.F12*
SnpEff	Ensembl	frameshift_variant	ENST00000380152.8:c.35_36delTT	ENSP00000369497.3:p.Phe12fs
	RefSeq	frameshift_variant	NM_000059.4:c.35_36delTT	NP_000050.3:p.Phe12fs
VEP	Ensembl	frameshift_variant	ENST00000380152.8:c.35_36del	ENSP00000369497.3:p.Phe12Ter
	RefSeq	frameshift_variant	NM_000059.4:c.35_36del	NP_000050.3:p.Phe12Ter

ACMG guideline(s)

- **Standards/guidelines for the interpretation of sequence variants**
 - Why guideline? We need a **consensus**.
- It was developed primarily as an **educational resource** for clinical laboratory geneticists with quality laboratory services. The recommendation describes a process for **classifying variants** into **five categories** based **types of variant evidence**.
- For **Mendelian disorder** only
- Not for the interpretation of
 - cancer somatic variation, pharmacogenomic (PGx) variants
 - variants associated with multigenic non-Mendelian complex disorders
- In the beginning, the guideline was intended to interpret known clinical relevant variants, it was **NOT for research purpose** on disease mapping.

	Benign Supporting Supporting Moderate Strong Very strong Pathogenic					
Population data	MAF is too high for disorder BA1/BS1 OR observation in controls inconsistent with disease penetrance BS2			Absent in population databases PM2	Prevalence in affecteds statistically increased over controls PS4	
Computational and predictive data		Multiple lines of computational evidence suggest no impact on gene /gene product BP4 Missense in gene where only truncating cause disease BP1 Silent variant with non predicted splice impact BP7 In-frame indels in repeat w/out known function BP3	Multiple lines of computational evidence support a deleterious effect on the gene /gene product PP3	Novel missense change at an amino acid residue where a different pathogenic missense change has been seen before PM5 Protein length changing variant PM4	Same amino acid change as an established pathogenic variant PS1	Predicted null variant in a gene where LOF is a known mechanism of disease PVS1
Functional data	Well-established functional studies show no deleterious effect BS3		Missense in gene with low rate of benign missense variants and path. missenses common PP2	Mutational hot spot or well-studied functional domain without benign variation PM1	Well-established functional studies show a deleterious effect PS3	
Segregation data	Nonsegregation with disease BS4		Cosegregation with disease in multiple affected family members PP1	Increased segregation data →		
De novo data				De novo (without paternity & maternity confirmed) PM6	De novo (paternity and maternity confirmed) PS2	
Allelic data		Observed in <i>trans</i> with a dominant variant BP2 Observed in <i>cis</i> with a pathogenic variant BP2		For recessive disorders, detected in <i>trans</i> with a pathogenic variant PM3		
Other database		Reputable source w/out shared data = benign BP6	Reputable source = pathogenic PP5			
Other data		Found in case with an alternate cause BP5	Patient's phenotype or FH highly specific for gene PP4			



Table 1 Population, disease-specific, and sequence databases

Population databases	
Exome Aggregation Consortium http://exac.broadinstitute.org/	Database of variants found during exome sequencing of 61,486 unrelated individuals sequenced as part of various disease-specific and population genetic studies. Pediatric disease subjects as well as related individuals were excluded.
Exome Variant Server http://evs.gs.washington.edu/EVS	Database of variants found during exome sequencing of several large cohorts of individuals of European and African American ancestry. Includes coverage data to inform the absence of variation.
1000 Genomes Project http://browser.1000genomes.org	Database of variants found during low-coverage and high-coverage genomic and targeted sequencing from 26 populations. Provides more diversity compared to the Exome Variant Server but also contains lower-quality data, and some cohorts contain related individuals.
dbSNP http://www.ncbi.nlm.nih.gov/snp	Database of short genetic variations (typically ≤ 50 bp) submitted from many sources. May lack details of the originating study and may contain pathogenic variants.
dbVar http://www.ncbi.nlm.nih.gov/dbvar	Database of structural variation (typically > 50 bp) submitted from many sources.
Disease databases	
ClinVar http://www.ncbi.nlm.nih.gov/clinvar	Database of assertions about the clinical significance and phenotype relationship of human variations.
OMIM http://www.omim.org	Database of human genes and genetic conditions that also contains a representative sampling of disease-associated genetic variants.
Human Gene Mutation Database http://www.hgmd.org	Database of variant annotations published in the literature. Requires fee-based subscription to access much of the content.
Locus/disease/ethnic/other-specific databases	
Human Genome Variation Society http://www.hgvs.org/dblist/dblist.html	The Human Genome Variation Society site developed a list of thousands of databases that provide variant annotations on specific subsets of human variation. A large percentage of databases are built in the Leiden Open Variation Database system.
Leiden Open Variation Database http://www.lovd.nl	
DECIPHER http://decipher.sanger.ac.uk	A molecular cytogenetic database for clinicians and researchers linking genomic microarray data with phenotype using the Ensembl genome browser.

BA1/BS1: MAF is **too high** for disorder

PM2: **Absent** in population databases

or

BS2: observation in controls **inconsistent** with disease penetrance

ClinVar: public archive of relationships among sequence variation and human phenotype

Melissa J. Landrum, Jennifer M. Lee, George R. Riley, Wonhee Jang, Wendy S. Rubinstein, Deanna M. Church and Donna R. Maglott*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD 20894, USA



About Us

Researchers & Clinicians

Patients & Families

Sign-In

ICCG » About Us » The ClinGen Resource

About Us

The ClinGen Resource

ClinGen is a National Institute of Health (NIH)-funded resource dedicated to harnessing both research data and the data from the hundreds of thousands of clinical genetics tests being performed each year, as well as supporting expert curation to determine which variants are most relevant to patient care. In 2013, the National Human Genome Research Institute (NHGRI) and the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD) (subsidiaries of NIH) awarded three grants totalling over \$25 million dollars to support a consortium of research groups to design and implement a framework for evaluating which variants play a role in disease and those that are relevant to patient care. The groups will work closely with the National Center for Biotechnology Information (NCBI) of the National Library of Medicine (NLM), which will distribute this information through its ClinVar database. Learn about each of the groups involved in the ClinGen Resource, their individual charges, and their collaborative efforts.

About Us

[Our Aims](#)

[The ClinGen Resource](#)

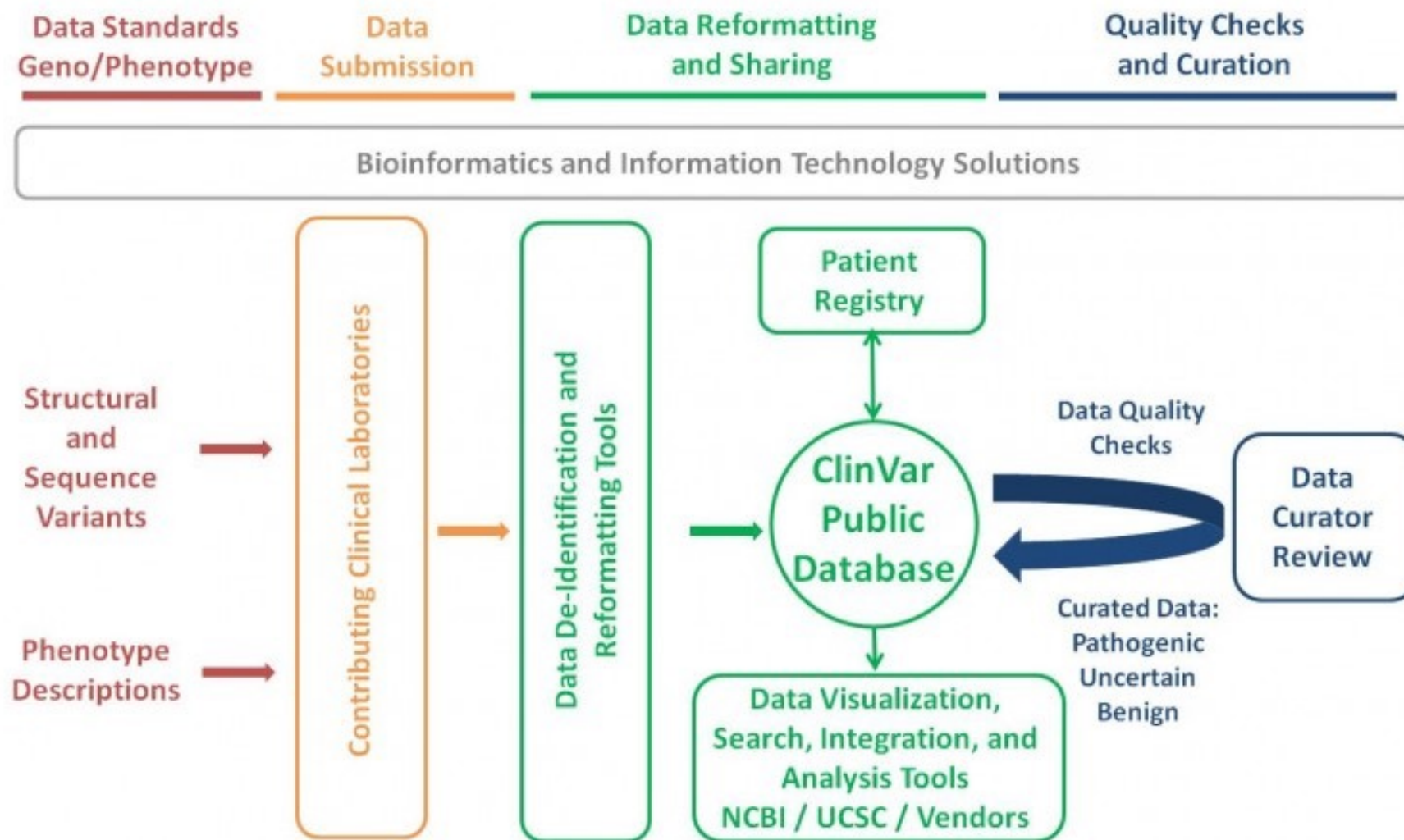
[News, Events & Publications](#)

[ICCG Conferences](#)

[Other Projects and Collaborations](#)



The aim of ClinVar



CFTR:c.1652G>A (p.Gly551Asp) AND Cystic fibrosis

Clinical significance: Pathogenic (Last evaluated: Mar 28, 2013)
Review status: ★ ★ ★ ★

Based on: 4 submissions [\[Details\]](#)
Record status: current
Accession: RCV000007540.7

Allele description

Gene: CFTR:cystic fibrosis transmembrane conductance regulator (ATP-binding cassette sub-family C, member 7) [\[Gene\]](#) - [OMIM](#)
Variant type: single nucleotide variant
Cytogenetic location: 7q31.2
Genomic location: Chr7:117587806 (on Assembly GRCh38)
Preferred name: CFTR:c.1652G>A (p.Gly551Asp)
HGVS: NC_000007.14:g.117587806G>A
NG_016465.3:g.127023G>A
NM_000492.3:c.1652G>A [...more](#)

Protein change: G551D; GLY551ASP
Links: OMIM: [602421.0013](#); dbSNP: [75527207](#)
1000Genome: [rs75527207](#)
Allele Frequency: 0.0014, [GO-ESP](#)
Molecular consequence: NM_000492.3:c.1652G>A: missense [Sequence Ontology: SO:0001583]
Suspect: Not available
Observations: 3

Condition(s)

Name: Cystic fibrosis (CF)
Synonyms: Mucoviscidosis; Classic Cystic Fibrosis
Identifiers: MedGen: [C0010674](#); OMIM: [219700](#); Orphanet: [586](#)
Age of onset: Neonatal/infancy
Prevalence:

- 1-5 / 10 000 Orphanet: [586](#)
- 1:3200 <http://www.ncbi.nlm.nih.gov/books/NBK1250/>
- 1 in 2000-4000 depending on the population studied.

* May be provided by NCBI

Interpretation

- Significance
- Review status *
 - submitter/criteria provided/etc.
- Accession.version *

Allele summary

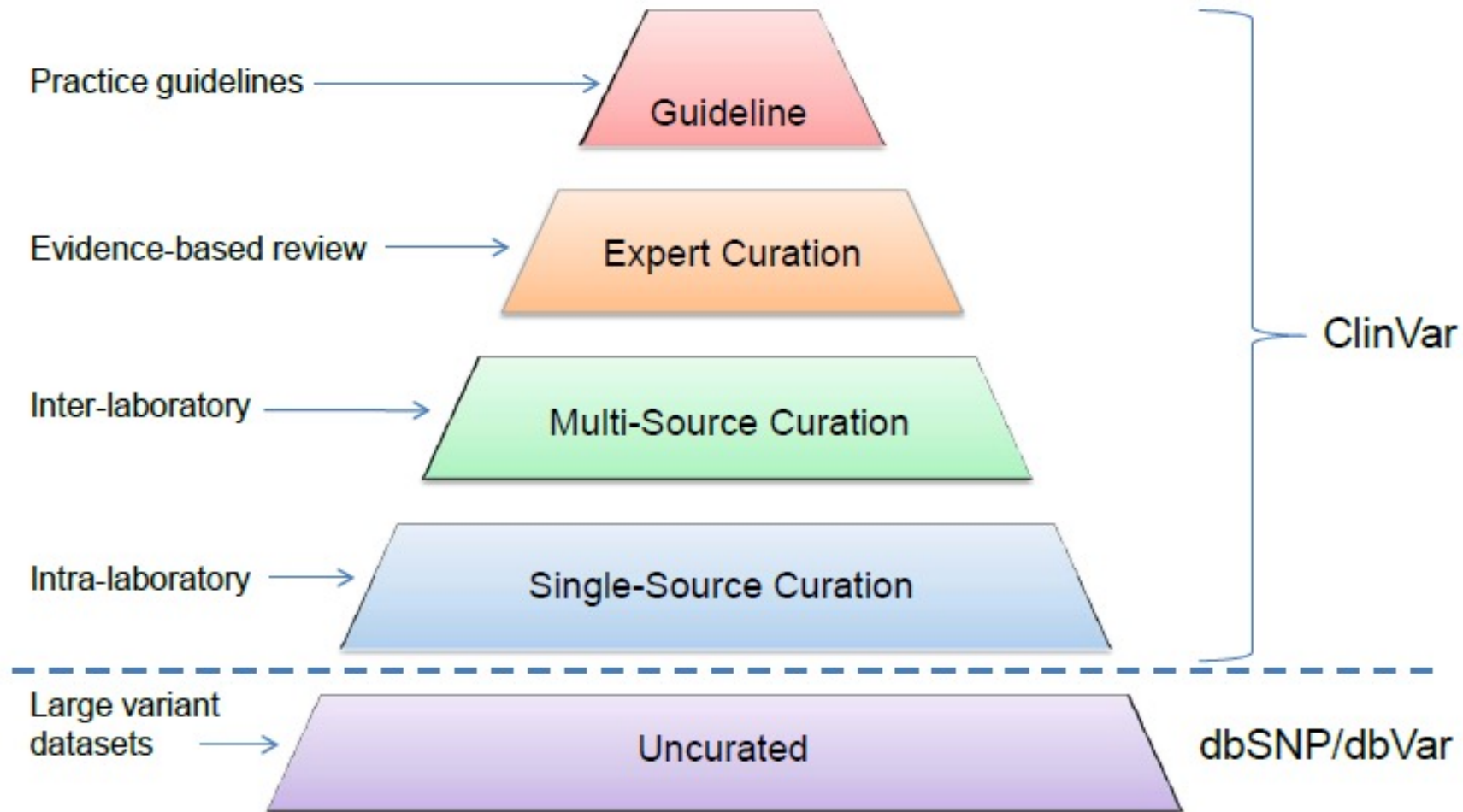
- Gene
- Variant type
- Genomic location
- HGVS expressions*
- Molecular consequence*
- Links*
- Frequency*

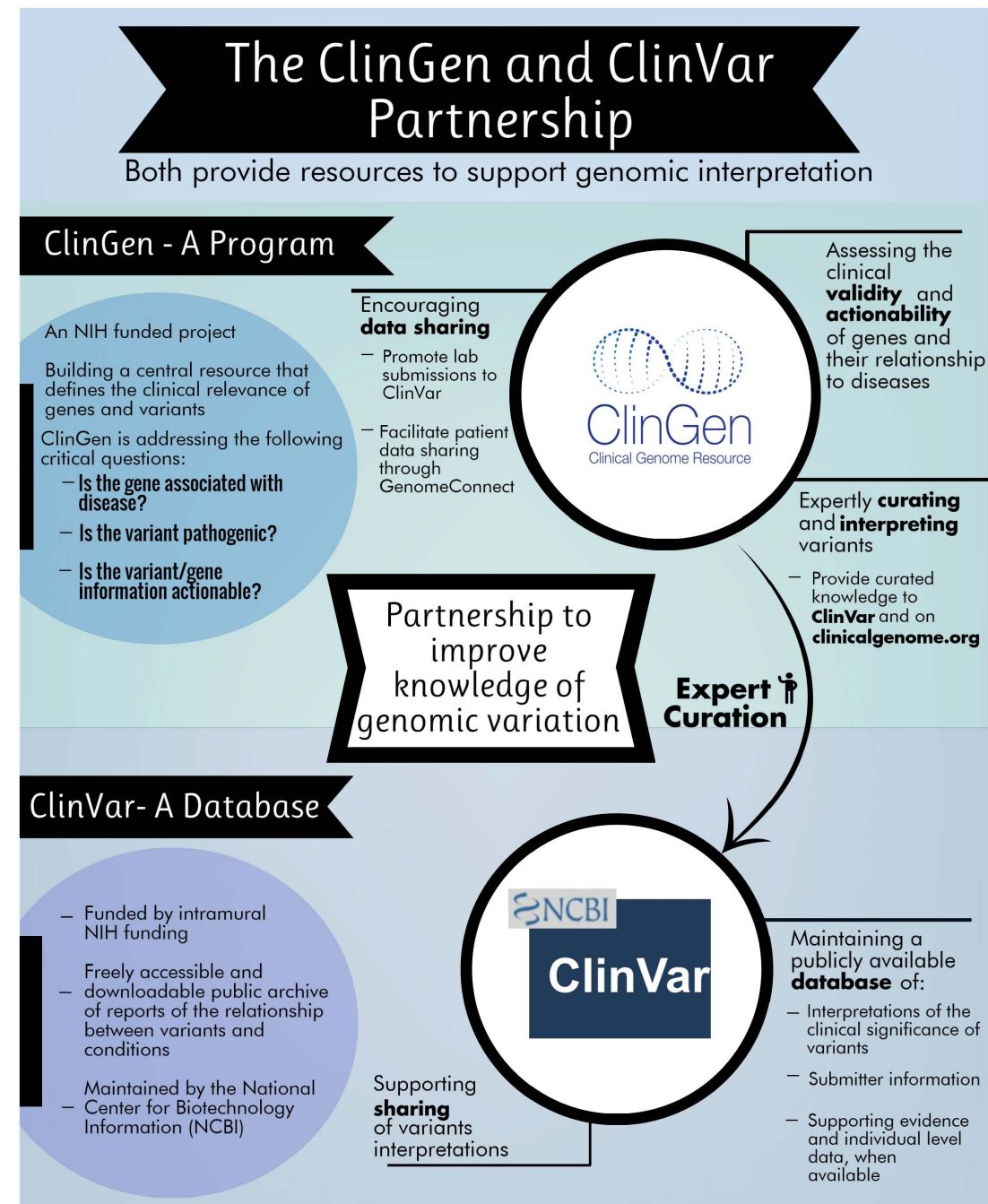
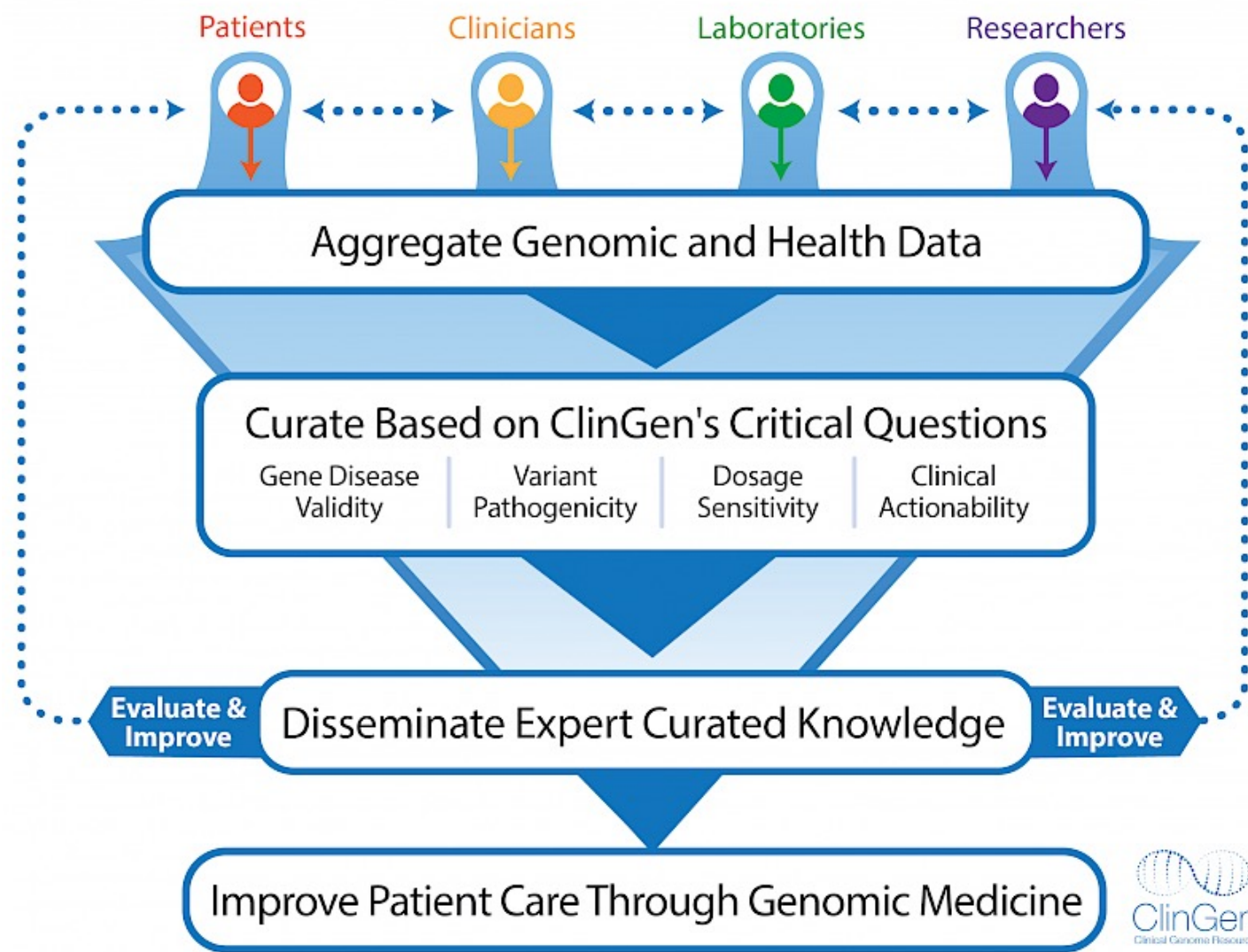
Phenotype summary

- Names
- Links*
- Age of onset *
- Prevalence *

Levels of Curation:

Maintaining quality and supporting multiple uses





varsome

GJB2:V37I

×

hg19

▼

Search

Editions

About

Community

News

Demo

Sign in

Join

chr13-20763612-C-T (GJB2:p.V37I)

1 user classified this variant as Pathogenic.

Link a publication

Classify

Community Contributions (1)

Favorites

Copy Shortlink

API Link

Submit to ClinVar

Upload FASTQ/VCF

Download Variant

Variant

Protein Viewer

Chromosome

chr13

UCSC genome browser

Mastermind

TraP Score

Position

20763612

REF Sequence

C

ALT Sequence

T

Variant type

SNV

Cytoband

13q12.11

HGVS

GJB2(NM_004004.6):c.109G>A (p.Val37Ile)

RS ID

rs72474224 | dbSNP

Gene symbol

GJB2

This variant has been viewed 1996 times on VarSome.

Connect with past and future viewers of this variant...

⚠ VarSome.com is for research use only. Find out about our clinically certified platform: VarSome Clinical.

ACMG Classification - Educational use only

Version: 11.2.8

Terms of use

Documentation

Options

Verdict

Pathogenic

Variant basic info

ACMG classification

Region browser

Community contributions1

Publications134

Structural Variants

Transcripts

ClinVar

UniProt Variants

Frequencies

Pathogenicity Scores

Conservation Scores

GTEx

PharmGKB

Beacon network

Gene GJB2

2025-11-05

#3192268

基因體暨蛋白體醫學研究所

Graduate Institute of Medical Genomics and Proteomics

124

https://varsome.com/variant/hg19/GJB2%3AV37I?annotation-mode=germline

Sequence Variant Interpretation

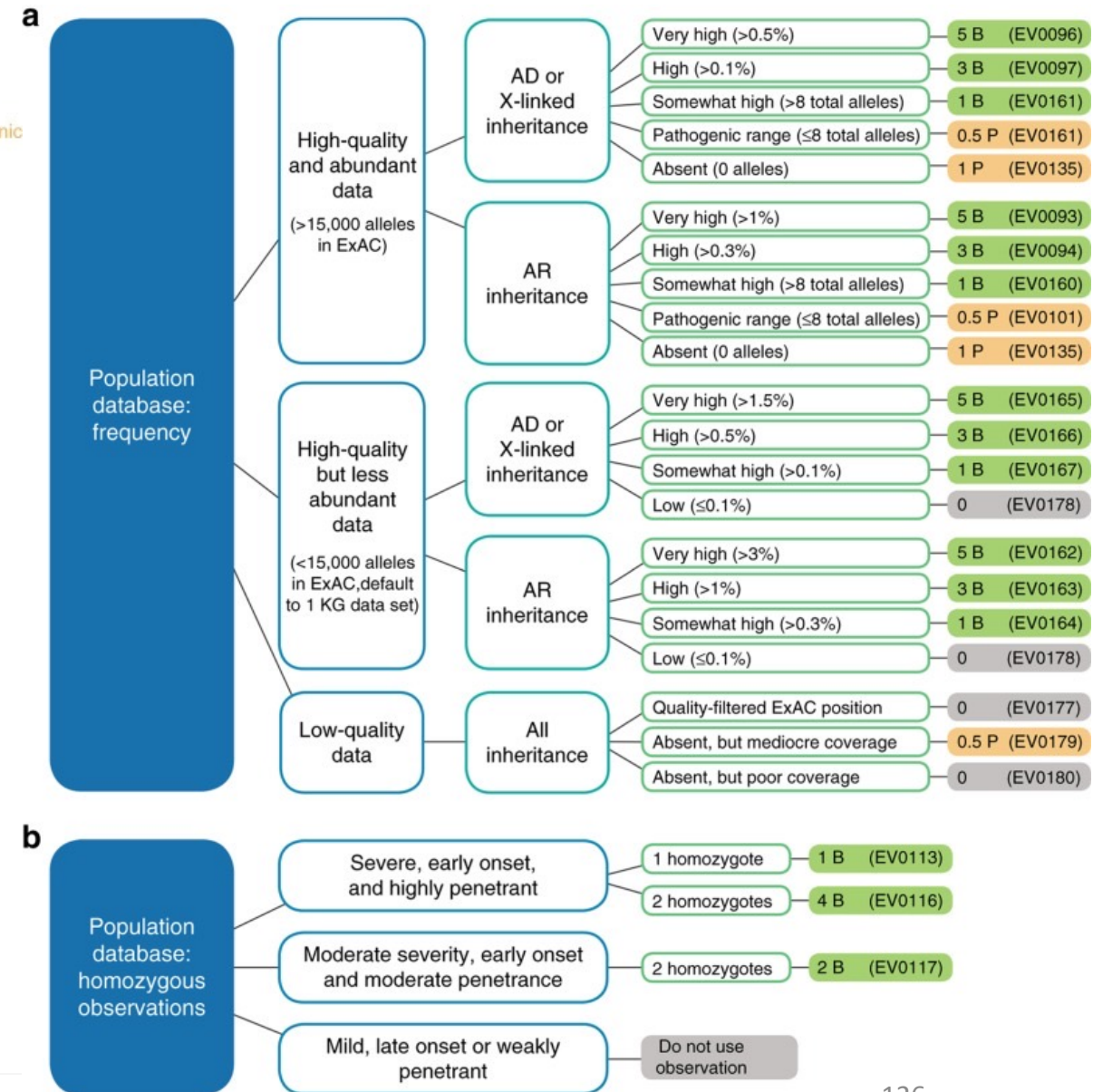
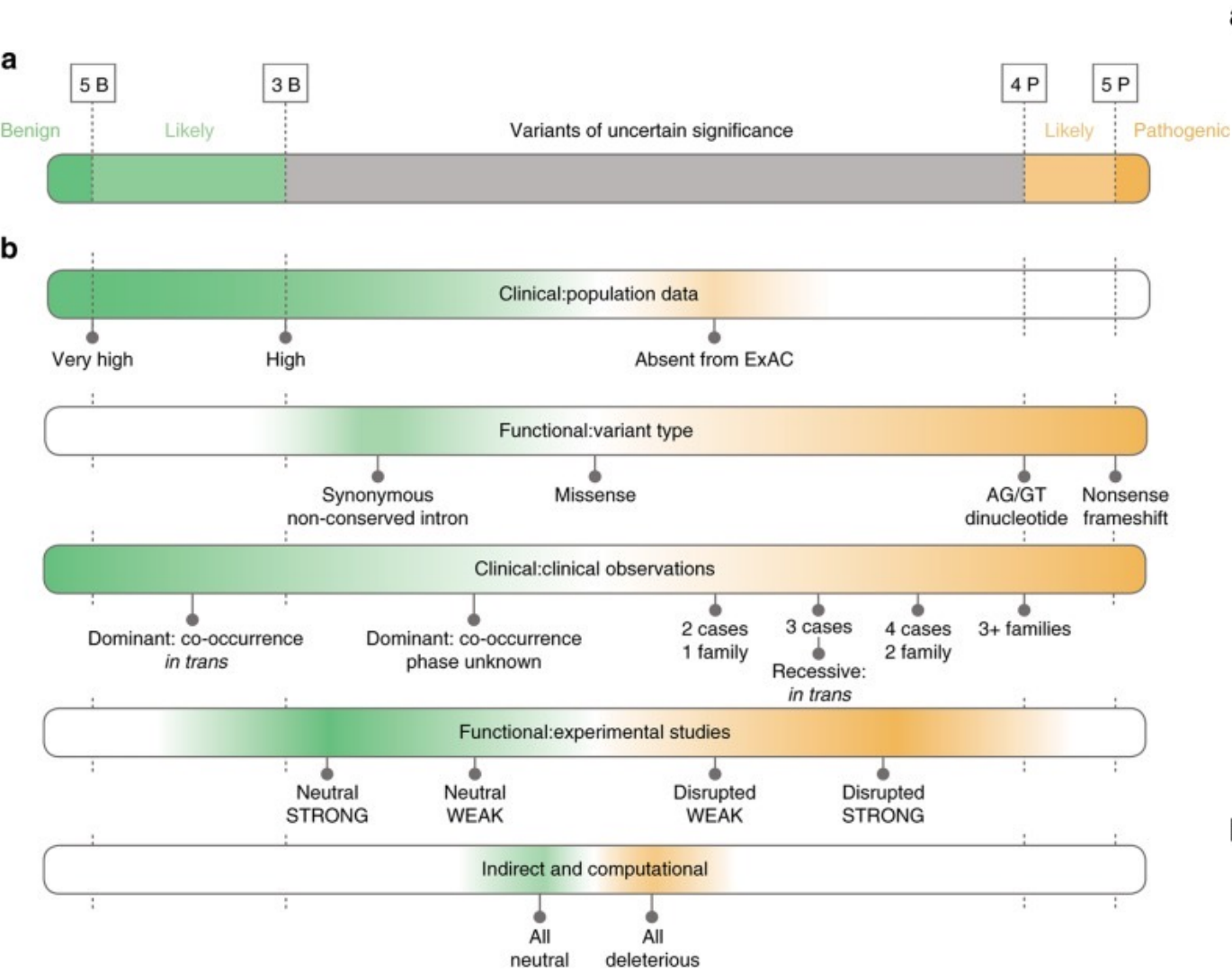
The goal of the Sequence Variant Interpretation Working Group (SVI WG) is to support the refinement and evolution of the [ACMG/AMP Interpreting Sequence Variant Guidelines](#) to develop quantitative approaches to variant interpretation.

[Subgroups](#)[Documents](#)[Tools](#)[Membership](#)

- **BA1**: Updated Recommendation for the ACMG/AMP Stand Alone Pathogenicity Criterion for Variant Classification
 - BA1 Exception List (July 2018)
 - BA1 Exception List Nomination Form
- **PVS1**: Recommendations for Interpreting the Loss of Function PVS1 ACMG/AMP Variant Criteria
 - ClinGen SVI PVS1 Decision Tree Editable
- **PVS1, PS1, PP3, BP4, BP7**: Application of the ACMG/AMP Framework to Capture Evidence Relevant to Predicted and Observed Impact on Splicing: Recommendations from the ClinGen SVI Splicing Subgroup
 - ClinGen SVI Splicing Subgroup feedback form
 - ClinGen SVI Splicing Subgroup - Response to feedback (March 2024)
- **PS2/PM6**: Recommendation for de novo PS2 and PM6 ACMG/AMP criteria (Version 1.1)
- **PS3/BS3**: Recommendations for application of the functional evidence PS3/BS3 criterion using the ACMG/AMP sequence variant interpretation framework
 - SVI Functional Assay Documentation Worksheet
- **PM2**: Recommendation for Absence/Rarity Criterion PM2 (Version 1.0)
- **PM3**: Recommendation for in trans Criterion PM3 (Version 1.0)
- **PP1/BS4, PP4**: ClinGen guidance for use of the PP1/BS4 co-segregation and PP4 phenotype specificity criteria for sequence variant pathogenicity classification
- **PP3/BP4**: Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4
- **PP5/BP6**: Recommendation for reputable source PP5 and BP6 ACMG/AMP criteria



SHERLOC refinement (2017)

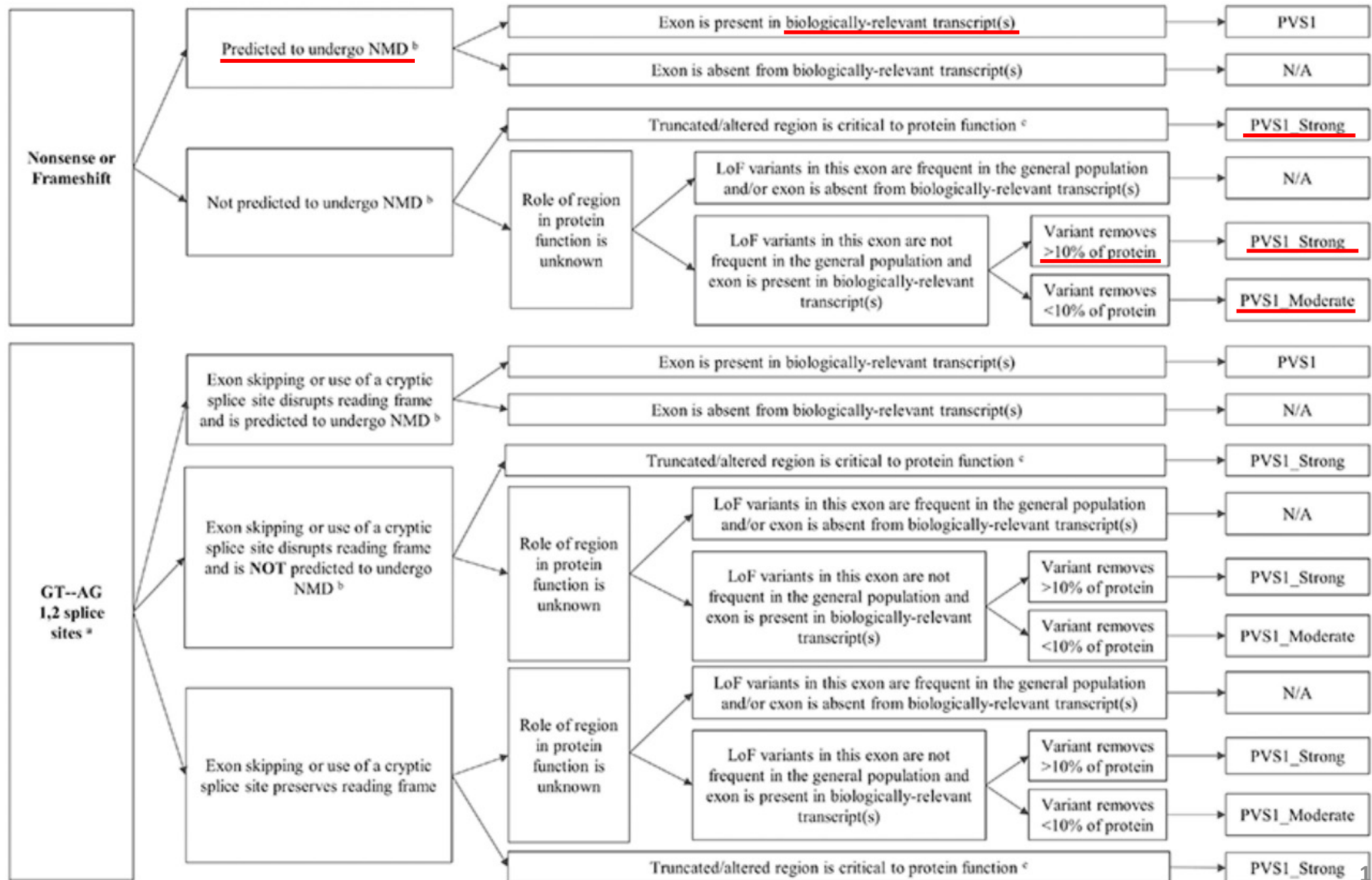


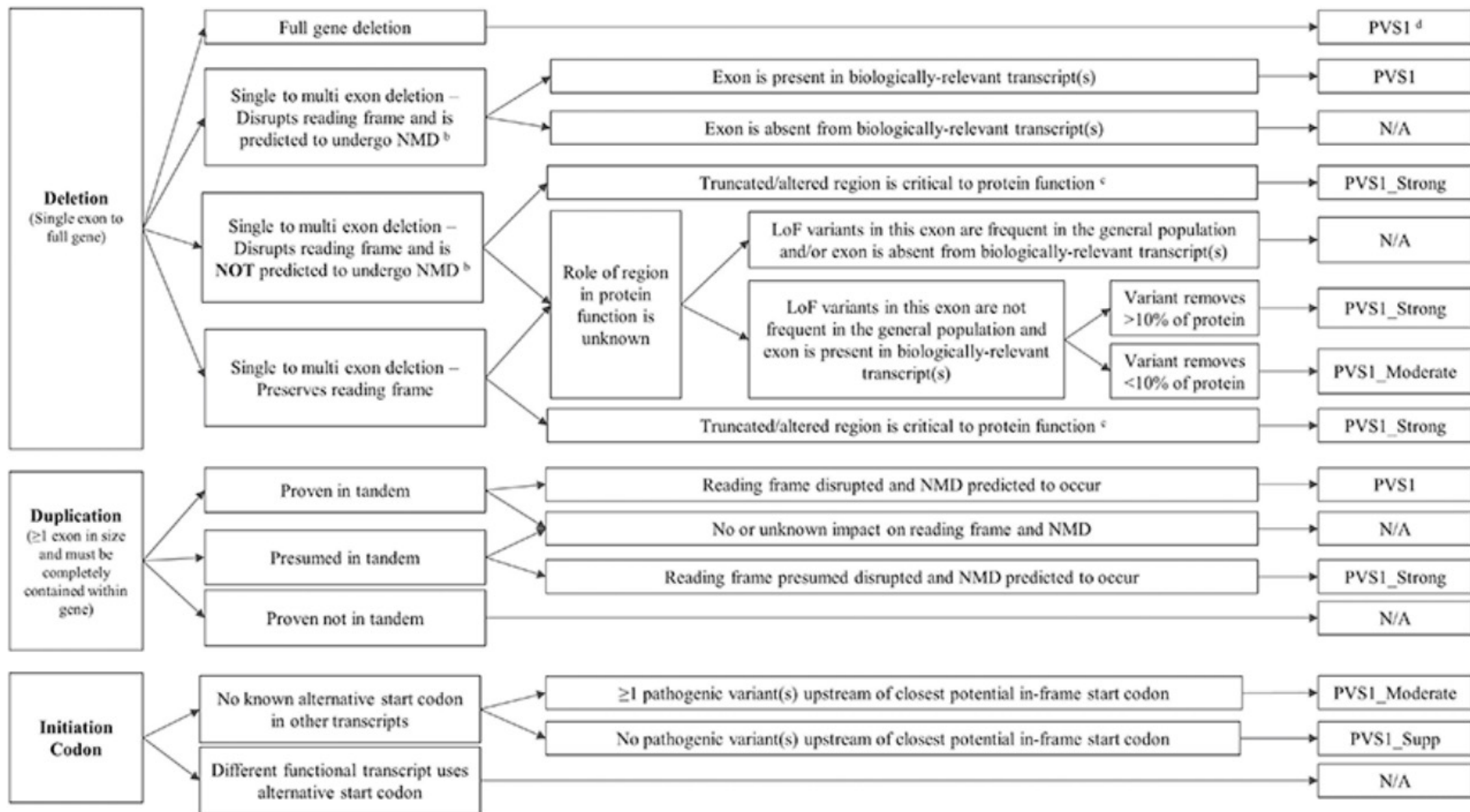
	Benign		Pathogenic			
	Strong	Supporting	Supporting	Moderate	Strong	Very strong
Population data	MAF is too high for disorder BA1/BS1 OR observation in controls inconsistent with disease penetrance BS2			Absent in population databases PM2	Prevalence in affecteds statistically increased over controls PS4	
Computational and predictive data		Multiple lines of computational evidence suggest no impact on gene /gene product BP4 Missense in gene where only truncating cause disease BP1 Silent variant with non predicted splice impact BP7 In-frame indels in repeat w/out known function BP3	Multiple lines of computational evidence support a deleterious effect on the gene /gene product PP3	Novel missense change at an amino acid residue where a different pathogenic missense change has been seen before PM5 Protein length changing variant PM4	Same amino acid change as an established pathogenic variant PS1	Predicted null variant in a gene where LOF is a known mechanism of disease PVS1
Functional data	Well-established functional studies show no deleterious effect BS3		Missense in gene with low rate of benign missense variants and path. missenses common PP2	Mutational hot spot or well-studied functional domain without benign variation PM1	Well-established functional studies show a deleterious effect PS3	
Segregation data	Nonsegregation with disease BS4		Cosegregation with disease in multiple affected family members PP1	Increased segregation data →		
De novo data				De novo (without paternity & maternity confirmed) PM6	De novo (paternity and maternity confirmed) PS2	
Allelic data		Observed in trans with a dominant variant BP2 Observed in cis with a pathogenic variant BP2		For recessive disorders, detected in trans with a pathogenic variant PM3		
Other database		Reputable source w/out shared data = benign BP6	Reputable source = pathogenic PP5			
Other data		Found in case with an alternate cause BP5	Patient's phenotype or FH highly specific for gene PP4			

Table 5 Rules for combining criteria to classify sequence variants

Pathogenic	(i) 1 Very strong (PVS1) <u>AND</u> (a) ≥1 Strong (PS1–PS4) OR (b) ≥2 Moderate (PM1–PM6) OR (c) 1 Moderate (PM1–PM6) and 1 supporting (PP1–PP5) OR (d) ≥2 Supporting (PP1–PP5) (ii) ≥2 Strong (PS1–PS4) OR (iii) 1 Strong (PS1–PS4) <u>AND</u> (a) ≥3 Moderate (PM1–PM6) OR (b) 2 Moderate (PM1–PM6) <u>AND</u> ≥2 Supporting (PP1–PP5) OR (c) 1 Moderate (PM1–PM6) <u>AND</u> ≥4 supporting (PP1–PP5)
------------	---







	Benign		Pathogenic			
	Strong	Supporting	Supporting	Moderate	Strong	Very strong
Population data	MAF is too high for disorder BA1/BS1 OR observation in controls inconsistent with disease penetrance BS2			Absent in population databases PM2	Prevalence in affecteds statistically increased over controls PS4	
Computational and predictive data		Multiple lines of computational evidence suggest no impact on gene /gene product BP4 Missense in gene where only truncating cause disease BP1 Silent variant with non predicted splice impact BP7 In-frame indels in repeat w/out known function BP3	Multiple lines of computational evidence support a deleterious effect on the gene /gene product PP3	Novel missense change at an amino acid residue where a different pathogenic missense change has been seen before PM5 Protein length changing variant PM4	Same amino acid change as an established pathogenic variant PS1	Predicted null variant in a gene where LOF is a known mechanism of disease PVS1
Functional data	Well-established functional studies show no deleterious effect BS3		Missense in gene with low rate of benign missense variants and path. missenses common PP2	Mutational hot spot or well-studied functional domain without benign variation PM1	Well-established functional studies show a deleterious effect PS3	
Segregation data	Nonsegregation with disease BS4		Cosegregation with disease in multiple affected family members PP1	Increased segregation data →		
De novo data				De novo (without paternity & maternity confirmed) PM6	De novo (paternity and maternity confirmed) PS2	
Allelic data		Observed in trans with a dominant variant BP2 Observed in cis with a pathogenic variant BP2		For recessive disorders, detected in trans with a pathogenic variant PM3		
Other database		Reputable source w/out shared data = benign BP6	Reputable source = pathogenic PP5			
Other data		Found in case with an alternate cause BP5	Patient's phenotype or FH highly specific for gene PP4			

Table 5 Rules for combining criteria to classify sequence variants

Pathogenic	(i) 1 Very strong (PVS1) AND (a) ≥ 1 Strong (PS1–PS4) OR (b) ≥ 2 Moderate (PM1–PM6) OR (c) 1 Moderate (PM1–PM6) and 1 supporting (PP1–PP5) OR (d) ≥ 2 Supporting (PP1–PP5) (ii) ≥ 2 Strong (PS1–PS4) OR (iii) 1 Strong (PS1–PS4) AND (a) ≥ 3 Moderate (PM1–PM6) OR (b) 2 Moderate (PM1–PM6) AND ≥ 2 Supporting (PP1–PP5) OR (c) 1 Moderate (PM1–PM6) AND ≥ 4 supporting (PP1–PP5)
------------	--

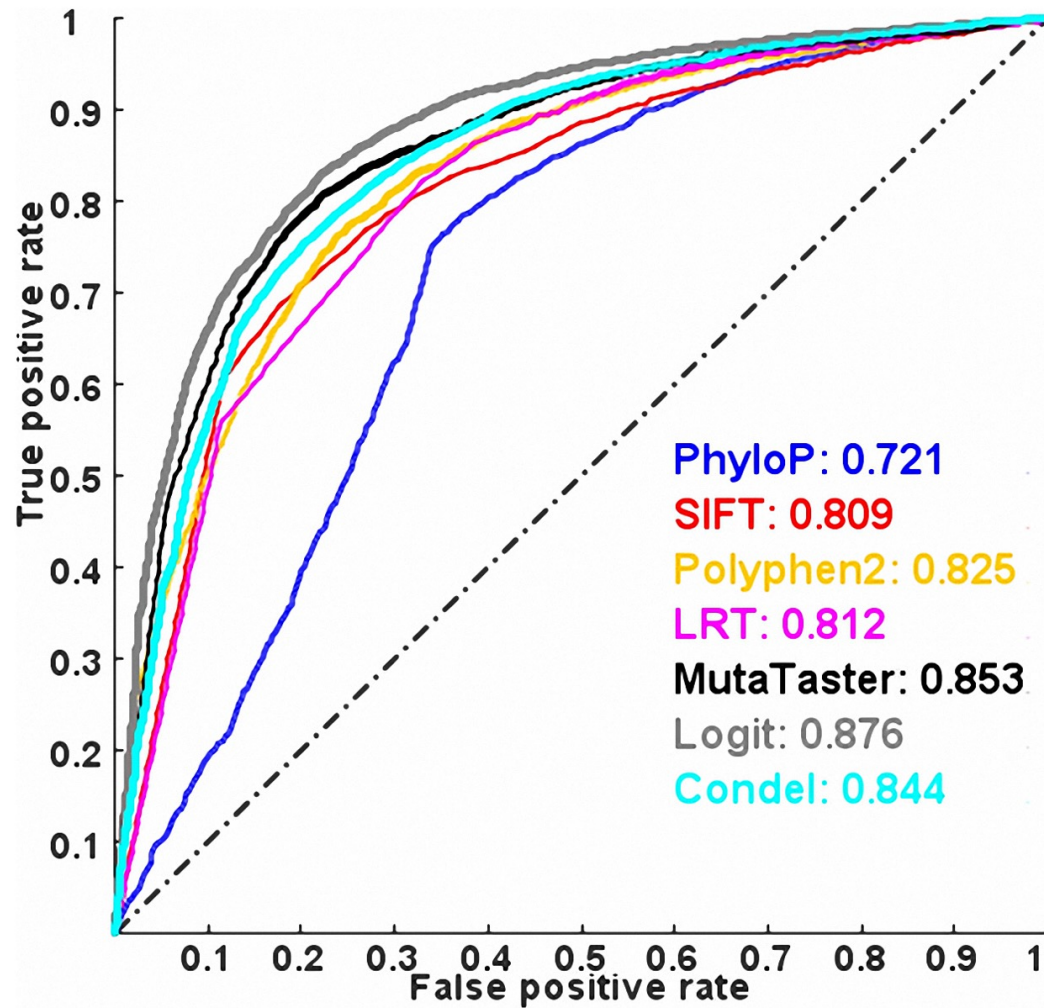


Missense prediction (PP3)

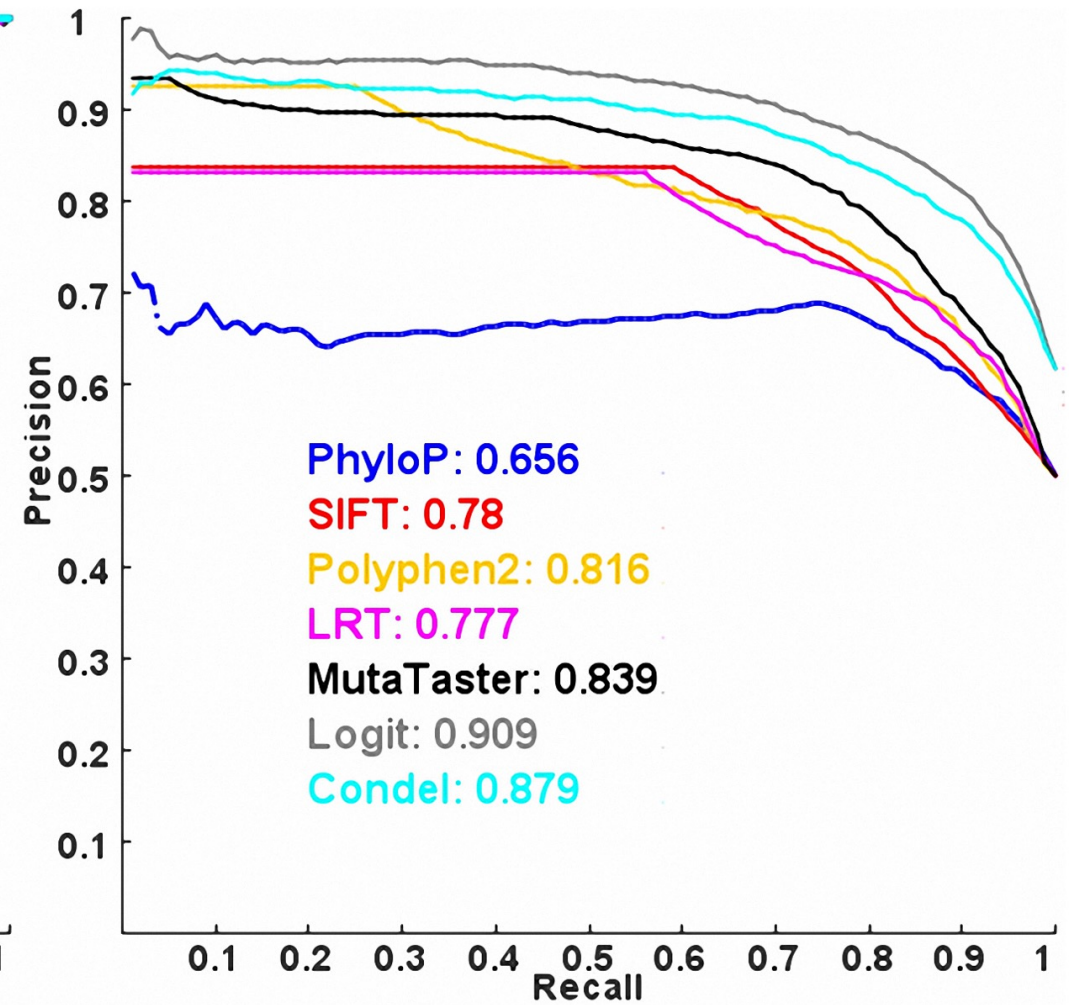
1. Evolutionary conservation
 2. Protein structure/function
 3. Both 1. & 2.
 4. Alignment and the similarity between sequence homolog
 5. 2. & 4.
 6. multiple 1. 2. 3.
 7. Intergration, combination
- If **all** of the in silico programs tested agree on the prediction, then this evidence can be counted as supporting.
 - If in silico predictions disagree, however, then this evidence should not be used in classifying a variant.

Name	Website	Basis
ConSurf	http://consurftest.tau.ac.il	Evolutionary conservation
FATHMM	http://fathmm.biocompute.org.uk	Evolutionary conservation
MutationAssessor	http://mutationassessor.org	Evolutionary conservation
PANTHER	http://www.pantherdb.org/tools/csnpscoreForm.jsp	Evolutionary conservation
PhD-SNP	http://snps.biofold.org/phd-snp/phd-snp.html	Evolutionary conservation
SIFT	http://sift.jcvi.org	Evolutionary conservation
SNPs&GO	http://snps-and-go.biocomp.unibo.it/snps-and-go	Protein structure/function
Align GVGD	http://agvgd.iarc.fr/agvgd_input.php	Protein structure/function and evolutionary conservation
MAPP	http://mendel.stanford.edu/SidowLab/downloads/MAPP/index.html	Protein structure/function and evolutionary conservation
MutationTaster	http://www.mutationtaster.org	Protein structure/function and evolutionary conservation
MutPred	http://mutpred.mutdb.org	Protein structure/function and evolutionary conservation
PolyPhen-2	http://genetics.bwh.harvard.edu/pph2	Protein structure/function and evolutionary conservation
PROVEAN	http://provean.jcvi.org/index.php	Alignment and measurement of similarity between variant sequence and protein sequence homolog
nsSNPAnalyzer	http://snpanalyzer.uthsc.edu	Multiple sequence alignment and protein structure analysis
Condel	http://bg.upf.edu/fannssdb/	Combines SIFT, PolyPhen-2, and MutationAssessor
CADD	http://cadd.gs.washington.edu	Contrasts annotations of fixed/nearly fixed derived alleles in humans with simulated variants

Machine Learning Method



(a)



(b)

CADD-Combined Annotation Dependent Depletion

- Mutation rate and sequence evolution, sequence conservation metrics like GERP, phastCons, and phyloP
- Functional genomic data like DNase hypersensitivity and transcription factor binding
- Transcript information like distance to exon-intron boundaries and Expression levels in commonly studied cell lines
- Protein-level scores like Grantham, SIFT, and PolyPhen
- PHRED-like ($-10 \cdot \log_{10}(\text{rank}/\text{total})$) scaled C-score:
 - C-score ≥ 10 indicates that it's predicted to be the 10% most deleterious substitutions
 - C-score ≥ 20 indicates that it's predicted to be the 1% most deleterious substitutions

Research | [Open Access](#) | Published: 22 February 2021

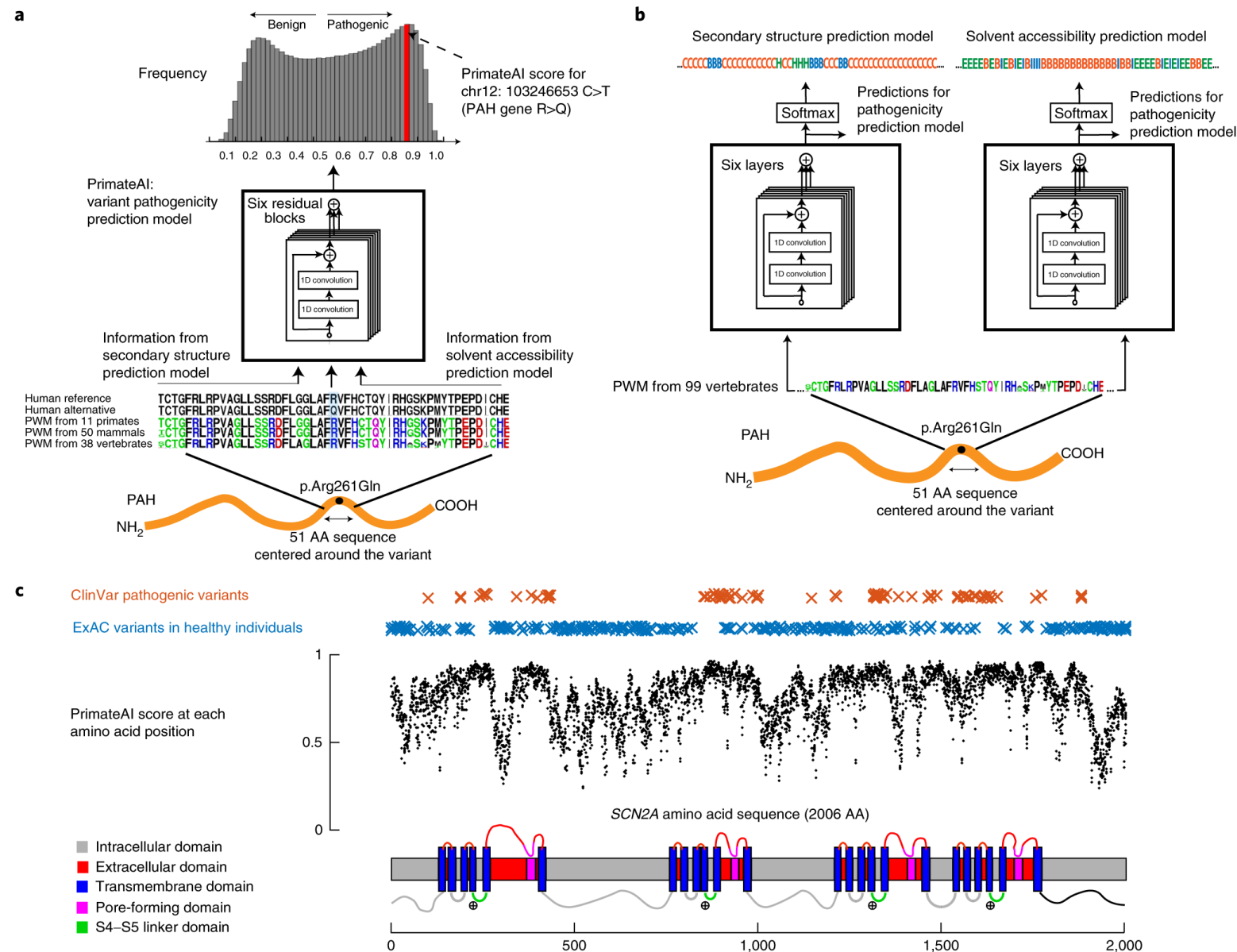
CADD-Splice—improving genome-wide variant effect prediction using deep learning-derived splice scores

[Philipp Rentzsch](#), [Max Schubach](#), [Jay Shendure](#) & [Martin Kircher](#) 

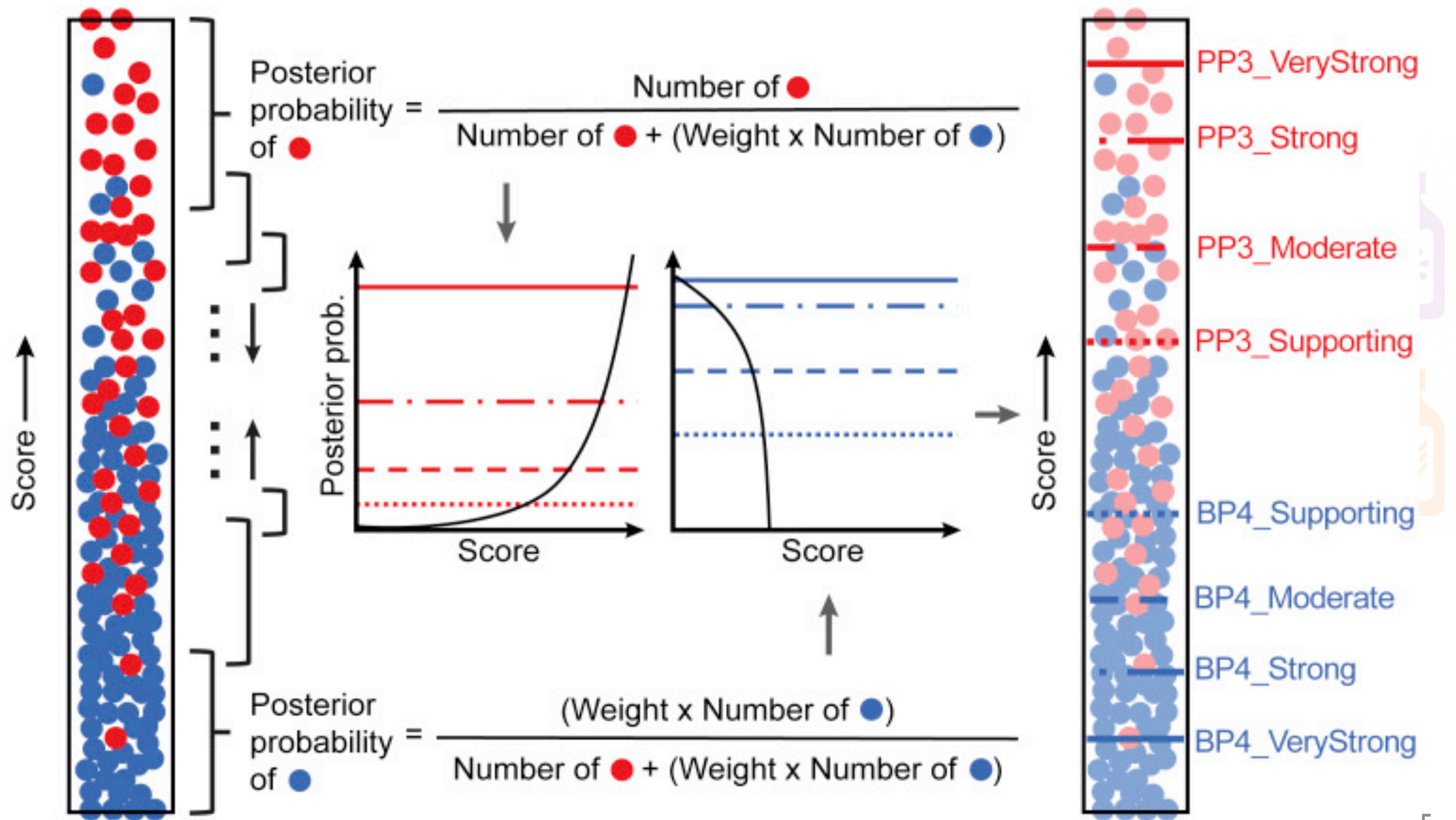
[Genome Medicine](#) **13**, Article number: 31 (2021) | [Cite this article](#)

PrimateAI

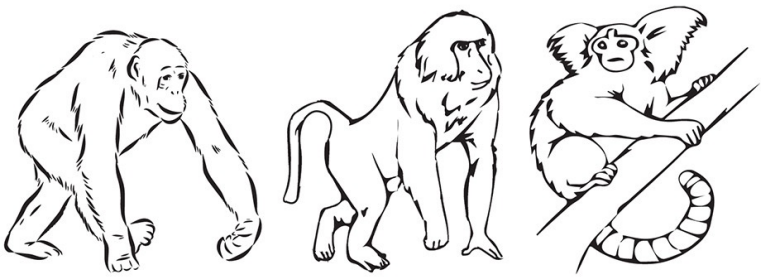
- Common missense variants in other primate species are largely clinically benign in human (deep neural network)
- The network takes human amino acid (AA) reference and alternate sequence (51 AAs) centered at the variant, the position weight matrix (PWM) conservation profiles calculated from 99 vertebrate species
- secondary structure and solvent accessibility



● Pathogenic/likely pathogenic variant ● Benign/likely benign variant

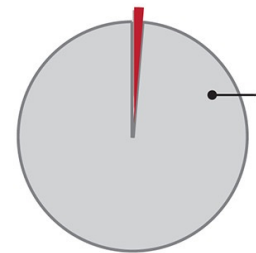


PrimateAI-3D

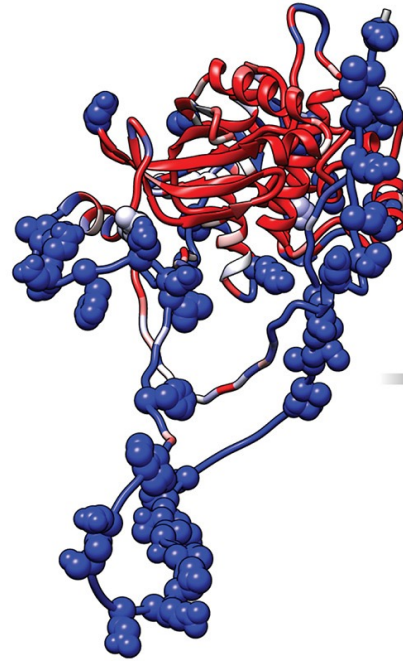


4.3 million common benign variants from 233 primate species

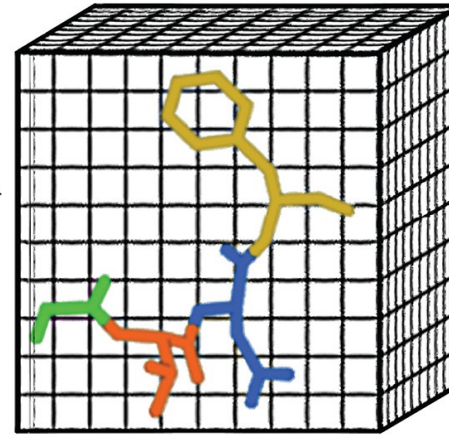
Validation of primate variants in human clinical variant database



98.7% of common primate variants in ClinVar are benign

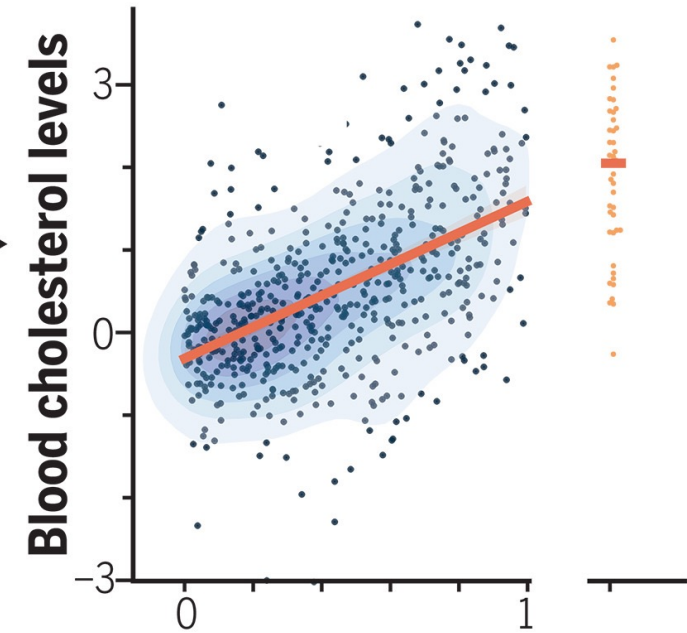


Benign primate variants superimposed on 3D protein structures



3D convolutions + deep learning protein language models

Individuals carrying LDLR variants



PrimateAI-3D score

LoF

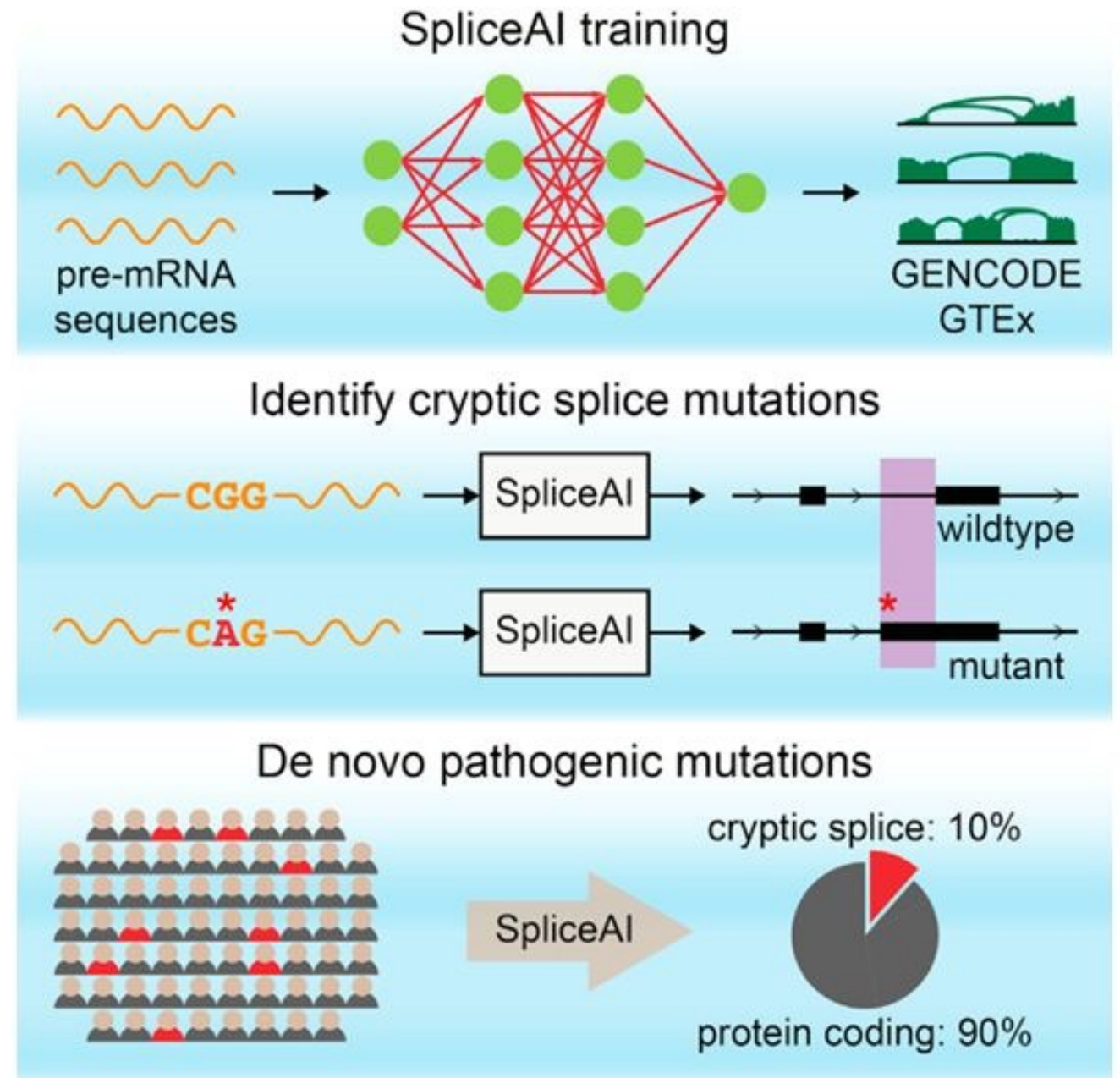
Validation of variant effect predictions in clinical cohorts

Splice site and others

Splice site prediction	GeneSplicer	http://www.cbcb.umd.edu/software/GeneSplicer/gene_spl.shtml	Markov models
	Human Splicing Finder	http://www.umd.be/HSF/	Position-dependent logic
	MaxEntScan	http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html	Maximum entropy principle
	NetGene2	http://www.cbs.dtu.dk/services/NetGene2	Neural networks
	NNSplice	http://www.fruitfly.org/seq_tools/splice.html	Neural networks
Nucleotide conservation prediction	FSPLICE	http://www.softberry.com/berry.phtml?topic=fsplce&group=programs&subgroup=gfind	Species-specific predictor for splice sites based on weight matrices model
	GERP	http://mendel.stanford.edu/sidowlab/downloads/gerp/index.html	Genomic evolutionary rate profiling
	PhastCons	http://compugen.bscb.cornell.edu/phast/	Conservation scoring and identification of conserved elements
	PhyloP	http://compugen.bscb.cornell.edu/phast/	Alignment and phylogenetic trees: Computation of <i>P</i> values for conservation or acceleration, either lineage-specific or across all branches

SpliceAI

- SpliceAI, a 32-layer deep neural network, predicts splicing from a pre-mRNA sequence
- 75% of predicted cryptic splice variants validate on RNA-seq
- Cryptic splicing may yield ~10% of pathogenic variants in neurodevelopmental disorders
- Cryptic splice variants frequently give rise to alternative splicing



AlphaMissense

Google DeepMind, London, UK.

- (i) Unsupervised protein language modeling to learn amino acid distributions conditioned on sequence context;
- (ii) Incorporating structural context by using an AlphaFold-derived system
- (iii) Fine-tuning on weak labels from population frequency data, thereby avoiding bias from human-curated annotations.
- (iv) MAVE benchmarked data

Input

Reference:

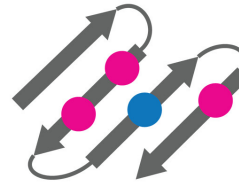
DNA: CAG
Protein: MDVVAMVNQTVATMIS

Missense variant:

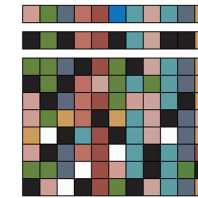
DNA: CGG
Protein: MDVVAMVNR TVATMIS

AlphaMissense

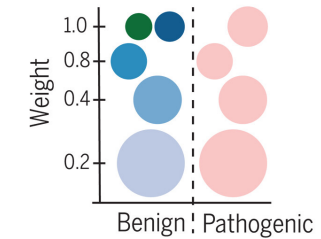
① Structure context



② Protein language modeling

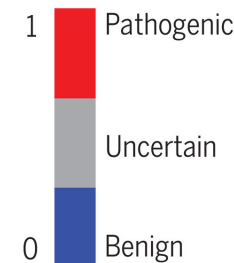


③ Training variants

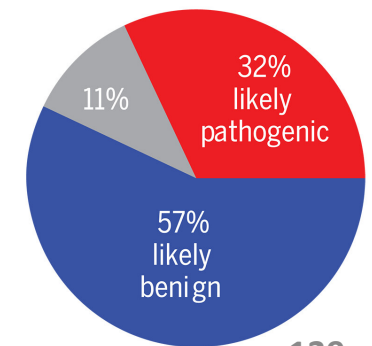


Output

Alpha Missense pathogenicity:



For all 71M possible missense variants in the human proteome:

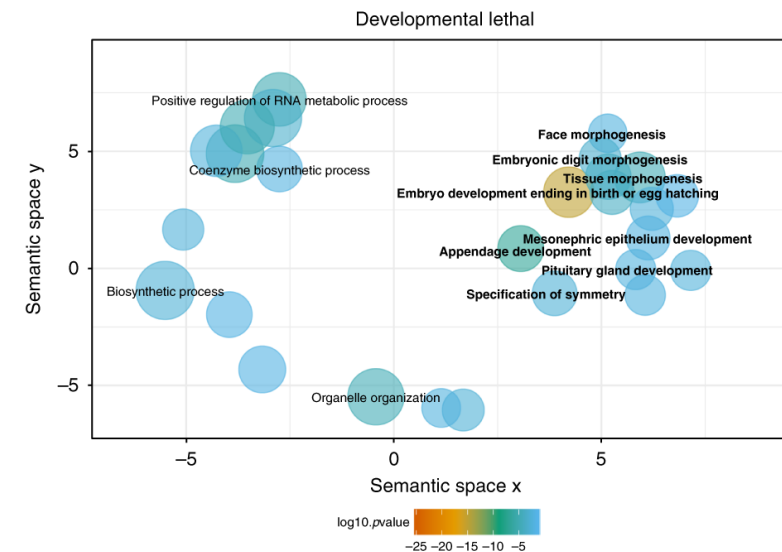
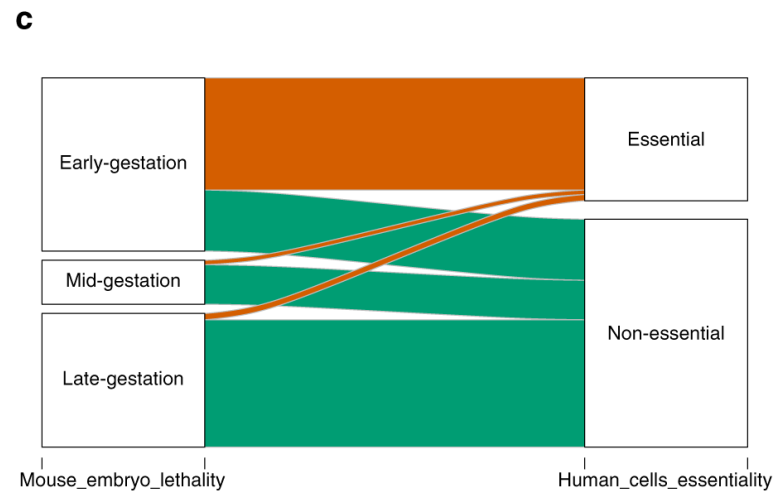
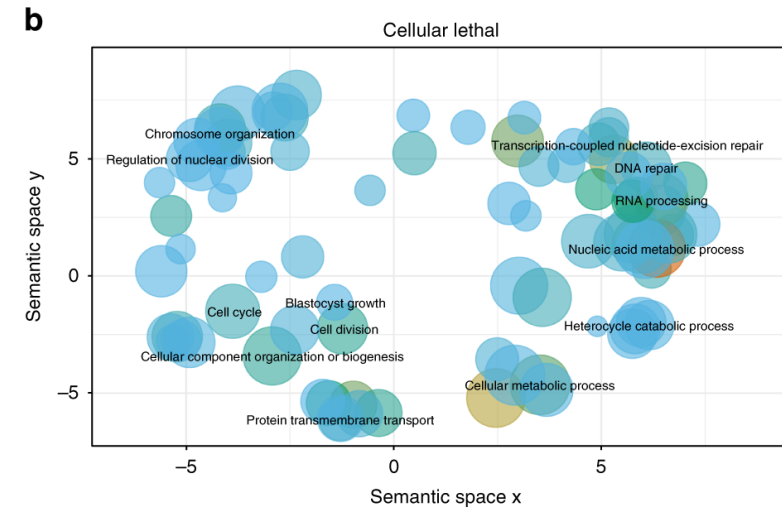
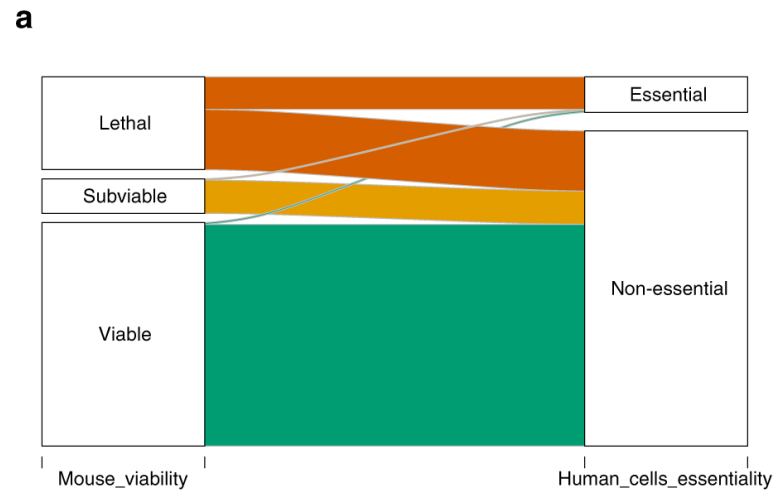


139

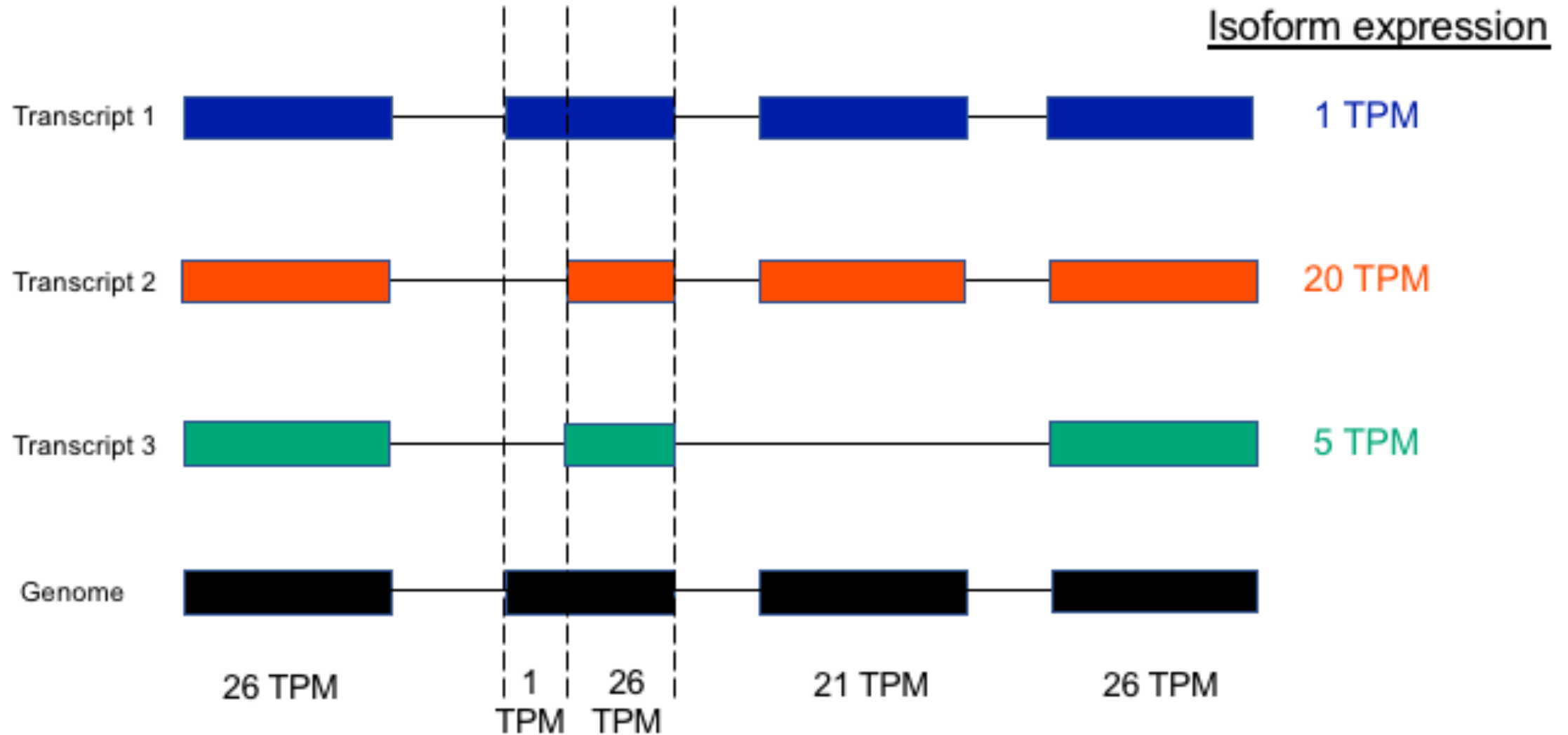


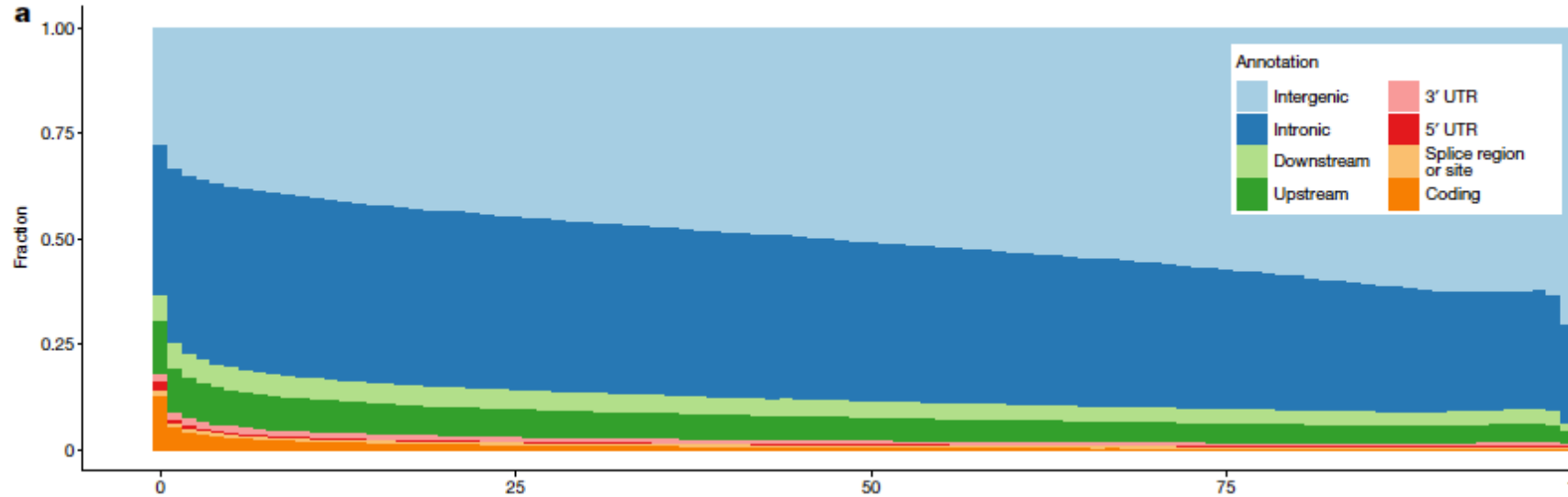
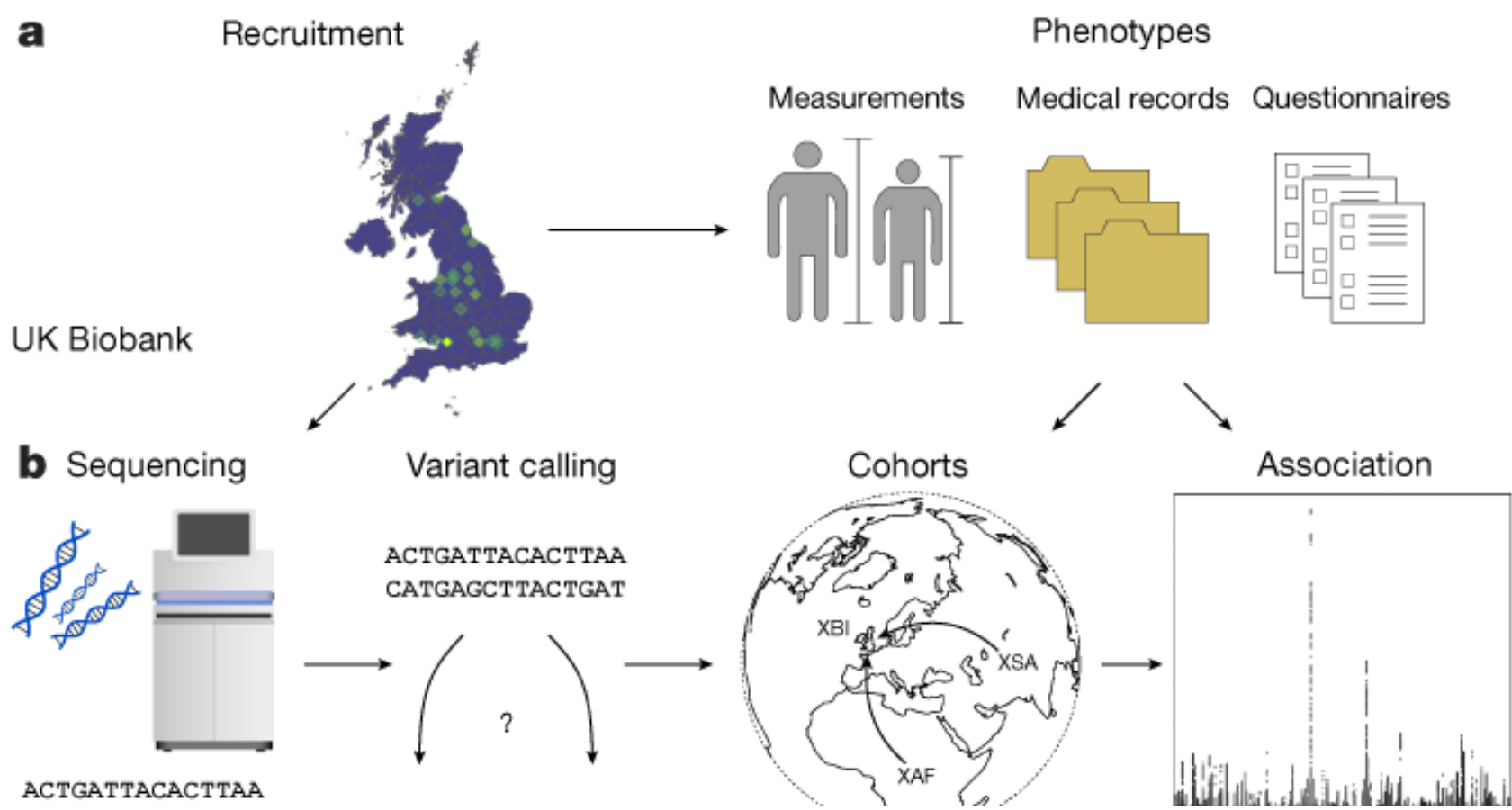
基因體暨蛋白質醫學研究所
Graduate Institute of Medical Genomics and Proteomics

Cross-species intolerance to LoF ()

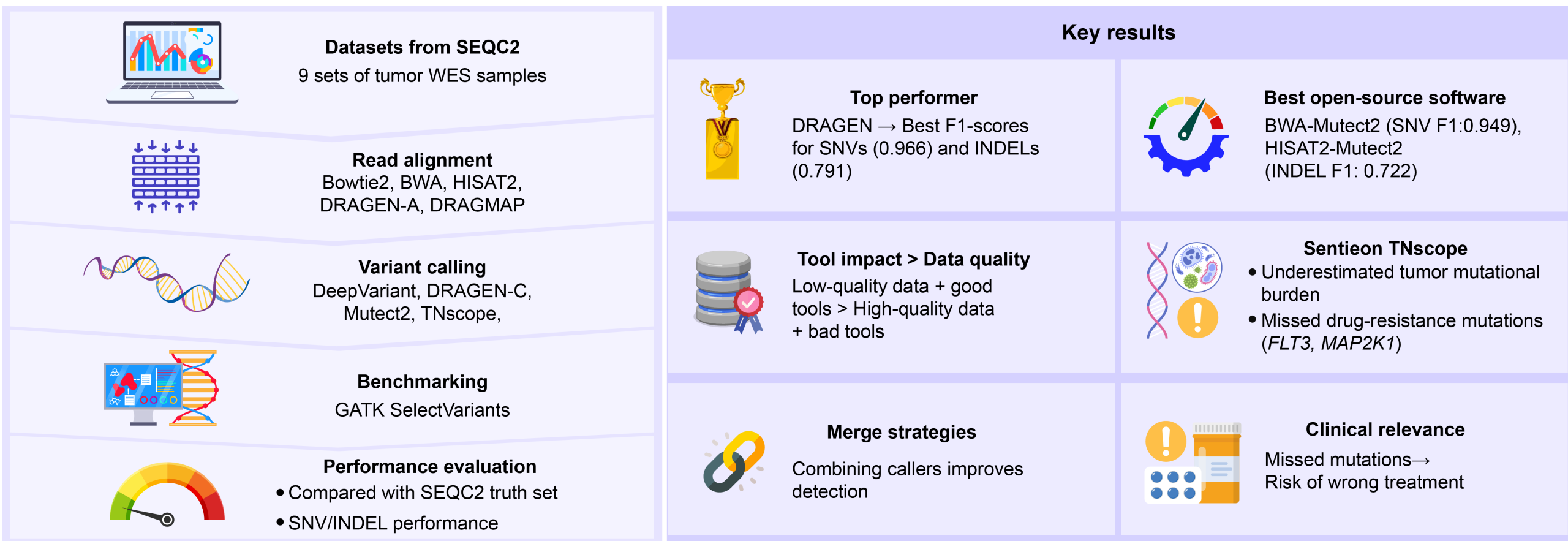


Proportion expressed across transcripts (pext) score





Benchmarking ctDNA WES analysis workflow



BWA: Burrows-wheeler aligner; DRAGEN: Dynamic read analysis for genomics; DRAGEN-A: DRAGEN-aligner; DRAGEN-C: DRAGEN-caller; DRAGMAP: DRAGEN open-source mapper; FLT3: FMS-like tyrosine kinase 3; FMS: Feline McDonough Sarcoma; GATK: Genome analysis toolkit; HISAT2: Hierarchical indexing for spliced alignment of transcripts 2; INDEL: Insertion-deletion; MAP2K1: Mitogen-activated protein kinase 1; SEQC2: Sequencing quality control phase 2; SNV: Single nucleotide variant; WES: Whole exome sequencing

Different analytical methods applied to the same data will have an impact on the final molecular diagnosis and treatment

Thank you for your attention

NTUH & NTU

陳信孚醫師

許家郎博士

陳慧玲醫師

陳沛隆醫師

吳振吉醫師

陳倩瑜教授

林靜嫻醫師

楊雅倩教授

張以承醫師

郭柏秀教授

陳達慶醫師

馮嫻臻教授

彭彥菱

陳品瑄

衛品妃

賴鴻昇

王世仔

黃益泓

黃冠達

李宜哲

曾宜萱

許鈞

李婕瑜

劉萬騏

李振宇

鄭敬譯

張廷宇

洪崇凱博士

陳薇安

莊惠文

朱家淇

蕭毅

周庭暄

賴勝凱

簡培苗

蔡心縈

吳東祈

江鈺婷

林俐伶

林弘曄



台大醫學院
National Taiwan University
College of Medicine



NSTC 國家科學及技術委員會
National Science and Technology Council



LIONS
國網生科雲

