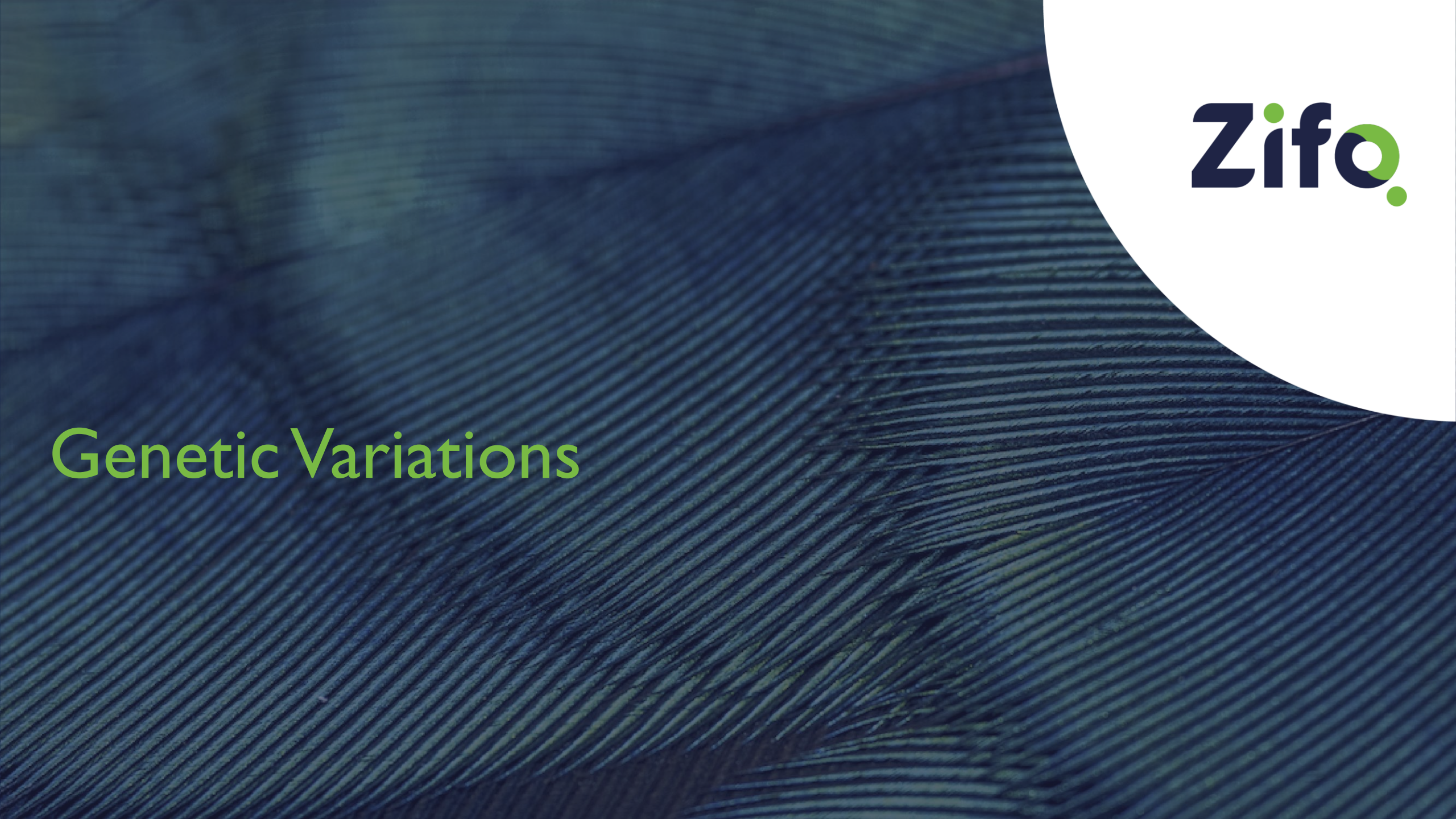
Genomic Variation

09/11/22

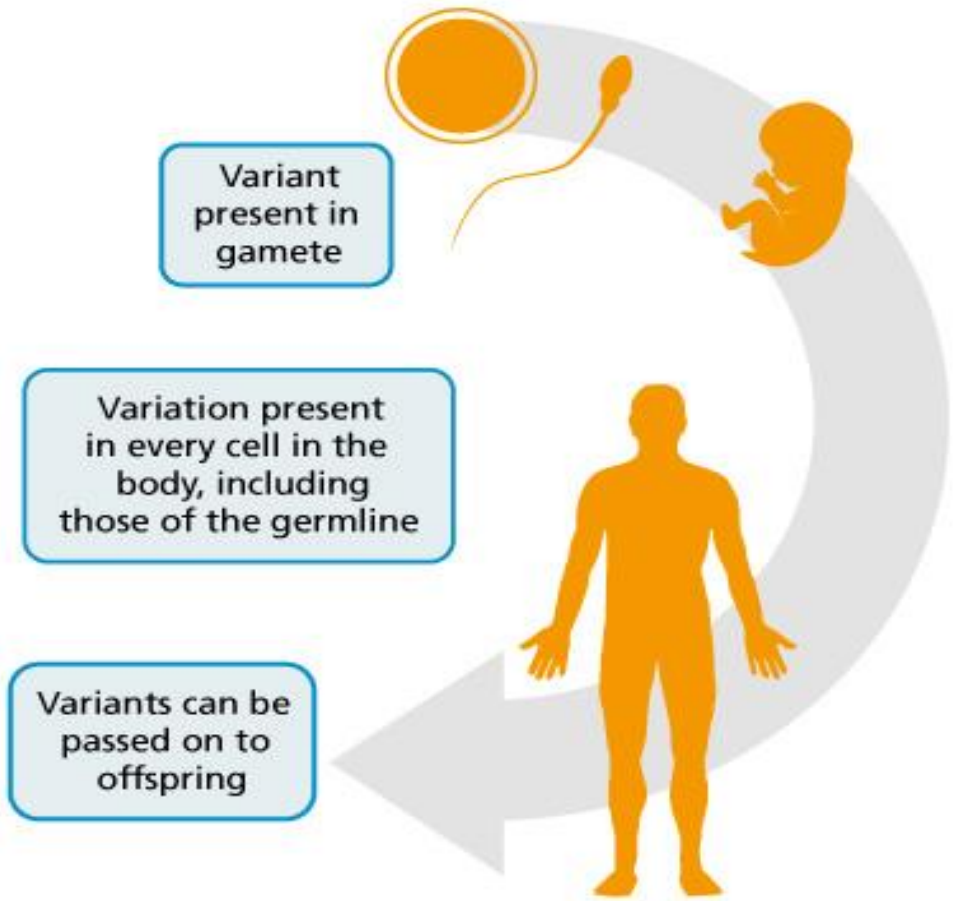
TOPICS COVERED

- What is genetic variation?
- Types of genetic variation studies
- Variant identification and analysis

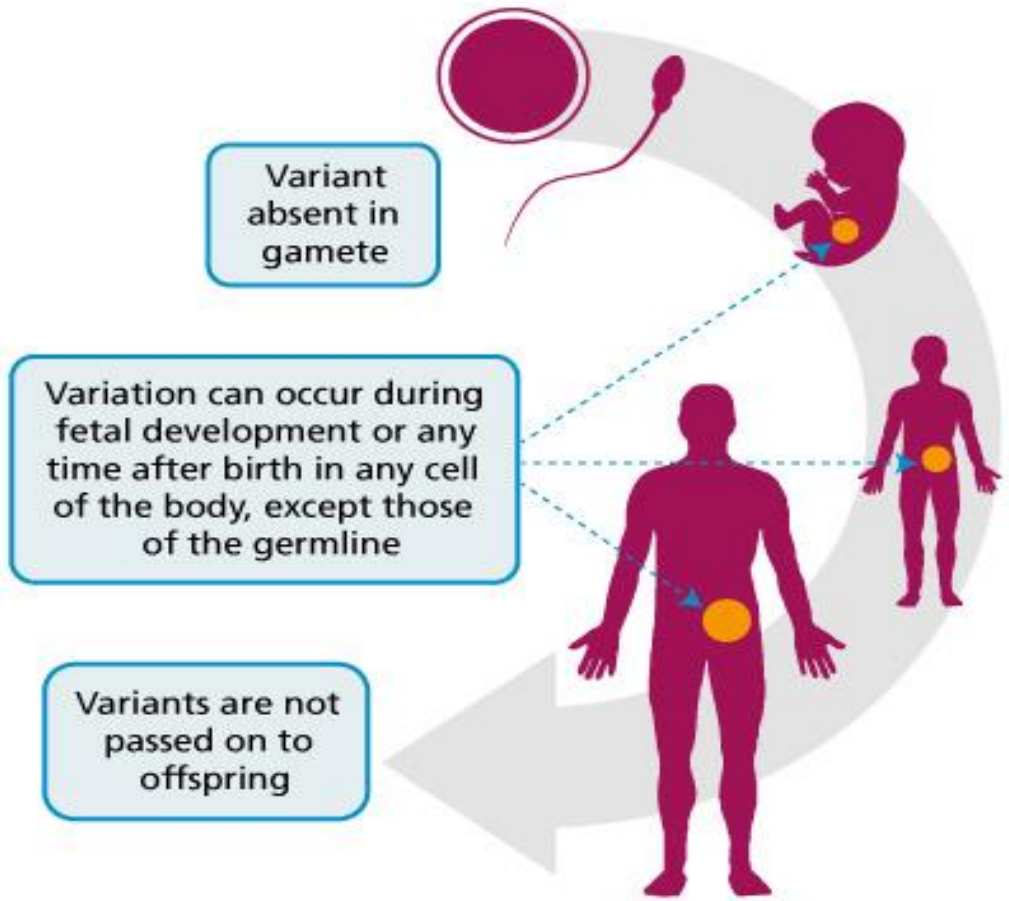


Genetic Variations

GERMLINE vs. SOMATIC VARIANTS



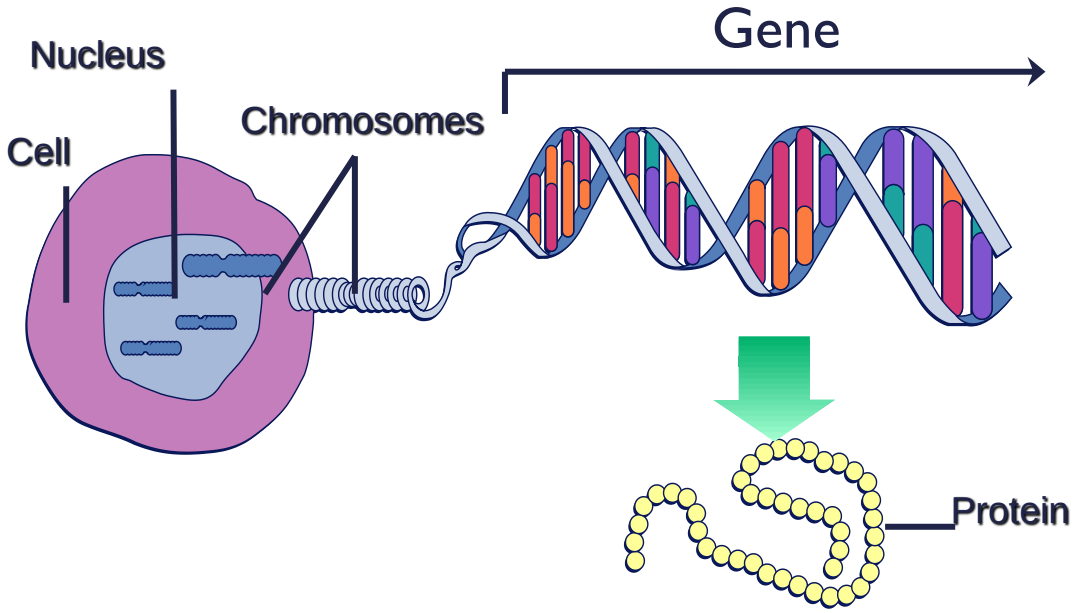
Germline variants



Somatic variants

https://www.genomicseducation.hee.nhs.uk/cancer-genomics/

WHAT MAKES US UNQIUE?



99.9% similar in genetic makeup
OR
0.1% Different




TYPES OF VARIATION

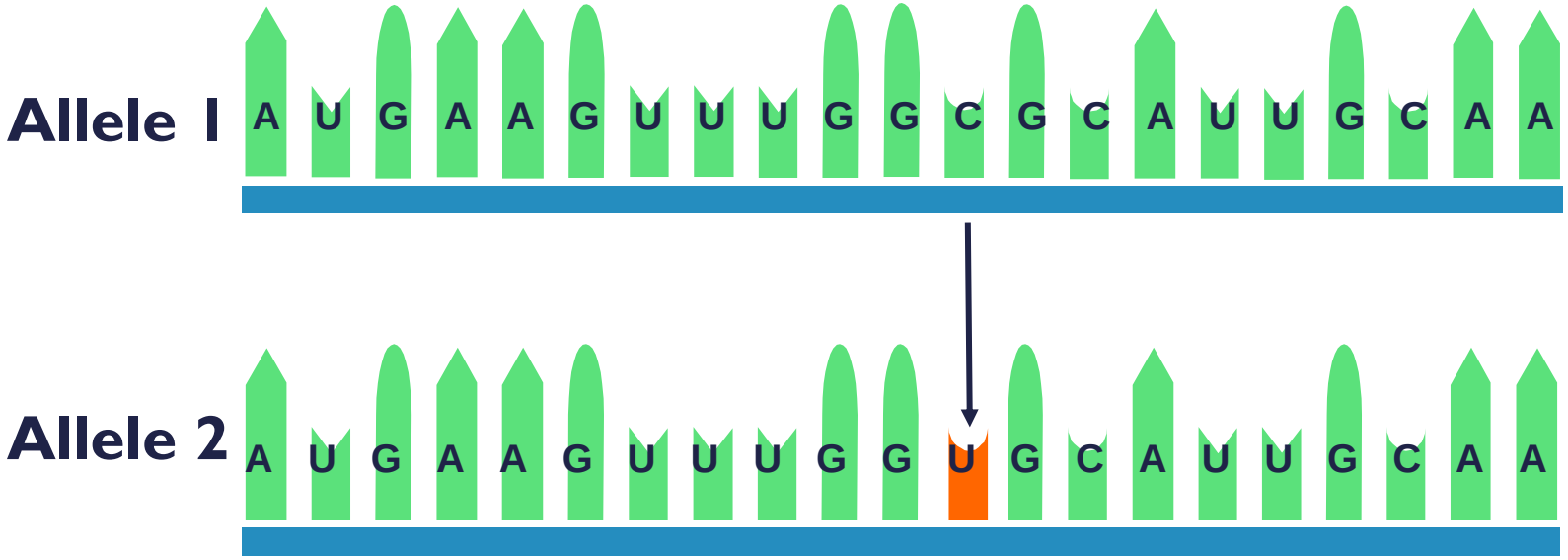
- **RFLP: Restriction Fragment Length Polymorphism**
- **VNTR: Variable Number of Tandem Repeats**
  - or minisatellite
  - ~10-100 bp core unit
- **SSR : Simple Sequence Repeat**
  - or STR (simple tandem repeat)
  - or microsatellite
  - ~1-5 bp core unit
- **SNP: Single Nucleotide Polymorphism**
  - Commonly used to also include rare variants (SNVs)
- **Insertions or deletions**
  - INDEL – small (few nucleotides) insertion or deletion
- **Rearrangement** (inversion, duplication, complex rearrangement)
  - CNV: **C**opy **N**umber **V**ariation

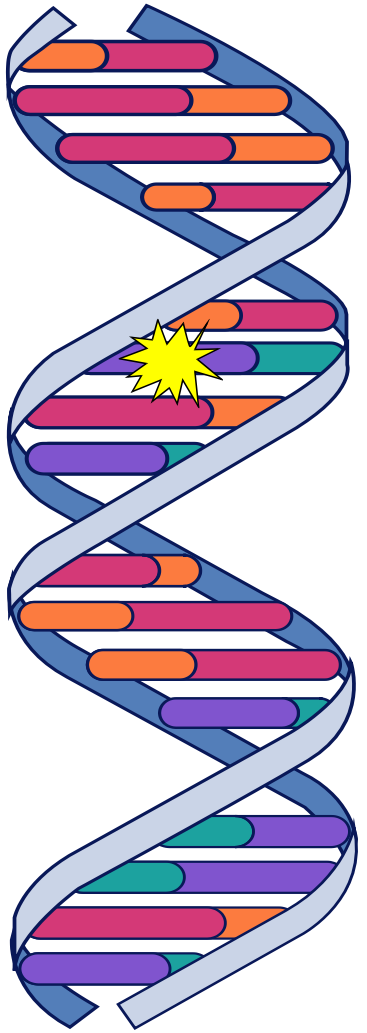
SINGLE NUCLEOTIDE POLYMORPHISM AND MUTATION

- **Genetic Polymorphism**
  - Common variation in the population:
    - Phenotype (eye color, height, etc.)
    - Genotype (DNA sequence polymorphism)
  - Frequency of minor allele(s) >= 1%
- **DNA sequence variation**
  - Most common <= 0.99 (Polymorphism)
  - Minor allele >= 1%
  - Rare variant < 0.01%
- **DNA mutation** – any change in DNA sequence
  - Silent vs. amino acid substitution vs. other
  - Neutral vs. disease-causing
  - 1X10-8/bp/generation (~70 new mutations/individual)
- **Common but incorrect usage**
  - “Mutation” vs. ”Polymorphism”

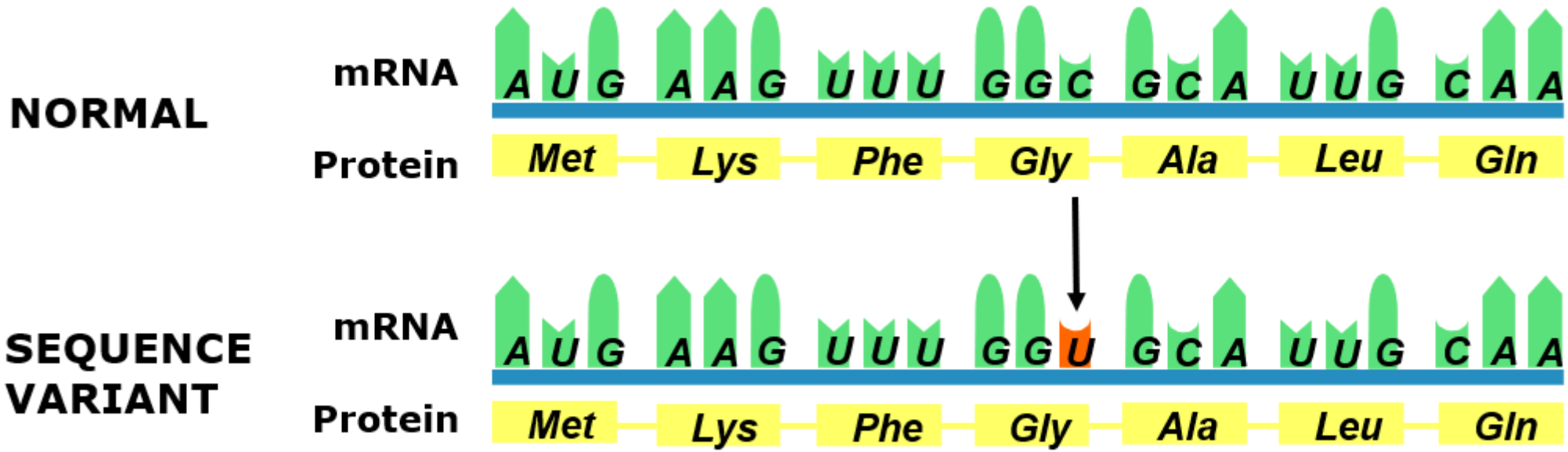
MUTATION

- A mutation is a change in the “normal” base pair sequence
- Can be:
  - A single base pair substitution
  - A deletion or insertions of 1 or more base pairs (indel)
  - A larger deletion/insertion or rearrangement



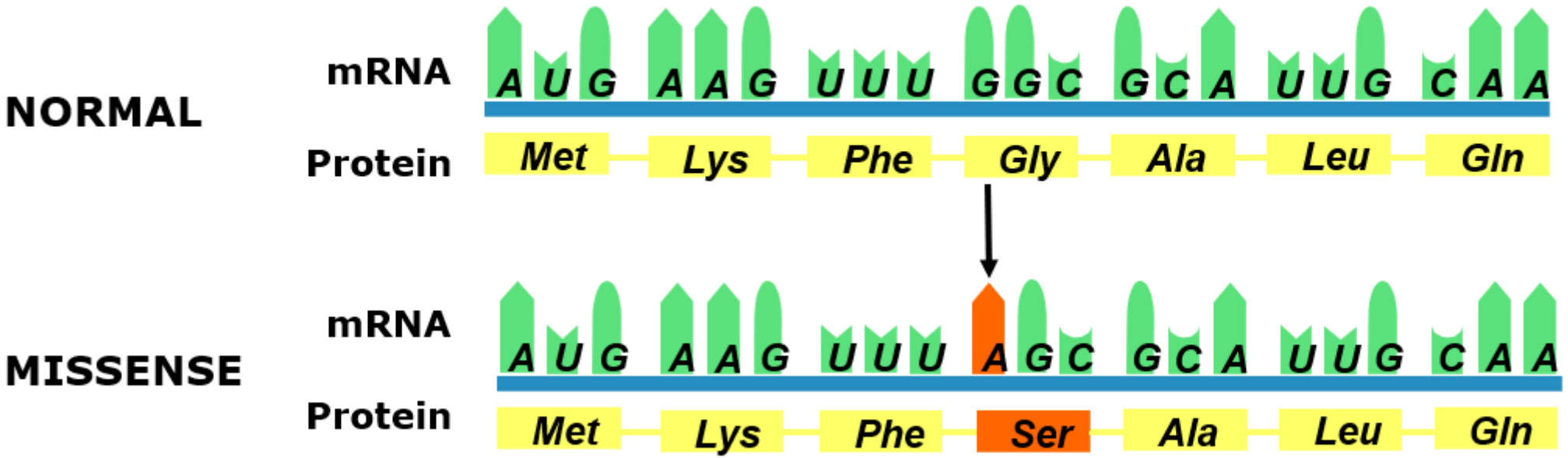


SILENT SEQUENCE CHANGE (Synonymous SNP)



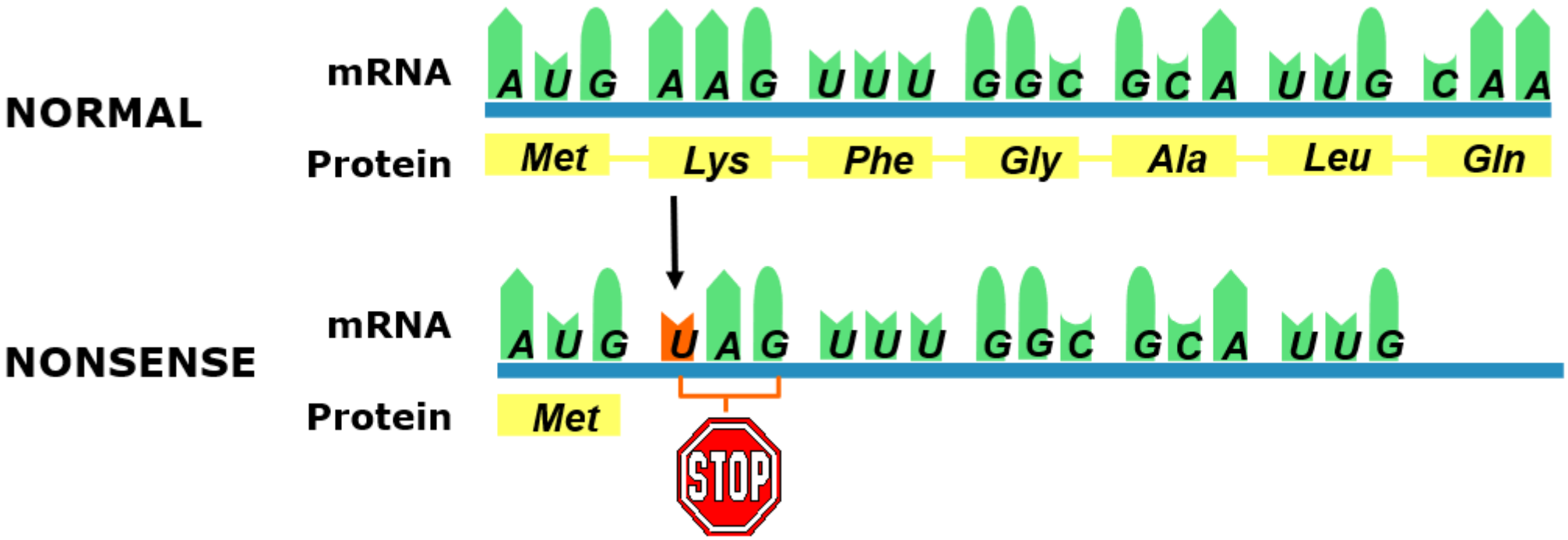
Changes that do not alter the encoded amino acid

MISSENSE MUTATION (Non-Synonymous SNP)



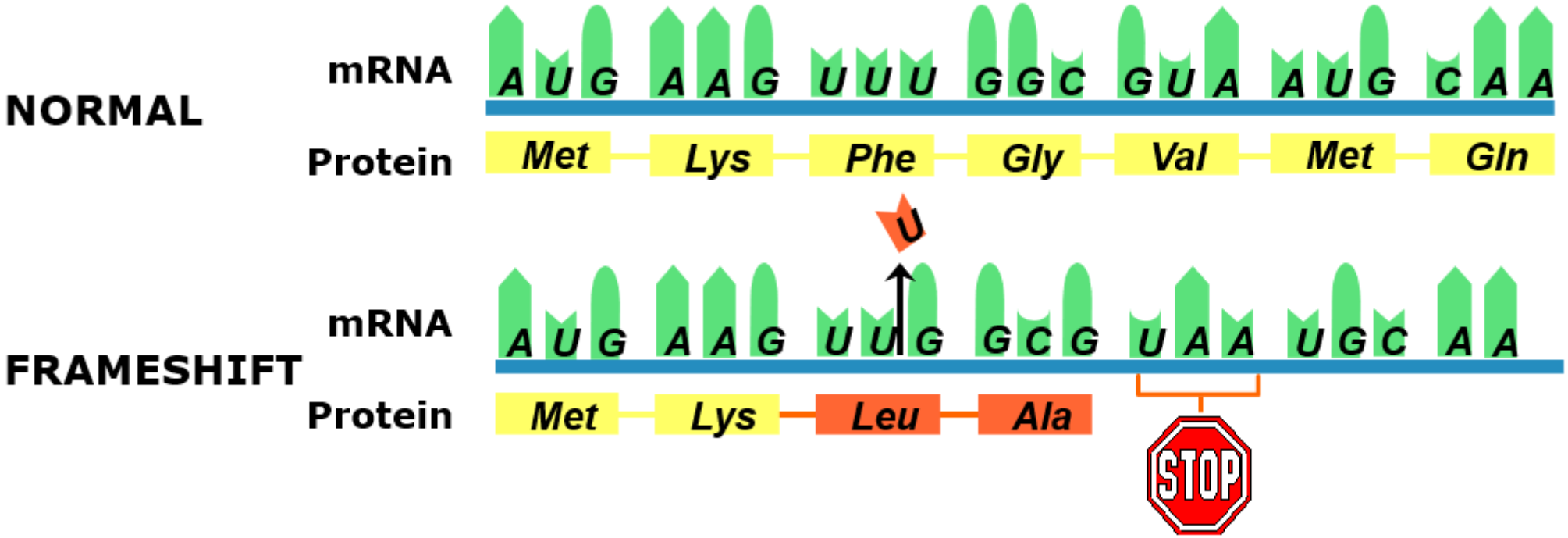
Missense: changes to a codon for another amino acid (can be harmful mutation or neutral variant)

NONSENSE MUTATION (Non-Synonymous SNP)



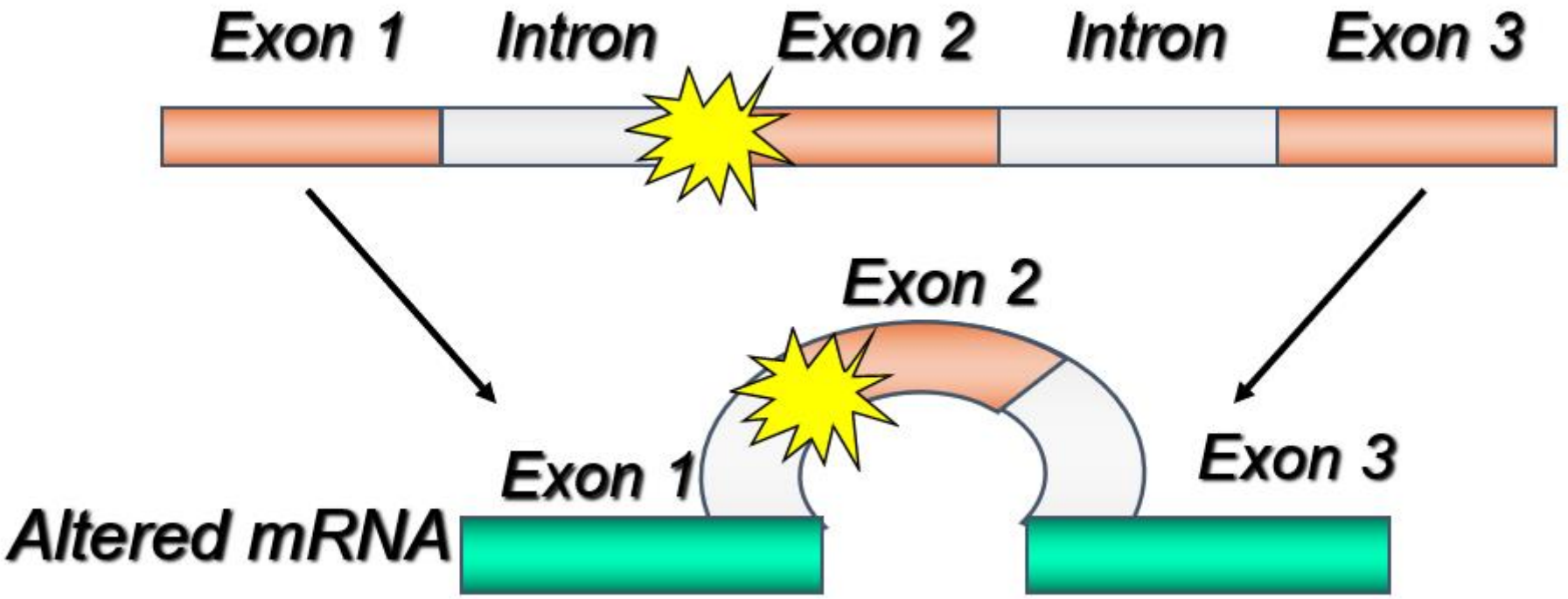
Nonsense: change from an amino acid codon to a stop codon, producing a shortened protein

FRAMESHIFT MUTATION (Non-Synonymous SNP)



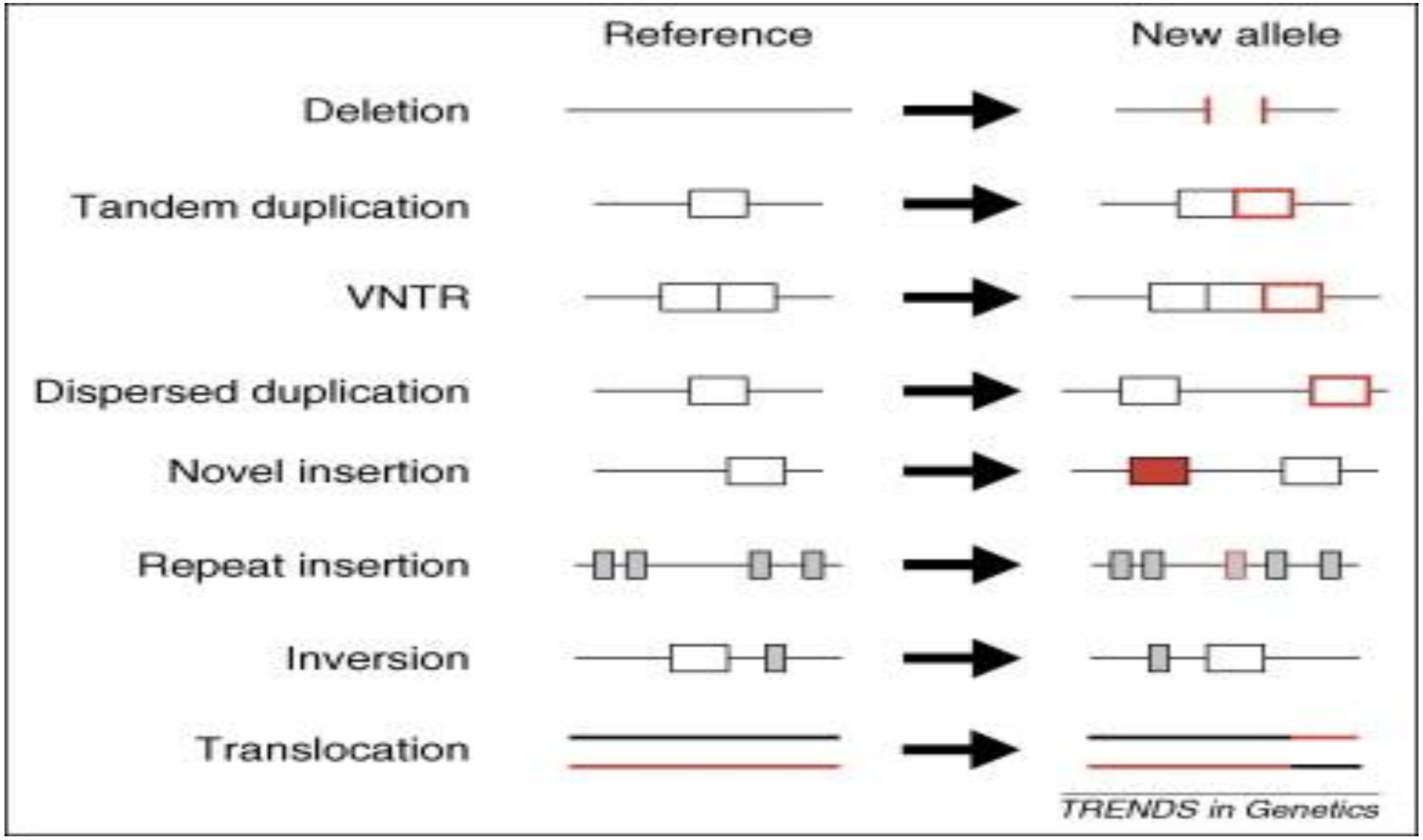
Frameshift: insertion or deletion of base pairs, producing a stop codon downstream and (usually) shortened protein

SPLICE SITE MUTATION

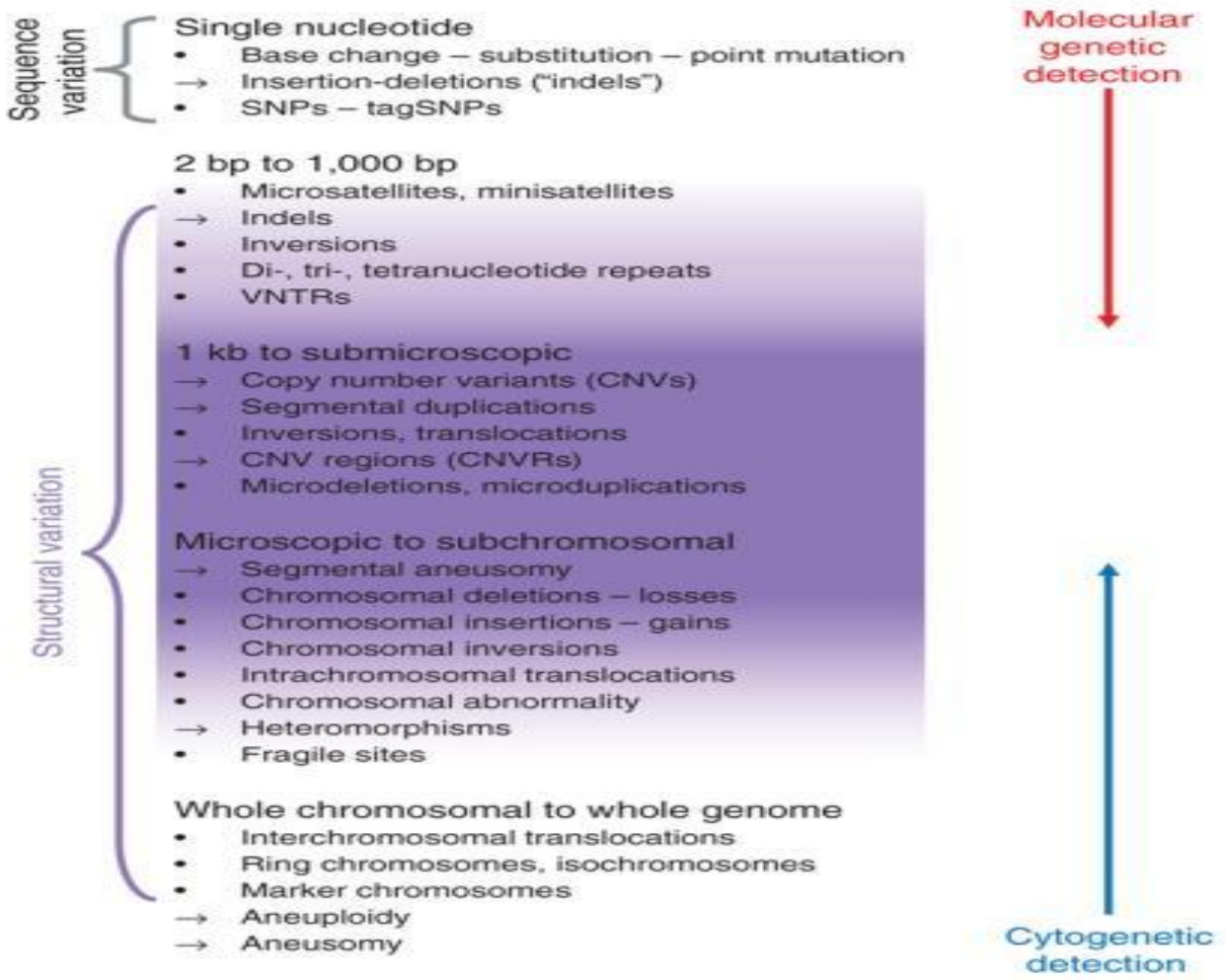


Splice-site mutation: a change that results in altered RNA sequence

OTHER VARIANT TYPES



SIZE SPECTRUM OF HUMAN SEQUENCE VARIATION

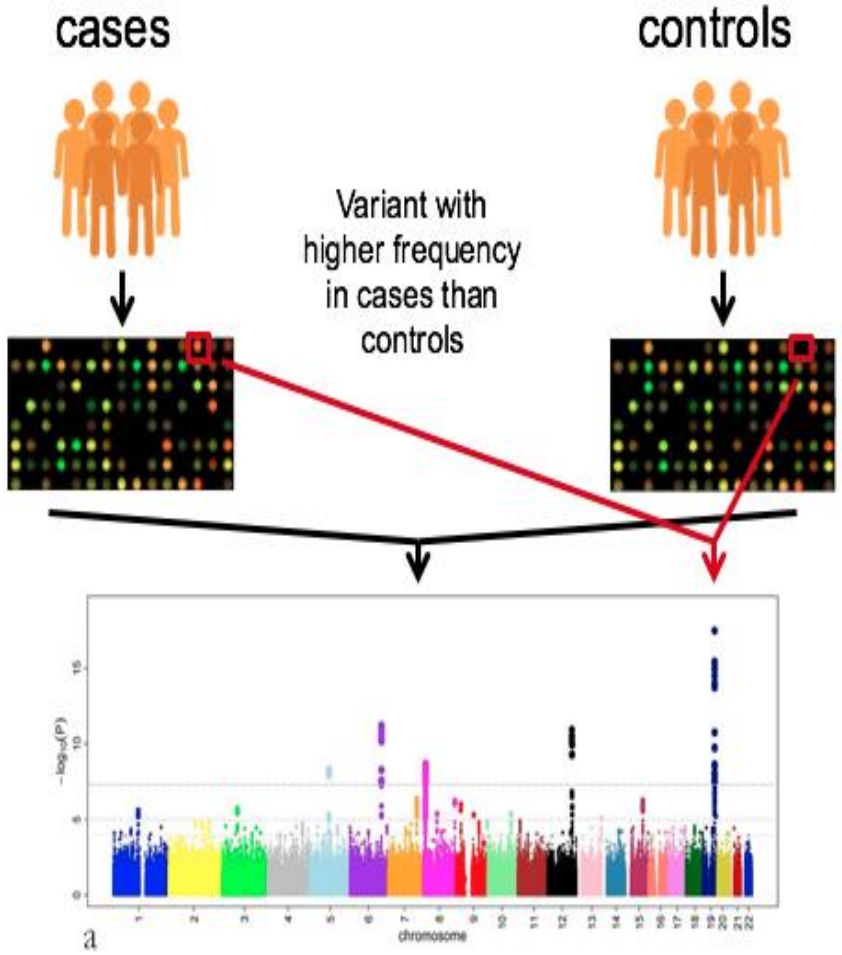


https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2698291/

Genetic Variation Studies

Genome Wide Association Studies (GWAS)

- Genotyping individuals at common variants across the genome using genome wide SNP arrays.
- Variants associated with trait, or within the same haplotype as a variant associated with a trait, will be found at a higher frequency in cases than controls.
- Statistical analysis is carried out to indicate how likely a variant is to be associated with a trait. The p-value of the association indicates how likely the variant is to be associated with the trait.

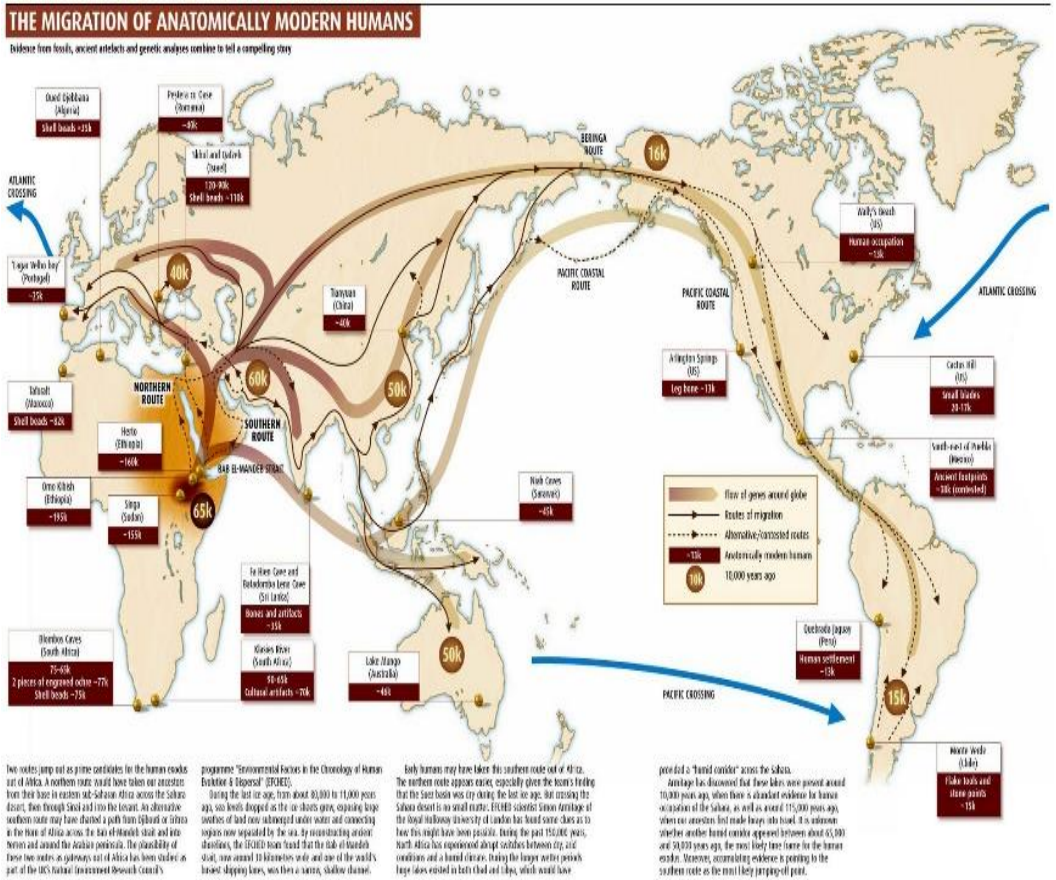


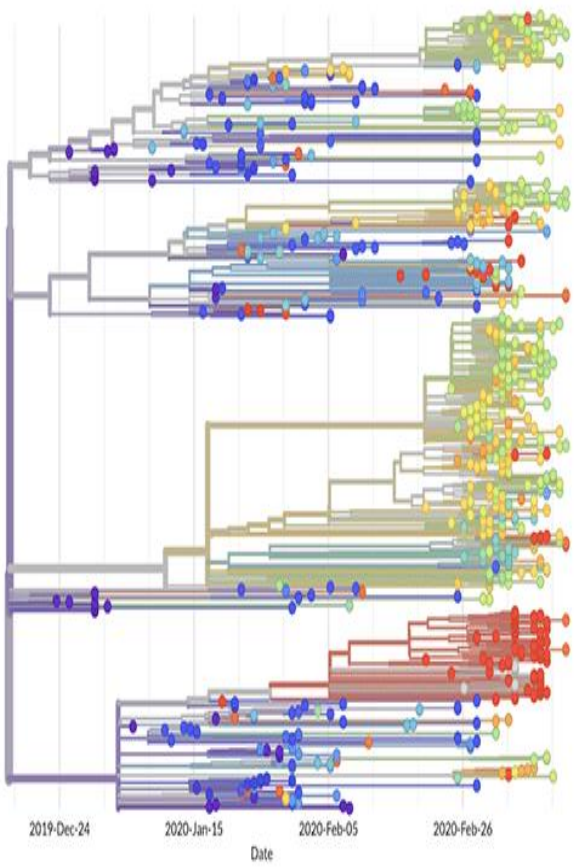
Functional Genetic Variation Studies

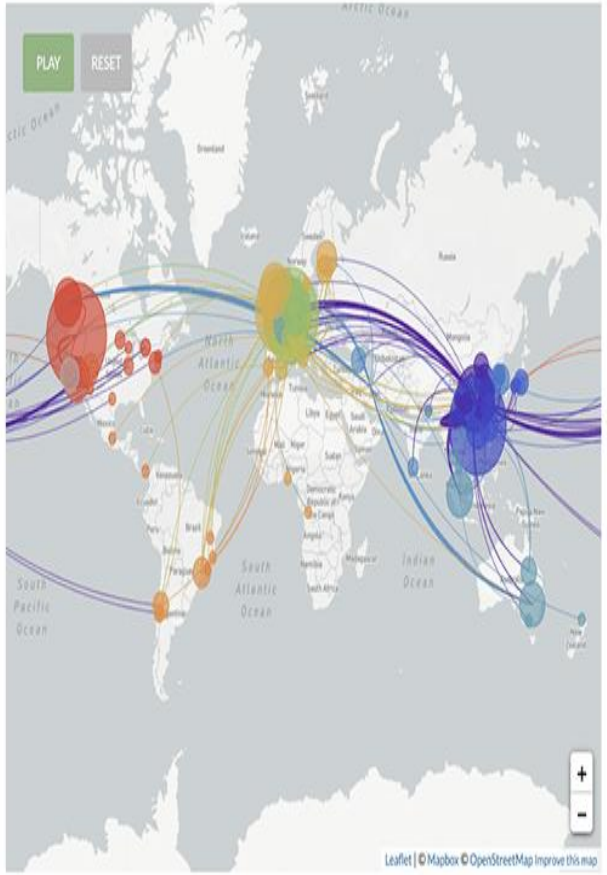
- **Aim:** understand the molecular mechanisms and pathways that link genotype to phenotype.
- Simple variants that alter the translated protein sequence, such as, missense, splice site variant, stop gained, stop lost variants, can cause functional consequences by:
  - Altering ligand and/or co-factor binding sites
  - Alter the natural protein structure by:
    - Removing or adding additional cysteine reduces that can alter disulfide bond patterns
    - Alter normal formation of secondary structure elements or their interaction (sickle cell anaemia is an example of this)
    - Disrupt the normal interactions between proteins' tertiary protein complexes or other cellular components
  - Remove or add post-translational modification sites.
- Personalize medicine, precision medicine, ACMG guidelines

Population Genetics

- Study of variation within populations of individuals.
- Data from genome-scale population genetics studies has been used to:







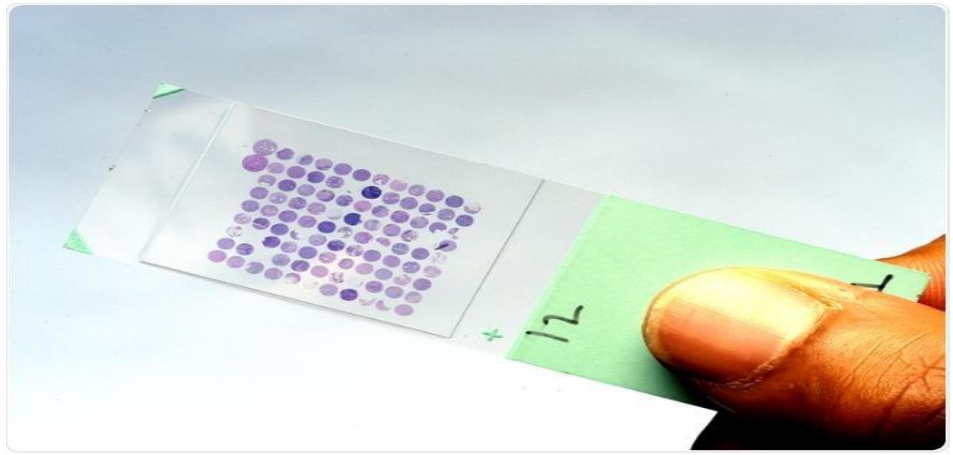
Variant Identification and Analysis

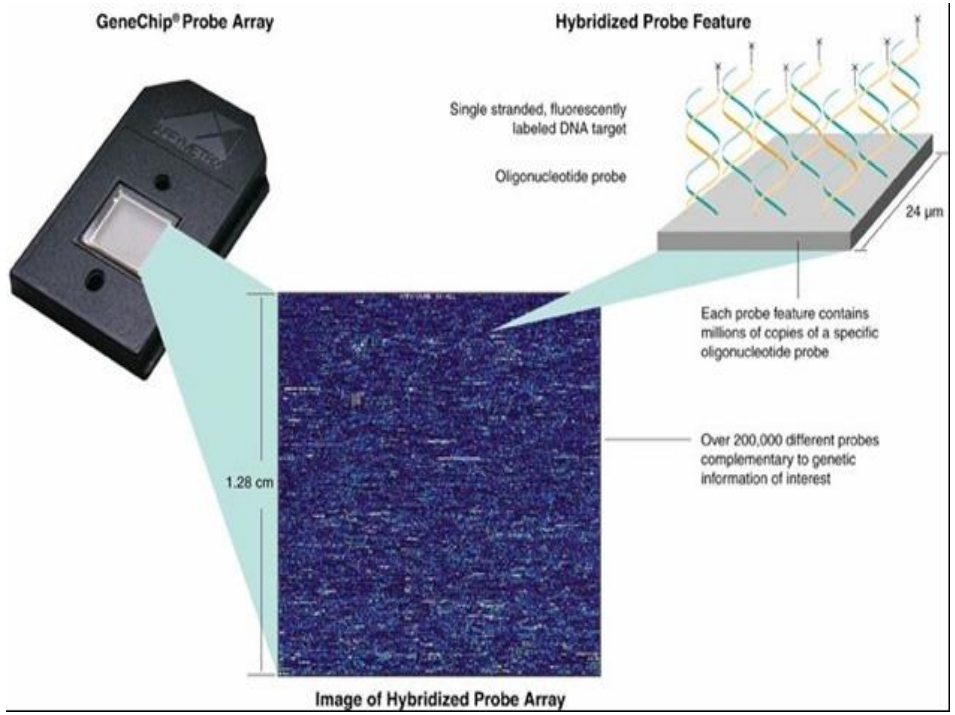
TECHNOLOGIES

- **SNP Array**
- **Next Generation Sequencing**
  - Gene Panel Sequencing
  - Whole Exome Sequencing (WES)
  - Whole Genome Sequencing (WGS)

MICROARRAY

- **Microscopic slide usually made of glass, silicon chip or nylon membrane.**
- **Surface provided with thousands of minute pores in defined positions.**
- **Able researchers analyze thousands of genes in a single reaction.**
- **Various types:**
  - DNA microarrays, MMChips, Protein microarrays, Peptide microarrays, Tissue microarrays, Cellular microarrays, Chemical compound microarrays, Antibody microarrays, Carbohydrate microarrays, Phenotype microarrays, Reverse phase protein microarrays, Interferometric reflectance imaging sensor or IRIS

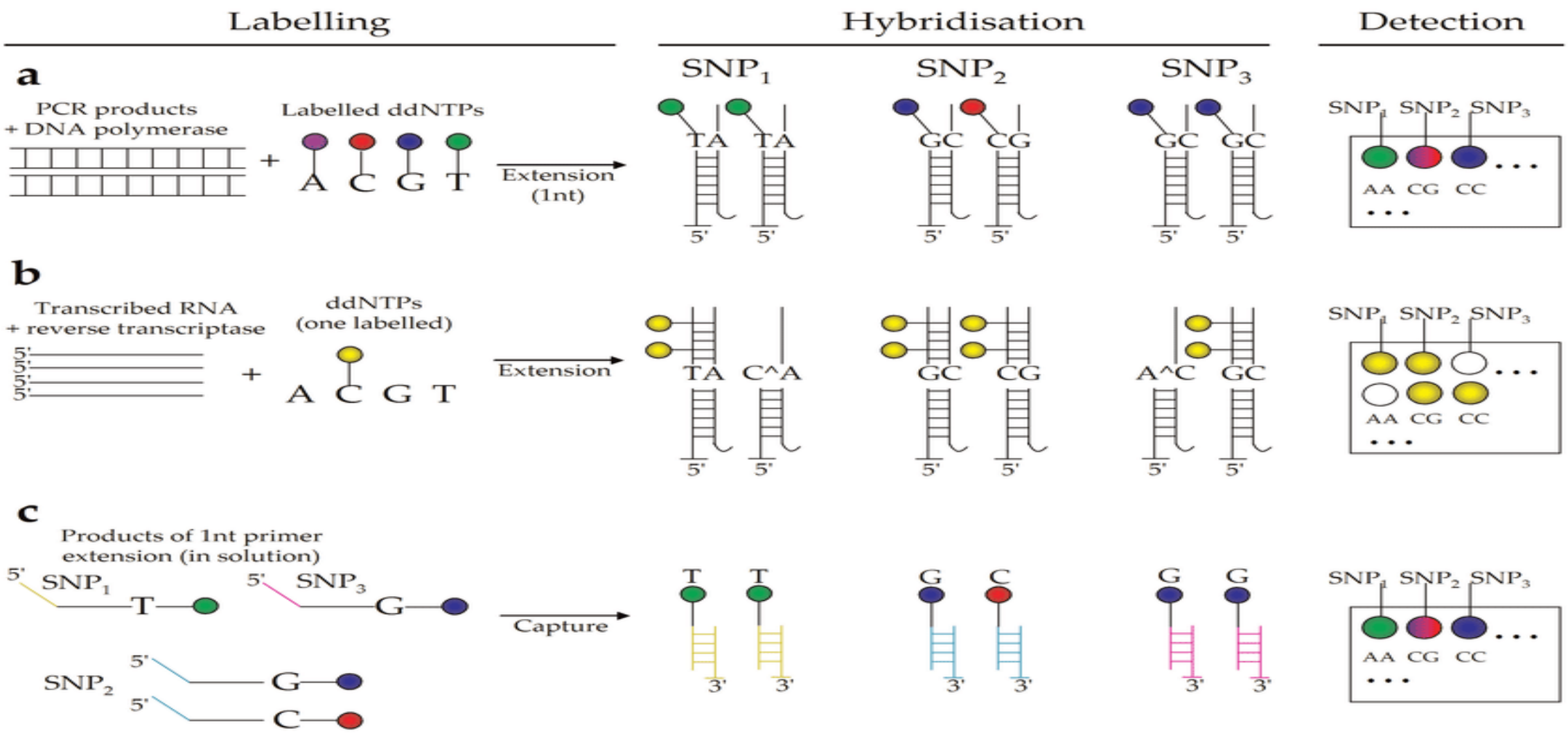




DNA MICROARRAY

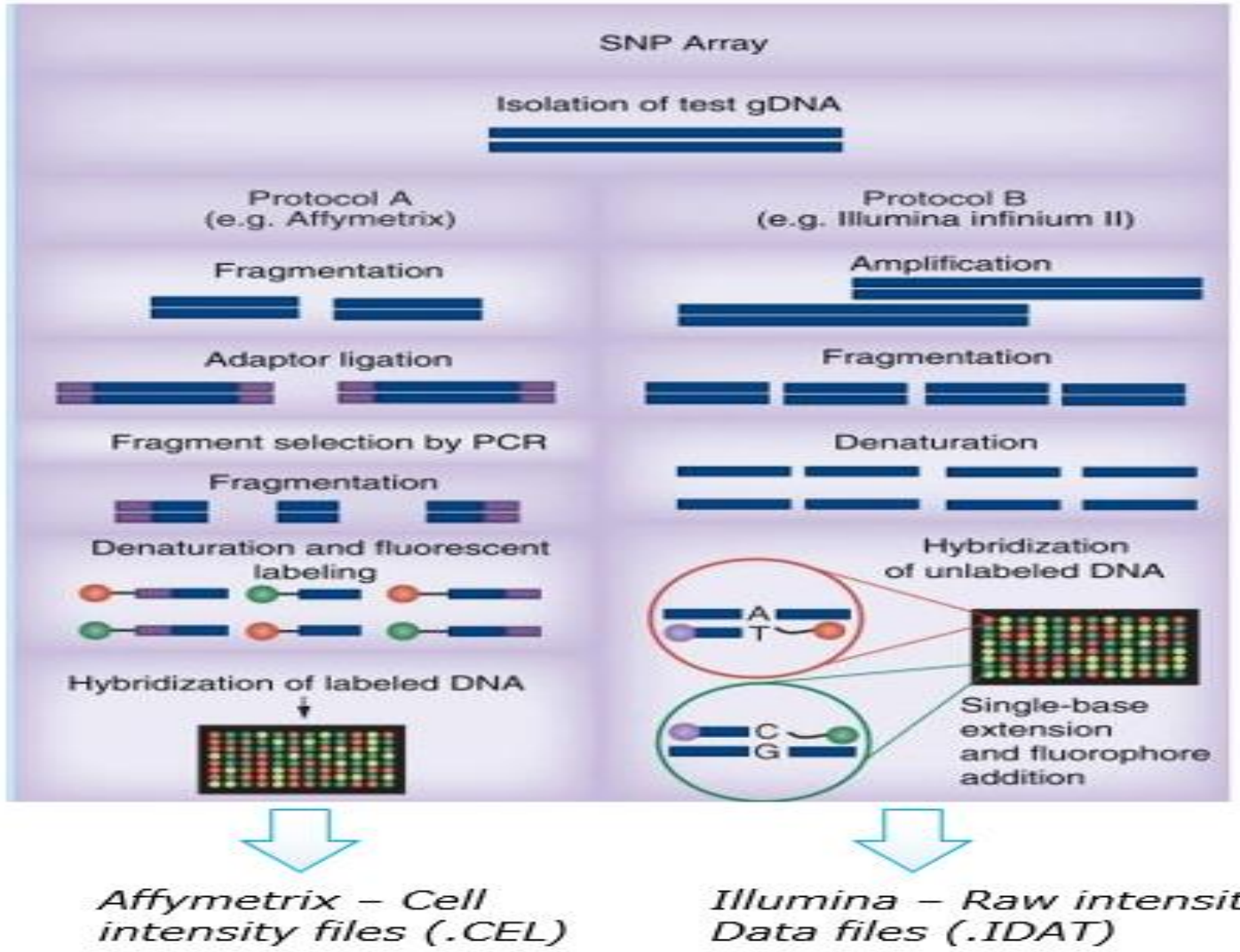
- **Types:** cDNA microarrays, oligo DNA microarrays, BAC microarrays and SNP microarrays.
- SNP microarray ***works on the principle of DNA hybridization in which a single base change can be detected through fluorescence chemistry.***
- Application:
  - Haplotype and gene mapping
  - Cancer research
  - Personalized genetic research
  - Genetic medicine research
  - Genome-wide association studies
- SNP array completes in three common steps:
  - Immobilization oligonucleotides/probes (make a chip)
  - Fragmentation and labelling nucleic acid
  - Hybridization

MAJOR TECHNIQUES FOR DETECTION OF SNPS USING MICROARRAYS



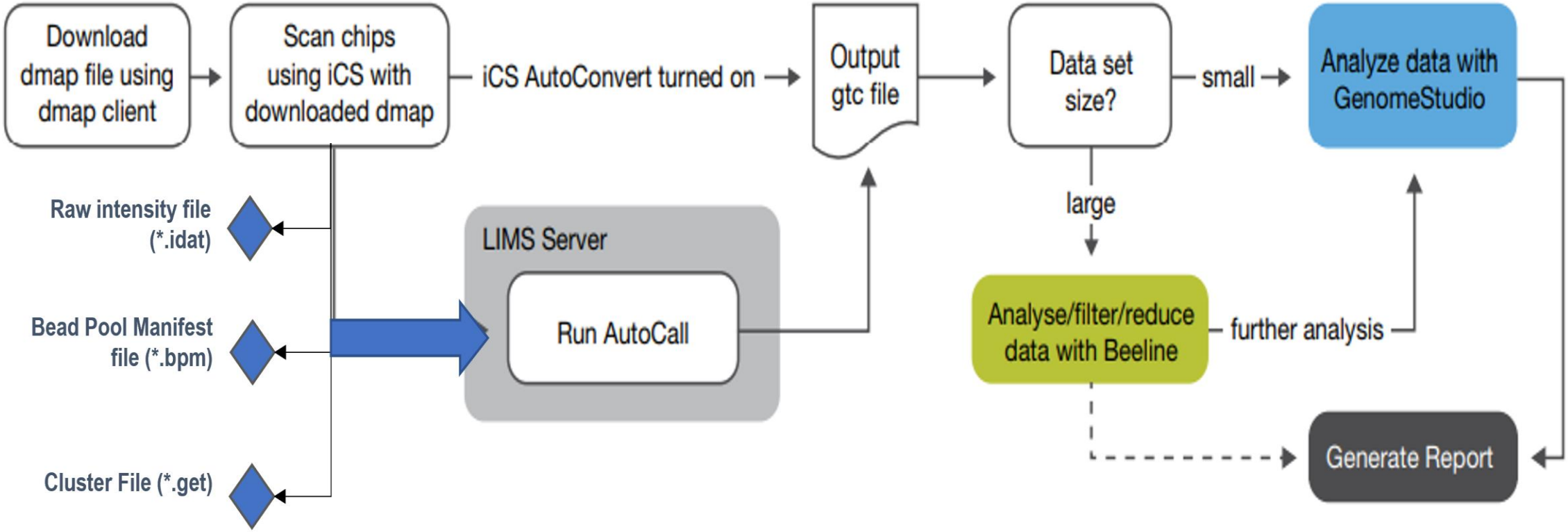
https://molmed.biomedcentral.com/articles/10.2119/2006-00107.Trevino

SNP ARRAY PROCESS FLOW



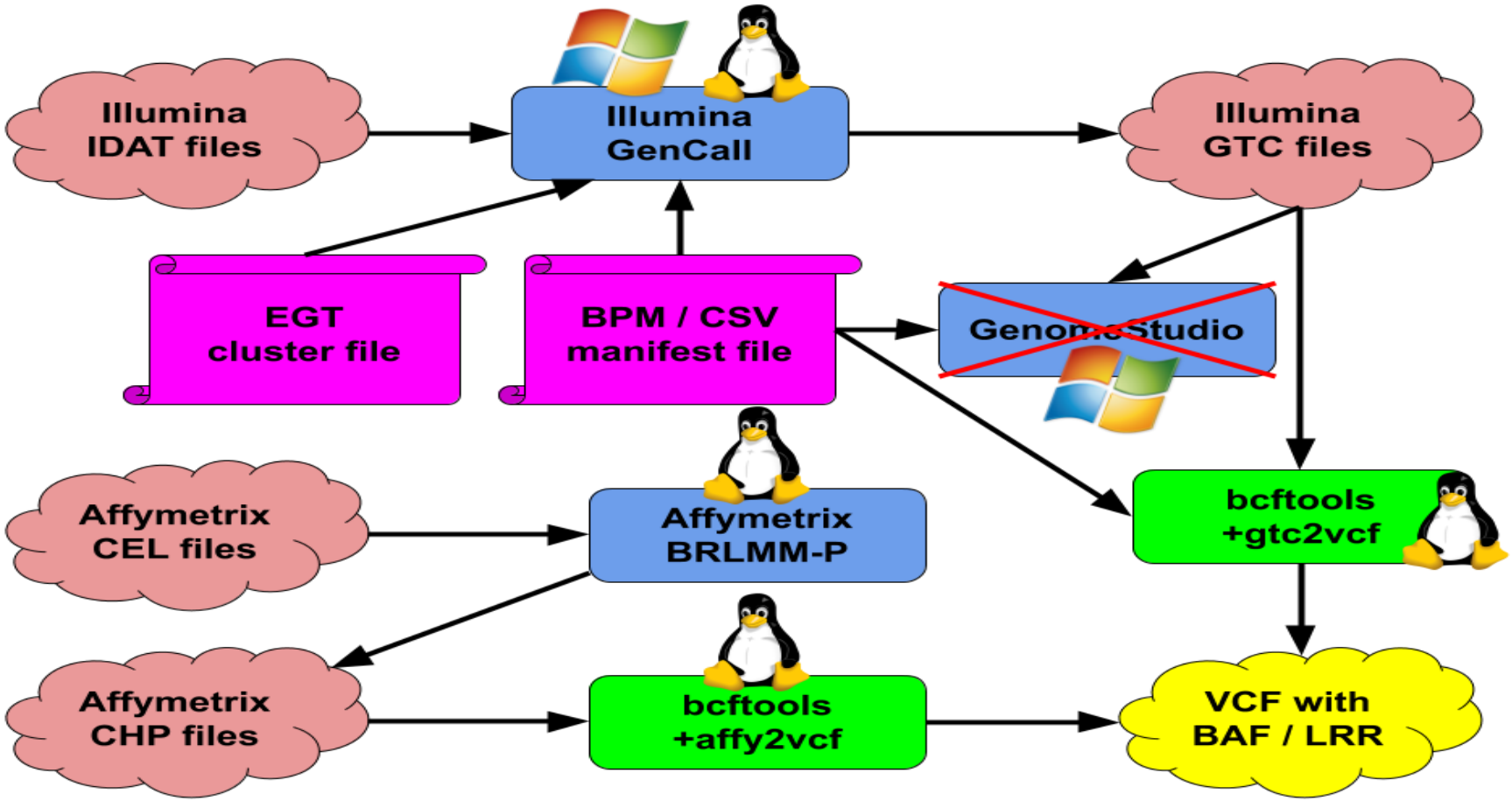
Clinical application of targeted and genome-wide technologies: can we predict treatment responses in chronic lymphocytic leukemia? (nih.gov)

RAW INTENSITY FILE TO VCF (ILLUMINA)



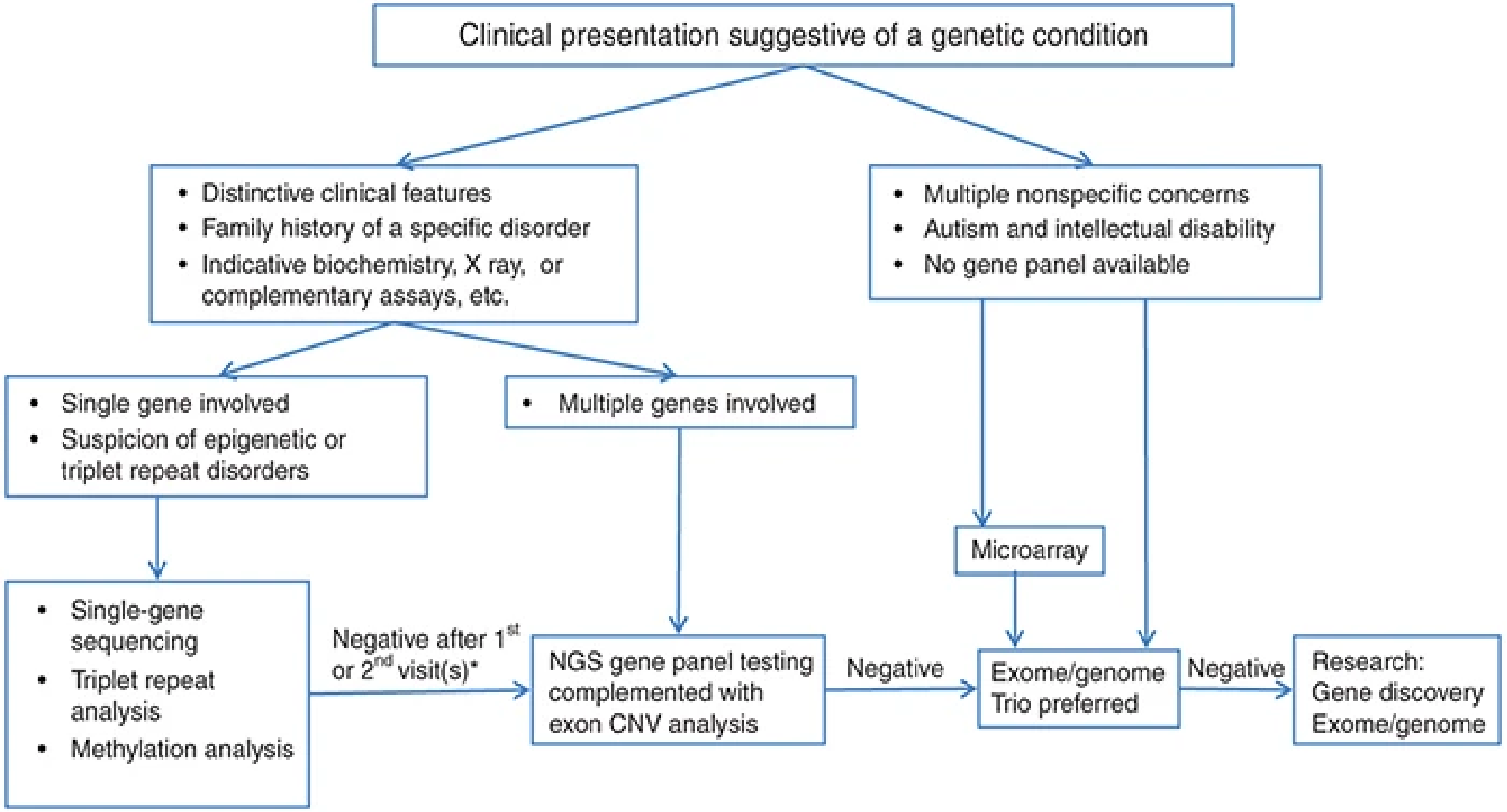
technote_array_analysis_workflows.pdf (illumina.com)

RAW INTENSITY FILE TO VCF (ILLUMINA & AFFYMETRIX)



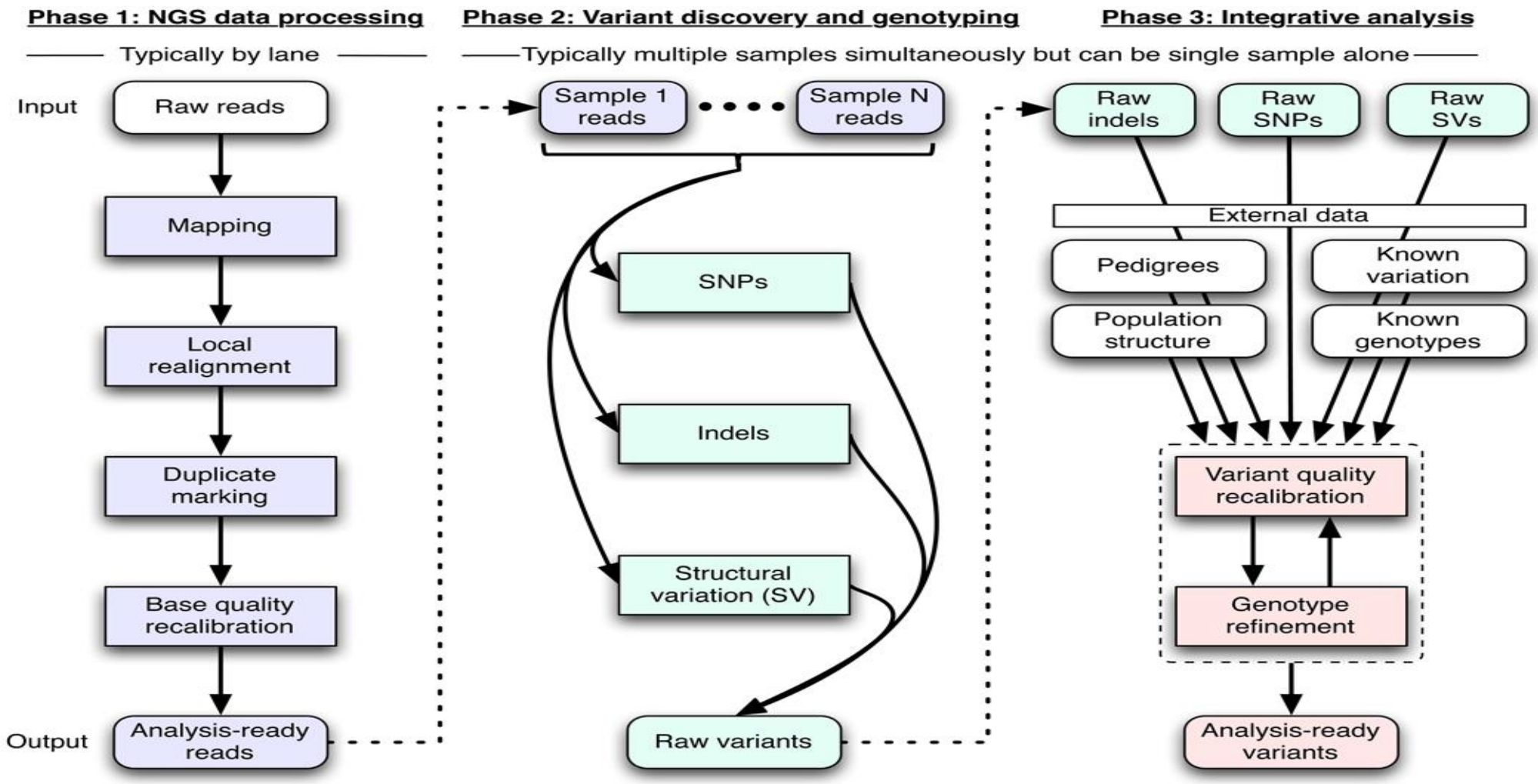
freeseek/gtc2vcf: Tools to convert Illumina IDAT/BPM/EGT/GTC and Affymetrix CEL/CHP files to VCF (github.com)

NEXT GENERATION SEQUENCING (NGS)



https://www.nature.com/articles/gim2014122

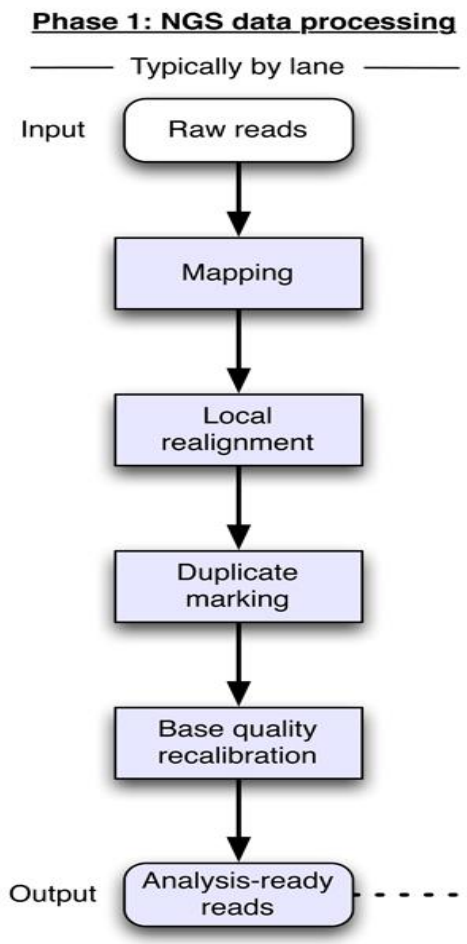
FRAMEWORK FOR VARIANT DISCOVERY (NGS)



 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3083463/

MAPPING (NGS)

- **Place reads with an initial alignment on the reference genome using mapping algorithms.**
- **Refine initial alignments**
  - local realignment around indels
  - molecular duplicates are eliminated
- **Generate the technology-independent SAM/BAM alignment map format.**

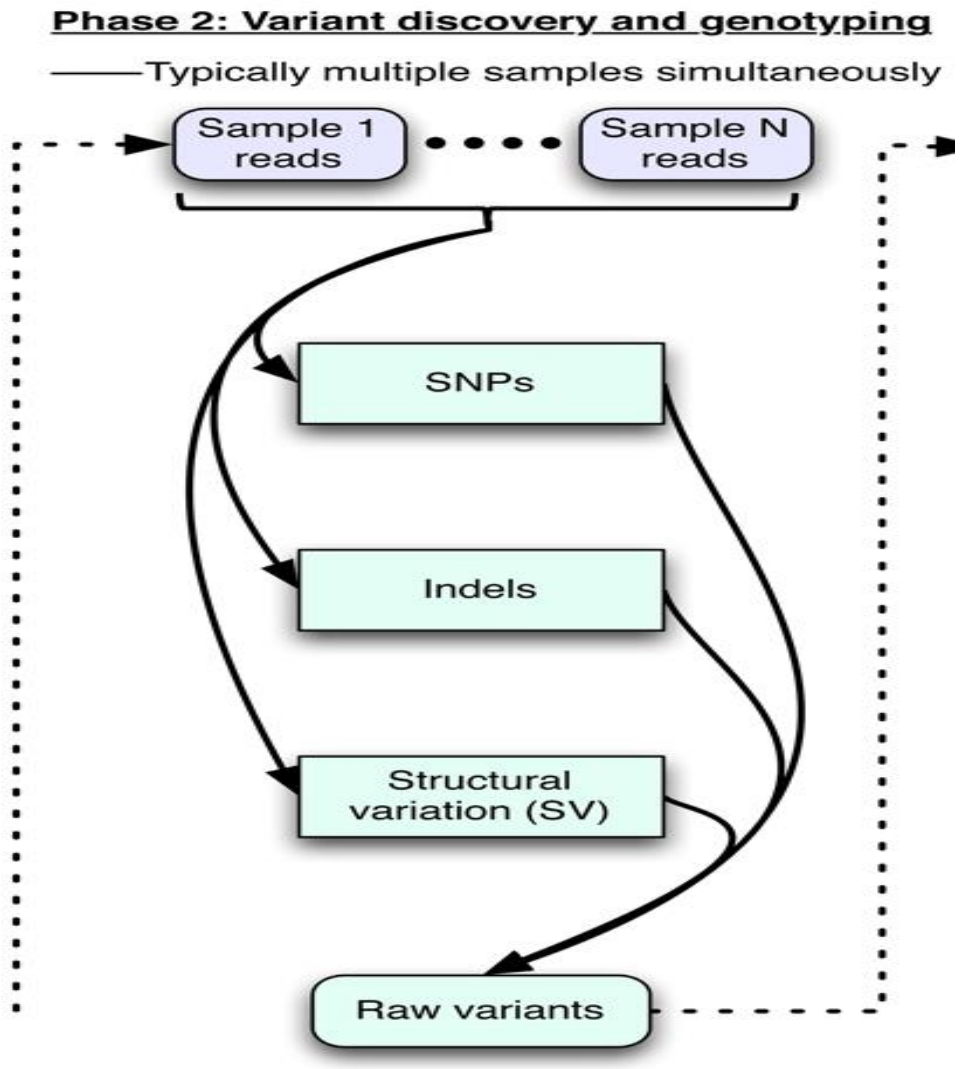


Accurate mapping crucial for variation discovery

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3083463/

DISCOVERY OF RAW VARIANTS

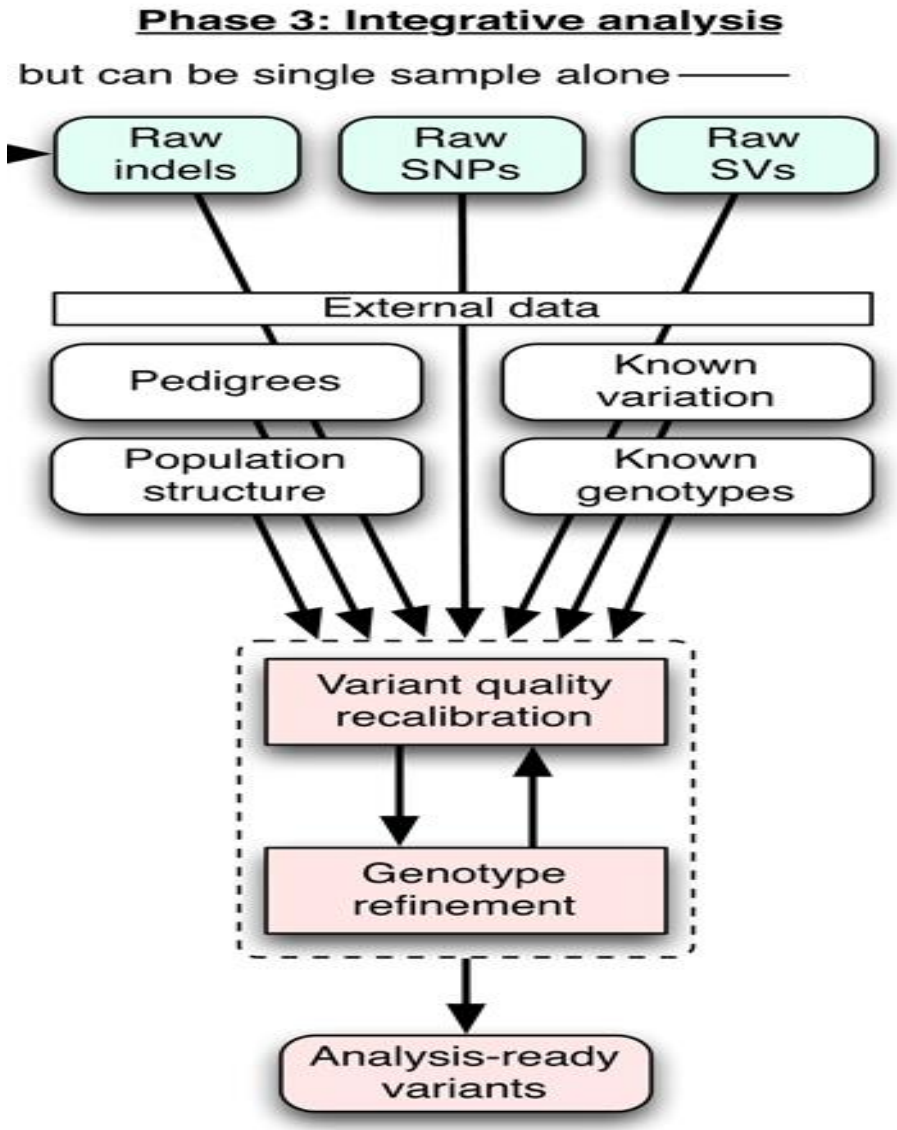
- **Analysis-ready SAM/BAM files are analyzed to discover all sites with statistical evidence for an alternate allele present among the samples.**
- **SNPs, SNVs, short indels, and SVs.**



https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3083463/

DISCOVERY OF ANALYSIS READY VARIANTS

- **Technical covariates, known sites of variation, genotypes for individuals, linkage disequilibrium, and family and population structure are integrated with the raw variant calls from Phase 2 to separate true polymorphic sites from machine artifacts.**
- **At these sites high-quality genotypes are determined for all samples.**



https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3083463/

VARIANT CALL FORMAT (VCF)

Header

```
##fileformat=VCFv4.2
##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">
#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
```

(b) SNP

Alignment

```
1234
ACGT
ATGT
```

VCF representation

```
POS REF ALT
2   C   T
```

(c) Insertion

```
12345     POS REF ALT
AC-GT     2   C   CT
ACTGT
```

(d) Deletion

```
1234      POS REF ALT
ACGT      1   ACG A
A--T
```

(e) Replacement

```
1234      POS REF ALT
ACGT      1   ACG AT
A-TT
```


VCFv4.2.pdf (samtools.github.io)

Header Line

- The header line names the 8 fixed, mandatory columns;

1. #CHROM
2. POS
3. ID
4. REF
5. ALT
6. QUAL
7. FILTER
8. INFO

- If genotype data is present in the file, these are followed by a FORMAT column header, then an arbitrary number of sample IDs.
- The header line is tab-delimited.

ARRAY VS. NGS

ARRAY		NGS	
Pros	***Cons***	***Pros***	***Cons***
Relatively Inexpensive	High background, low sensitivity	Low background, very sensitive	Expensive
Easy Sample Prep.	Limited dynamic range	Large dynamic range	Complex sample preparation
Mature Informatics & Stats.	Not quantitative	Quantitative	Limited bioinformatics
	Competitive hybridization		Massive information technology infrastructure required
	Annotation probes		

TASKS (NGS)

- **Article reading and discuss**
  - DePristo, M.A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 43(5):491-8. PMID: 21478889 (2011).
  - Narendra M. et al. A Bioinformatics Pipeline for Whole Exome Sequencing: Overview of the Processing and Steps from Raw Data to Downstream Analysis .BioRxiv (2017).
- **Hands on “Disease causing mutation” (NGS)**

TERMINALOGIES

- **Variation:** any difference between individuals of a particular species.
- **Mutation:** alteration in the nucleotide sequence of a gene.
- **Alleles:** Different versions of the same variant.
- **Reference allele:** to the base that is found in the reference genome.
- **Alternative allele:** any base, other than the reference allele found at that locus (position).
- **Major allele:** most common allele for a given SNP.
- **Minor allele:** less common allele for a given SNP. MAF (Minor Allele Frequency)
- **Genotype:** genetic make-up of an individual.
- **Phenotype:** physical traits and characteristics of an individual and are influenced by their genotype and the environment.

"The important thing is not to stop questioning. Curiosity has its own reason for existing." - Albert Einstein

Thank You!