A Clinician's Guide to the Human Genome

Unraveling Genomic Variants for Patient Care

Created by:
Maryam Vaseghi-Shanjani (MSc)
Emilie Théberge (MSc)
Dr. Phillip Richmond (PhD)
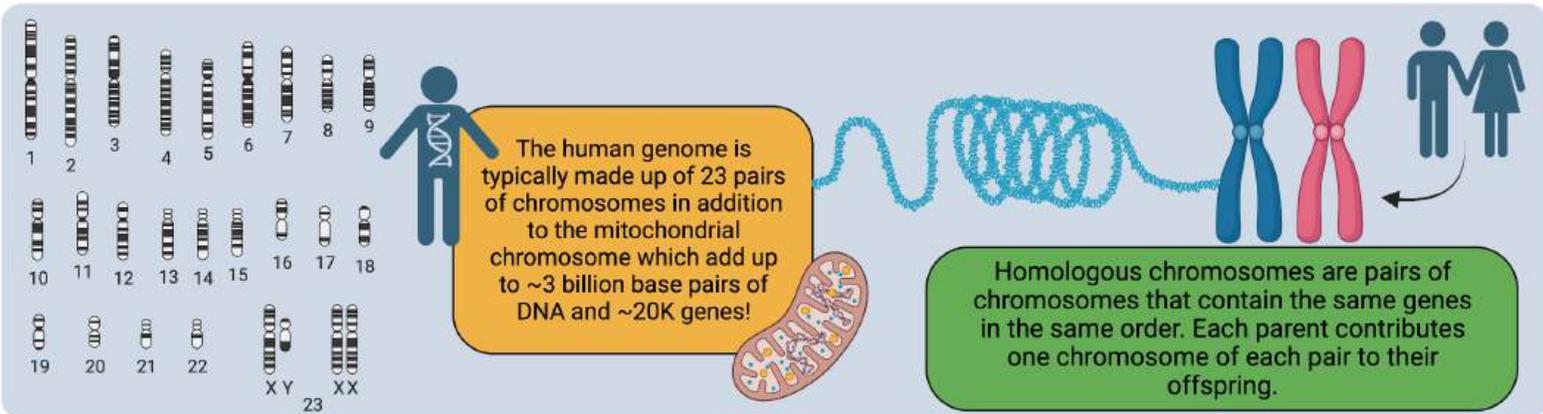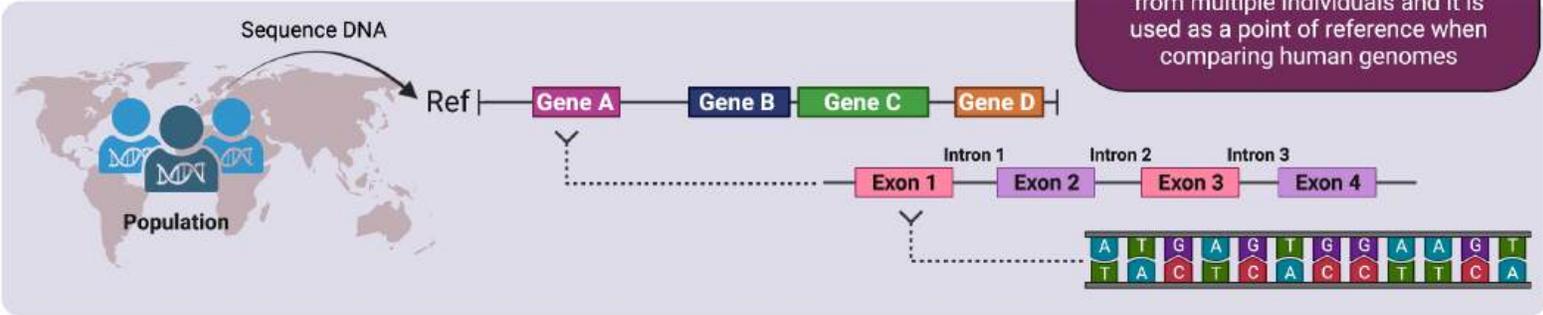
THE UNIVERSITY OF BRITISH COLUMBIA

Genome BritishColumbia

BC Children's Hospital Research Institute

Understanding the Human Genome and Genomic Variants

Sequence DNA

Population

Ref | Gene A | Gene B | Gene C | Gene D |

Intron 1 | Intron 2 | Intron 3

Exon 1 | Exon 2 | Exon 3 | Exon 4

ATGAGTGGAAGT
TACTCACCTTCA

The reference genome is assembled from multiple individuals and it is used as a point of reference when comparing human genomes

1 2 3 4 5 6 7 8 9
10 11 12 13 14 15 16 17 18
19 20 21 22
X Y XX
23

The human genome is typically made up of 23 pairs of chromosomes in addition to the mitochondrial chromosome which add up to ~3 billion base pairs of DNA and ~20K genes!

Homologous chromosomes are pairs of chromosomes that contain the same genes in the same order. Each parent contributes one chromosome of each pair to their offspring.

Genomic variants are differences in the DNA sequence of an individual compared to a reference genome and these variants come in different shapes and sizes, here are a few examples of them:

Chromosomal Variants

Aneuploidy

Aneuploidy is a condition in which the affected individual has an abnormal number of chromosomes

Chromosomal Translocation

Chromosomal translocation occurs when a part of one chromosome breaks off and attaches to another chromosome

Large Structural Variants

Inversion
Ref

Large Insertion
Ref

Tandem Duplication
Ref

Deletion
Ref

Small Genomic Variants

Single Nucleotide Variant (SNVs)

Ref A C G C T G A

A C G T T G A

Insertion

A C G C T G A

A C G T G C T G A

Deletion

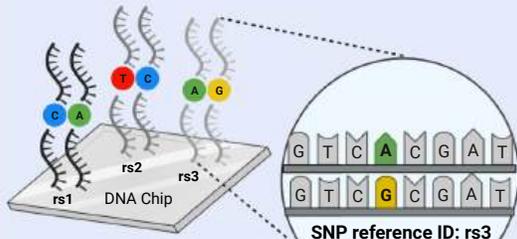A C G C T G A C T

A C G _ _ _ A C T

**Repeat Expansion Variant**

A C G C T G A

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

These variations can have a range of effects, from having no noticeable impact to causing disease.

Created with Biorender.com

Technologies for the Detection of Genomic Variants

DNA Microarray

① DNA Preparation & Tagging

Patient Sample

When patient DNA is washed over the chip, it binds to a sequence match and emits measurable fluorescence

Shear

Denature

Dye Labelling

② Chip Hybridization

rs1 rs2 rs3

rs1 rs2 rs3 DNA Chip

SNP reference ID: rs3

GTC**A**CGAT
GTC**G**CGAT

The chip's surface has thousands of synthetic single-stranded DNA probes representing reference genome sequences or known common single nucleotide polymorphisms (SNPs) in the population
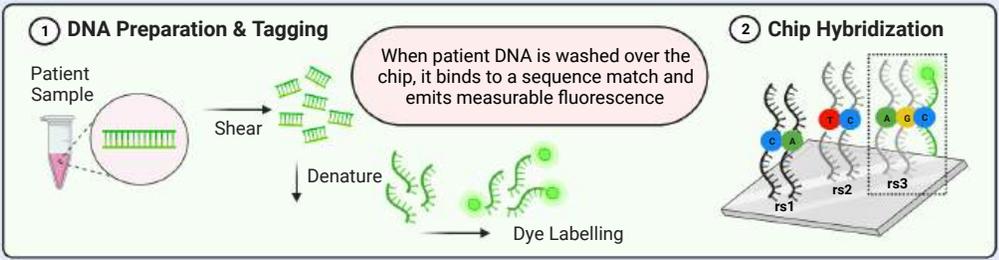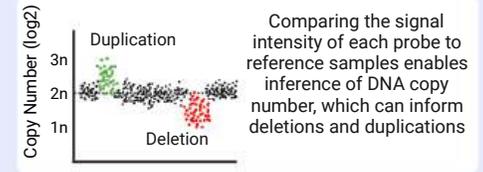
Detect single nucleotide polymorphisms (SNPs)

rs	Chr	Pos	Allele 1	Allele 2	Patient
rs3	12	100	A	G	G / G

Based on the fluorescence intensity for the allele 1 and 2 probes for each SNP, the patient genotype is determined
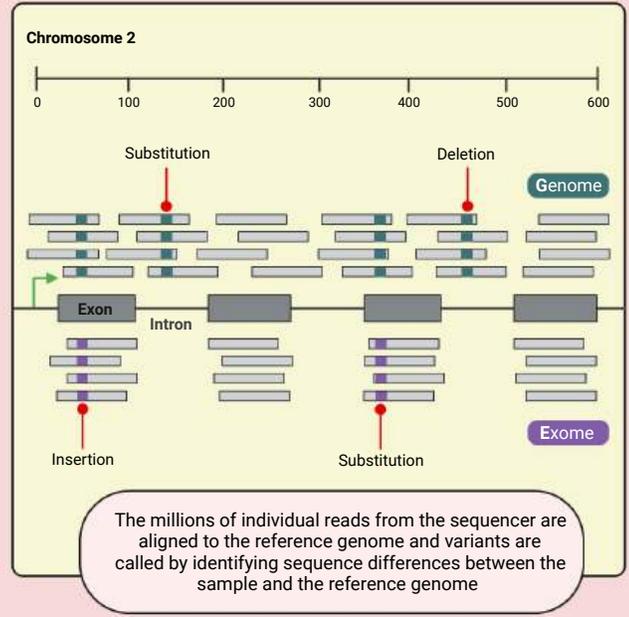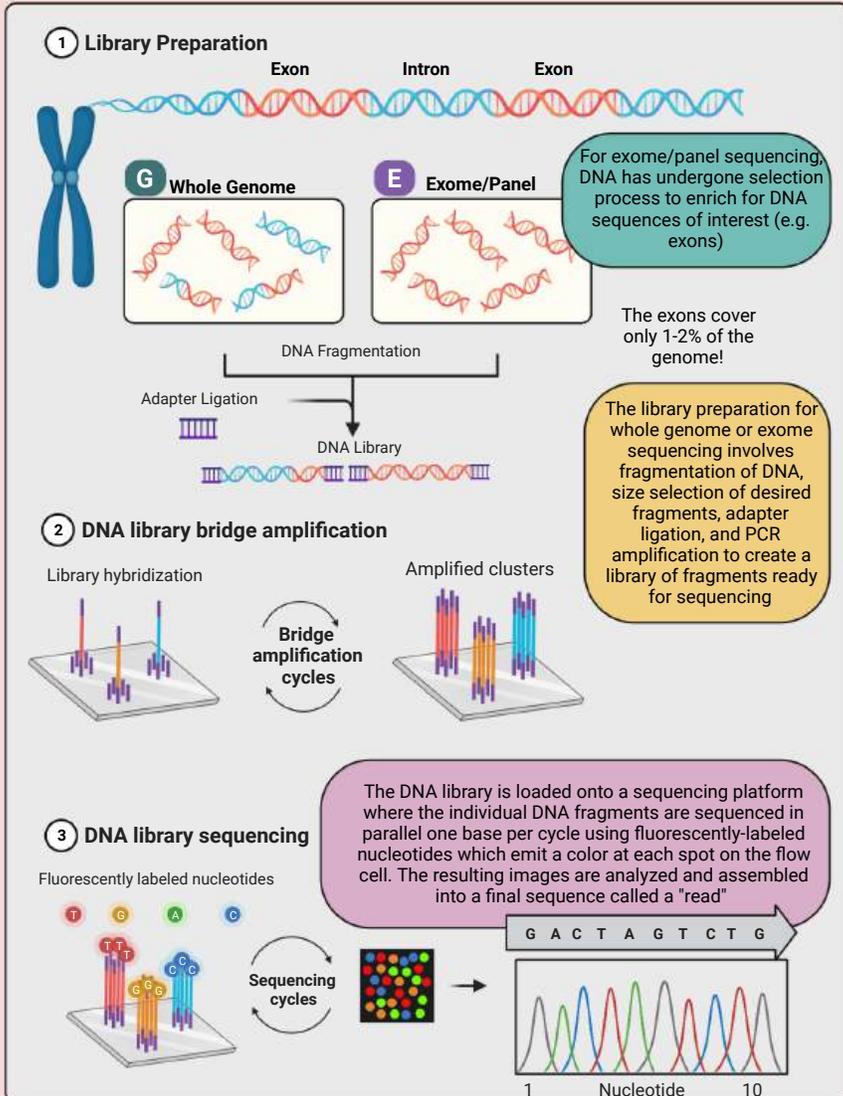
Detect copy number variants (CNVs)

Copy Number (log2)

3n — Duplication
2n
1n — Deletion

Comparing the signal intensity of each probe to reference samples enables inference of DNA copy number, which can inform deletions and duplications

👍 Can detect large CNVs (>50 kb) and known pre-targetted SNPs represented by the probes

👎 Unable to detect new or rare variants that are not represented by a probe on the DNA chip

Next Generation Sequencing

① Library Preparation

Exon Intron Exon

G Whole Genome

E Exome/Panel

For exome/panel sequencing, DNA has undergone selection process to enrich for DNA sequences of interest (e.g. exons)

DNA Fragmentation

Adapter Ligation

DNA Library

The exons cover only 1-2% of the genome!

The library preparation for whole genome or exome sequencing involves fragmentation of DNA, size selection of desired fragments, adapter ligation, and PCR amplification to create a library of fragments ready for sequencing

② DNA library bridge amplification

Library hybridization

Amplified clusters

Bridge amplification cycles

③ DNA library sequencing

Fluorescently labeled nucleotides

Sequencing cycles

G A C T A G T C T G

1 Nucleotide 10

The DNA library is loaded onto a sequencing platform where the individual DNA fragments are sequenced in parallel one base per cycle using fluorescently-labeled nucleotides which emit a color at each spot on the flow cell. The resulting images are analyzed and assembled into a final sequence called a "read"

Chromosome 2

0 100 200 300 400 500 600

Substitution   Deletion

Genome

Exon   Intron

Insertion   Substitution

Exome

The millions of individual reads from the sequencer are aligned to the reference genome and variants are called by identifying sequence differences between the sample and the reference genome

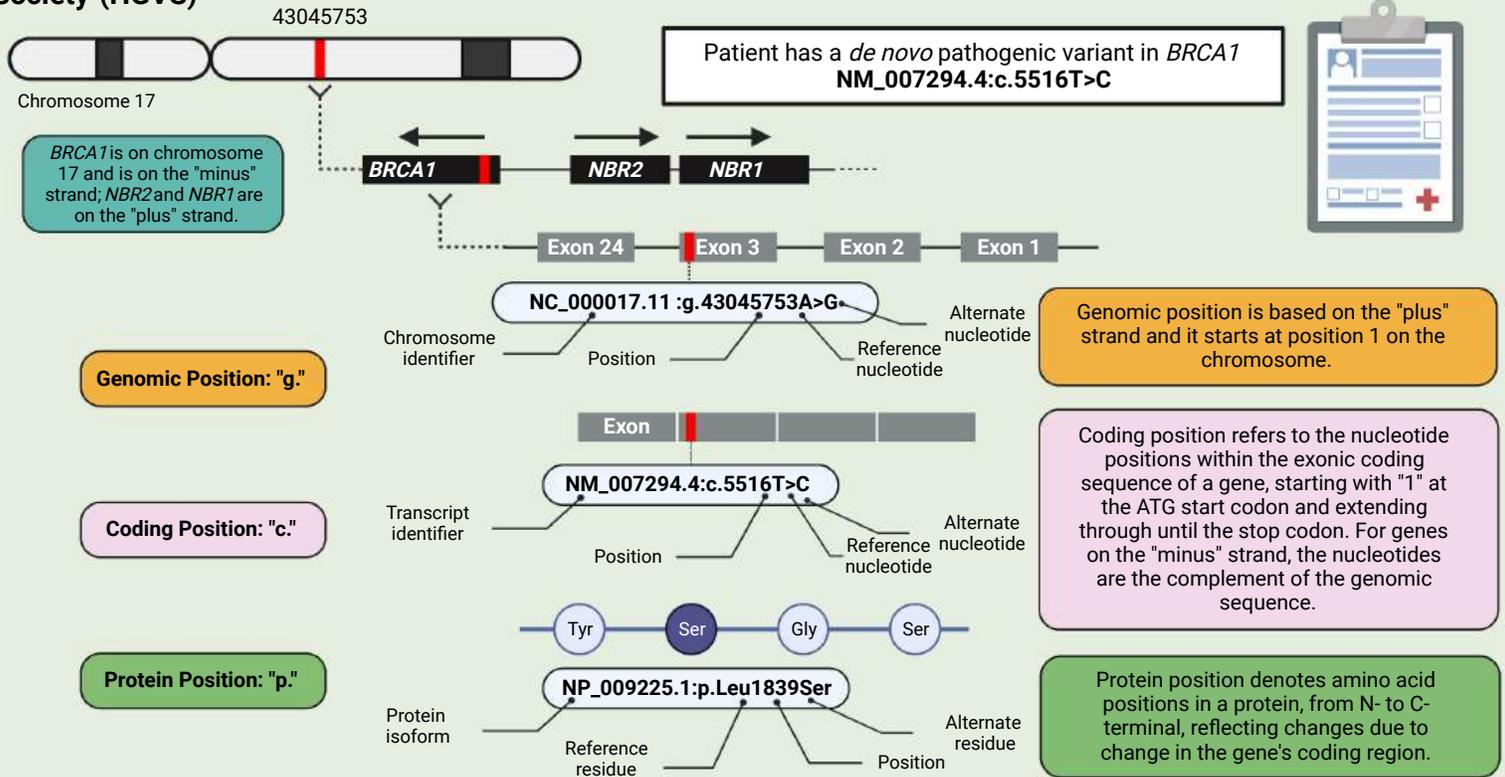|	Detect	Chr.	Pos.	Ref.	Alt.
Variant 1	G E	2	55	A	AC
Variant 2	G	2	130	C	T
Variant 3	G E	2	380	G	A
Variant 4	G	2	470	T	–

Exome sequencing is generally less expensive than whole genome sequencing, but it may miss variants in non-coding regions that could be relevant to disease

👍 Exome sequencing can detect novel rare SNVs and Indels over exonic regions. Genome sequencing can detect novel and rare SNVs, Indels, large CNVs and simple structural variants in both exonic/intronic and intergenic regions

👎 Both technologies have difficulty accurately characterizing complex structural variants such as inversions, translocations, large insertions, and repeat expansions in part due to the limited length of the sequencing reads

Interpreting Genomic Variants

Reporting variants with standardized nomenclature: Human Genome Variation Society (HGVS)
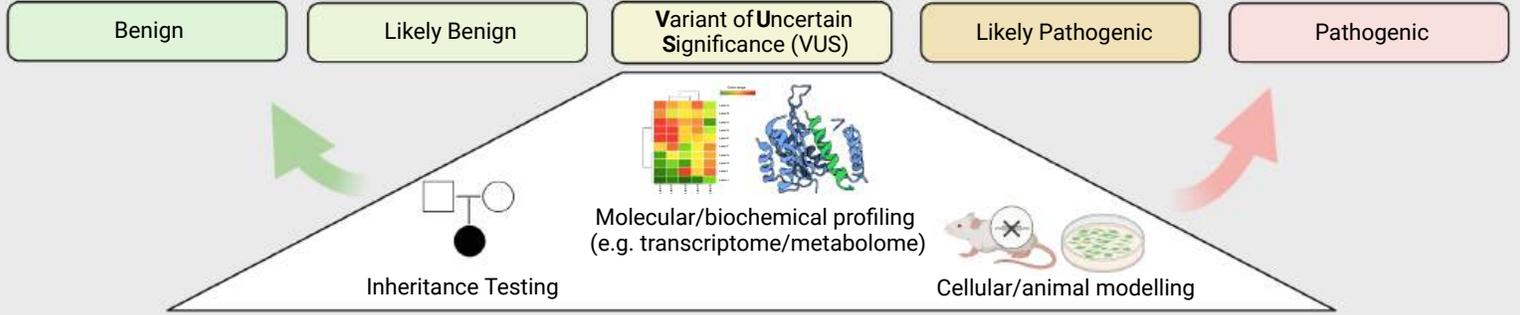
43045753

Chromosome 17

BRCA1 is on chromosome 17 and is on the "minus" strand; *NBR2* and *NBR1* are on the "plus" strand.

BRCA1 NBR2 NBR1

Patient has a *de novo* pathogenic variant in *BRCA1*
NM_007294.4:c.5516T>C

Exon 24 Exon 3 Exon 2 Exon 1

NC_000017.11 :g.43045753A>G

Chromosome identifier
Position
Alternate nucleotide
Reference nucleotide

Genomic Position: "g."

Genomic position is based on the "plus" strand and it starts at position 1 on the chromosome.

Exon

NM_007294.4:c.5516T>C

Transcript identifier
Position
Alternate nucleotide
Reference nucleotide

Coding Position: "c."

Coding position refers to the nucleotide positions within the exonic coding sequence of a gene, starting with "1" at the ATG start codon and extending through until the stop codon. For genes on the "minus" strand, the nucleotides are the complement of the genomic sequence.

Tyr Ser Gly Ser

NP_009225.1:p.Leu1839Ser

Protein isoform
Reference residue
Alternate residue
Position

Protein Position: "p."

Protein position denotes amino acid positions in a protein, from N- to C-terminal, reflecting changes due to change in the gene's coding region.

! The exact positions may differ between genome builds and transcript variants; for this example, here we use the GRCh38 (hg38) human reference genome assembly and the NM_007294.3 RefSeq transcript for *BRCA1*, which is the MANE select transcript model. Learn more at: https://www.ncbi.nlm.nih.gov/refseq/MANE/

Each variant is assessed based on the following characteristics: whether it has previously been identified as causing disease in a clinical database, its frequency within the population, the potential it has to damage gene function or expression, and how it segregates with disease occurrence in a familial context.

Clinical Databases
- Has the variant been previously identified as causing disease?
- Is the gene of relevance to the disease in question?

ClinVar OMIM

Population Databases
- What is the frequency of variant allele in the healthy population?
- Is this gene commonly mutated within the population with impactful variants?

gnomAD UKBB

In silico Predictions
- Does the variant alter the protein sequence or introduce a stop codon?
- Is the base pair evolutionarily conserved across species?

CADD SpliceAI

Inheritance
- Is the variant *de novo*?
- Does the variant segregate with the phenotype?
- If the condition is compound heterozygous, are the variants present on different alleles ?

Pedigree Trio Sequencing

Through this process, candidate variants are categorized on a scale ranging from benign to pathogenic. However, despite these efforts, many variants may still remain classified as "VUS". To further categorize these as either benign or pathogenic, additional evidence is required. Techniques such as familial segregation analysis, cellular/animal modeling, and molecular/biochemical profiling can provide the necessary information to clarify the status of a VUS as either benign or pathogenic.

Benign | Likely Benign | **V**ariant of **U**ncertain **S**ignificance (VUS) | Likely Pathogenic | Pathogenic

Inheritance Testing

Molecular/biochemical profiling (e.g. transcriptome/metabolome)

Cellular/animal modelling

Common Disease Risk via Genome-wide Association Studies and Polygenic Risk Scores

Common diseases caused by common variants

Complex diseases result from the cumulative effects of 100s to 1000s of common genetic variants of relatively small effect sizes interacting with the environment over time to manifest the trait.

*For example: Alzheimer's disease, asthma, coronary artery disease, obesity*

Generally, a "common" variant is seen in >5% of the population.

⚠ Variant frequencies differ by genetic ancestry, so what is considered "common" is relative to the referent population.

Variant Effect Size

Rare variants, rare disease

Common variants, common disease

0.001 — Allele Frequency in Population — 0.05
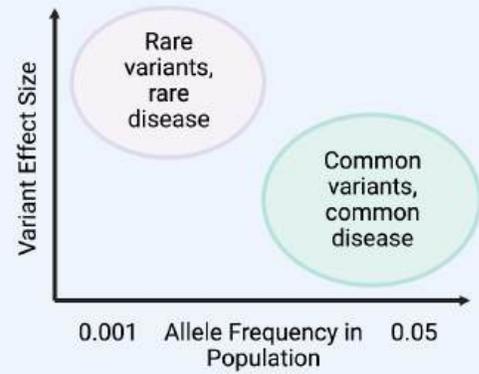
How do researchers find common variants associated with disease? Genome-Wide Association Studies (GWAS)!
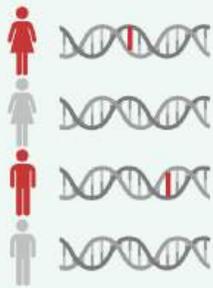
1. Genetic variant information is collected from a large population sample (early days 10,000-100,000; now 100,000-1 million)

2. Each person is labeled as either a case (have disease) or control (no disease)

3. Effect sizes and associated p values for all genetic variants are calculated based on how often the variants are present in cases and not in controls

each dot is a genetic variant from the SNP-chip

$-log_{10}(P)$

Chromosome
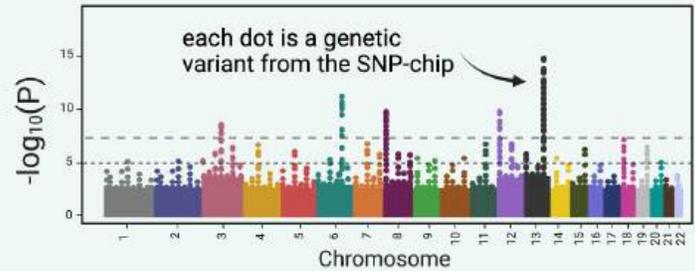
Higher negative log-transformed p value (y axis) means higher probability that this variant is associated with the disease

## Clinical applications using GWAS

Many stakeholders are hoping to use information from GWAS to inform population-based screening for common complex diseases to help guide primary prevention interventions.

You can calculate an individual's relative risk of developing disease using GWAS results:

Polygenic Risk Scores (PRS) computing methods calculate a score integrating the number and effect sizes of the risk variants they carry, adding them up to estimate overall risk from many genetic risk variants.

Low Risk   Medium Risk   High Risk
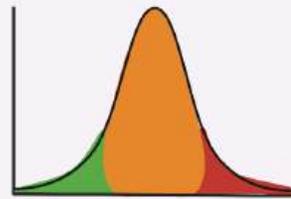
Risk assessment studies for many common diseases are suggesting that individuals in the highest PRS risk group (i.e., 10th percentile) would benefit the most from preventive interventions.
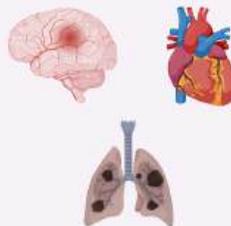
Researchers can also identify potential drug targets using GWAS results:

Applying gene network analyses, they can identify molecular pathways shared by many variants, and in what tissue(s) these pathways are expressed.

Concerns about the relative utility and generalizability of PRSs.

Many GWASs are made from mostly "White European" population samples. PRSs made from these GWASs are less accurate in assessing risk for individuals with non-European genetic ancestry.

Prediction Accuracy Relative to European

1.0

Middle Eastern   East Asian   African
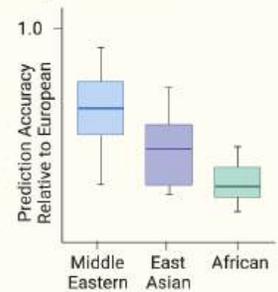
Many efforts are underway internationally to develop diverse-ancestry-informed GWAS.

Several common diseases show pronounced sex differences in prevalence, age of onset, presentation (pathology), comorbid risk factor profiles, etc.

However, many GWASs regress sex out as a covariate (ignore within the model) and exclude X- and Y-chromosome genes from analysis. This does not allow for the genetic variants encoding mechanisms behind these sex differences to be identified.

More sex-disaggregated GWAS need to be done to identify which disease pathways are sex-specific and shared between both sexes.