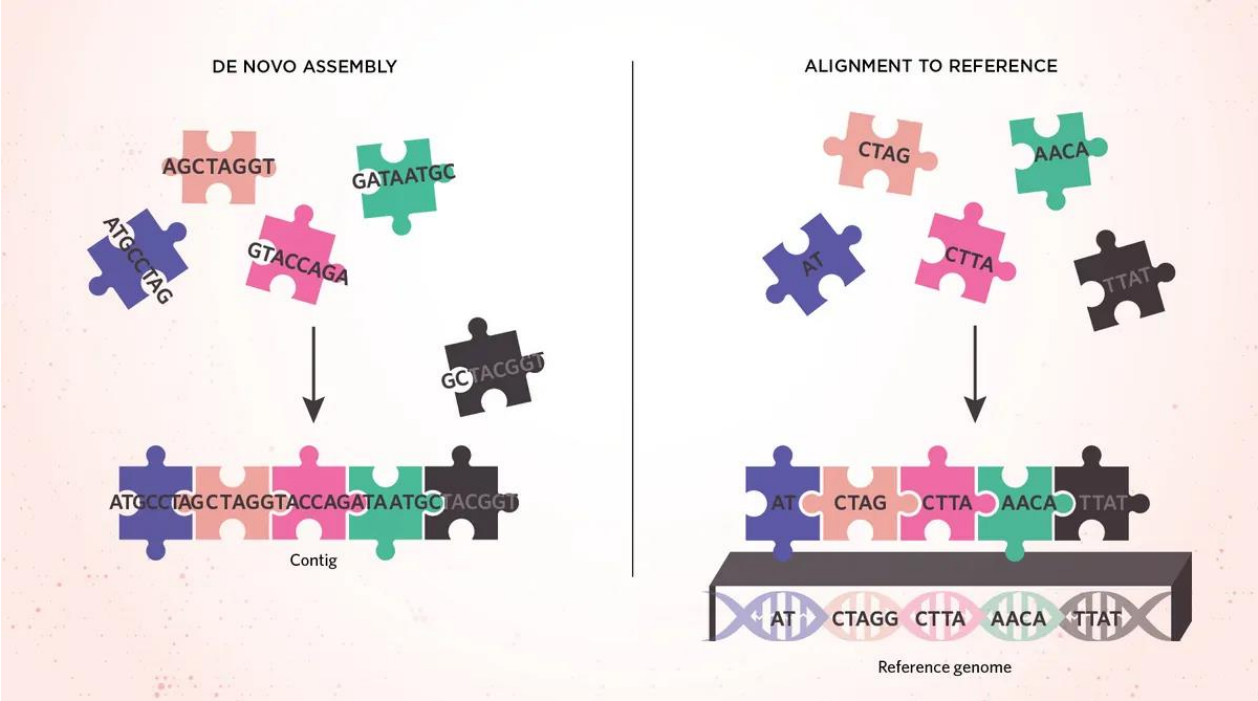
Learning Your Data: Data Processing of Genetic Data.

Benjamin Kaufman

PhD Student in Human Genetics

What is Genome Build?

A human genome reference build is essentially a "standard map" of human DNA - it's the agreed-upon sequence of all 3 billion DNA letters that serves as the universal reference point for comparing everyone else's genomes



MODIFIED FROM © ISTOCK.COM, FILO

What is a Genome Build?

As sequencing technology improves and we discover more about human genetic variation, the Human Reference Genome is updated to fix errors and fill in gaps. Each iteration of the Reference Human Genome is called a build!

hg18 (GRCh 36)
Debut: 2006

→

hg19 (GRCh 37)
Debut: 2009

→

GRCh3 8
Debut: 2013

→

Next Build?
- Stay tune d!

- The builds essentially represent snapshots of our best knowledge at the time.
- As technology gets better, we can sequence harder regions, fix mistakes, and create more accurate reference maps for everyone to use.

Variant Calling

```
##fileformat=VCFv4.2
##fileDate=20250712
##source=PLINKv1.90
##contig=<ID=1,length=249218993>
##contig=<ID=2,length=243048761>
##contig=<ID=3,length=197833759>
##contig=<ID=4,length=190939666>
##contig=<ID=5,length=180696890>
##contig=<ID=6,length=170919471>
##contig=<ID=7,length=159119221>
##contig=<ID=8,length=146296415>
##contig=<ID=9,length=141066492>
##contig=<ID=11,length=14940417>
```

Mandatory Columns for VCFs to be correctly read

ID: The chromosome/contig identifier

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	BenK_ID001
1	734462	rs12564807	G	A	.	.	PR	GT	0/0
1	752721	rs3131972	A	C	.	.	PR	GT	0/0
1	760998	rs148828841	C	A	.	.	PR	GT	0/0
1	776546	rs12124819	A	G	.	.	PR	GT	0/0
1	787173	rs115093905	G	A	.	.	PR	GT	./.

chromosomal positions and associated metadata, making it the go-to format for sharing genomic variation data

```
##contig=<ID=24,length=59032[illegible]0>
##contig=<ID=26,length=165[illegible]6>
##INFO=<ID=PR,Number=0,Type=Flag,Description="Provisional reference allele, may not be based on real reference genome">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
#CHROM  POS  ID  REF ALT QUAL  FILTER INFO  FORMAT BenK_ID001
1  734462  rs12564807  G  A  .  .  PR  GT  0/0
1  752721  rs3131972  A  C  .  .  PR  GT  0/0
1  760998  rs148828841 C  A  .  .  PR  GT  0/0
1  776546  rs12124819  A  G  .  .  PR  GT  0/0
1  787173  rs115093905 G  A  .  .  PR  GT  ./.
```

How do I know what Build my dataset is in?

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	BenK_ID001
1	734462	rs12564807	G	A	.	.	PR	GT	0/0
1	752721	rs3131972	A	C	.	.	PR	GT	0/0
1	760998	rs148828841	C	A	.	.	PR	GT	0/0
1	776546	rs12124819	A	G	.	.	PR	GT	0/0
1	787173	rs115093905	G	A	.	.	PR	GT	./.
1	798959	rs11240777	G	.	.	.	PR	GT	0/0

How do I know what Build my dataset is in?

An official website of the United States government Here's how you know

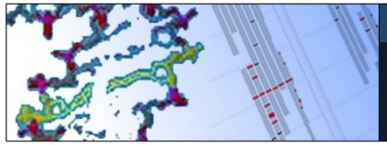
NIH National Library of Medicine
National Center for Biotechnology Information

Log in

Service Alert: Planned Maintenance beginning July 25th

Most services will be unavailable for 24+ hours starting 9 PM EDT. Learn more about the maintenance.

dbSNP SNP Search
Advanced Help



dbSNP

dbSNP contains human single nucleotide variations, microsatellites, and small-scale insertions and deletions along with publication, population frequency, molecular consequence, and genomic and RefSeq mapping information for both common variations and clinical mutations.

Getting Started
- dbSNP 25th Anniversary
- Overview of dbSNP
- About Reference SNP (rs)
- Factsheet
- FAQ
- Entrez Updates (May 26, 2020)

Submission
- How to Submit
- Hold Until Published (HUP) Policies
- Submission Search

Access Data
- Web Search
- eUtils API
- Variation Services
- FTP Download
- Tutorials on GitHub

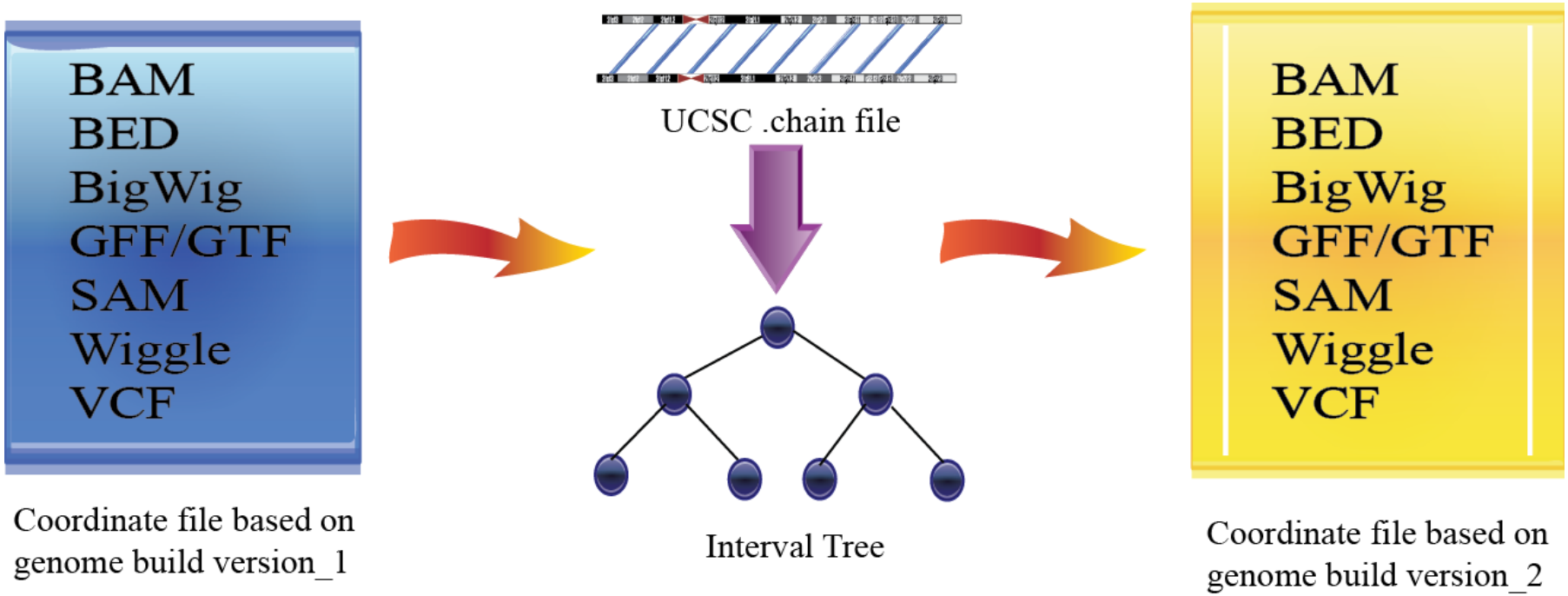
Important: When using dbSNP, please cite the resource using the following publication: The evolution of dbSNP: 25 years of impact in genomic research.

ALFA Project Release 4 with over 900M variants from 400K subjects is now available (May 15, 2025)

The goal is to provide allele frequency from more than 1 million dbGaP subjects with regular updates. Visit the project page for more information or view the introduction video below.

dbSNP

What is liftover?

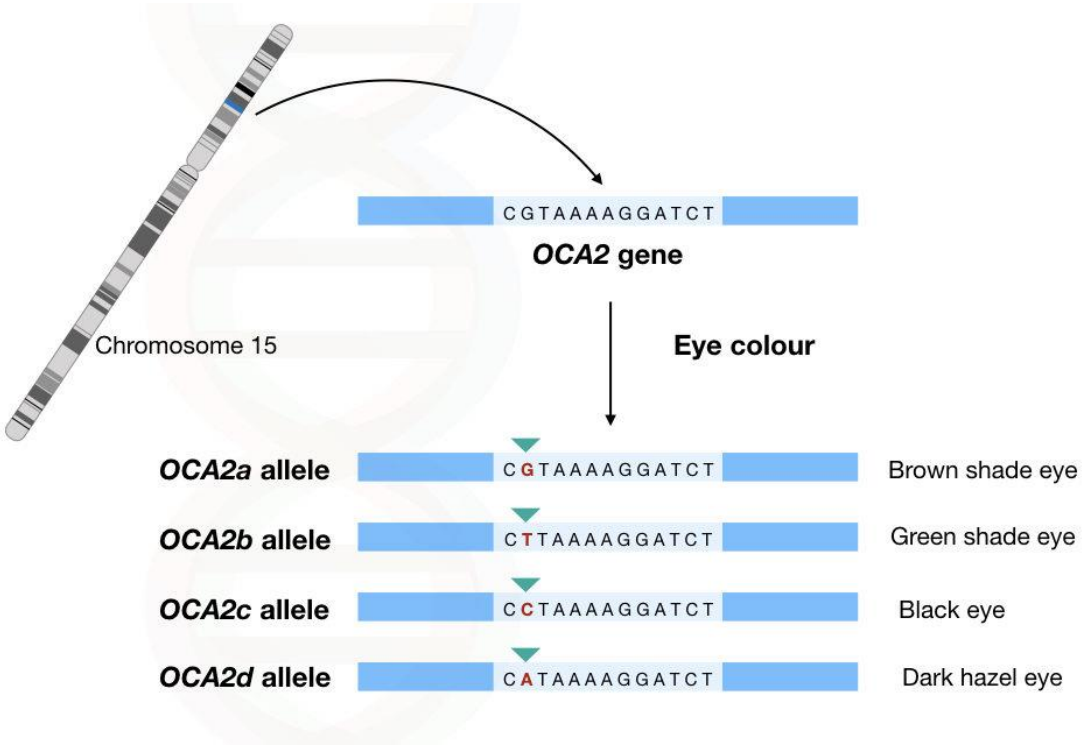


Coordinate file based on genome build version_1

Coordinate file based on genome build version_2

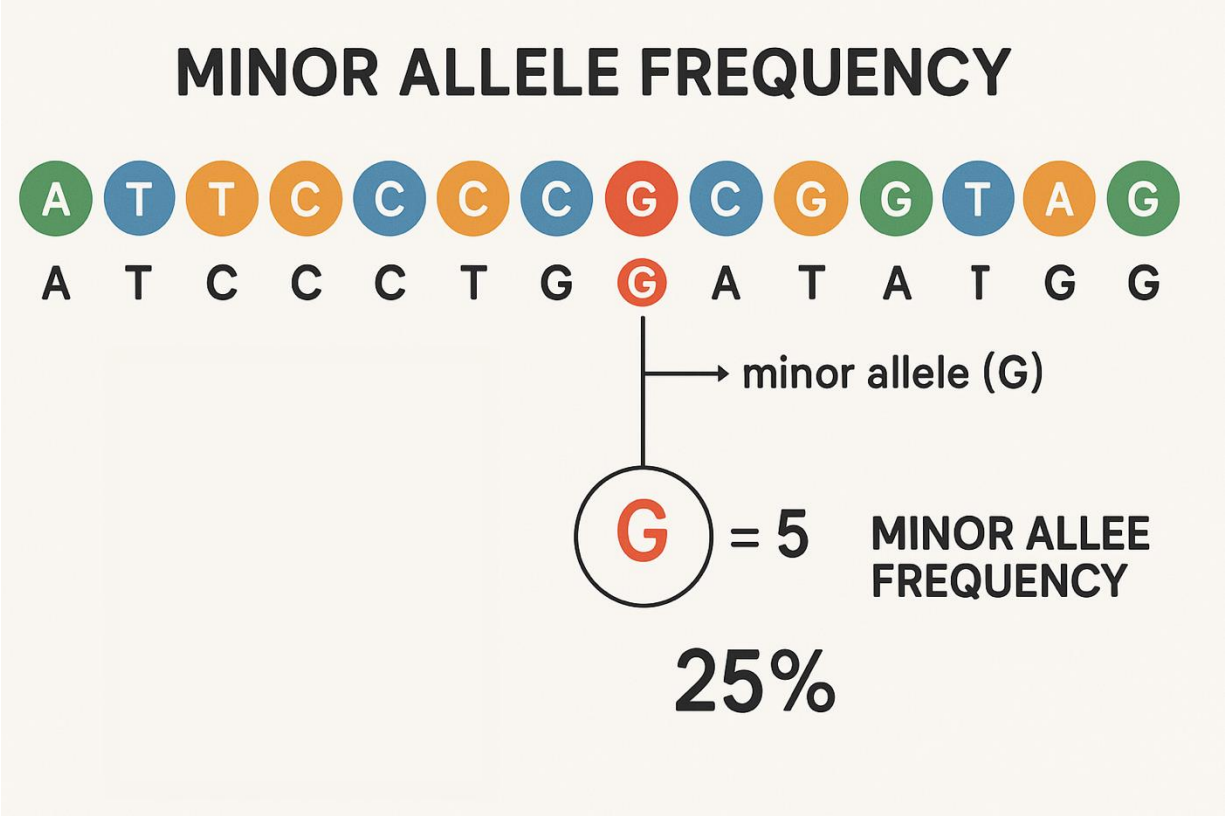
CrossMap

What do we mean by reference allele?



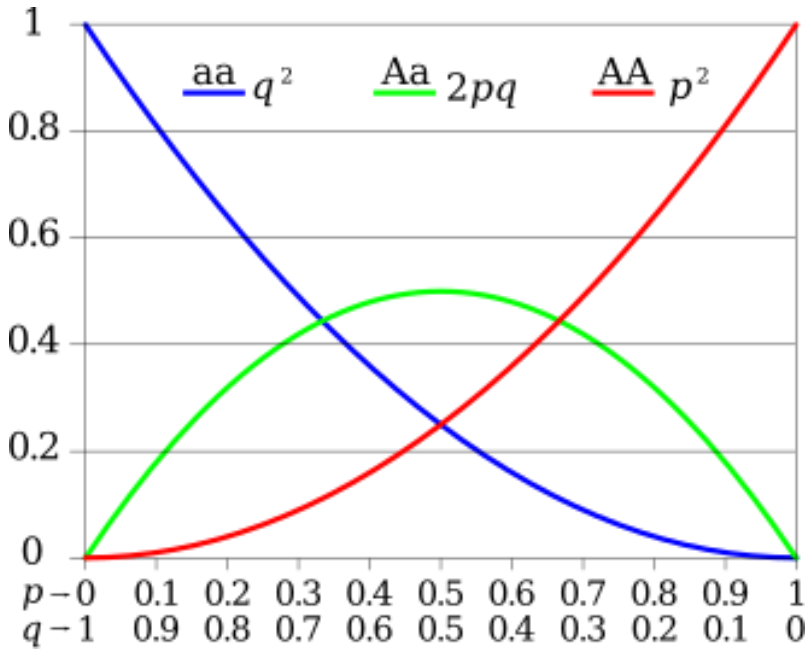
https://geneticeducation.co.in/gene-vs-allele/

What is Minor Allele Frequency?



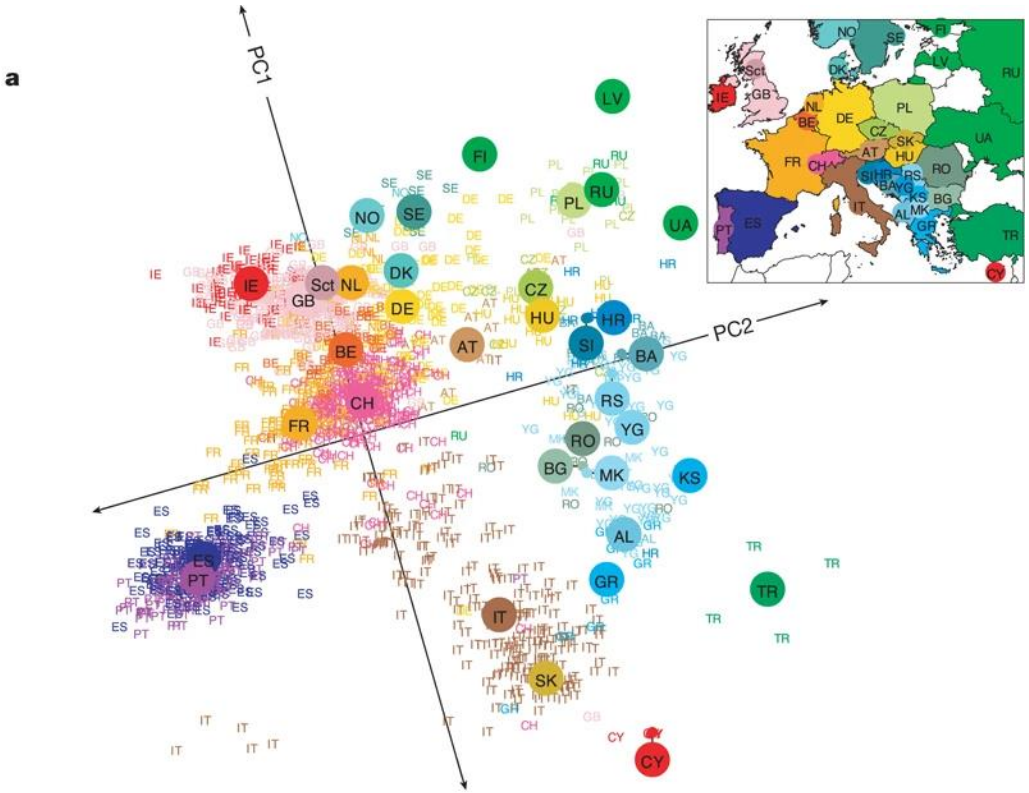
Hardy-Weinberg Equilibrium

- Hardy-Weinberg equilibrium describes the relationship between allele frequencies and genotype frequencies in an idealized population
  - Does it mean we shouldn't use statistically Hardy-Weinberg Equilibrium, indicate technical issues or modeling biological phenomena
    - Mutations
    - Non-random mating
    - Infinite population size
    - No natural selection



Principal Component Analysis (PCA)

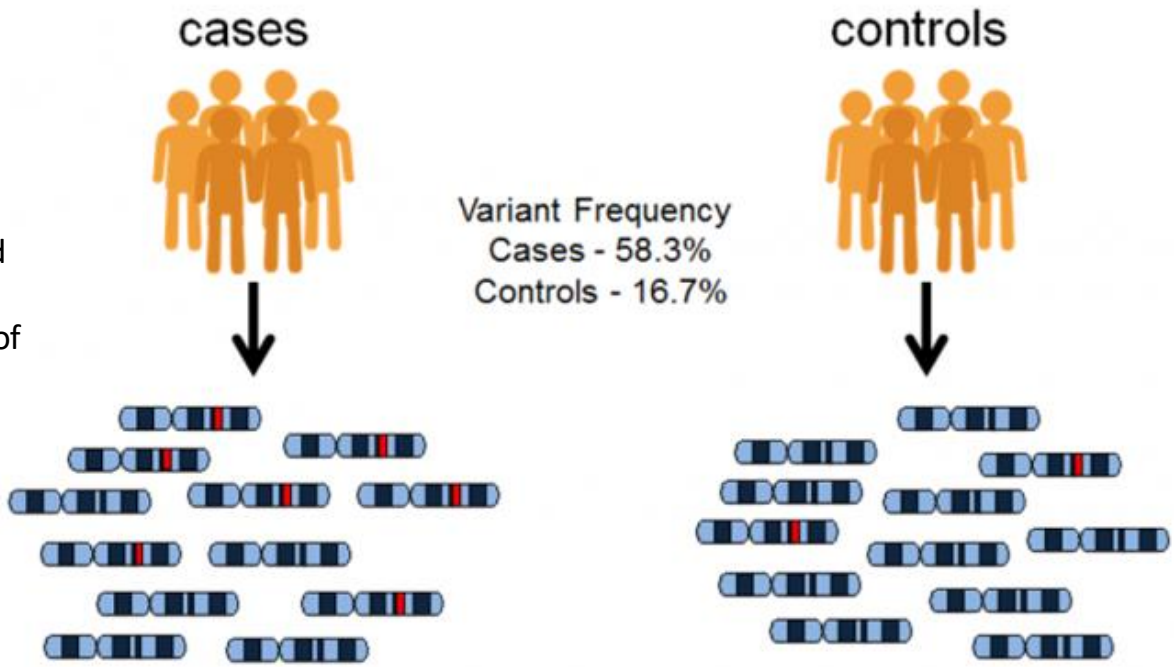
- PCA is a statistical method that reduces the complexity of datasets by finding the main patterns of variation.
- PCA is often used to examine population structure and ancestry patterns from genome-wide SNP data



10.1038/nature07331

Genome-Wide Association Studies (GWAS)

- Genome-Wide Association Studies (GWAS) scan the entire genome to identify genetic variants associated with diseases or traits
- Because we are making hundreds of thousands to millions of comparisons, we use a different p value threshold



https://www.ebi.ac.uk/training/online/courses/gwas-catalogue-exploring-snp-trait-associations/what-is-gwas-catalog/what-are-genome-wide-association-studies-gwas/

Manhattan Plots