Genome-Wide Association Studies (GWAS)

GWAS 1: Case-Control Association Testing

GWAS 2: Quantitative Traits

Nov 07 2025

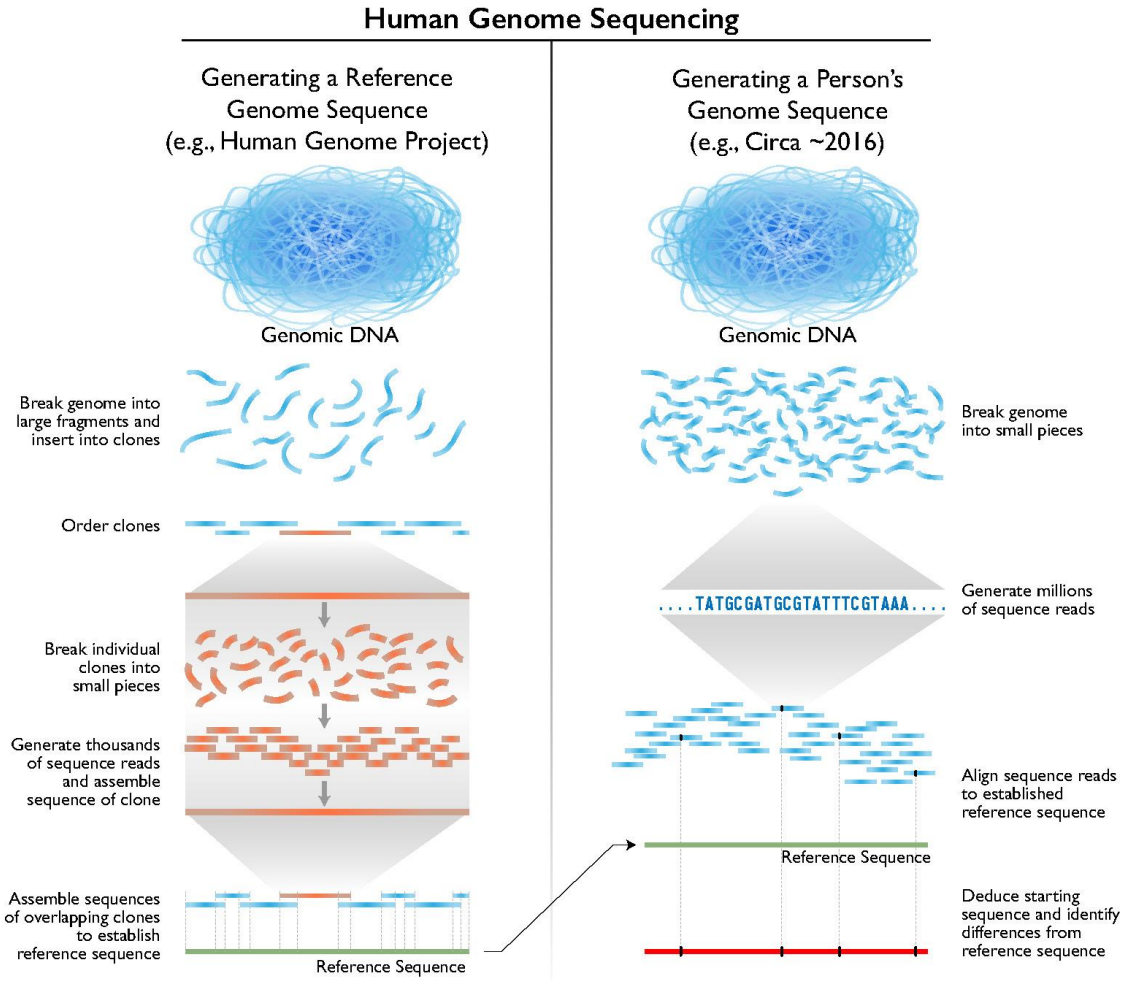
Phuc-Loi Luu, PhD

Email: luu.p.loi@googlemail.com

Zalo: 0901802182

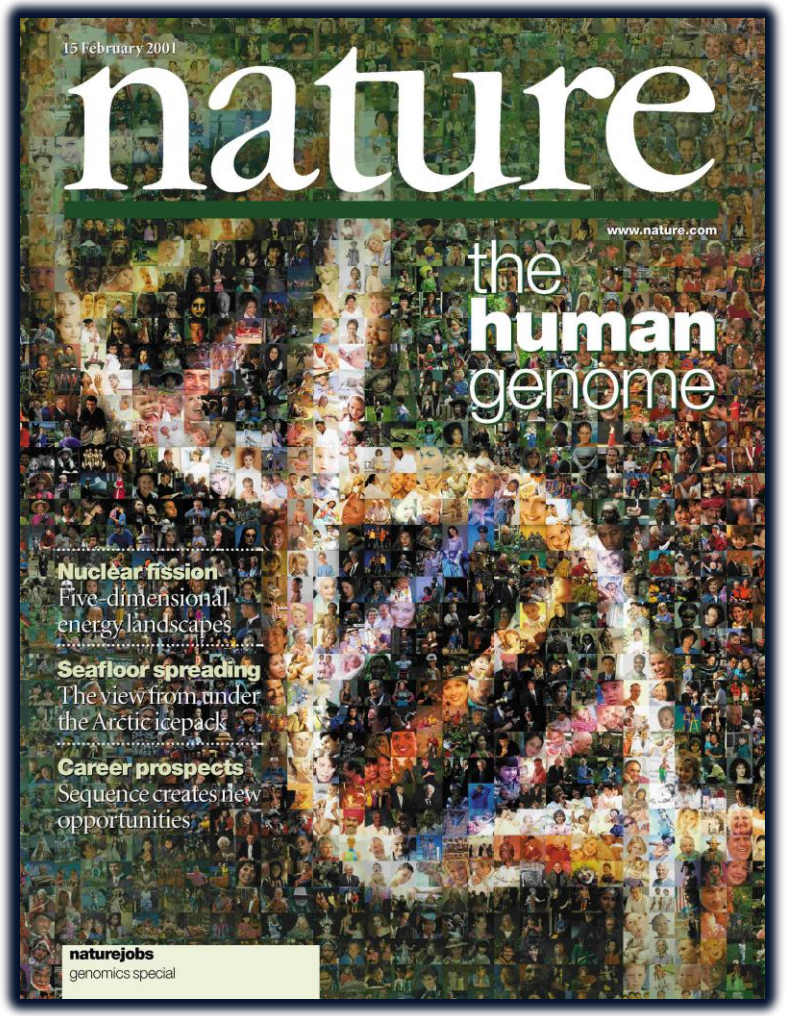
Human Genome Project: shotgun method

Once significant human genome sequencing began for the HGP, a 'draft' human genome sequence (as described above) was produced over a 15-month period (from April 1999 to June 2000). The estimated cost for generating that initial 'draft' human genome sequence is **~$300** million worldwide

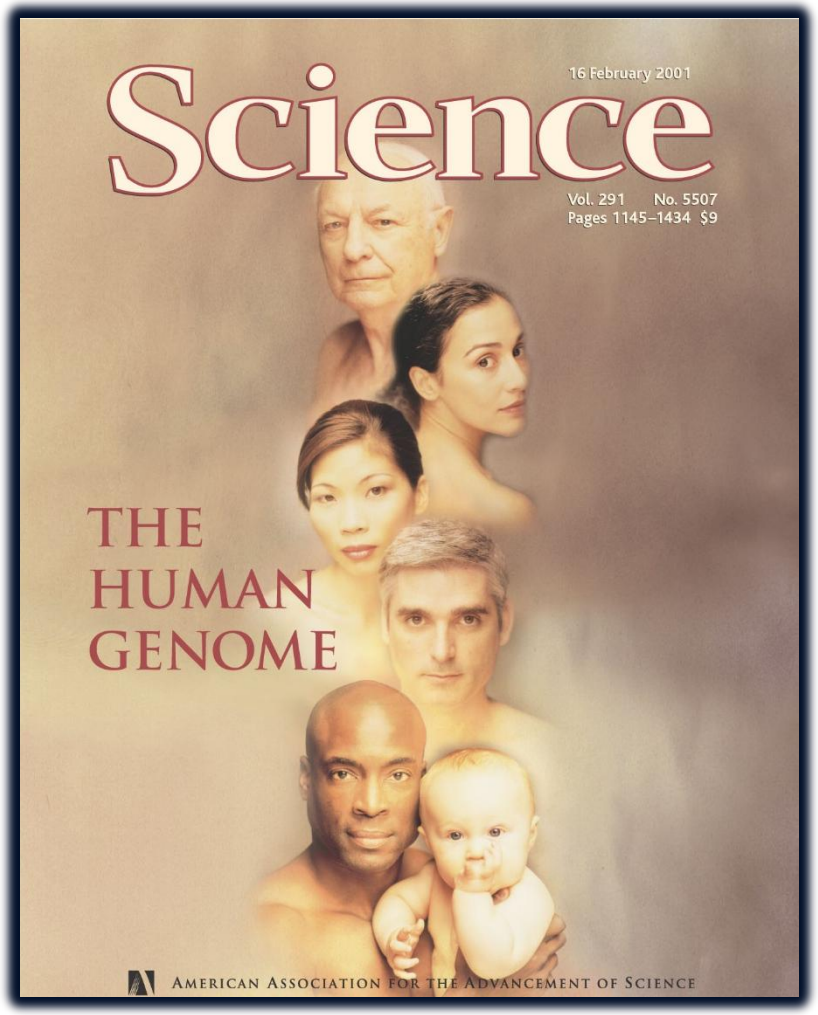


https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

February 2001: Papers Reporting Draft Sequence of Human Genome



HGP Paper



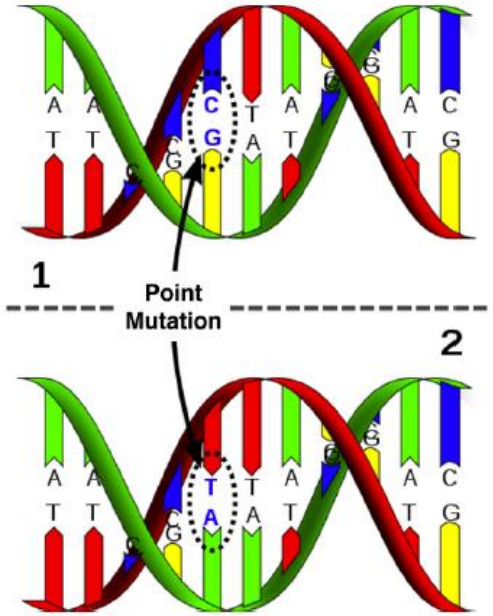
Venter/Celera Paper

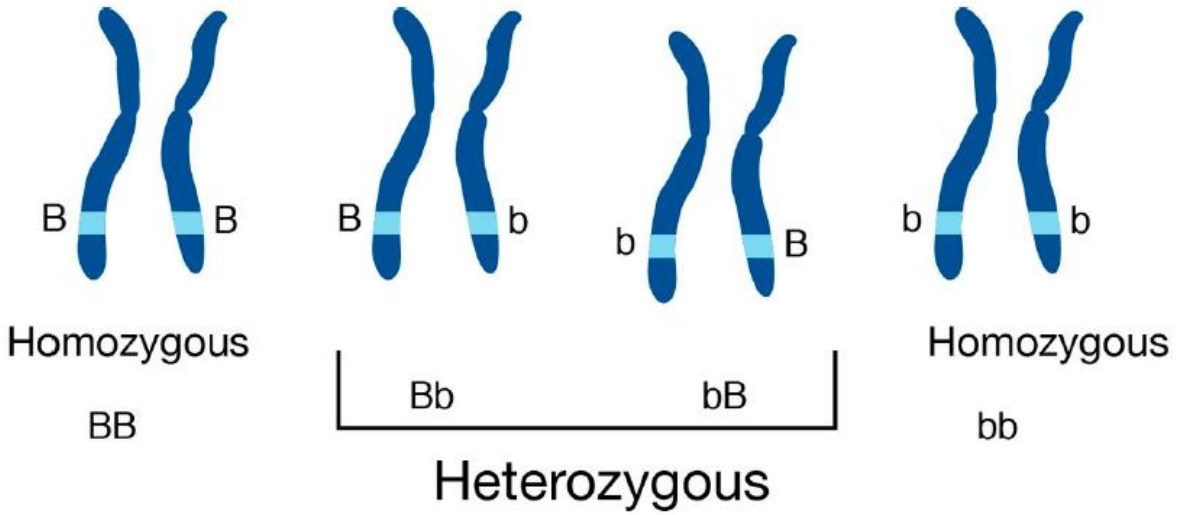

```
48541 agcccttcaa agaaatgttc tcagcaggca tggagcccag gacttgctcc ctttgggtag
48601 agagccgggt tgaaggtgac tgaagtgaaa tgggacagta gaggcggggg gggtggtgag
48661 ttcctggagg tgggggggtgt gggaacctgc tgtgtactga gatgcacccc tgccagttct
48721 gcctgaagat ttgaggcggg gggcaggggg gcggagtgaa gtcatttac tggtaagtaa
48781 ttttaaacct tttaatatta aagcaaacgt ggatatgtaa tgaatgaaat tcattctgga
48841 atgaaaaatt cacgtgatgt tgaaaaataa cacggggctt cagagaggac tttctggctg
48901 gcagcagact ccagattccc agggcccctg caccctcctc tgcccacagg gcaccttaat
48961 ggagaaggtg tgggaggaga gccaggccgg agtcagagca cactggtgac tccacatttg
49021 cagcgtgccc tgcctctctc ctgaggcttg gcaacgtgca atatgctaag caaactcccc
49081 ctgtccccgt ccagtttctg aggacaagag ccaccacctg tagcaaataa agacccagca
49141 acccttgac tcatctttgt gagtctctgg aatcagaggg tagccacatc gctgagaggt
49201 ggagtgaagc actcgggtga aaaggtacaa ggaagtcagg gacaggagtg tggggacatc
49261 acctagacaa tgacagagaa gaggggcaca gccgagtgag gggagagggg ccggcagtcc
49321 tacatcccct ggcctgaagc acgctccagg gcagaaggaa aaacactgtc tttggggtcc
49381 aagagacctg agttcaaatt ctggctccac cactgaccac ctgtgtaacc ttgaactgct
49441 gctgcctgaa cctcaggttc cccttctaaa aatagaggag aaaaggatgc atttctcctt
49501 gcccctgtga gaacgaaatg gtgcaagcac caaggagcct cagcaaaggt cgggcctgcc
49561 cccgcctggc caaacctttc ctcttcagga ggccacggca accgtagttt gacagaagag
49621 cagcaccttg atttaatgct tcccagcatg tgtccttgag caagtcacct aacctctctg
49681 ggctgcttcc tcattgggaa aatatggctg ccagtaaaac ctgccctgtc cacctcctgg
49741 ggcacttggc aaacagcaaa agagtccaaa tgtgcaggct gggccaggcg cagtggctca
49801 tgcctgtaat cccagcaatt taggaagcca aggtgggcgg atcacctgag gtcaggagtt
49861 tgagaccagc ctggccaaca tggtgaaacc ttgtctctac aaaaatacaa aaattagccg
49921 ggcatgatgg cgggtgcctg taatcccagt tactcgggag gctgaggcaa gagaatcgct
49981 tgaacccgga aggggaaggt tgcagtgagc caagattgtg ccactgcact ccagcctggg
50041 caacagagcg agactctgtc tcaaaaaaaa aaaaaaaaaa aaacaatgca gagctggctg
50101 tgtaaaaaac ctgttccact gcagggccca gtgtccacca ggctggggtg caggcctatg
50161 gggtgggggc ccagcatcag cctctcagca gccctgggag gcggggcgca tcccgtgccc
50221 ctcgtggtct ggatgtgttc tagcccaagt cctaggttac acctgccgtc gcctggcctc
50281 tcaggagagg cccagggtga ggaggagcat ggtaaaggtg aagctgattg ggaagtcagc
50341 tgttgggaaa gcaactcctt gcacattgga ggaaccgaga aagactgacc ccgaggacag
50401 cagccagcat ggcccttcct gggagcccat gttgggggat tcctgctgca gccaaggctc
50461 agcccttgtg gtcgcaggtg ctggtcctgg cctcttcccc tcccatgcag gagcacagga
50521 gagatggctt ctgaggacct gttgcagctg tggccctggg aatagatttg ccagggagct
50581 ttaaagcagc tgagtgtgtc atccagctaa gcctggggaa ggagcttggc tcaggtcctg
50641 acaggtgtga cagggatggg gactgggaag taagagatga aaccctggct ggaggctgtg
50701 agcttccaca gccagcgctg gacagggagg gtccagatat acccactagt gccctcacca
```

https://www.ncbi.nlm.nih.gov/nuccore/806904736

Human Genome Variation

- Humans genomes are >99% similar by sequence
- A typical human genome has ~5 million variants with 3-4 million single-nucleotide variants
- Humans are diploid

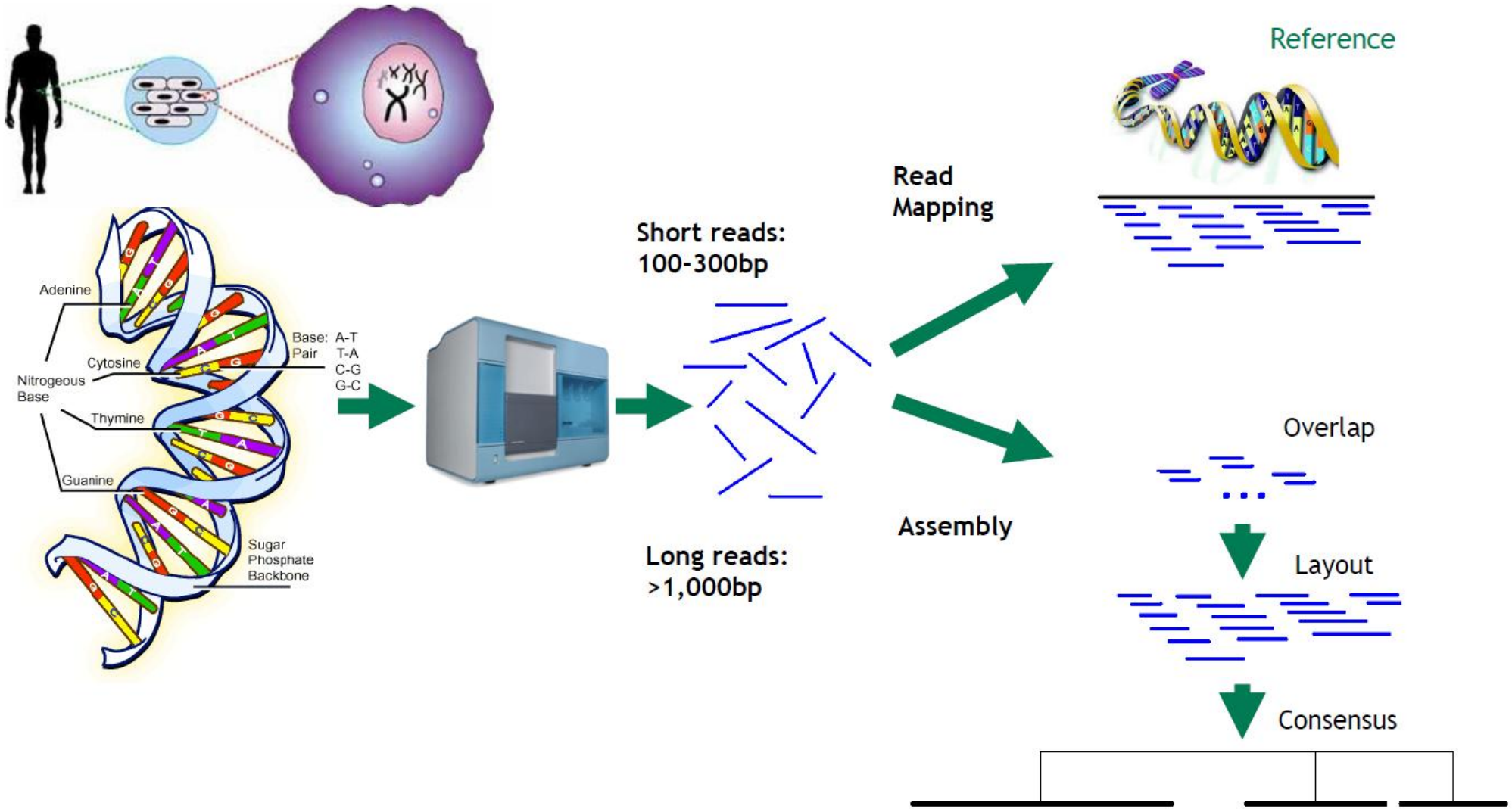




Source: https://rosalind.info/media/point_mutation.png, https://en.wikipedia.org/wiki/Zygosity

Tobias Rausch

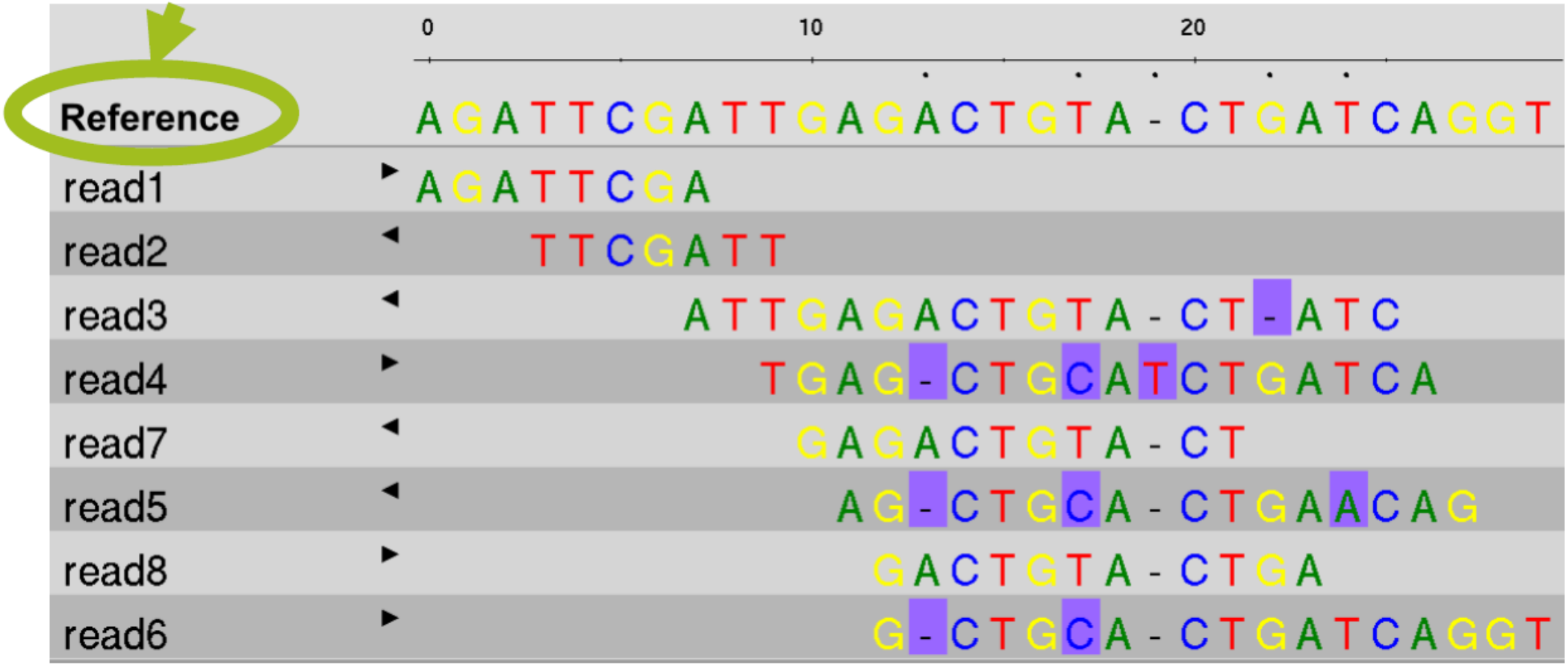
Sequencing DNA



Tobias Rausch

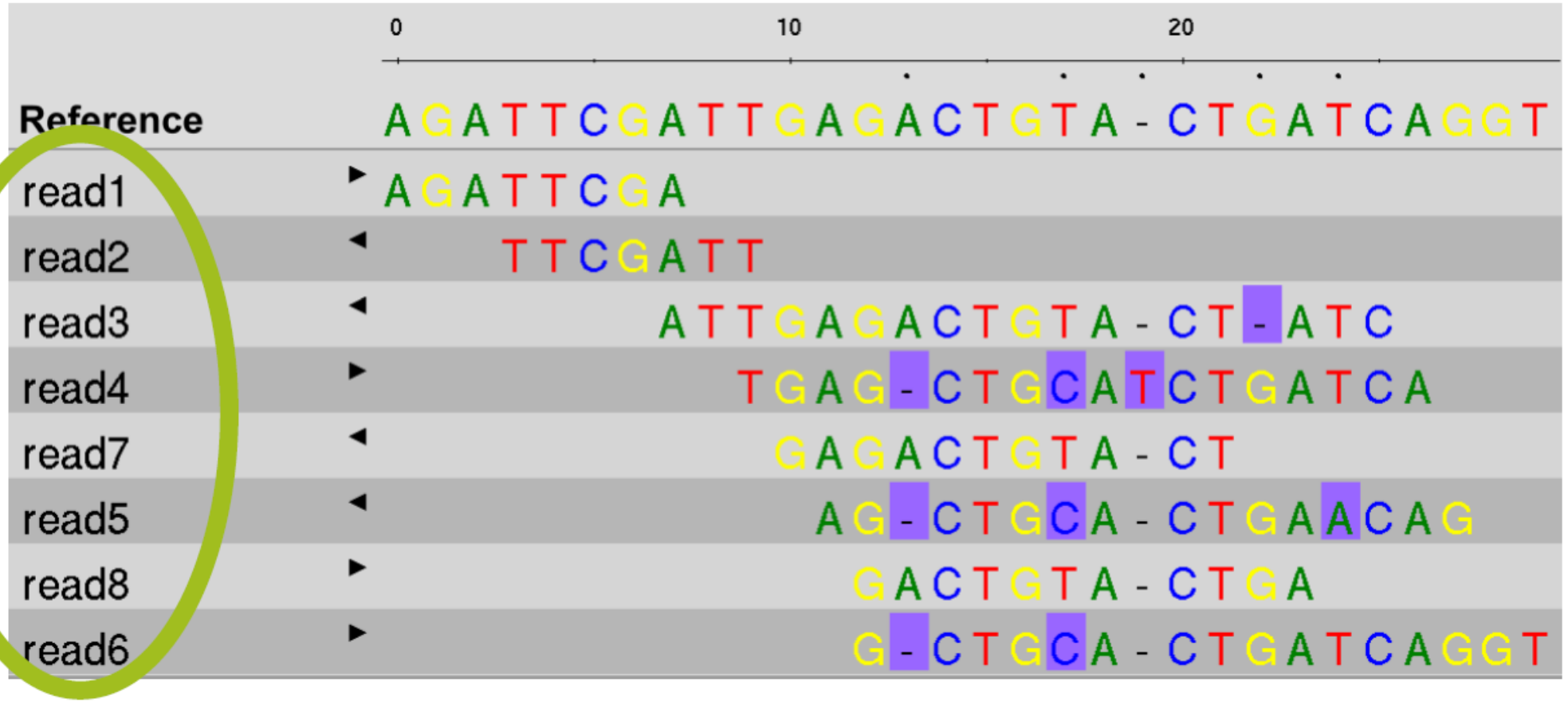
Read Alignment

Reference sequence



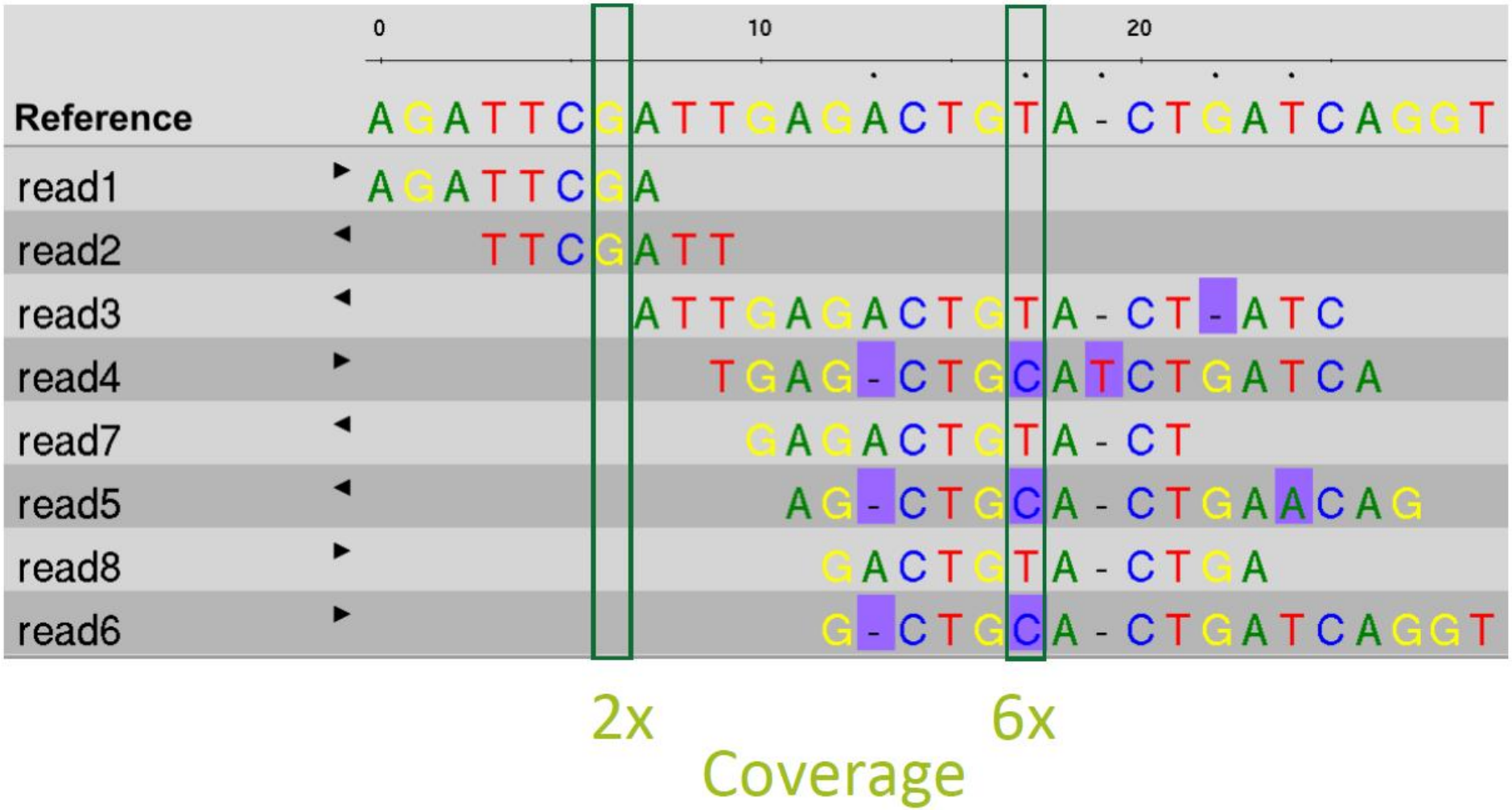
Tobias Rausch

Read Alignment



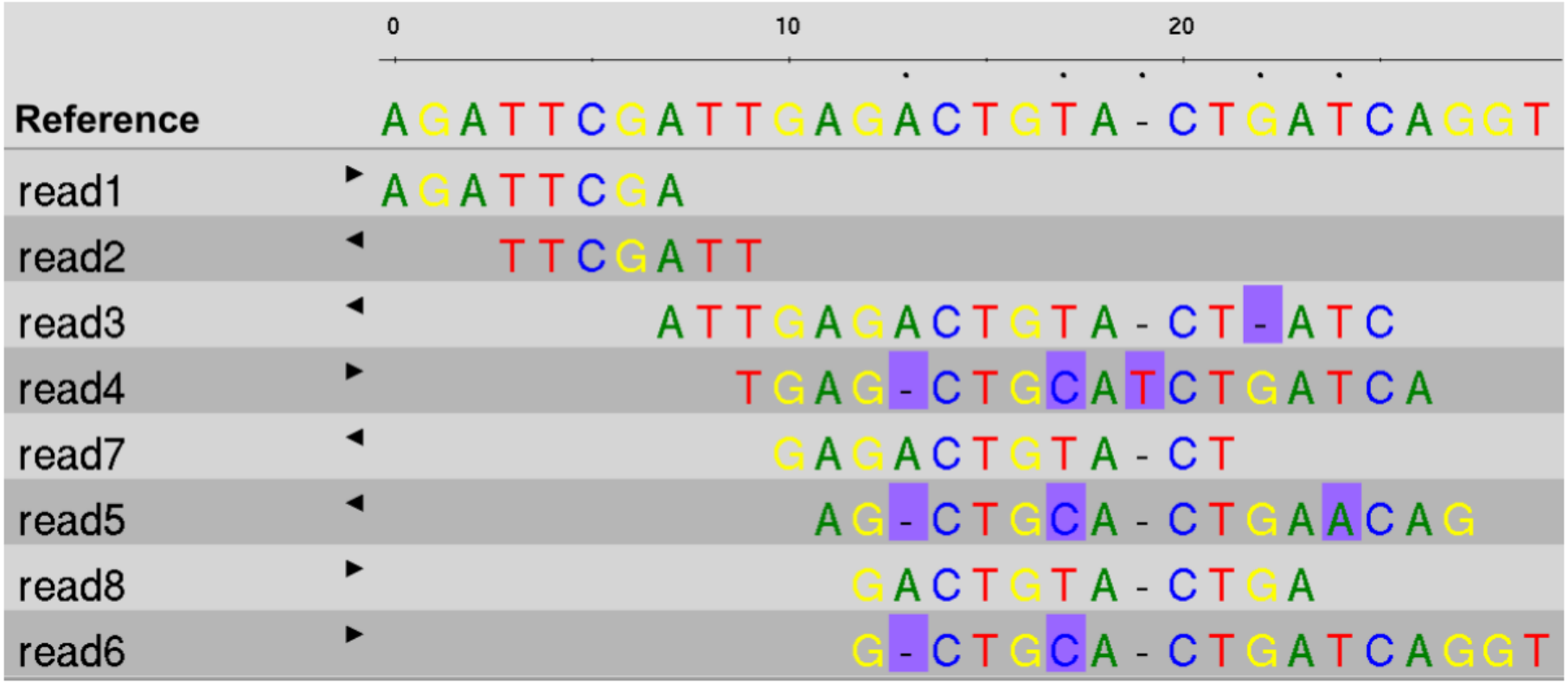
Tobias Rausch

Sequencing Coverage



Tobias Rausch

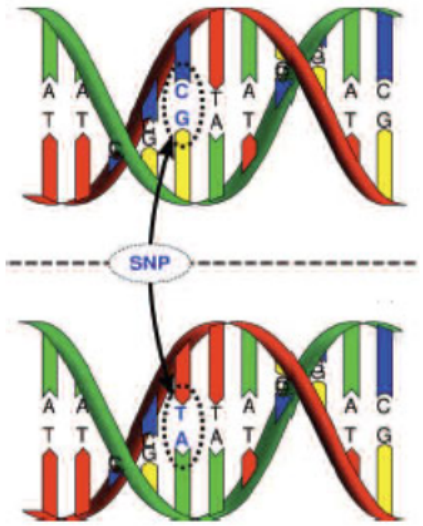
DNA Variant Detection



Variations: Deletion & Single-nucleotide variant

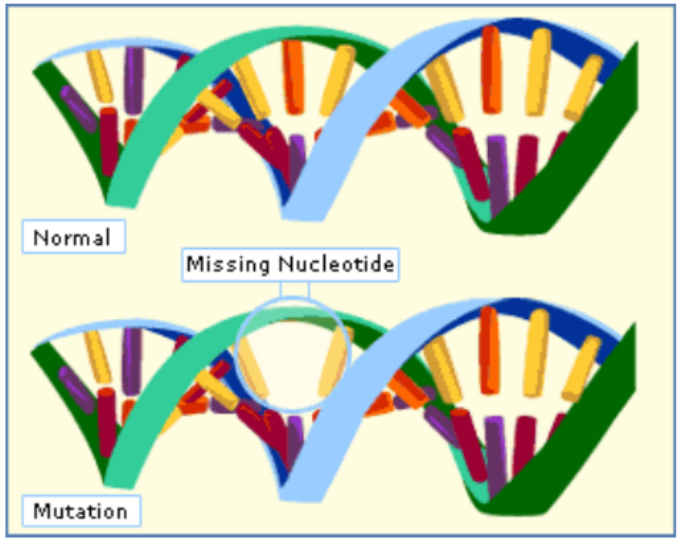
Types of Variants

Single-nucleotide variants (SNVs)



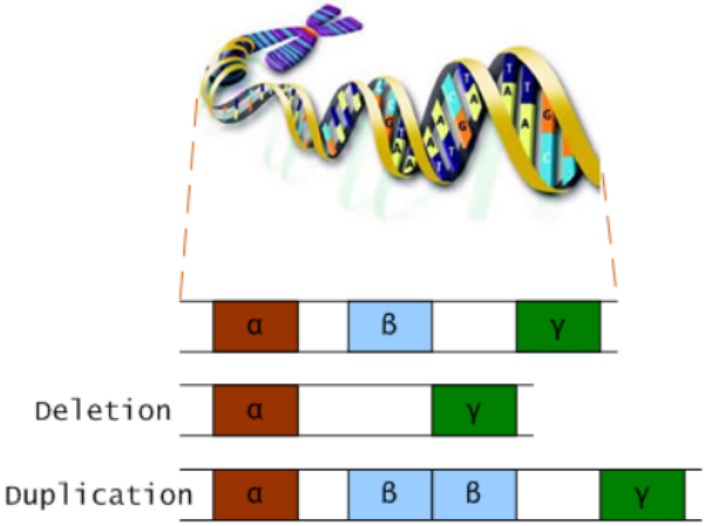
Size: 1bp

Short insertions & deletions (InDels)



1-50bp

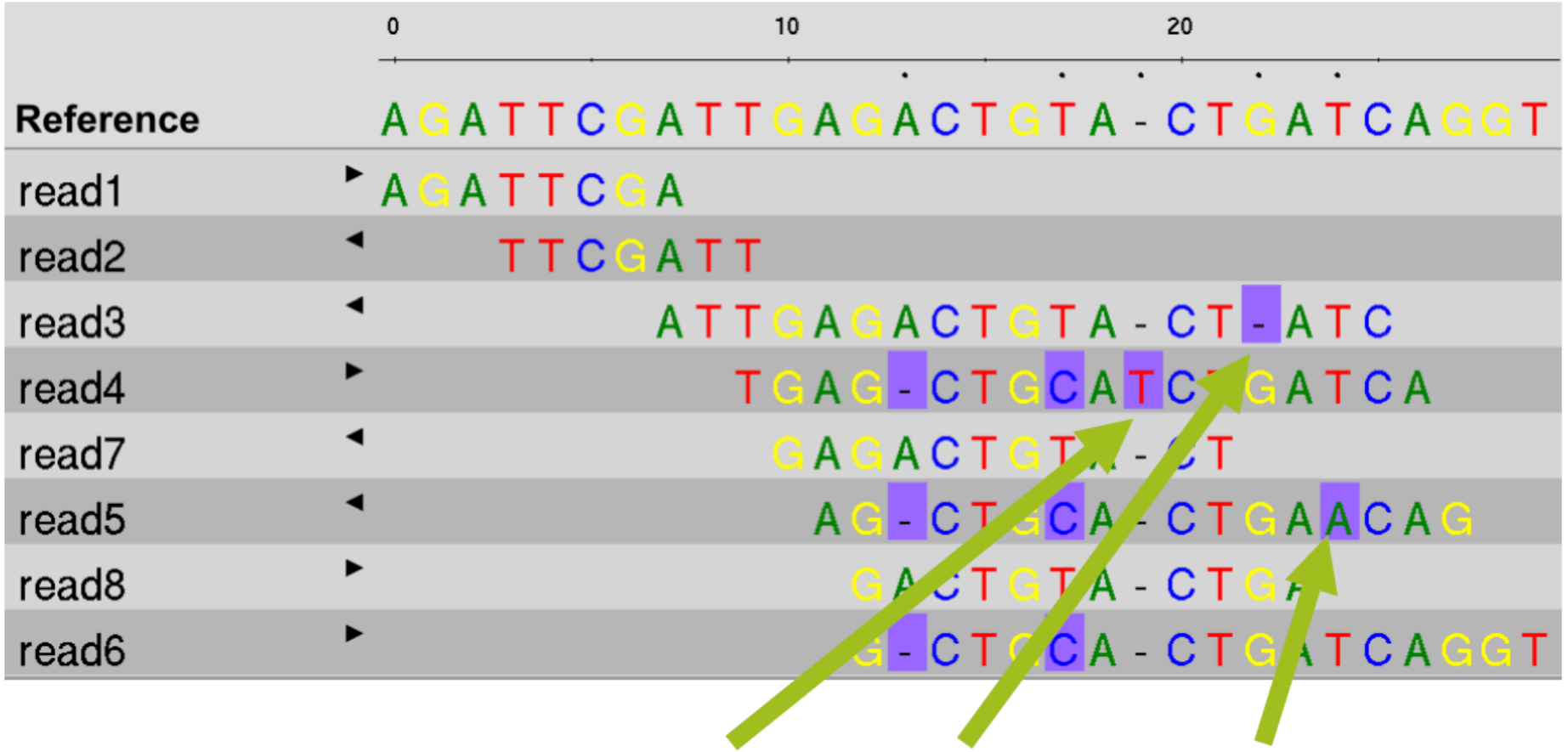
Copy-Number Variants (CNVs) & Structural Variants (SVs)



>50bp

Different methods are used to discover and genotype SNVs, InDels, SVs and CNVs.

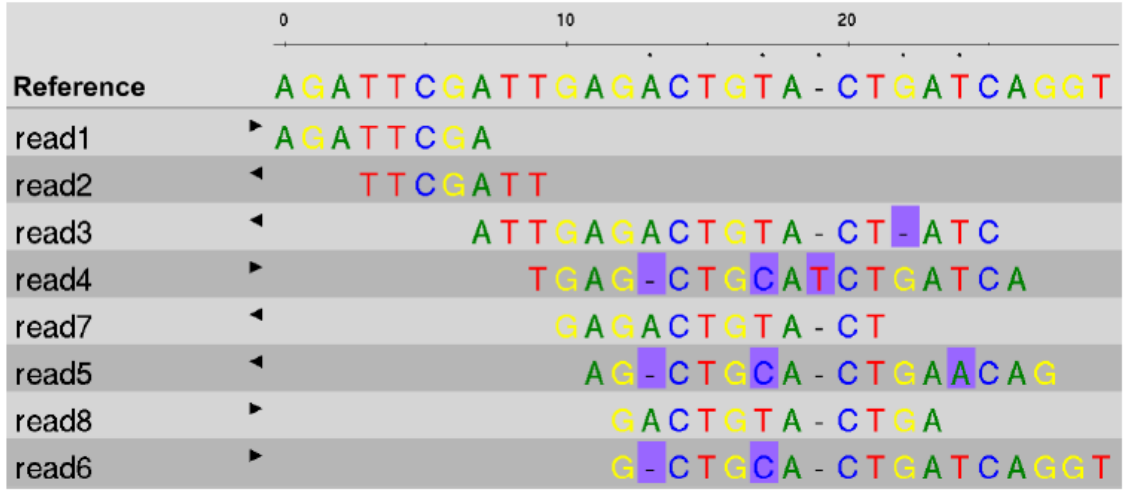
Sequencing Errors



Sequencing errors: Insertions, deletions & basecalling errors

Variant Calling - Data Transformation

Alignment



SAM/BAM file

List of Variants

CHR	POS	ID	REF	ALT	GT
chr1	12	.	GA	G	0/1
chr1	17	rs123	T	C	0/1

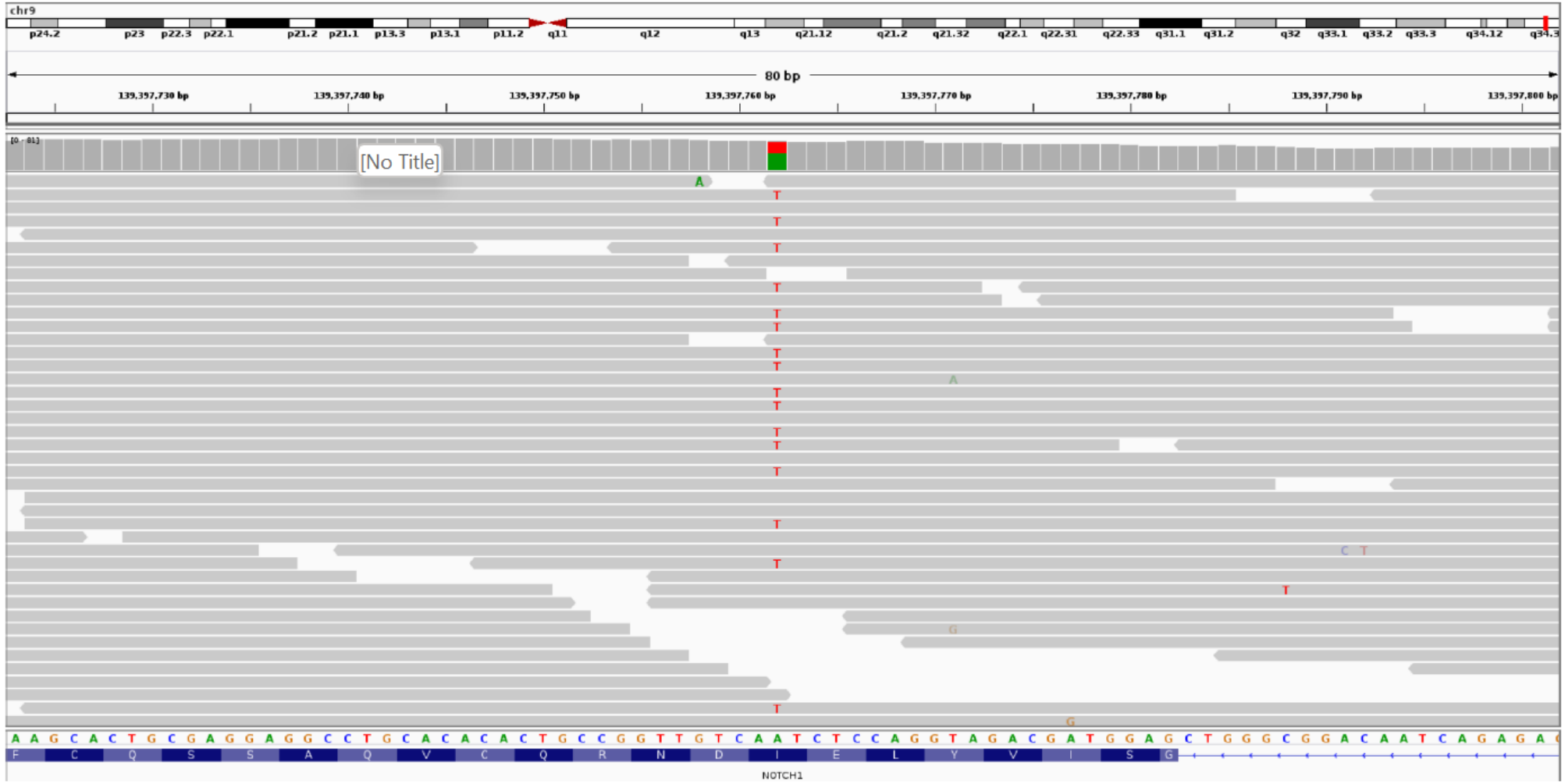
[No Title]

Genotype (GT):

0/0: hom. reference
0/1: heterozygous
1/1: hom. alternative

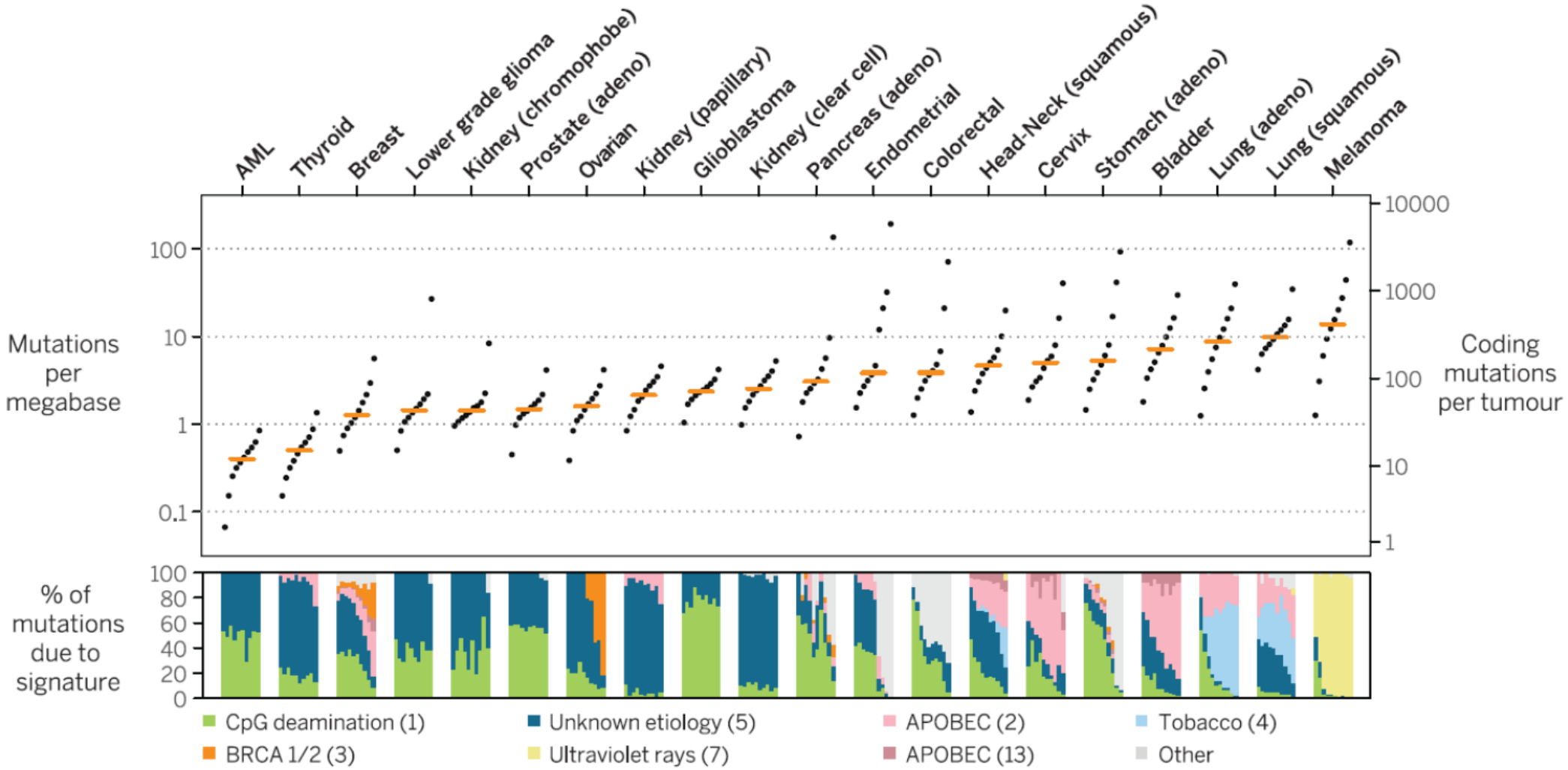
VCF/BCF file

Alignment and variant viewers



Tobias Rausch

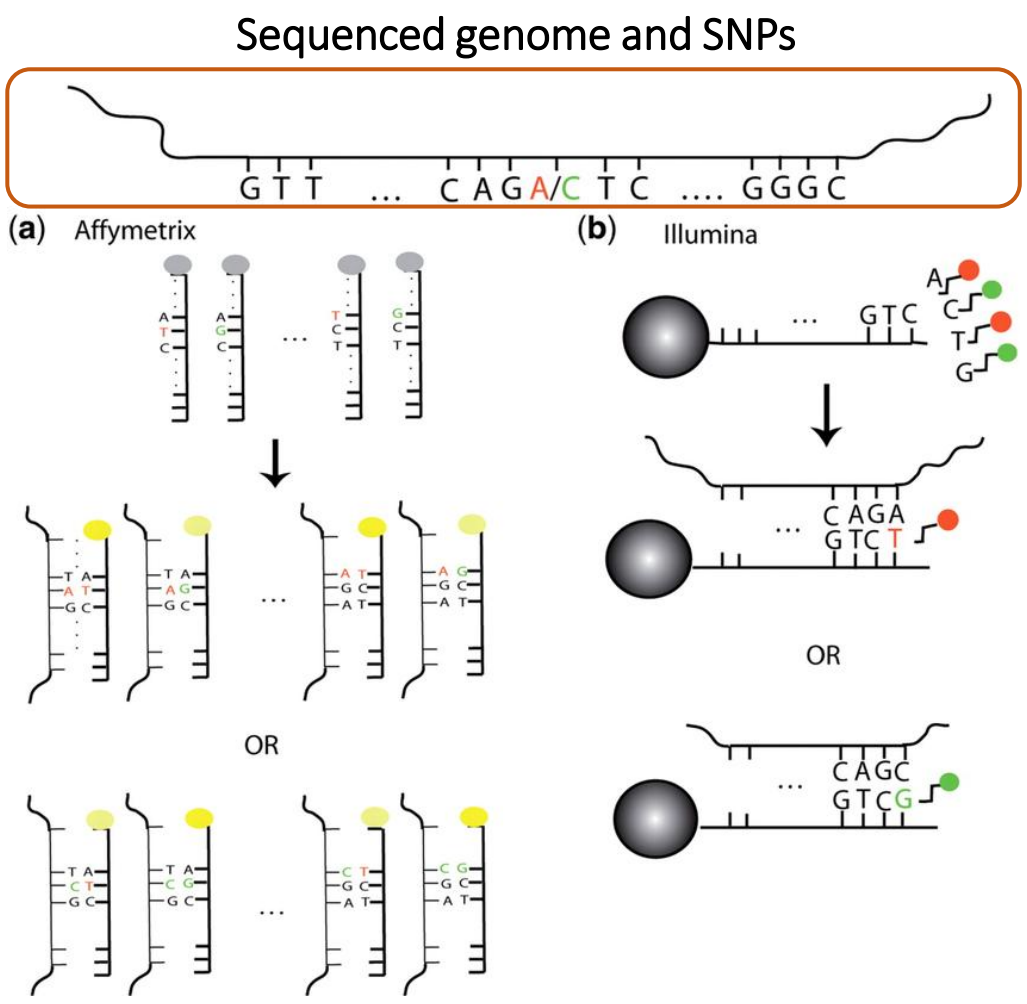
Mutational Signatures across Human Cancer Types



Source: Somatic mutation in cancer and normal cells. Martincorena I and Campbell PJ, Science. 2015 Sep 25;349(6255):1483-9.

Tobias Rausch

From genome sequence to SNP array



Overview of SNP array technology. At the top is the fragment of DNA harboring an A/C SNP to be interrogated by the probes shown.

(a) In the Affymetrix assay, there are 25-mer probes for both alleles, and the location of the SNP locus varies from probe to probe. The DNA binds to both probes regardless of the allele it carries, but it does so more efficiently when it is complementary to all 25 bases (bright yellow) rather than mismatching the SNP site (dimmer yellow). This impeded binding manifests itself in a dimmer signal.

(b) Attached to each Illumina bead is a 50-mer sequence complementary to the sequence adjacent to the SNP site. The single-base extension (T or G) that is complementary to the allele carried by the DNA (A or C, respectively) then binds and results in the appropriately-colored signal (red or green, respectively). For both platforms, the computational algorithms convert the raw signals into inferences regarding the presence or absence of each of the two alleles.

DOI:10.1093/nar/gkp552

Content

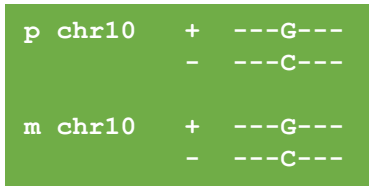
- Terms used in GWAS
- What is GWAS?
- How to do GWAS with Case-Control Association Testing?
- What is QTL?
- GWAS database
- Why do we do GWAS?

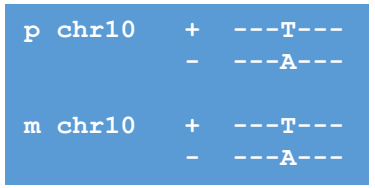
Terms

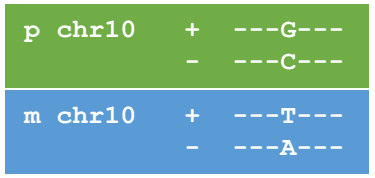
- **Gene**. The most obvious way how genetic variation can affect phenotypes is through variation in how **genes function**. Genes are segements of DNA that code for proteins (see yourgenome.org) and variation in the physical structure of the protein or in the time and place where the protein is made can have phenotypic consequences. Therefore, we are very interested in how **genetic variation** can affect the function of genes, and a lot of this is still unknown. **Protein coding genes** cover less than 2% of the whole human genome, but the remaining 98% affects the **regulation of genes (intergenic/ enhancer/ promoter)** in many ways.
- **Locus** (pl. loci). A continuous region of the genome is called a locus (plural loci). It can be of any size (e.g. a single nucleotide site of length 1 bp or a region of 10 milion base pairs, 10 Mbp). GWAS loci are regions that include a clear statistical association with the phenotype of interest.

Terms

- For example, my **paternal** chromosome can have a base **A** and **maternal** chromosome can have a base **G** (on the +strand of the DNA) at that position. Such a **one-nucleotide variation** is called a **single-nucleotide variant** (SNV) and the two versions are called **alleles**. So in the example case, I would be carrying both an allele A and an allele G at that SNV, whereas you might be carrying two copies of allele A at the same SNV.
- My genotype would be AG and yours AA. An individual having different **alleles** on his/her two genomes is **heterozygous** at that locus, and an individual having two copies of the same allele is **homozygous** at that locus.
- Since each individual has two copies of the genome, there are individuals with three possible genetic types (called **genotypes**) at this SNP. We denote each genotype by the number of copies of allele 1 that the genotype contains (i.e. genotype can be 0, 1 or 2).







Genotype vs phenotype

Genotype	Refers broadly to the genetic makeup of an organism—its complete set of genes. Sometimes used in a narrower definition, (as in this article), **genotype** refers to the specific alleles found on each chromosome.
Phenotype	The physical/observed traits determined or "expressed" by a given genotype; for example, the purple or white petals of a pea flower seen in Figure 3.

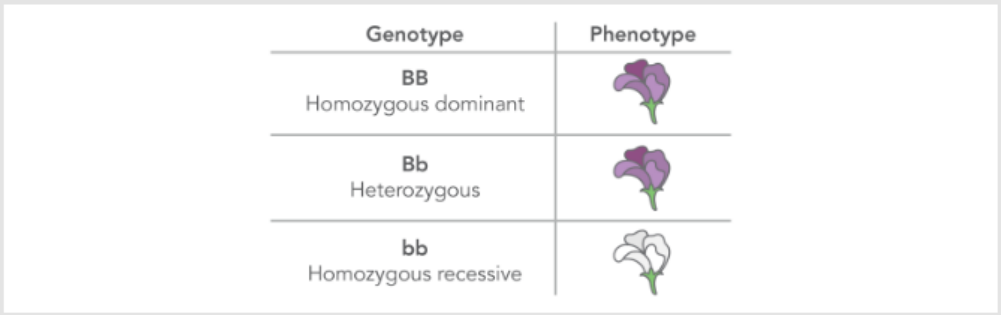


Figure 2. Different genotypes give rise to distinct phenotypes.

https://sg.idtdna.com/pages/education/decoded/article/genotyping-terms-to-know

Allele Frequency

SNV: 16-79619835-G-A(GRCh37)

Copy variant ID

Dataset: gnomAD v2.1.1

	Exomes	Genomes	Total
Filters	Pass	Pass	
Allele Count	25725	6378	32103
Allele Number	107734 *	31314	139048 *
Allele Frequency	0.2388	0.2037	0.2309
Popmax Filtering AF (95% confidence)	0.4018	0.3917	
Number of homozygotes	3402	709	4111
Mean depth of coverage	11.4	30.4	

External Resources

- dbSNP (rs60533944)
- UCSC
- ClinGen Allele Registry (CA8183922)

Feedback

Report an issue with this variant

Warning This variant is covered in fewer than 50% of individuals in gnomAD v2.1.1 exomes. Allele frequency estimates may not be reliable.

Population Frequencies

Population	Allele Count	Allele Number	Number of Homozygotes	Allele Frequency
East Asian	4186	10112	844	0.4140
Latino/Admixed American	6109	18448	1027	0.3311
Ashkenazi Jewish	1730	7464	204	0.2318
Other	1028	4442	122	0.2314
South Asian	4050	17832	498	0.2271
European (non-Finnish)	11777	58018	1185	0.2030
European (Finnish)	1461	8408	123	0.1738
African/African American	1762	14324	108	0.1230
XX	14510	62098	1909	0.2337
XY	17593	76950	2202	0.2286
Total	**32103**	**139048**	**4111**	**0.2309**

* Allele frequencies for some sub-continental populations were not computed for genome samples.

Include: Exomes Genomes

https://gnomad.broadinstitute.org/variant/16-79619835-G-A

Minor Allele Frequency (MAF) in Single Nucleotide Variants

- **Minor allele frequency** (**MAF**) is the frequency at which the *second most common* allele occurs in a given population.
- Single nucleotide polymorphisms (SNPs) with a minor allele frequency of 0.05 (5%) or greater were targeted by the HapMap project.[2]
- MAF is widely used in population genetics studies because it provides information to differentiate between common and rare variants in the population.
- As an example, a 2015 study sequenced the whole genomes of 2,120 Sardinian individuals. The authors classified the variants found in the study in three classes according to their MAF. It was observed that rare variants (MAF < 0.05) appeared more frequently in coding regions than common variants (MAF > 0.05) in this population.[3]

Minor Allele Frequency (MAF) in Single Nucleotide Variants

1. Introduce the reference of a SNP of interest, as an example: rs429358, in a database (dbSNP or other).

$$\text{MAF} = \frac{\text{Alleles positive for the variant}}{\text{Total alleles screened}}$$

2. Find MAF/MinorAlleleCount link.

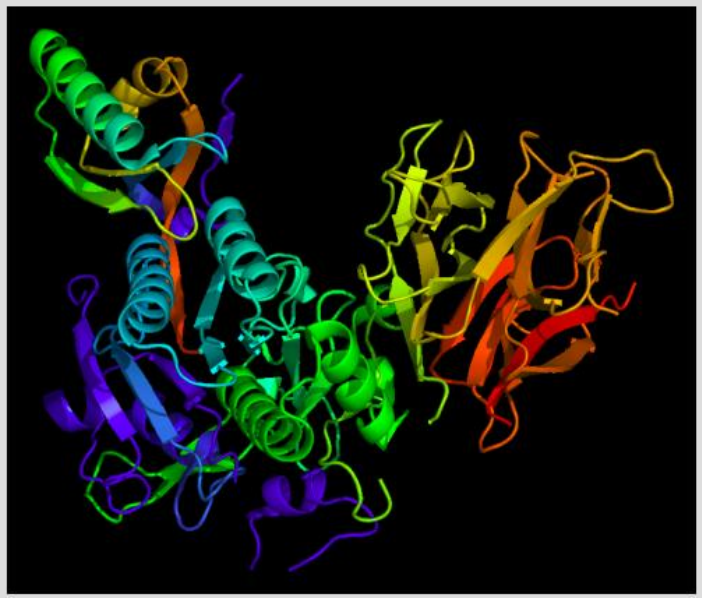
MAF/MinorAlleleCount: C=0.1506/754 where **C** is the minor allele for that particular locus; **0.1506** is the frequency of the C allele (MAF), i.e. 15% within the 1000 Genomes database; and **754** is the number of times this SNP has been observed in the population of the study.

https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/minor-allele-frequency

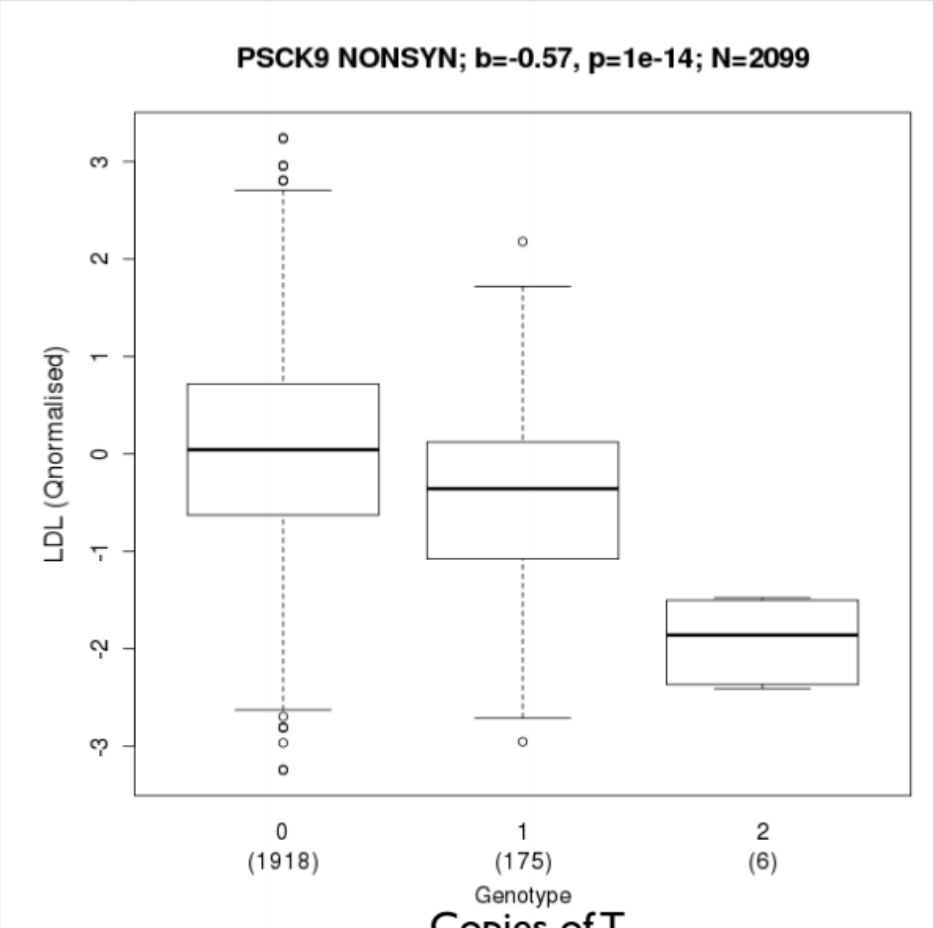
WHY STUDY GENOME?
A STORY ABOUT *PCSK9*

Gene *PCSK9* on chr 1

Codes for protein
692 amino acids long



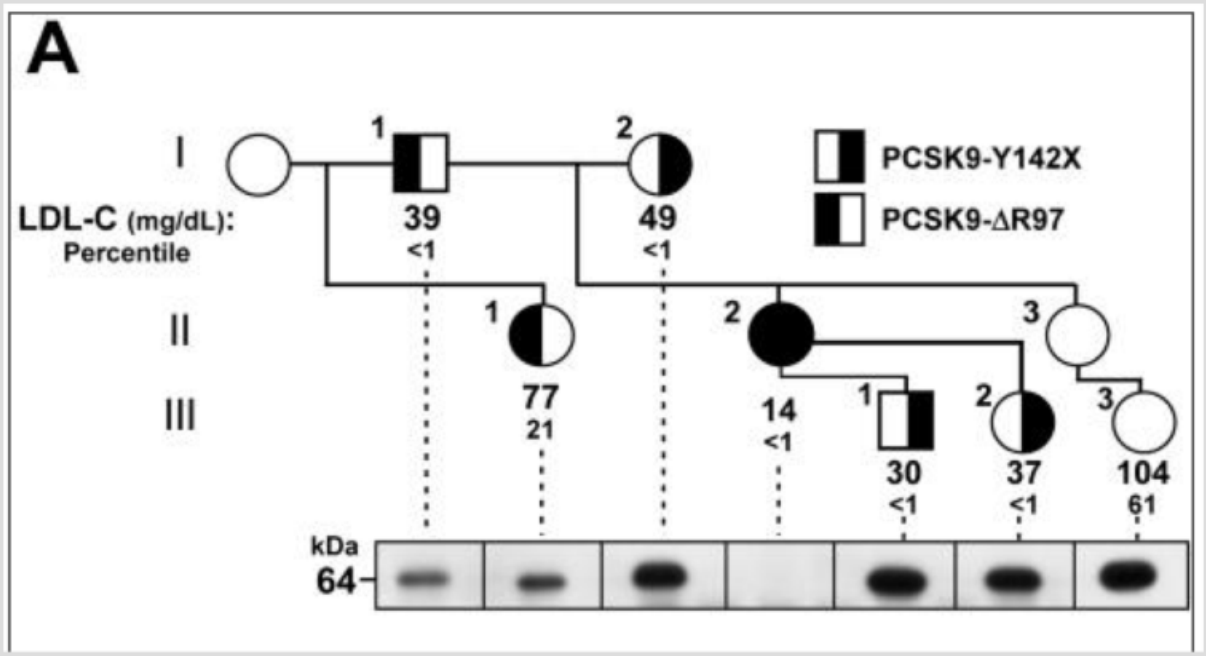
GENETIC VARIANT "RS11591147" IN *PCSK9*



- Carriers of T variant have lower levels of LDL cholesterol than carriers of G variant
- LDL is a strong risk factor for heart disease

2099 Finnish individuals

A HUMAN KNOCK-OUT OF *PCSK9* (2006)



Zhao et al. AJHG 2006

- Individual II.2 has zero working copies of *PCSK9* gene
  - no circulating *PCSK9* and an LDL-C of only 14 mg/dL
  - apparently healthy, fertile, normotensive, college-educated woman with normal liver and renal function tests who works as an aerobics instructor
  - Why is this very interesting observation?
    - Inhibiting *PCSK9* might be a **safe** way to reduce LDL

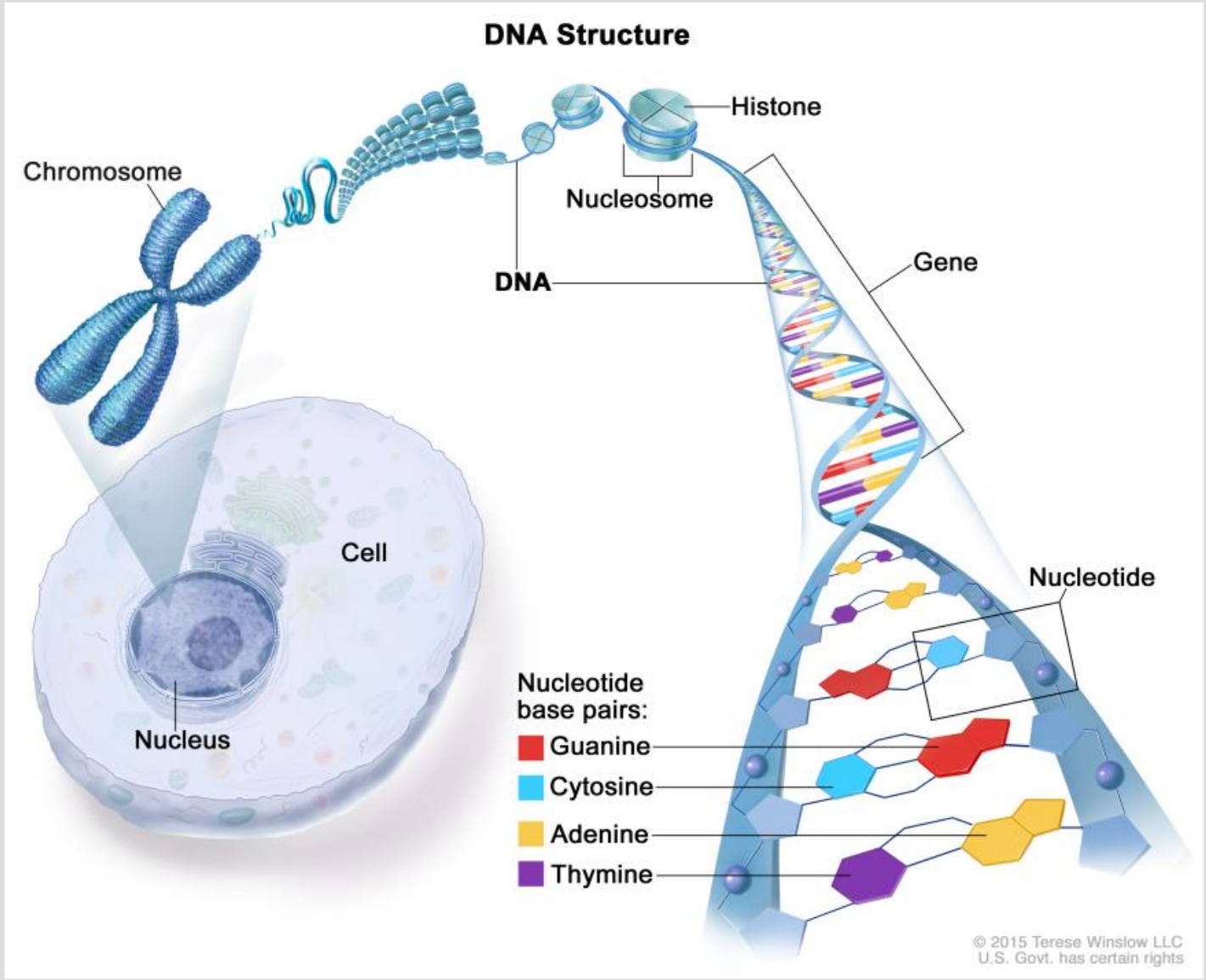
HUMAN GENOME

- Sequence of 3×10^9 letters from alphabet { A, C, G, T }

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

You have two genomes: maternal and paternal.

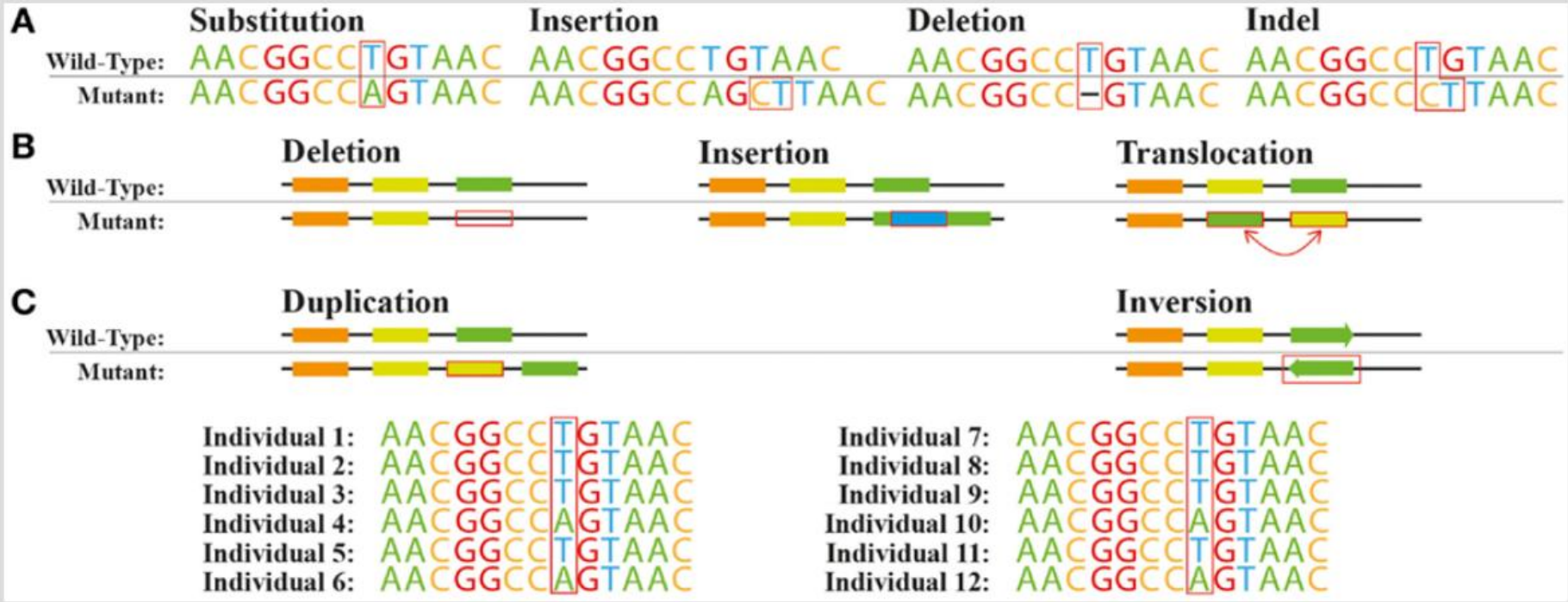
Your genomes are physically divided into 22 pairs of autosomal chromosomes and
1 pair of sex chromosomes (males XY, females: XX)



Most DNA is found inside the nucleus of a cell, where it forms the chromosomes. Chromosomes have proteins called histones that bind to DNA. DNA has two strands that twist into the shape of a spiral ladder called a helix. DNA is made up of four building blocks called nucleotides: adenine (A), thymine (T), guanine (G), and cytosine (C). The nucleotides attach to each other (A with T, and G with C) to form chemical bonds called base pairs, which connect the two DNA strands. Genes are short pieces of DNA that carry information for creating proteins.

https://siteman.wustl.edu/glossary/cdr0000046470/

TYPES OF VARIATION



Nucleotide level variation

Structural variation

SNP A/T in population

Cardoso et al. 2015
Front. Bioeng. Biotechnol., 16 February 2015 | https://doi.org/10.3389/fbioe.2015.00013

SINGLE NUCLEOTIDE POLYMORPHISM (SNP)

On average, 1:300 positions in genome has common (MAF>1%) variation in population; these are called "SNPs"

Genomes in population

... G C G T T ... 96%

... G C T T T ... 4%

Only forward strand of genomes is shown here

This is a SNP, with alleles: G / T, minor allele frequency (MAF) = 4%

Genotypes at this SNP in population

0: GG ~ 92.1%
1: GT ~ 7.7 %
2: TT ~ 0.2 %

```
p chr10    +  ---G---
              -  ---C---

m chr10    +  ---G---
              -  ---C---
```

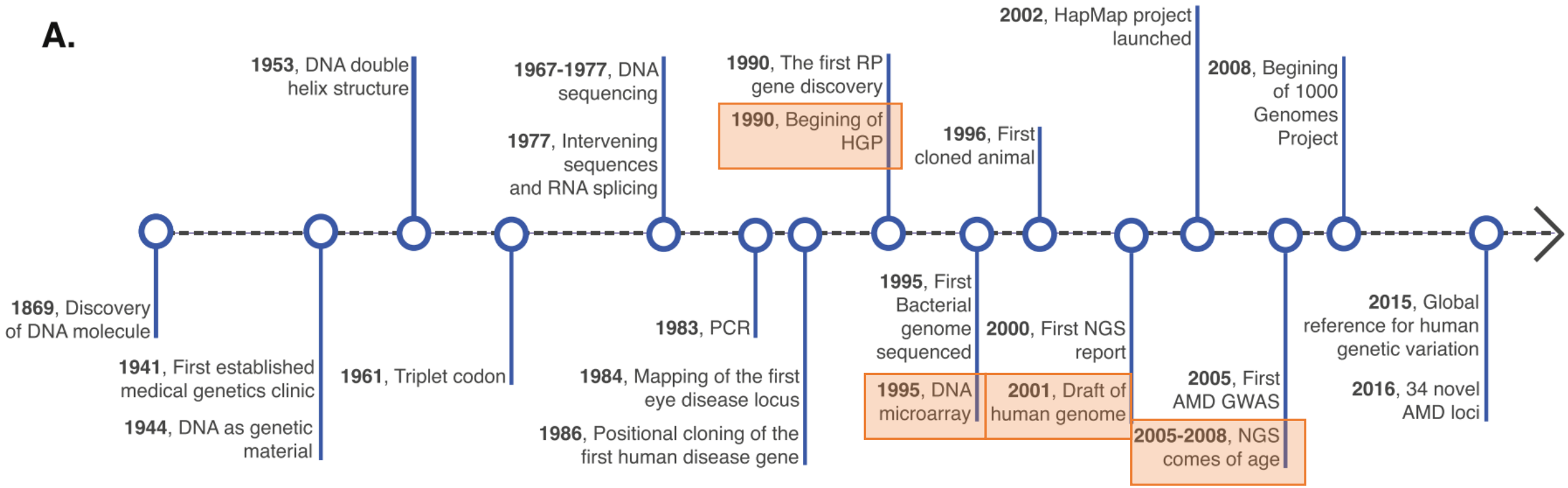
```
p chr10    +  ---T---
              -  ---A---

m chr10    +  ---T---
              -  ---A---
```

```
p chr10    +  ---G---
              -  ---C---

m chr10    +  ---T---
              -  ---A---
```

Timeline of human genetics and genomic technologies

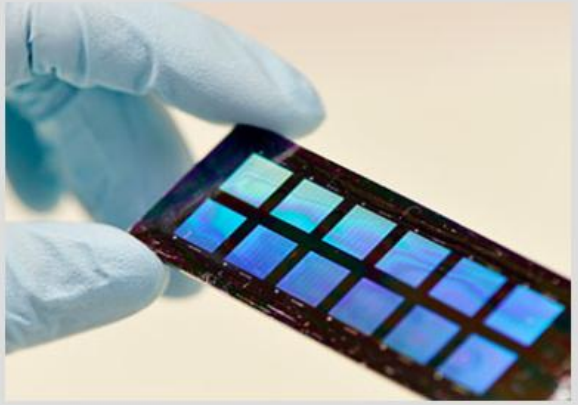


Age-related macular degeneration (AMD)
Human Genome Project (HGP)

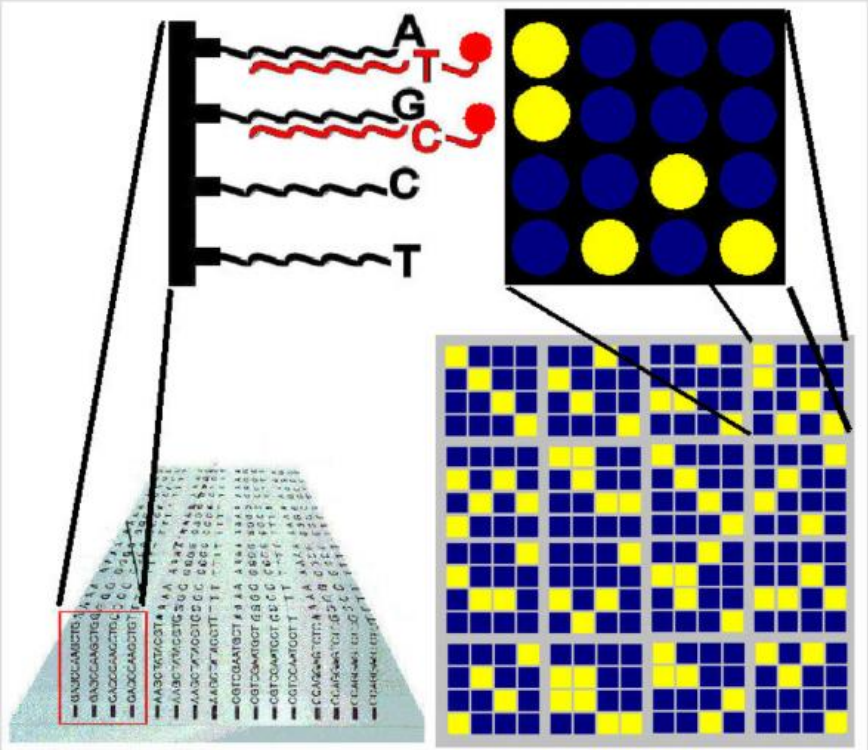
http://dx.doi.org/10.1016/j.preteyeres.2016.06.001

Profiling SNPs

- Human SNP array can measure 10^6 SNPs
- Cost per individual ~30 euros



This array can genotype 12 individuals at 10^6 SNPs

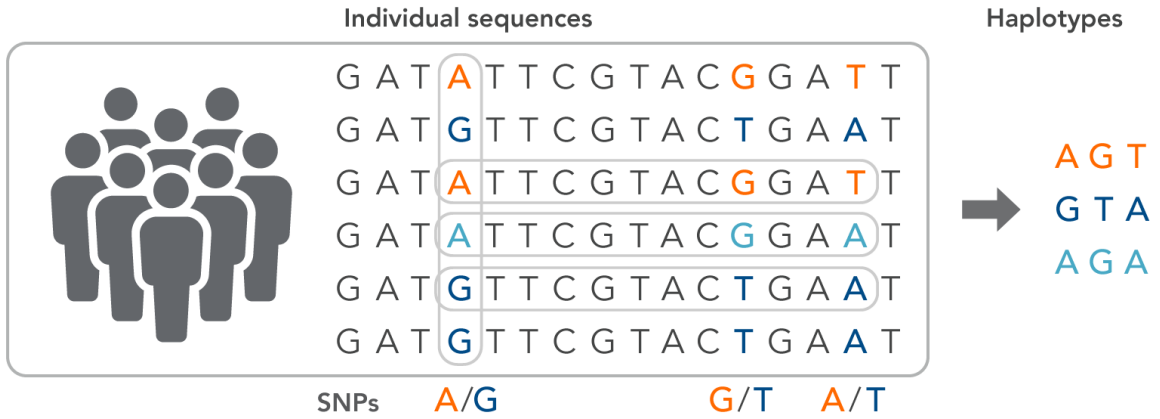


Steven M. Carr
www.mun.ca/biology/scarr/DNA_Chips.html

Whole genome sequencing (WGS), Whole Exome Sequencing (WES) and Target Sequencing with NGS

One individual

Multi individual



https://online.stat.psu.edu/stat555/node/106/

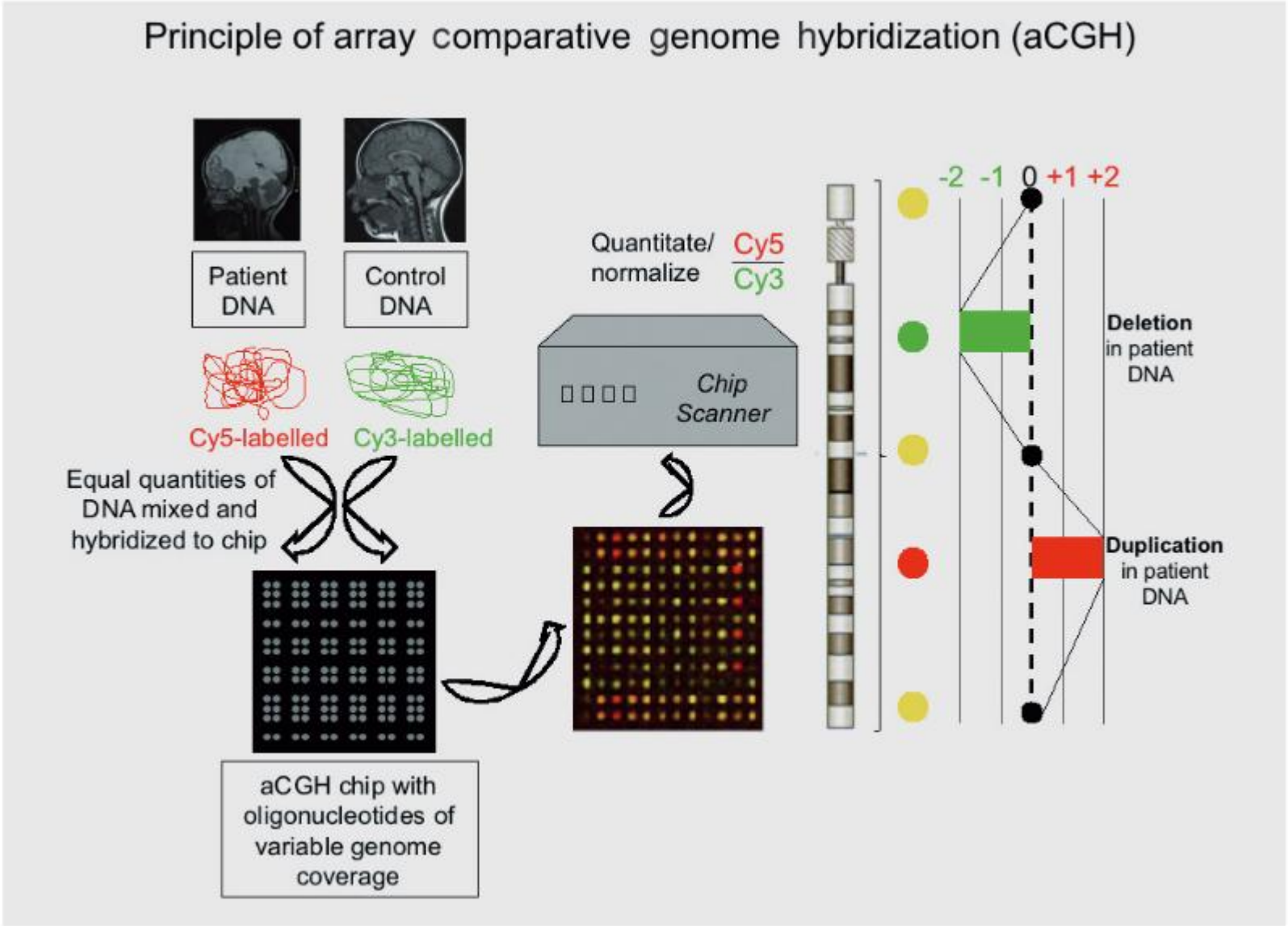
SNP Arrays

GPL6975	"KKI/Pevsnerlab Human 550K SNP"	"oligonucleotide beads"	"Homo sapiens"	555352
GPL21480	"[Axiom_GW_Hu-CHB_SNP] Affymetrix Axiom Genome-Wide CHB 1 Array Plate"	"in situ oligonucleotide"	"Homo sapiens"	657178
GPL22678	"HumanOmniExpress-24 v1.1 BeadChip [SNP_ID version]"	"oligonucleotide beads"	"Homo sapiens"	713014
GPL31092	"HumanOmniExpress-24 v1.2 BeadChip [SNP_ID version]"	"oligonucleotide beads"	"Homo sapiens"	713599
GPL21168	"HumanOmniExpress-24 v1.0 BeadChip [SNP_ID version]"	"oligonucleotide beads"	"Homo sapiens"	716503
GPL19718	"IIllumina Infinium human CytoSNP-850K BeadChip"	"oligonucleotide beads"	"Homo sapiens"	851274
GPL6804	"[GenomeWideSNP_5] Affymetrix Genome-Wide Human SNP 5.0 Array"	"in situ oligonucleotide"	"Homo sapiens"	949274
GPL26040	"Illumina Human1Mv1 DNA Analysis BeadChip (Human1Mv1_C) SNP ID"	"oligonucleotide beads"	"Homo sapiens"	1072820
GPL16209	"Affymetrix Human Exon 1.0 ST Array [CDF: HuEx_1_0_st_v2_core_na31_hg19_HB20110919_noSNPs]"	"in situ oligonucleotide"	"Homo sapiens"	1432143
GPL6801	"[GenomeWideSNP_6] Affymetrix Genome-Wide Human SNP 6.0 Array"	"in situ oligonucleotide"	"Homo sapiens"	1880794
GPL23136	"Illumina HumanOmni 2.5M 4v1-H SNP array"	"oligonucleotide beads"	"Homo sapiens"	2443177
GPL23135	"Illumina HumanOmni 2.5M 4v1-D SNP array"	"oligonucleotide beads"	"Homo sapiens"	2443179

GEO Browser - GEO - NCBI (nih.gov)

Comparative genome hybridization (CGH)

CGH is a molecular-cytogenetic method for the analysis of copy number changes (gains or losses) in the DNA content of a given individual's DNA.



[A glossary of relevant genetic terms (tandfonline.com)]

CNV Arrays

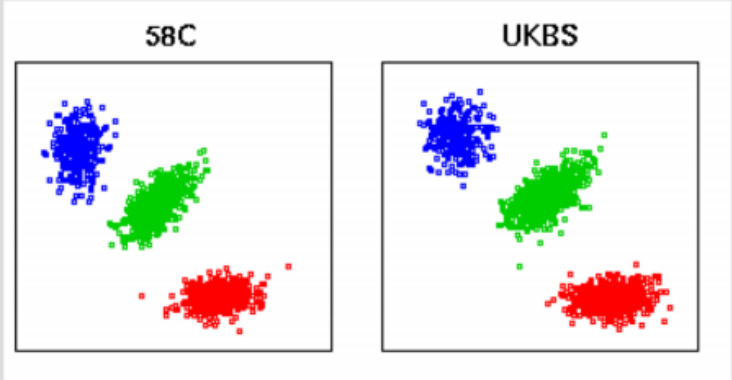
GPL20879	"Agilent-073748 PPCD1 CNV 17.3-34.0 (Probe Name version)"	"in situ oligonucleotide"	"Homo sapiens"	57806
GPL20878	"Agilent-073748 PPCD1 CNV 17.3-34.0 (Feature Number version)"	"in situ oligonucleotide"	"Homo sapiens"	62976
GPL8890	"Illumina HumanCNV100W v1.0 BeadChip"	"oligonucleotide beads"	"Homo sapiens"	95484
GPL9681	"Eichler-NimbleGen Human 135K Custom CNV array"	"in situ oligonucleotide"	"Homo sapiens"	135000
GPL22557	"Agilent-044360 mCNV1_design1"	"in situ oligonucleotide"	"Homo sapiens"	180880
GPL22558	"Agilent-044459 mCNV2_design1"	"in situ oligonucleotide"	"Homo sapiens"	180880
GPL21558	"[OncoScan_CNV] Affymetrix OncoScan CNV FFPE Assay"	"in situ oligonucleotide"	"Homo sapiens"	239038
GPL20252	"Agilent-018897 Unrestricted Human Genome CNV Microarray 244K (1 of 2)"	"in situ oligonucleotide"	"Homo sapiens"	243504
GPL20253	"Agilent-018898 Unrestricted Human Genome CNV Microarray 244K (2 of 2)"	"in situ oligonucleotide"	"Homo sapiens"	243504
GPL6944	"Agilent Technologies Human Genome CNV microarray set G4423B (AMADID 018897)"	"in situ oligonucleotide"	"Homo sapiens"	243504
GPL6945	"Agilent Technologies Human Genome CNV microarray set G4423B (AMADID 018898)"	"in situ oligonucleotide"	"Homo sapiens"	243504
GPL6986	"Illumina HumanCNV370-Duov1 DNA Analysis BeadChip (HumanCNV370v1)"	"oligonucleotide beads"	"Homo sapiens"	370404
GPL6985	"Illumina HumanCNV370-QuadV3 DNA Analysis BeadChip (HumanCNV370-QuadV3_C)"	"oligonucleotide beads"	"Homo sapiens"	373397
GPL20249	"Agilent-021365 SurePrint G3 Human CNV Microarray 2x400K (Probe Name version)"	"in situ oligonucleotide"	"Homo sapiens"	415622
GPL10154	"Agilent-021365 SurePrint G3 Human CNV Microarray 2x400K (Feature Number version)"	"in situ oligonucleotide"	"Homo sapiens"	420288
GPL19515	"Nimblegen Homo sapiens HG18 2.1M CNV Array [090518_HG18_CNV_v1_HX1]"	"in situ oligonucleotide"	"Homo sapiens"	2159817

GEO Browser - GEO - NCBI (nih.gov)

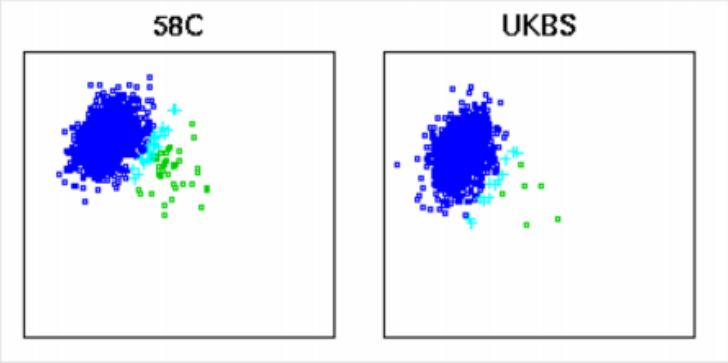
Experiment design

	Microarrays	Whole genome sequencing
Variant types	SNPs, CNVs	~All
Loci probed	500k - 5m	~All (3b)
% variation captured	60%-85% of MAF>1% (r^2>0.8)	~90%+ of genome
Cost per sample (approx, 2014)	$100-$500	$1000-5000
Data file size	<100Mb	100Gb
Data storage cost (Amazon cloud 2014, 1000 samples)	$50 / year	$50,000 / year

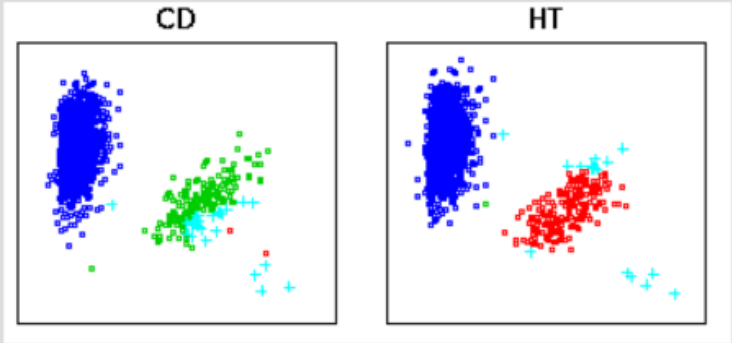
GENOTYPE CALLING FROM SNP ARRAY DATA



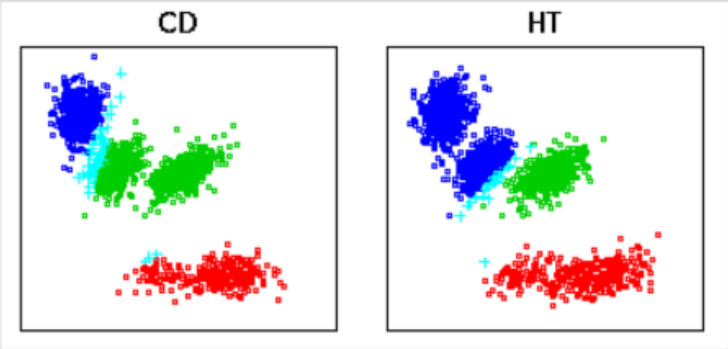
GOOD calling!



ERROR, rare variant has less than 3 clusters



ERROR, clustering algorithm performs differently in two cohorts



ERROR, structural variant has more than three genotypes

From D.Phil thesis of Damjan Vukcevic, Oxford, 2009

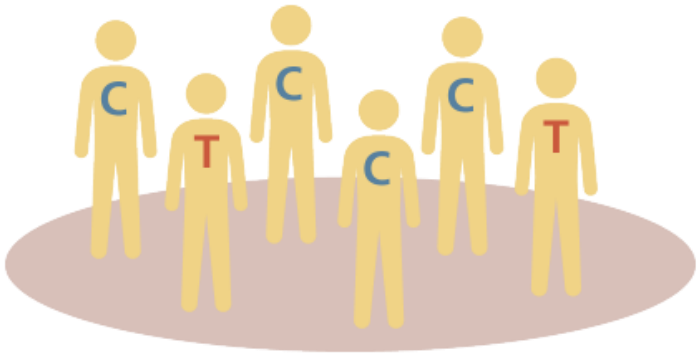
The calling algorithm tries to find the three genotype clusters.

Figures shows how an algorithm has clustered individuals into three groups

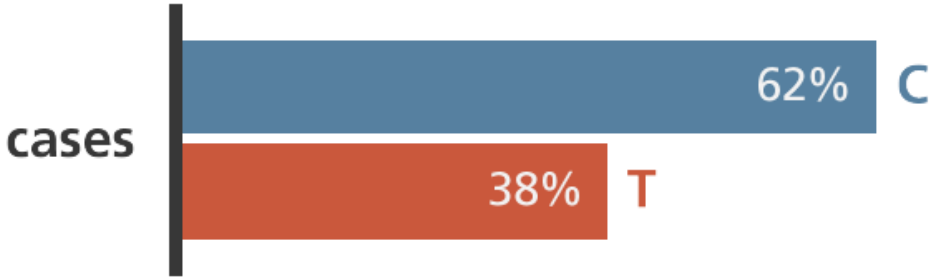
Light blue means algorithm has made no call.

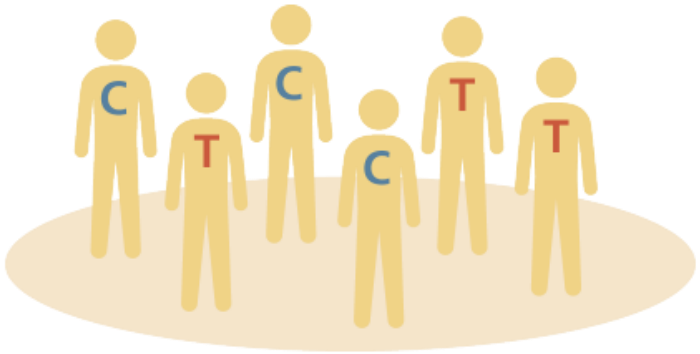
Bottom line errors would likely fail HWE test.

Example of heart disease at Chr1 Pos 12345



cases (n=1,000)
people with heart disease

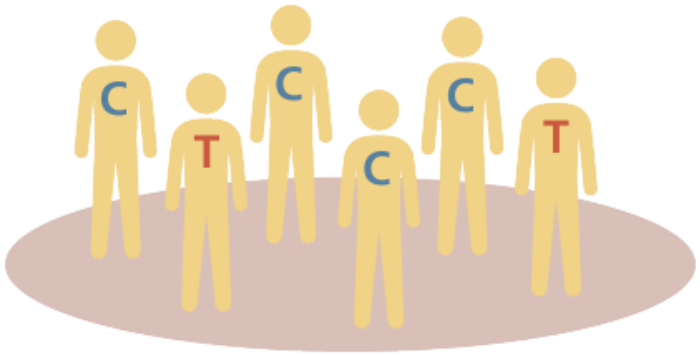




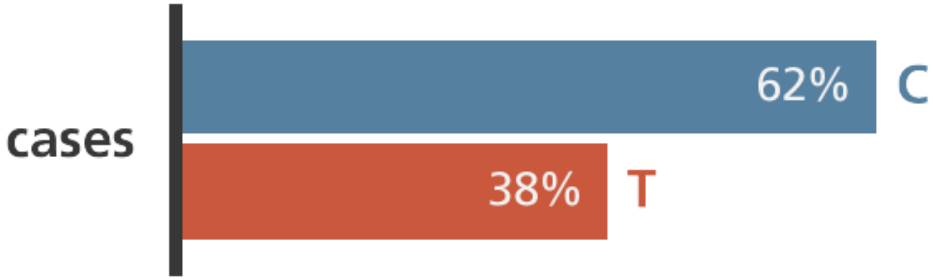
controls (n=1,000)
people without heart disease

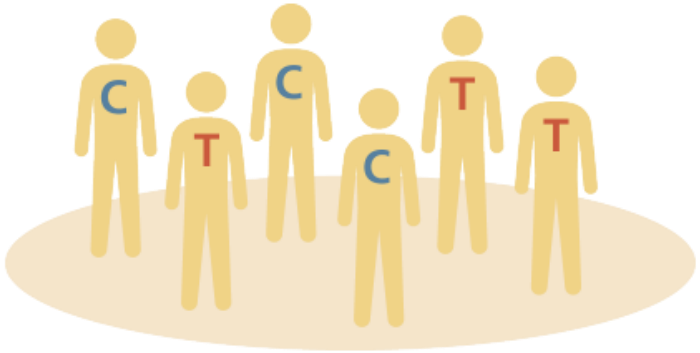
https://www.yourgenome.org/stories/genome-wide-association-studies

Example of heart disease at Chr1 Pos 543256

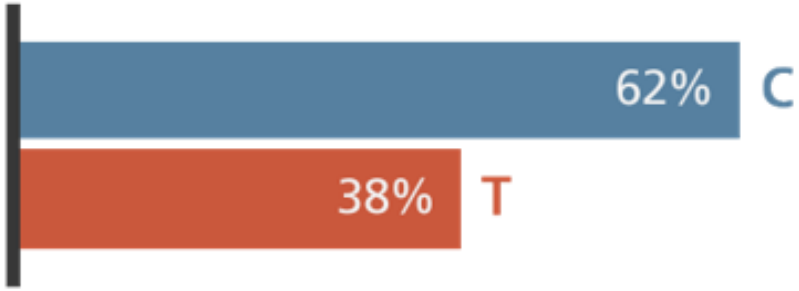


cases (n=1,000)
people with heart disease





controls (n=1,000)
people without heart disease



https://www.yourgenome.org/stories/genome-wide-association-studies

Review statistics

Chr1-10	C	T	Total
Case	620	380	1000
Control	490	510	1000
Total	1110	890	2000

```
>dat <- data.frame(
  "C" = c(620,490),
  "T" = c(380,510),
  row.names = c("Case", "Control"),
  stringsAsFactors = FALSE
)
>colnames(dat) <- c("C", "T")
>chisq.test(dat)

        Pearson's Chi-squared test with Yates' continuity correction

data:  dat
X-squared = 33.69, df = 1, p-value = 6.464e-09

> fisher.test(dat)

        Fisher's Exact Test for Count Data

data:  dat
p-value = 6.15e-09
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 1.415601 2.037173
sample estimates:
odds ratio
  1.697752
```

Chr1-20	A	G	Total
Case	620	380	1000
Control	610	390	1000
Total	1230	770	2000

```
> dat <- data.frame(
  "A" = c(620,610),
  "G" = c(380,390),
  row.names = c("Case", "Control"),
  stringsAsFactors = FALSE
)
colnames(dat) <- c("C", "T")
> chisq.test(dat)

        Pearson's Chi-squared test with Yates' continuity correction

data:  dat
X-squared = 0.17105, df = 1, p-value = 0.6792

> fisher.test(dat)

        Fisher's Exact Test for Count Data

data:  dat
p-value = 0.6792
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.8676031 1.2542032
sample estimates:
odds ratio
  1.043142
```

Review statistics

Chr1-10	C	T	Total
Case	620	380	1000
Control	490	510	1000
Total	1110	890	2000

Chr1-20	A	G	Total
Case	620	380	1000
Control	610	390	1000
Total	1230	770	2000

Hypotheses

The hypotheses of the Fisher's exact test are the same than for the Chi-square test, that is:

- H_0 : the variables are independent, there is **no** relationship between the two categorical variables. Knowing the value of one variable does not help to predict the value of the other variable
- H_1 : the variables are dependent, there is a relationship between the two categorical variables. Knowing the value of one variable helps to predict the value of the other variable

Since pvalue < 0.05
Accept H1: There is a relation between disease case/control with allelic C/T at Chr1-10

Since pvalue > 0.05
Reject H1: Disease case/control and allelic C/T are independent at Chr1-20

https://statsandr.com/blog/fisher-s-exact-test-in-r-independence-test-for-a-small-sample/

EXAMPLE GWAS

- Let's next look at two examples GWAS
- Body-mass index GWAS by Locke et al. (Nature 2015) as an example of a quantitative trait analysis
- Migraine GWAS by Gormley et al. (Nature Genetics 2016) as an example of case-control analysis.

GWAS ON MIGRAINE (GORMLEY ET AL. 2016) (1/3)

- 60,000 cases and 315,000 controls from 22 studies
- 38 loci with convincing association
- Highlights vascular system

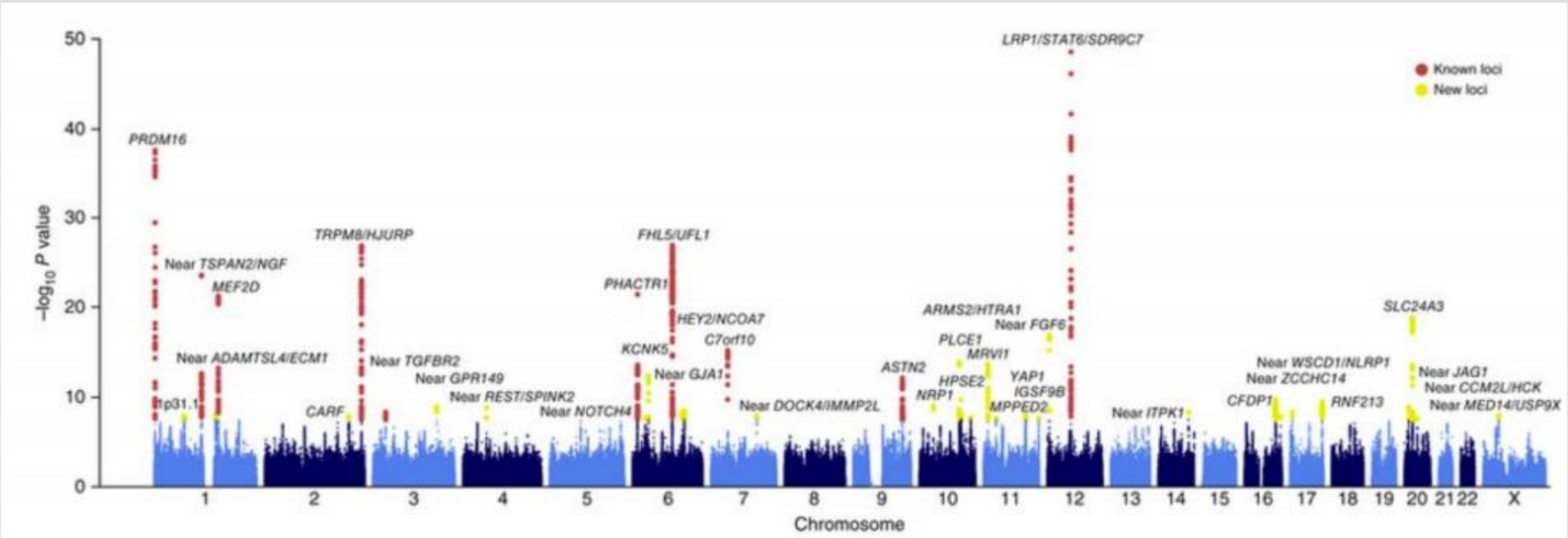


Fig. 1 of Gormley et al.
Manhattan plot of results.

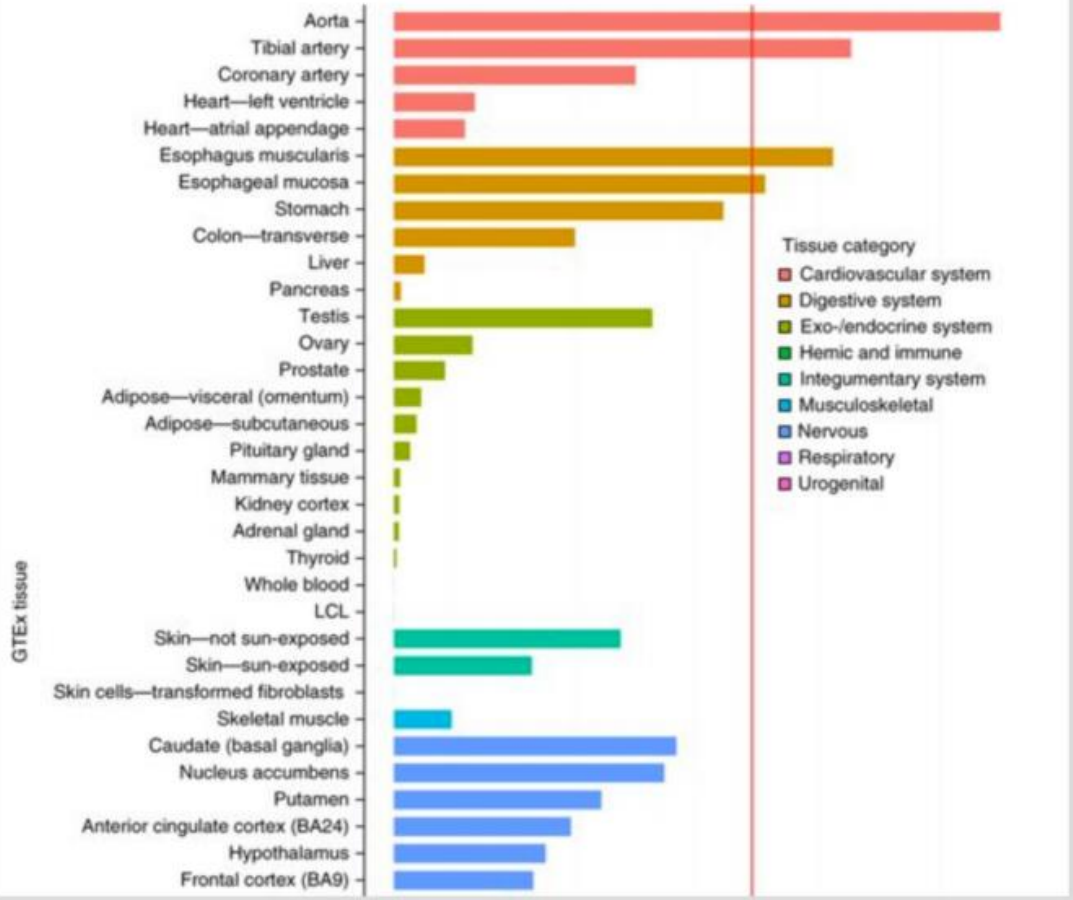
GWAS ON MIGRAINE (GORMLEY ET AL. 2016) (2/3)

Locus rank	Locus	Chr	Index SNP	Minor allele	MAF	All migraine		Secondary signal		Migraine without aura		Previous publication PMID
						OR (95% CI)	*P*	Index SNP	*P*	Index SNP	*P*	
1	*LRP1–STAT6–SDR9C7*	12	rs11172113	C	0.42	0.90 (0.89–0.91)	5.6×10^{-49}	rs11172055	1.3×10^{-9}	rs11172113	4.3×10^{-16}	21666692
2	*PRDM16*	1	rs10218452	G	0.22	1.11 (1.10–1.13)	5.3×10^{-38}	rs12135062	3.7×10^{-10}	-	-	21666692
3	*FHL5–UFL1*	6	rs67338227	T	0.23	1.09 (1.08–1.11)	2.0×10^{-27}	rs4839827	5.7×10^{-10}	rs7775721	1.1×10^{-12}	23793025
4	Near *TSPAN2–NGF*	1	rs2078371	C	0.12	1.11 (1.09–1.13)	4.1×10^{-24}	rs7544256	8.7×10^{-9}	rs2078371	7.4×10^{-9}	23793025
5	*TRPM8–HJURP*	2	rs10166942	C	0.20	0.94 (0.89–0.99)	1.0×10^{-23}	rs566529	2.5×10^{-9}	rs6724624	1.1×10^{-9}	21666692
6	*PHACTR1*	6	rs9349379	G	0.41	0.93 (0.92–0.95)	5.8×10^{-22}	-	-	rs9349379	2.1×10^{-9}	22683712
7	*MEF2D*	1	rs1925950	G	0.35	1.07 (1.06–1.09)	9.1×10^{-22}	-	-	-	-	22683712
8	*SLC24A3*	20	rs4814864	C	0.26	1.07 (1.06–1.09)	2.2×10^{-19}	-	-	-	-	-

Table 1 shows summary statistics for each locus.
Index SNP is the one with lowest P-value and 2nd signal is conditionally independent signal in the same locus.

GWAS ON MIGRAINE (GORMLEY ET AL. 2016) (3/3)

Figure 2: Expression enrichment of genes from the migraine loci in GTEx tissue samples.



We @FIMM, UH
are currently expanding
this migraine GWAS to over
100,000 cases and 750,000
controls and will report over
nearly 100 new migraine loci.

Case-Control Association Testing

	Cases	Controls	Total
Allele T	158	392	550
Allele C	20	86	106
Total	178	478	656

- Allelic Association Tests
  - Allele is treated as the sampling unit
  - Typically make an assumption of Hardy-Weinberg equilibrium (HWE). Alleles within an individual are conditionally independent, given the trait value.
- Genotypic Association Tests
  - Individual is the sampling unit
  - Does not assume HWE

	CC	CT	TT
Cases	6	8	75
Controls	10	66	163

Chi-square test for Allelic Association

- Below is a 2×2 contingency table for trait and allelic type

	Cases	Controls	Total
Allele 1	n_1^{ca}	n_1^{co}	n_1
Allele 2	n_2^{ca}	n_2^{co}	n_2
Total	$2N_{ca}$	$2N_{co}$	T

- n_1^{ca} is the number of type 1 alleles in the cases and $n_1^{ca} = 2 \times$ the number of homozygous $(1,1)$ cases $+$ the number of heterozygous (1,2) cases
- n_2^{co} is the number of type 2 alleles in the controls and $n_2^{co} = 2 \times$ the number of homozygous $(2,2)$ controls $+$ the number of heterozygous (1,2) controls

Chi-square test for Allelic Association

Chr1-10	C	T	Total
Case	620	380	1000
Control	490	510	1000
Total	1110	890	2000

```
>dat <- data.frame(
  "C" = c(620,490),
  "T" = c(380,510),
  row.names = c("Case", "Control"),
  stringsAsFactors = FALSE
)
>colnames(dat) <- c("C", "T")
>chisq.test(dat)

        Pearson's Chi-squared test with Yates' continuity correction

data:  dat
X-squared = 33.69, df = 1, p-value = 6.464e-09

> fisher.test(dat)

        Fisher's Exact Test for Count Data

data:  dat
p-value = 6.15e-09
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 1.415601 2.037173
sample estimates:
odds ratio
  1.697752
```

Chr1-20	A	G	Total
Case	620	380	1000
Control	610	390	1000
Total	1230	770	2000

```
> dat <- data.frame(
  "A" = c(620,610),
  "G" = c(380,390),
  row.names = c("Case", "Control"),
  stringsAsFactors = FALSE
)
colnames(dat) <- c("C", "T")
> chisq.test(dat)

        Pearson's Chi-squared test with Yates' continuity correction

data:  dat
X-squared = 0.17105, df = 1, p-value = 0.6792

> fisher.test(dat)

        Fisher's Exact Test for Count Data

data:  dat
p-value = 0.6792
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.8676031 1.2542032
sample estimates:
odds ratio
  1.043142
```

Chi-square test for Allelic Association

Chr1-10	C	T	Total
Case	620	380	1000
Control	490	510	1000
Total	1110	890	2000

Chr1-20	A	G	Total
Case	620	380	1000
Control	610	390	1000
Total	1230	770	2000

Hypotheses

The hypotheses of the Fisher's exact test are the same than for the Chi-square test, that is:

- H_0 : the variables are independent, there is **no** relationship between the two categorical variables. Knowing the value of one variable does not help to predict the value of the other variable
- H_1 : the variables are dependent, there is a relationship between the two categorical variables. Knowing the value of one variable helps to predict the value of the other variable

Since p-value < 0.05
Accept H1: There is a relation between disease case/control with allelic C/T at Chr1-10

Since p-value > 0.05
Reject H1: Disease case/control and allelic C/T are independent at Chr1-20

https://statsandr.com/blog/fisher-s-exact-test-in-r-independence-test-for-a-small-sample/

The Armitage Trend Test for Genotypic Association

- The most common genotypic test for unrelated individuals is the Armitage trend test (Sasieni 1997)
- Consider a single marker with 2 allelic types (e.g., a SNP) labeled "1" and "2"
- Let $Y_i = 2$ if individual i is homozygous (1,1), 1 if the i is heterozygous, and 0 if i is homozygous (2,2)
- Let $X_i = 1$ if i is a case and 0 if i is a control.
- A simple linear regression model of

$$Y = \beta_0 + \beta_1 X + \epsilon$$

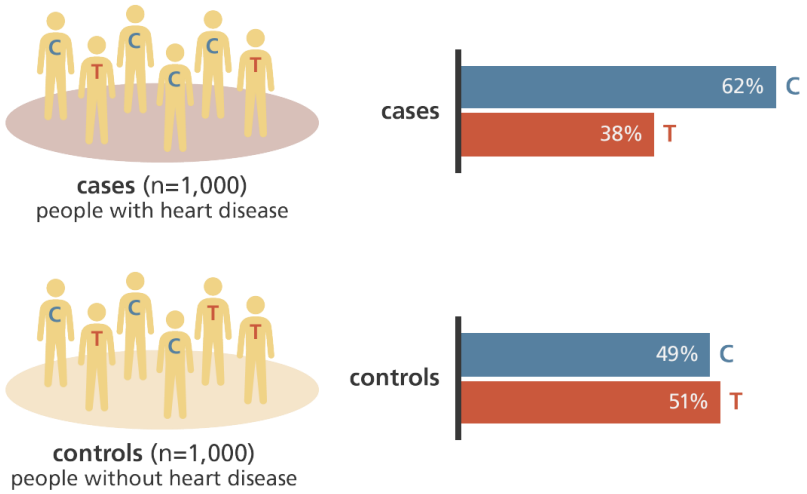
- $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$

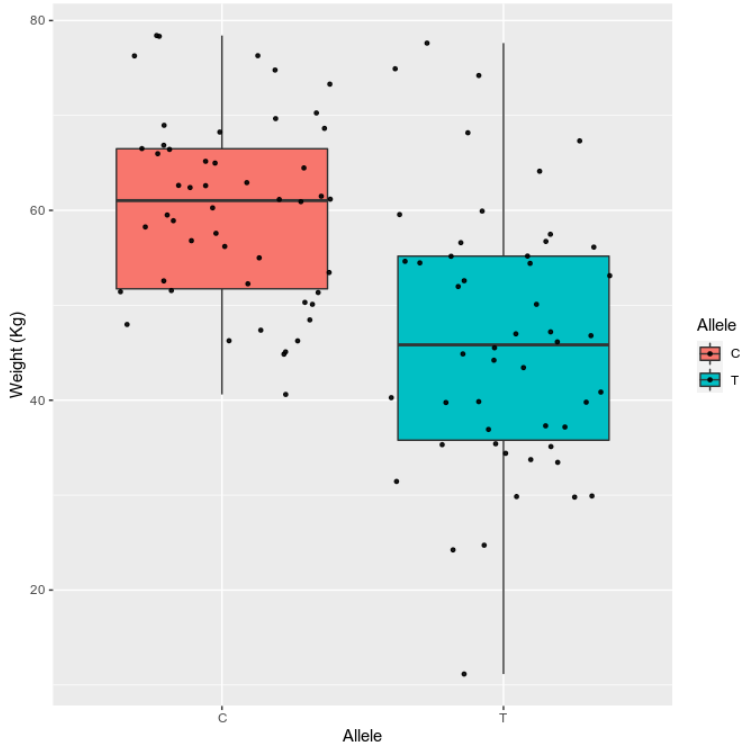
Leber Hereditary Optic Neuropathy (LHON) disease and genotypes for marker rs6767450:

	CC	CT	TT
Cases	6	8	75
Controls	10	66	163

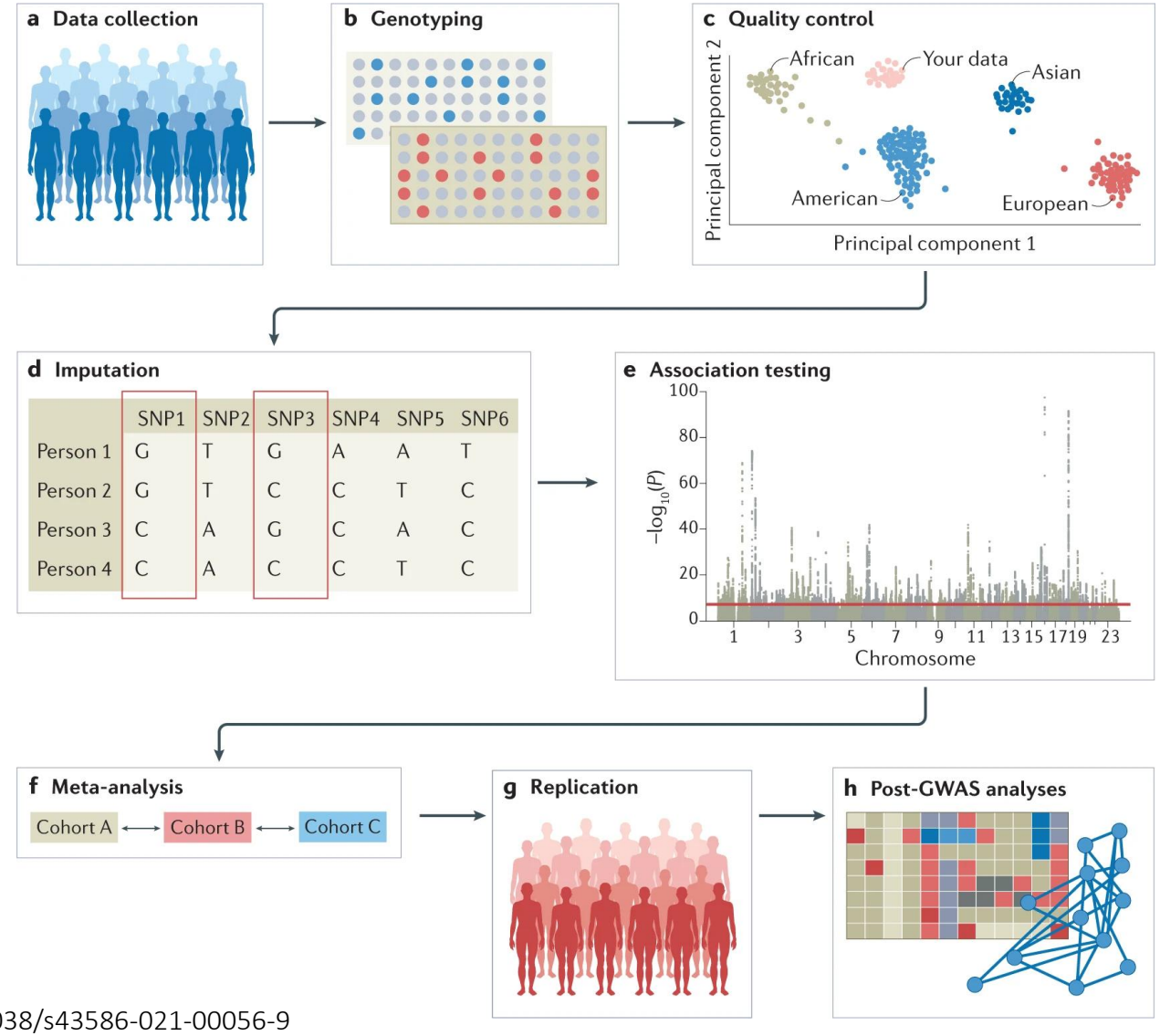
Quantitative trait loci (QTL)

- An overview of association testing methods when the trait of interest is binary or case/control (e.g. 1/0, affected/unaffected, dead/alive).
- Phenotypes of interest are often quantitative.
- The field of **quantitative genetics** is the study of the inheritance of continuously measured traits and their mechanisms.
- **Quantitative trait loci** (QTL) mapping involves identifying genetic loci that influence the phenotypic variation of a quantitative trait.





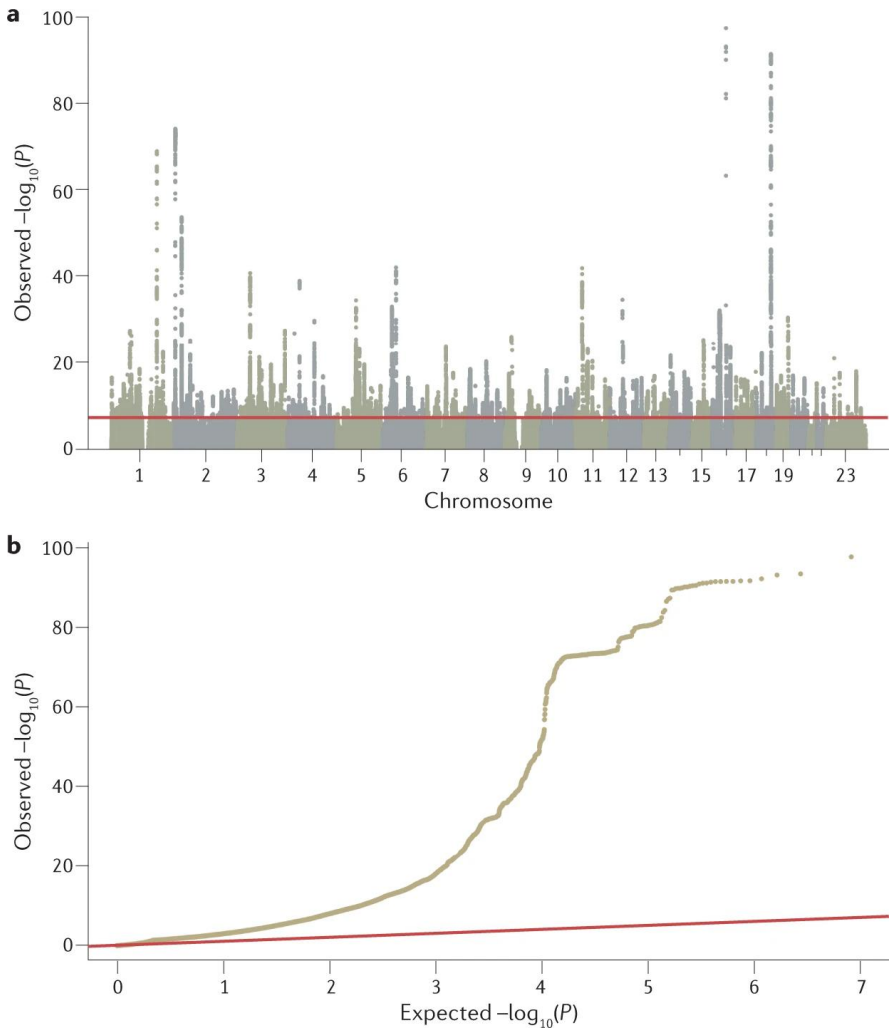
Overview of steps for conducting GWAS



a | Data can be collected from study cohorts or available genetic and phenotypic information can be used from biobanks or repositories. Confounders need to be carefully considered and recruitment strategies must not introduce biases such as collider bias.
b | Genotypic data can be collected using microarrays to capture common variants, or next-generation sequencing methods for whole-genome sequencing (WGS) or whole-exome sequencing (WES).
c | Quality control includes steps at the wet-laboratory stage, such as genotype calling and DNA switches, and dry-laboratory stages on called genotypes, such as deletion of bad single-nucleotide polymorphisms (SNPs) and individuals, detection of population strata in the sample and calculation of principle components. Figure depicts clustering of individuals according to genetic substrata.
d | Genotypic data can be phased, and untyped genotypes imputed using information from matched reference populations from repositories such as 1000 Genomes Project or TopMed. In this example, genotypes of SNP1 and SNP3 are imputed based on the directly assayed genotypes of other SNPs.
e | Genetic association tests are run for each genetic variant, using an appropriate model (for example, additive, non-additive, linear or logistic regression). Confounders are corrected for, including population strata, and multiple testing needs to be controlled. Output is inspected for unusual patterns and summary statistics are generated.
f | Results from multiple smaller cohorts are combined using standardized statistical pipelines.
g | Results can be replicated using internal replication or external replication in an independent cohort. For external replication, the independent cohort must be ancestrally matched and not share individuals or family members with the discovery cohort.
h | In silico analysis of genome-wide association studies (GWAS), using information from external resources. This can include in silico fine-mapping, SNP to gene mapping, gene to function mapping, pathway analysis, genetic correlation analysis, Mendelian randomization and polygenic risk prediction. After GWAS, functional hypotheses can be tested using experimental techniques such as CRISPR or massively parallel reporter assays, or results can be validated in a human trait/disease model (not shown).

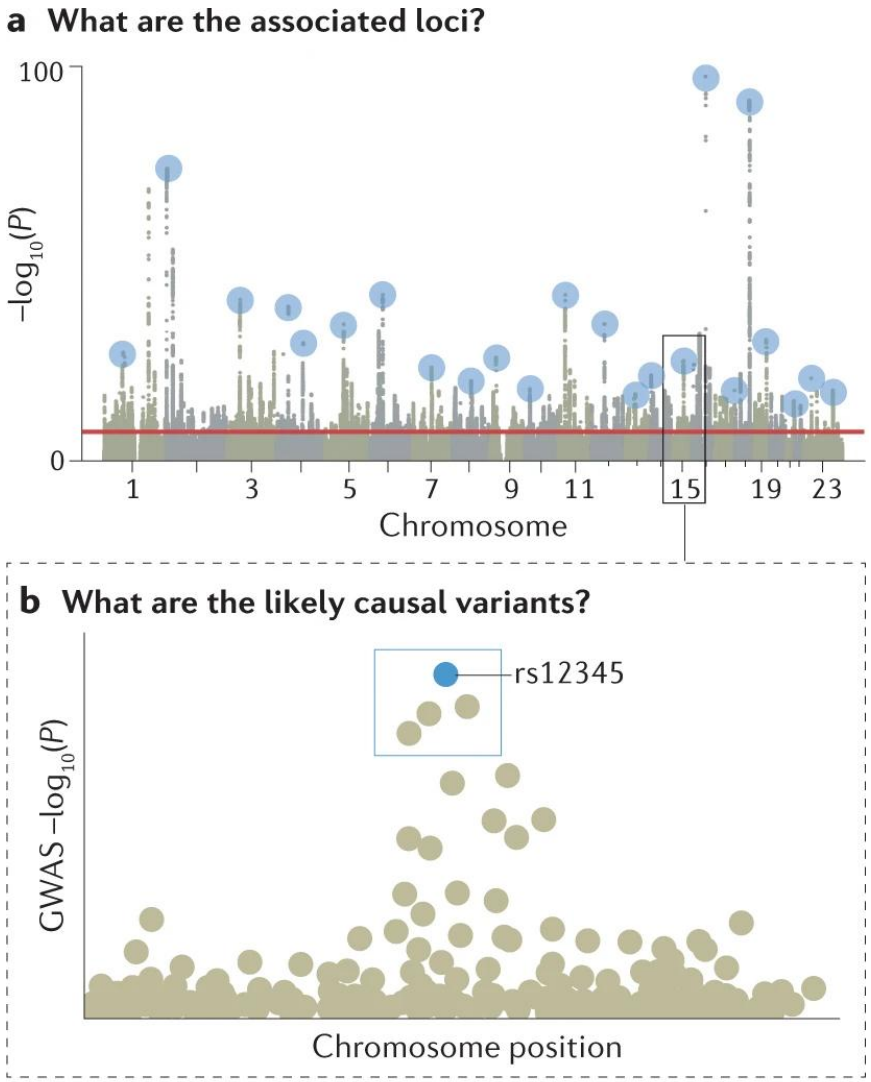
10.1038/s43586-021-00056-9

Manhattan plot and quantile–quantile plot to visualize GWAS results



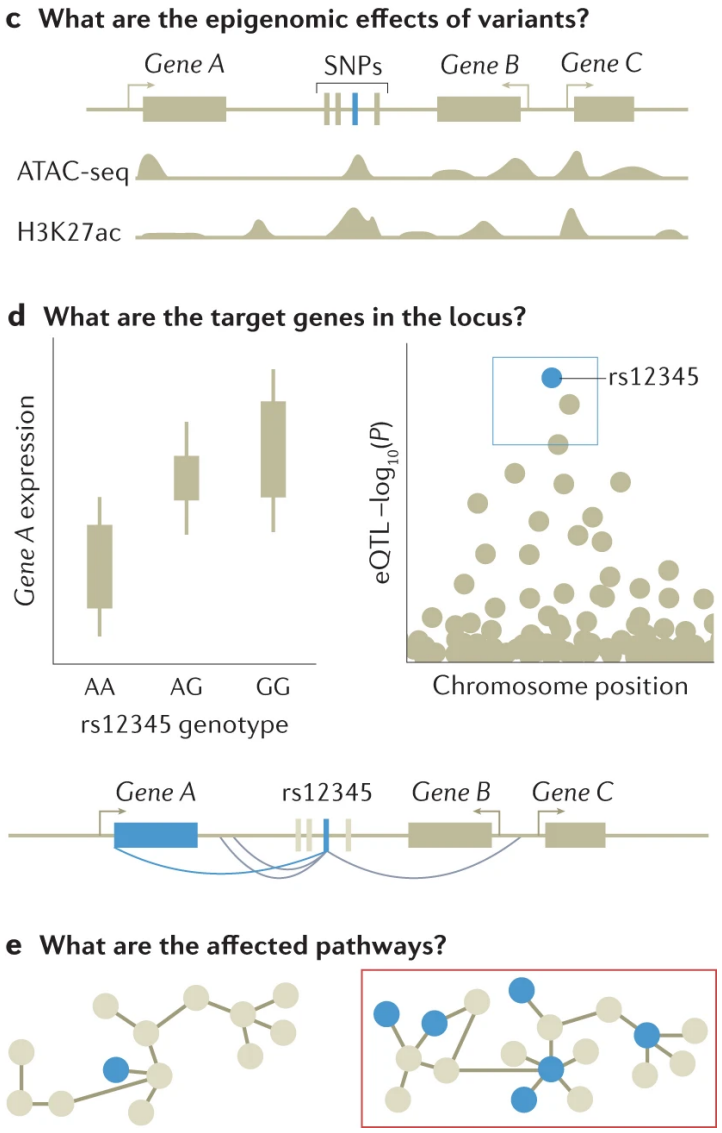
a | Manhattan plot showing significance of each variant's association with a phenotype (body mass index in this case77). Each dot represents a single-nucleotide polymorphism (SNP), with SNPs ordered on the x axis according to their genomic position. y axis represents strength of their association measured as –log10 transformed P values. Red line marks genome-wide significance threshold of P<5×10–8. b | Quantile–quantile plot showing distribution of expected P values under a null model of no significance versus observed P values. Expected –log10 transformed P values (x axis) for each association are plotted against observed values (y axis) to visualize the enrichment of association signal. Deviation from the expectation under the null hypothesis (red line) indicates the presence of either true causal effects or insufficiently corrected population stratification. In the case of true causal effects, one would expect to observe this deviation mostly at the right side of the plot, whereas population stratification causes the deviation to start closer to the origin. In this case, BMI is extremely polygenic and the genome-wide association study (GWAS) was highly powered, which may also cause the deviation to start close to the origin, making it difficult to visually spot stratification. LDSC may be used to assess whether this inflation is due to bias or polygenicity.

Illustration of functional follow-up of GWAS



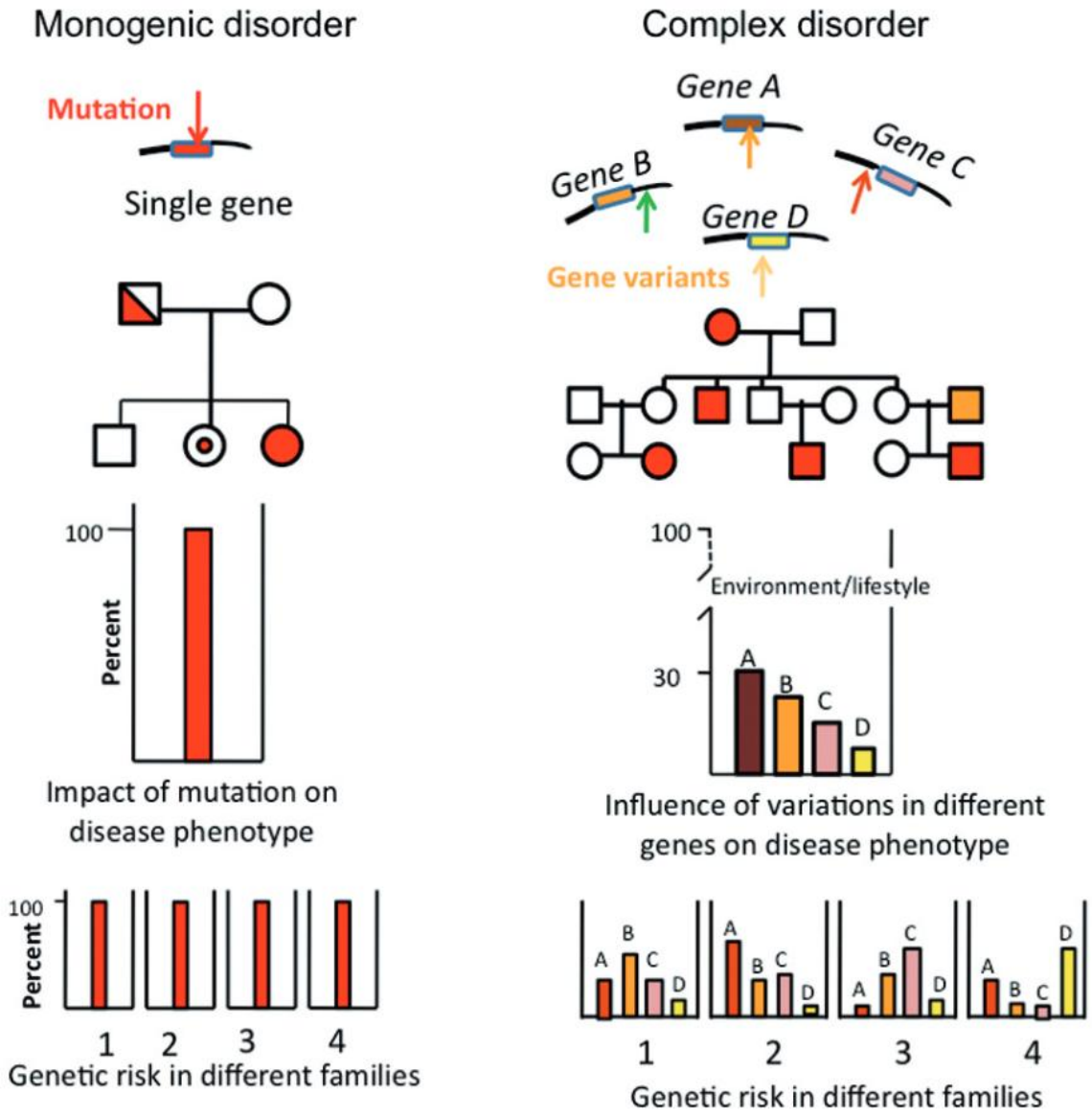
a | Genome-wide association studies (GWAS) are conducted to identify associated variants, often visualized as a Manhattan plot to show their genomic positions and strength of association. b | To prioritize likely causal variants, statistical fine-mapping is applied to identify a set of variants that are likely to include the causal variant (blue box) as well as the most likely causal variant (rs12345; blue dot). Massively parallel reporter assays can be used to measure whether alleles differ in their ability to drive gene expression or other molecular activity for each variant (not shown).

Illustration of functional follow-up of GWAS

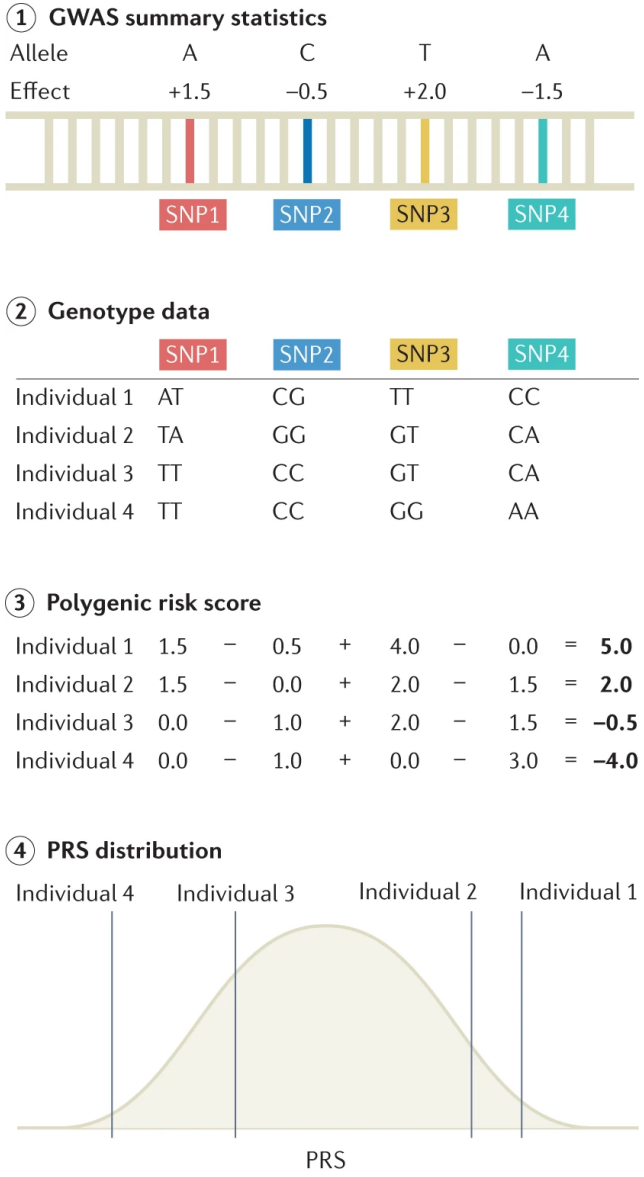


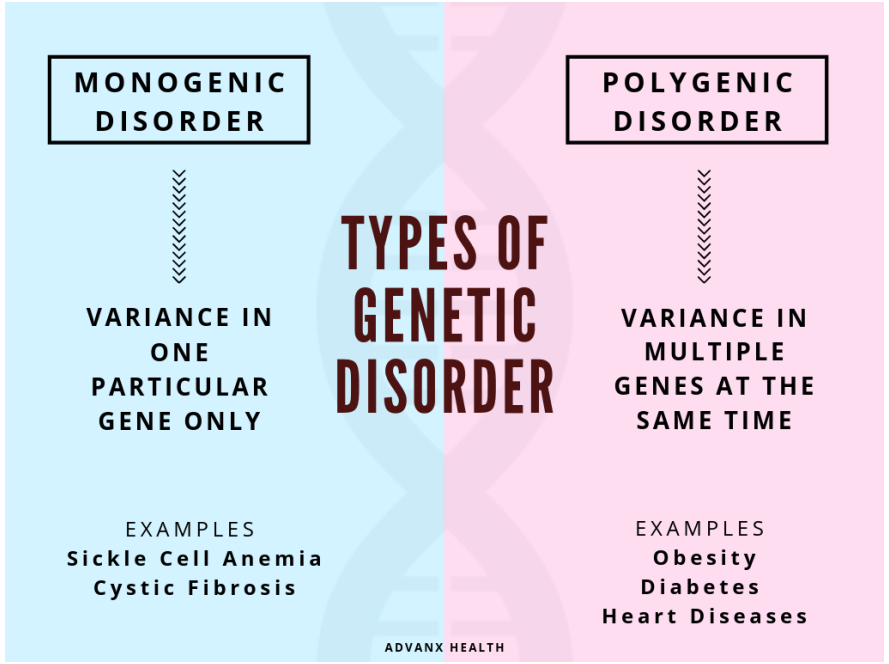
c | Functional annotations of the genome can be integrated with GWAS data to identify epigenetic mechanisms that may be perturbed by the causal variant, including enhancers, promoters or other functional elements. Additional approaches include mapping molecular quantitative trait loci (molQTL) or in vitro assays (not shown). d | Target gene for a GWAS locus can be prioritized by mapping expression quantitative trait loci (eQTLs) (left) and their co-localization (right) to identify loci where the causal variant from GWAS is also a causal variant affecting gene expression. For GWAS variants in enhancers, high-throughput chromosome conformation capture (Hi-C) data and maps of enhancer target genes can be used together with simple prioritization by distance to identify genes affected by the causal variant (below). e | To identify pathways whose perturbation may mediate the trait in question (red box), one can analyse the enrichment of multiple GWAS-implicated genes in predefined pathways. Additional approaches include trans-eQTL mapping and CRISPR perturbation of GWAS loci/genes followed by cellular phenotyping (not shown). For these analyses, the context of a relevant tissue, cell type and cell state needs to be carefully considered and analysed. ATAC-seq, assay for transposase-accessible chromatin using sequencing; H3K27Ac, histone H3 acetylated at K27; SNP, single-nucleotide polymorphism.

Polygenic Diseases



Overview of the steps necessary for calculating PRSs



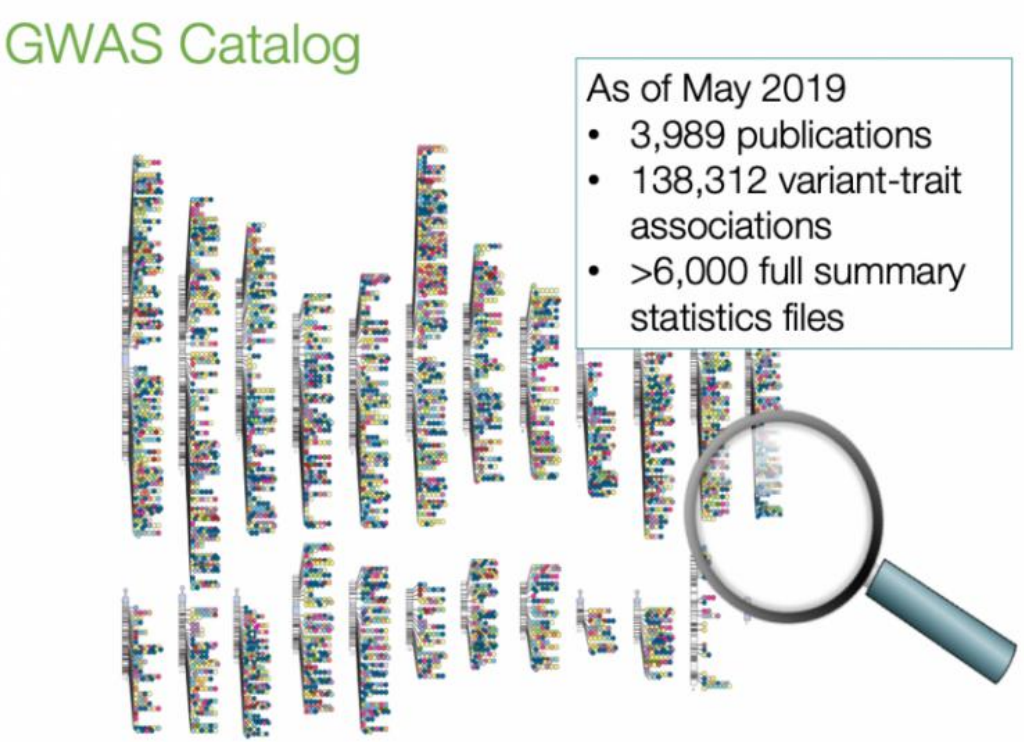


Step 1: genome-wide association studies (GWAS) summary statistics are obtained, which detail the effect of each single-nucleotide polymorphism (SNP) on the phenotype of interest.
Step 2: genotype data for a set of individuals are referenced against GWAS summary statistics. Here, genotype data for four SNPs are shown for four individuals.
Step 3: polygenic risk scores (PRSs) can be calculated for each individual by summing up the effect sizes of all risk alleles for each individual.
Step 4: linear regression analysis is performed on the calculated PRS to assess the effect of the PRS on the outcome measure.

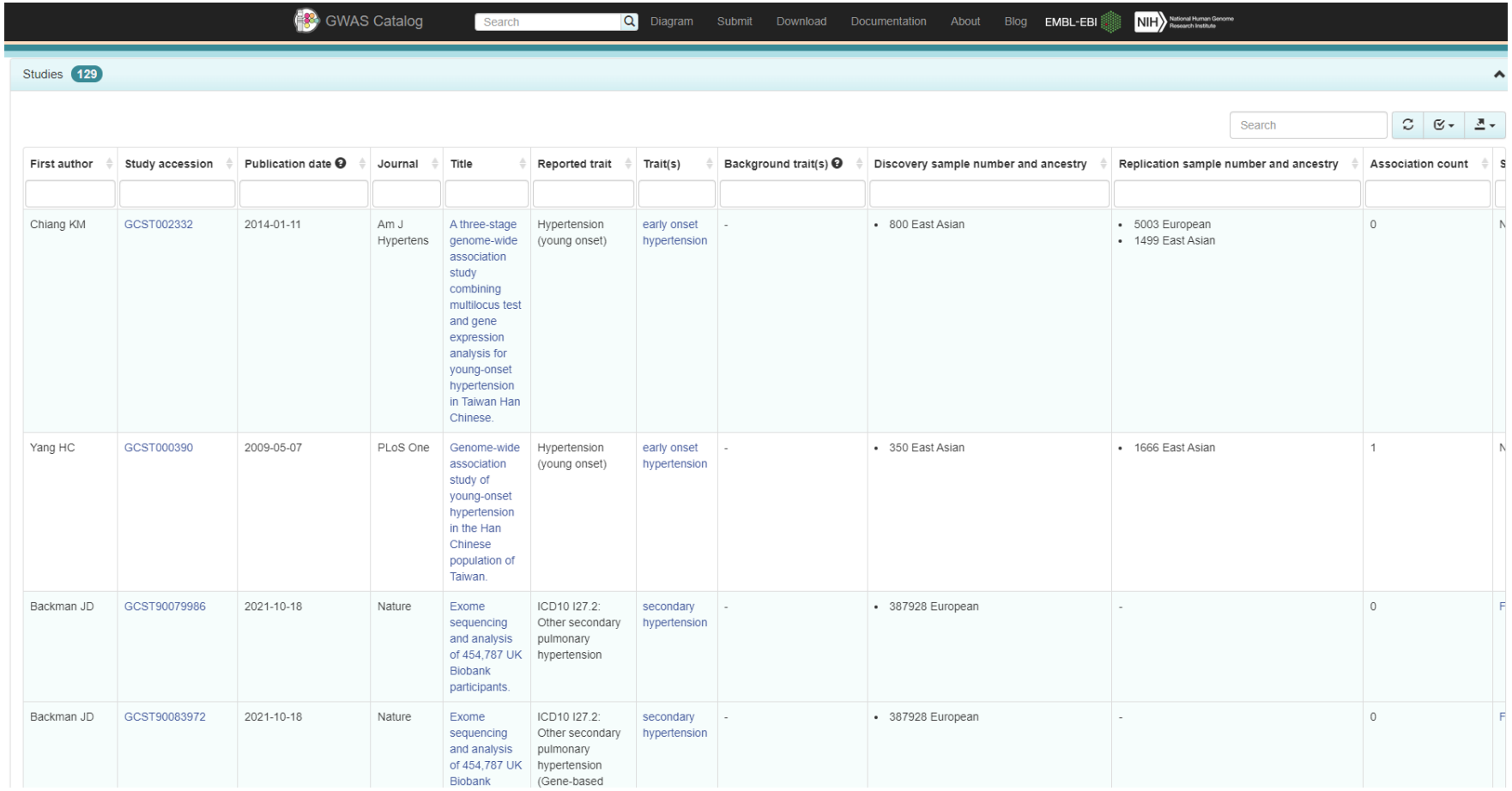
GWAS catalog

a structured repository which provides summary data from all published human GWAS studies, in a consistent, searchable format.

- The NHGRI-EBI GWAS Catalog is a publicly available resource of Genome Wide Association Studies (GWAS) and their results.
- It was originally founded in 2008 by the National Human Genome Research Institute (NHGRI), and since 2010 has been a collaboration between the EBI and the NHGRI
- The Catalog contains a vast amount of data and is designed to be easily accessible to scientists who want to use the data.
- It is searchable and contains useful visualisations of the variant-trait associations, which are mapped onto their chromosomal positions on the human genome.
- All traits and diseases are mapped onto an ontology to improve searchability.



GWAS catalog: A structured repository which provides summary data from all published human GWAS studies, in a consistent, searchable format.



First author	Study accession	Publication date	Journal	Title	Reported trait	Trait(s)	Background trait(s)	Discovery sample number and ancestry	Replication sample number and ancestry	Association count
Chiang KM	GCST002332	2014-01-11	Am J Hypertens	A three-stage genome-wide association study combining multilocus test and gene expression analysis for young-onset hypertension in Taiwan Han Chinese.	Hypertension (young onset)	early onset hypertension	-	• 800 East Asian	• 5003 European • 1499 East Asian	0
Yang HC	GCST000390	2009-05-07	PLoS One	Genome-wide association study of young-onset hypertension in the Han Chinese population of Taiwan.	Hypertension (young onset)	early onset hypertension	-	• 350 East Asian	• 1666 East Asian	1
Backman JD	GCST90079986	2021-10-18	Nature	Exome sequencing and analysis of 454,787 UK Biobank participants.	ICD10 I27.2: Other secondary pulmonary hypertension	secondary hypertension	-	• 387928 European	-	0
Backman JD	GCST90083972	2021-10-18	Nature	Exome sequencing and analysis of 454,787 UK Biobank	ICD10 I27.2: Other secondary pulmonary hypertension (Gene-based	secondary hypertension	-	• 387928 European	-	0

https://www.ebi.ac.uk/gwas/efotraits/EFO_0000537

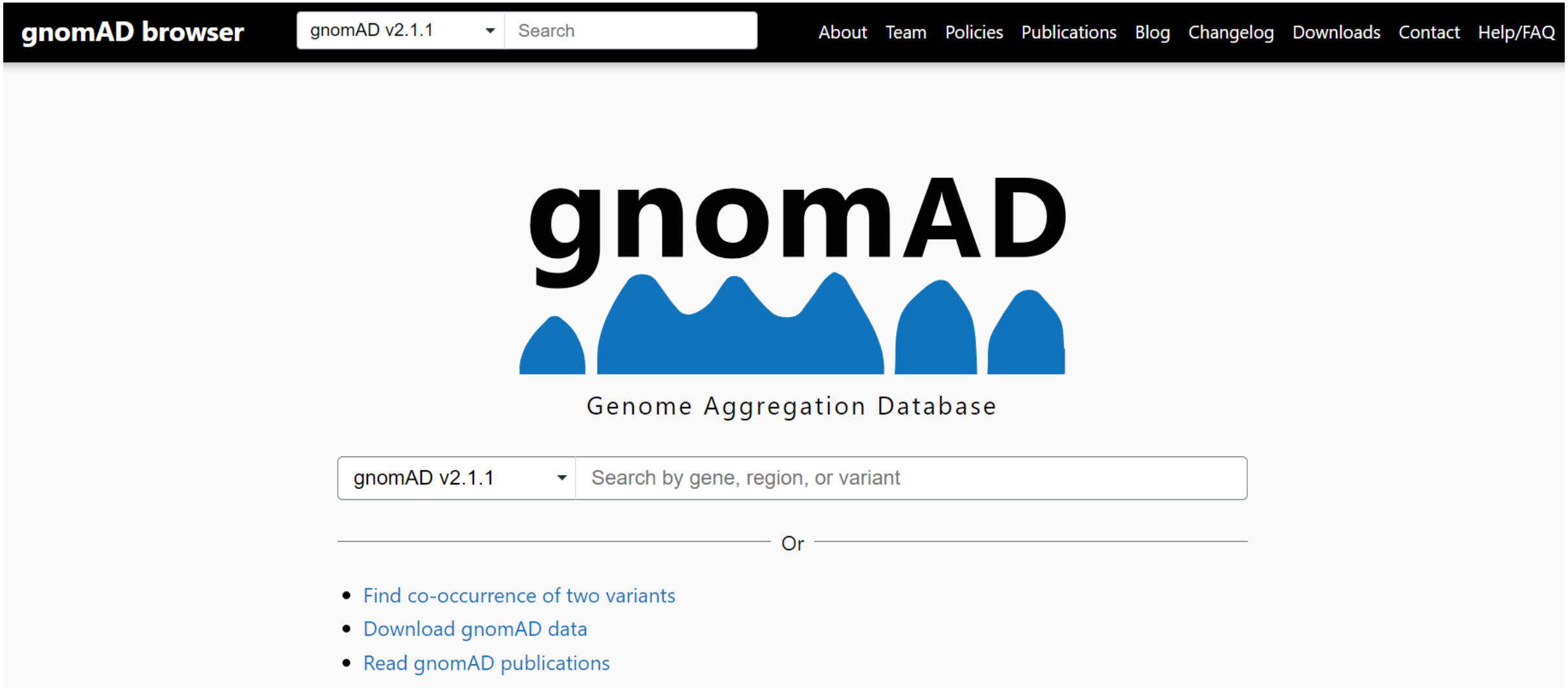
Table 3 | **Databases of GWAS summary statistics**

Database	Content
GWAS Catalog[110]	GWAS summary statistics and GWAS lead SNPs reported in GWAS papers
GeneAtlas[8]	UK Biobank GWAS summary statistics
Pan UKBB	UK Biobank GWAS summary statistics
GWAS Atlas[273]	Collection of publicly available GWAS summary statistics with follow-up in silico analysis
FinnGen results	GWAS summary statistics released from FinnGen, a project that collected biological samples from many sources in Finland
dbGAP	Public depository of National Institutes of Health-funded genomics data including GWAS summary statistics
OpenGWAS database	GWAS summary data sets
Pheweb.jp	GWAS summary statistics of Biobank Japan and cross-population meta-analyses

For a comprehensive list of genetic data resources, see REF.[13]. GWAS, genome-wide association studies; SNP, single-nucleotide polymorphism.

10.1038/s43586-021-00056-9

https://genome.vinbigdata.org/



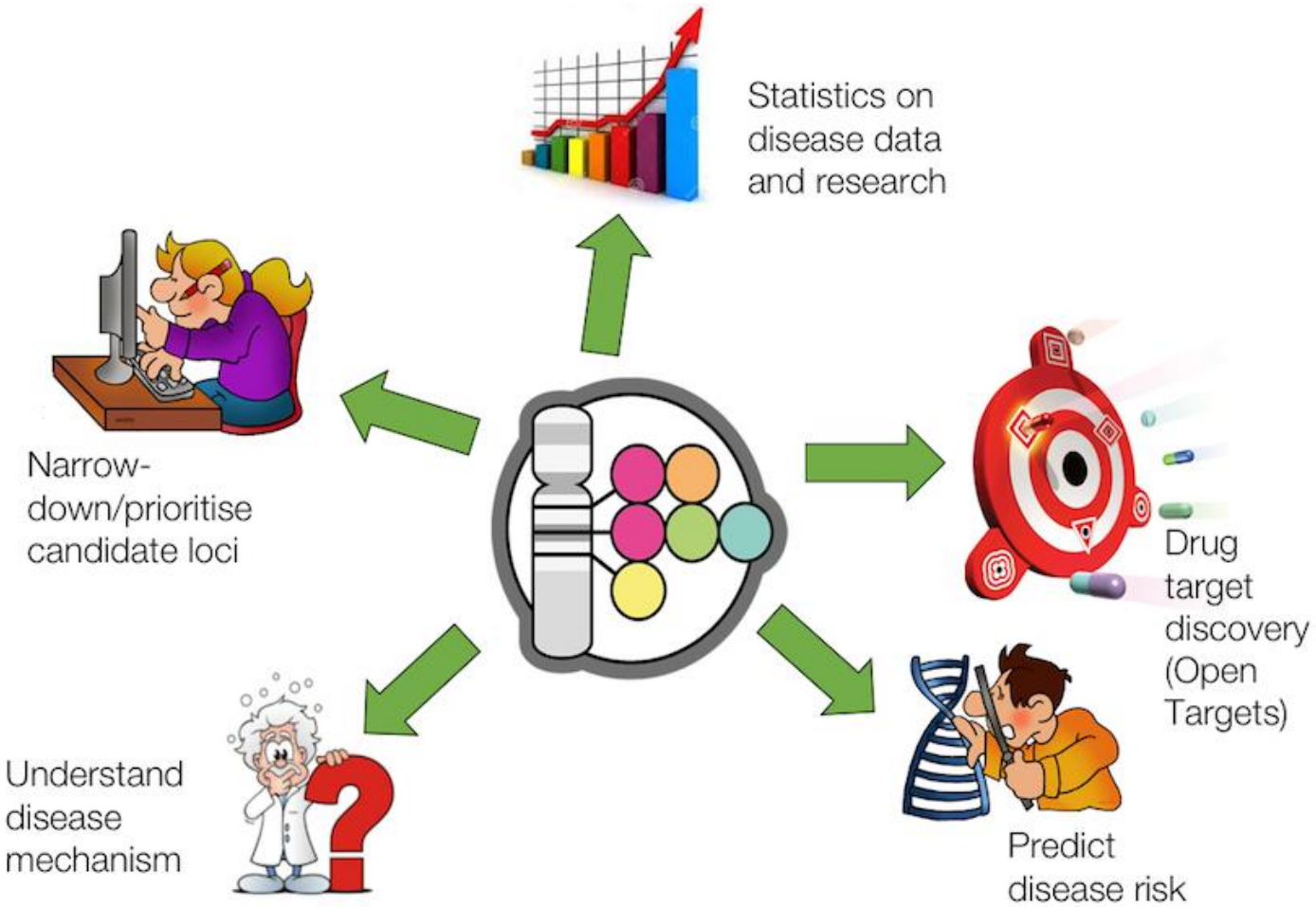
https://gnomad.broadinstitute.org/

https://ow.ly/qeIv50Pzr0R

Why do we do GWAS?

- We do GWAS because a statistical association between a particular physical region of the genome and the phenotype
  - can point to biological **mechanisms** affecting the phenotype
  - can allow **prediction** of the phenotype from genomic information
- These results may further benefit
  - **medicine** by leading to molecular or environmental interventions against harmful phenotypes
  - **biotechnology** by improving the ways we utilize microbes, plants or animals
  - **forensics** by more accurate identification of an individual from a DNA sample
  - **biogeographic ancestry inference** of individuals, populations and species
  - our **understanding of the role of natural selection** and other evolutionary forces in the living world

Why do we do GWAS?



https://blog.opentargets.org/everything-you-need-to-know-about-the-new-gwas-catalog/