3000788 Intro to Comp Molec Biol

Lecture 6: Variant calling and analysis

Fall 2025

Sira Sriswasdi, PhD

- Research Affairs
- Center of Excellence in Computational Molecular Biology (CMB)
- Center for Artificial Intelligence in Medicine (CU-AIM)

Today's agenda

- Variant calling (and filtering)

- Variant annotation / interpretation

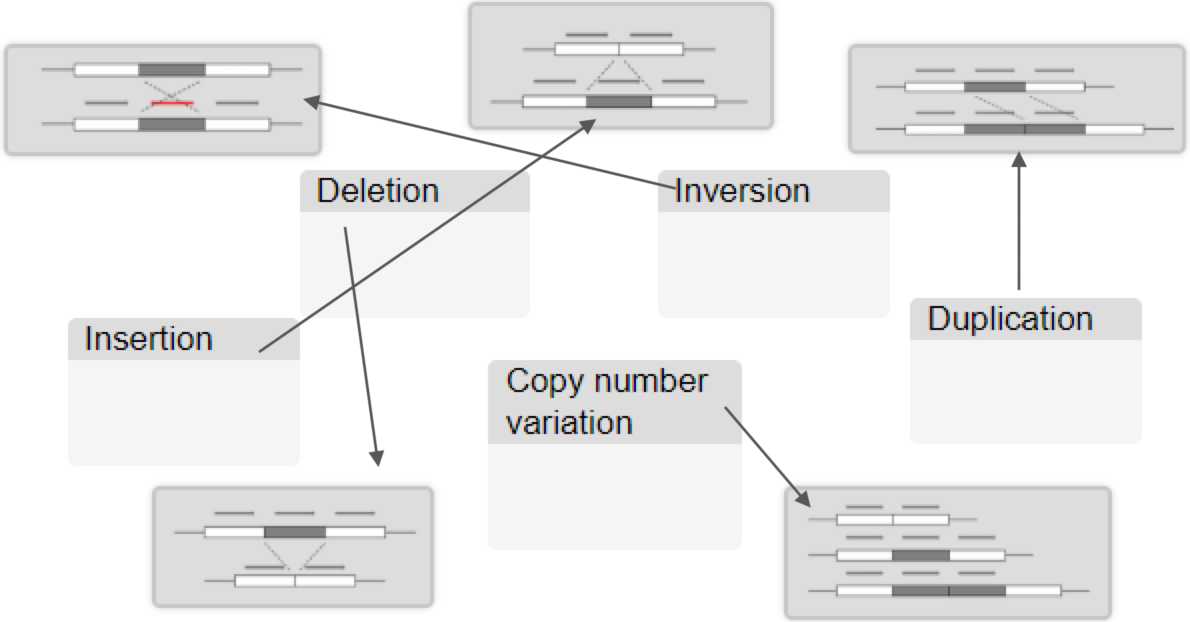
- Genotype-phenotype association
  - Quantitative trait loci (QTL)
  - Genome-wide association study (GWAS)

Variant calling

What we knew so far

- We can align sequencing reads to reference genome
- We can assemble reads into contigs and scaffolds
- **We can identify differences between sequence data and reference genomes**

Type of variants



Deletion

Inversion

Insertion

Duplication

Copy number variation

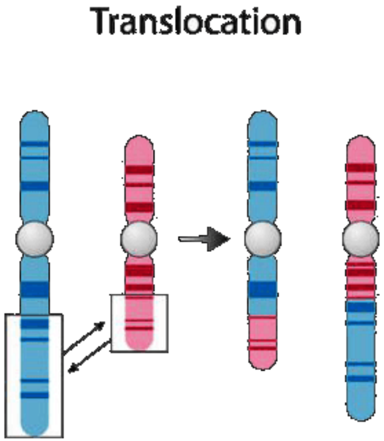


Image from wikipedia

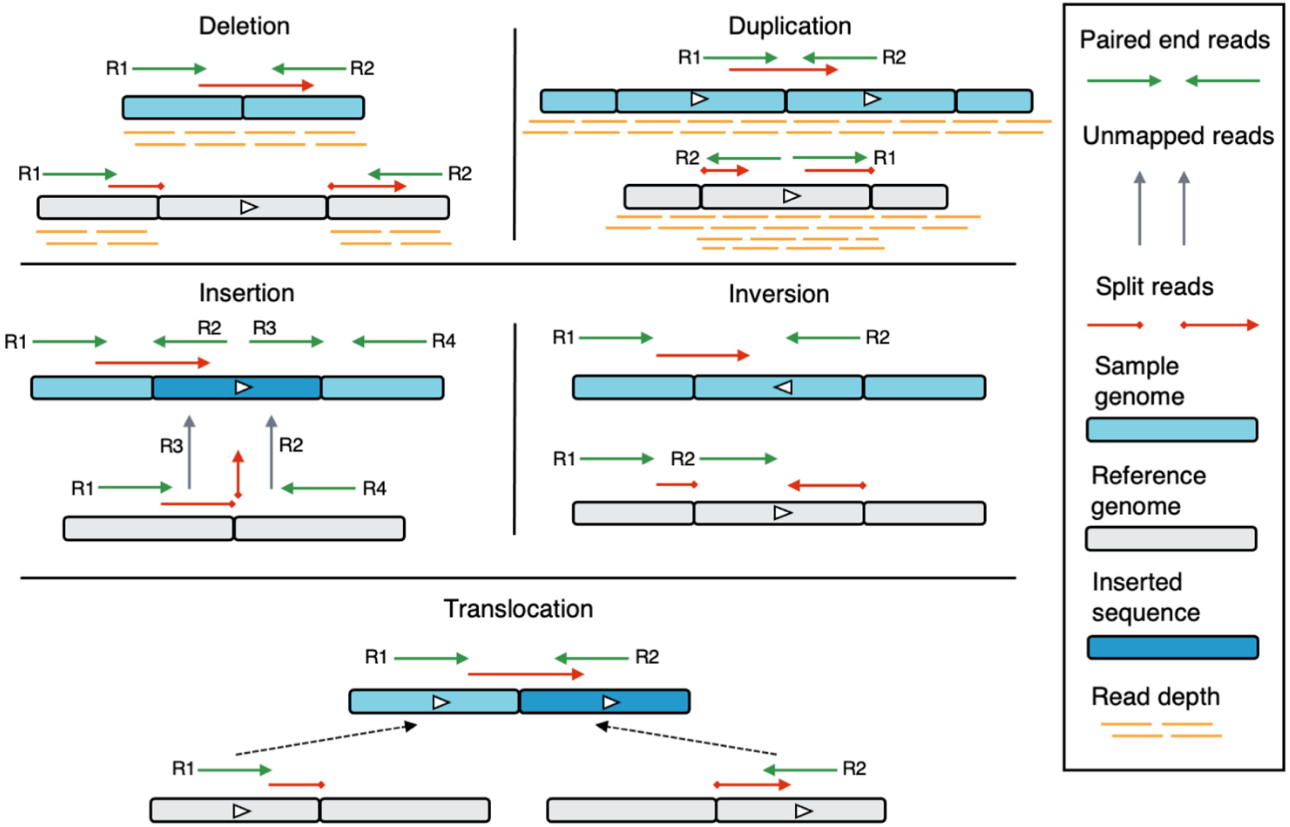
Image from EBI

Variant calling from read alignment & assembly

- Small-scale variants: SNV and small indel
    - Identified from alignments and read frequencies
- Copy number variations
    - Genomic loci with relatively higher or lower frequencies
- Chromosomal translocation and inversion
    - Paired-end reads whose forward and reverse mapped to different regions
    - *De novo* assembly

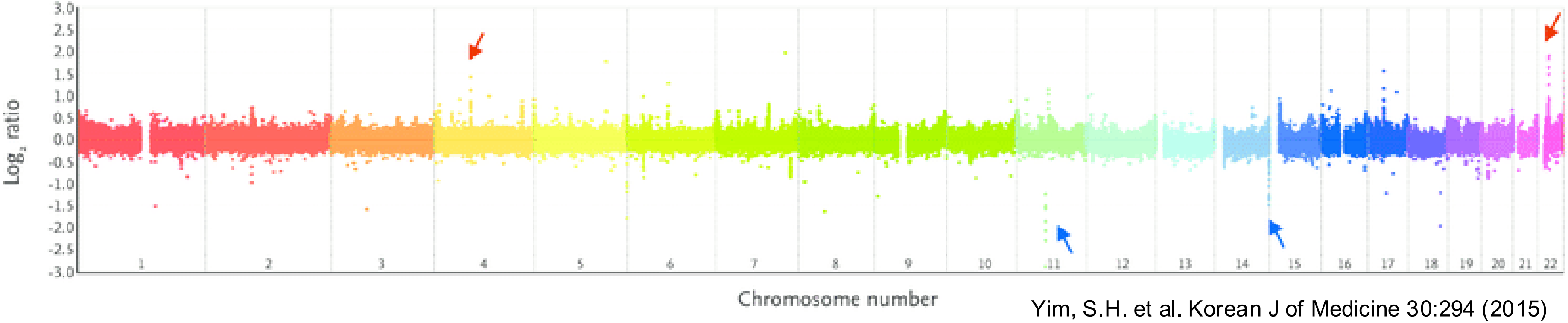
Short read signatures of structural variants

Legend



https://gatk.broadinstitute.org/hc/en-us/articles/9022476791323-Structural-Variants

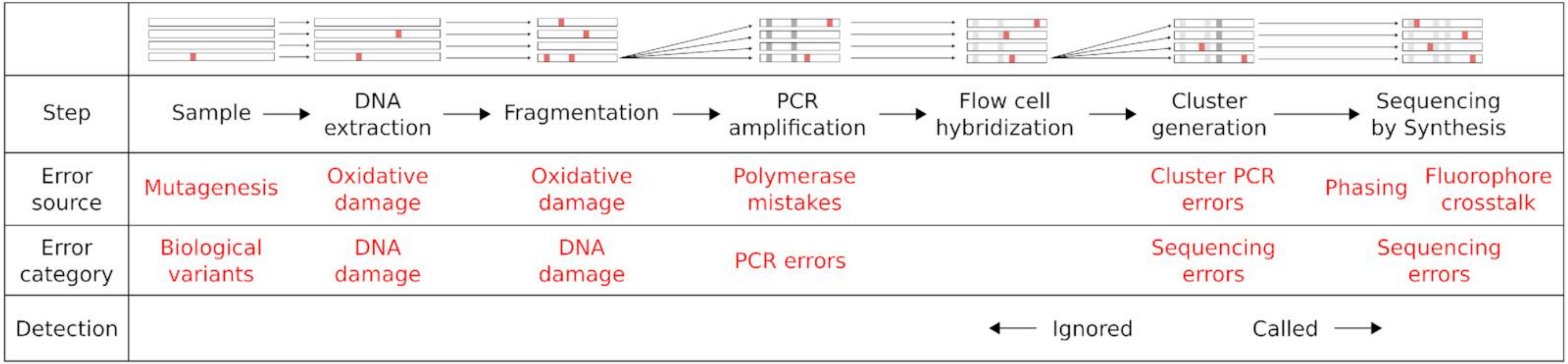
Copy number variations



Yim, S.H. et al. Korean J of Medicine 30:294 (2015)

- Usually clearly seen by elevated (gain) or diminished (loss) number of reads mapped to some genomic regions
- Can be observed with shallow sequencing

Not all differences are true variants



Step	Sample →	DNA extraction →	Fragmentation →	PCR amplification →	Flow cell hybridization →	Cluster generation →	Sequencing by Synthesis
Error source	Mutagenesis	Oxidative damage	Oxidative damage	Polymerase mistakes		Cluster PCR errors	Phasing Fluorophore crosstalk
Error category	Biological variants	DNA damage	DNA damage	PCR errors		Sequencing errors	Sequencing errors
Detection					← Ignored	Called →	

- Errors are random
- High sequencing depth can resolve (but costly)

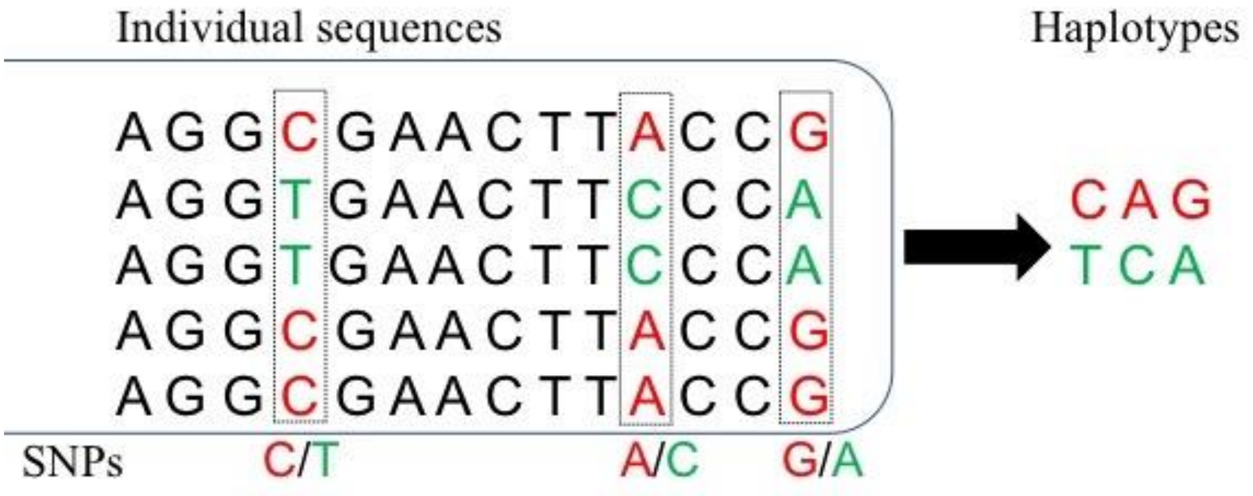
Germline versus somatic variants

Germline vs somatic variants

- **Germline** = inherited
  - Appear in every cell
  - Compare DNA from normal tissues to reference genomes, **or parental**
  - Heterozygous or homozygous

- **Somatic** = occurred during lifetime
  - Compare DNA from diseased tissues to normal tissues, **and other healthy people**
  - Variable frequency (different cells can have different mutations)

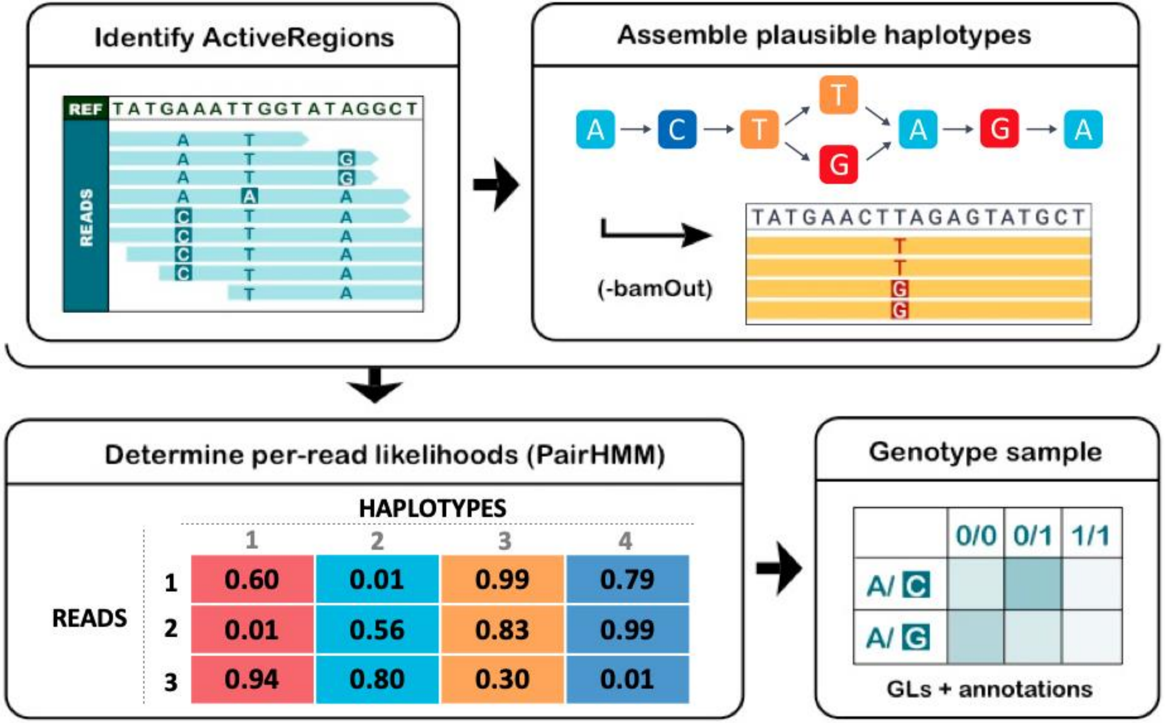
Haplotype



https://openpress.wheatoncollege.edu/molecularecologyv1/chapter/haplotype-networks/

- Certain regions of the genome segregate together, also called “linkage”
- Linked variants should be called together

Germline variant calling: identify haplotype



Identify ActiveRegions

Assemble plausible haplotypes

Determine per-read likelihoods (PairHMM)

HAPLOTYPES

READS	1	2	3	4
1	0.60	0.01	0.99	0.79
2	0.01	0.56	0.83	0.99
3	0.94	0.80	0.30	0.01

Genotype sample

	0/0	0/1	1/1
A/ C			
A/ G			

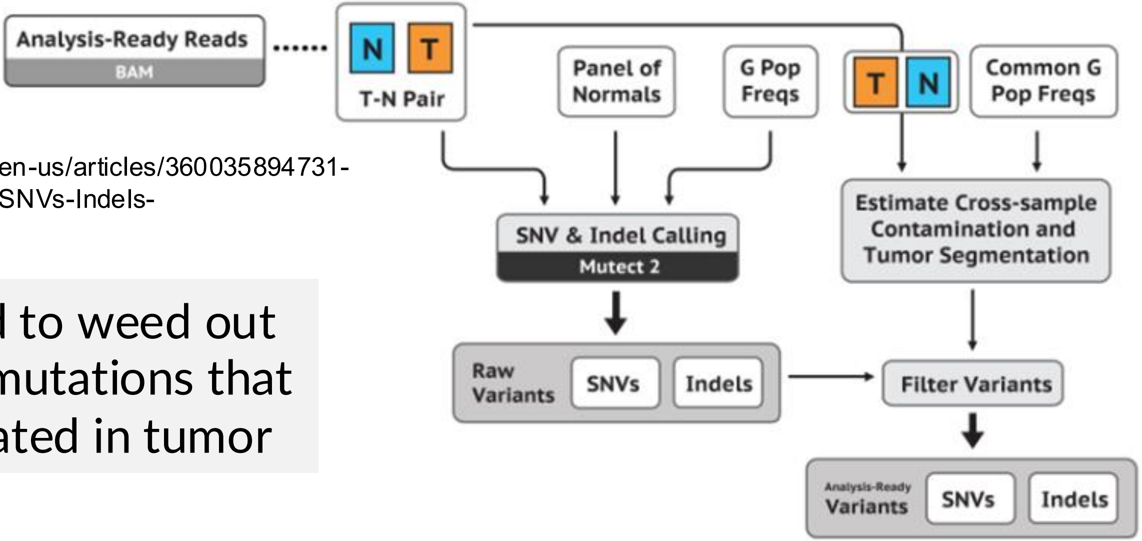
GLs + annotations

https://gatk.broadinstitute.org/hc/en-us/articles/360035531412-HaplotypeCaller-in-a-nutshell

PairHMM

- **Hidden states**: Position on the haplotype
- **Observations**: Position on the sequence reads
- Find haplotype that maximize: **P(read | haplotype)**

Somatic variant calling: tumor example



https://gatk.broadinstitute.org/hc/en-us/articles/360035894731-Somatic-short-variant-discovery-SNVs-Indels-

Designed to weed out spurious mutations that accumulated in tumor

- Use matched normal (N), panels of healthy individual (PoN), and allele frequency in the general population (G Pop Freqs) to remove
- Estimate **contamination** = fraction of normal cells in tumor sample

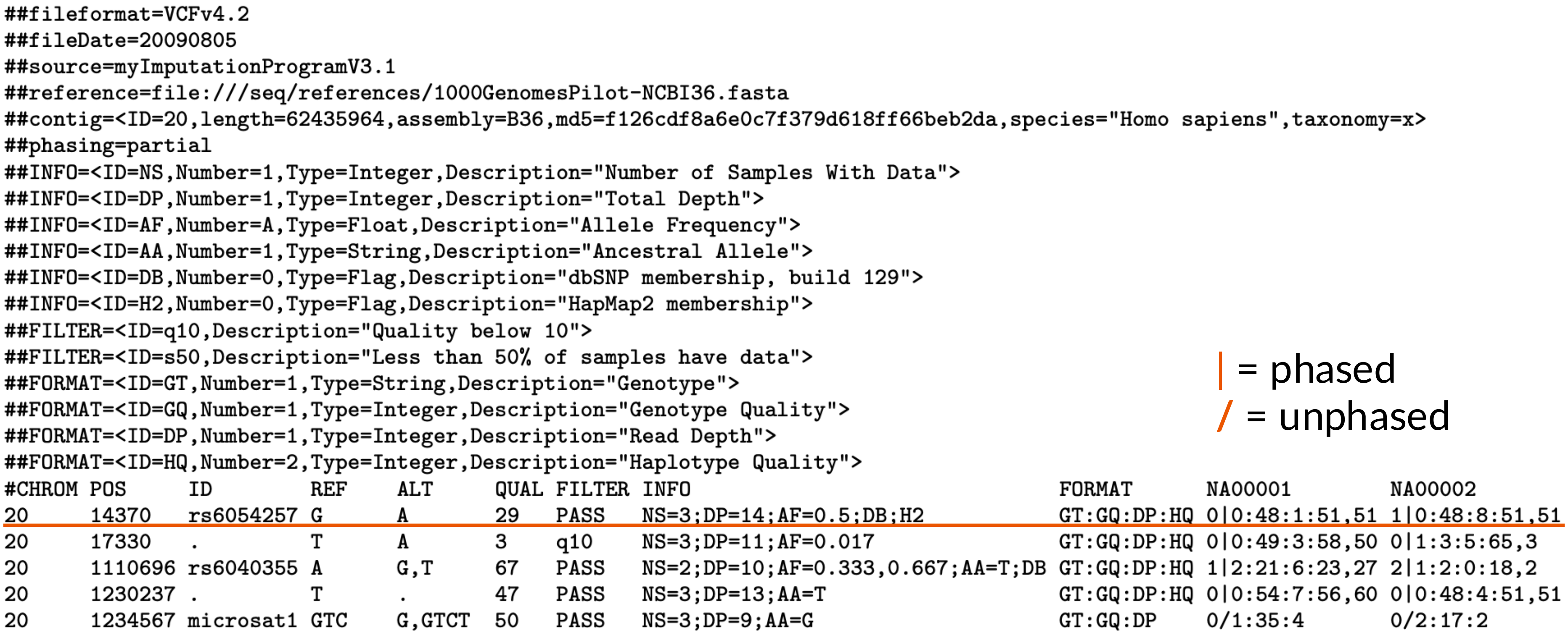
Somatic variant calling algorithm sketch

- Call variants on tumor and normal samples separately

- For each tumor variant, check whether it is present in normal, panel of normal, or general population → remove

- Usually still too many variants
    - Filtering
    - Intersecting variants called by multiple tools
        - **Examples**: Mutect2 (GATK), Strelka, Varscans2, DeepVariant
        - Machine learning models

Variant Call Format (VCF)

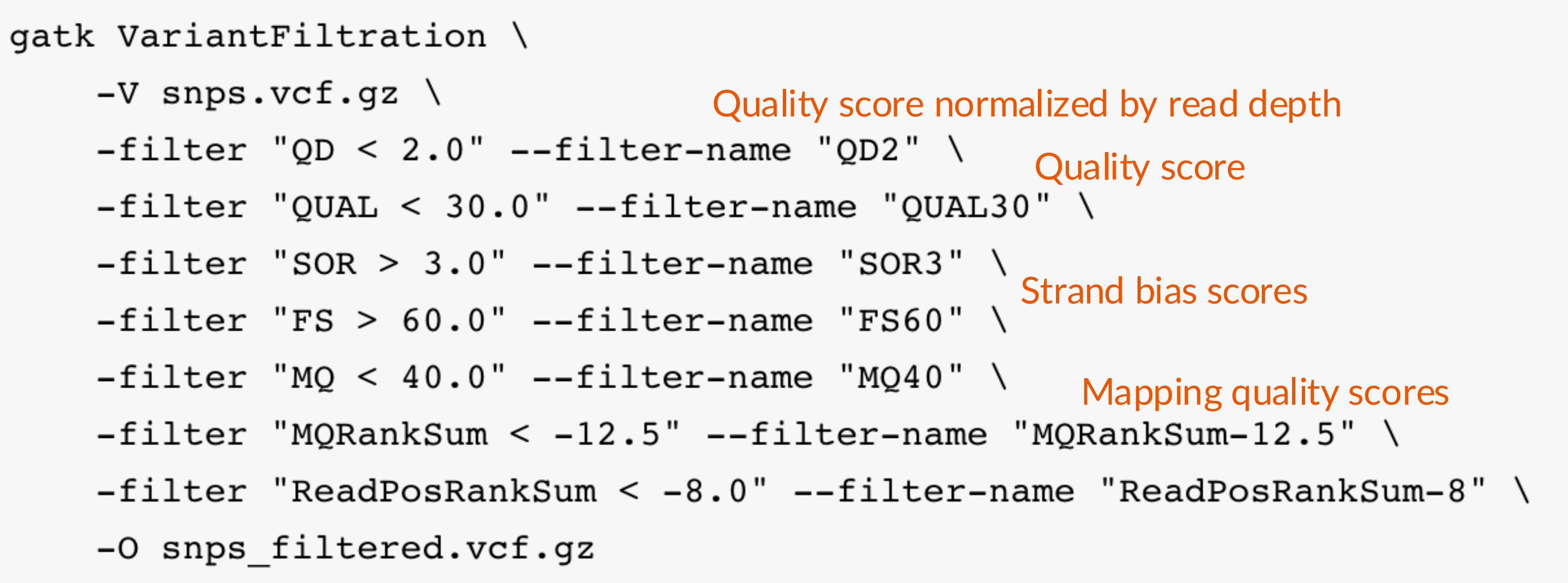


```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number 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
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
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
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
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
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
```

| = phased

/ = unphased

Variant filtering



```
gatk VariantFiltration \
    -V snps.vcf.gz \
    -filter "QD < 2.0" --filter-name "QD2" \
    -filter "QUAL < 30.0" --filter-name "QUAL30" \
    -filter "SOR > 3.0" --filter-name "SOR3" \
    -filter "FS > 60.0" --filter-name "FS60" \
    -filter "MQ < 40.0" --filter-name "MQ40" \
    -filter "MQRankSum < -12.5" --filter-name "MQRankSum-12.5" \
    -filter "ReadPosRankSum < -8.0" --filter-name "ReadPosRankSum-8" \
    -O snps_filtered.vcf.gz
```

Quality score normalized by read depth

Quality score

Strand bias scores

Mapping quality scores

- More of an art than science, follow publications and guidelines

Variant annotation / interpretation

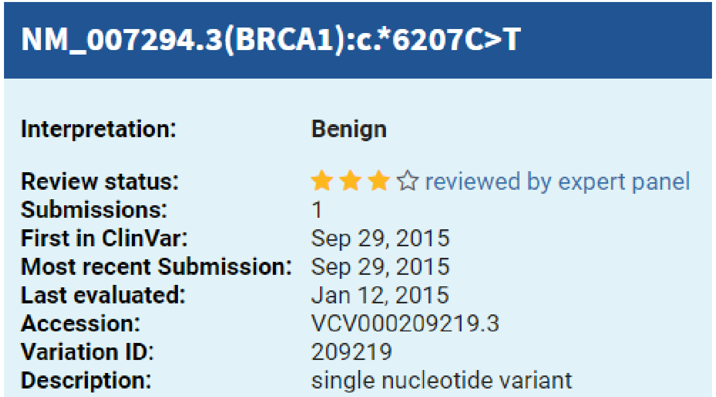
Functional impact of a variant

- Does it impact protein integrity and expression?
  - Non-synonymous/synonymous, frameshift
  - Splice site, regulatory element

- Is it a known or common variant?
  - Genome Aggregation Database: gnomAD, dbSNP

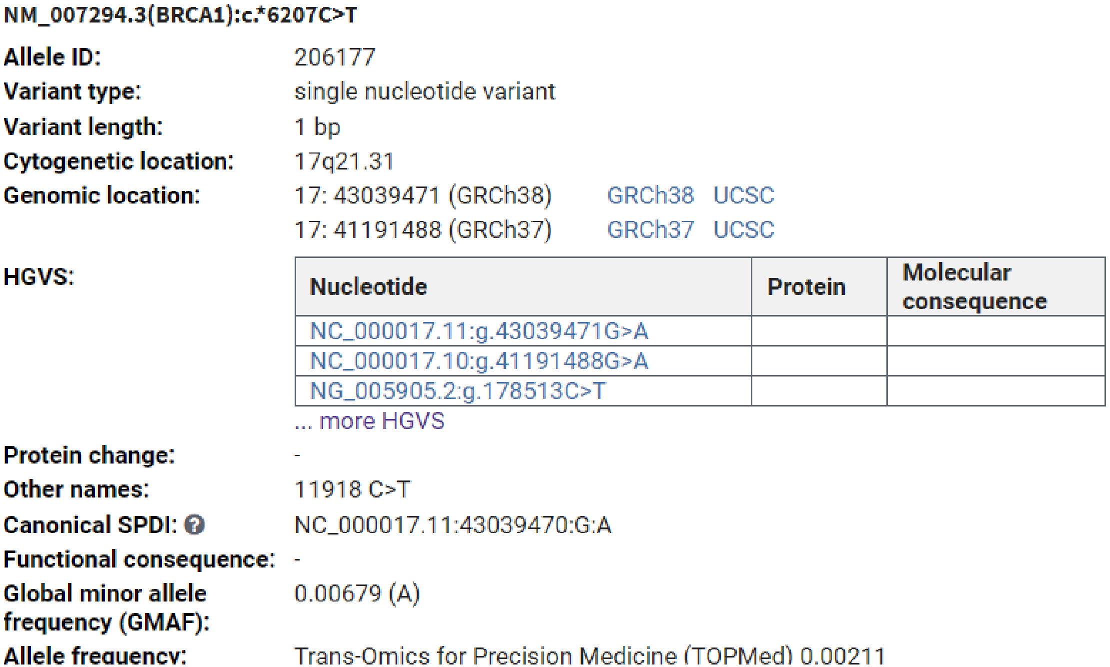
- Does it have clinical implication?
  - Frequency in patients
  - Association with drugs/treatments
  - ClinVar, COSMIC, PharmGKB

ClinVar



NM_007294.3(BRCA1):c.*6207C>T

Interpretation:	Benign
Review status:	★★★☆ reviewed by expert panel
Submissions:	1
First in ClinVar:	Sep 29, 2015
Most recent Submission:	Sep 29, 2015
Last evaluated:	Jan 12, 2015
Accession:	VCV000209219.3
Variation ID:	209219
Description:	single nucleotide variant



NM_007294.3(BRCA1):c.*6207C>T

Allele ID:	206177
Variant type:	single nucleotide variant
Variant length:	1 bp
Cytogenetic location:	17q21.31
Genomic location:	17: 43039471 (GRCh38) GRCh38 UCSC 17: 41191488 (GRCh37) GRCh37 UCSC

HGVS:

Nucleotide	Protein	Molecular consequence
NC_000017.11:g.43039471G>A		
NC_000017.10:g.41191488G>A		
NG_005905.2:g.178513C>T		

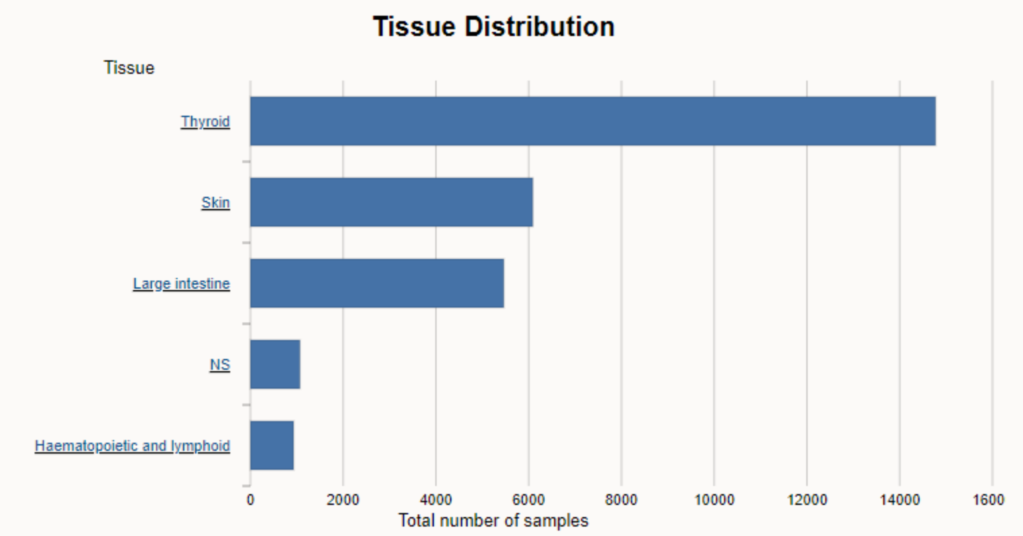
... more HGVS

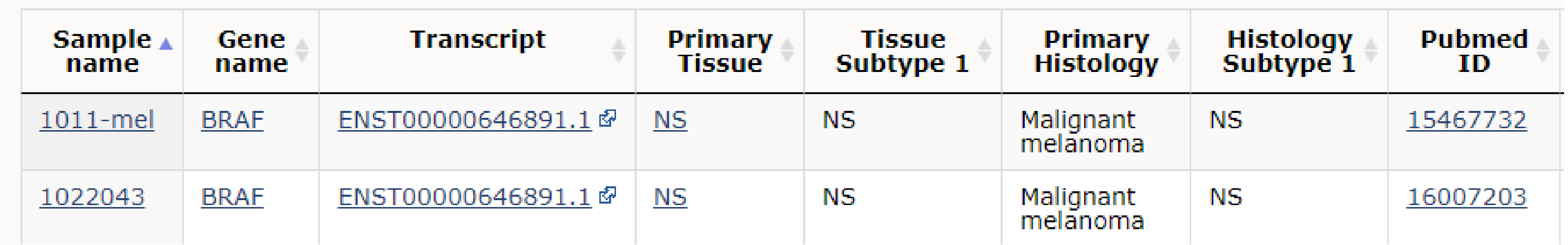
Protein change:	-
Other names:	11918 C>T
Canonical SPDI: ?	NC_000017.11:43039470:G:A
Functional consequence:	-
Global minor allele frequency (GMAF):	0.00679 (A)
Allele frequency:	Trans-Omics for Precision Medicine (TOPMed) 0.00211

Catalog of Somatic Mutations in Cancer (COSMIC)

Mutation

COSV56056643

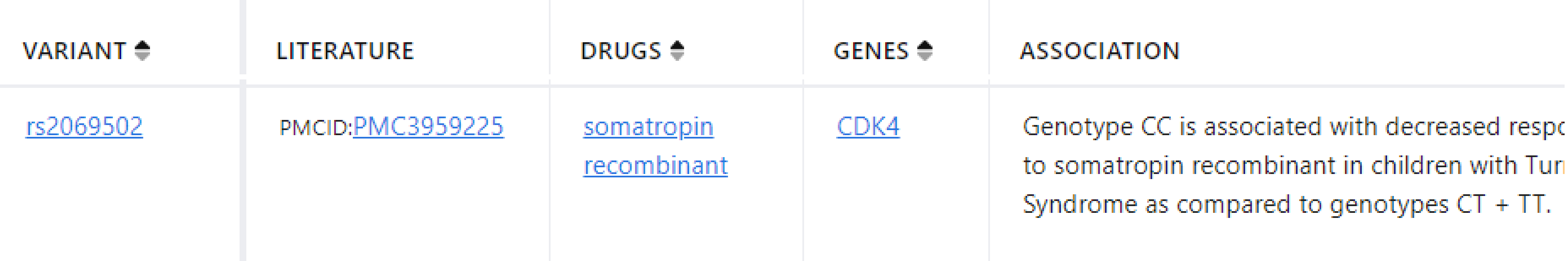




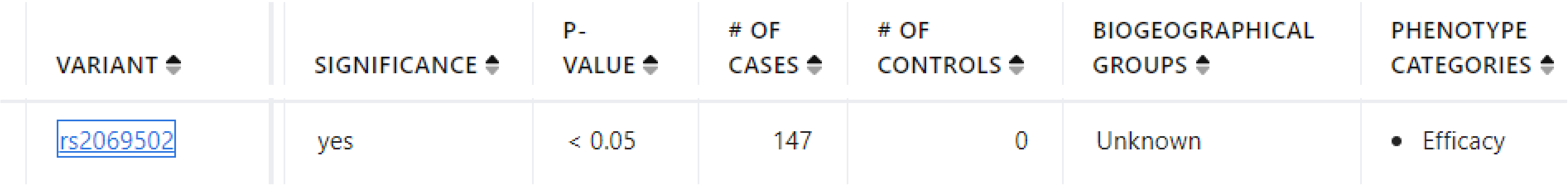
Sample name	Gene name	Transcript	Primary Tissue	Tissue Subtype 1	Primary Histology	Histology Subtype 1	Pubmed ID
1011-mel	BRAF	ENST00000646891.1	NS	NS	Malignant melanoma	NS	15467732
1022043	BRAF	ENST00000646891.1	NS	NS	Malignant melanoma	NS	16007203

Pharmocogenomics knowledgebase

PHARMGKB

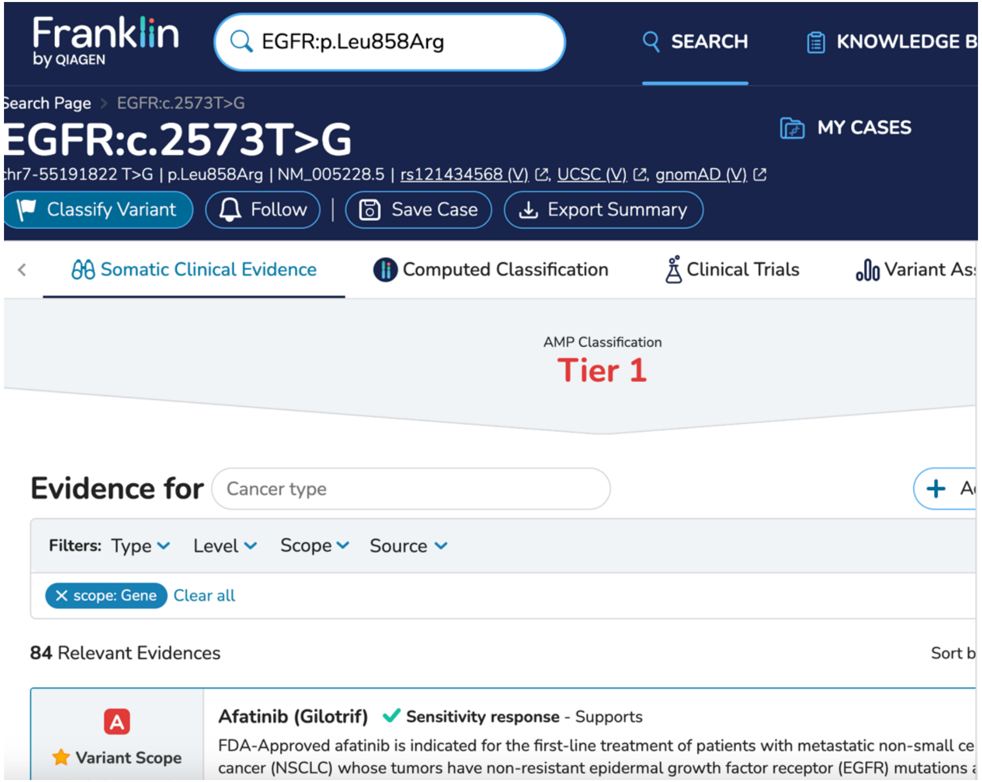


VARIANT	LITERATURE	DRUGS	GENES	ASSOCIATION
rs2069502	PMCID:PMC3959225	somatropin recombinant	CDK4	Genotype CC is associated with decreased resp to somatropin recombinant in children with Tur Syndrome as compared to genotypes CT + TT.



VARIANT	SIGNIFICANCE	P-VALUE	# OF CASES	# OF CONTROLS	BIOGEOGRAPHICAL GROUPS	PHENOTYPE CATEGORIES
rs2069502	yes	< 0.05	147	0	Unknown	• Efficacy

Franklin



earch Page > EGFR:c.2573T>G

EGFR:c.2573T>G

hr7-55191822 T>G | p.Leu858Arg | NM_005228.5 | rs121434568 (V), UCSC (V), gnomAD (V)

AMP Classification

Tier 1

Evidence for

84 Relevant Evidences

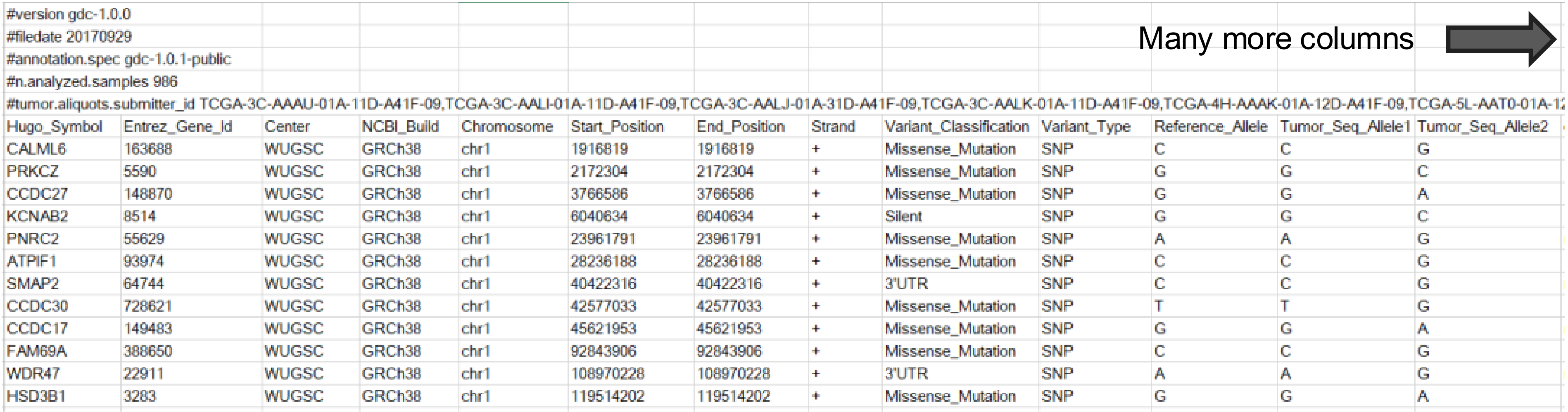
Afatinib (Gilotrif) ✓ **Sensitivity response** - Supports

FDA-Approved afatinib is indicated for the first-line treatment of patients with metastatic non-small ce

cancer (NSCLC) whose tumors have non-resistant epidermal growth factor receptor (EGFR) mutations a

- Proprietary algorithm for prioritizing variants
- Assess variant confidence and association with clinical information
- Classify variants according to clinical guideline

Mutation Annotation Format (MAF)



#version gdc-1.0.0
#filedate 20170929
#annotation.spec gdc-1.0.1-public
#n.analyzed.samples 986
#tumor.aliquots.submitter_id TCGA-3C-AAAU-01A-11D-A41F-09,TCGA-3C-AALI-01A-11D-A41F-09,TCGA-3C-AALJ-01A-31D-A41F-09,TCGA-3C-AALK-01A-11D-A41F-09,TCGA-4H-AAAK-01A-12D-A41F-09,TCGA-5L-AAT0-01A-1

Hugo_Symbol	Entrez_Gene_Id	Center	NCBI_Build	Chromosome	Start_Position	End_Position	Strand	Variant_Classification	Variant_Type	Reference_Allele	Tumor_Seq_Allele1	Tumor_Seq_Allele2
CALML6	163688	WUGSC	GRCh38	chr1	1916819	1916819	+	Missense_Mutation	SNP	C	C	G
PRKCZ	5590	WUGSC	GRCh38	chr1	2172304	2172304	+	Missense_Mutation	SNP	G	G	C
CCDC27	148870	WUGSC	GRCh38	chr1	3766586	3766586	+	Missense_Mutation	SNP	G	G	A
KCNAB2	8514	WUGSC	GRCh38	chr1	6040634	6040634	+	Silent	SNP	G	G	C
PNRC2	55629	WUGSC	GRCh38	chr1	23961791	23961791	+	Missense_Mutation	SNP	A	A	G
ATPIF1	93974	WUGSC	GRCh38	chr1	28236188	28236188	+	Missense_Mutation	SNP	C	C	G
SMAP2	64744	WUGSC	GRCh38	chr1	40422316	40422316	+	3'UTR	SNP	C	C	G
CCDC30	728621	WUGSC	GRCh38	chr1	42577033	42577033	+	Missense_Mutation	SNP	T	T	G
CCDC17	149483	WUGSC	GRCh38	chr1	45621953	45621953	+	Missense_Mutation	SNP	G	G	A
FAM69A	388650	WUGSC	GRCh38	chr1	92843906	92843906	+	Missense_Mutation	SNP	C	C	G
WDR47	22911	WUGSC	GRCh38	chr1	108970228	108970228	+	3'UTR	SNP	A	A	G
HSD3B1	3283	WUGSC	GRCh38	chr1	119514202	119514202	+	Missense_Mutation	SNP	G	G	A

Many more columns

https://cloud.tencent.com/developer/article/2116172

- Tabular file, one row per variant
- Containing numerous columns, one for each annotation system

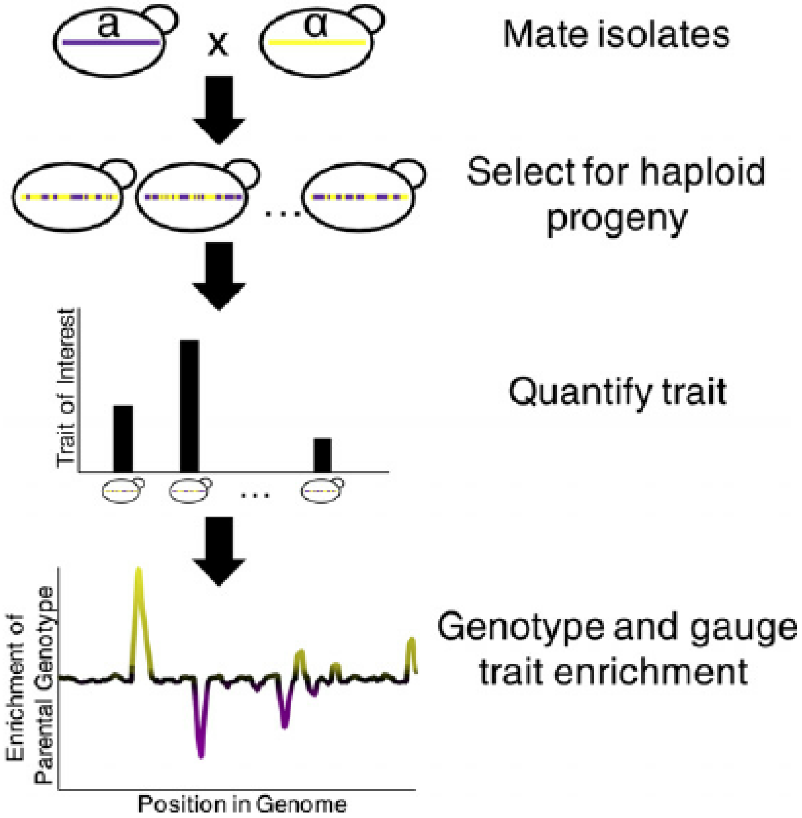
Genotype-Phenotype analysis

Quantitative trait loci

- Measure a trait in a population, or disease status

- Call variants

- Single-locus, biallelic, QTL
  - ANOVA (or Kruskal-Wallis) test of trait scores across variants
  - **Linked locus** can have similar significance → need biological knowledge



Mate isolates

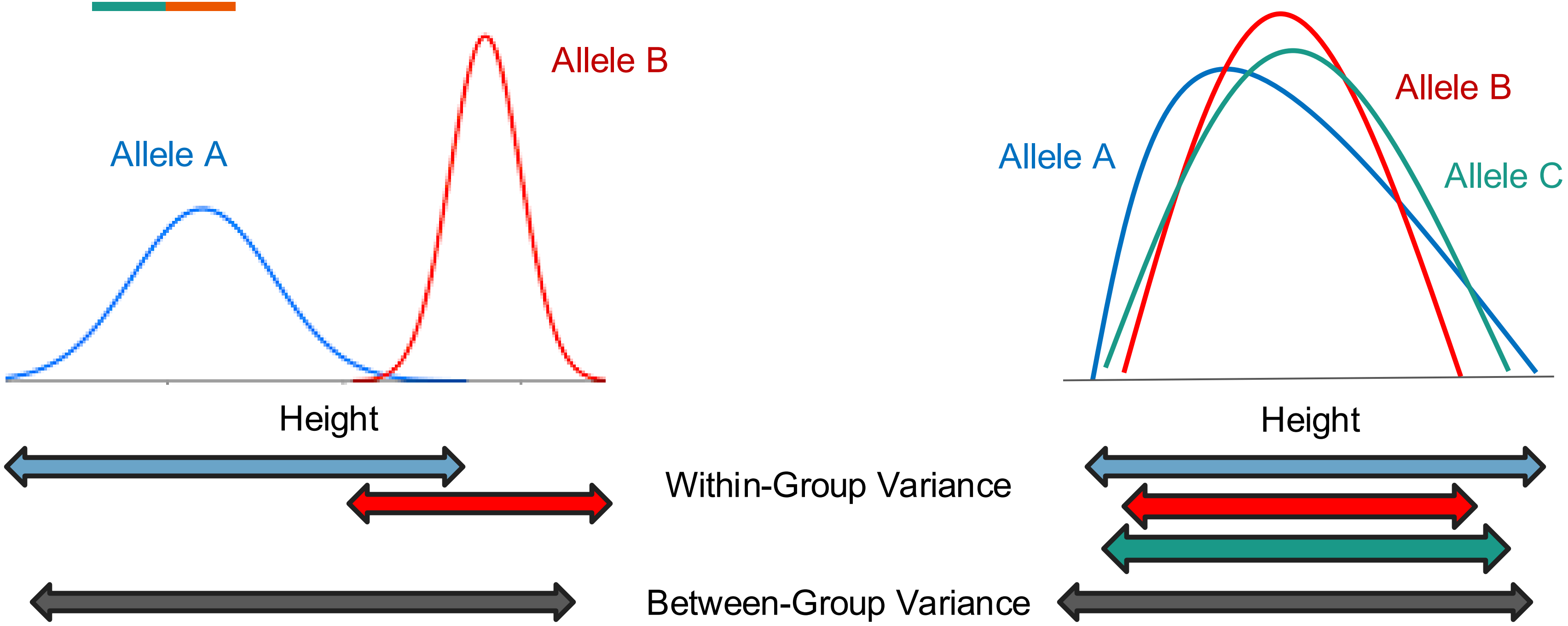
Select for haploid progeny

Quantify trait

Genotype and gauge trait enrichment

Hughes, T.R. and de Boer, C. Genetics 195:9-36

Analysis of Variant (ANOVA)



Within-Group Variance

Between-Group Variance

- **F-score** = Average Between-Group / Average Within-Group

Linear effect model

Alternative Model: Phenotype = Genotype x (effect size) + variance

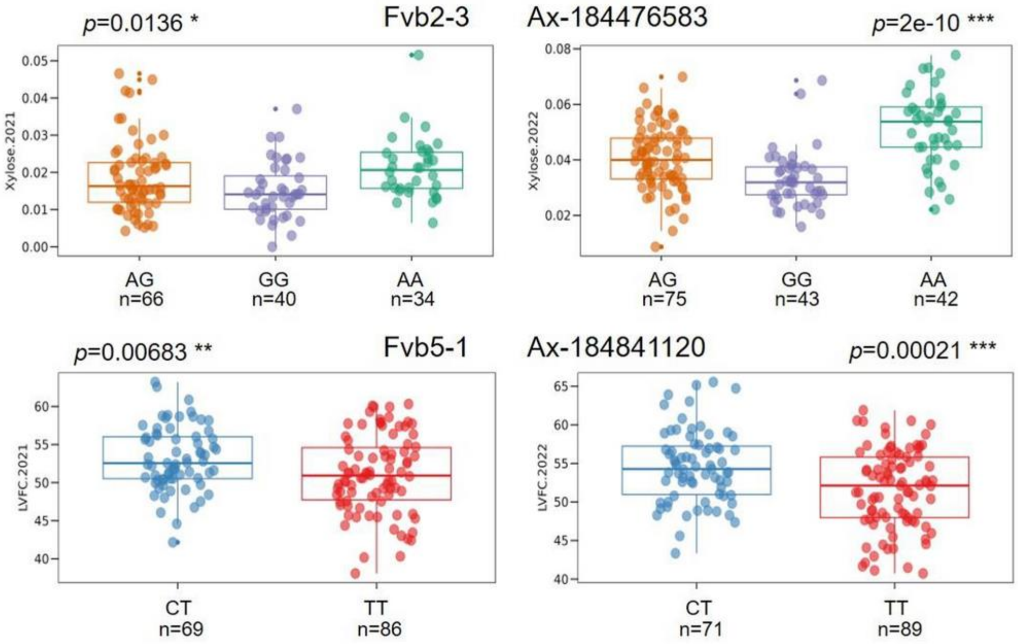
Null Model: Phenotype = (average effect size) + variance

- If **Alternative Model** can capture Phenotype significantly better than **Null Model**, then Genotype has some effects

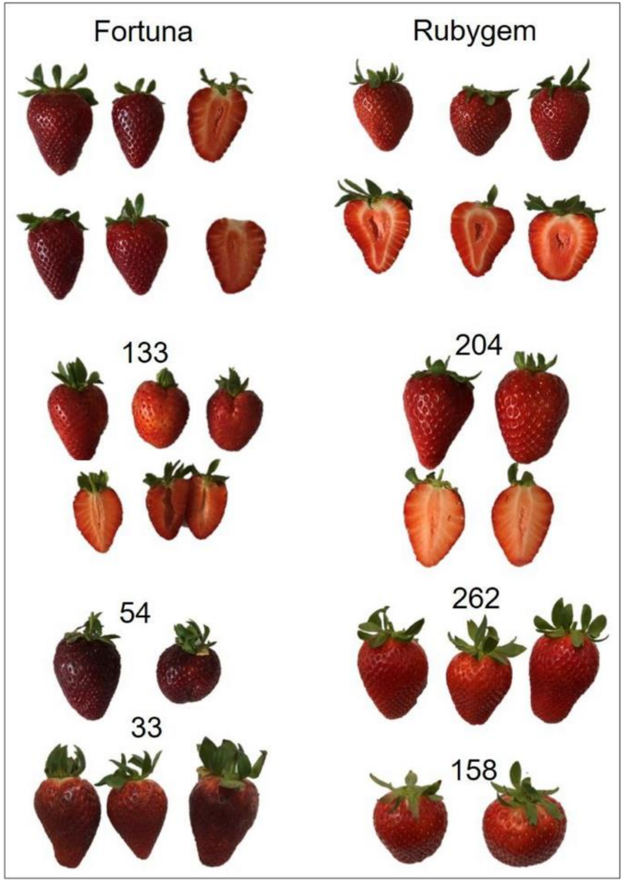
- If effect size is significantly different from **zero**, then Genotype has some effects

Example of QTL results

Genomic position



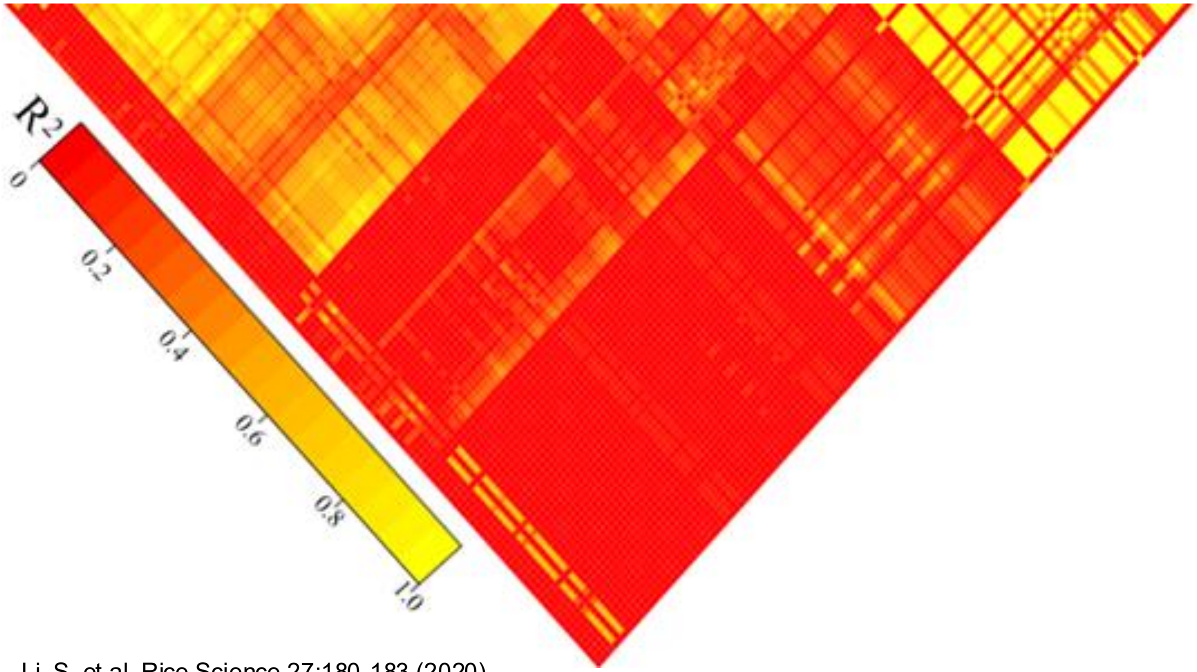
- QTL analyzed over two years



Urun, I, et al. Euphytica 221:48 (2025)

Linkage disequilibrium (LD)

Genomic position



Li, S, et al. Rice Science 27:180-183 (2020)

- Distal loci can appear together in a population with high frequency (linked)

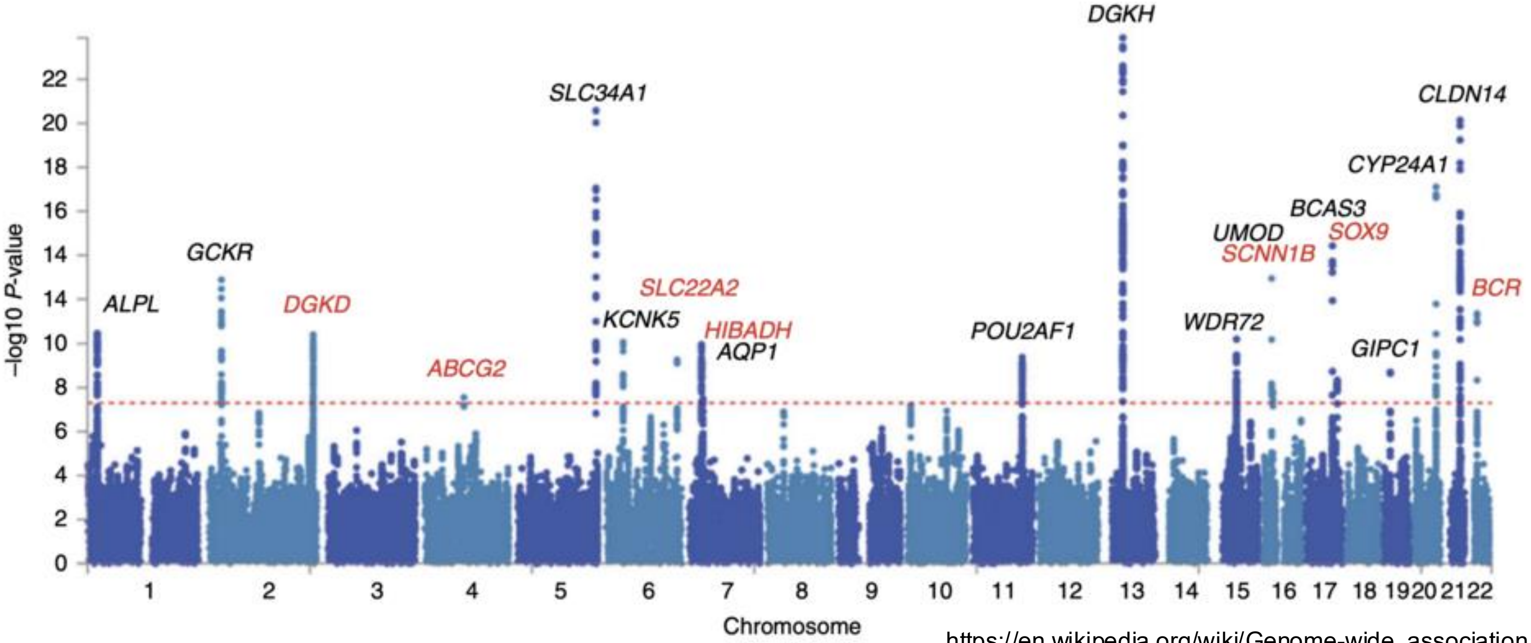
- Population analysis cannot distinguish between linked loci

A basic measurement for LD

Locus 1	Locus 2	Frequency
A	B	g_1
A	b	g_2
a	B	g_3
a	b	g_4

- Linkage between A-B = frequency of AB – expected frequency
  = g_1 – (frequency of A *times* frequency of B)
  = $g_1 - (g_1 + g_2)$ x $(g_1 + g_3)$

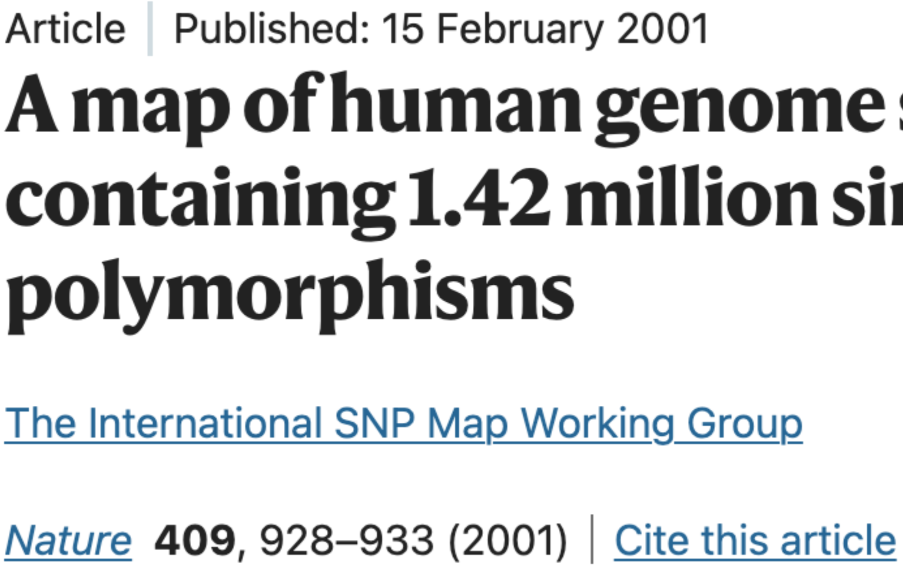
Genome-Wide Association Study (GWAS)



https://en.wikipedia.org/wiki/Genome-wide_association_study

- >100,000 association tests covering all loci

The scope for GWAS



Article | Published: 15 February 2001

A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms

The International SNP Map Working Group

Nature **409**, 928–933 (2001) | Cite this article

- Imagine having to perform statistical correction for 1.42 million tests!

P-value

- P-value cutoff of 0.05 means that **if there is no association, there is a 5% random chance that you will observe an association that is at least as extreme as the data**

Example scenario:

- Out of 100 patients with disease X, 40 have allele *aa*
- **Null Hypothesis**: No association between X and allele *a*
- If there's no association, we expect 25 patients with *aa*
- P-value = 0.0007 (binomial distribution)
- Significant!

Why must we correct p-value for multiple tests?

- P-value cutoff of 0.05 means that **if there is no association, there is a 5% random chance that you will observe an association that is at least as extreme as the data**

Example scenario:

- Perform 1.4 million ANOVA tests on all SNPs in human genome
- Let's say there is no association between disease and 1 million SNPs
- How many of of them would pass the threshold of 0.05 by chance?
  - 0.05 x 1,000,000 = 50,000

☹

P-value correction methods

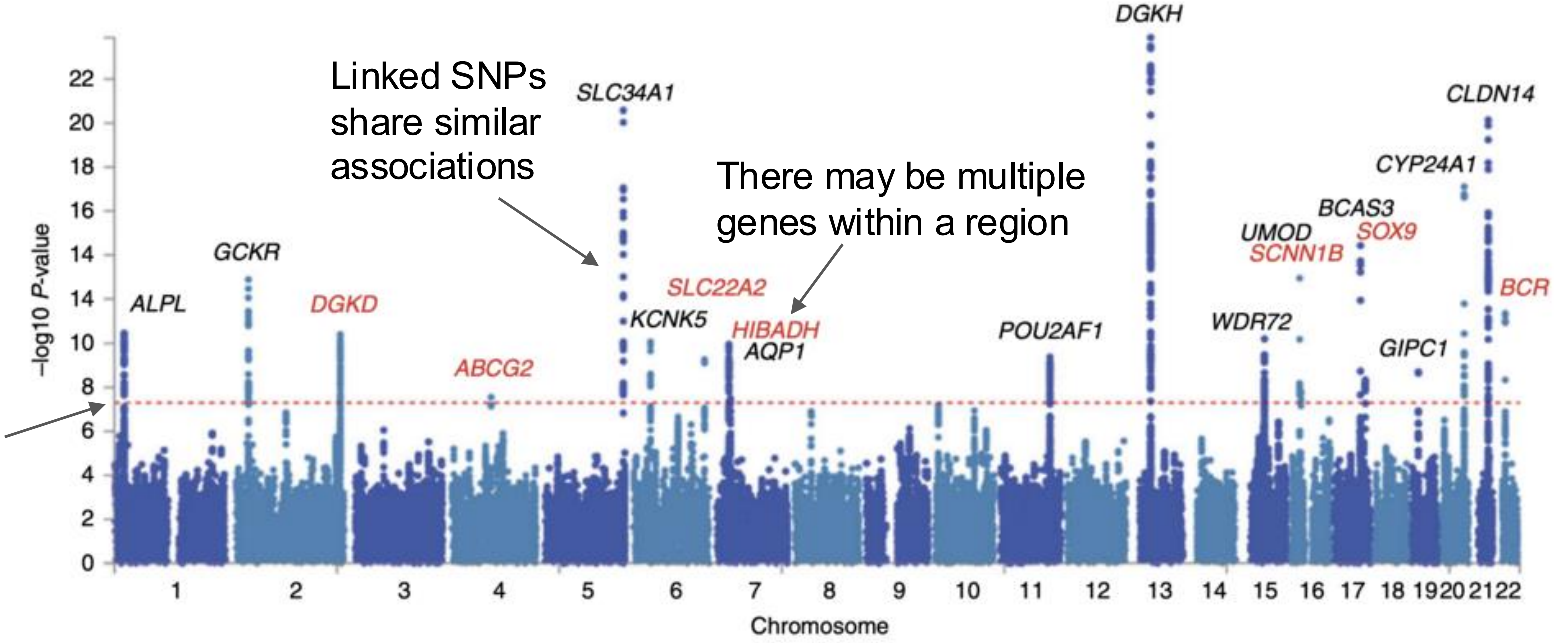
- **Bonferroni:** Divide the p-value threshold by the number of test

Example scenario:

- Perform 1.4 million ANOVA tests on all SNPs in human genome
- New p-value threshold = 0.05 / 1,400,000
- Let’s say there is no association between disease and 1 million SNPs
- How many of of them would pass the threshold by chance?
    - 0.05 / 1,400,000 x 1,000,000 = 0.036

☺

Manhatton Plot



Linked SNPs share similar associations

There may be multiple genes within a region

Adjusted p-value threshold

https://en.wikipedia.org/wiki/Genome-wide_association_study

Any question?

- See you next time