THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

CREATE CHANGE

Genetics and Genomics Winter School

Module 1- GWAS follow-up

June 2024

Fleur Garton

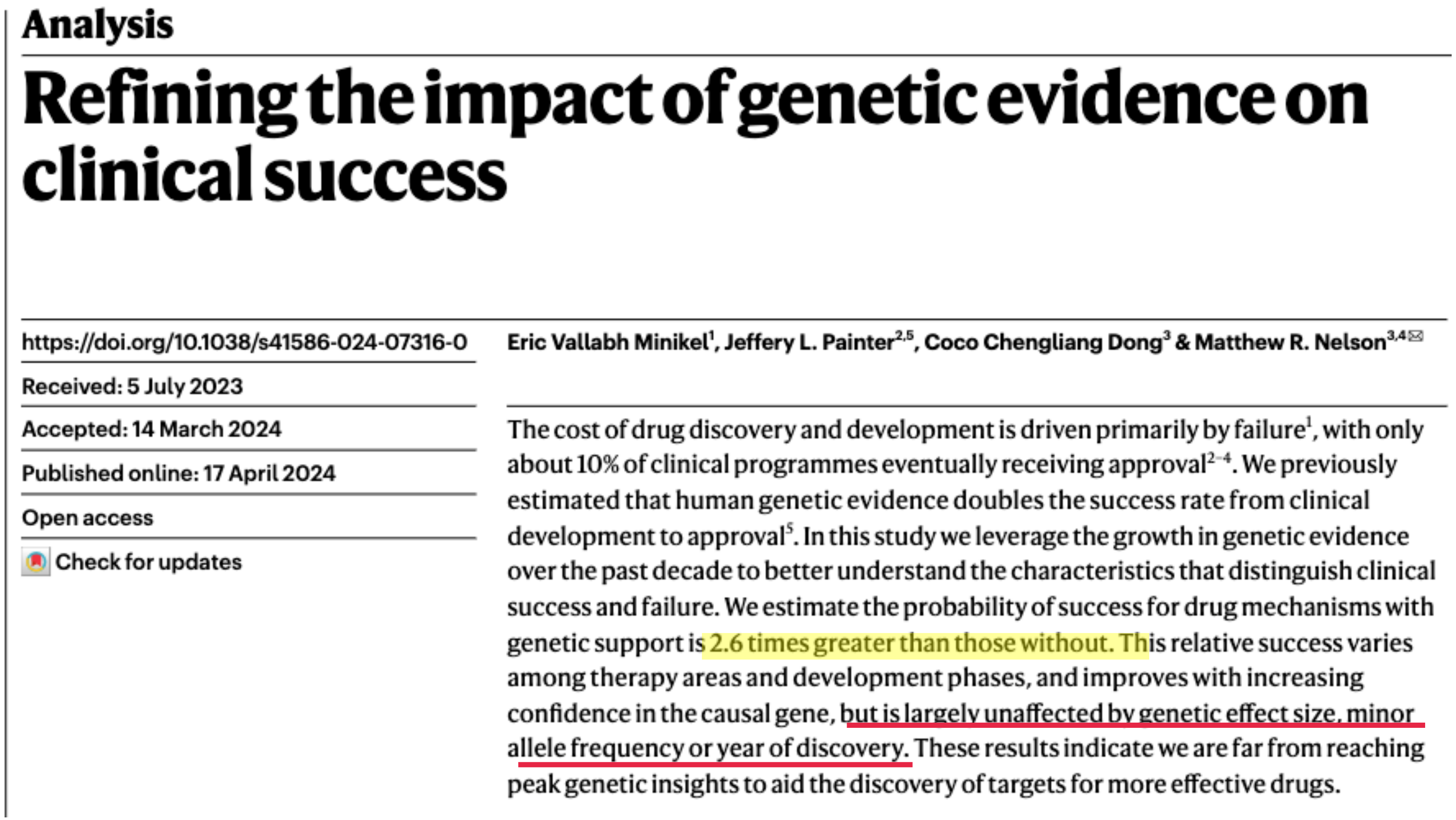
Introduction- Why

- Complex traits/diseases are generally highly polygenic

- “Significant loci” are regions of the genome

- To translate findings / biological insight a range of methods and complementary data can be used → covered in detail in **Module 6 - Systems Genomics and Pharmacogenomics**

- Huge area of growth – having an identified genetic links with disease (risk or cause) – is a significant predictor to success in the drug approval process (Nelson et al. 2015, Minikel et al. 2024)



Analysis

Refining the impact of genetic evidence on clinical success

https://doi.org/10.1038/s41586-024-07316-0

Received: 5 July 2023

Accepted: 14 March 2024

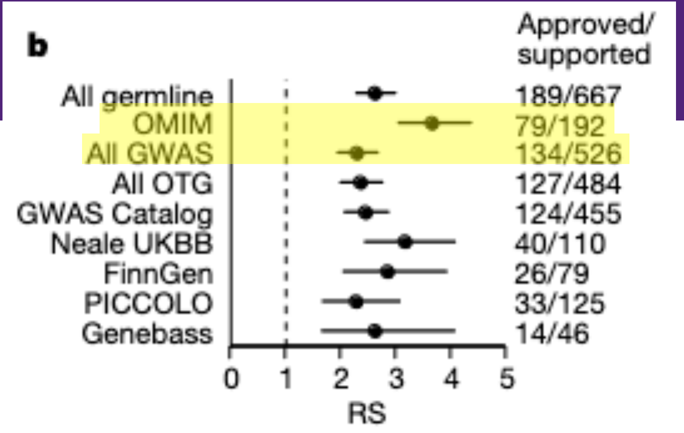
Published online: 17 April 2024

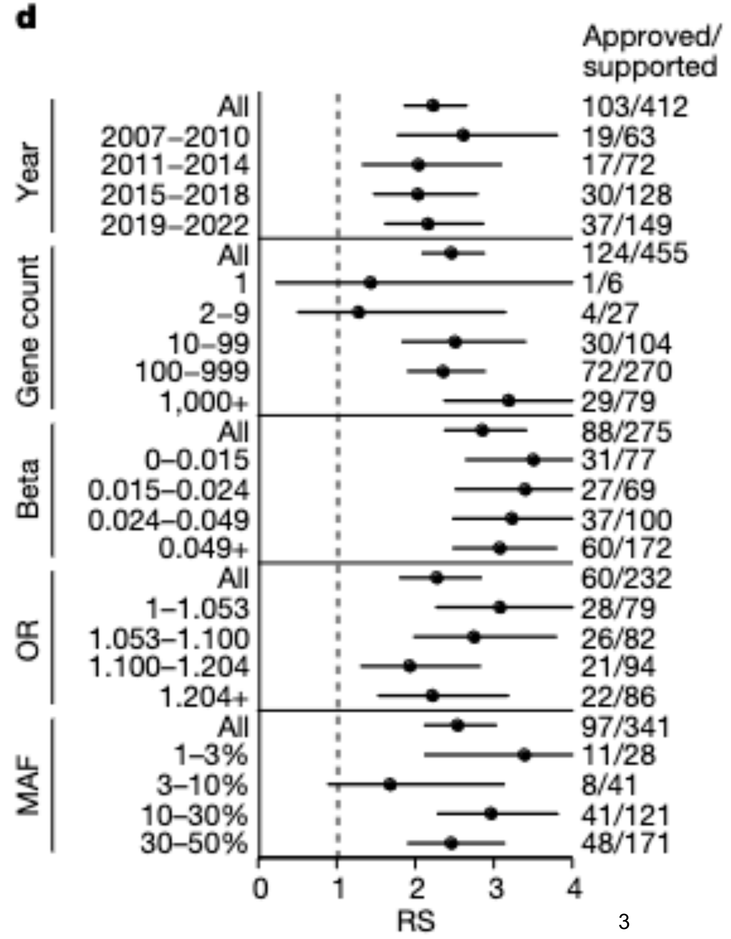
Open access

Check for updates

Eric Vallabh Minikel[1], Jeffery L. Painter[2,5], Coco Chengliang Dong[3] & Matthew R. Nelson[3,4] ✉

The cost of drug discovery and development is driven primarily by failure[1], with only about 10% of clinical programmes eventually receiving approval[2–4]. We previously estimated that human genetic evidence doubles the success rate from clinical development to approval[5]. In this study we leverage the growth in genetic evidence over the past decade to better understand the characteristics that distinguish clinical success and failure. We estimate the probability of success for drug mechanisms with genetic support is 2.6 times greater than those without. This relative success varies among therapy areas and development phases, and improves with increasing confidence in the causal gene, but is largely unaffected by genetic effect size, minor allele frequency or year of discovery. These results indicate we are far from reaching peak genetic insights to aid the discovery of targets for more effective drugs.





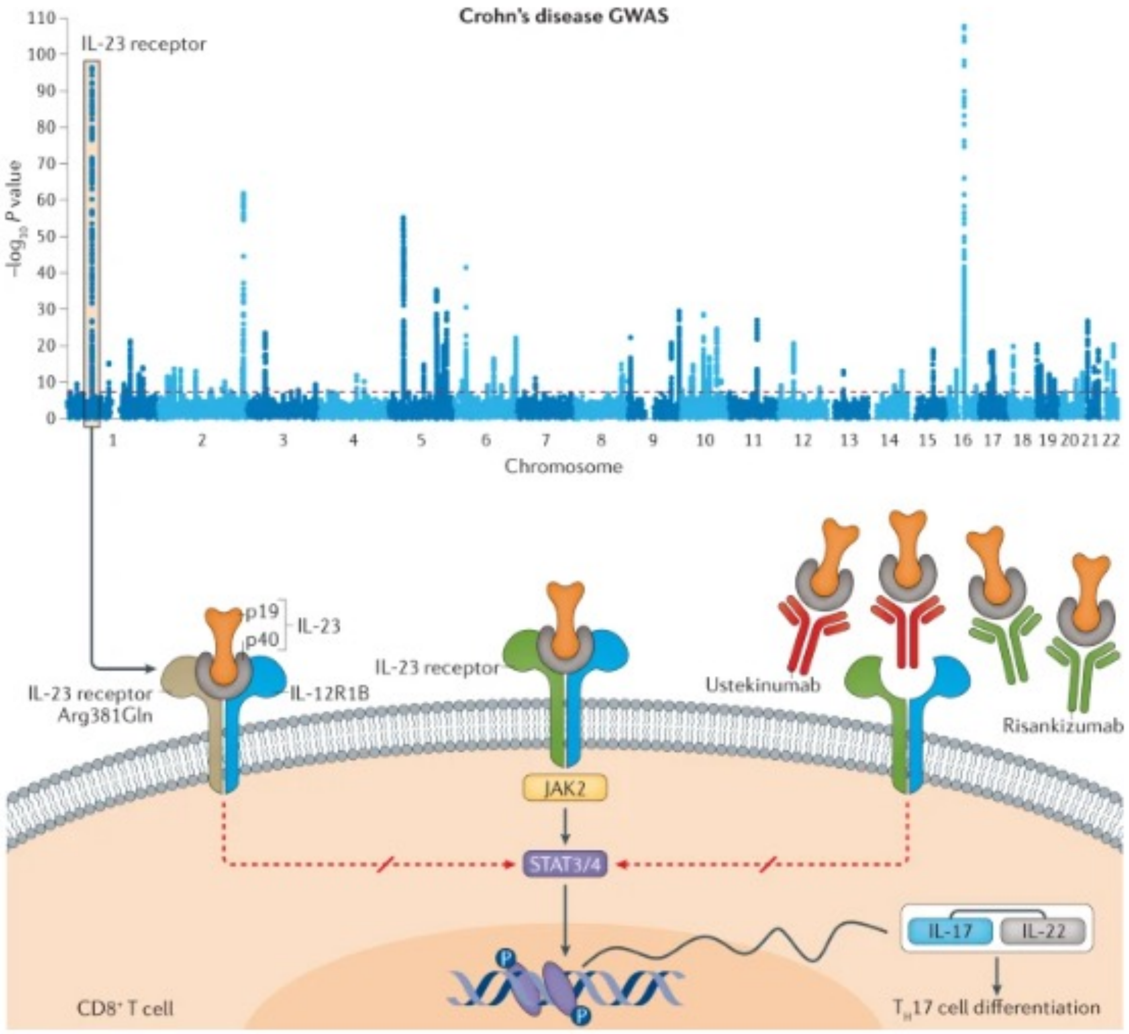
Aims

- Interrogate a locus that has been translated
- Understand ‘best practice’ nomenclature when describing human variation
- Be provided with tools and databases that support variant follow-up
- Carry out annotation in ANNOVAR for a list of variants

An example

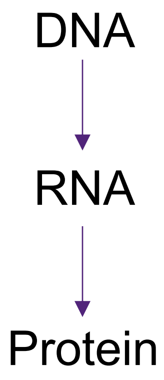
- Crohn’s disease GWAS
- One locus, top SNP, rs11209026
- Variant was coding (missense) in the IL23 receptor - protective effect in carriers
- Pharmacological inhibition of this gene of value to treat disease
- Two central monoclonal antibodies modulating IL-23 signalling were trialled -- ustekinumab and Risankizumab (psoriasis)
- Ustekinumab now approved in United States, Europe and Australia

Fig. 1: Genome-wide significant variants associated with Crohn’s disease spanning the IL-23 receptor provide drug repurposing opportunities.



Raey et al. 2021 https://www.nature.com/articles/s41576-021-00387-z

What do we mean when we say coding change?

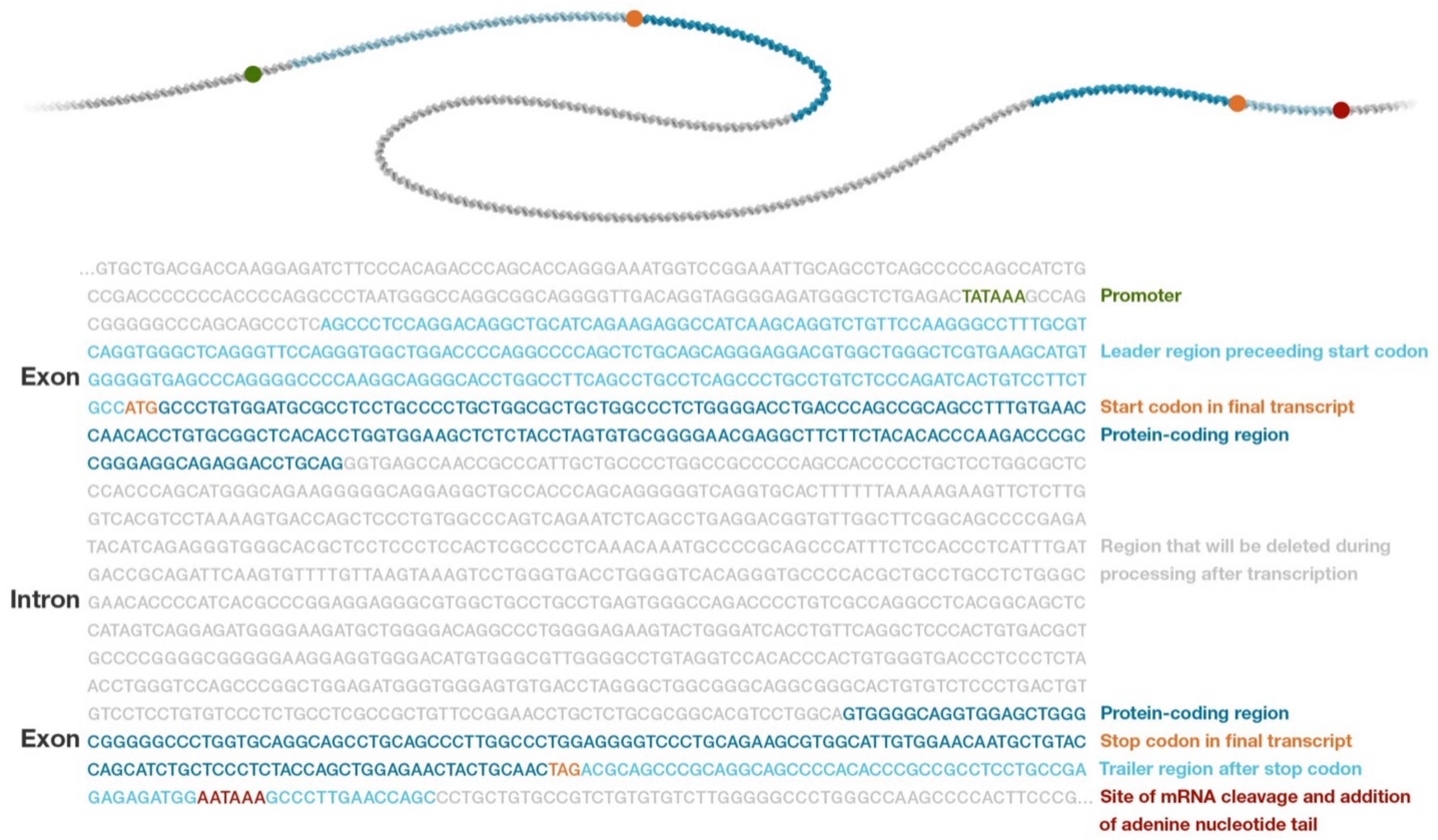


DNA

RNA

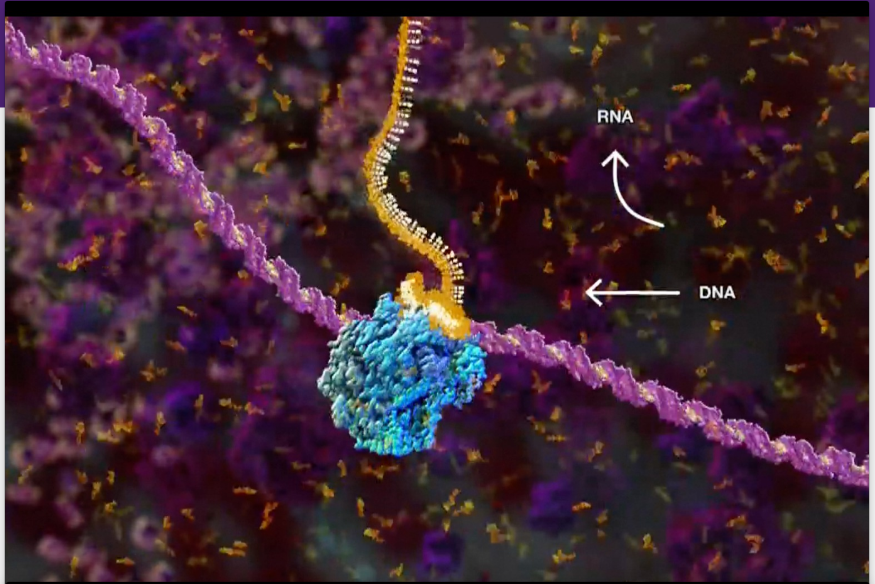
Protein

What do we mean when we say coding change?



```
      ...GTGCTGACGACCAAGGAGATCTTCCCACAGACCCAGCACCAGGGAAATGGTCCGGAAATTGCAGCCTCAGCCCCCAGCCATCTG
         CCGACCCCCCCACCCCAGGCCCTAATGGGCCAGGCGGCAGGGGTTGACAGGTAGGGGAGATGGGCTCTGAGACTATAAAGCCAG   Promoter
         CGGGGGCCCAGCAGCCCTCAGCCCTCCAGGACAGGCTGCATCAGAAGAGGCCATCAAGCAGGTCTGTTCCAAGGGCCTTTGCGT
         CAGGTGGGCTCAGGGTTCCAGGGTGGCTGGACCCCAGGCCCCAGCTCTGCAGCAGGGAGGACGTGGCTGGGCTCGTGAAGCATGT   Leader region preceeding start codon
Exon     GGGGGTGAGCCCAGGGGCCCCAAGGCAGGGCACCTGGCCTTCAGCCTGCCTCAGCCCTGCCTGTCTCCCAGATCACTGTCCTTCT
         GCCATGGCCCTGTGGATGCGCCTCCTGCCCCTGCTGGCGCTGCTGGCCCTCTGGGGACCTGACCCAGCCGCAGCCTTTGTGAAC   Start codon in final transcript
         CAACACCTGTGCGGCTCACACCTGGTGGAAGCTCTCTACCTAGTGTGCGGGGAACGAGGCTTCTTCTACACACCCAAGACCCGC   Protein-coding region
         CGGGAGGCAGAGGACCTGCAGGGTGAGCCAACCGCCCATTGCTGCCCCTGGCCGCCCCCAGCCACCCCCTGCTCCTGGCGCTC
         CCACCCAGCATGGGCAGAAGGGGGCAGGAGGCTGCCACCCAGCAGGGGGTCAGGTGCACTTTTTTAAAAAGAAGTTCTCTTG
         GTCACGTCCTAAAAGTGACCAGCTCCCTGTGGCCCAGTCAGAATCTCAGCCTGAGGACGGTGTTGGCTTCGGCAGCCCCGAGA
         TACATCAGAGGGTGGGCACGCTCCTCCCTCCACTCGCCCCTCAAACAAATGCCCCGCAGCCCATTTCTCCACCCTCATTTGAT   Region that will be deleted during
         GACCGCAGATTCAAGTGTTTTGTTAAGTAAAGTCCTGGGTGACCTGGGGTCACAGGGTGCCCCACGCTGCCTGCCTCTGGGC   processing after transcription
Intron   GAACACCCCATCACGCCCGGAGGAGGGCGTGGCTGCCTGCCTGAGTGGGCCAGACCCCTGTCGCCAGGCCTCACGGCAGCTC
         CATAGTCAGGAGATGGGGAAGATGCTGGGGACAGGCCCTGGGGAGAAGTACTGGGATCACCTGTTCAGGCTCCCACTGTGACGCT
         GCCCCGGGGCGGGGGAAGGAGGTGGGACATGTGGGCGTTGGGGCCTGTAGGTCCACACCCACTGTGGGTGACCCTCCCTCTA
         ACCTGGGTCCAGCCCGGCTGGAGATGGGTGGGAGTGTGACCTAGGGCTGGCGGGCAGGCGGGCACTGTGTCTCCCTGACTGT
         GTCCTCCTGTGTCCCTCTGCCTCGCCGCTGTTCCGGAACCTGCTCTGCGCGGCACGTCCTGGCAGTGGGGCAGGTGGAGCTGGG   Protein-coding region
Exon     CGGGGGCCCTGGTGCAGGCAGCCTGCAGCCCTTGGCCCTGGAGGGGTCCCTGCAGAAGCGTGGCATTGTGGAACAATGCTGTAC   Stop codon in final transcript
         CAGCATCTGCTCCCTCTACCAGCTGGAGAACTACTGCAACTAGACGCAGCCCGCAGGCAGCCCCACACCCGCCGCCTCCTGCCGA   Trailer region after stop codon
         GAGAGATGGAATAAAGCCCTTGAACCAGCCCTGCTGTGCCGTCTGTGTGTCTTGGGGGCCCTGGGCCAAGCCCCACTTCCCG...   Site of mRNA cleavage and addition
                                                                                                 of adenine nucleotide tail
```

Transcription



RNA
DNA
G
G
T
A
A
G
C
C
RNA polymerase
transcription mechanism

Majority of bases are associated with at least one primary transcript

Chromatin accessibility and histone-modification patterns are highly predictive of both the presence and activity of transcription start sites.

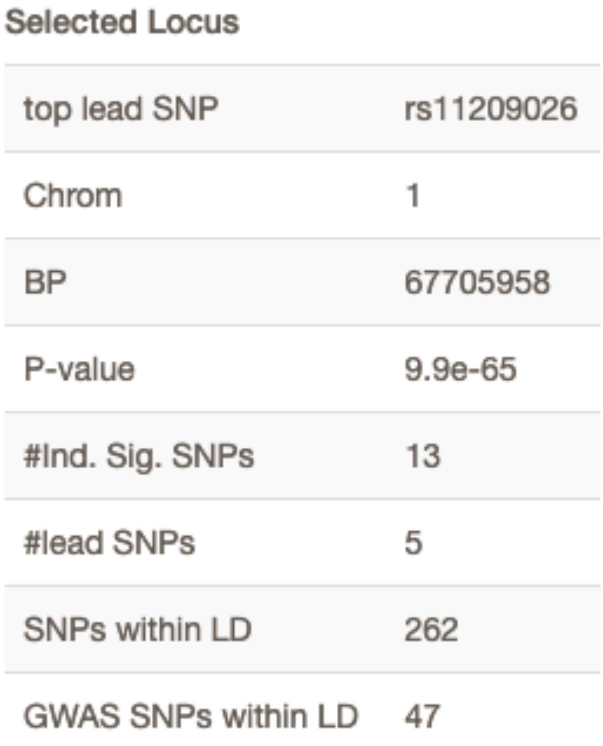
DNA-replication timing is correlated with chromatin structure.

Transcription controls;

- RNA polymerase cannot initiate transcription on their own; require regulatory factors, such as a promoter.
- Promotors are recognised and bound by transcription factors that guide and activate the RNA polymerase
- Transcription factors, act in *trans*, because they are produced by remote genes and then need to mitigate to sites of action
- Promoters are *cis*-acting because they located near the transcriptional start site
- Enhancer/silencer= a cluster of *cis*-acting short sequence elements that can alter the transcriptional activity of a gene

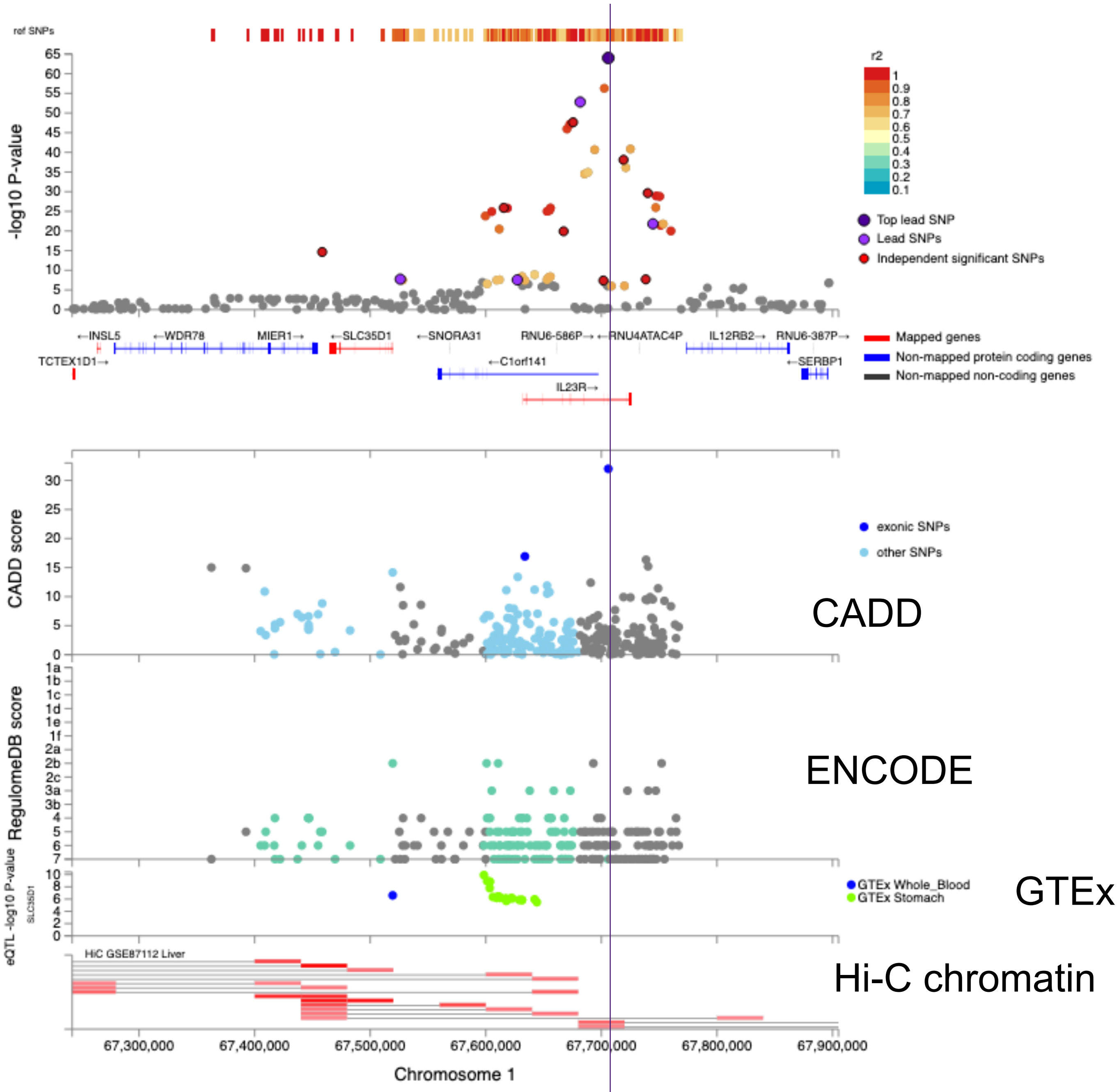
https://www.wehi.edu.au/wehi-tv/dna-central-dogma-part-1-transcription

The IL-23R locus in more detail….



Selected Locus

top lead SNP	rs11209026
Chrom	1
BP	67705958
P-value	9.9e-65
#Ind. Sig. SNPs	13
#lead SNPs	5
SNPs within LD	262
GWAS SNPs within LD	47



CADD

ENCODE

GTEx

Hi-C chromatin

Annotation support/scoring

In-silico prediction - evolving field

Meta-tools perform better (i.e. more sensitive) than a single score i.e. conservation

Fewer tools that score non-coding variants – (rely instead on regulatory data)

CADD - Combined Annotation Dependant Depletion (2014..updated)- based on diverse genomic features derived from surrounding sequence context, gene model annotations, evolutionary constraint, epigenetic measurements and functional predictions. Includes splice version and hg38 update.

VEP - Variant Effect Predictor (2016) - VEP determines the effect of your variants (SNPs, insertions, deletions, CNVs or structural variants) on genes, transcripts, and protein sequence, as well as regulatory regions.

BayesDel (2017..updated)- is a deleteriousness meta-score. It works for coding and non-coding variants, single nucleotide variants and small insertion / deletions. With and without allele frequency.

REVEL (2016) - (rare exome variant ensemble learner), an ensemble method for predicting the pathogenicity of missense variants on the basis of individual tools: MutPred, FATHMM, VEST, PolyPhen, SIFT, PROVEAN, MutationAssessor, MutationTaster, LRT, GERP, SiPhy, phyloP, and phastCons.

Alphamissense (2023)- a deep learning model that builds on the protein structure prediction tool AlphaFold2. Model is trained on population frequency data and uses sequence and predicted structural context, all of which contribute to its performance.

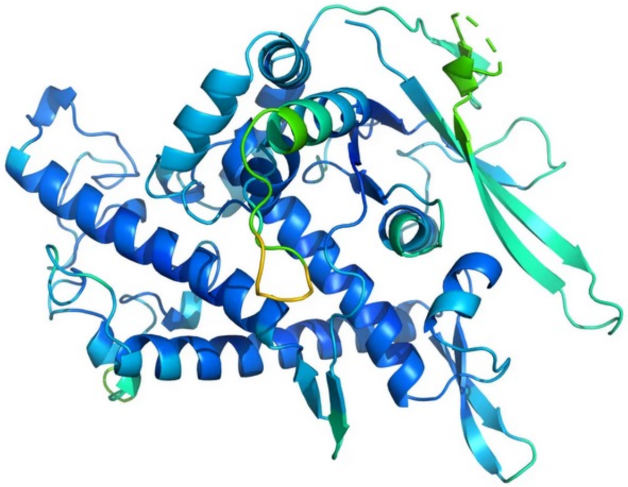
View the protein in 3D

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

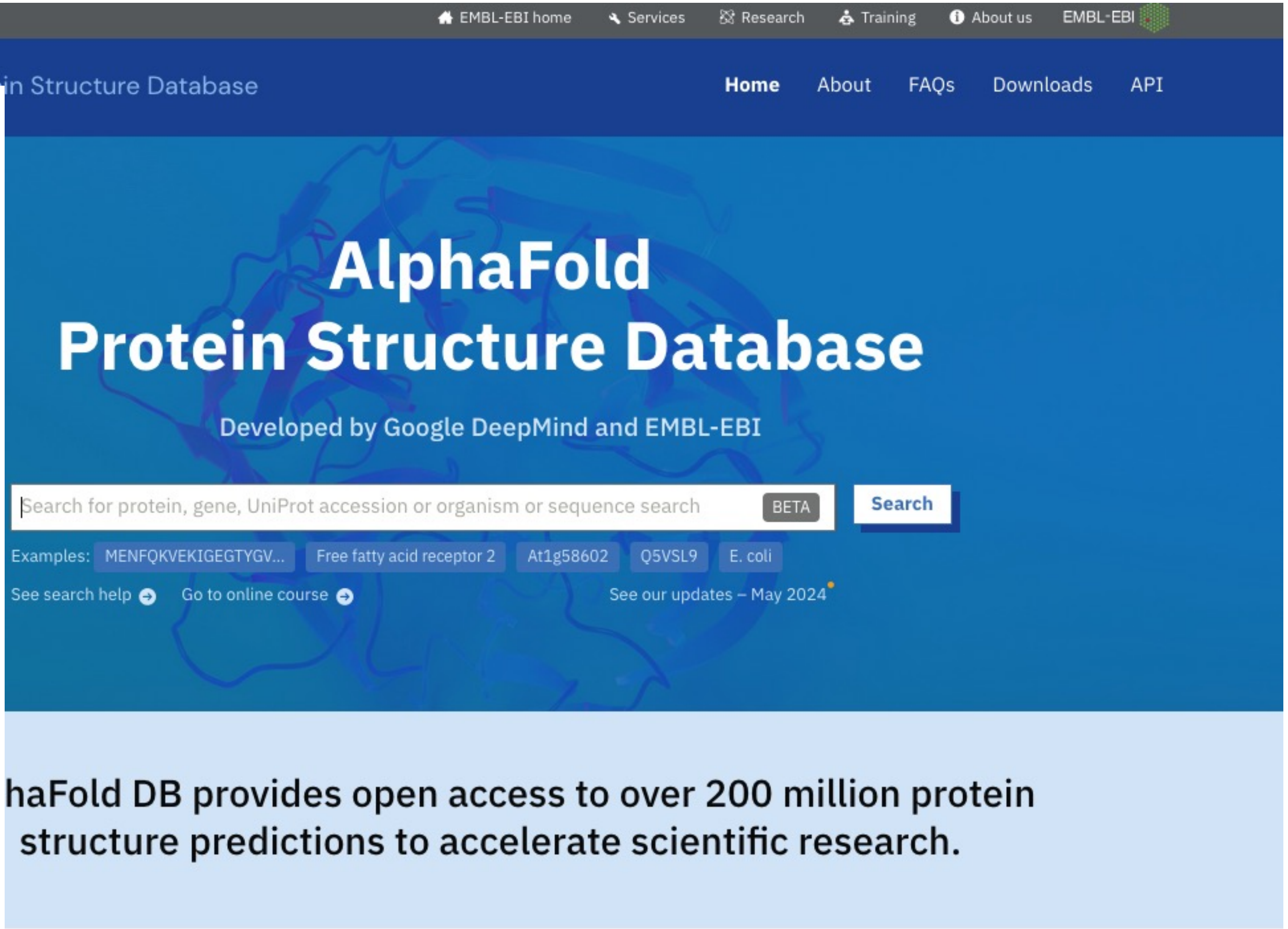
London A.I. Lab Claims Breakthrough That Could Accelerate Drug Discovery

Researchers at DeepMind say they have solved “the protein folding problem,” a task that has bedeviled scientists for more than 50 years.

Share full article 86



A computer model of folded protein targets studied by the DeepMind scientists. DeepMind



EMBL-EBI home
Services
Research
Training
About us
EMBL-EBI
AlphaFold Protein Structure Database
Home
About
FAQs
Downloads
API

AlphaFold
Protein Structure Database

Developed by Google DeepMind and EMBL-EBI

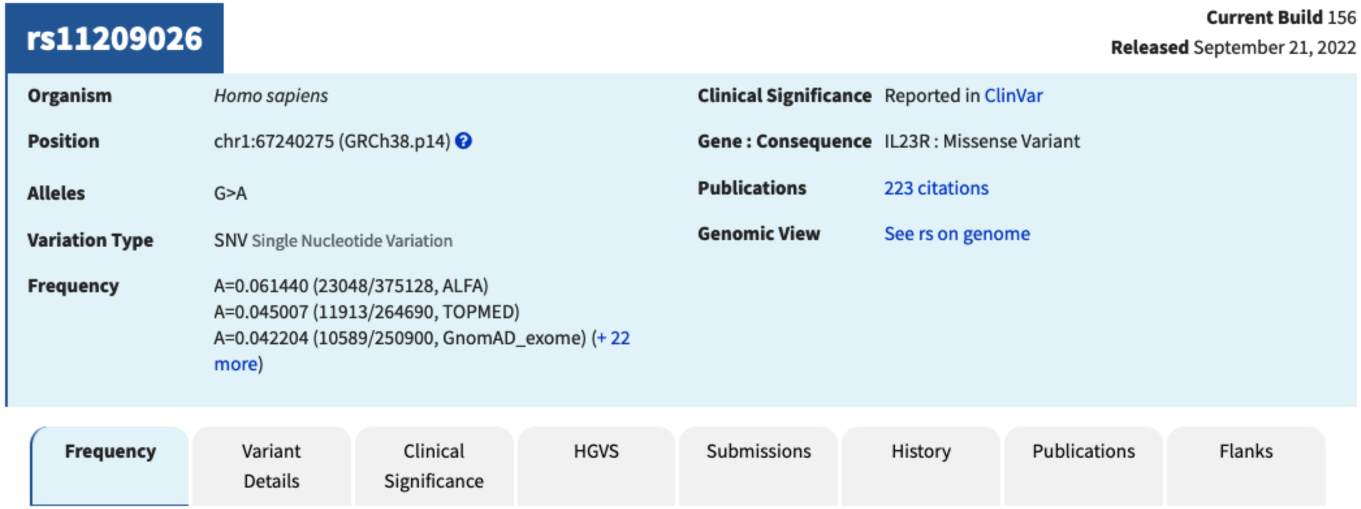
Search for protein, gene, UniProt accession or organism or sequence search
BETA
Search
Examples: MENFQKVEKIGEGTYGV... Free fatty acid receptor 2 At1g58602 Q5VSL9 E. coli
See search help
Go to online course
See our updates – May 2024

haFold DB provides open access to over 200 million protein structure predictions to accelerate scientific research.

https://alphafold.ebi.ac.uk/

dbSNP rs11209026

Perform lookup using dbSNP

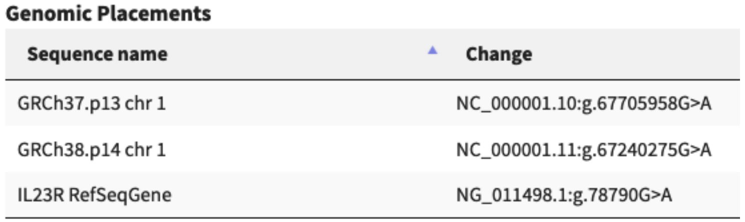


rs11209026

Current Build 156
Released September 21, 2022

Organism	*Homo sapiens*	**Clinical Significance**	Reported in ClinVar
Position	chr1:67240275 (GRCh38.p14)	**Gene : Consequence**	IL23R : Missense Variant
Alleles	G>A	**Publications**	223 citations
Variation Type	SNV Single Nucleotide Variation	**Genomic View**	See rs on genome

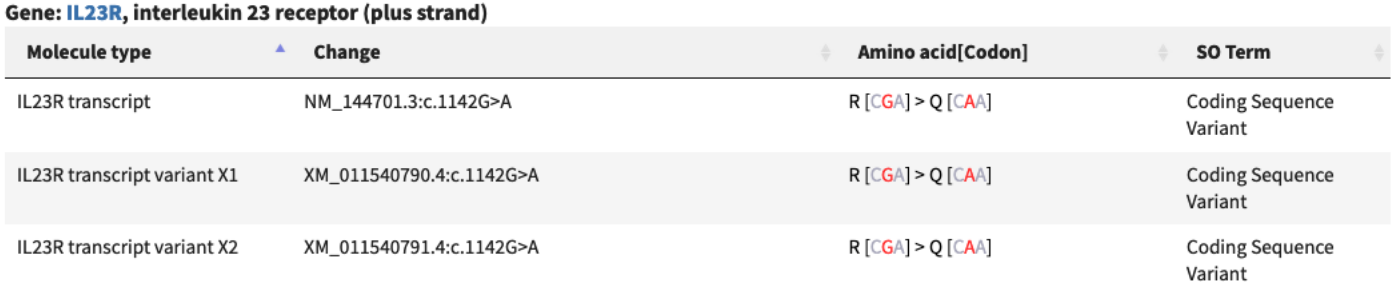
Frequency	A=0.061440 (23048/375128, ALFA)
	A=0.045007 (11913/264690, TOPMED)
	A=0.042204 (10589/250900, GnomAD_exome) (+ 22 more)		



Genomic Placements

Sequence name	Change
GRCh37.p13 chr 1	NC_000001.10:g.67705958G>A
GRCh38.p14 chr 1	NC_000001.11:g.67240275G>A
IL23R RefSeqGene	NG_011498.1:g.78790G>A

IL23R(NM_144701.3):c.1142G>A p.(Arg381Gln)

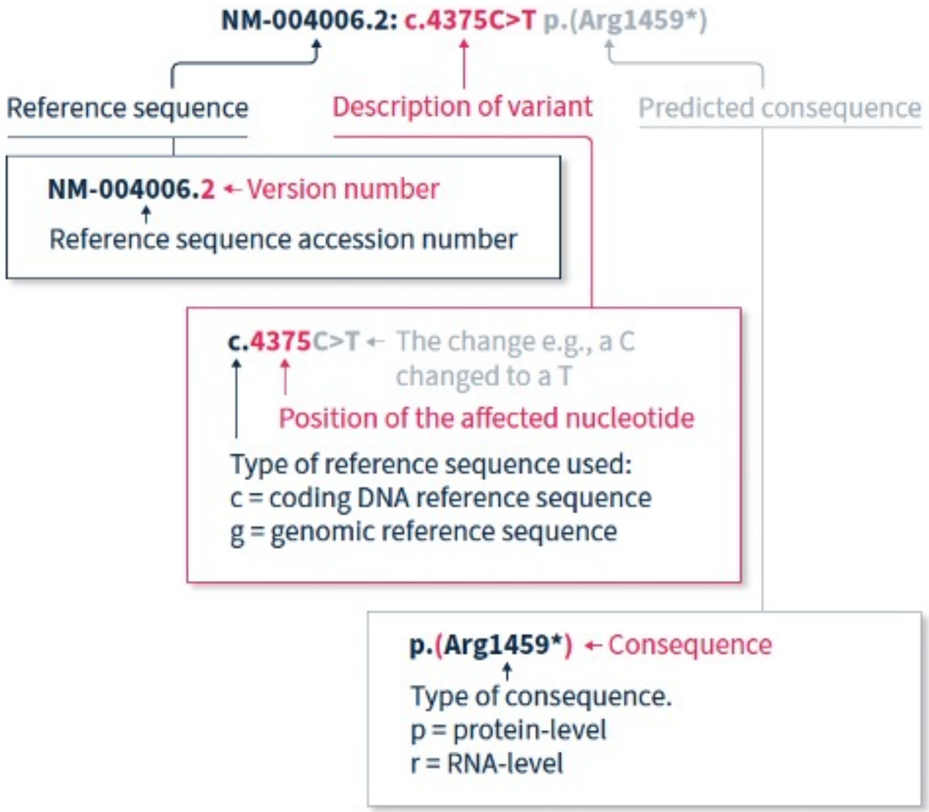


Gene: IL23R, interleukin 23 receptor (plus strand)

Molecule type	Change	Amino acid[Codon]	SO Term
IL23R transcript	NM_144701.3:c.1142G>A	R [CGA] > Q [CAA]	Coding Sequence Variant
IL23R transcript variant X1	XM_011540790.4:c.1142G>A	R [CGA] > Q [CAA]	Coding Sequence Variant
IL23R transcript variant X2	XM_011540791.4:c.1142G>A	R [CGA] > Q [CAA]	Coding Sequence Variant

Reading and writing

The HGVS recommendations for mutation nomenclature state that the format of a complete variant description should first include the reference sequence, followed by the variant description, and then the predicted consequence in parentheses. For example, NM-004006.2:c.4375C>T p.(Arg1459*) (**Figure 1**).



NM-004006.2: c.4375C>T p.(Arg1459*)

Reference sequence

Description of variant

Predicted consequence

NM-004006.2 ← Version number

Reference sequence accession number

c.4375C>T ← The change e.g., a C changed to a T

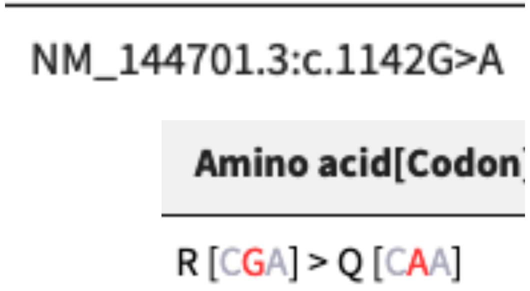
Position of the affected nucleotide

Type of reference sequence used:
c = coding DNA reference sequence
g = genomic reference sequence

p.(Arg1459*) ← Consequence

Type of consequence.
p = protein-level
r = RNA-level

Figure 1. Application of the HGVS nomenclature recommendations for sequence variants



NM_144701.3:c.1142G>A

Amino acid[Codon]
R [CGA] > Q [CAA]

Other useful nomenclature - transcripts

Transcript IDs

Multiple transcripts exist for a single gene / the longest transcript has *traditionally been chosen as the reference*, MANE Select transcripts and APPRIS principal transcripts are the best reference transcripts for clinical variation.

MANE: Matched Annotation from the NCBI and EMBL-EBI (MANE) converge on human gene + transcript →define a GW set of representative transcripts and corresponding proteins for human protein-coding genes.

Each MANE transcript represents an exact match in exonic regions between a Refseq transcript and its counterpart in the Ensembl/GENCODE annotation such that the two identifiers can be used synonymously.

MANE Select: The MANE Select set consists of one transcript at each protein-coding locus across the genome that is representative of biology at that locus.

MANE Plus Clinical: The MANE Plus Clinical set includes additional transcripts for genes where MANE Select alone is not sufficient to report all "Pathogenic (P)" or "Likely Pathogenic (LP)" clinical variants available in public resources.

RefSeq= A comprehensive, integrated, non-redundant, well-annotated set of reference sequences including genomic, transcript, and protein (eg. NG_029916.1)
https://www.ncbi.nlm.nih.gov/refseq/rsg/

Other useful nomenclature - Genes

input identifiers (Entrez Gene ID, RefSeq, Ensembl ID, UnProt ID or Symbol)

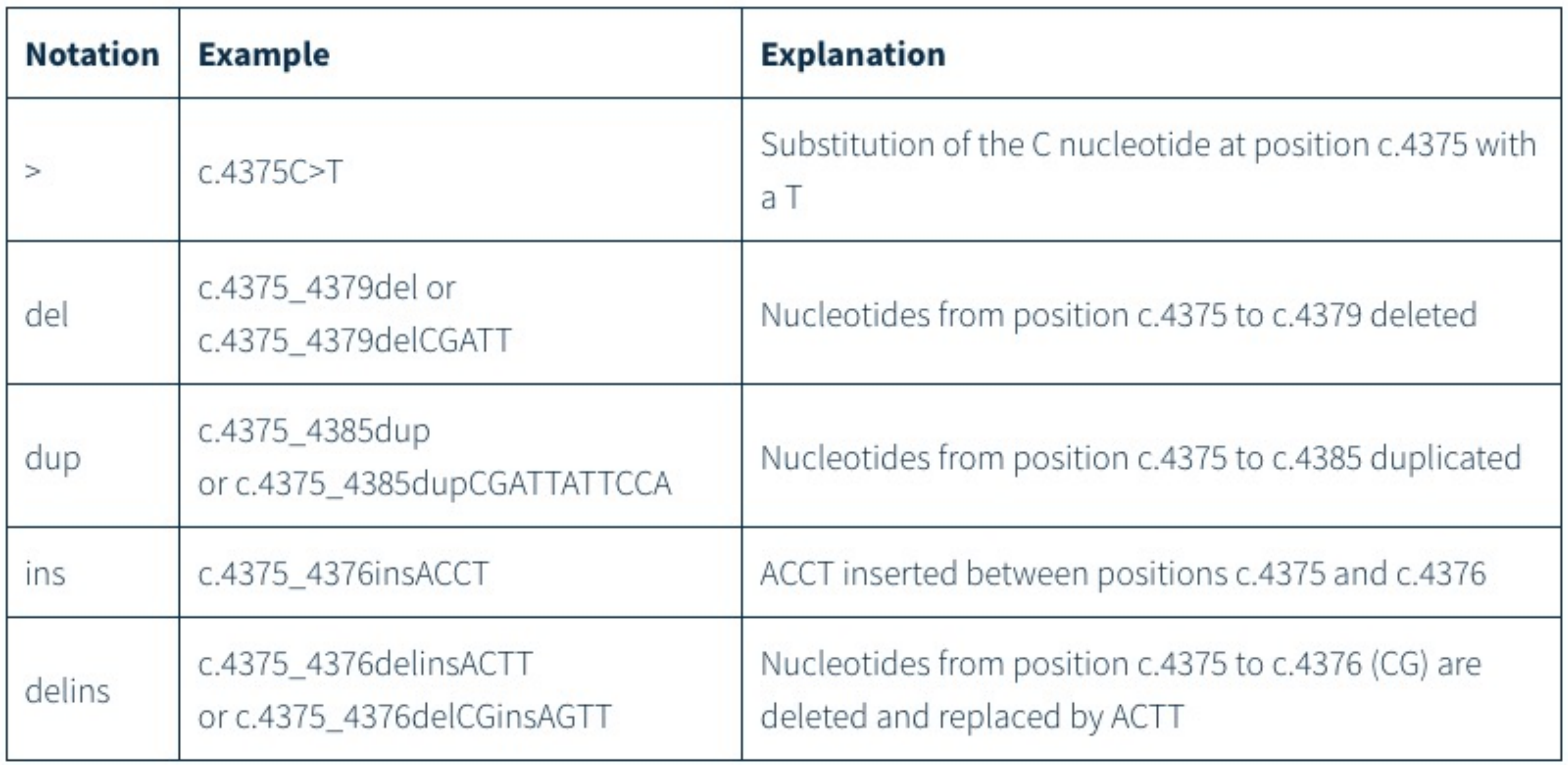
Symbol: Letters generally HUGO gene nomenclature committee (*IL23R*) and a full name

Entrez = unique integer identifiers for genes and other loci (such as officially named mapped markers) for a subset of model organisms (149233)

Ensembl = An Ensembl stable ID consists of five parts: ENS(species)(object type)(identifier)(version). Humans don't have a species code. (e.g. ENSG00000162594.17)

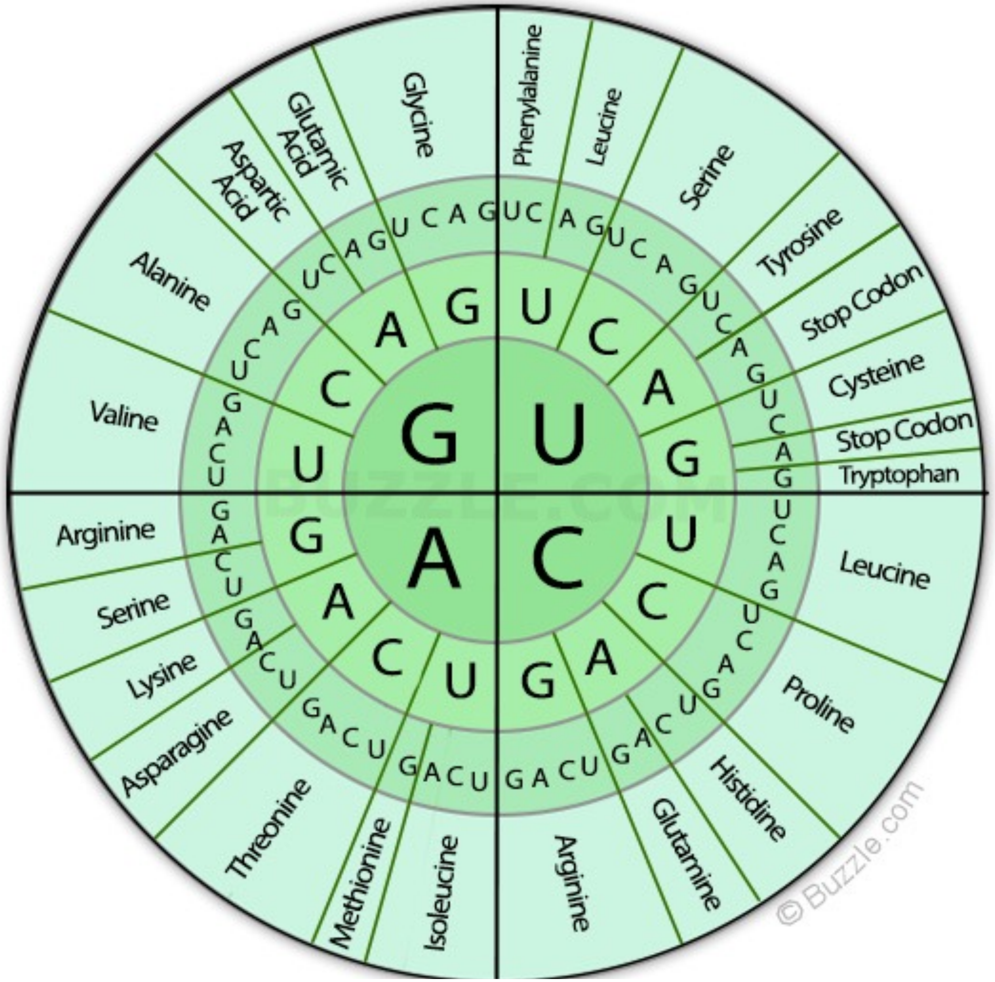
RefSeq= A comprehensive, integrated, non-redundant, well-annotated set of reference sequences including genomic, transcript, and protein (eg. NG_029916.1) https://www.ncbi.nlm.nih.gov/refseq/rsg/

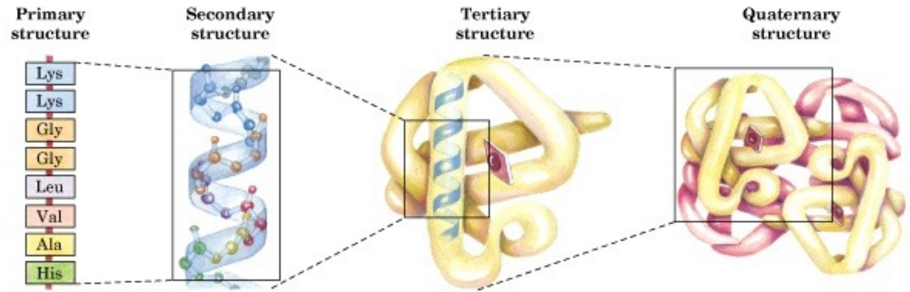
Other notation examples



Notation	Example	Explanation
>	c.4375C>T	Substitution of the C nucleotide at position c.4375 with a T
del	c.4375_4379del or c.4375_4379delCGATT	Nucleotides from position c.4375 to c.4379 deleted
dup	c.4375_4385dup or c.4375_4385dupCGATTATTCCA	Nucleotides from position c.4375 to c.4385 duplicated
ins	c.4375_4376insACCT	ACCT inserted between positions c.4375 and c.4376
delins	c.4375_4376delinsACTT or c.4375_4376delCGinsAGTT	Nucleotides from position c.4375 to c.4376 (CG) are deleted and replaced by ACTT

Amino Acid Wheel

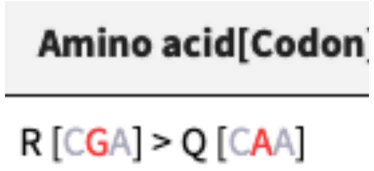




Start from the centre
Follow the RNA codons - 3 bases.
1 Amino acid from the mRNA codons.
(RNA translation)

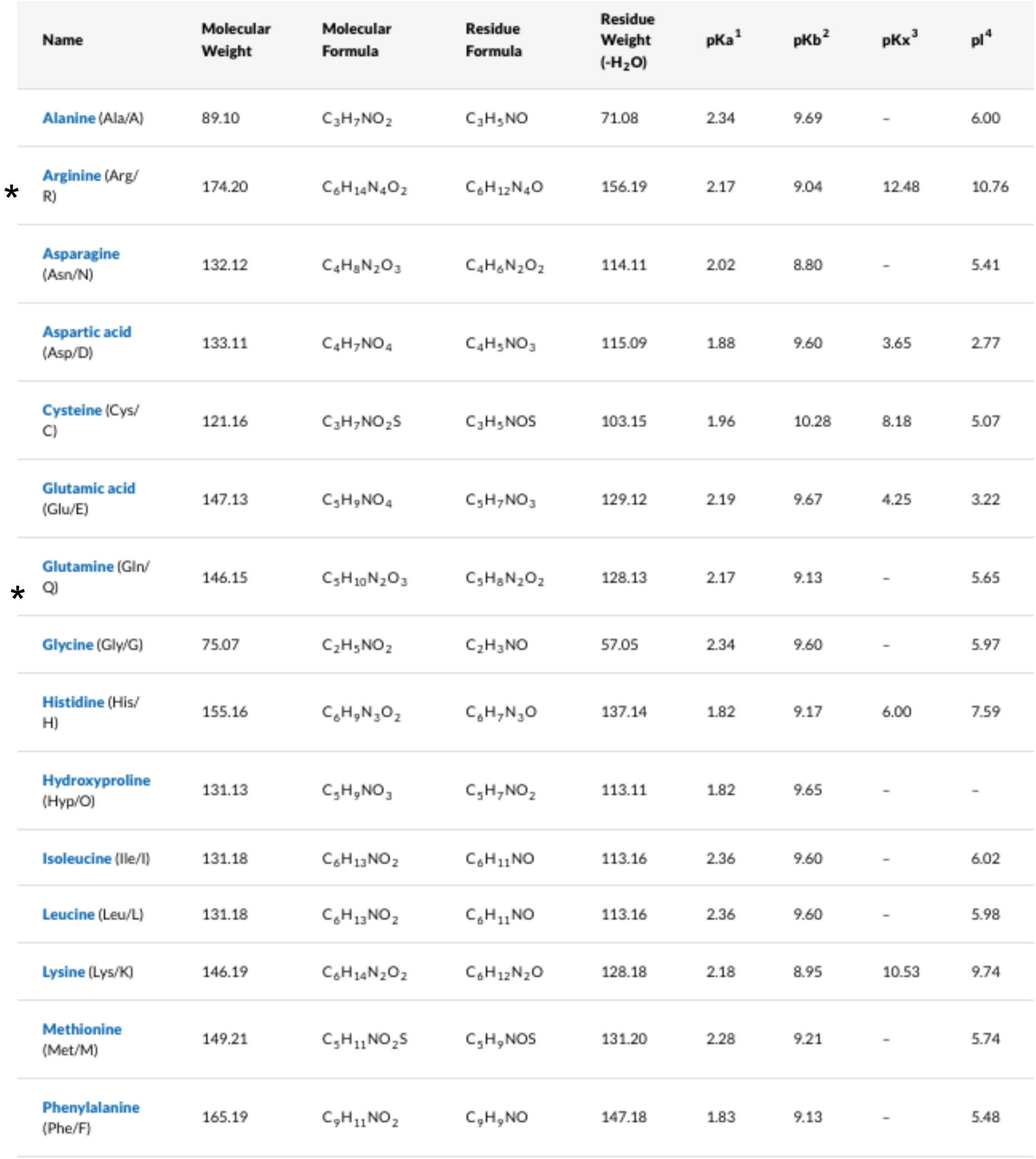
4 possible options (G, U, A, C)
4^3 codon multiples= 64..
Only 20 amino acids- (*Arginine has 6 combinations*)

IL23R(NM_144701.3):c.1142G>A
p.(Arg381Gln)



Amino acid[Codon
R [CGA] > Q [CAA]

AA properties- example

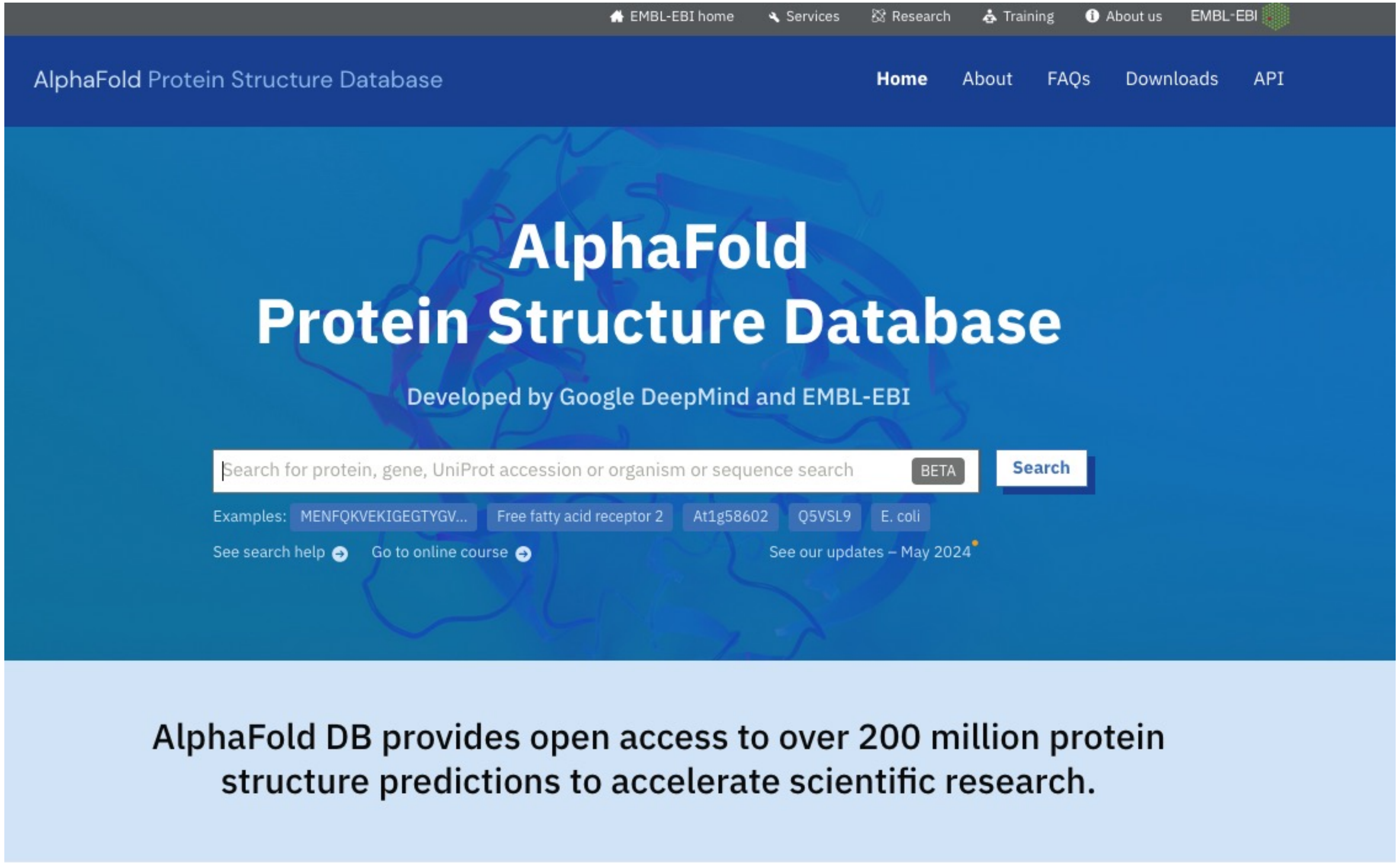


Name	Molecular Weight	Molecular Formula	Residue Formula	Residue Weight (-H_2O)	pKa[1]	pKb[2]	pKx[3]	pI[4]
Alanine (Ala/A)	89.10	$C_3H_7NO_2$	C_3H_5NO	71.08	2.34	9.69	-	6.00
* **Arginine** (Arg/ R)	174.20	$C_6H_{14}N_4O_2$	$C_6H_{12}N_4O$	156.19	2.17	9.04	12.48	10.76
Asparagine (Asn/N)	132.12	$C_4H_8N_2O_3$	$C_4H_6N_2O_2$	114.11	2.02	8.80	-	5.41
Aspartic acid (Asp/D)	133.11	$C_4H_7NO_4$	$C_4H_5NO_3$	115.09	1.88	9.60	3.65	2.77
Cysteine (Cys/ C)	121.16	$C_3H_7NO_2S$	C_3H_5NOS	103.15	1.96	10.28	8.18	5.07
Glutamic acid (Glu/E)	147.13	$C_5H_9NO_4$	$C_5H_7NO_3$	129.12	2.19	9.67	4.25	3.22
* **Glutamine** (Gln/ Q)	146.15	$C_5H_{10}N_2O_3$	$C_5H_8N_2O_2$	128.13	2.17	9.13	-	5.65
Glycine (Gly/G)	75.07	$C_2H_5NO_2$	C_2H_3NO	57.05	2.34	9.60	-	5.97
Histidine (His/ H)	155.16	$C_6H_9N_3O_2$	$C_6H_7N_3O$	137.14	1.82	9.17	6.00	7.59
Hydroxyproline (Hyp/O)	131.13	$C_5H_9NO_3$	$C_5H_7NO_2$	113.11	1.82	9.65	-	-
Isoleucine (Ile/I)	131.18	$C_6H_{13}NO_2$	$C_6H_{11}NO$	113.16	2.36	9.60	-	6.02
Leucine (Leu/L)	131.18	$C_6H_{13}NO_2$	$C_6H_{11}NO$	113.16	2.36	9.60	-	5.98
Lysine (Lys/K)	146.19	$C_6H_{14}N_2O_2$	$C_6H_{12}N_2O$	128.18	2.18	8.95	10.53	9.74
Methionine (Met/M)	149.21	$C_5H_{11}NO_2S$	C_5H_9NOS	131.20	2.28	9.21	-	5.74
Phenylalanine (Phe/F)	165.19	$C_9H_{11}NO_2$	C_9H_9NO	147.18	1.83	9.13	-	5.48

The majority of Mendelian phenotypes are currently associated with protein coding changes

- Impact depends on context in the protein and its role in the protein's function.
- It *can* lead to changes in charge interactions, hydrogen bonding, protein stability, and biological activity
- potentially resulting in significant functional consequences
- Some aa substitutions are much more significant than others

View the protein in 3D



EMBL-EBI home
Services
Research
Training
About us
EMBL-EBI
AlphaFold Protein Structure Database
Home
About
FAQs
Downloads
API

AlphaFold
Protein Structure Database

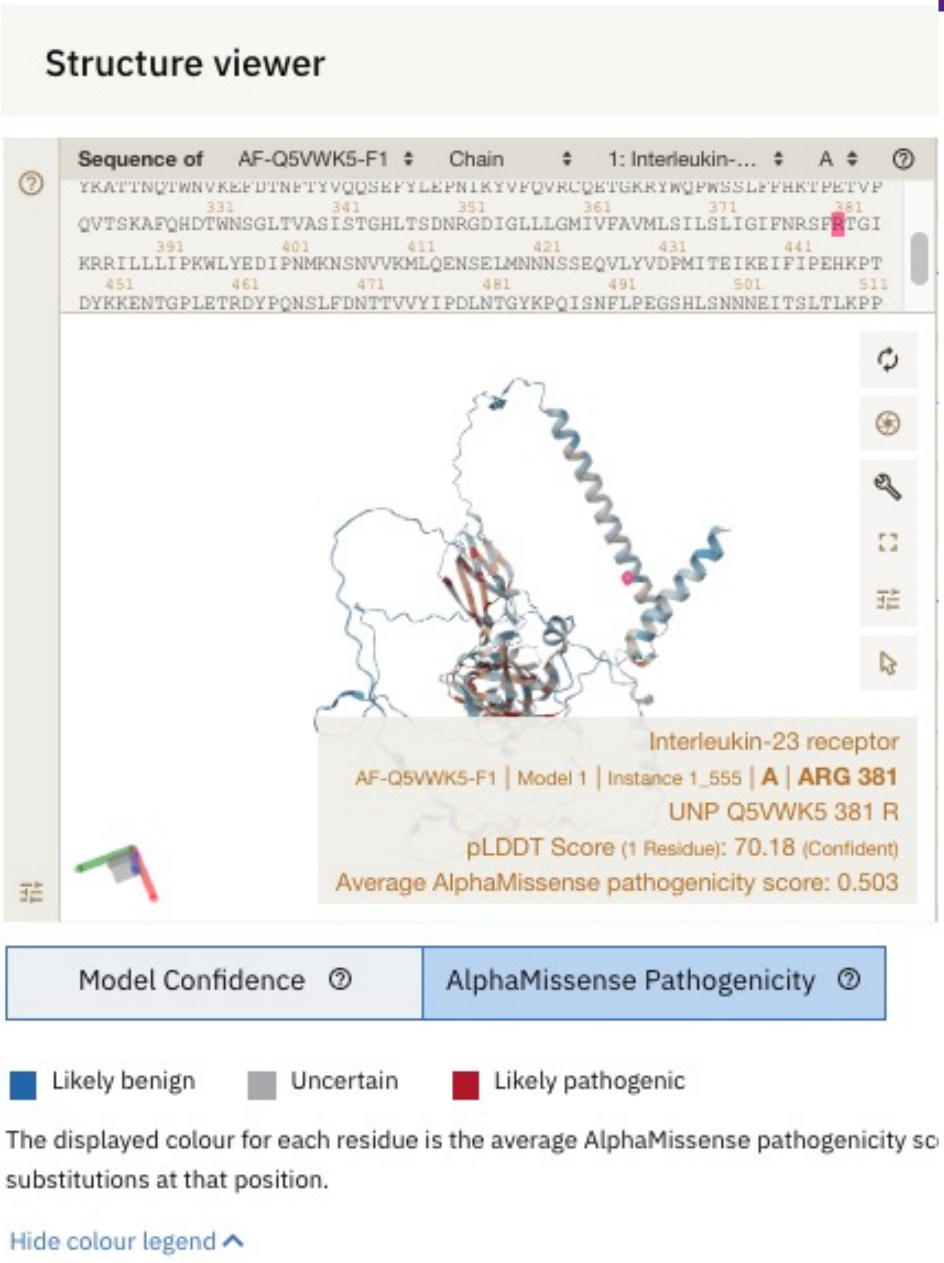
Developed by Google DeepMind and EMBL-EBI

Search for protein, gene, UniProt accession or organism or sequence search
BETA
Search
Examples: MENFQKVEKIGEGTYGV... Free fatty acid receptor 2 At1g58602 Q5VSL9 E. coli
See search help
Go to online course
See our updates – May 2024

AlphaFold DB provides open access to over 200 million protein structure predictions to accelerate scientific research.

https://alphafold.ebi.ac.uk/

IL23R



Structure viewer

Sequence of AF-Q5VWK5-F1 ⬍ Chain ⬍ 1: Interleukin-... ⬍ A ⬍ ⓘ

```
YKATTNQTWNVKEFDTNFTYVQQSEFYLEPNIKYVFQVRCQETGKRYWQPWSSLFFHKTPETVP
      331        341        351        361        371        381
QVTSKAFQHDTWNSGLTVASISTGHLTSDNRGDIGLLLGMIVFAVMLSILSLIGIFNRSFRTGI
    391        401        411        421        431        441
KRRILLLIPKWLYEDIPNMKNSNVVKMLQENSELMNNNSSEQVLYVDPMITEIKEIFIPEHKPT
  451        461        471        481        491        501        511
DYKKENTGPLETRDYPQNSLFDNTTVVYIPDLNTGYKPQISNFLPEGSHLSNNNEITSLTLKPP
```

Interleukin-23 receptor
AF-Q5VWK5-F1 | Model 1 | Instance 1_555 | A | ARG 381
UNP Q5VWK5 381 R
pLDDT Score (1 Residue): 70.18 (Confident)
Average AlphaMissense pathogenicity score: 0.503

Model Confidence ⓘ | AlphaMissense Pathogenicity ⓘ

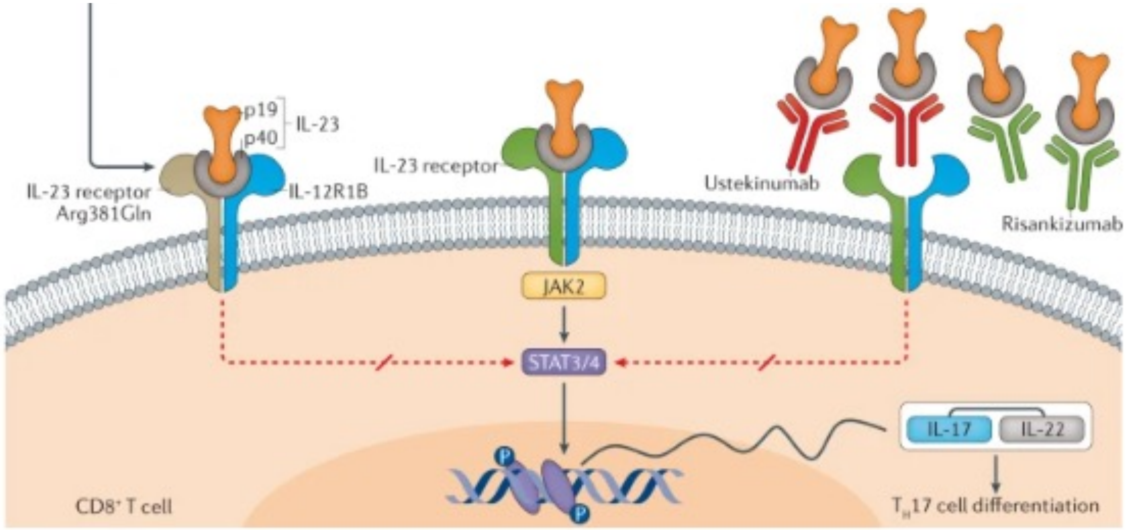
🟦 Likely benign ▨ Uncertain 🟥 Likely pathogenic

The displayed colour for each residue is the average AlphaMissense pathogenicity sc substitutions at that position.

Hide colour legend ^

IL23R binds with IL12R1B1.
Docking of IL23 mediates T-cells, NK cells and possibly certain macrophage/myeloid cells stimulation probably through activation of the Jak-Stat signaling cascade.

IL23 functions in innate and adaptive immunity and may participate in acute response to infection in peripheral tissues.



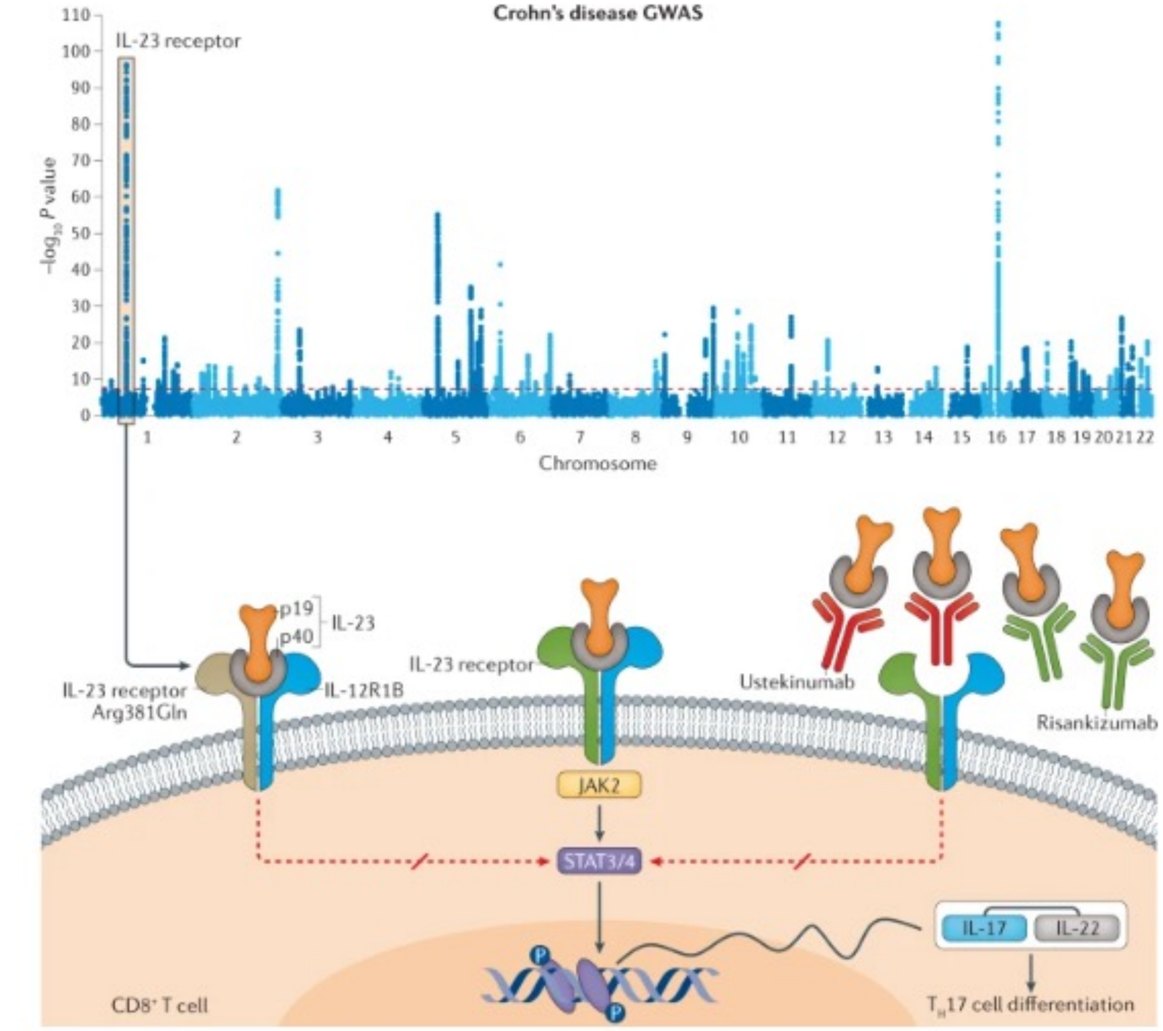
Crohn's Disease

- Need to consider GWAS contain multiple loci and hundreds of correlated SNPs

- Most will be in the non-coding regions

- Need an effective pipeline
- No gold standard
- Typically - bigger data is more powerful
- 4 main challenges…

Fig. 1: Genome-wide significant variants associated with Crohn's disease spanning the IL-23 receptor provide drug repurposing opportunities.



GWAS follow-up

Challenge 1 = correlated SNPs (LD)

Significant association P-values are distributed over blocks of correlated genetic variants: actual causal variant is unclear

Solution → fine-mapping (correlation structure modelled with association values to pinpoint the most likely causal SNPs, this can be integrated with functional information (e.g. tools FINEMAP, PAINTOR))

Solution → Annotation - provides orthogonal information that may help to distinguish the causal variant from the SNP in perfect LD with it (e.g. some are platforms i.e. FUMA, ANNOVAR, SNPEff with integrated data, or standalone – e.g. CADD, VEP)

Challenge 2. Many GWAS hits are in non-coding regions

The majority of GWAS hits are in non-coding regions. Do not directly lead to a different protein structure and their impact on protein function may be less straightforward to assess

Solution: link GWAS variants to genes via regulatory information from external resources, such as ENCODE, GTEX, eQTLGen (e-QTL), chromatin interactions, i.e. add information on the association of a variant with DNA transcription and RNA or protein levels

GWAS follow-up

Challenge 3. Many traits are polygenic

Multiple genetic variants of small effect contribute. A single genetic variant, even if it is known to be causal, might not be informative for biology

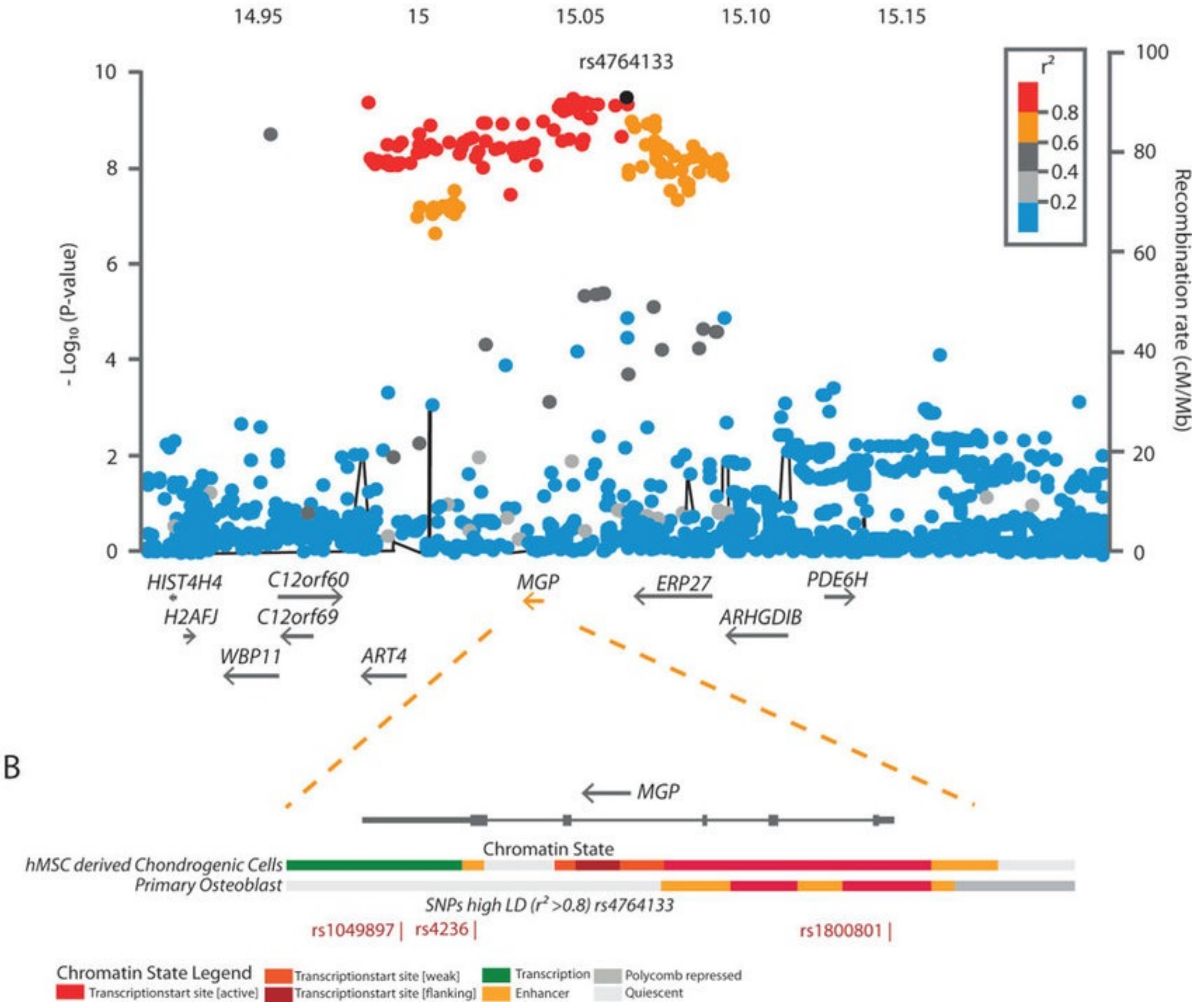
Solution: map associated SNPs to genes and look for convergence in biological pathways, shared cellular or synaptic function, co-localization, co-expression in tissue or cell types (e.g. tools MAGMA, Ldscore regression)

Challenge 4. Unobserved variation

If SNPs are not imputed or observed – they will not be considered- its effect may be captured through LD by an SNP that has a different annotation from the causal variant

Solution: Better imputation and/or sequencing (whole genome) – esp. for CNV/SV calling or methylation data

Typical locus



- multiple correlated SNPs
- multiple closely-located genes

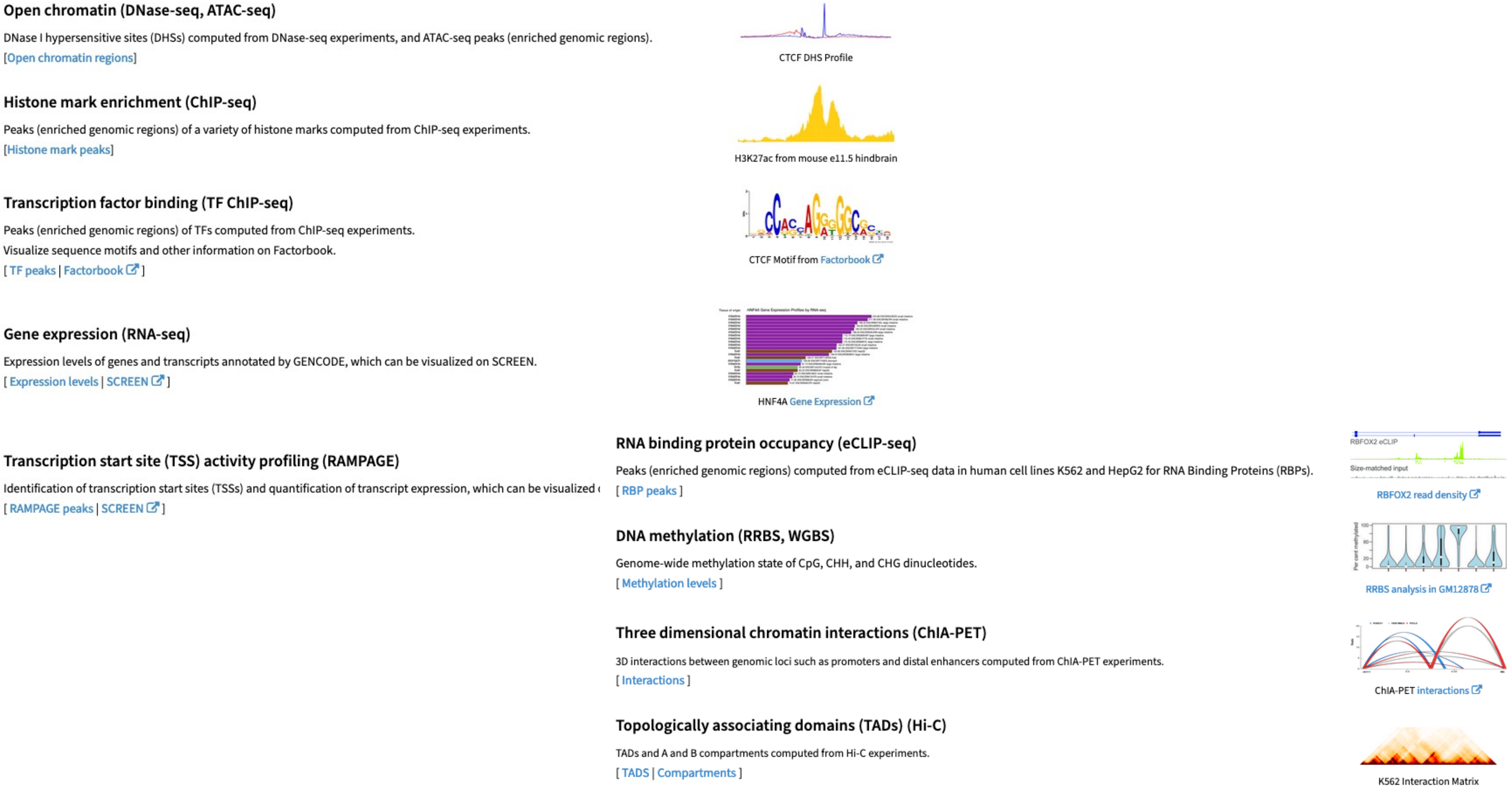
MULTIPLE PAPERS TO INTERROGATE

- Consider prioritising regulatory regions in cells relevant to disease
- Models that can recapitulate the condition

- rs4764133

https://arthritis-research.biomedcentral.com/articles/10.1186/s13075-019-1934-7

ENCODE



Open chromatin (DNase-seq, ATAC-seq)

DNase I hypersensitive sites (DHSs) computed from DNase-seq experiments, and ATAC-seq peaks (enriched genomic regions).
[Open chromatin regions]

CTCF DHS Profile

Histone mark enrichment (ChIP-seq)

Peaks (enriched genomic regions) of a variety of histone marks computed from ChIP-seq experiments.
[Histone mark peaks]

H3K27ac from mouse e11.5 hindbrain

Transcription factor binding (TF ChIP-seq)

Peaks (enriched genomic regions) of TFs computed from ChIP-seq experiments.
Visualize sequence motifs and other information on Factorbook.
[TF peaks | Factorbook]

CTCF Motif from Factorbook

Gene expression (RNA-seq)

Expression levels of genes and transcripts annotated by GENCODE, which can be visualized on SCREEN.
[Expression levels | SCREEN]

HNF4A Gene Expression

Transcription start site (TSS) activity profiling (RAMPAGE)

Identification of transcription start sites (TSSs) and quantification of transcript expression, which can be visualized
[RAMPAGE peaks | SCREEN]

RNA binding protein occupancy (eCLIP-seq)

Peaks (enriched genomic regions) computed from eCLIP-seq data in human cell lines K562 and HepG2 for RNA Binding Proteins (RBPs).
[RBP peaks]

RBFOX2 read density

DNA methylation (RRBS, WGBS)

Genome-wide methylation state of CpG, CHH, and CHG dinucleotides.
[Methylation levels]

RRBS analysis in GM12878

Three dimensional chromatin interactions (ChIA-PET)

3D interactions between genomic loci such as promoters and distal enhancers computed from ChIA-PET experiments.
[Interactions]

ChIA-PET interactions

Topologically associating domains (TADs) (Hi-C)

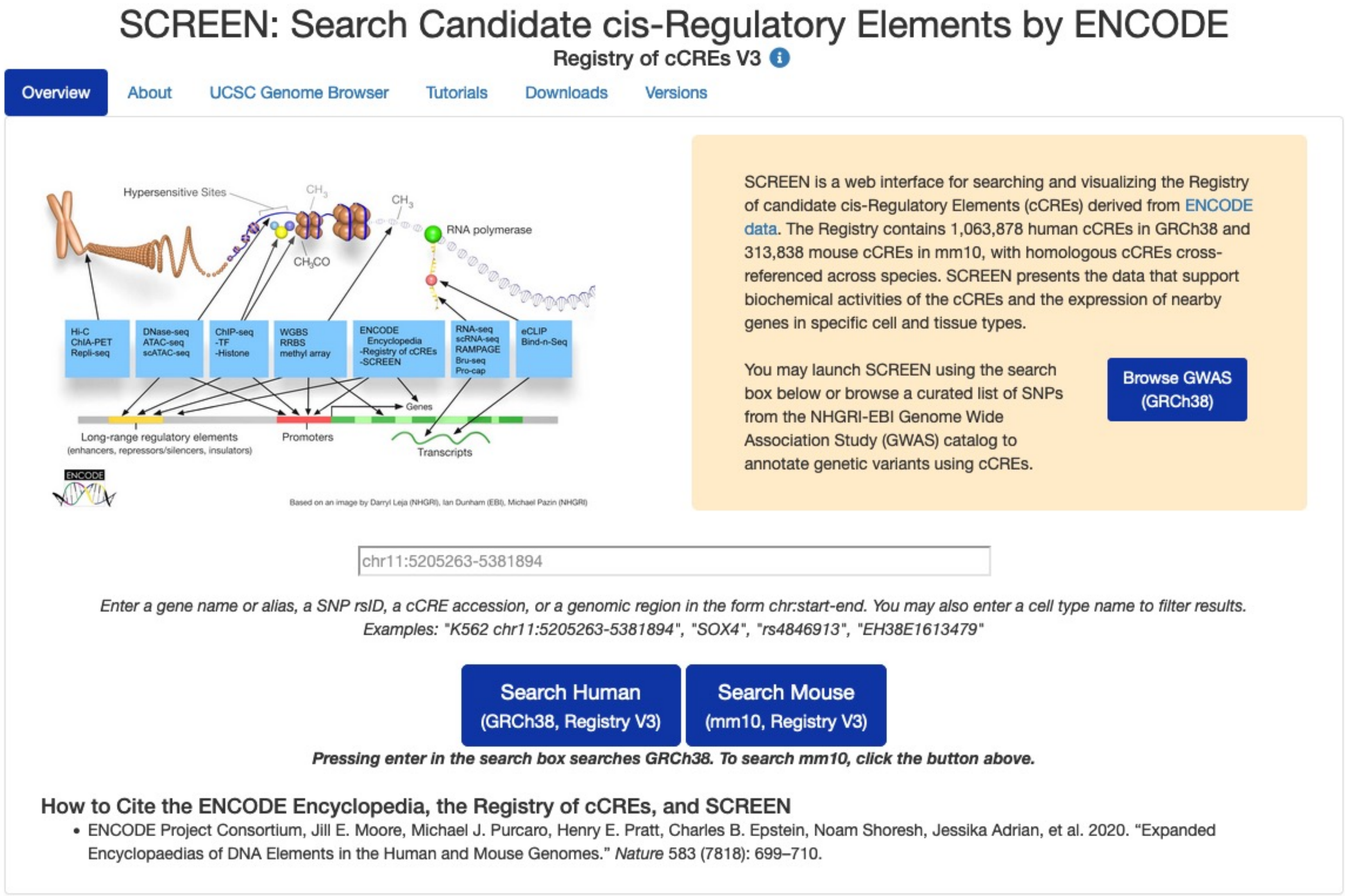
TADs and A and B compartments computed from Hi-C experiments.
[TADS | Compartments]

K562 Interaction Matrix

ENCODE

Eg. Rs4764133

- Look at osteoclast



SCREEN: Search Candidate cis-Regulatory Elements by ENCODE

Registry of cCREs V3

Based on an image by Darryl Leja (NHGRI), Ian Dunham (EBI), Michael Pazin (NHGRI)

SCREEN is a web interface for searching and visualizing the Registry of candidate cis-Regulatory Elements (cCREs) derived from ENCODE data. The Registry contains 1,063,878 human cCREs in GRCh38 and 313,838 mouse cCREs in mm10, with homologous cCREs cross-referenced across species. SCREEN presents the data that support biochemical activities of the cCREs and the expression of nearby genes in specific cell and tissue types.

You may launch SCREEN using the search box below or browse a curated list of SNPs from the NHGRI-EBI Genome Wide Association Study (GWAS) catalog to annotate genetic variants using cCREs.

chr11:5205263-5381894

Enter a gene name or alias, a SNP rsID, a cCRE accession, or a genomic region in the form chr:start-end. You may also enter a cell type name to filter results. Examples: "K562 chr11:5205263-5381894", "SOX4", "rs4846913", "EH38E1613479"

Pressing enter in the search box searches GRCh38. To search mm10, click the button above.

How to Cite the ENCODE Encyclopedia, the Registry of cCREs, and SCREEN

- ENCODE Project Consortium, Jill E. Moore, Michael J. Purcaro, Henry E. Pratt, Charles B. Epstein, Noam Shoresh, Jessika Adrian, et al. 2020. "Expanded Encyclopaedias of DNA Elements in the Human and Mouse Genomes." *Nature* 583 (7818): 699–710.

PRACTICAL

https://wannovar.wglab.org/

TASK -- Use a resource of your choice to annotate 5 SNPs

Use a consistent genome alignment – i.e. hg19 or hg38

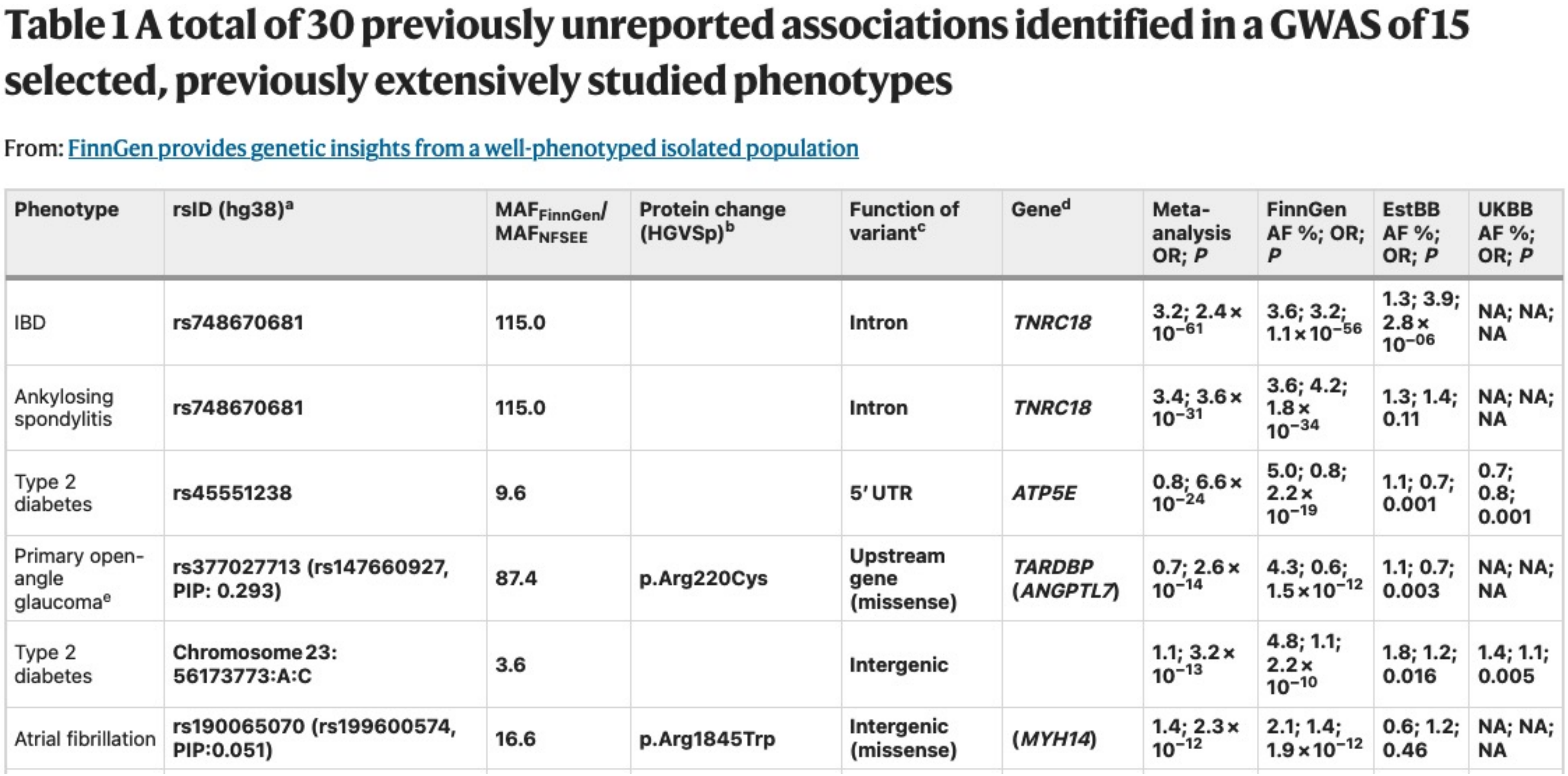


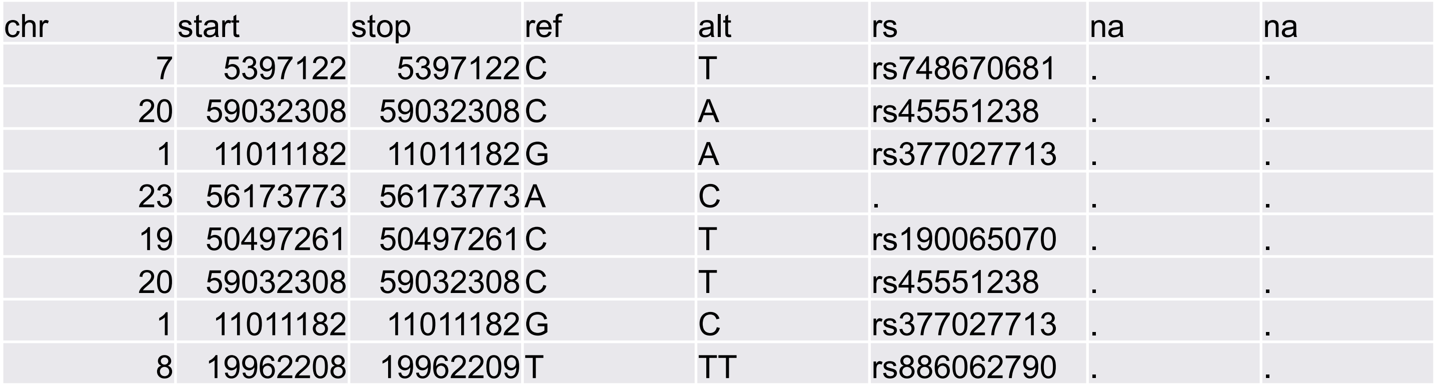
Table 1 A total of 30 previously unreported associations identified in a GWAS of 15 selected, previously extensively studied phenotypes

From: FinnGen provides genetic insights from a well-phenotyped isolated population

Phenotype	rsID (hg38)[a]	$\mathrm{MAF}_{\mathrm{FinnGen}}$/ $\mathrm{MAF}_{\mathrm{NFSEE}}$	Protein change (HGVSp)[b]	Function of variant[c]	Gene[d]	Meta-analysis OR; *P*	FinnGen AF %; OR; *P*	EstBB AF %; OR; *P*	UKBB AF %; OR; *P*
IBD	**rs748670681**	**115.0**		**Intron**	***TNRC18***	**3.2; 2.4×10^{-61}**	**3.6; 3.2; 1.1×10^{-56}**	**1.3; 3.9; 2.8×10^{-06}**	**NA; NA; NA**
Ankylosing spondylitis	**rs748670681**	**115.0**		**Intron**	***TNRC18***	**3.4; 3.6×10^{-31}**	**3.6; 4.2; 1.8×10^{-34}**	**1.3; 1.4; 0.11**	**NA; NA; NA**
Type 2 diabetes	**rs45551238**	**9.6**		**5′ UTR**	***ATP5E***	**0.8; 6.6×10^{-24}**	**5.0; 0.8; 2.2×10^{-19}**	**1.1; 0.7; 0.001**	**0.7; 0.8; 0.001**
Primary open-angle glaucoma[e]	**rs377027713 (rs147660927, PIP: 0.293)**	**87.4**	**p.Arg220Cys**	**Upstream gene (missense)**	***TARDBP* (*ANGPTL7*)**	**0.7; 2.6×10^{-14}**	**4.3; 0.6; 1.5×10^{-12}**	**1.1; 0.7; 0.003**	**NA; NA; NA**
Type 2 diabetes	**Chromosome 23: 56173773:A:C**	**3.6**		**Intergenic**		**1.1; 3.2×10^{-13}**	**4.8; 1.1; 2.2×10^{-10}**	**1.8; 1.2; 0.016**	**1.4; 1.1; 0.005**
Atrial fibrillation	**rs190065070 (rs199600574, PIP:0.051)**	**16.6**	**p.Arg1845Trp**	**Intergenic (missense)**	**(*MYH14*)**	**1.4; 2.3×10^{-12}**	**2.1; 1.4; 1.9×10^{-12}**	**0.6; 1.2; 0.46**	**NA; NA; NA**

Input format

Annovar.txt



chr	start	stop	ref	alt	rs	na	na
7	5397122	5397122	C	T	rs748670681	.	.
20	59032308	59032308	C	A	rs45551238	.	.
1	11011182	11011182	G	A	rs377027713	.	.
23	56173773	56173773	A	C	.	.	.
19	50497261	50497261	C	T	rs190065070	.	.
20	59032308	59032308	C	T	rs45551238	.	.
1	11011182	11011182	G	C	rs377027713	.	.
8	19962208	19962209	T	TT	rs886062790	.	.

Output format

Chr	Start	End	Ref	Alt	Func.refGene	Gene.refGene	GeneDetail.refGene	ExonicFunc.refGene	AAChange.refGene	1000G_ALL
7	5397122	5397122	C	T	intronic	TNRC18				.
20	59032308	59032308	C	A	ncRNA_intronic	SLMO2-ATP5E				
1	11011182	11011182	G	A	intergenic	C1orf127;TARDBP	dist=29145;dist=1440			0.0014
1	11011182	11011182	G	C	intergenic	C1orf127;TARDBP	dist=29145;dist=1440			0.0008
1	11193760	11193760	C	T	exonic	ANGPTL7		nonsynonymous SNV	ANGPTL7:NM_021146:exon3:c.C658T:p.R220C	0.001
23	56173773	56173773	A	C	intergenic	NONE;NONE	dist=NONE;dist=NONE			
19	50497261	50497261	C	T	intergenic	EMC10;JOSD2	dist=13735;dist=8736			0.001
19	50301724	50301724	C	T	exonic	MYH14		nonsynonymous SNV	MYH14:NM_024729:exon38:c.C5410T:p.R1804W,MYH14:NM_001077186:exon39:c.C5434T:p.R1812W,MYH14:NM_001145809:exon40:c.C5533T:p.R1845W	0.0008
20	59032308	59032308	C	T	ncRNA_intronic	SLMO2-ATP5E				0.0038
8	19962208	19962208	-	T	exonic	LPL		stopgain	LPL:NM_000237:exon9:c.1416_1417insT:p.K473*	.

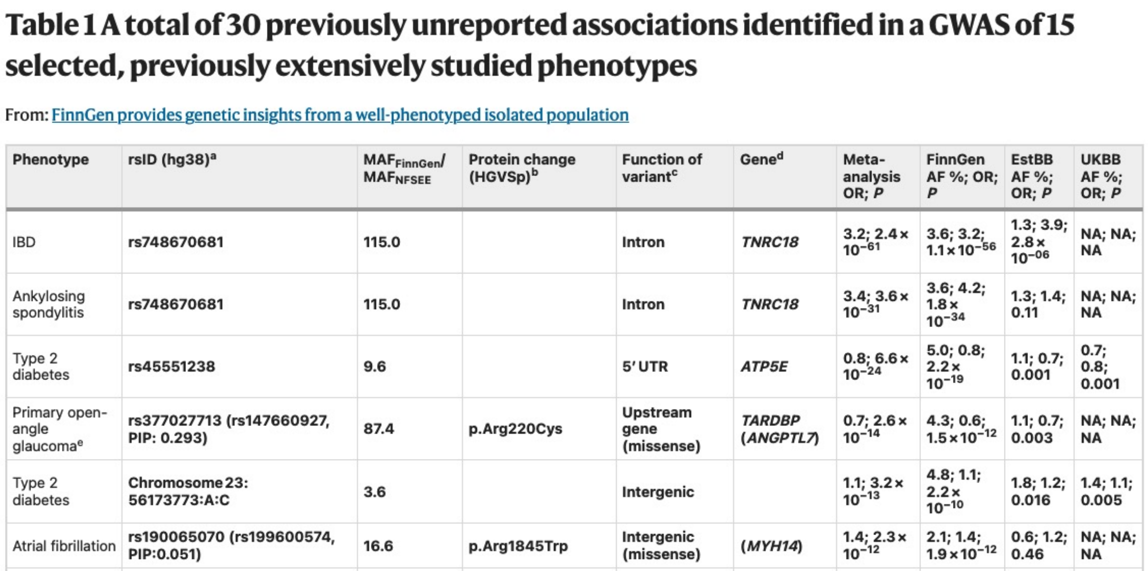


Table 1 A total of 30 previously unreported associations identified in a GWAS of 15 selected, previously extensively studied phenotypes

From: FinnGen provides genetic insights from a well-phenotyped isolated population

Phenotype	rsID (hg38)[a]	$MAF_{FinnGen}$/ MAF_{NFSEE}	Protein change (HGVSp)[b]	Function of variant[c]	Gene[d]	Meta-analysis OR; *P*	FinnGen AF %; OR; *P*	EstBB AF %; OR; *P*	UKBB AF %; OR; *P*
IBD	rs748670681	115.0		Intron	*TNRC18*	3.2; 2.4×10^{-61}	3.6; 3.2; 1.1×10^{-56}	1.3; 3.9; 2.8×10^{-06}	NA; NA; NA
Ankylosing spondylitis	rs748670681	115.0		Intron	*TNRC18*	3.4; 3.6×10^{-31}	3.6; 4.2; 1.8×10^{-34}	1.3; 1.4; 0.11	NA; NA; NA
Type 2 diabetes	rs45551238	9.6		5′ UTR	*ATP5E*	0.8; 6.6×10^{-24}	5.0; 0.8; 2.2×10^{-19}	1.1; 0.7; 0.001	0.7; 0.8; 0.001
Primary open-angle glaucoma[e]	rs377027713 (rs147660927, PIP: 0.293)	87.4	p.Arg220Cys	Upstream gene (missense)	*TARDBP* (*ANGPTL7*)	0.7; 2.6×10^{-14}	4.3; 0.6; 1.5×10^{-12}	1.1; 0.7; 0.003	NA; NA; NA
Type 2 diabetes	Chromosome 23: 56173773:A:C	3.6		Intergenic		1.1; 3.2×10^{-13}	4.8; 1.1; 2.2×10^{-10}	1.8; 1.2; 0.016	1.4; 1.1; 0.005
Atrial fibrillation	rs190065070 (rs199600574, PIP:0.051)	16.6	p.Arg1845Trp	Intergenic (missense)	(*MYH14*)	1.4; 2.3×10^{-12}	2.1; 1.4; 1.9×10^{-12}	0.6; 1.2; 0.46	NA; NA; NA

Example databases to load in Annovar

Database	Explanation
refGene	FASTA sequences for all annotated transcripts in RefSeq Gene
cytoBand	Identify Giemsa-stained chromosomes bands (cytogenetic band)
exac03	ExAC 65000 exome allele frequency data for ALL, AFR (African), AMR (Admixed American), EAS (East Asian), FIN (Finnish), NFE (Non-finnish European), OTH (other), SAS (South Asian)). version 0.3. Left normalization done.
avsnp150	dbSNP150 with allelic splitting and left-normalization
dbnsfp30a	whole-exome SIFT, PolyPhen2 HDIV, PolyPhen2 HVAR, LRT, MutationTaster, MutationAssessor, FATHMM, MetaSVM, MetaLR, VEST, CADD, GERP++, DANN, fitCons, PhyloP and SiPhy scores from dbNSFP version 3.0a
clinvar_20220320	Clinvar version 20220320 with separate columns (CLNALLELEID CLNDN CLNDISDB CLNREVSTAT CLNSIG)
dbnsfp42c	whole-exome SIFT, PolyPhen2 HDIV, PolyPhen2 HVAR, LRT, MutationTaster, MutationAssessor, FATHMM, PROVEAN, MetaSVM, MetaLR, VEST, M-CAP, CADD, GERP++, DANN, fathmm-MKL, Eigen, GenoCanyon, fitCons, PhyloP and SiPhy scores from dbNSFP version 3.3a
intervar_20180118	InterVar: clinical interpretation of missense variants (indels not supported)
gnomad211_genome	gnomAD genome collection with "AF AF_popmax AF_male AF_female AF_raw AF_afr AF_sas AF_amr AF_eas AF_nfe AF_fin AF_asj AF_oth non_topmed_AF_popmax non_neuro_AF_popmax non_cancer_AF_popmax controls_AF_popmax" header
1000g2015aug (ALL.sites.2015_08)	alternative allele frequency data in 1000 Genomes Project for autosomes (ALL, AFR (African), AMR (Admixed American), EAS (East Asian), EUR (European), SAS (South Asian)). Based on 201409 collection v5 (based on 201305 alignment) but including chrX and chrY data finally!

Cluster-based annotation

```
table_annovar.pl \
    FL_denovo_anno/unzipped_vcf/${input}.vcf \
    humandb/ \
    -buildver hg19 \
    -out FL_denovo_anno/${input}_anno \
    -vcfinput    -nastring .  -polish \
    -xref  humandb/hg19_refGene.txt  \
    -protocol  refGene,cytoBand,exac03,avsnp150,dbnsfp30a,clinvar_20220320,dbnsfp42c,
               intervar_20180118,gnomad211_genome,ALL.sites.2015_08  \

    -operation gx,r,f,f,f,f,f,f,f,f
```

g -- Gene-based Annotation

r – Region-based Annotation

f -- Filter-based Annotation

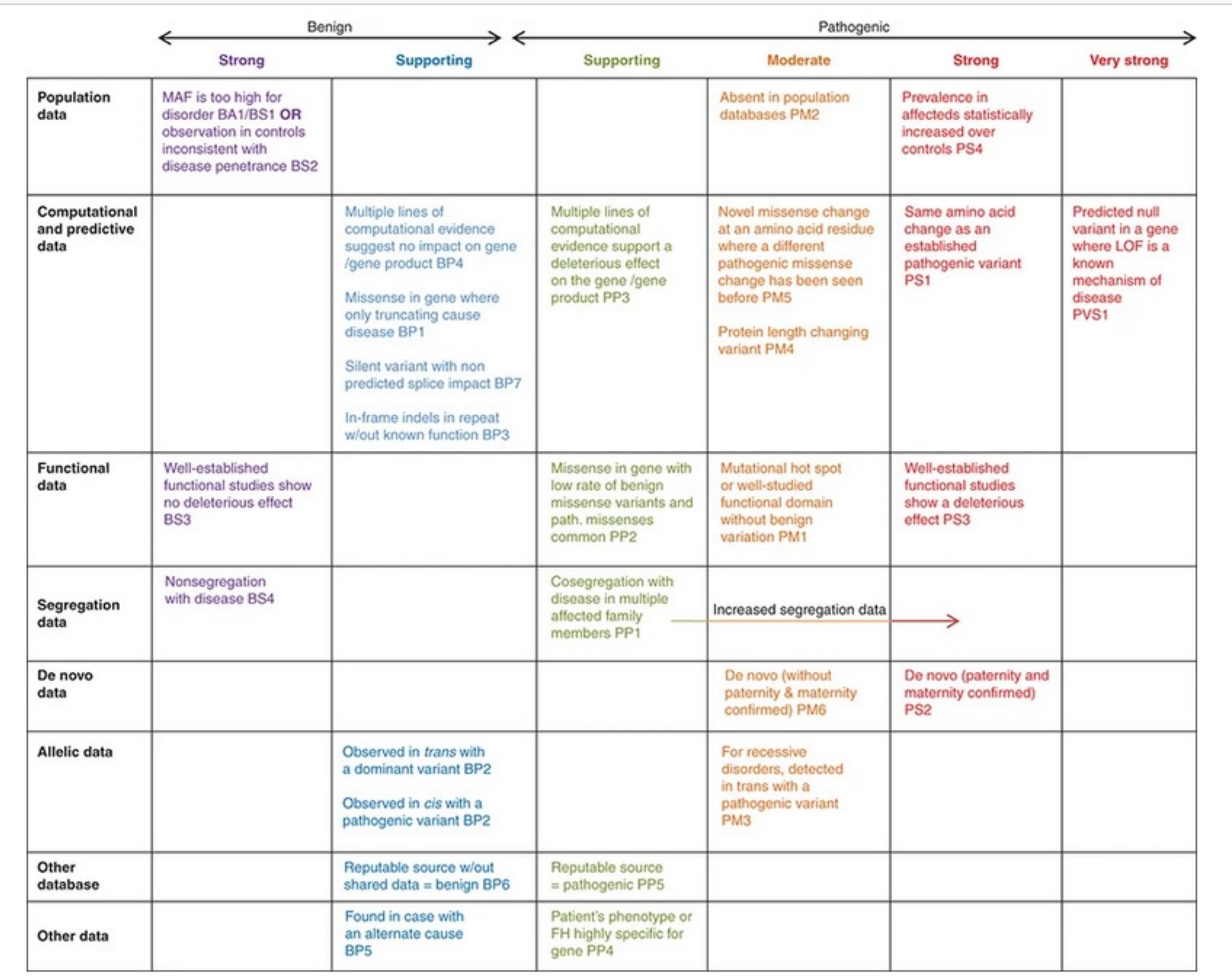
gx -- Gene-based with cross-reference annotation

Clinvar = clinically relevant variants

ACMG criteria
Richards et al. 2015

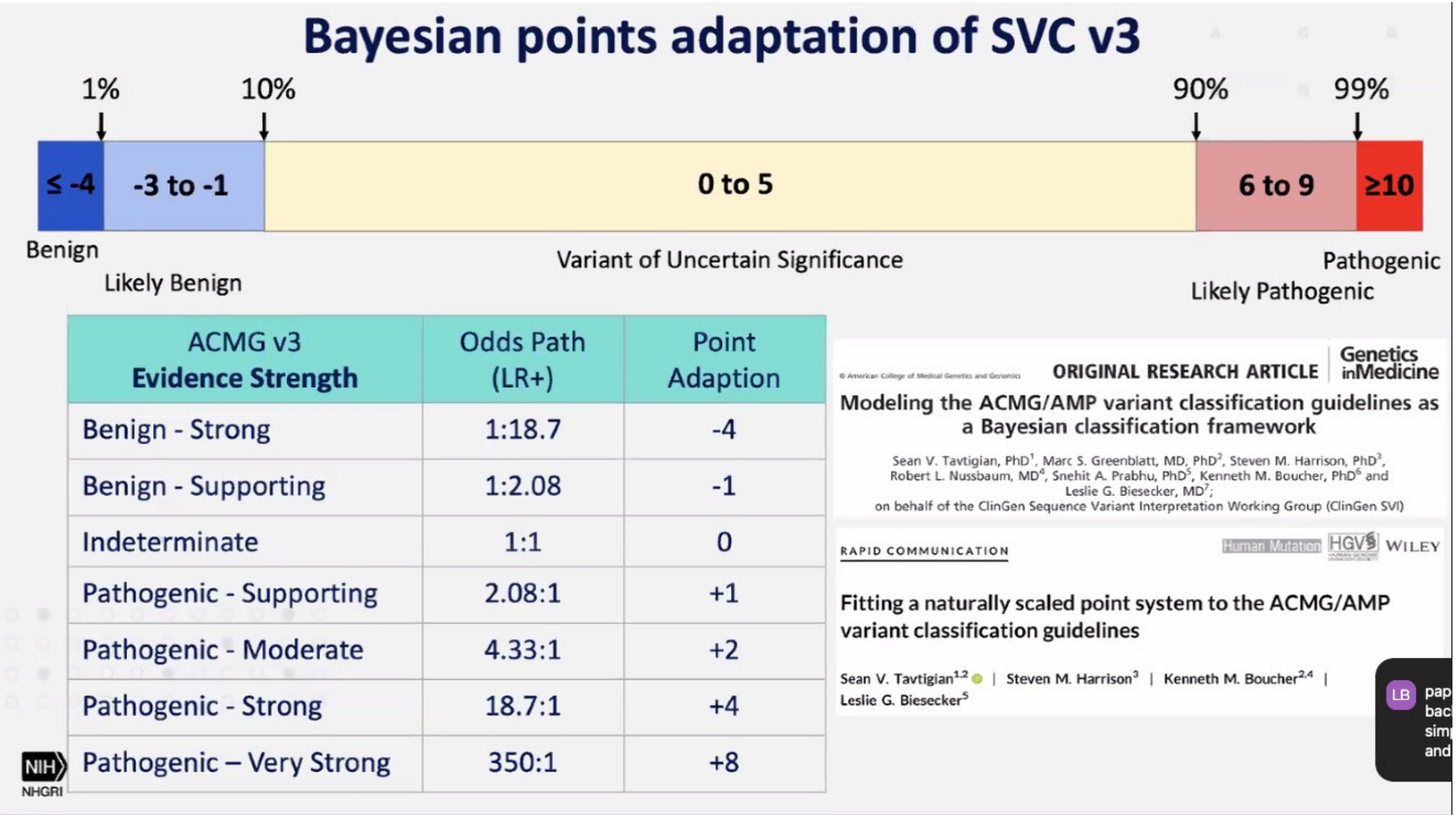
coding variants

Version 4 of the ACMG guidelines
To be released 2024…
Significant updates



	Benign		Pathogenic			
	Strong	Supporting	Supporting	Moderate	Strong	Very strong
Population data	MAF is too high for disorder BA1/BS1 **OR** observation in controls inconsistent with disease penetrance BS2			Absent in population databases PM2	Prevalence in affecteds statistically increased over controls PS4	
Computational and predictive data		Multiple lines of computational evidence suggest no impact on gene /gene product BP4 Missense in gene where only truncating cause disease BP1 Silent variant with non predicted splice impact BP7 In-frame indels in repeat w/out known function BP3	Multiple lines of computational evidence support a deleterious effect on the gene /gene product PP3	Novel missense change at an amino acid residue where a different pathogenic missense change has been seen before PM5 Protein length changing variant PM4	Same amino acid change as an established pathogenic variant PS1	Predicted null variant in a gene where LOF is a known mechanism of disease PVS1
Functional data	Well-established functional studies show no deleterious effect BS3		Missense in gene with low rate of benign missense variants and path. missenses common PP2	Mutational hot spot or well-studied functional domain without benign variation PM1	Well-established functional studies show a deleterious effect PS3	
Segregation data	Nonsegregation with disease BS4		Cosegregation with disease in multiple affected family members PP1	Increased segregation data		
De novo data				De novo (without paternity & maternity confirmed) PM6	De novo (paternity and maternity confirmed) PS2	
Allelic data		Observed in *trans* with a dominant variant BP2 Observed in *cis* with a pathogenic variant BP2		For recessive disorders, detected in trans with a pathogenic variant PM3		
Other database		Reputable source w/out shared data = benign BP6	Reputable source = pathogenic PP5			
Other data		Found in case with an alternate cause BP5	Patient's phenotype or FH highly specific for gene PP4			

V4 is moving towards a quantitative approach



Bayesian points adaptation of SVC v3

1% 10% 90% 99%

≤ -4 | -3 to -1 | 0 to 5 | 6 to 9 | ≥10

Benign

Likely Benign

Variant of Uncertain Significance

Pathogenic

Likely Pathogenic

ACMG v3 **Evidence Strength**	Odds Path (LR+)	Point Adaption
Benign - Strong	1:18.7	-4
Benign - Supporting	1:2.08	-1
Indeterminate	1:1	0
Pathogenic - Supporting	2.08:1	+1
Pathogenic - Moderate	4.33:1	+2
Pathogenic - Strong	18.7:1	+4
Pathogenic – Very Strong	350:1	+8

NIH NHGRI

© American College of Medical Genetics and Genomics **ORIGINAL RESEARCH ARTICLE** Genetics inMedicine

Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework

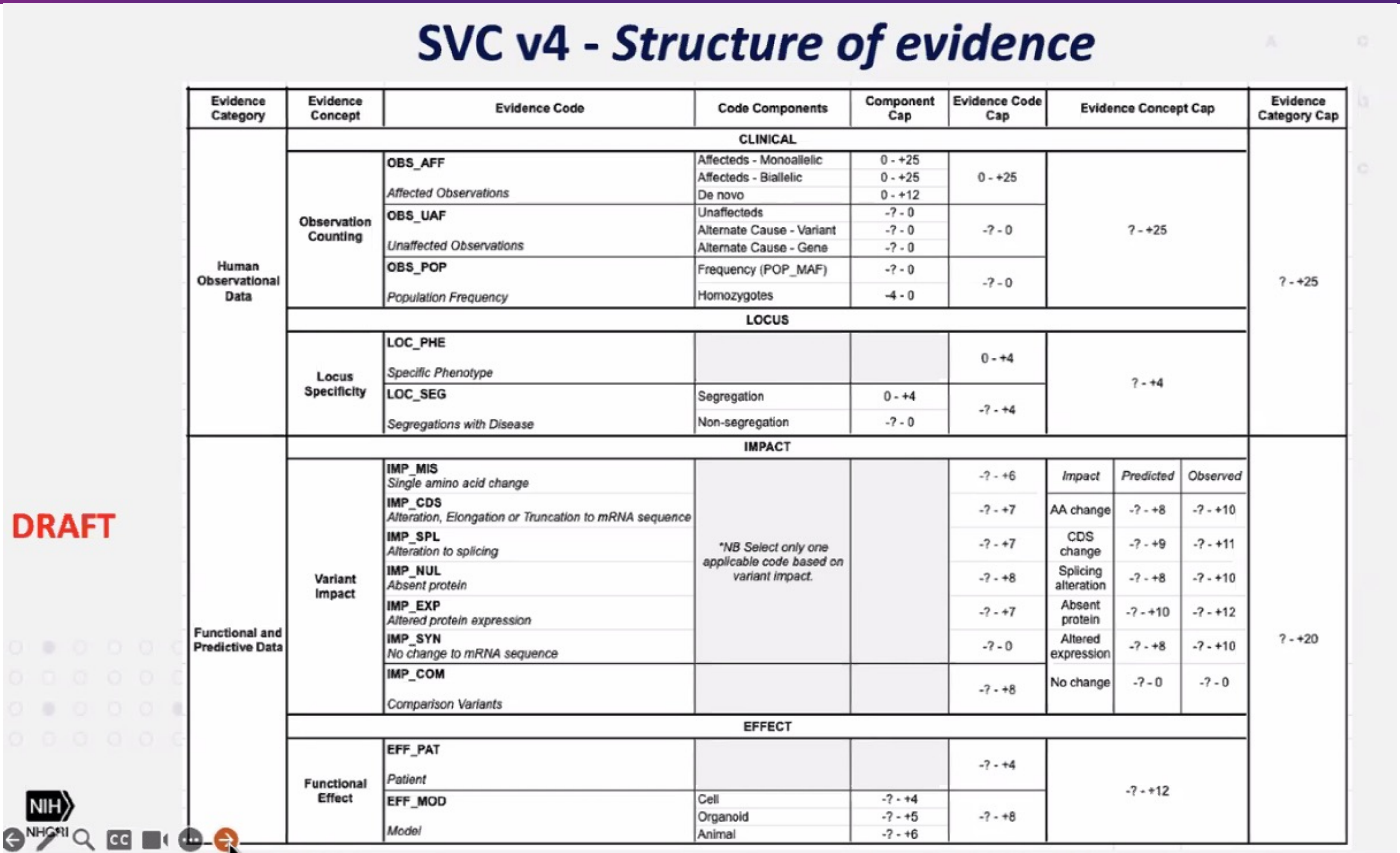
Sean V. Tavtigian, PhD[1], Marc S. Greenblatt, MD, PhD[2], Steven M. Harrison, PhD[3], Robert L. Nussbaum, MD[4], Snehit A. Prabhu, PhD[5], Kenneth M. Boucher, PhD[6] and Leslie G. Biesecker, MD[7];
on behalf of the ClinGen Sequence Variant Interpretation Working Group (ClinGen SVI)

RAPID COMMUNICATION Human Mutation HGVS WILEY

Fitting a naturally scaled point system to the ACMG/AMP variant classification guidelines

Sean V. Tavtigian[1,2] | Steven M. Harrison[3] | Kenneth M. Boucher[2,4] |
Leslie G. Biesecker[5]

ACMG – v4. DRAFT

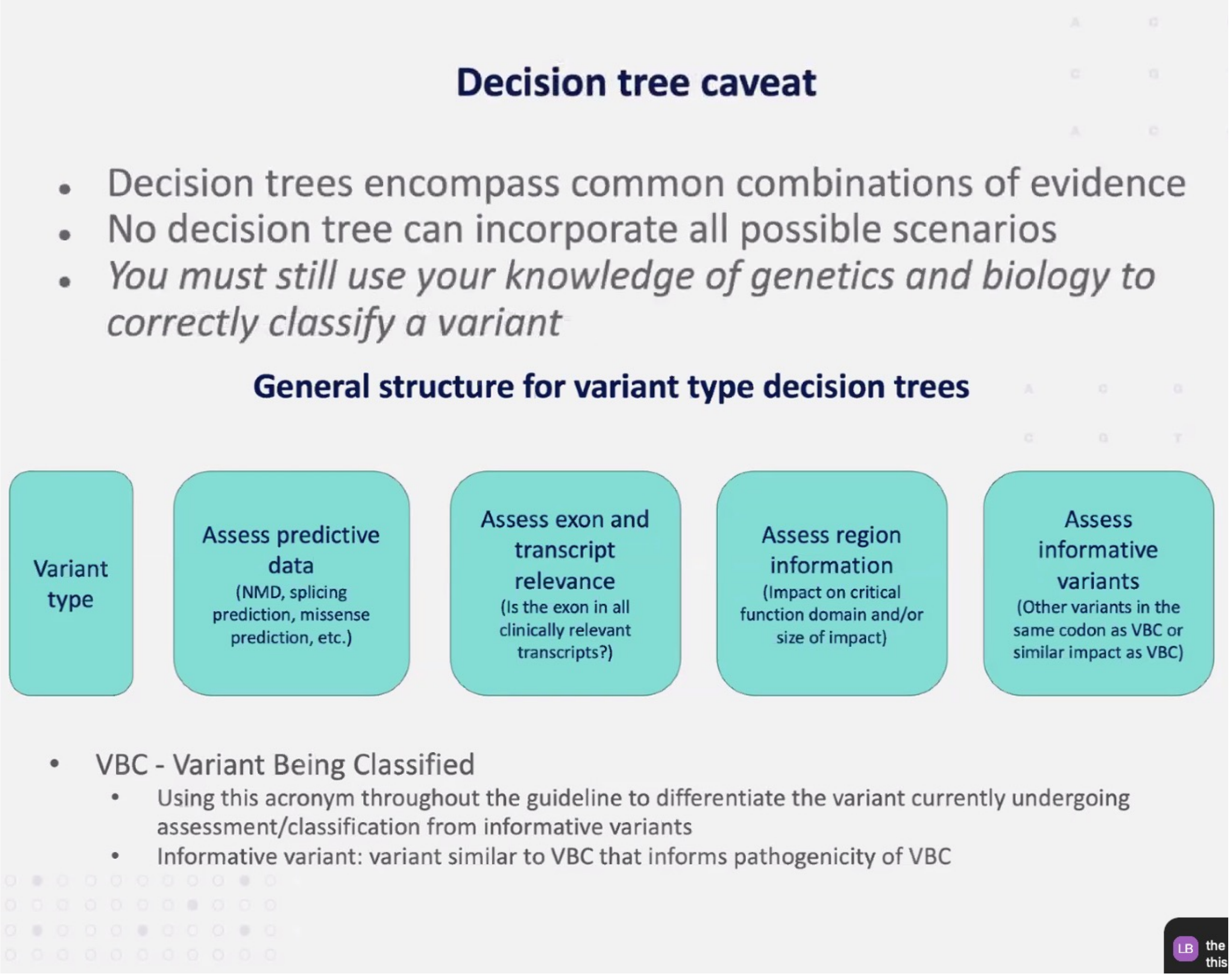


SVC v4 - *Structure of evidence*

Evidence Category	Evidence Concept	Evidence Code	Code Components	Component Cap	Evidence Code Cap	Evidence Concept Cap			Evidence Category Cap
	CLINICAL								
Human Observational Data	Observation Counting	OBS_AFF *Affected Observations*	Affecteds - Monoallelic	0 - +25	0 - +25	? - +25			? - +25
			Affecteds - Biallelic	0 - +25					
			De novo	0 - +12					
		OBS_UAF *Unaffected Observations*	Unaffecteds	-? - 0	-? - 0				
			Alternate Cause - Variant	-? - 0					
			Alternate Cause - Gene	-? - 0					
		OBS_POP *Population Frequency*	Frequency (POP_MAF)	-? - 0	-? - 0				
			Homozygotes	-4 - 0					
	LOCUS								
	Locus Specificity	LOC_PHE *Specific Phenotype*			0 - +4	? - +4			
		LOC_SEG *Segregations with Disease*	Segregation	0 - +4	-? - +4				
			Non-segregation	-? - 0					
	IMPACT								
Functional and Predictive Data	Variant Impact	IMP_MIS *Single amino acid change*	*NB Select only one applicable code based on variant impact.*		-? - +6	*Impact*	*Predicted*	*Observed*	? - +20
		IMP_CDS *Alteration, Elongation or Truncation to mRNA sequence*			-? - +7	AA change	-? - +8	-? - +10	
		IMP_SPL *Alteration to splicing*			-? - +7	CDS change	-? - +9	-? - +11	
		IMP_NUL *Absent protein*			-? - +8	Splicing alteration	-? - +8	-? - +10	
		IMP_EXP *Altered protein expression*			-? - +7	Absent protein	-? - +10	-? - +12	
		IMP_SYN *No change to mRNA sequence*			-? - 0	Altered expression	-? - +8	-? - +10	
		IMP_COM *Comparison Variants*			-? - +8	No change	-? - 0	-? - 0	
	EFFECT								
	Functional Effect	EFF_PAT *Patient*			-? - +4	-? - +12			
		EFF_MOD *Model*	Cell	-? - +4	-? - +8				
			Organoid	-? - +5					
			Animal	-? - +6					

DRAFT

ACMG – v4. updated criteria – DRAFT version



Decision tree caveat

- Decision trees encompass common combinations of evidence
- No decision tree can incorporate all possible scenarios
- *You must still use your knowledge of genetics and biology to correctly classify a variant*

General structure for variant type decision trees

Variant type

Assess predictive data (NMD, splicing prediction, missense prediction, etc.)

Assess exon and transcript relevance (Is the exon in all clinically relevant transcripts?)

Assess region information (Impact on critical function domain and/or size of impact)

Assess informative variants (Other variants in the same codon as VBC or similar impact as VBC)

- VBC - Variant Being Classified
  - Using this acronym throughout the guideline to differentiate the variant currently undergoing assessment/classification from informative variants
  - Informative variant: variant similar to VBC that informs pathogenicity of VBC

Clinvar - clinically relevant variants

Non-coding variants

Guideline | Open Access | Published: 19 July 2022

Recommendations for clinical interpretation of variants found in non-coding regions of the genome

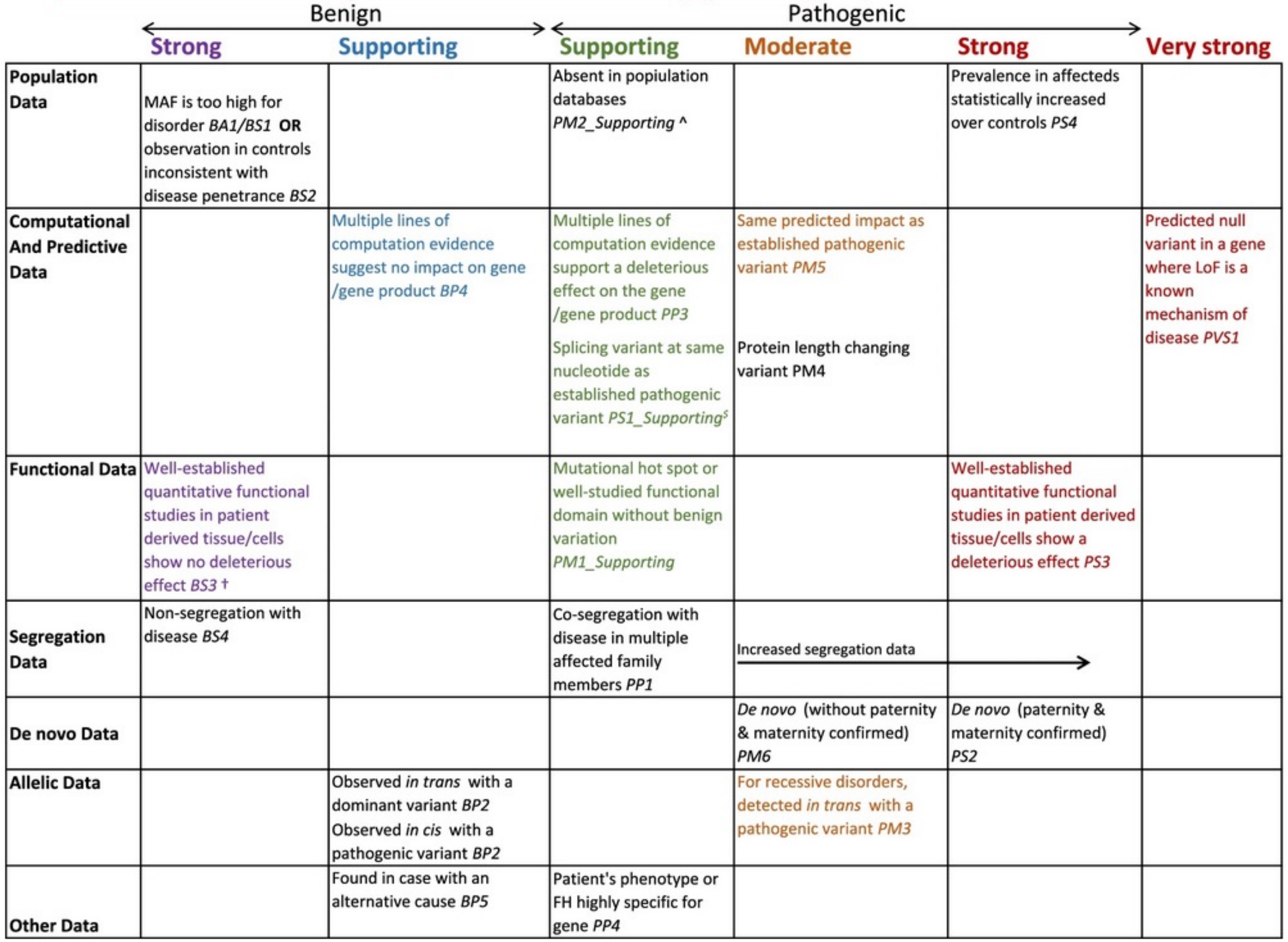
Jamie M. Ellingford, Joo Wook Ahn, ... Nicola Whiffin + Show authors

Genome Medicine **14**, Article number: 73 (2022) | Cite this article

3 Accesses | **92** Altmetric | Metrics

Fig. 3

From: Recommendations for clinical interpretation of variants found in non-coding regions of the genome



	Benign		Pathogenic			
	Strong	Supporting	Supporting	Moderate	Strong	Very strong
Population Data	MAF is too high for disorder *BA1/BS1* **OR** observation in controls inconsistent with disease penetrance *BS2*		Absent in popiulation databases *PM2_Supporting ^*		Prevalence in affecteds statistically increased over controls *PS4*	
Computational And Predictive Data		Multiple lines of computation evidence suggest no impact on gene /gene product *BP4*	Multiple lines of computation evidence support a deleterious effect on the gene /gene product *PP3* Splicing variant at same nucleotide as established pathogenic variant *PS1_Supporting$*	Same predicted impact as established pathogenic variant *PM5* Protein length changing variant PM4		Predicted null variant in a gene where LoF is a known mechanism of disease *PVS1*
Functional Data	Well-established quantitative functional studies in patient derived tissue/cells show no deleterious effect *BS3* †		Mutational hot spot or well-studied functional domain without benign variation *PM1_Supporting*		Well-established quantitative functional studies in patient derived tissue/cells show a deleterious effect *PS3*	
Segregation Data	Non-segregation with disease *BS4*		Co-segregation with disease in multiple affected family members *PP1*	Increased segregation data →		
De novo Data				*De novo* (without paternity & maternity confirmed) *PM6*	*De novo* (paternity & maternity confirmed) *PS2*	
Allelic Data		Observed *in trans* with a dominant variant *BP2* Observed *in cis* with a pathogenic variant *BP2*		For recessive disorders, detected *in trans* with a pathogenic variant *PM3*		
Other Data		Found in case with an alternative cause *BP5*	Patient's phenotype or FH highly specific for gene *PP4*			

Summary

- GWAS evidence is robust and is one of the most useful / relevant pieces of preclinical evidence for translation

- Most of the genome is made up of regulatory regions
- >90% of GWAS loci are situated in these regions

- Coding variants = low-hanging fruit – more data, tools to assess their impact
- Non-coding can still be interrogated, especially if the gene or genes being regulated is clear and direction is known..

- To keep in mind…
  - Moving field - updates in data, rs numbers, genome builds, predictions
  - Genetic information of an organism can be differentially expressed over time and in different tissues
  - This is influenced by DNA (G), the environment (E) and their interaction (GxG ,GxE)
  - Story-telling is easier if there is existing literature /this can also bias a conclusion
  - **Much to discover**