Prediction of missense variant effect

Applications

- Disease gene discovery
- Clinical sequencing // ~11,000 nsSNVs per individual, including rare
- Evolutionary biology
- Protein design

Missense effect is diverse; experiment is not feasible. **What experiment?**

In vivo:

- Clinical impact // rare, context-dependent, inheritance mode
- Model organisms // applicability?

In vitro:

- Functional assay // applicability?

In silico: Damaging | Tolerated, Benign

- Data sources and features
- Prediction methods
- Evaluation

Prediction of missense variant effect

Data sources

Clinical impact . . . . . pathogenic
- ClinVar, HGMD

Biochemical assays. . . . . .functional
- Papers, Protein Mutant Database

Deep mutational scans. . . . . .functional
- Papers, MAVEdb

Population data. . . . . deleterious
- dbSNP, ExAC/gnomAD, other species

Phylogenetic data. . . . . deleterious
- NCBI nr, UniPto UCSC MultiZ

Prediction of missense variant effect

Features

---

1. Substitution

- Conservative / radical (BLOSUM, Grantham score)
- Volume, hydrophobicity change

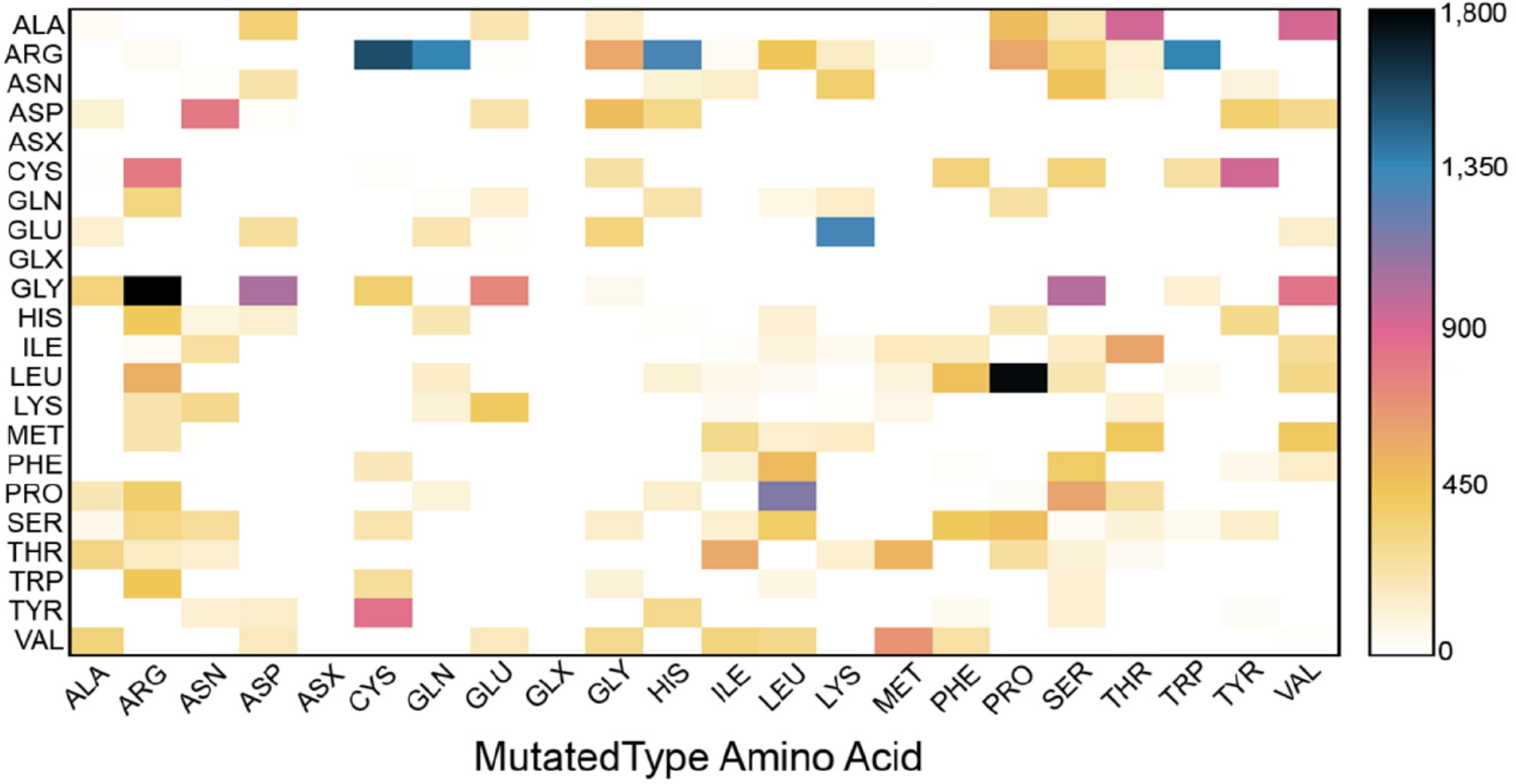
2. Site

- Conservation
- Location: core / surface (Relative Surface Area)
- Contacts: protein, ligand, DNA/RNA
- Secondary structure, disorder
- B-factor

3. Protein

- Number of interactions
- Number of PubMed references

Missense variants in human disease

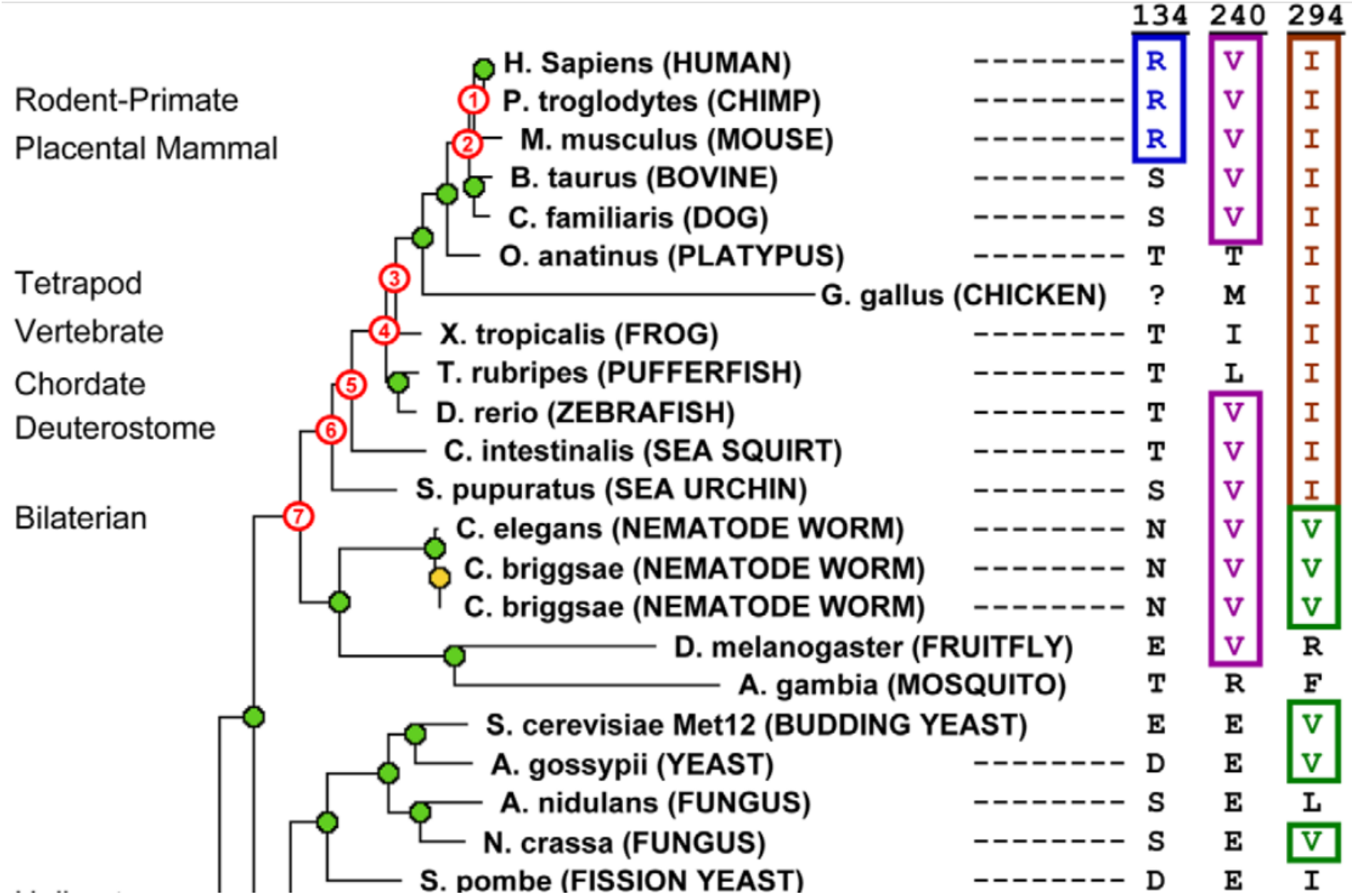


Exercise: list top 10 most frequent disease-causing missense variants

Peterson (2013) *J Mol Biol*

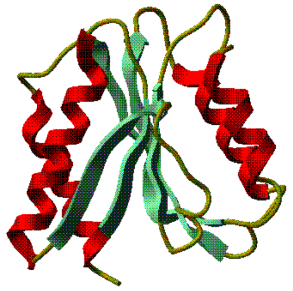
Prediction of missense variant effect

Multiple Sequence Alignment: evolutionary record



Marini (2010) *PLOS Genet*

Prediction of missense variant effect



Protein

Multiple Sequence Alignment

N	E	L	V	T	L	T	C	L	A	R	G	F	S	-	P	K	D	V	L	V	R	W	L
R	E	S	A	T	I	T	C	L	V	T	G	F	S	-	P	A	D	V	F	V	Q	W	M
G	G	S	L	R	L	S	C	V	A	S	G	I	T	-	F	S	G	Y	D	M	Q	W	V
T	P	G	L	T	L	T	C	T	V	S	G	F	S	-	L	S	S	Y	D	M	G	W	V
G	Q	K	A	K	M	R	C	I	P	E	-	-	-	-	K	G	H	P	V	V	F	W	Y
G	Q	E	A	T	L	W	C	E	P	I	-	-	-	-	S	G	H	S	A	V	F	W	Y
G	Q	Q	V	T	L	S	C	F	P	I	-	-	-	-	S	G	H	L	S	L	Y	W	Y
R	K	D	V	S	L	T	C	L	V	V	G	F	N	-	P	G	D	I	S	V	E	W	T
G	Q	K	L	T	L	K	C	Q	Q	N	-	-	-	-	F	N	H	D	T	M	Y	W	Y
R	D	K	A	T	F	T	C	F	V	V	G	S	D	-	L	K	D	A	H	L	T	W	E
S	K	S	A	T	L	T	C	R	V	S	N	M	V	N	A	D	G	L	E	V	S	W	W
G	A	R	T	S	L	N	C	T	F	S	D	-	-	-	S	A	S	Q	Y	F	W	W	Y
G	A	S	L	Q	L	R	C	K	Y	S	Y	-	-	-	S	A	T	P	Y	L	F	W	Y
N	G	A	P	K	L	T	C	L	V	V	D	L	E	S	E	K	N	V	N	V	T	W	N
E	A	T	V	T	L	T	C	V	V	S	N	-	-	A	P	Y	G	V	N	V	S	W	T

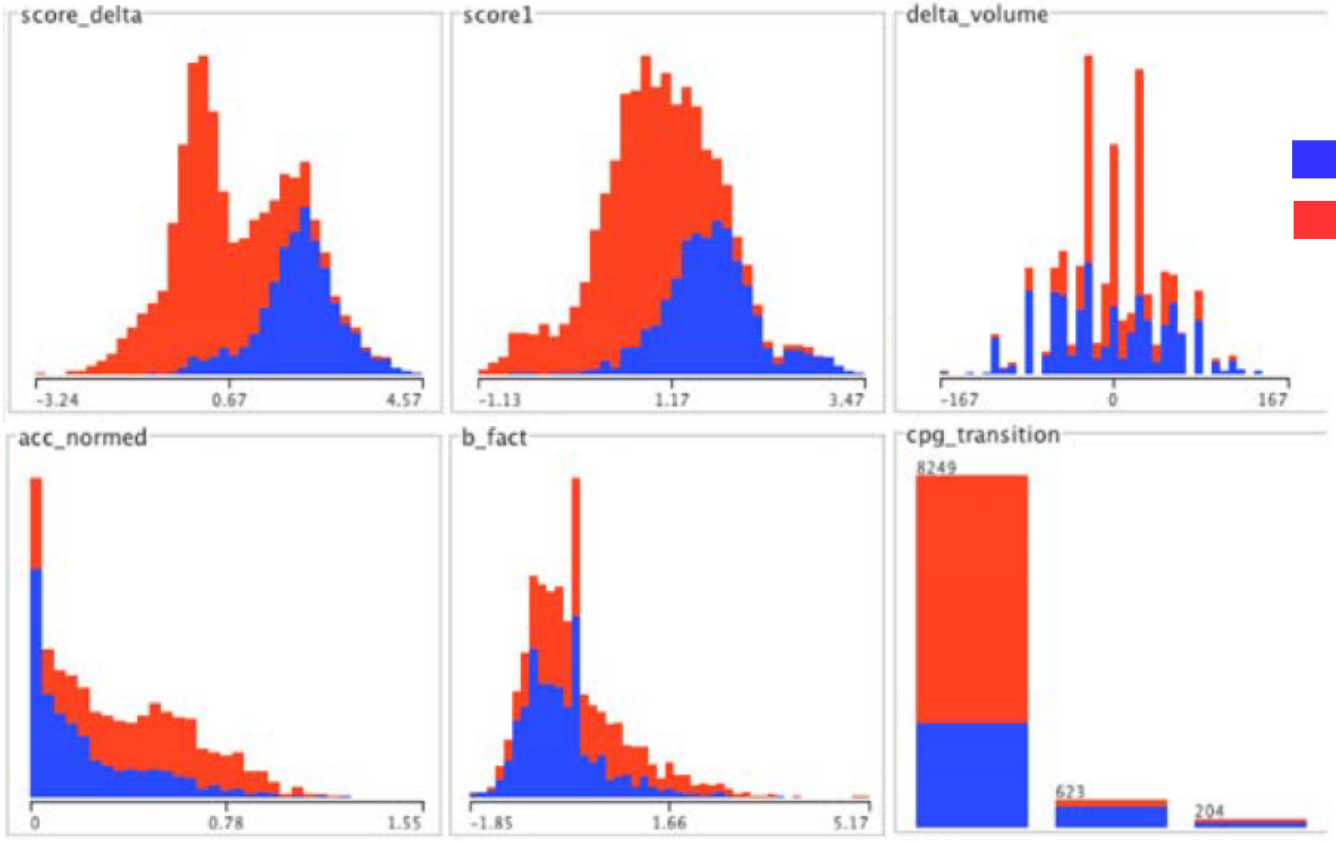
Profile

Ala	-1.2	1.1	-0.6	-0.8	0.3	...	...
Arg	0.6	-0.3	-0.3	-0.5	0.6	...	...
Asn	-1.1	-0.5	-0.5	-0.7	0.4	...	...
Asp	-0.9	-0.3	-0.3	-0.5	0.6	...	...
Cys	0.4	-0.5	0.6	0.8	-0.3	...	...
Gln	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...

PSIC (Position Specific Independent Counts) profile scores matrix

Sunyaev (1999) *Protein Eng*

Prediction of missense variant effect

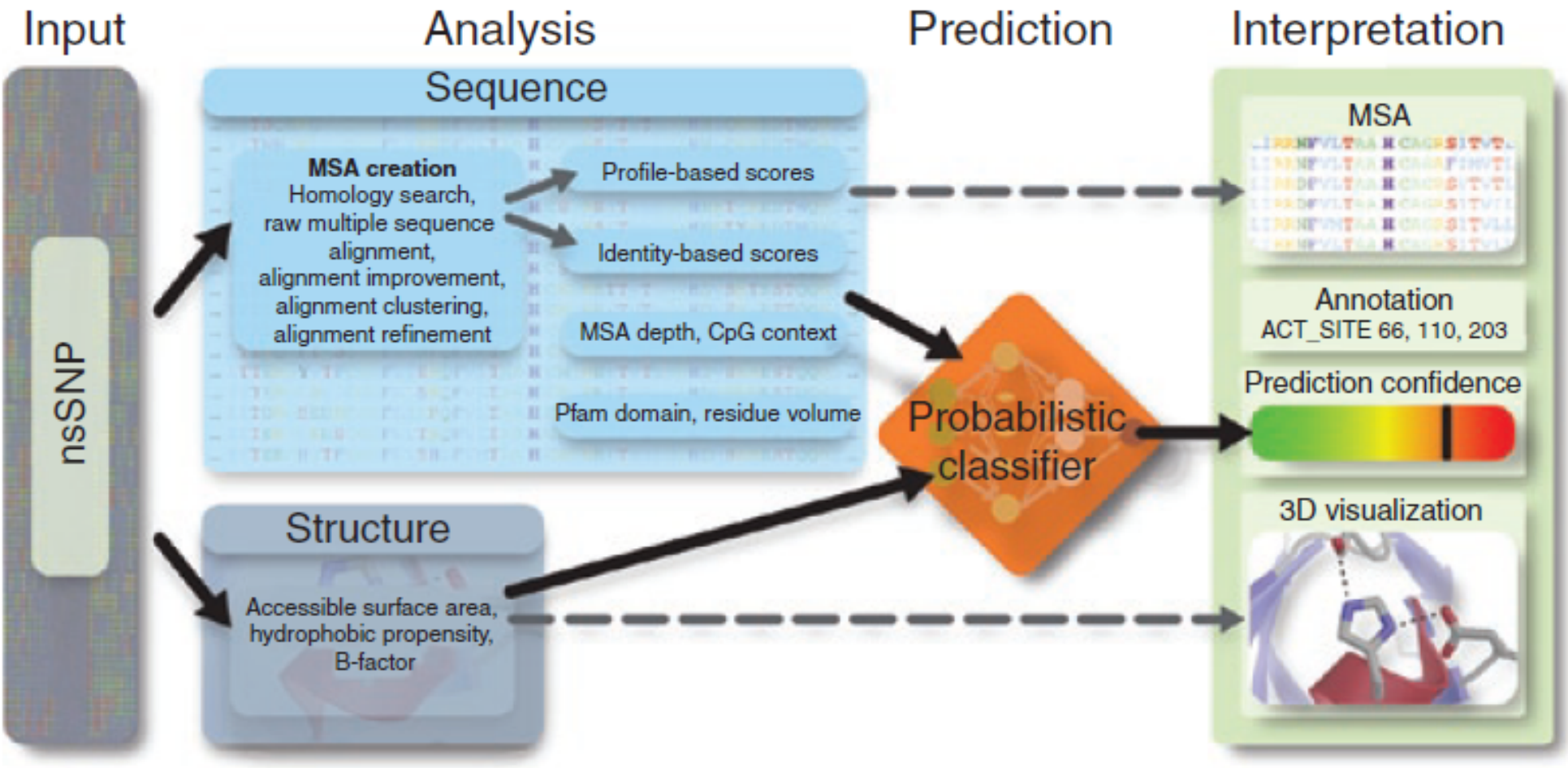


Examples of predictive features used by PolyPhen-2

score_delta: PSIC(AA1)-PSIC(AA2)
score1: PSIC(AA1)
delta_volume: change in side chain volume
cpg_transition: CpG context (0:no, 1: removes CpG, 2:creates)
acc_normed:* normalized accessible surface area // if 3D structure available
b_fact:* average temperature factor

Adzhubei (2010) *Nat Methods*

Prediction of missense variant effect

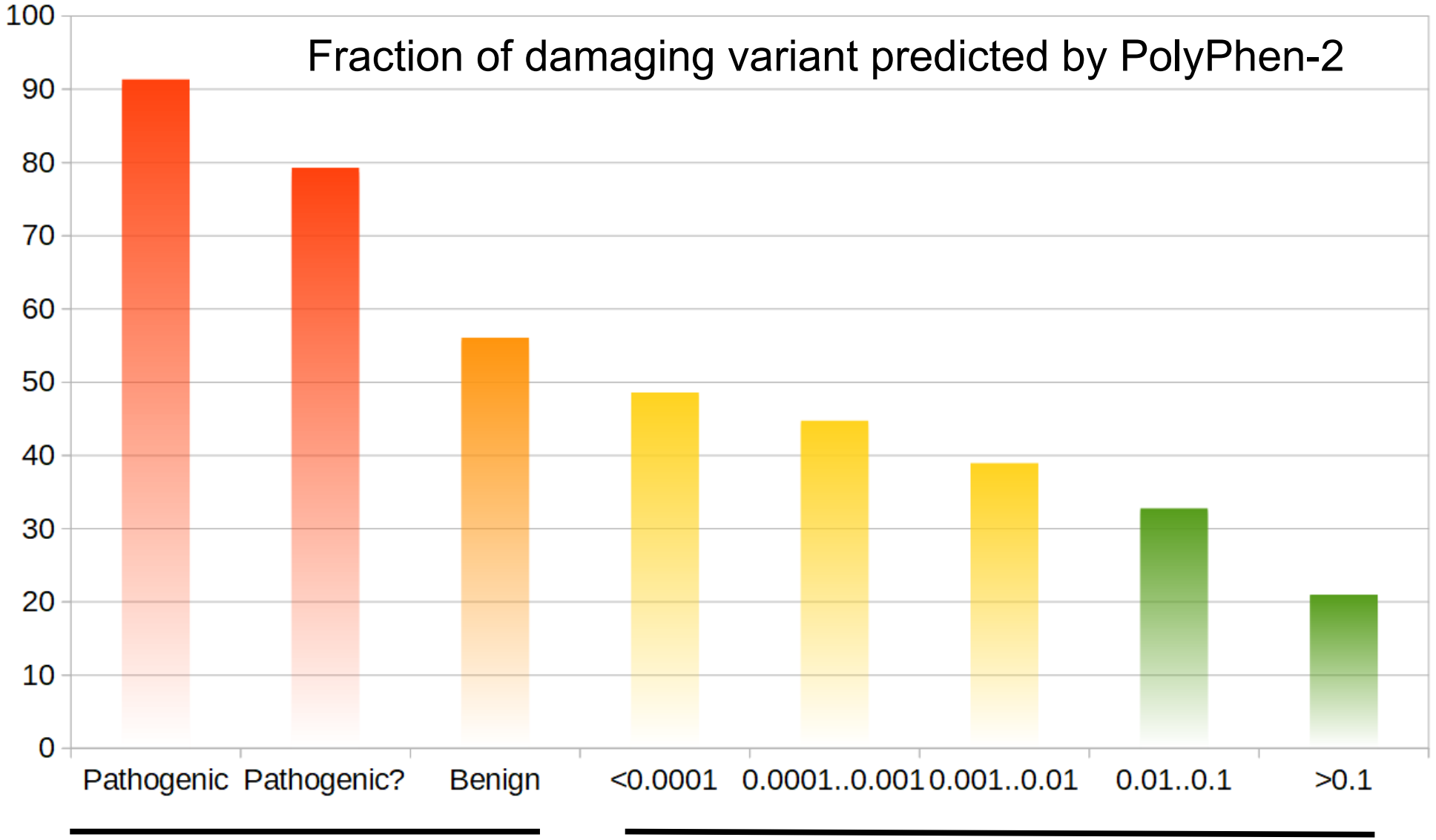


PolyPhen-2 prediction pipeline

Training set (HumDiv): 3,155 disease mutations, 6,321 human-ortholog subst
Performance: FPR=10%, TPR=77%; FPR=20%, TPR=92%

Adzhubei (2010) *Nat Methods*

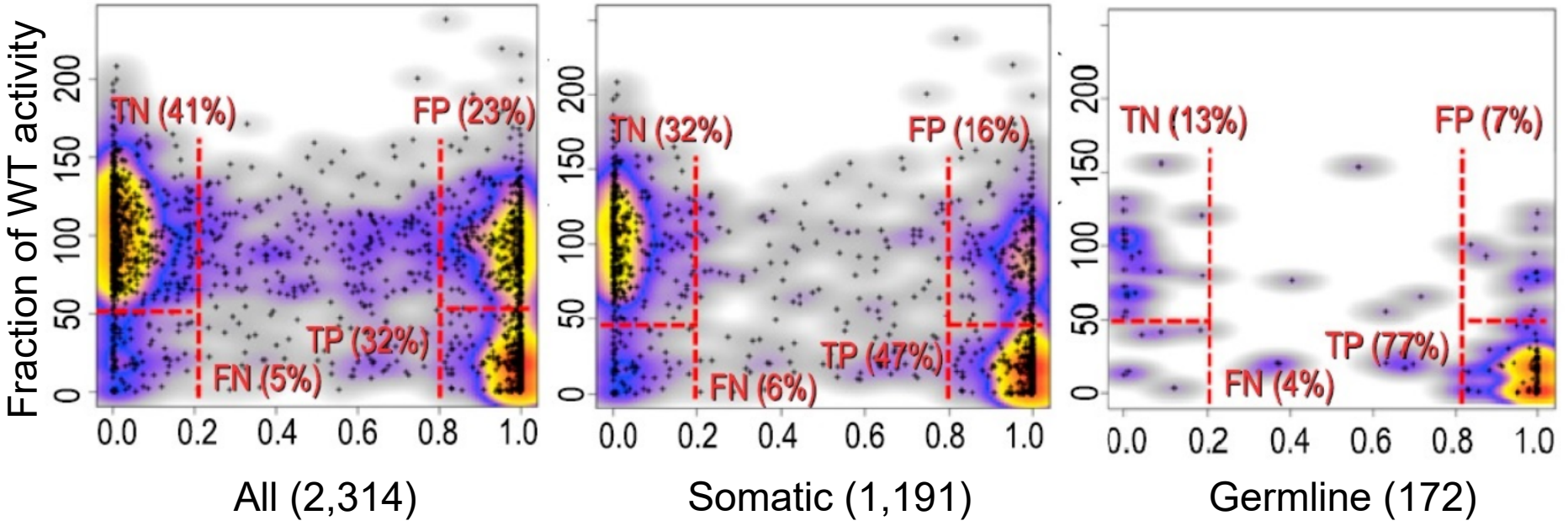
Prediction of missense variant effect



ClinVar: disease mutations

ExAC: population variants by AAF

Prediction of missense variant effect



- Experiment: *in vitro* activity of TP53 compared with predictions by PolyPhen and other tools, threshold: 50% of WT activity
- Low false negative prediction rate, but
- 42% of mutations predicted by PolyPhen2 to be damaging had little measurable consequence for TP53-promoted transcription
- The predictions do not effectively differentiate between mutations that are immediately clinically relevant (ablate or markedly reduce function), and those that are nearly neutral (decrease the function of the corresponding protein by 10%)

What do we predict?

Miosge (2015) *PNAS*

Damaging does not mean pathogenic

Variant prioritization tools such as SIFT (Sorts Intolerant From Tolerant) and PolyPhen2 (Polymorphism Phenotyping v2) use the terms **damaging** and **tolerated** to describe whether a variant is predicted to affect protein function or be functionally neutral, respectively. We emphasize that the term damaging should never be logically equated with causal for a disease phenotype, because a variant that damages a gene is not necessary damaging to an individual's health.

The term **pathogenic** has become widely used to describe a damaging variant that is (potentially) disease-causing. This is straightforward for dominant Mendelian disorders for which pathogenic variants typically cause the disease phenotype but more complex for recessive disorders for which both copies of the gene must harbour variants for pathogenicity (see the figure). Consider a variant producing a stop codon, p.Arg510Ter, in hexosaminidase subunit-α (*HEXA*), which is a gene that is implicated in Tay–Sachs disease. Obviously, this variant changes the transcript in which it resides: the resulting protein is probably nonfunctional due to truncation and may be subject to nonsense-mediated decay. However, this does not mean that it will necessarily be pathogenic to the individual, as many Mendelian diseases such as Tay–Sachs disease, are recessive. Cystic fibrosis is another well-known example, for which the genomes of approximately 1 in 20 healthy Western Europeans contain a damaging variant in the cystic fibrosis transmembrane conductance regulator (*CFTR*) gene. As the disease is recessive, there are no negative health consequences to carriers of damaging variants. For recessive diseases, two copies of the pathogenic variant must be present, or it must be in trans to another pathogenic variant, as a so-called compound heterozygote (see the figure).

Eilbeck (2017) *Nat Rev Genet*

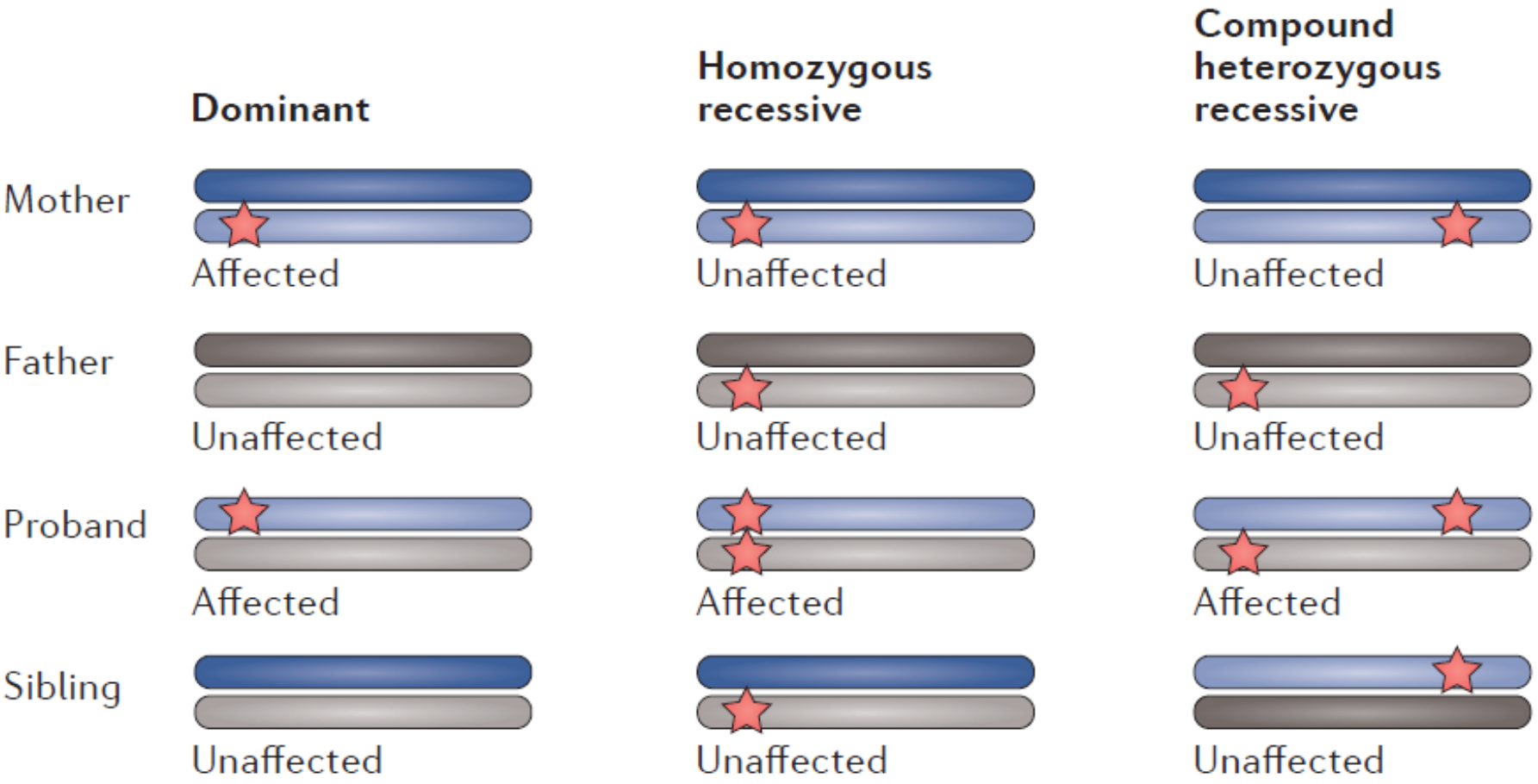
Damaging does not mean pathogenic

The association of damaging variants with pathogenicity has other pitfalls as well. A variant elsewhere in the genome may introduce a seemingly minor and conservative amino acid substitution that may nonetheless damage the patient's health, thereby causing a dominant Mendelian disease. For example, the semi-conservative amino acid-changing variant p.Arg143Gln in the gap junction protein-β2 (*GJB2*) gene is implicated with non-syndromic hearing loss. This variant has been shown in functional studies to encode a protein with impaired function and curated by multiple laboratories in the ClinVar database to be pathogenic.

In a study from 2010, variants implicated in cystic fibrosis and related disorders were assessed using three prediction tools[107]. This study shed light on the differences between predictions and causative alleles. For example, the CFTR variant p.Arg75Gln is predicted to be damaging because it alters a highly conserved position in the protein, but the phenotypic effect is mild. The converse was shown by p.Val520Phe, a deleterious mutation at a non-conserved position in the CFTR protein. In another example, the truncating breast cancer type 2 susceptibility protein (BRCA2) variant p.Tyr791Phe is seemingly damaging — it causes the loss of the 93 C-terminal amino acids of the protein implicated in hereditary breast cancer, but does not cause the disease phenotype (see ClinVar database where it is curated as benign by multiple laboratories and an expert panel). BRCA2 provides another example of the complex relationship between damaging and pathogenic variants. Damaging BRCA2 alleles are typically classified as pathogenic, but they are not immediately disease-causing; instead, they increase cancer risk over a lifetime.

Eilbeck (2017) *Nat Rev Genet*

Damaging does not mean pathogenic



Eilbeck (2017) *Nat Rev Genet*

Prediction of missense variant effect

https://genomeinterpretation.org/vipdb

NEW!

Prediction of missense variant effect

- **Predictions for the whole proteome**: dbNSFP, 84 mln missense and splicing site SNVs
- **Ensemble (meta-) predictors**: MetaSVM, MetaLR, ReVel, M-CAP, etc
- **Neural networks and other ML techniques**: PrimateAI, ~380,000 common missense variants from humans and primates, gradient boosting tree classifier
- **Covariation**: EVmutation accounts for epistasis by explicitly modeling interactions between all the pairs of residues
- **Prediction of quantitative effect**: Envision 21,026 variant effect measurements from 9 large-scale experimental mutagenesis datasets
- **Clinical applicability**: M-CAP, 9 tools, 7 conservation scores, 298 features derived from MSA, gradient boosting tree classifier

Prediction of missense variant effect

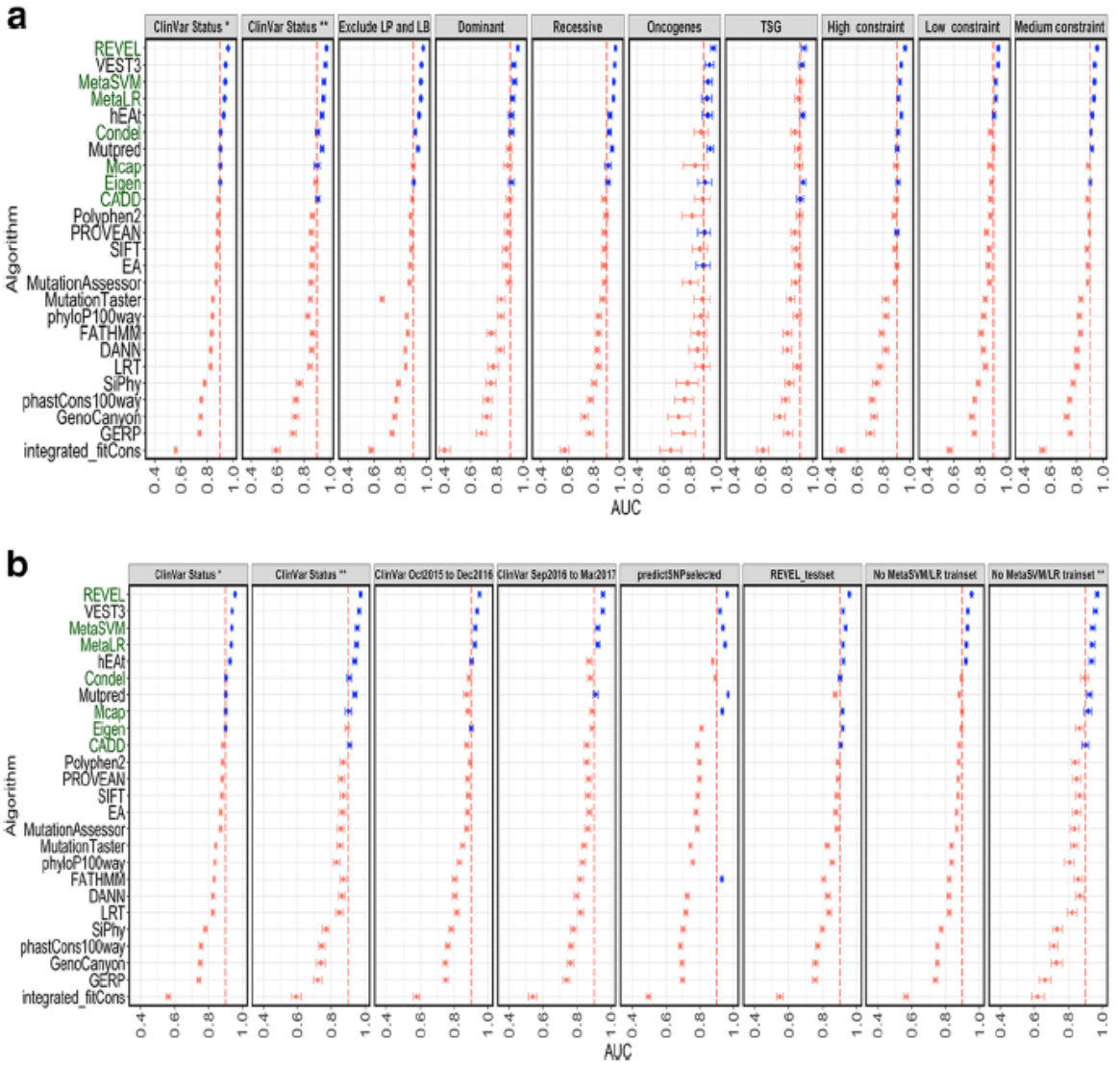
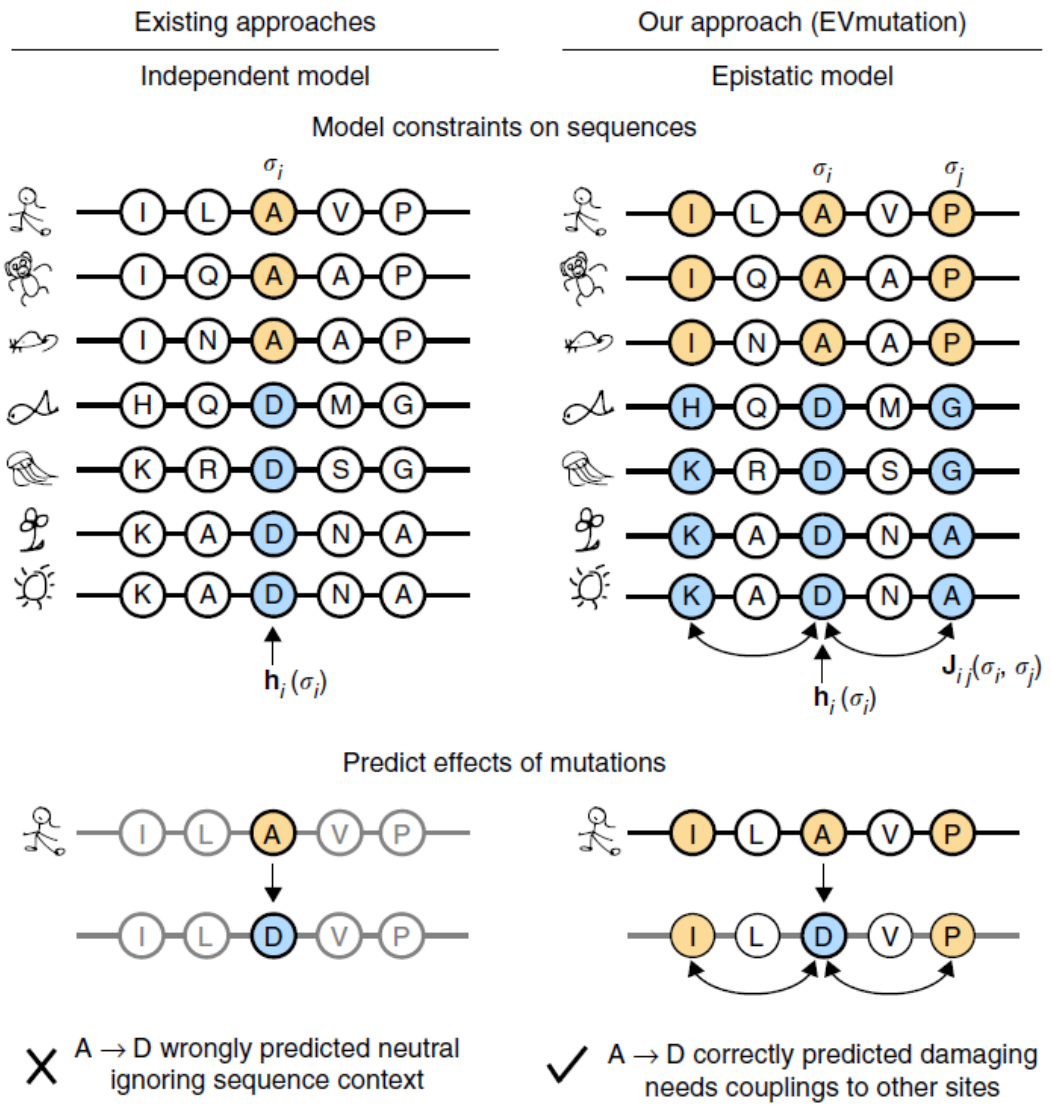


Fig. 3 Performance analysis of algorithms. The AUC of a ROC are plotted for 25 algorithms. *Vertical dotted line* indicates an AUC of 0.9 and 99% confidence intervals for each AUC are shown. *Blue dots* indicate AUC > 0.89. **a** AUCs of the algorithms across different datasets shown in the panels and described in text. **b** AUCs of the algorithms across different datasets (represented in panels) to address type I circularity as described in text. The same plots for ClinVar Status * and ClinVar Status ** as in Fig. 3a are used in 3b for comparison. Any instance of ** represents variants with ClinVar review status of two stars or above. Ensemble predictors are indicated by *dark green labels* on the *y-axis*

Ghosh (2017) *Genome Biol*

Prediction of missense variant effect

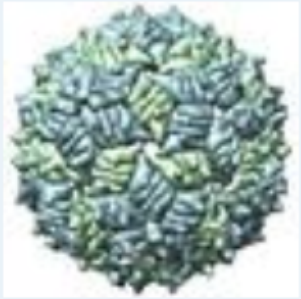


Inferring context-dependent effects of mutations from sequences. Evolution has generated diverse families of proteins and RNAs with varied sequences that perform a common function. An unsupervised probabilistic model trained to generate the natural diversity in a multiple sequence alignment of a family can be used to predict the relative favorability of unseen mutations. Existing models describe functional constraints on each position i in a sequence σ independently, averaging over the effect of background positions j. This can lead to incorrect predictions of neutrality. Our approach infers a global probability model with pairwise interactions between positions i and j (J_{ij}) as well as background biases at single positions (h_i).

Hopf (2017) *Nat Biotech*

Prediction of inframe indels effect

MS2 COAT PROTEIN



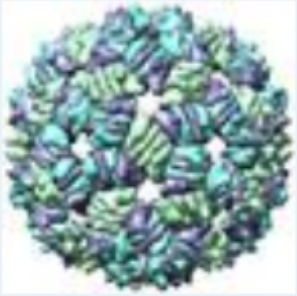
Query:

PDB ID: **2BU1**

Chain ID: A

EC number:

BACTERIOPHAGE FR CAPSID



Subject:

PDB ID: **1FR5**

Chain ID: A

EC number:


```
2BU1.A   61 KVEVPKVATQTVGGVELPVAAWRSYLNMELTIPIFATNSDCELIVKAMQGLLKDGNPIPS 120
            ||||||||||    |||||||||||.|||||||.|||| || |||||.||  | |||| .
1FR5.A   61 KVEVPKVAT----GVELPVAAWRSYMNMELTIPVFATNDDCALIVKALQGTFKTGNPIAT 116
```

Prediction of inframe indels effect

	Insertions, duplications	Deletions
ClinVar, 21 Oct 2019 (hg38)		
Pathogenic, Likely pathogenic	303	1,193
Benign, Likely benign	306	483
Other	1,291	3,566
GnomAD 2.1.1 (hg38)		
AF_POPMAX<1%	30,489	79,023
AF_POPMAX≥1%	742	1,517
Unknown	7,389	10,640
Individual exome (GiaB)	228	275

Q: what is the most “famous” disease-causing inframe indel?

Prediction of inframe indels effect

Gene	*ClinVar*	*gnomAD*
KCNH2 Potassium Voltage-Gated Channel Subfamily H Member 2	Pathogenic (4) Unknown (8)	Rare (11)
PHOX2B Paired Like Homeobox 2B	Benign (7) Pathogenic (4) Unknown (2)	Common (2) Rare/Unknown (14)
CACNA1A Calcium Voltage-Gated Channel Subunit Alpha1 A	Benign (5) Pathogenic (2)	Common (4) Rare/Unknown (42)
FOXC1 Forkhead Box C1	Benign (5) Pathogenic (3) Unknown (4)	Common (2) Rare/Unknown (49)

Prediction of inframe indels effect

Method	Genome version	Coordinates	Implemen-tation	Publi-cation	Last update
VEST-Indel	37, 38	Genome	Web / Local	2016	2019
CADD	37, 38	Genome	Web / Local	2013	2019
SIFT Indel	37, 38	Genome	Web / Local	2013	2016
MutPred-Indel	37 ?	Protein	Web / Local	2019	-
DDIG-in	37	Genome	Web	2013	2017
PROVEAN	37	Genome	Web / Local	2012	2015

Prediction of inframe indels effect

Method	ML	Best features
VEST-Indel	Random forest	Log10 of count of publications in PubMed where gene name is mentioned, Exon Conservation, protein local regional sequence composition
CADD	SVM	cDNApos, ProtPos, PolyPhenVal, SIFTVal, Relative position in coding sequence
SIFT Indel	Decision tree	Repeat, DNA Conservation score, Protein disorder region, Fraction of all Pfam domains affected due to indel
MutPred-Indel	Neural Network	PSSM*, sequence conservation indices, number of homologs in the human and mouse genomes, relative position in protein
DDIG-in	SVM	Disorder, ASA*, DNA Conservation, Neff*, Probabylity of sheet
PROVEAN	Not ML	PROVEAN score

* PSSM - position-specific scoring matrix, ASA - solvent accessible surface area, Neff - number of effective homologous sequences aligned to residues

Prediction of inframe indels effect

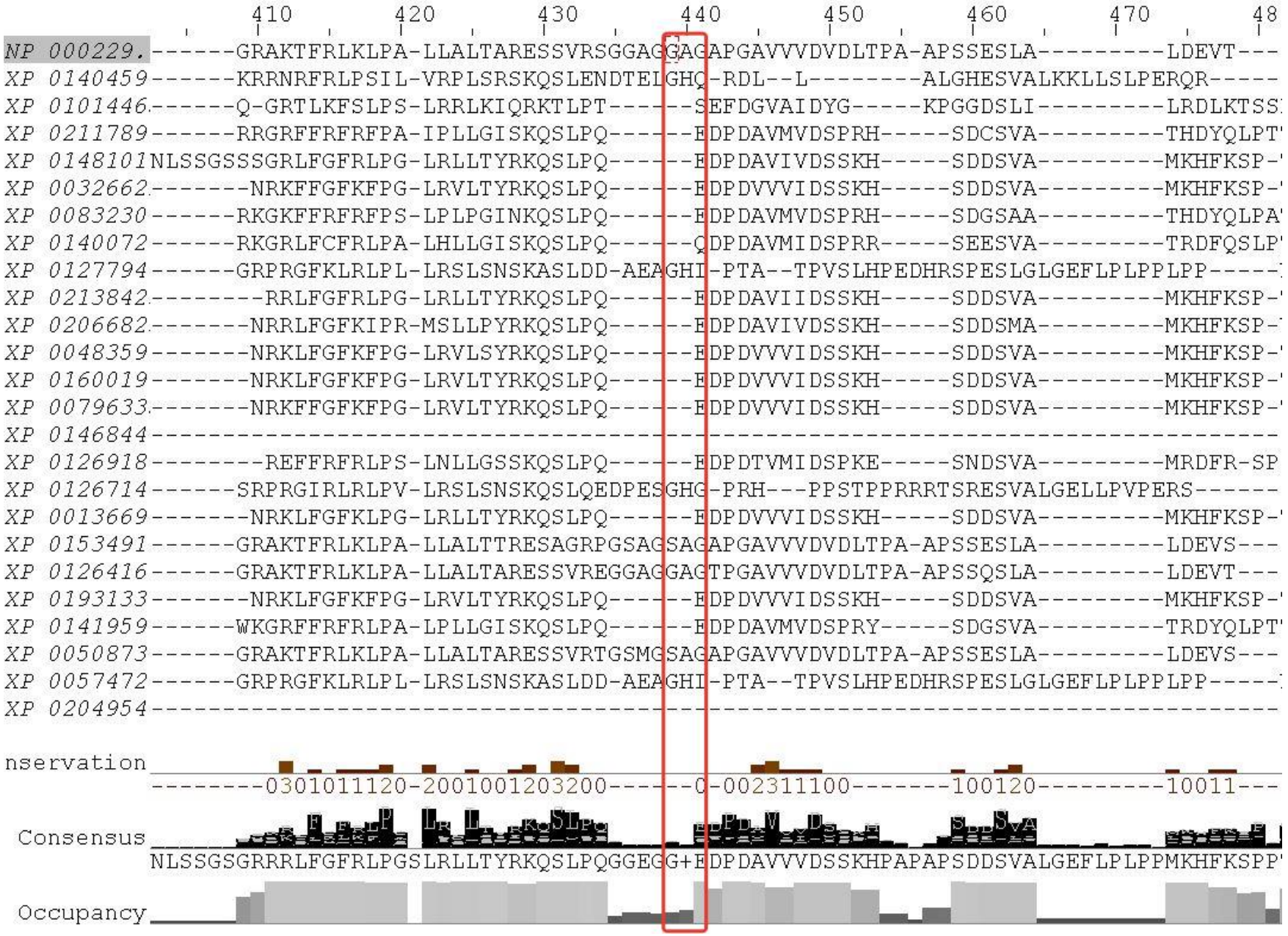
Meta-Predictors that Combine Classifications of Multiple Methods

In these Boolean expressions, each method is represented by a variable X_i, which is set to TRUE when the method classifies an example as pathogenic and FALSE when the method classifies an example as benign. For combinations of two methods, candidate meta-predictors were $(X_1$ and $X_2)$ and $(X_1$ or $X_2)$. For combinations of three methods, candidate meta-predictors $(X_1$ and X_2 and $X_3)$, $(X_1$ or X_2 or $X_3)$,$(X_1$ or X_2 or $X_3)$, $((X_1$ and $X_2)$ or $X_3)$, $((X_1$ or $X_2)$ and $X_3)$, $((X_1$ and $X_3)$ or $X_2)$,$((X_1$ or $X_3)$ and $X_2)$, $((X_2$ and $X_3)$ or $X_1)$, $((X_2$ or $X_3)$ and $X_1)$. For combinations of four methods, there are 64 possible combinations (Supp. Table S4). We used a brute-force approach and limited the number of methods in the meta-predictor to a maximum of four to avoid a combinatorial explosion. All possible four-way combinations of the five methods were explored.

Method	Sensitivity	Specificity	Balanced Accuracy
(VEST-indel AND PROVEAN) OR (CADD AND DDIG-in)	0.930	0.974	0.952
(VEST-indel OR CADD) AND PROVEAN	0.947	0.955	0.951
(VEST-indel OR CADD) AND (PROVEAN OR DDIG-in)	0.947	0.949	0.948
VEST-indel OR (CADD AND PROVEAN AND DDIG-in	0.930	0.955	0.942
VEST-indel OR (CADD AND DDIG-in)	0.930	0.949	0.939
VEST-indel OR (DDIG-in AND CADD)	0.930	0.949	0.939
VEST-indel OR (CADD AND PROVEAN)	0.947	0.929	0.938
(VEST-indel OR DDIG-in) AND PROVEAN	0.930	0.942	0.936

Douville (2016) *Hum Mutation*

Prediction of inframe indels effect

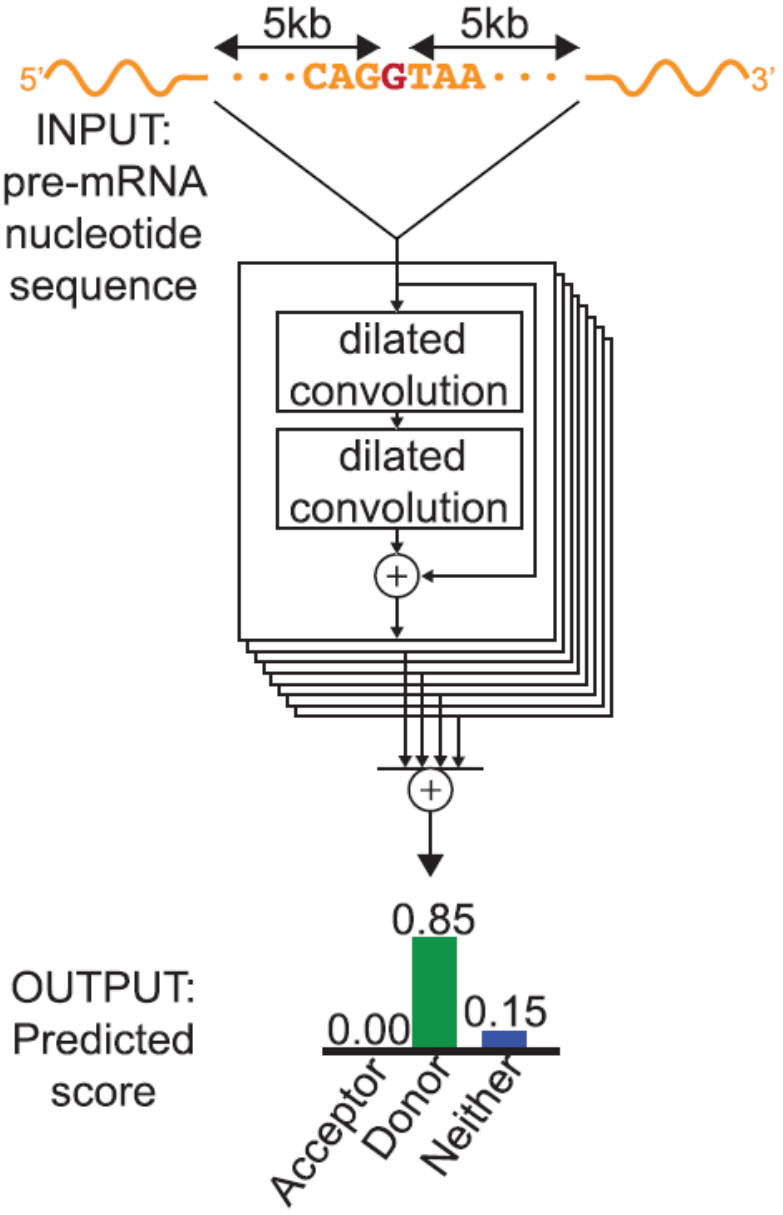


SpliceAI: predicting splicing from sequence

Essential splice variants disrupt canonical splice sites (GT, AG)
Cryptic splice variants: noncoding (intronic, synonymous) variants *outside* the canonical splice sites that disrupt the normal pattern of mRNA splicing

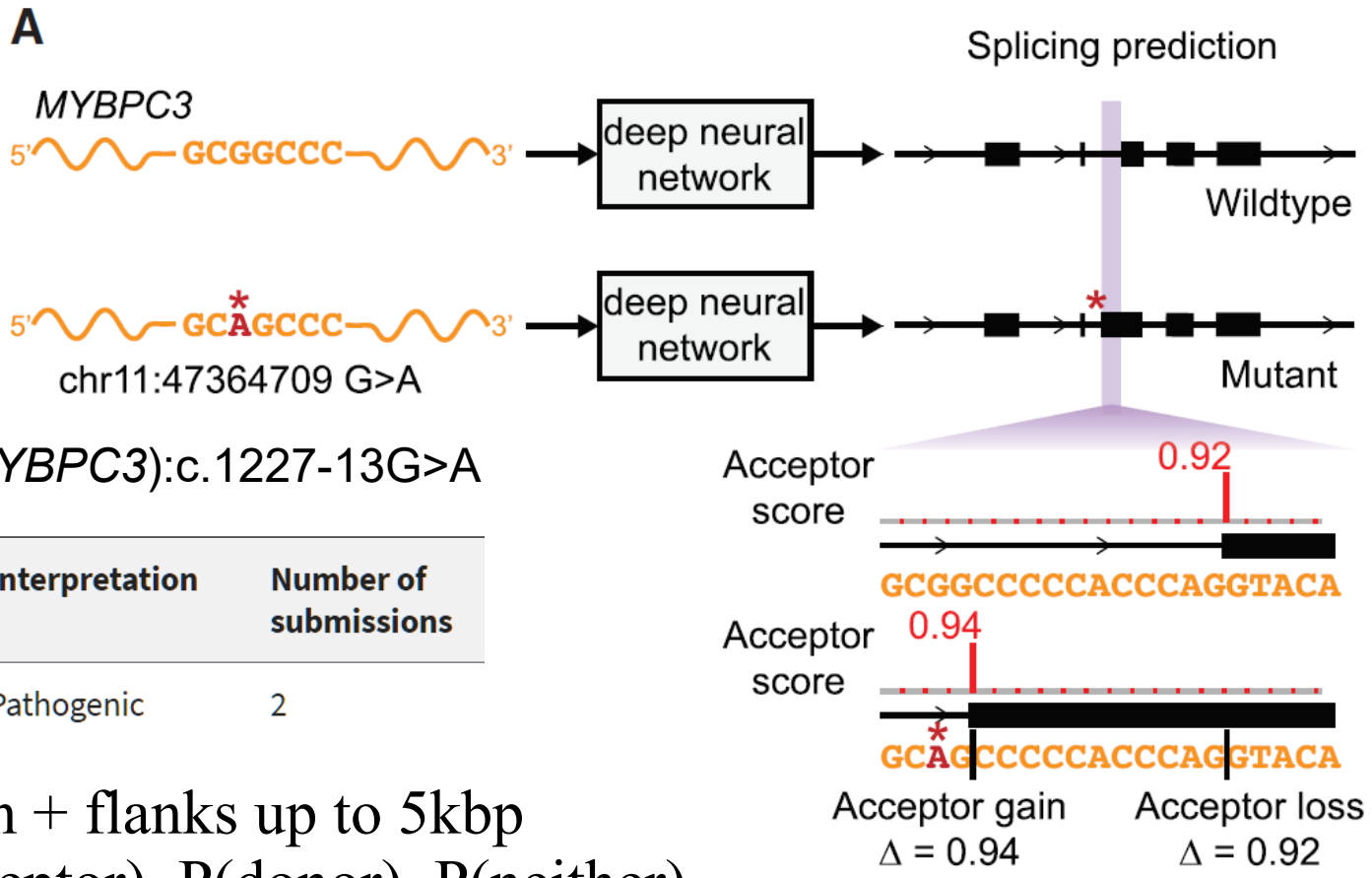
SpliceAI: a 32-layer deep neural network that accurately predicts splice junctions from an arbitrary pre-mRNA transcript sequence

Training set: pre-mRNA transcripts; algorithm learns the context of actual splicing sites



Jaganathan (2019) *Cell*

SpliceAI: predicting splicing from sequence



NM_000256.3(*MYBPC3*):c.1227-13G>A

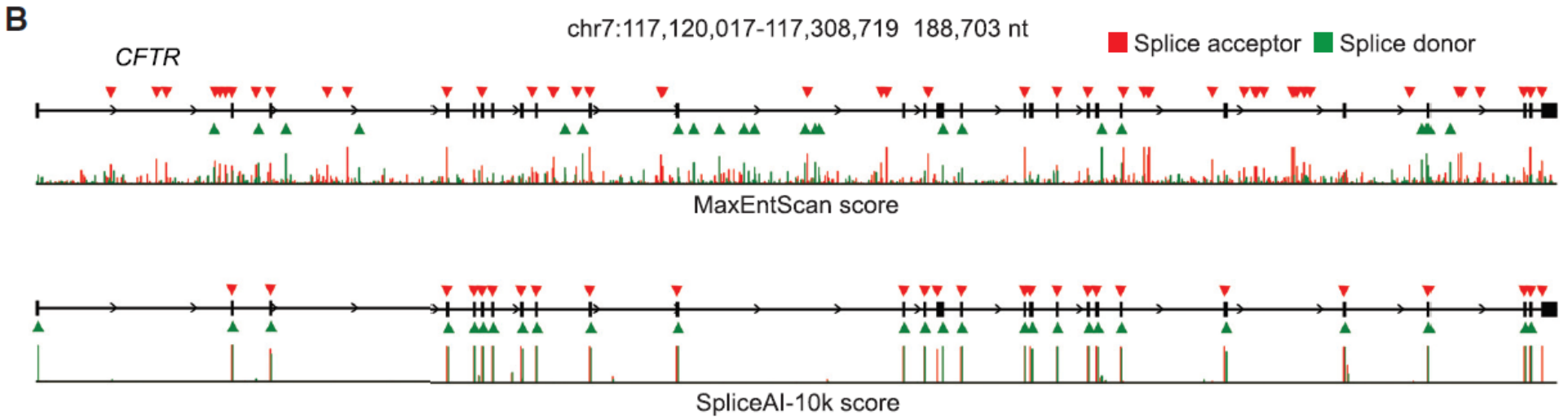
Interpreted condition	Interpretation	Number of submissions
Hypertrophic cardiomyopathy	Pathogenic	2

Input: position + flanks up to 5kbp
Output: P(acceptor), P(donor), P(neither)

SpliceAI-10k predicts acceptor and donor scores at each position in the pre-mRNA sequence of the gene with and without the mutation, as shown here for rs397515893, a pathogenic cryptic splice variant in the MYBPC3 intron associated with cardiomyopathy. The D score value for the mutation is the largest change in splice prediction scores within 50 nt from the variant.

Jaganathan (2019) *Cell*

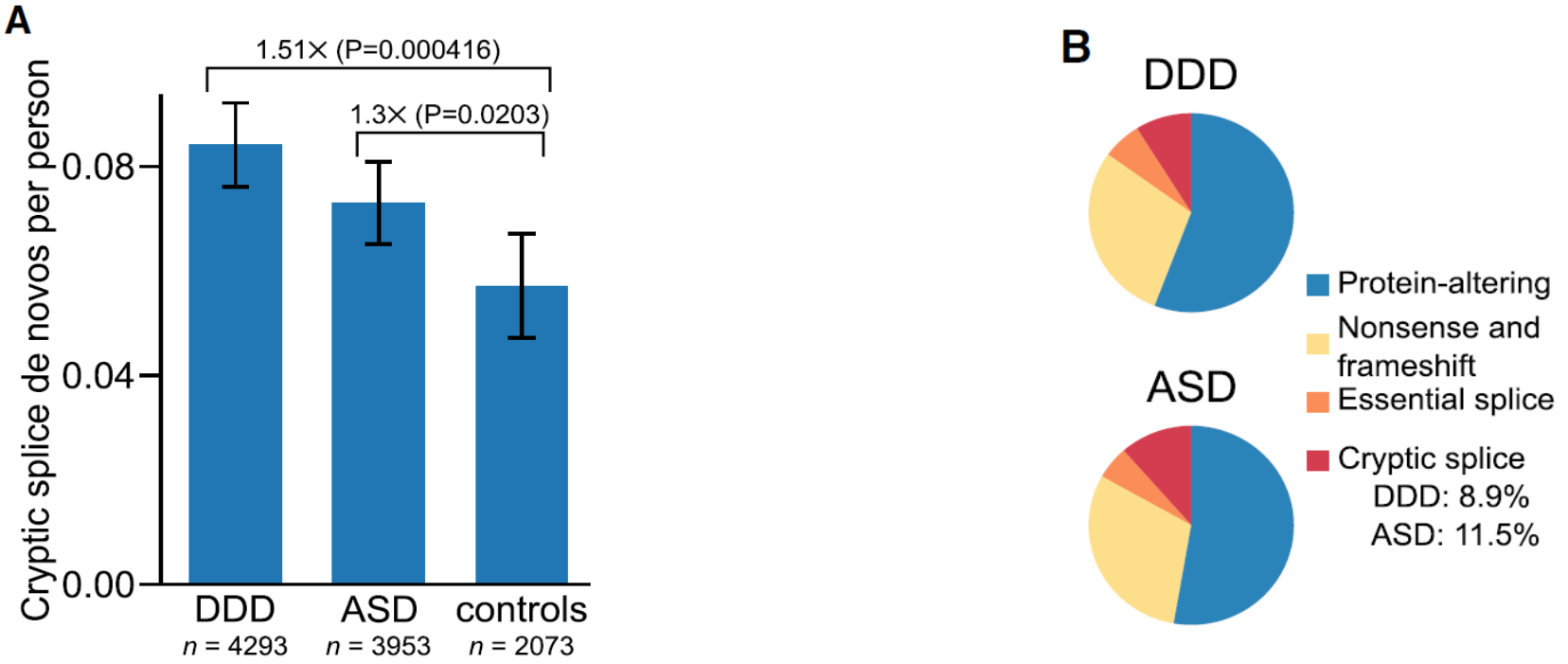
SpliceAI: predicting splicing from sequence



The full pre-mRNA transcript for the *CFTR* gene scored using MaxEntScan (top) and SpliceAI-10k (bottom) is shown, along with predicted acceptor (red arrows) and donor (green arrows) sites and the actual positions of the exons (black boxes). For each method, we applied the threshold that made the number of predicted sites equal to the total number of actual sites.

Jaganathan (2019) *Cell*

SpliceAI: predicting splicing from sequence



(A) Predicted cryptic splice de novo mutations per person for patients from the Deciphering Developmental Disorders cohort (DDD), individuals with autism spectrum disorders (ASDs) from the Simons Simplex Collection and the Autism Sequencing Consortium, as well as healthy controls.
(B) Estimated proportion of pathogenic de novo mutations by functional category for the DDD and ASD cohorts, based on comparison to controls.
Cryptic splicing may yield up to 10% of pathogenic variants in neurodevelopmental disorders

Jaganathan (2019) *Cell*

Regulatory elements in the human genome

Promoter: region (100-1000 bp) at the 5' end of genes where transcription factors and RNA polymerase bind to initiate transcription.

- Proximal promoters typically contain a CpG island
- Methylation of CpG islands silences genes

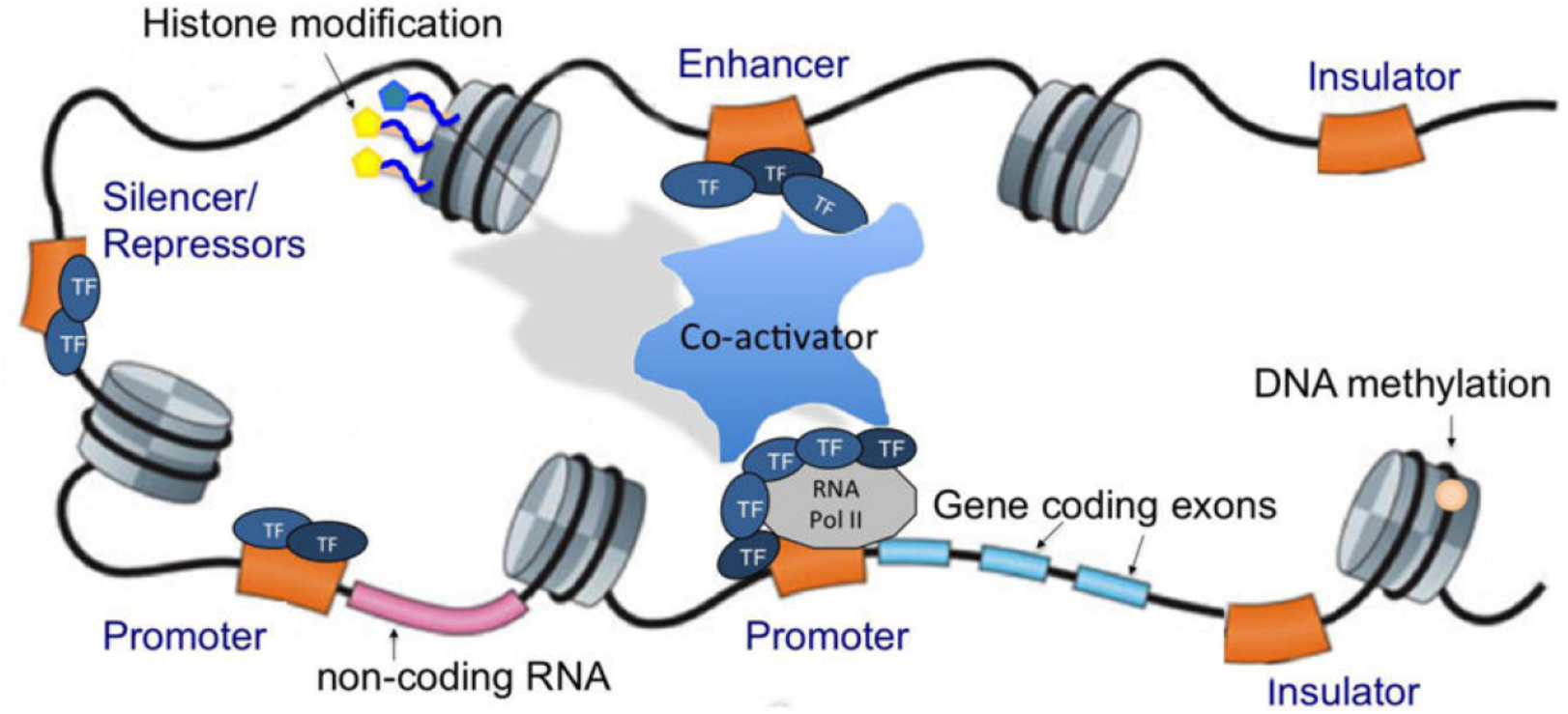
Enhancer: region (50-1500 bp) that binds transcription factors and interact with promoters to stimulate transcription of distant genes (<1Mbp)

- ~10^5 in the human genome (Penacchio 2013 *Nat Rev Genet*)
- Tissue-, time- or cell-specific
- Highly variable location (e.g., intron of an other distant gene)

Transcription factor binding motif/site: short genomic sequence that is known to bind to a particular transcription factor

- 1000-2000 TFs in the human genome
- 400-800 TFBS models (HOCOMOCO v.11)

Regulatory elements in the human genome

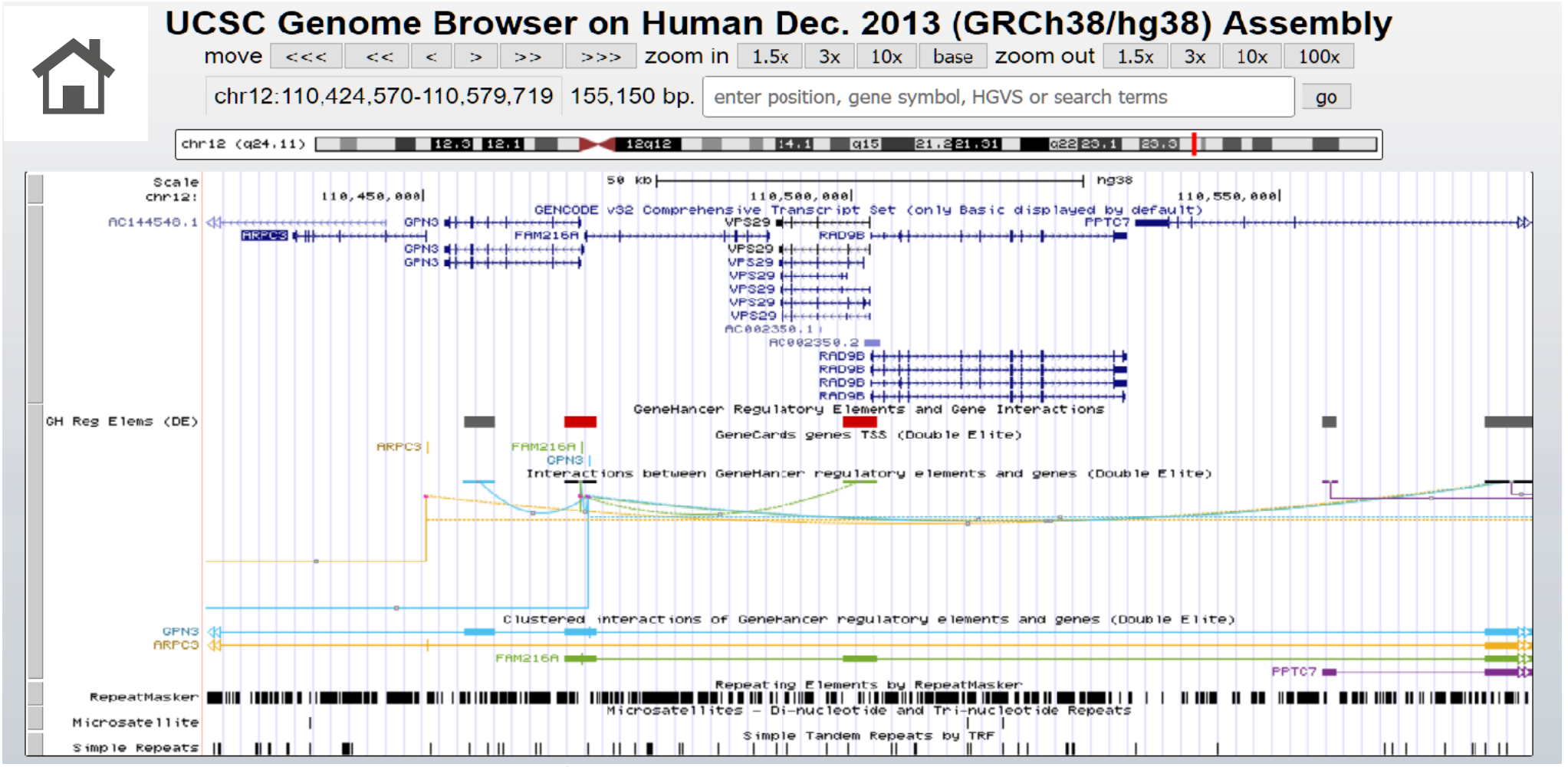


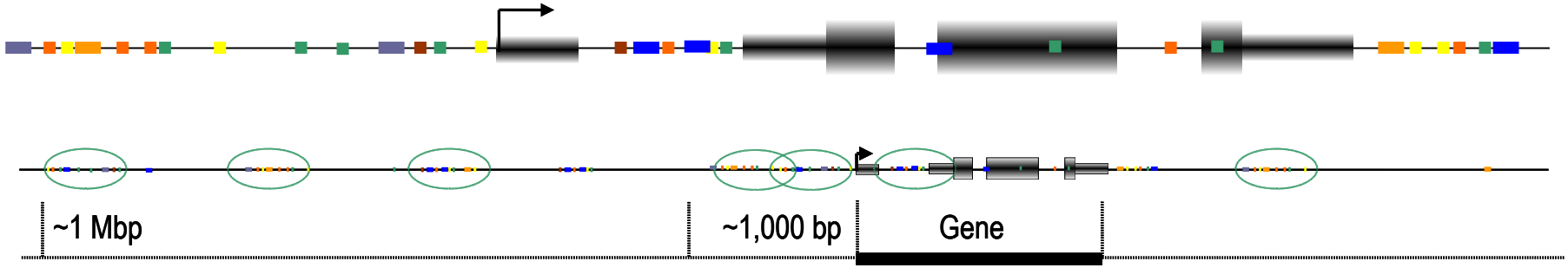
Cis-regulatory elements: **promoters** (100–1000bp) initiate the transcription of a target gene and are located immediately upstream of transcription start sites.

Distal DNA regulatory elements: Enhancers (50–1500bp), silencers, and insulators are DNA regulatory sequences, where transcription factors can bind and regulate expression rates of target genes. A complex of transcription factor and co-activators, mediated by **enhancers**, induce a conformational change of the chromatin structure, allowing the rapid production of specific genes depending on tissue/ cell-type and development-specific contexts. This lies in contrast to co-repressors, which serve to reduce gene expression by attaching to **silencers**. **Insulators** (300–2000bp) establish boundaries of gene expression by mediating loop formation and nucleosome modifications and thus prevent unneeded interactions of both enhancers and silencers with promoters

Lee (2018) *Hum Genet*

Regulatory elements in the human genome





Examples of non-coding functional variants

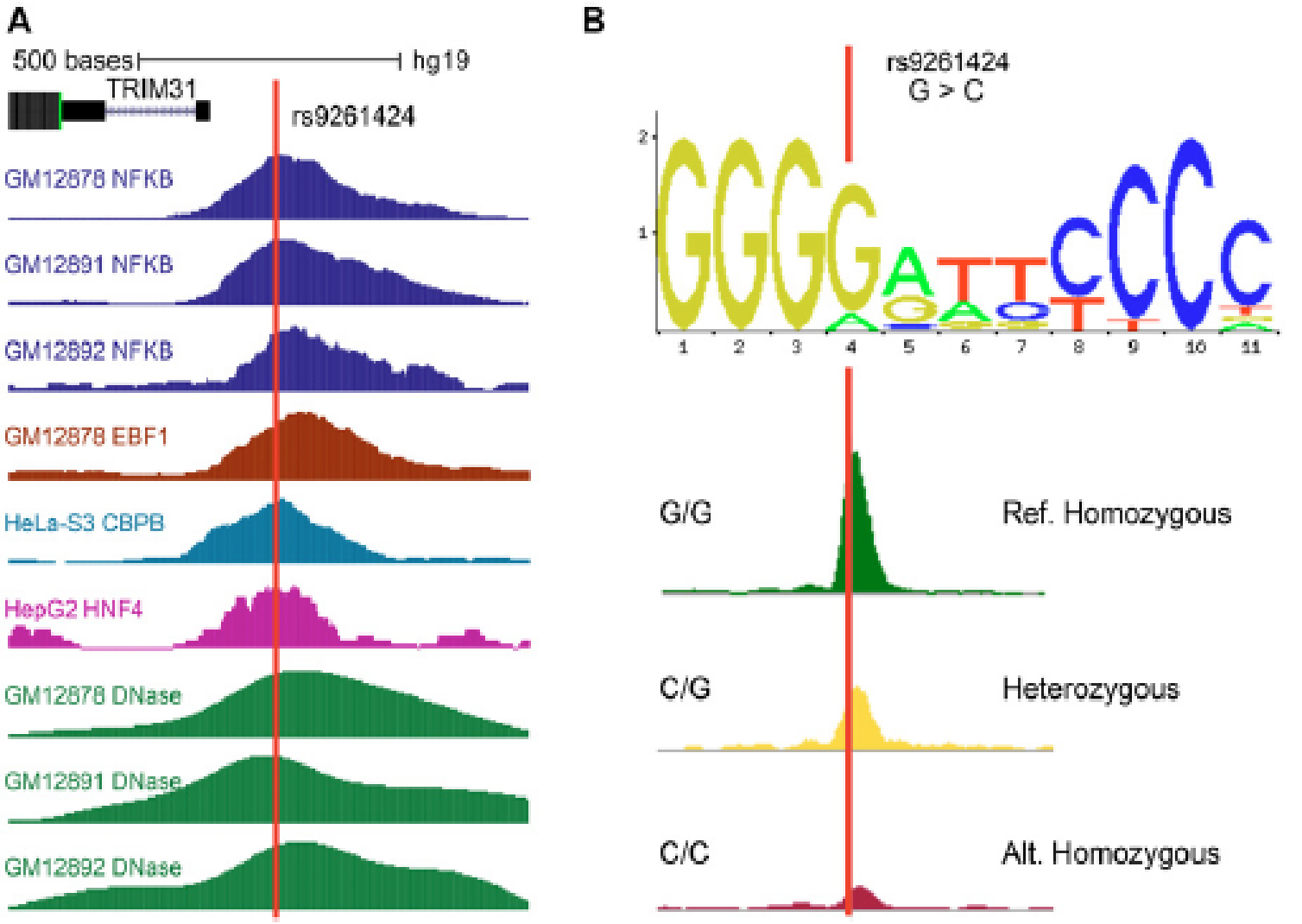
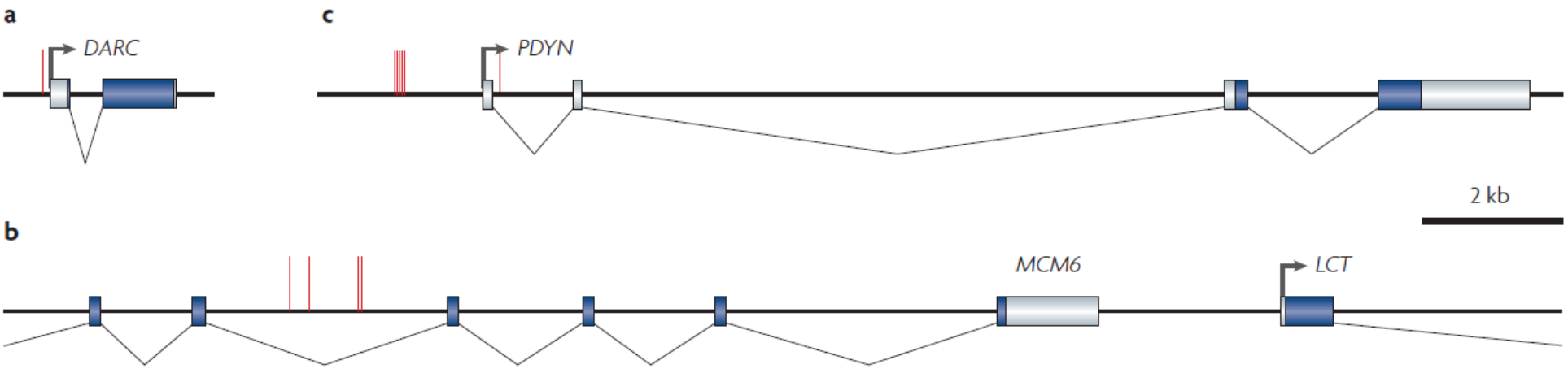


Figure 1. A SNV (rs9261424) overlapping many regulatory features. (*A*) This SNV falls within peak regions for many ChIP-seq factors as well as DNase-seq peaks from multiple cell lines. (*B*) The same SNV overlaps a motif match to the NFKB motif and has been shown to alter binding. The signal tracks represent ChIP-seq peaks of NFKB at the SNV site for three individuals: homozygous to reference allele (*G*), heterozygous, and homozygous to alternate allele (*C*) (Kasowski et al. 2010).

Boyle (2012) *Genome Res*

Examples of non-coding functional variants



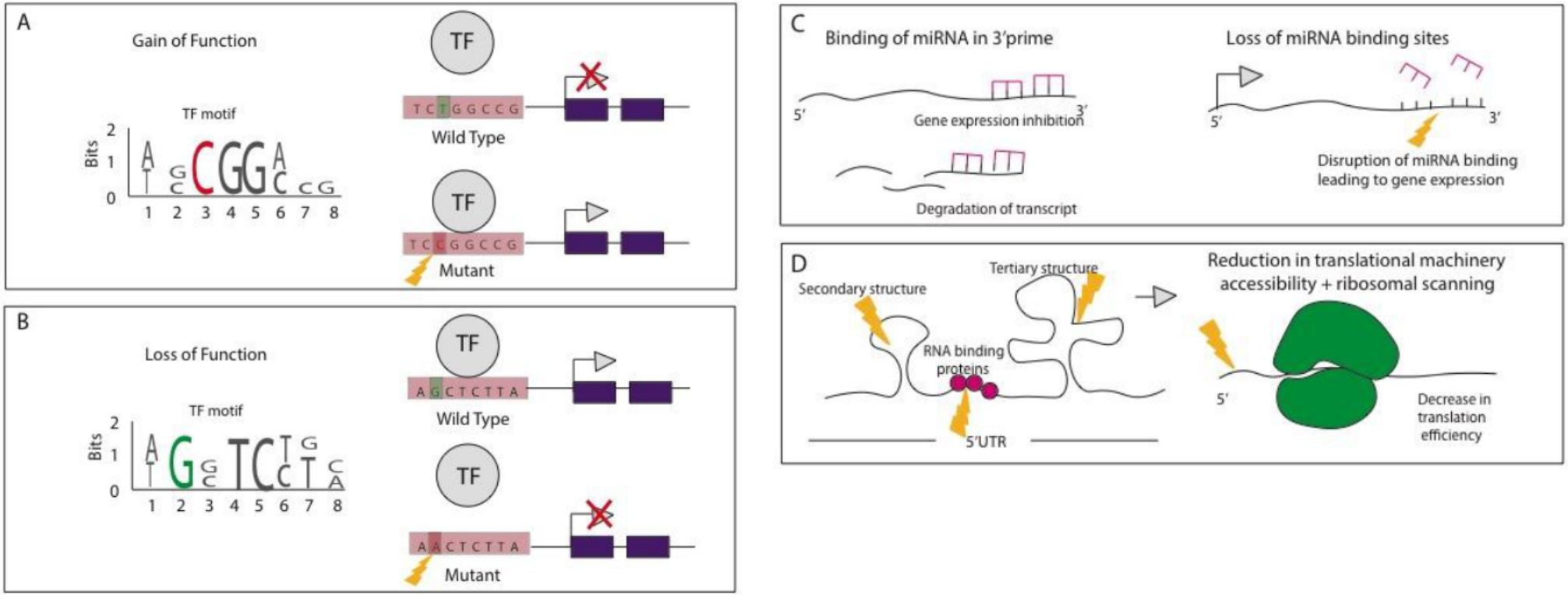
(a) Atypical chemokine receptor 1 *ACKR1 (DARC)*: mutations disrupt *GATA1* binding site ⟹ no expression in erythrocytes ⟹ no point of entry for the malarial parasite *Plasmodium vivax*

(b) Lactase *LCT*: mutations in *MCM6* intron elevate *LCT* transcription, allowing digestion of lactose

(c) Prodynorphin *PDYN*: precursor of neuropeptide dynorphin, implicated in SCZ, BP, temporal lobe epilepsy. Human-branch specific mutations (5+1) regulate constitutive and induced expression, respectively

Wray (2007) *Nat Rev Genet*

Examples of non-coding functional variants



(**A**) Mutations within promoter (e.g., *TERT*) and enhancer regions (*TAL1*) can create transcription factor (TF) binding motifs in a gain-of-function manner allowing the binding of transcriptional activators (**B**) Alternatively, mutations within regulatory regions can create the loss of transcription factor binding sites, leading to transcriptional repression (**C**) miRNA binding within the 3' UTR control gene expression, by inhibiting translation or marking transcripts for degradation. Mutations that disrupt these binding sites can lead to over-expression (*NFKBIE* and *NOTCH1* genes in cancer) (**D**) Mutations within the 5' UTR can alter the secondary and tertiary structures, as well as trans-acting RNA binding protein sites. These alterations can affect translation efficiency and mRNA stability (*BRCA1* and *CDKN2A* genes)

Patel (2018) *High-Throughput*

Examples of non-coding functional variants

The *NOS1AP* gene on human chromosome 1q has been long known to be associated with variability of **QT interval and cardiac repolarization**, whereas the underlying mechanism was unclear. A recent study utilized high-coverage resequencing and regional association for fine mapping in the GWAS locus for QT interval variation, which identified **210 common non-coding risk variants**. Further enhancer/suppressor analysis of 12 selected variants located in cardiac phenotype associated DNaseI hypersensitivity sites assisted in the identification of an upstream enhancer variant (rs7539120) associated with QT interval. This variant can affect cardiac function by increasing *NOS1AP* transcript expression in cardiomyocyte-intercalated discs and increase risk of cardiac arrhythmias.

Similar evidence for functional enhancer SNPs has also been observed at many other loci, including the intronic enhancer SNPs at the *MEIS1* gene associated with **restless legs syndrome** and at the *BCL11A* gene associated with fetal hemoglobin levels, the intergenic enhancer SNP upstream to the *MYB* gene that is a critical regulator of erythroid development and fetal hemoglobin levels, and the recessive mutations in a distal enhancer located 25 kb downstream of *PTF1A* that is associated with **isolated pancreatic agenesis**.

Zhang (2015) *Hum Mol Genet*

Examples of non-coding functional variants

A recent study on the **schizophrenia**-associated locus at 1p21.3 identified a rare enhancer SNP (chr1:98515539A>T, hg19) with increased risk. The chromatin conformation capture assay showed that this risk allele has no obvious influence on the neighboring genes such as *DPYD*, but can reduce the expression of non-coding genes MIR137/MIR2682.

In some instances, such functional variants are located in either the 5′ or 3′ untranslated region (UTR) of the disease-associated genes. A recent study identified the association of rs11603334 (a SNP located in the 5′ UTR of *ARAP1*) with **fasting proinsulin and type 2 diabetes**. The allele-specific expression assay in human pancreatic islet samples showed that the risk allele of rs11603334 can upregulate gene expression of *ARAP1* by 2-fold, which is also supported by the observation of decreased binding of pancreatic beta cell transcriptional regulators *PAX6* and *PAX4* to the rs11603334 risk allele and its corresponding increased promoter activity.

In the case of **hypertriglyceridemia**-associated *APOA5*, the 3′ UTR SNP rs2266788 was predicted to create a potential miRNA binding site for liver-expressed miR-485-5p. Luciferase reporter assays in both HEK293T cells with a miR485-5p precursor and in HuH-7 cells with endogenously expressed miR-485-5p suggested that the mutant allele of rs2266788 is involved in the miR-485-5p-mediated downregulation of *APOA5*.

Zhang (2015) *Hum Mol Genet*

Prediction of non-coding variant effect

CADD: Combined Annotation–Dependent Depletion integrates diverse genome annotations and scores *any possible* human single-nucleotide variant (SNV) or small insertion-deletion (indel) event

«Deleterious variants—that is, variants that reduce organismal fitness—are depleted by natural selection in fixed but not simulated variation»

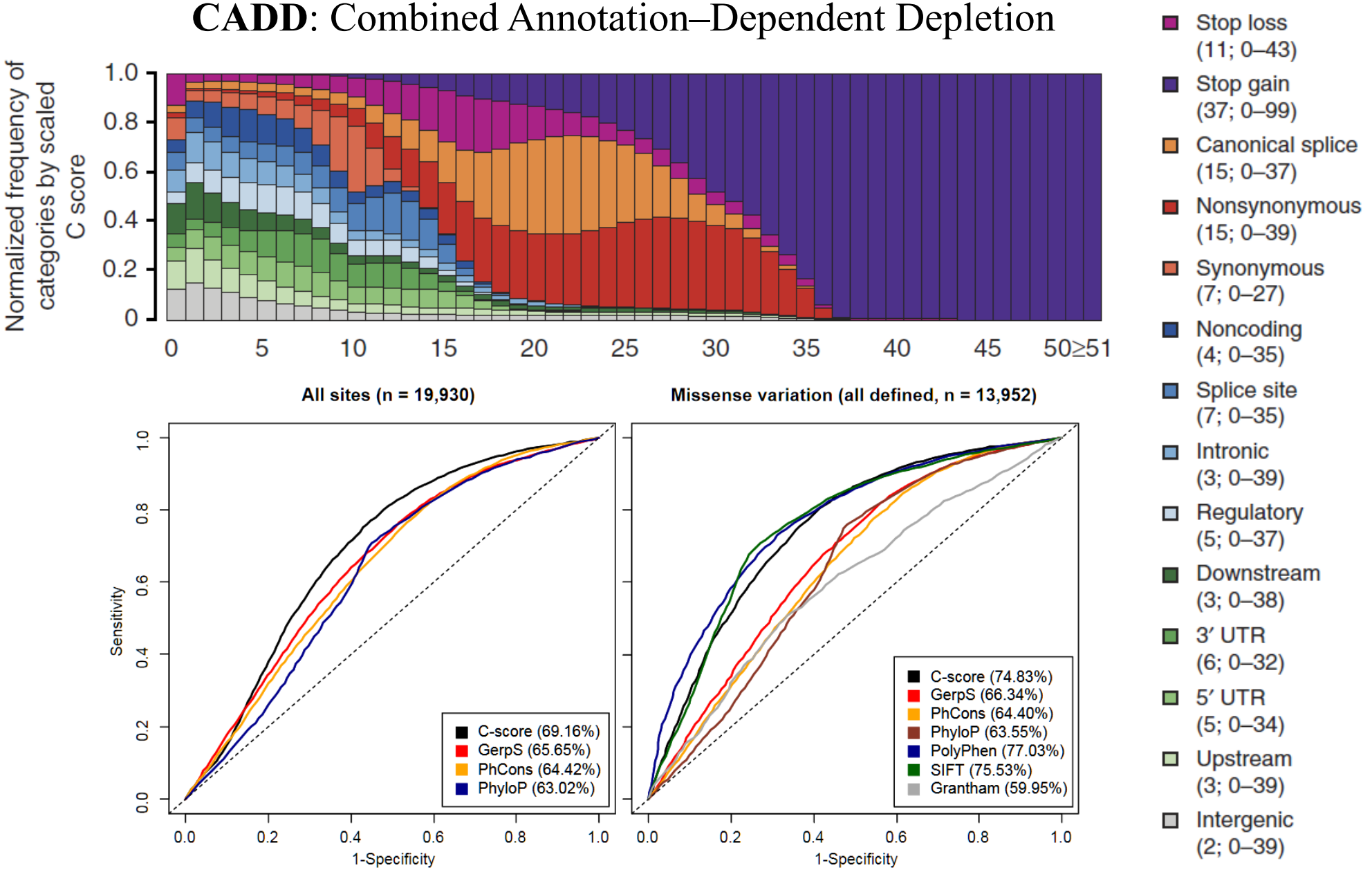
Observed variants (15 mln SNVs, 0.63 mln insertions and 1.1 mln deletions):
– human-chimp differences; SNPs with MAF>5% excluded
– SNPs with DAF (derived allele frequency) > 95% (<5% of total)
Simulated variants (44 mln SNVs, 2.1 mln insertions and 3.1 mln deletions):
– a fully empirical model of sequence evolution with a separate rate for CpG dinucleotides and local adjustment of mutation rates

Features: VEP annotation, SIFT, PolyPhen-2, conservation scores, ENCODE methylation and histone modification annotation in various cell/tissue types, TF binding sites, etc.
Output: C-scores that measure deleteriousness for 8.6×10^9 variants

Kircher (2014) *Nat Genet*

Prediction of non-coding variant effect

CADD: Combined Annotation–Dependent Depletion



ClinVar pathogenic vs population variants with matched annotation

Kircher (2014) *Nat Genet*

Prediction of non-coding variant effect

Score	Data sources	Approach
Eigen	• Uses data from the ENCODE and Roadmap Epigenomics projects	• Weighted linear combination of individual annotations • Unsupervised learning method
FunSeq2	• Inter- and Intra-species conservation • Loss- and gain-of-function events for transcription factor binding • Enhancer–gene linkage	• Weighted scoring system
LINSIGHT	• Conservation scores (phastCons, phylopP), predicted binding sites (TFBS, RNA), regional annotations (ChIP-seq, RNA-seq)	• Graphical model • Selection parameter fitting using generalized linear model based on 48 genomic features
CADD	• Ensembl variant effect predictor • Protein-level scores: Grantham, SIFT, PolyPhen • DNase hypersensitivity, TFBS, transcript information • GC content, CpG content, histone methylation	• Support vector machine
FATHMM	• 46-way sequence conservation • ChIP-seq, TFBS, DNase-seq • FAIRE, footprints, GC content	• Hidden Markov models
ReMM	• Predict potential of non-coding variant to cause a Mendelian disease if mutated • 26 features: PhastCons, PhyloP, CpG, GC, regulation annotations	• Random forest classifier
Orion	• Predict potential of non-coding variant to cause a Mendelian disease if mutated • Independent from annotation and features	• Expected and observed site-frequency spectrum of a given stretch of sequence
CDTS	• Identify constrained non-coding regions in the human genome and deleteriousness of variants • Independent from annotation and features. Uses *k*-mers	• Expected and observed site-frequency spectrum of a given heptamer

Telenti (2018) *Hum Mol Genet*

Prediction of non-coding variant effect

Table 2 Summary of genomic features used for LINSIGHT scores

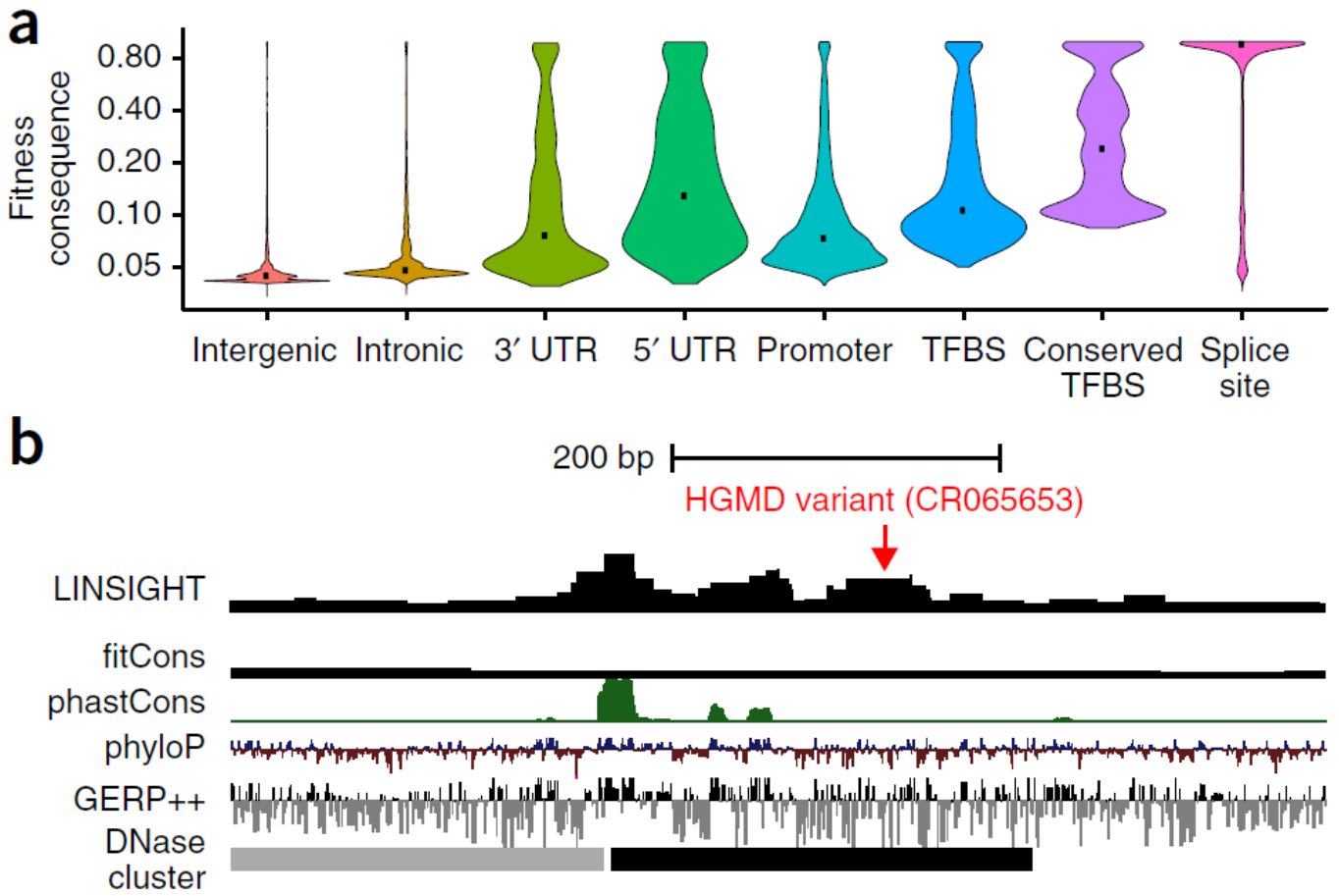
Class	Genomic feature[a]	Spatial resolution
Conservation	phyloP score	High
	phastCons element	High
	SiPhy element	High
	CEGA element	High
Binding site	Conserved TFBS	High
	rVISTA TFBS	High
	SwissRegulon TFBS	High
	Predicted TFBS within ChIP-seq peak	High
	Conserved miRNA binding site	High
	Splicing site predicted by SPIDEX	High
Regional annotation	ChIP-seq peak of transcription factor	Low
	DNase-I hypersensitive site	Low
	UCSC FAIRE peak	Low
	RNA-seq signal	Low
	Histone modification peak	Low
	FANTOM5 enhancer	Low
	Predicted distal regulatory module	Low
	Distance to nearest TSS	Low

[a]Each 'genomic feature' listed here may actually correspond to multiple features in the model. For example, four features are derived from phyloP scores: two from the mammalian phyloP scores and two from the vertebrate phyloP scores. See **Supplementary Table 3** for complete details.

LINSIGHT integrates functional genomic data together with conservation scores and other features to provide a high-powered, high-resolution measure of potential function.

Huang (2017) *Nat Genet*

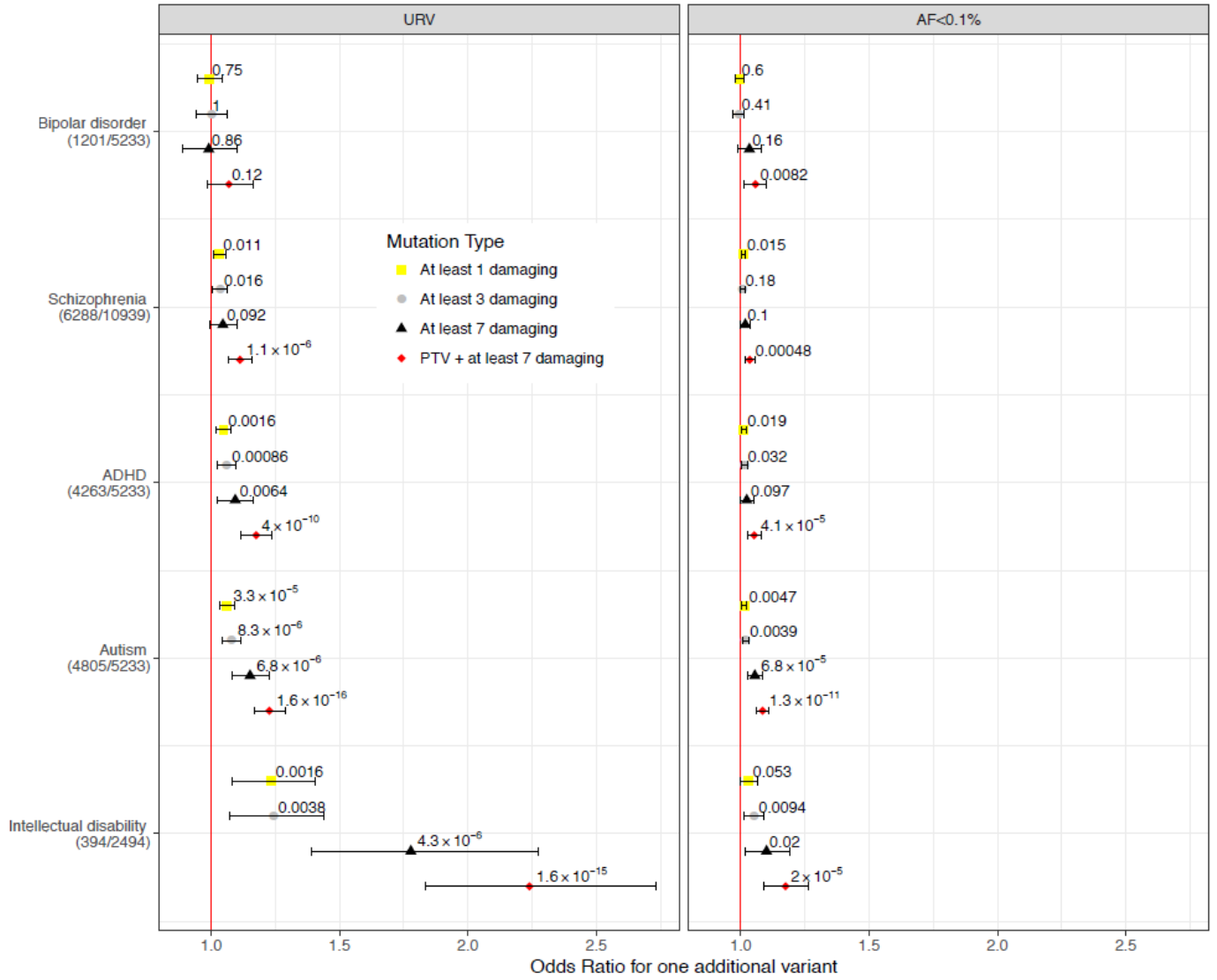
Prediction of non-coding variant effect



(**a**) Distributions of LINSIGHT scores for various genomic regions. Intergenic regions, intronic regions, UTRs, and 1-kb promoters: GENCODE 19; TFBSs: ChIP-seq peaks (Ensembl Regulatory Build); conserved TFBSs: UCSC Genome Browser. (**b**) LINSIGHT is the only method to highlight a variant from HGMD (CR065653) that is associated with upregulation of the *TERT* gene.

Huang (2017) *Nat Genet*

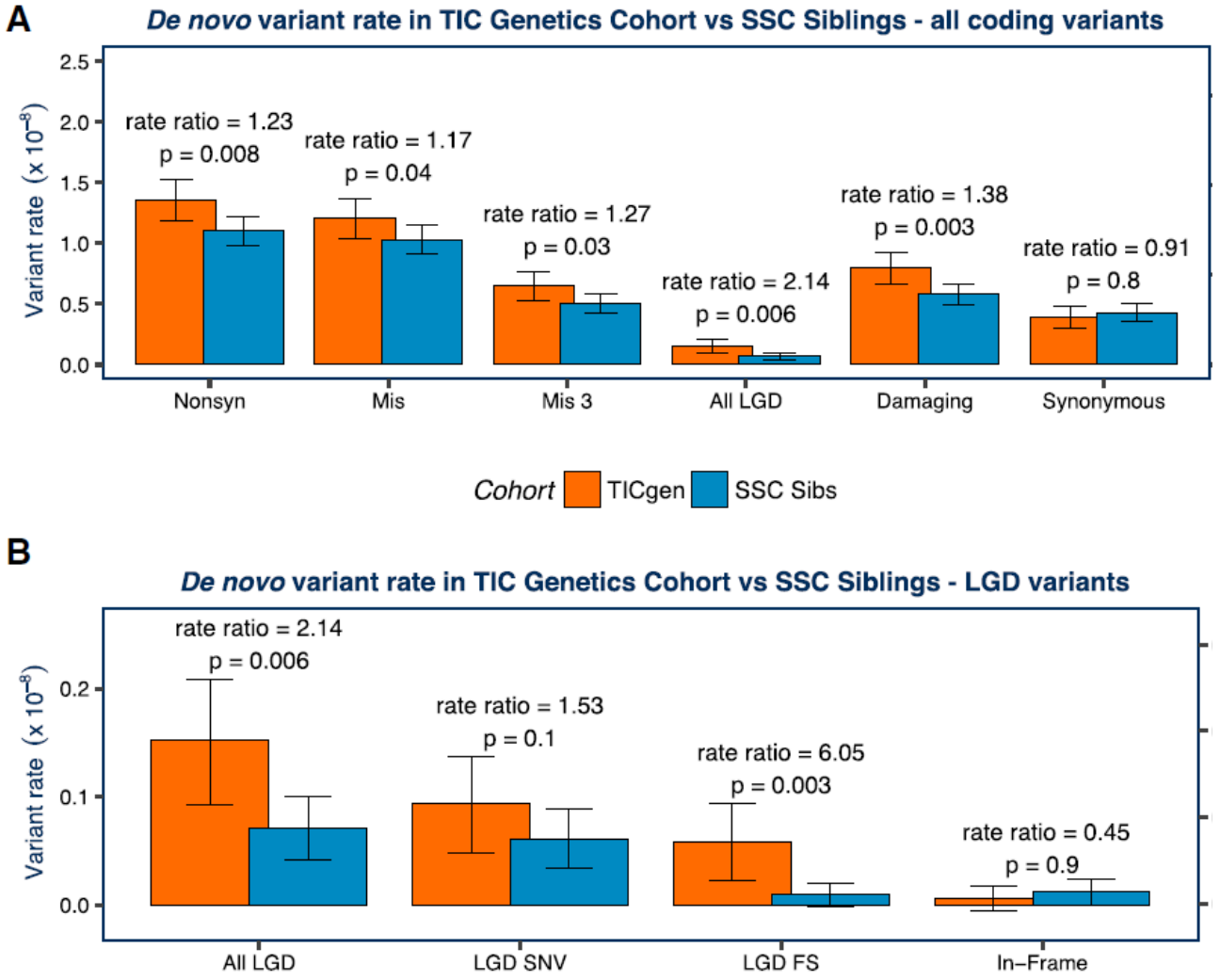
Variant effect and association with phenotypes



Meta-analyzed association between ultra-rare and rare damaging missense variants in PTV-intolerant genes and 5 diseases. **The strength of the association increases as function of the number of algorithms and is particularly strong among ultra-rare variants**

Ganna (2018) *Am J Hum Genet*

Variant effect and association with phenotypes



All classes of *de novo* non-synonymous variants show a higher mutation rate in Tourette disorder probands (orange) versus SSC siblings (controls, blue). LGD: likely gene disrupting variants: insertion of premature stop codon, frameshift, or canonical splice-site variant; **FS**: frameshift indels; **Damaging**: variants predicted by PolyPhen2; **Mis3**: LGD or damaging; **Nonsyn**: missense or nonsense

Willsey (2017) *Neuron*

Summary

- Human genome sequence is still being updated. We may soon switch from a single reference sequence to multiple ones
- Protein-coding genes represent only a minor fraction of all human genes and a tiny fraction of the genome
- Roughly one half of human genome are repetitive sequences
- Human gene structure and processing is quite diverse and complicated
- There are multiple sequence regions that assist in gene splicing: exonic and intronic splicing enhancers and silencers. A significant fraction of human disease mutations are believed to be splicing-related
- Epigenetics provide heritable phenotype changes that do not involve alterations in the DNA sequence: DNA methylation at CpG nucleotides, covalen modification of histone proteins. Noncoding RNAs are considered as part of epigenetic machinery.

Summary

- Approximately 100 genes on various chromosomes are subject to chromosomal imprinting
- Variant annotation is a proccedure that determines variant consequence for a gene/protein based on its location relative to the gene sequence. It is governed and complicated by transcript structure complexity.
- Variant effect prediction determines potential functional impact of a particular variant based on its features.
- There are numerous prediction algorithms for major types of variants. Their performance and domain of applicability is a debated question, however, phenotype-associated variants are typically enriched with functional predictions.

Further reading

- Strachan, Read – *Human Molecular Genetics*, Chapter 13

- Rivas, M.A., Pirinen, M., Conrad, D.F., Lek, M., et al. (2015). Effect of predicted protein-truncating genetic variants on the human transcriptome. *Science* 348, 666–669.

- Saleheen, D., Natarajan, P., et al. (2017). Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. *Nature* 544, 235–239

- Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J.F., et al. (2019). Predicting Splicing from Primary Sequence with Deep Learning. *Cell* 176, 535-548.e24.

- Niroula, A., and Vihinen, M. (2016). Variation Interpretation Predictors: Principles, Types, Performance, and Choice. *Human Mutation* 37, 579–597.

Further reading

- Li, J., Zhao, T., Zhang, Y., Zhang, K., Shi, L., Chen, Y., Wang, X., and Sun, Z. (2018). Performance evaluation of pathogenicity-computation methods for missense variants. *Nucleic Acids Res* 46, 7793–7804.

- DePristo, M.A., Weinreich, D.M., and Hartl, D.L. (2005). Missense meanderings in sequence space: a biophysical view of protein evolution. *Nat. Rev. Genet* 6, 678–687.

- Park, E., Pan, Z., Zhang, Z., Lin, L., and Xing, Y. (2018). The Expanding Landscape of Alternative Splicing Variation in Human Populations. *Am. J. Hum. Genet.* 102, 11–26.

- Lee, P., Lee, C., Li, X., Wee, B., Dwivedi, T., and Daly, M. (2018). Principles and methods of in-silico prioritization of non-coding regulatory variants. *Hum Genet* 137, 15–30.

- Eilbeck, K., Quinlan, A., and Yandell, M. (2017). Settling the score: variant prioritization and Mendelian disease. *Nature Reviews Genetics* 18, 599.