

Case Series of Open Data Utilization in Disease Genetics and Genomics



Atsushi Takata

Generated by DALL-E

Laboratory for Molecular Pathology of Psychiatric Disorders

RIKEN Center for Brain Science

01/27/2025

主要論文

2. Nakamura T,^{*} Ueda J,^{*†} Mizuno S,^{*} Honda K, Kazuno A-a, Yamamoto H, Hara T, and Takata A[†]

Topologically associating domains define the impact of de novo promoter variants on autism spectrum disorder risk

Cell Genomics 2024 <https://doi.org/10.1016/j.xgen.2024.100488>

SFARI Baseの**大規模公的ゲノムデータの解析**と実験的検証を組み合わせ、ラボの総力を結集して仕上げました。プレスリリースは**こちら**。

Analysis of publicly available large genome data

6. Takata A[†], Hamanaka K, Matsumoto N[†]

Refinement of the clinical variant interpretation framework by statistical evidence and machine learning

Med 2021 <https://doi.org/10.1016/j.medj.2021.02.003>

世界標準の遺伝子診断指針であるACMG/AMPガイドラインを統計学と人工知能でカイゼンする方法を提示しました。**計算機の初期投資以外は電気**

代しかかかっていません。プレスリリースは**こちら**。GitHubページは**こちら**。 **We have only spent the electricity cost except for the initial investment in the computing system.**

10. Takata A[†], Matsumoto N, Kato T[†]

Genome-wide identification of splicing QTLs in the human brain and their enrichment among schizophrenia-associated loci

Nature Communications 2017

CommonMindコンソーシアムのデータを使って、網羅的に脳内で選択的スプライシングに関わる遺伝子変異を同定し、それらが統合失調症リスクに関与することを示しました。**計算機の初期投資以外は電気代しかかかっていません。**プレスリリースは**こちら**。

12. Takata A, Ionita-Laza I, Gogos JA, Xu B, Karayiorgou M

De Novo Synonymous Mutations in Regulatory Elements Contribute to the Genetic Etiology of Autism and Schizophrenia

Neuron 2016

De novo変異のうち、タンパク質のアミノ酸配列を変えない同義置換変異（シノニマス変異、サイレント変異ともよばれる）の中にも自閉スペクトラム症や統合失調症リスクに関与するものがあることを明らかにしました。**計算機の初期投資以外は電気代しかかかっていません。**

Case 1

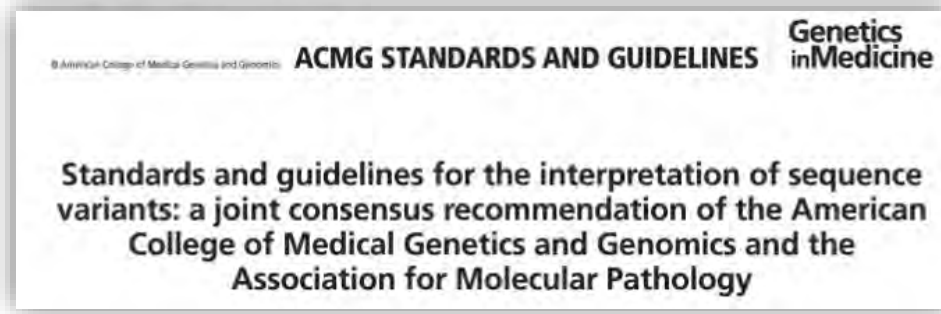
Clinical and Translational Article

Refinement of the clinical variant interpretation framework by statistical evidence and machine learning

STAR★METHODS

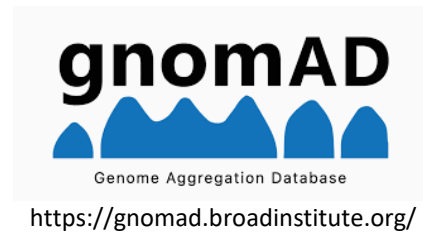
KEY RESOURCED TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
ACMG/AMP guideline	Richards et al. ¹	
gnomAD	Karczewski et al. ⁹ and Lek et al. ¹⁰	https://gnomad.broadinstitute.org/
HGMD	Stenson et al. ¹¹	http://www.hgmd.cf.ac.uk/ac/index.php
ClinVar	Landrum et al. ¹²	https://www.ncbi.nlm.nih.gov/clinvar/ Accessed in August 2018
Repeating Elements by RepeatMasker in the UCSC Table Browser	http://genome.ucsc.edu/cgi-bin/hgTrackUi?db=hg19&g=rmsk	
1000 Genomes Project ancestral allele data	The 1000 Genomes Project Consortium, 2012 ⁴⁰	ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/ancestral_alignments/human_ancestor_GRCh37_e59.tar.bz2
GTEx	https://www.gtportal.org/home/	GTEx Analysis V8
TOPMed	Taliun et al. ²⁹	https://bravo.sph.umich.edu/freeze5/hg38/get_authorized
GenomeAsia 100K	GenomeAsia100K Consortium ³⁰	https://browser.genomeasia100k.org/#tid=download
De novo variants in DD, ASD, and controls	Satterstrom et al. ²⁴ and Kaplanis et al. ²⁵	Table S1 of the Satterstrom et al. ²⁴ study and Table S1 of the Kaplanis et al. ²⁵ study
Software and algorithms		
bcftools	https://samtools.github.io/bcftools/bcftools.html	Version 1.3.1
Snpeff (v4.2)	Cingolani et al. ⁶	http://snpeff.sourceforge.net/
dbNSFP (v3.0a or 3.5a)	Liu et al. ⁴¹	http://varianttools.sourceforge.net/Annotation/DbNSFP
SIFT	Kumar et al. ¹⁶	http://provean.jcvi.org
PolyPhen-2	Adzhubei et al. ¹⁷	http://genetics.bwh.harvard.edu/pph2/
LRT	Chun and Fay ¹⁸	http://www.genetics.wustl.edu/flab/data5.html
MutationTaster	Schwarz et al. ¹⁹	http://www.mutationtaster.org/
Mutation Assessor	Reva et al. ²⁰	http://mutationassessor.org/r3/
PROVEAN	Choi et al. ²¹	http://provean.jcvi.org/index.php
bedtools	Quinlan and Hall ⁴²	https://bedtools.readthedocs.io/en/latest/index.html
Ensembl BioMart	https://grch37.ensembl.org/info/data/biomart/index.html	GRCh37.p13
SLiM (v3.2)	Haller and Messer ¹⁴	https://messerlab.org/slim/
pROC	Robin et al. ⁴³	https://web.expasy.org/pROC/
TITER	Zhang et al. ²⁶	https://github.com/zhangsathu/titer
randomForest	https://cran.r-project.org/web/packages/randomForest/	Version 4.6-14
CADD	Kircher et al. ²²	https://cadd.gs.washington.edu/
Eigen	Ionita-Laza et al. ²⁴	http://www.columbia.edu/~ni2135/eigen.html
phyloP score	Pollard et al. ⁴⁵	
dddMAPS	Short et al. ⁴⁶	https://github.com/pjshort/dddMAPS
DDG2P	https://decipher.sanger.ac.uk/ddd/ddgenes	Accessed in July 2019
UCSC Genome Browser	https://genome.ucsc.edu/	GRCh37/hg19



The target

The ACMG/AMP guideline refers to the **standardized framework developed by ACMG and the AMP to classify genetic variants based on their clinical significance.**



A **publicly** accessible resource that aggregates and harmonizes **genetic variation data from general populations worldwide.**



A comprehensive resource that catalogs published **genetic mutations associated with human inherited diseases.**



A **publicly** accessible database that links **genetic variants to their clinical significance and associated conditions.**

Pathogenic Criteria

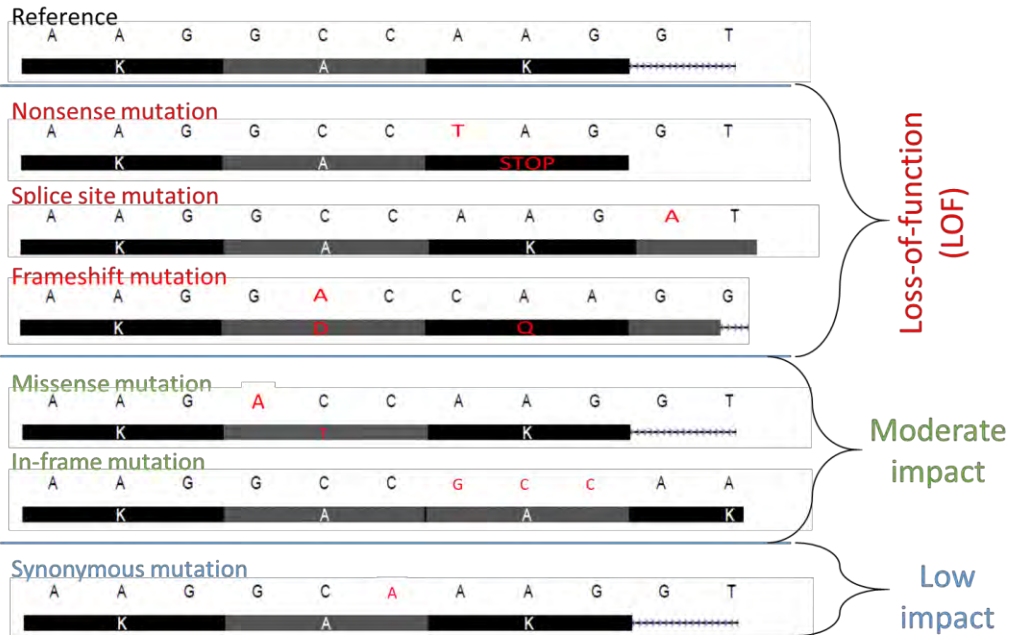
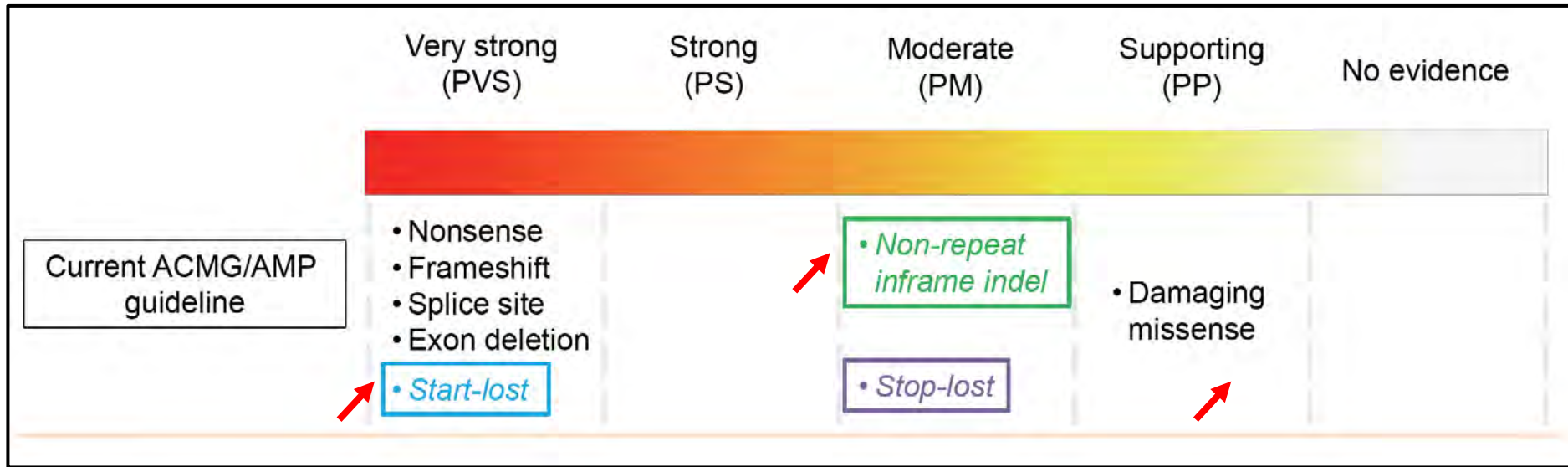
	Rule	Modification Type	Rule Description
STRONG	PV51	RE	Null variant in gene with established LOF as disease mechanism
	PS1	NC	Different nucleotide change (same amino acid) as a previously established pathogenic variant
	PS2	DG	<i>De novo</i> (paternity confirmed) in a patient with disease and no family history
	PS3	DG	Functional studies of mammalian knock-in models supportive of a damaging effect on the gene or gene product
	PS4	DG	Prevalence of the variant in affected individuals is significantly increased compared to the prevalence in controls -OR- Variant identified in ≥ 15 probands with consistent phenotypes
	PP1_Strong	MS	Variant segregates with ≥ 7 meioses
MODERATE	PM1	DG	Hotspot/est. functional domain (amino acids 181-937) without benign variation
	PM2	DG	Absent/extremely rare ($< 0.004\%$) from large population studies
	PM3	RE	Detected in trans with a pathogenic variant (recessive)
	PM4	DG	Protein length changes due to in-frame deletions/insertions of any size in a non-repeat region or stop-loss variants
	PM5	NC	Missense change at an amino acid residue where a different missense change previously established as pathogenic
	PM6	DG	Confirmed <i>de novo</i> without confirmation of paternity
	PVS1_Moderate	MS	Null variant in gene with evidence supporting LOF as disease mechanism
	PS4_Moderate	MS	Variant identified in ≥ 6 probands with consistent phenotypes
	PP1_Moderate	MS	Variant segregates in ≥ 5 meioses
SUPPORTING	PP1	DG	Variant segregates in ≥ 3 meioses
	PP2	RE	Missense variant in a gene that has a low rate of benign missense variation and where missense variants are a common mechanism of disease
	PP3	NC	Multiple lines of computational evidence support a deleterious effect on the gene or gene product
	PP4	RE	Phenotype specific for disease with single genetic etiology
	PP5	RE	Reputable source reports as pathogenic
	PS4_Supporting	MS	Variant identified in ≥ 2 probands with consistent phenotypes

The ACMG Guideline in Brief:

- The pathogenicity of a variant is assessed based on various criteria, which are categorized into four distinct levels: **Very Strong**, **Strong**, **Moderate**, and **Supporting**.
- For instance, loss-of-function mutations (e.g., nonsense, splice site, or frameshift variants) in genes known to cause diseases are classified as "**Very Strong**", while *de novo* (newly occurring) mutations are classified as "**Strong**."
- The final pathogenicity of a variant is determined by the combination of criteria that are met.

Table 5 Rules for combining criteria to classify sequence variants

Pathogenic	Likely pathogenic
(i) 1 Very strong (PVS1) AND (a) ≥ 1 Strong (PS1-PS4) OR (b) ≥ 2 Moderate (PM1-PM6) OR (c) 1 Moderate (PM1-PM6) and 1 supporting (PP1-PP5) OR (d) ≥ 2 Supporting (PP1-PP5)	(i) 1 Very strong (PVS1) AND 1 moderate (PM1-PM6) OR (ii) 1 Strong (PS1-PS4) AND 1-2 moderate (PM1-PM6) OR (iii) 1 Strong (PS1-PS4) AND ≥ 2 supporting (PP1-PP5) OR (iv) ≥ 3 Moderate (PM1-PM6) OR (v) 2 Moderate (PM1-PM6) AND ≥ 2 supporting (PP1-PP5) OR (vi) 1 Moderate (PM1-PM6) AND ≥ 4 supporting (PP1-PP5)
(ii) ≥ 2 Strong (PS1-PS4) OR (iii) 1 Strong (PS1-PS4) AND (a) ≥ 3 Moderate (PM1-PM6) OR (b) 2 Moderate (PM1-PM6) AND ≥ 2 Supporting (PP1-PP5) OR (c) 1 Moderate (PM1-PM6) AND ≥ 4 supporting (PP1-PP5)	



		Second letter					
		U	C	A	G		
First letter	U	UUU } Phe UUC } UUA } Leu UUG }	UCU } Ser UCC } UCA } UCG }	UAU } Tyr UAC } UAA Stop UAG Stop	UGU } Cys UGC } UGA Stop UGG Trp	U C A G	
	C	CUU } Leu CUC } CUA } CUG }	CCU } Pro CCC } CCA } CCG }	CAU } His CAC } CAA } Gln CAG }	CGU } Arg CGC } CGA } CGG }	U C A G	
	A	AUU } Ile AUC } AUA } AUG Met	ACU } Thr ACC } ACA } ACG }	AAU } Asn AAC } AAA } Lys AAG }	AGU } Ser AGC } AGA } Arg AGG }	U C A G	
	G	GUU } Val GUC } GUA } GUG }	GCU } Ala GCC } GCA } GCG }	GAU } Asp GAC } GAA } Glu GAG }	GGU } Gly GGC } GGA } GGG }	U C A G	

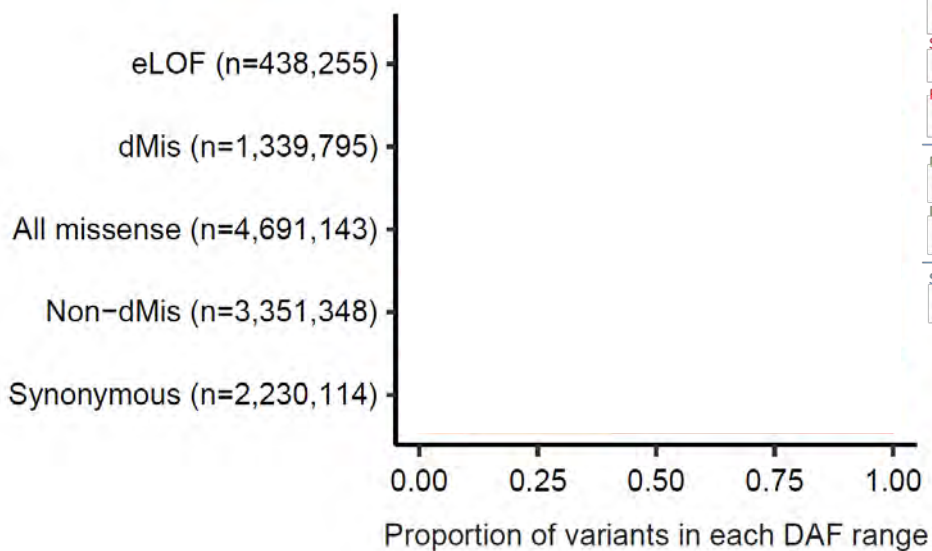


Use of gnomAD open data of human variants to estimate average deleteriousness across variant types

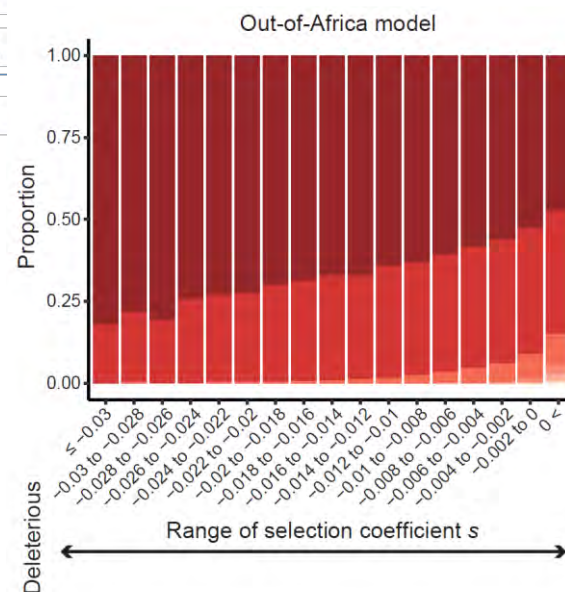
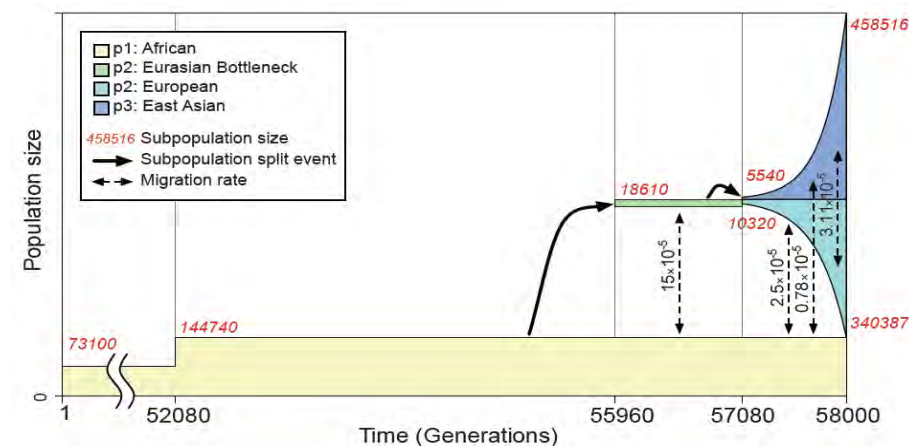


Real world data of variants in protein coding regions from **123,136*** individuals (*at the time of our previous study)

gnomad.broadinstitute.org



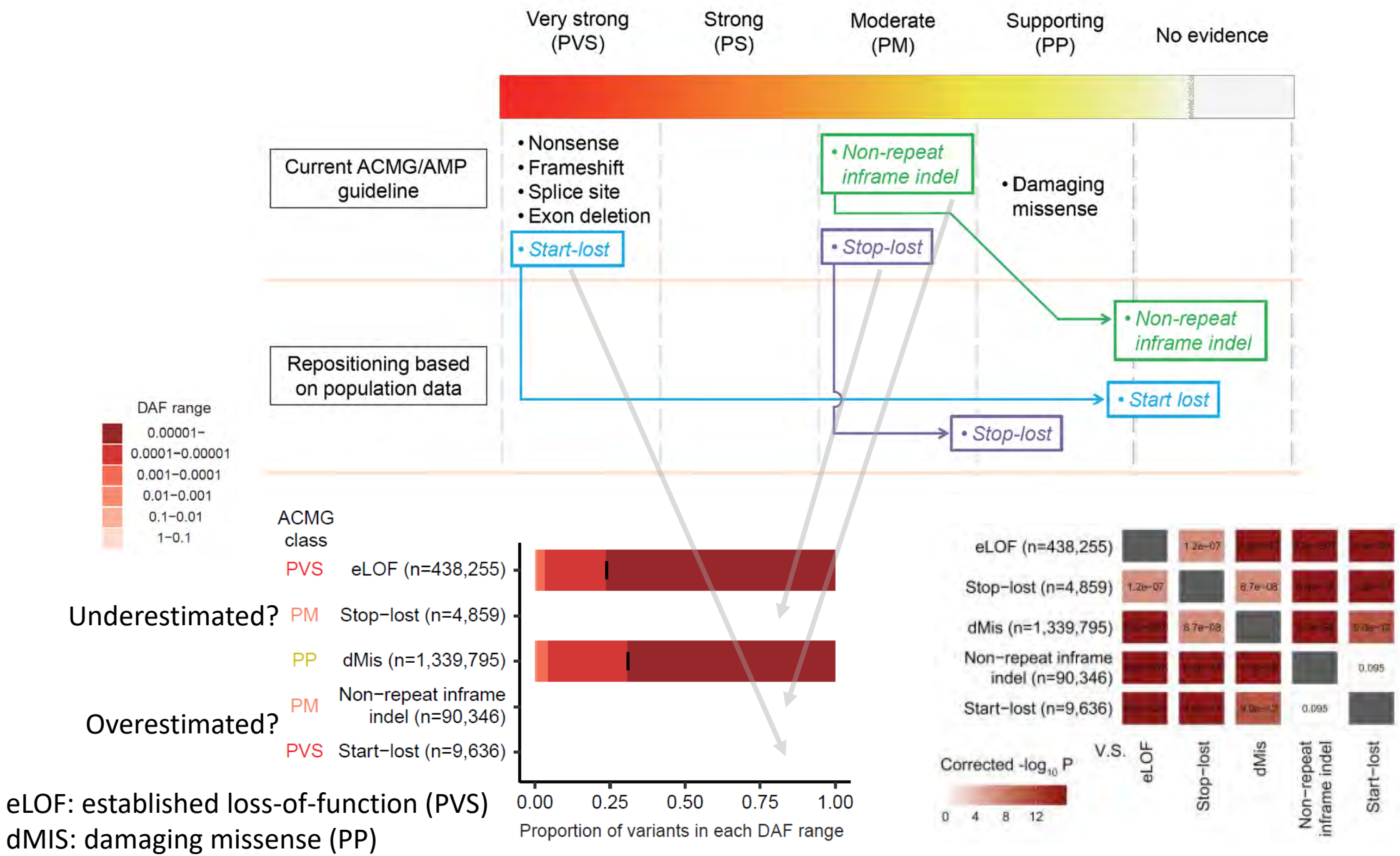
Forward genetic simulation



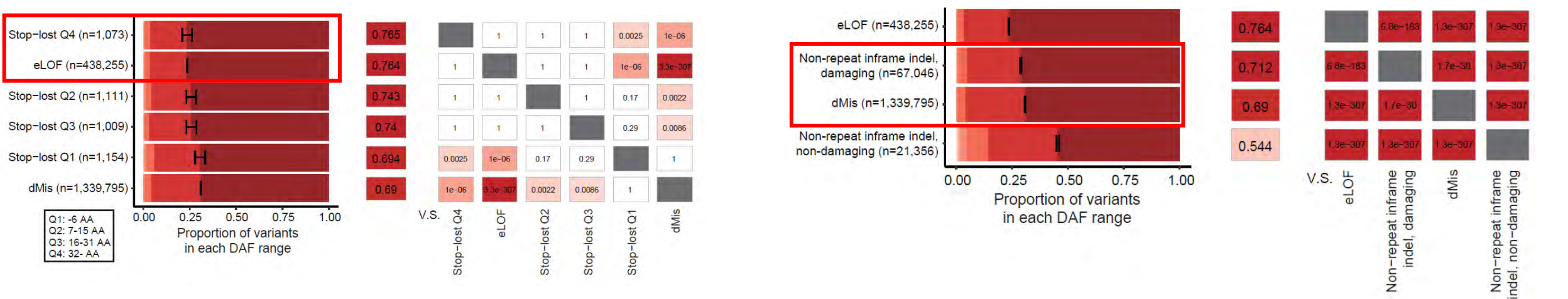
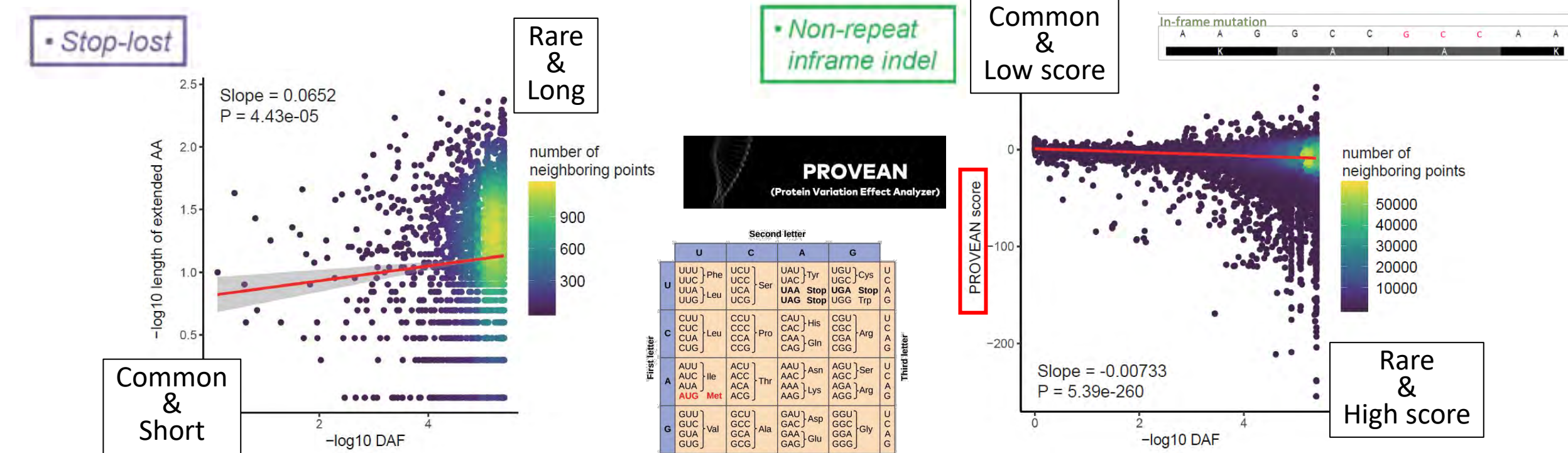
Deleterious types of variants are **enriched for rare variants** in population



eLOF: established loss-of-function, defined as nonsense, canonical splice site, and frameshift variants
 dMIS: damaging missense



Extraction of deleterious variant subtypes using the proportion of rare variants as an indicator (1)



eLOF: established loss-of-function (PVS)

dMIS: damaging missense (PS)

Extraction of deleterious variant subtypes using the proportion of rare variants as an indicator (2)

• Start-lost

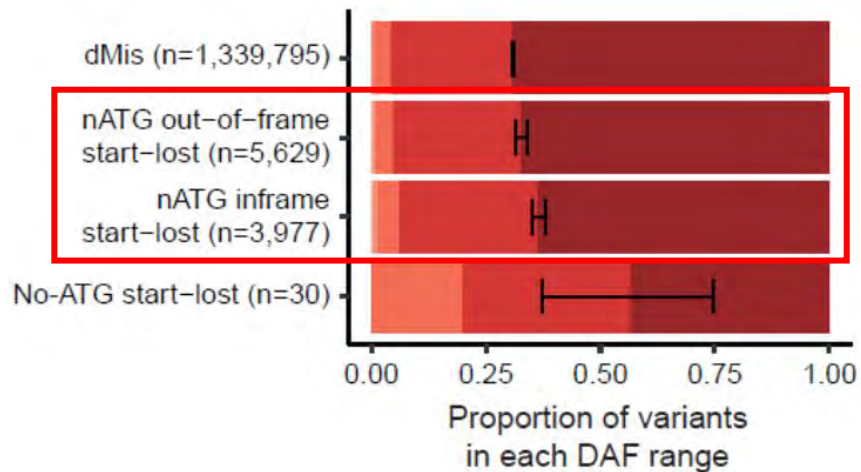
The **ATG** sequence closest to the known start codon (i.e. a potential start codon) is...

Inframe

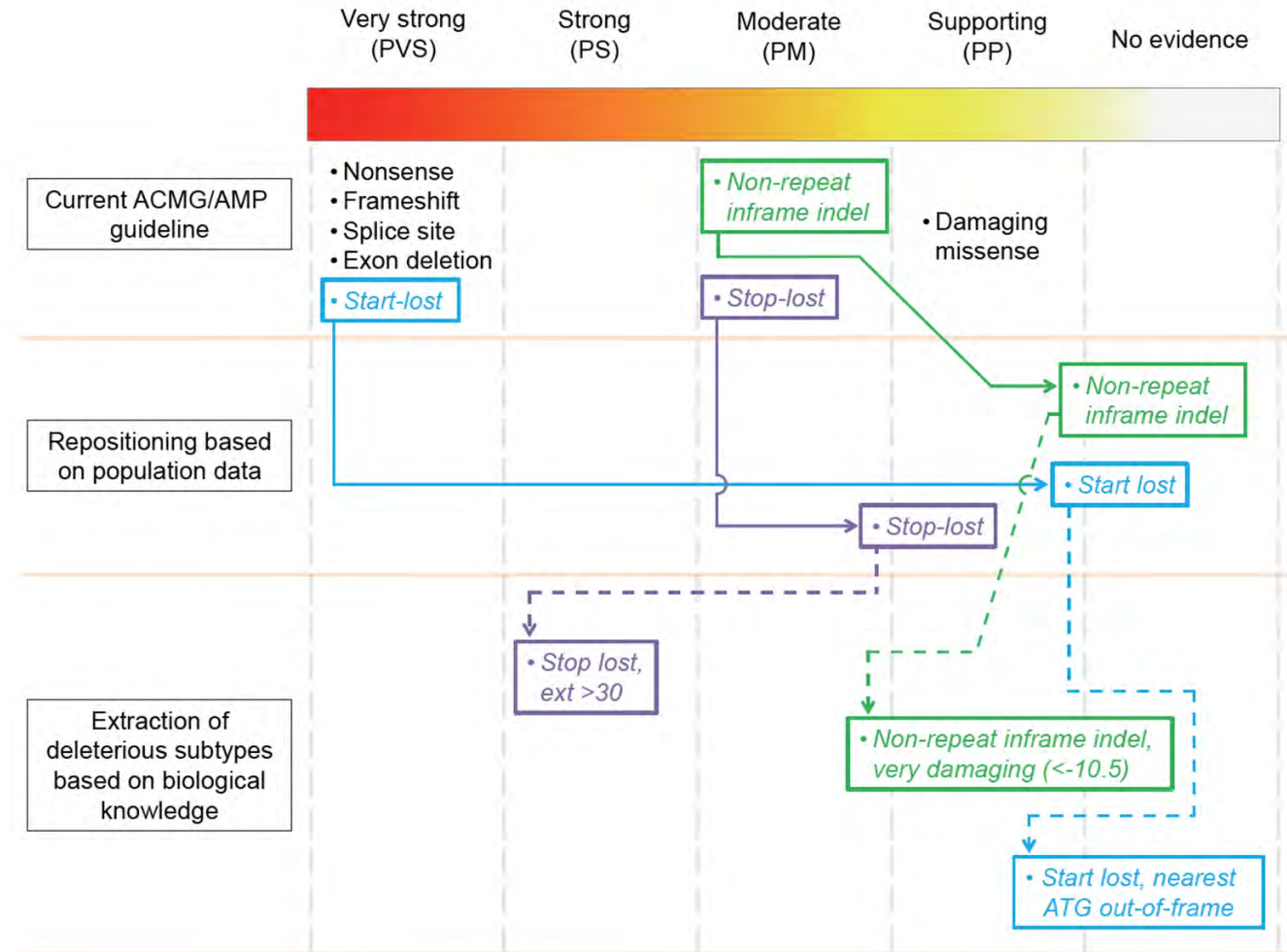
gcatgctagc**ATG** | GCT | AGC | TAG | TCA | **ATG** | CAT | AG
 → *Less deleterious ?*

Out-of-frame

acgtcgatcg**ATG** | CCG | CTG | **CAT** | **GCT** | AGC | TAG | TC
 → *More deleterious ?*



nATG: nearest ATG



Extraction of deleterious variant subtypes using the proportion of rare variants as an indicator (2)

• Start-lost

The ATG sequence nearest to the known start codon (i.e. a potential start codon) is...

Inframe

gcatgctagcATG | GCT | AGC | TAG | TCA | ATG | CAT | AG
 → Less deleterious ?

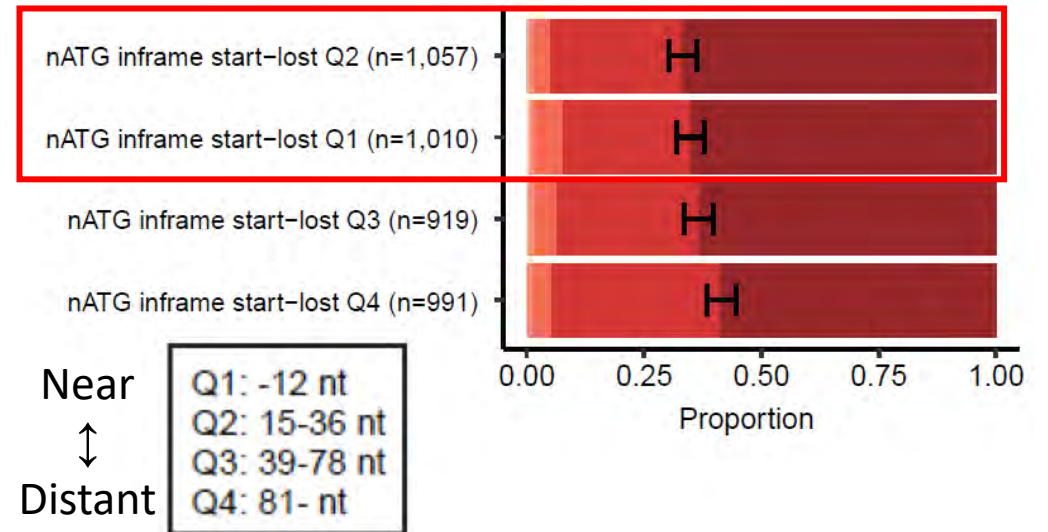
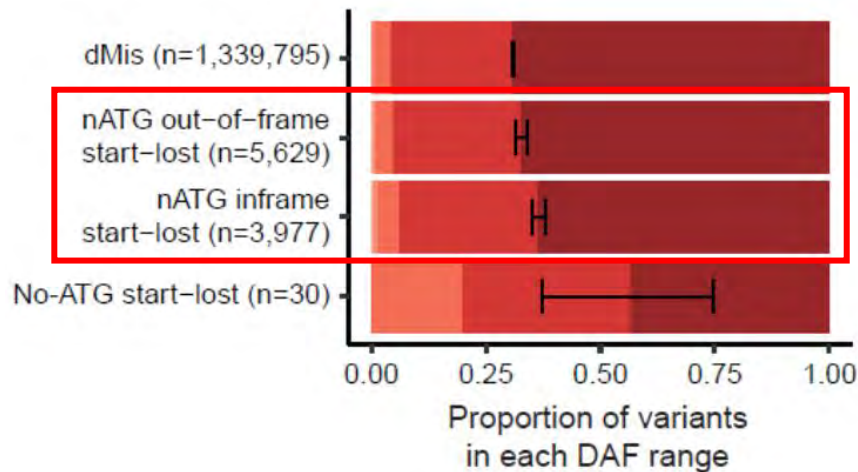
Out-of-frame

acgtcgatcgATG | CCG | CTG | CAT | GCT | AGC | TAG | TC
 → More deleterious ?

Stratification of nearest ATG inframe start-lost variants based on the distance between the known start codon and the nearest ATG.

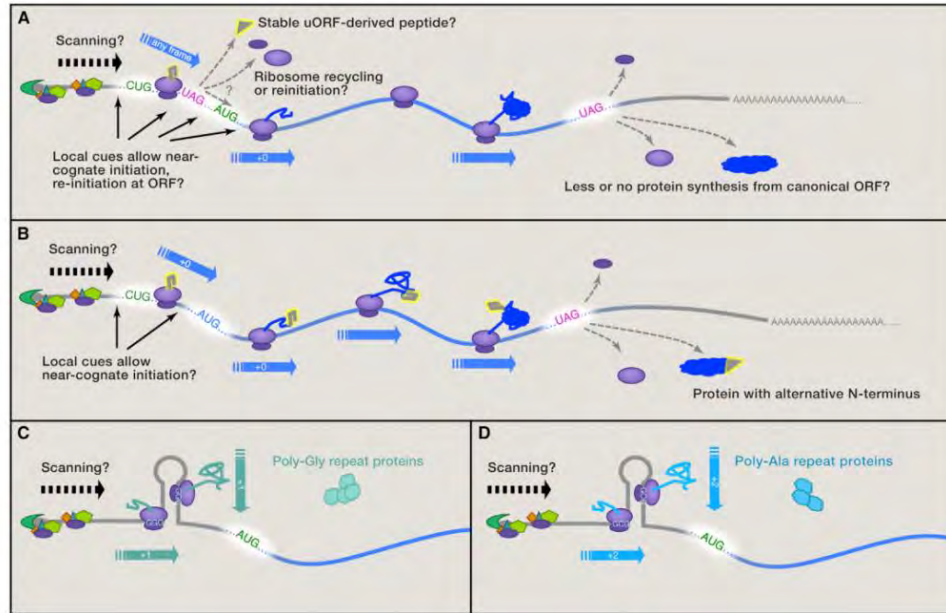
gcatgctagcATG | GCT | ATG | TAG | TCA | AGC | CAT | AG
 → Less deleterious ?

acgtcgatcgATG | CCG | CTG | | AGC | ATG | TC
 → More deleterious ?



nATG: nearest ATG

Extraction of deleterious variant subtypes using the proportion of rare variants as an indicator (2)



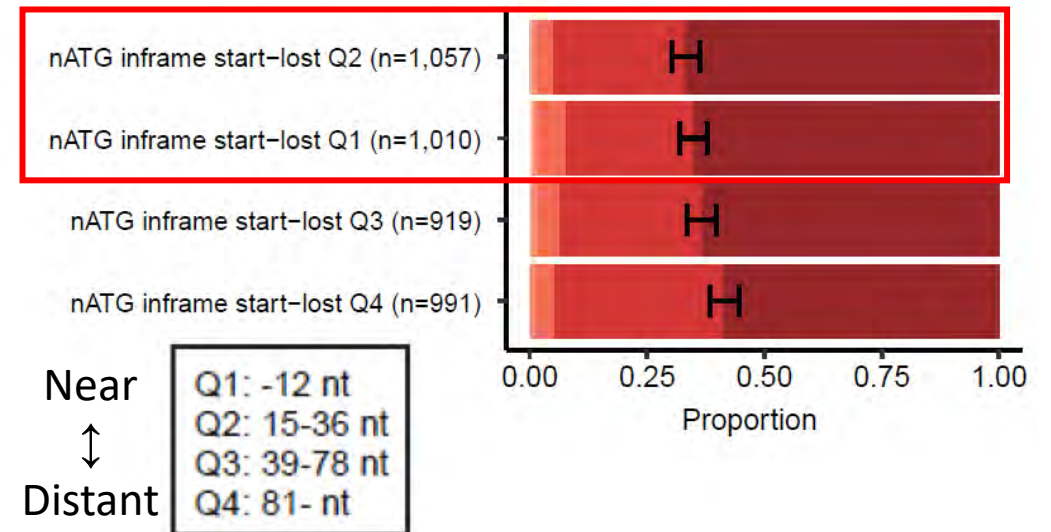
There is accumulating evidence indicating that

- Protein translation can occur at various locations within a transcript (Ingolia, Cell 2016).
- Translation can be initiated from ATG-like sequences, such as CTG, and such translation may occur more frequently than previously thought (Brar, Cell 2016).

Stratification of *nearest ATG inframe start-lost variants* based on the distance between the **known start codon** and the nearest **ATG**.

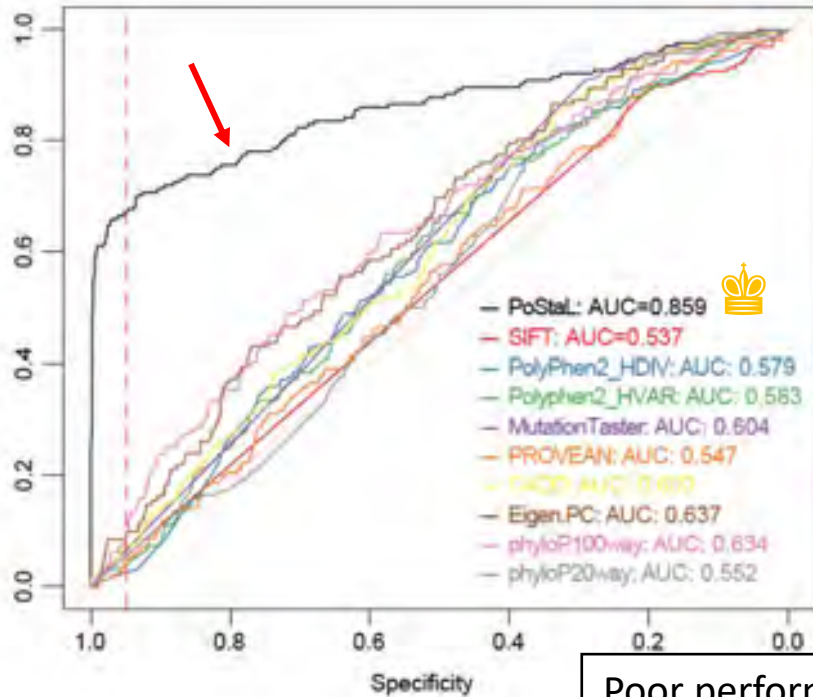
gcatgctagc**ATG** | GCT | **ATG** | TAG | TCA | AGC | CAT | AG
 → *Less deleterious?*

acgtcgatcg**ATG** | CCG | CTG | | AGC | **ATG** | TC
 → *More deleterious?*



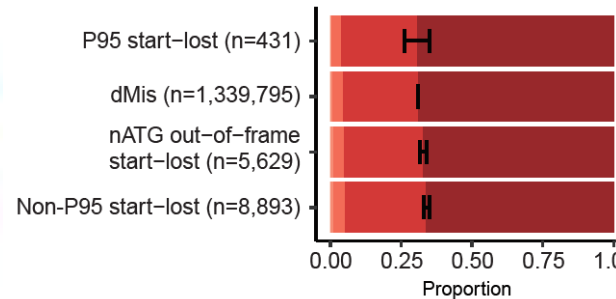
Construction of a machine learning model to predict pathogenic start-lost variants

Good performer

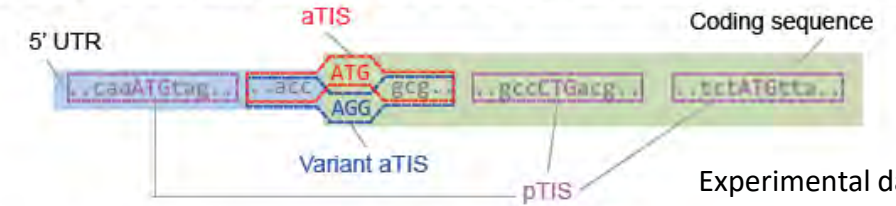


Poor performer

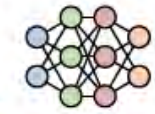
TIS: translation initiation site
aTIS: annotated (known) TIS
pTIS: potential (not known) TIS



Extraction of the sequences surrounding aTIS, variant aTIS and pTIS



Deep learning-based prediction of TIS scores from sequence features (TITER, Zhang et al.)



Experimental data of translation initiating ribosome sequencing

Variant annotation with TIS scores and other information



Variant	Score	Codon	Position	Frame
Variant 1				
aTIS	3.2	ATG	NA	NA
Variant aTIS	0.5	AGG	NA	NA
Up to ten (Downstream ATG pTIS1)	1.5	ATG	49	Out
Up to ten (Upstream ATG pTIS1)	2.1	ATG	-33	In
Up to ten (Downstream non-ATG pTIS1)	1.8	CTG	21	In
Up to ten (Upstream non-ATG pTIS1)	0.7	GTG	-16	Out

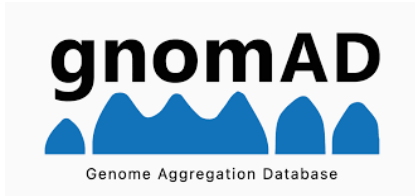


Construction and evaluation of the PoStaL (Pathogenicity of Start-Lost) scoring system

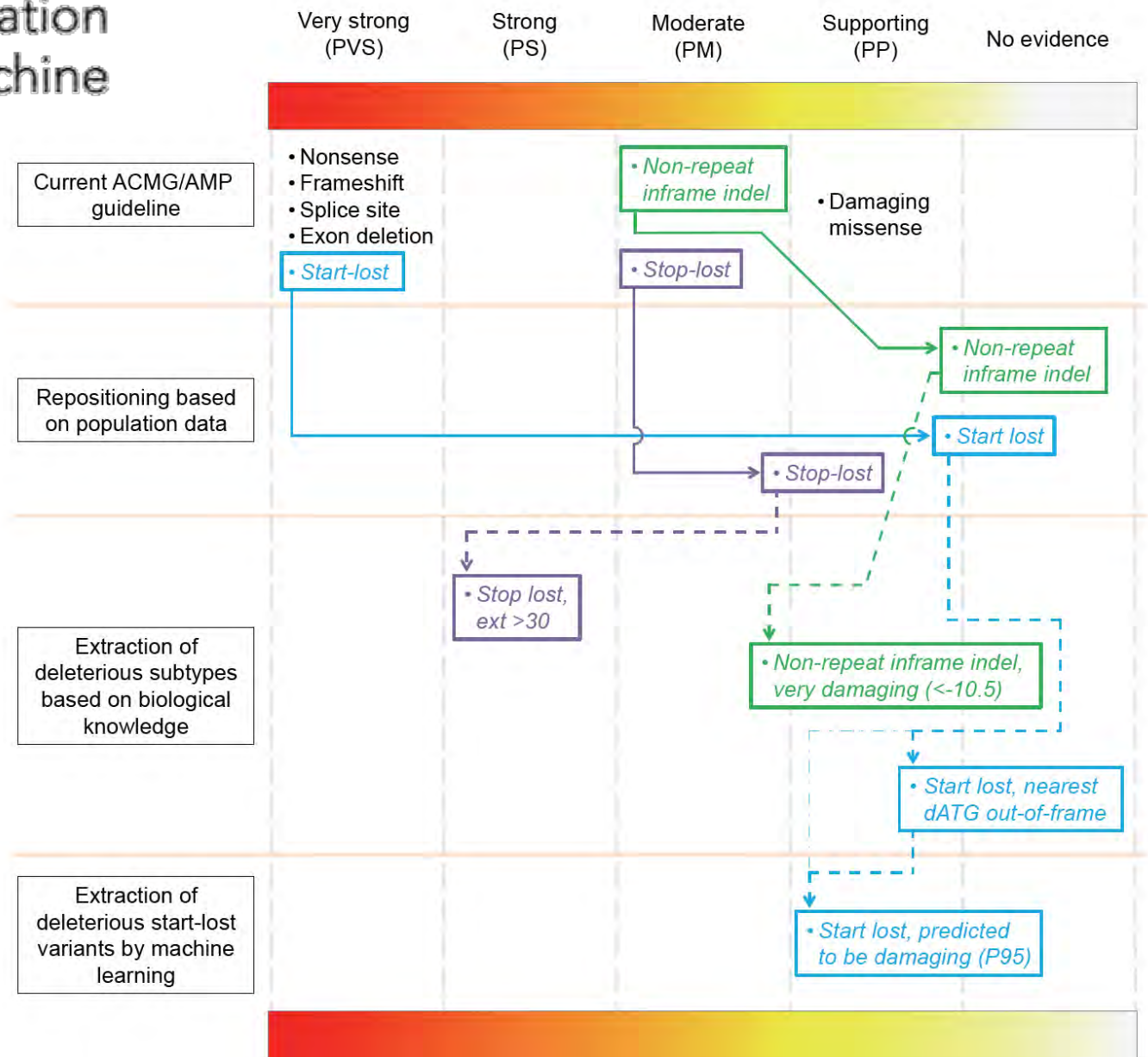
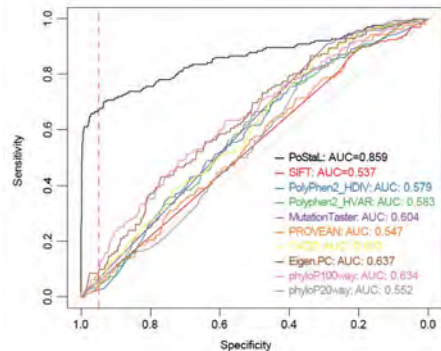


- PoStaL outperforms other tools.
- Start-lost variants with a PoStaL score showing the >95% specificity are as enriched for very rare variants as damaging missense variants.
- Our results indicate the importance of constructing a model optimized for the variant type of interest, as other tools are not designed specifically for start-lost variants.

Refinement of the clinical variant interpretation framework by statistical evidence and machine learning



Human Gene Mutation Database



主要論文



2. Nakamura T,^{*} Ueda J,^{*†} Mizuno S,^{*} Honda K, Kazuno A-a, Yamamoto H, Hara T, and Takata A[†]

Topologically associating domains define the impact of de novo promoter variants on autism spectrum disorder risk

Cell Genomics 2024 <https://doi.org/10.1016/j.xgen.2024.100488>

SFARI Baseの**大規模公的ゲノムデータの解析**と実験的検証を組み合わせ、ラボの総力を結集して仕上げました。プレスリリースは**こちら**。

Analysis of publicly available large genome data

6. Takata A[†], Hamanaka K, Matsumoto N[†]

Refinement of the clinical variant interpretation framework by statistical evidence and machine learning

Med 2021 <https://doi.org/10.1016/j.medj.2021.02.003>

世界標準の遺伝子診断指針であるACMG/AMPガイドラインを統計学と人工知能でカイゼンする方法を提示しました。**計算機の初期投資以外は電気**

代しかかかっていません。プレスリリースは**こちら**。GitHubページは**こちら**。 **We have only spent the electricity cost except for the initial investment in the computing system.**

10. Takata A[†], Matsumoto N, Kato T[†]

Genome-wide identification of splicing QTLs in the human brain and their enrichment among schizophrenia-associated loci

Nature Communications 2017

CommonMindコンソーシアムのデータを使って、網羅的に脳内で選択的スプライシングに関わる遺伝子変異を同定し、それらが統合失調症リスクに関与することを示しました。**計算機の初期投資以外は電気代しかかかっていません。**プレスリリースは**こちら**。

12. Takata A, Ionita-Laza I, Gogos JA, Xu B, Karayiorgou M

De Novo Synonymous Mutations in Regulatory Elements Contribute to the Genetic Etiology of Autism and Schizophrenia

Neuron 2016

De novo変異のうち、タンパク質のアミノ酸配列を変えない同義置換変異（シノニマス変異、サイレント変異ともよばれる）の中にも自閉スペクトラム症や統合失調症リスクに関与するものがあることを明らかにしました。**計算機の初期投資以外は電気代しかかかっていません。**

Topologically associating domains define the impact of *de novo* promoter variants on autism spectrum disorder risk

STAR METHODS
KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, peptides, and recombinant proteins		
iMatrix-511 silk	Takara Bio Inc.	Cat# B92011
TrypLE™ SELECT	Thermo Fisher Scientific	Cat# 12569029
0.5 mM EDTA/PBS solution	NACALAI TESQUE, INC.	Cat# 13567-84
StemFit AK02N	RIPROCELL	Cat# RKAK02N
Culture Sure Y-27632	Fujifilm	Cat# 036-24023
Alt-R S.p. HiFi Cas9 Nuclease V3	Integrated DNA Technologies	Cat# 1081060
7-AAD	BD Biosciences	Cat# 559925
Primocin(TM)	NACALAI TESQUE	Cat# 14860-36
Accutase	NACALAI TESQUE	Cat# 12679-54
BigDye Terminator V3.1	Thermo Fisher Scientific	Cat# 4337455
STEM CELL BANKER GMP grade	Takara Bio Inc.	Cat# 11924
Trizol reagent	Thermo Fisher Scientific	Cat# 15596026
Recombinant DNaseI (Rnase-free)	Takara Bio Inc.	Cat# 2270A
Critical commercial assays		
Guide-IT™ sgRNA <i>in vitro</i> Transcription Kit	Takara Bio Inc.	Cat# 632635
NEBNext Ultra RNA Library Prep Kit for Illumina	New England Biolabs Inc.	Cat # E7530
Deposited data		
RNA-seq data of wild-type/mutant iPSCs	This study	The NDBC Human Database, Japan (Accession number: JGA: JGAS000651)
Population datasets	SFARI Base	https://www.sfari.org/resources/sfari-base/
TAD list of DLPCF	PsychENCODE Integrative Analysis resource	http://resource.psychencode.org/Datasets/DevInd/DER-18_TAD_adultbrain.bed
TAD list of iPSC-derived neurons	GEO database	GSE79965, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE79965
TAD lists of germinal zone and cortical plate	GEO database	GSE77565, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE77565
Other TAD lists	3D Genome Browser ⁵⁶	http://3dgenome.fgw.northwestern.edu/
Data of reference epigenomics	Roadmap Epigenomics Project ⁵⁷	http://www.roadmapgenomics.org/
Data of enhancer regions	PsychENCODE ⁵⁴	http://resource.psychencode.org/Datasets/DevInd/DER-18_TAD_adultbrain.bed
QWAS sumstat data		
	Psychiatric Genomics Consortium; Matoba et al.; Alkes Price's lab	https://www.med.unc.edu/pgc/downloads-results/ ; https://github.com/alkesprice/spark_and_sumstats/master ; https://www.sprk.org/ ; https://alkesprice.org/brainstat/alkesprice/ldsc/dependent-sumstats/
Experimental models: Organisms/strains		
Human: 201B7-F1	RIKEN BioResource Research Center	HPS4290
Oligonucleotides		
Oligo lists	This paper Tables S6, S10, and S11	N/A
Software and algorithms		
R	The R Foundation	https://www.r-project.org/
BWA-MEM (v.0.7.15)	Highnam et al. ⁶⁰	https://github.com/aleksga/bwa-MEM/
Continued		
REAGENT or RESOURCE	SOURCE	IDENTIFIER
GSEA (v.4.3.2)	Subramanian et al. ¹¹³	https://www.gsea-msigdb.org/gsea/
Analysis codes used in this manuscript	This paper	https://doi.org/10.5281/zenodo.1045728 or upon request
Other		
Neon™ Transfection System	Thermo Fisher Scientific	Cat# MPK1025
BD FACSAria II	BD Biosciences	N/A
ABI 3730xl sequencer	Life Technologies	Cat# 3730xL-100
Agilent 2100 Bioanalyzer	Agilent Technologies, Ltd.	N/A
Illumina NovaSeq 6000 System	Illumina, Inc.	N/A

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Continued		
GATK (v.3.5 and v4.1.9.0)	DePristo et al. ⁷⁷	https://gatk.broadinstitute.org/
bcftools (v.1.10.2)	Danecek et al. ⁷⁸	http://samtools.github.io/bcftools/
vcftools (v.0.1.17)	Danecek et al. ⁷⁹	https://vcftools.sourceforge.net/
TrioDeNovo (v.0.06)	Wei et al. ⁸⁰	https://genome.sph.umich.edu/wiki/TrioDeNovo
SnpSift (v.5.0c)	Cingolani et al. ⁸¹	http://pcingola.github.io/SnpSift/
RepeatMasker	UCSC	https://www.repeatmasker.org/
PLINK (v.1.90c6.24)	Purcell et al. ⁸² ; Chang et al. ⁸³	https://www.cog-genomics.org/plink/
3D Genome Browser	Wang et al. ⁵⁶	http://3dgenome.fgw.northwestern.edu/
Ensembl Variant Effect Predictor	McLaren et al. ⁸⁴	https://asia.ensembl.org/info/docs/tools/vep/index.html
bedtools	Quinlan et al. ⁸⁵	https://bedtools.readthedocs.io/
meta package (v5.5-0)	Balduzzi et al. ⁸⁶	https://cran.r-project.org/
poolr package (v.1.1.1)	Cinar et al. ⁸⁷	https://cran.r-project.org/
regBase (v.1.1.1)	Zhang et al. ⁸⁸	https://github.com/mullerlab/regBase/
CADD (v.1.3, 1.4, and 1.6)	Kircher et al. ⁸⁹ ; Rentzsch et al. ⁹⁰	https://cadd.gs.washington.edu/
CDTS		
di lullo et al. ⁹¹		http://www.hi-ncid.org/noncoding/
Cocape	Rogers et al. ⁹²	https://cocape.biocompute.org.uk/
Cocape_Somatic	Rogers et al. ⁹³	http://cocape-somatic.biocompute.org.uk/
DANN	Quang et al. ⁹⁴	https://ccb.ics.uc.edu/public_data/DANN/
DVAR	Yang et al. ⁹⁵	https://www.vumc.org/cpg/dvar/
Eigen/Eigen_PC	Ionita-Laza et al. ⁹⁶	http://www.columbia.edu/~li2135/eigen.html
FATHMM (MKL and XF)	MKL, Shahab et al. ⁹⁷ ; XF, Rogers et al. ⁹⁸	http://fathmm.biocompute.org.uk/ ; http://fathmm.biocompute.org.uk/fathmm-xf/
FIRE	Ioannidis et al. ⁹⁹	https://sites.google.com/site/firegulatoryvariation/
fitCons	Gulko et al. ¹⁰⁰	http://compgen.cahf.edu/fitCons/
fitCons2	Gulko et al. ¹⁰¹	https://github.com/CahSepellab/fitCons2/
FunSeq2	Fu et al. ¹⁰²	http://funseq2.genetec.org/
GenoCanyon	Lu et al. ¹⁰³	https://zhao-center.org/GenoCanyon_index.html
LINSIGHT	Huang et al. ¹⁰⁴	https://github.com/CahSepellab/LINSIGHT/
ncER	Wells et al. ¹⁰⁵	https://github.com/Telestlab/ncER_datasets
Orion	Gussow et al. ¹⁰⁶	https://github.com/gm-team/orion-public
PAFA	Zhou et al. ¹⁰⁷	https://199.236.67.237-2020/pafa
ReMM	Smedley et al. ¹⁰⁸	https://remm.bwh.harvard.edu/
qamran package	Tuner et al. ¹⁰⁹	https://cran.r-project.org/
LDSC (v.1.0.1)	Bulk-Sullivan et al. ¹¹⁰ ; Finucane et al. ¹¹¹	https://github.com/bulik/ldsc
trim_galore (v.0.6.6)	Krueger et al. ¹¹²	https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
STAR (v.2.7.9a)	Dobin et al. ¹¹³	https://github.com/alexdobin/STAR/
Samtools (v.1.3.1)	Li et al. ¹¹⁴	https://www.htlib.org/
featureCounts (v.2.0.1)	Liao et al. ¹¹⁵	https://sourceforge.net/projects/featureCounts/
DESeq2 package (v.1.36.0)	Love et al. ¹¹⁶	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
sva package (v.3.44.0)	Leek et al. ¹¹⁷	https://bioconductor.org/packages/release/bioc/html/sva.html
Metascape	Zhou et al. ¹¹⁸	https://metascape.org/
Cytoscape (v.3.7.2)	Shannon et al. ¹¹⁹	https://cytoscape.org/
CHOPCHOP	Labun et al. ¹²⁰	https://chopchop.cbu.uib.no/
Cas-OffFinder	Bae et al. ¹²¹	http://www.genomesset/cas-offfinder

(Continued on next page)

Nakamura, Ueda, Mizuno et al.,
Cell Genomics 2024

SFARI Base



SFARI Base serves as a centralized repository for autism and autism-related research data and biospecimens, along with an online portal facilitating research recruitment requests.

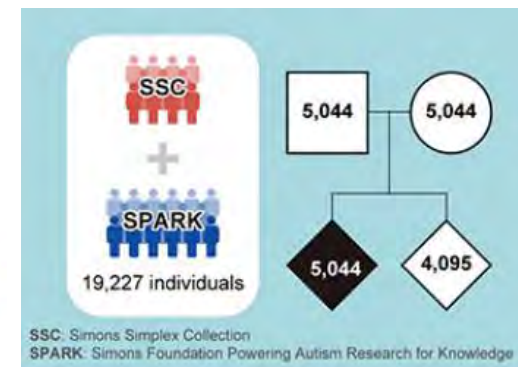
Available Resources

Learn More

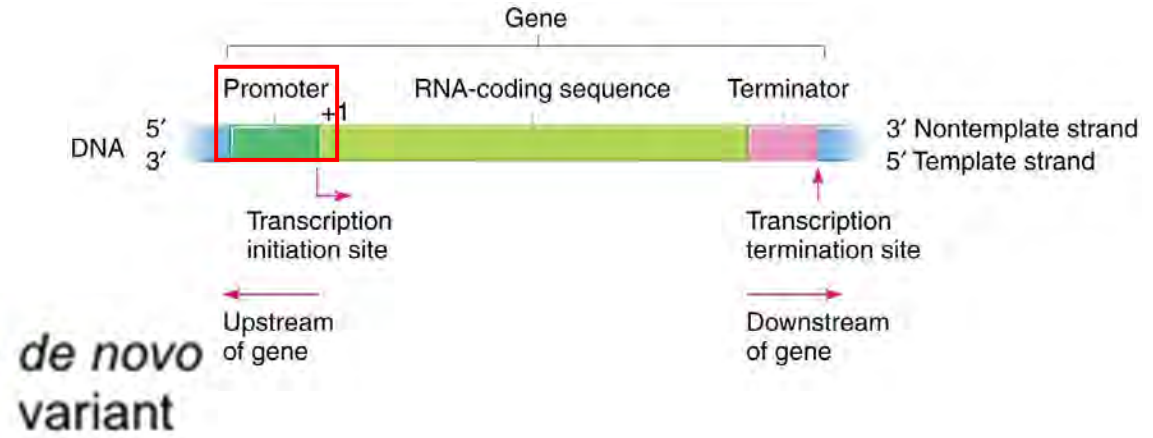
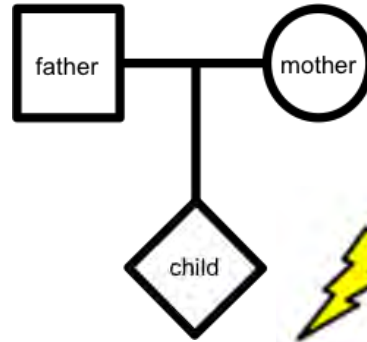
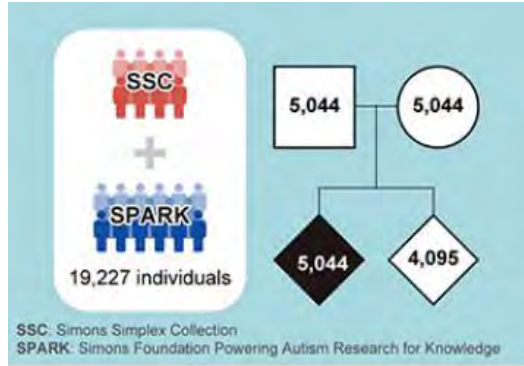
<https://base.sfari.org/>

Base Accession ID	Type	Name	Cohort ID
SFARI_DS341427	Genetic Data	Autism BrainNet (ABN) Genetic Data (WES & WGS)	ABN
	Genetic Data	ABN data by C. Walsh lab: NatNeuro2021	ABN
SFARI_DS824270	Genetic Data	AIC integrated WES (WES) (June 2024)	AIC
SFARI_DS822016	Genetic Data	AIC Biosensor Data (Imbirba et al., 2023)	AIC
SFARI_DS229125	Genetic Data	Genotypes from 25,746 individuals from SSC, SPARK, and A. Mixed	Mixed
SFARI_DS340922	Genetic Data	GATK reprocessed SNV/indel VCFs for SSC and SPARK (Pilot)	Mixed

Whole genome sequencing data

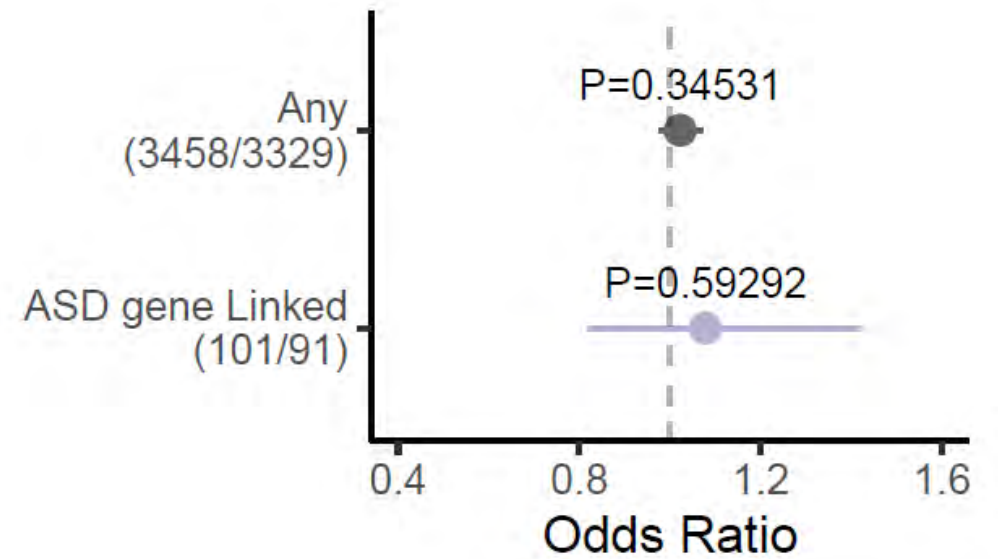
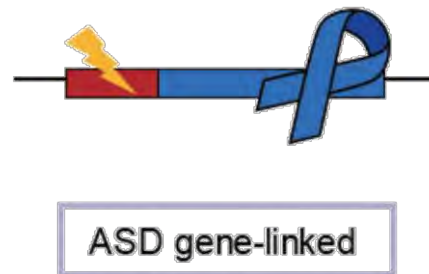
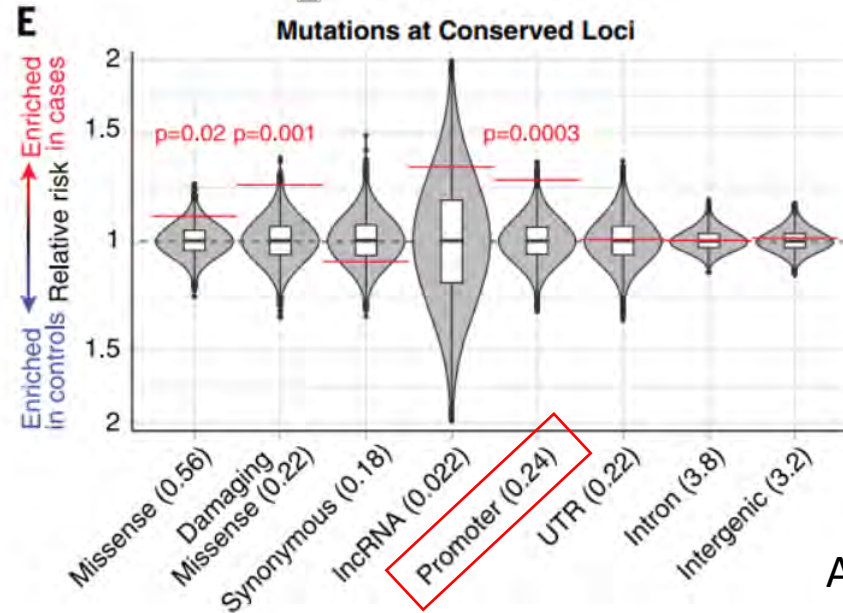


Whole genome sequencing data



PSYCHIATRIC GENOMICS

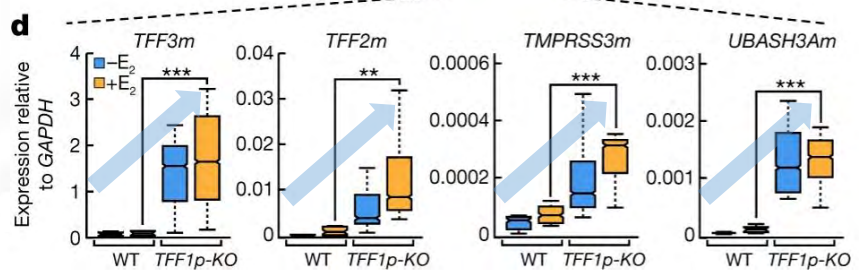
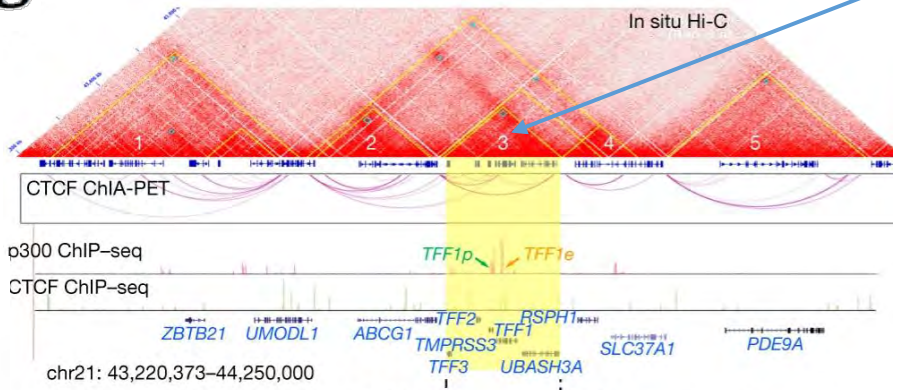
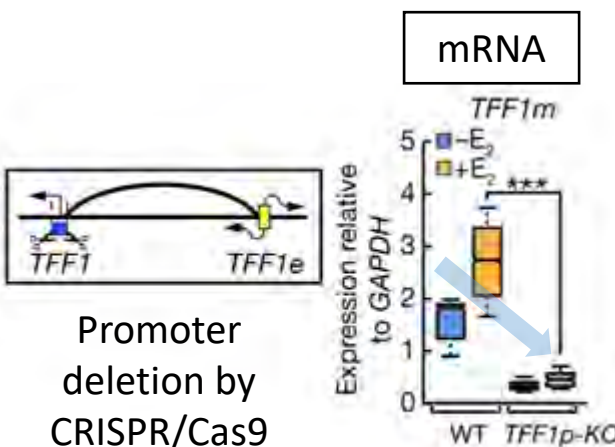
Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder



Unexpected **no enrichment** of ASD gene-linked promoter de novo variants in ASD probands

Enhancer release and retargeting activates disease-susceptibility genes

Oh et al., Nature 2021



mRNA of other genes in the same TAD

TAD (topologically associating domain)

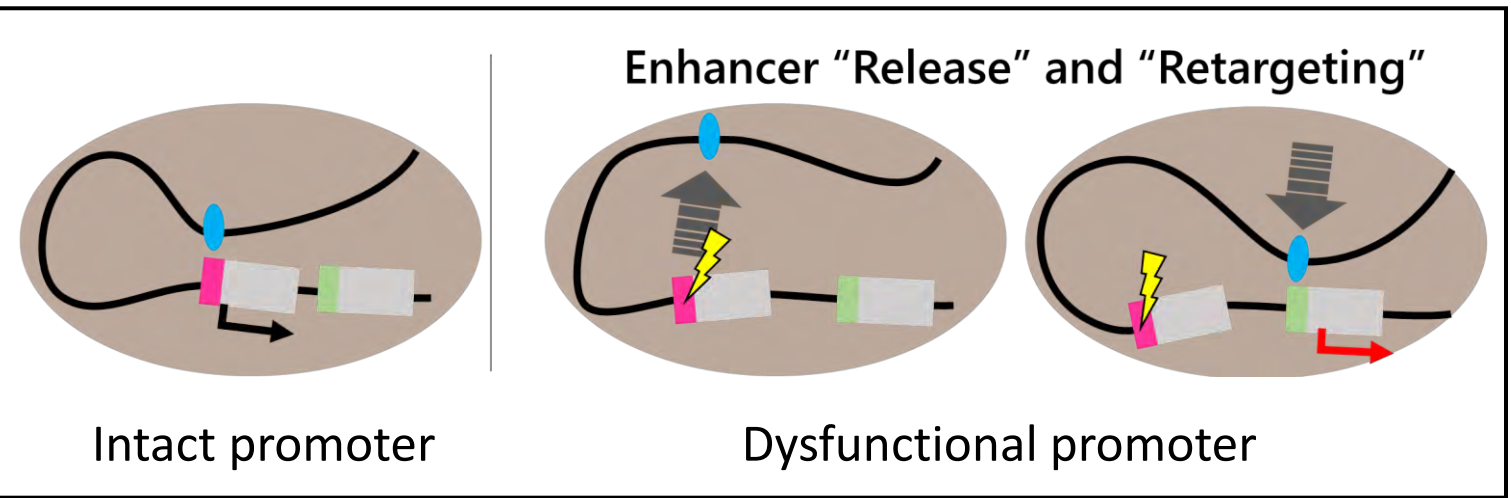
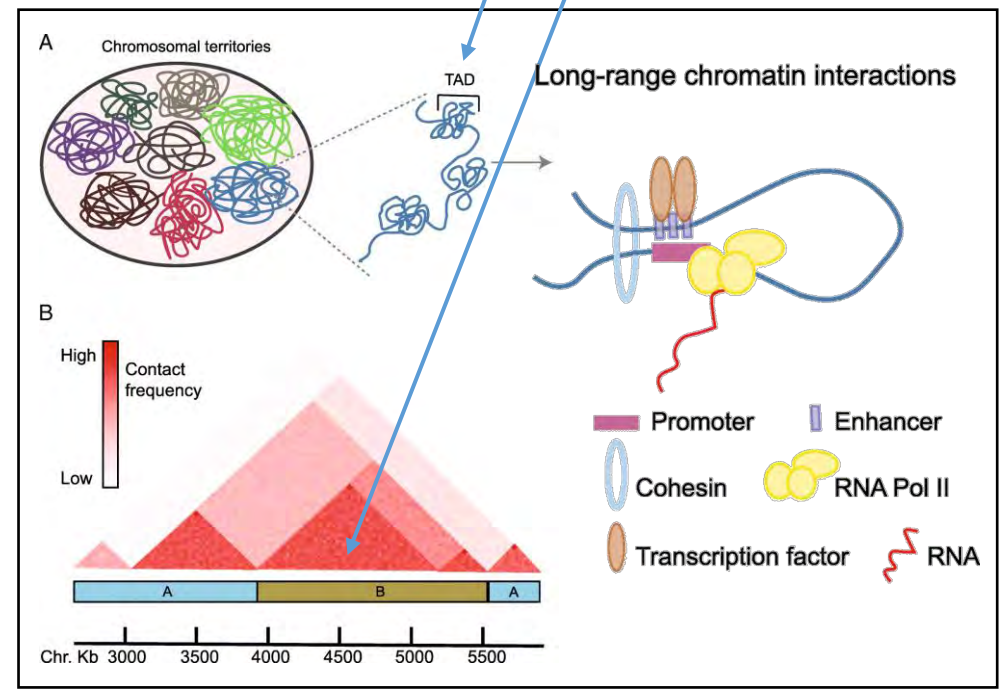
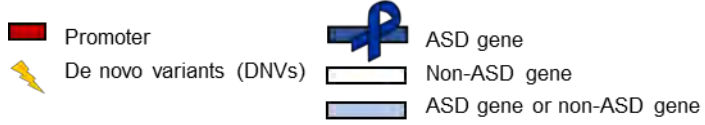
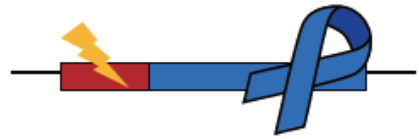


Diagram comparing ASD gene-linked (left) and ASD gene TAD (right) configurations. The ASD gene-linked configuration shows a gene and its enhancer on the same chromosome. The ASD gene TAD configuration shows a gene and its enhancer on different chromosomes but within the same TAD. The ASD gene TAD configuration is associated with a red question mark, indicating a hypothesis or area of investigation.

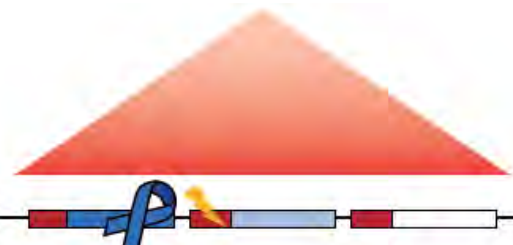
Prefrontal cortex public TAD data @ PsychENCODE Knowledge Portal



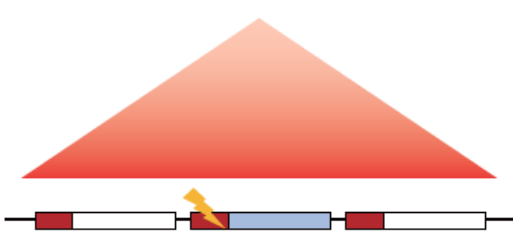
DNV: de novo variant



ASD gene-linked



ASD gene TAD



Non-ASD gene TAD



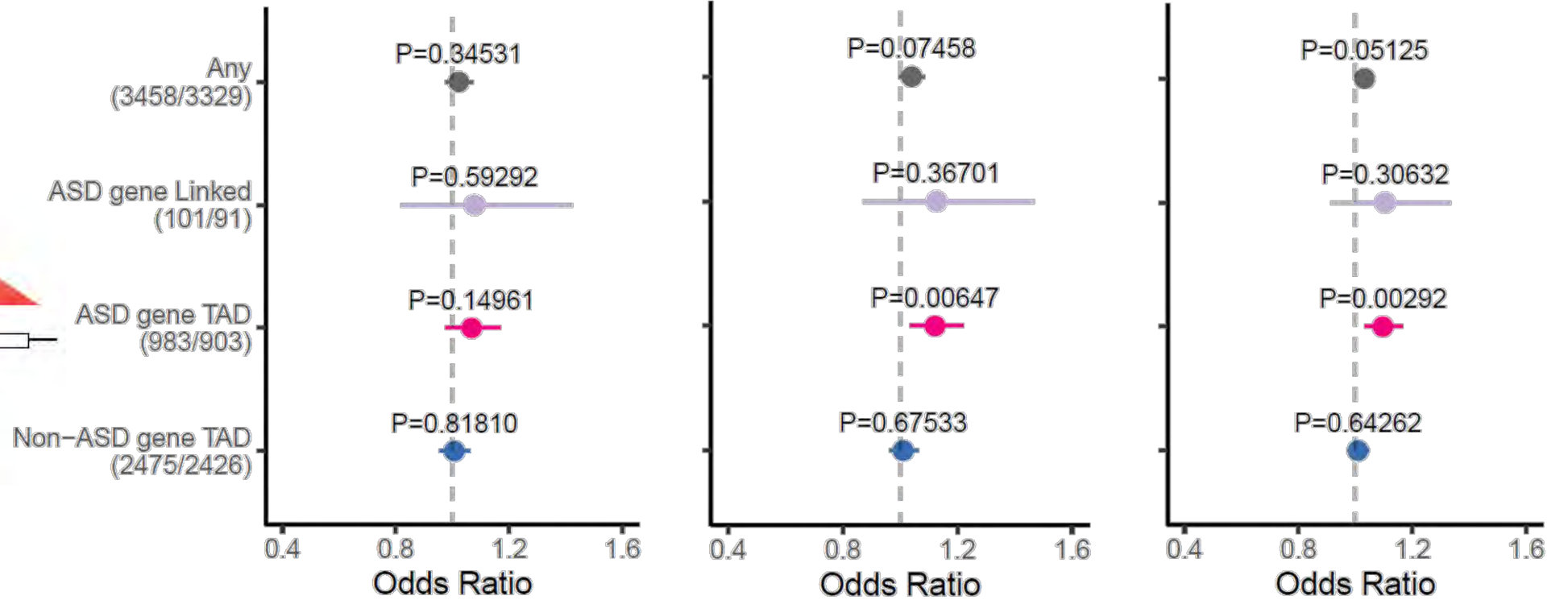
SSC
(1,902 ASD vs. 1,902 unaff sib)



SPARK
(3,142 ASD vs. 2,193 unaff sib)

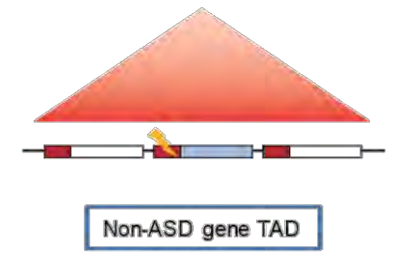
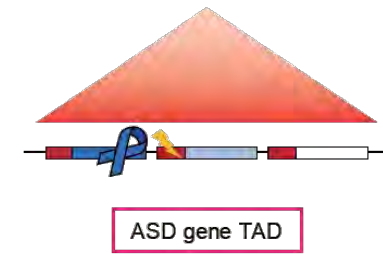
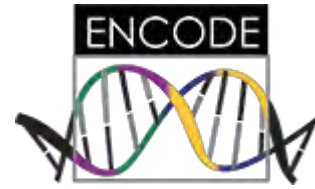
Meta analysis

SSC + SPARK
(5,044 ASD vs. 4,095 unaff sib)



Topologically associating domains (TADs) define the impact of de novo promoter variants on autism spectrum disorder risk

Analyses using TAD data from various tissues and cell lines

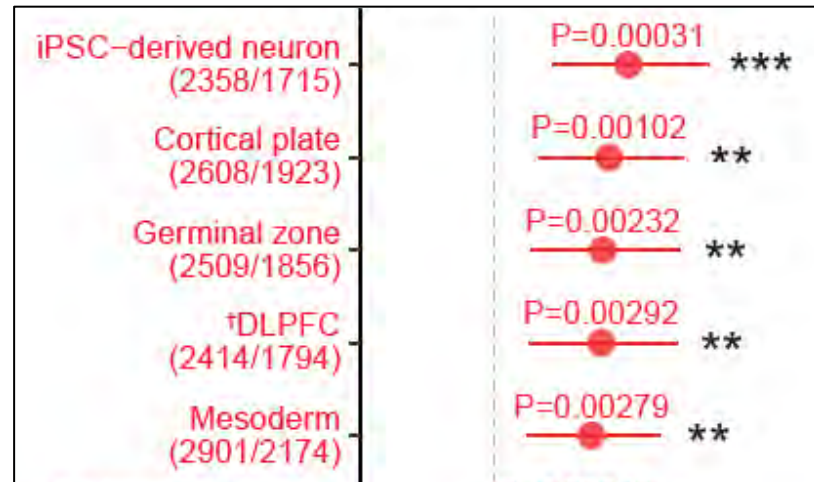
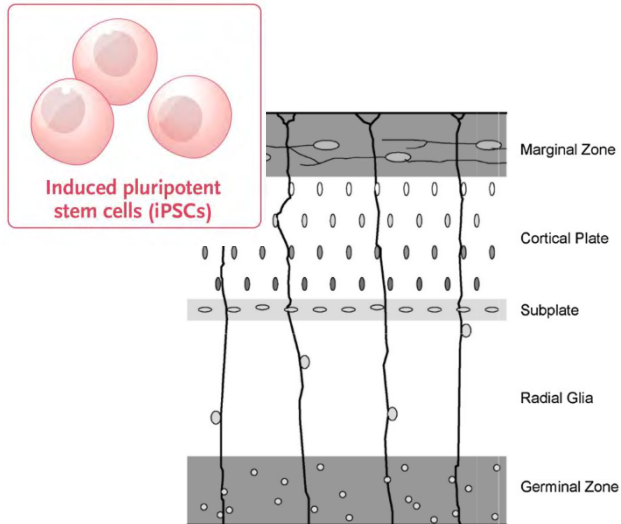


The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions

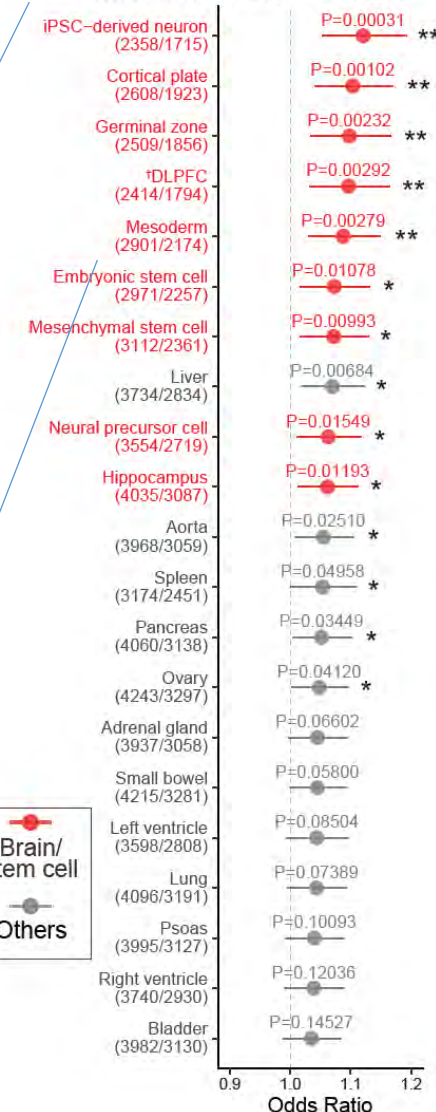
<https://www.encodeproject.org/>

Yanli Wang, Fan Song, Bo Zhang, Lijun Zhang, Jie Xu, Da Kuang, Daofeng Li, Mayank N. K. Choudhary, Yun Li, Ming Hu, Ross Hardison, Ting Wang & Feng Yue

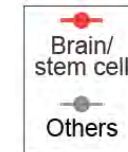
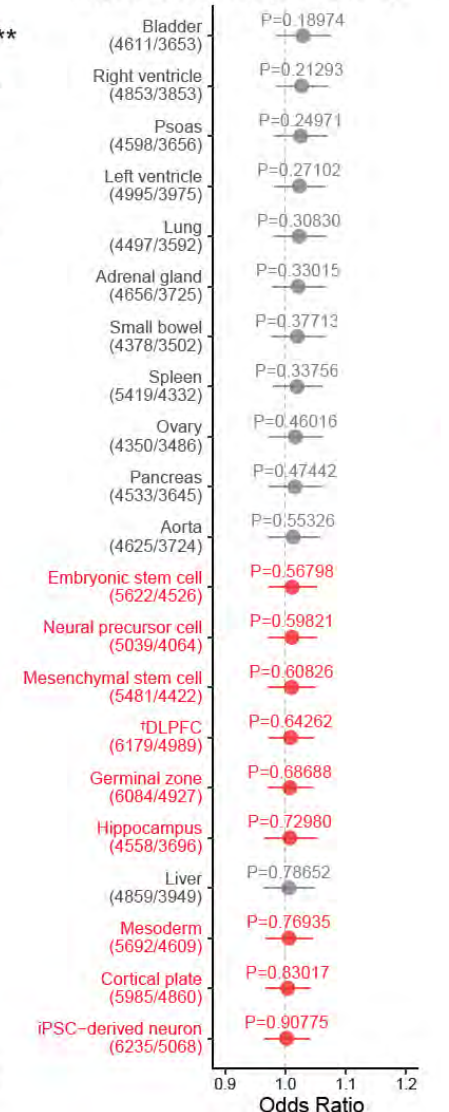
Genome Biology 19, Article number: 151 (2018)



Wilcoxon rank sum test, $P = 2.72 \times 10^{-5}$

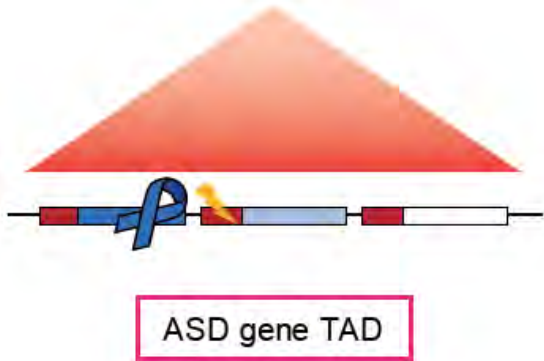


Wilcoxon rank sum test, $P = 2.04 \times 10^{-4}$



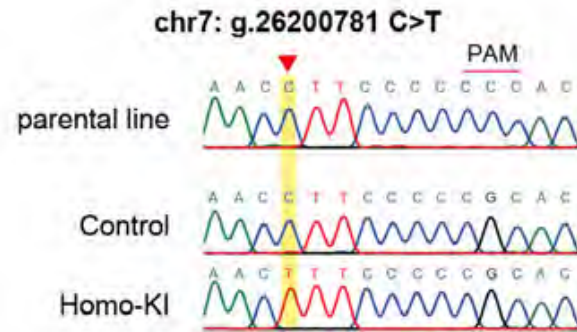
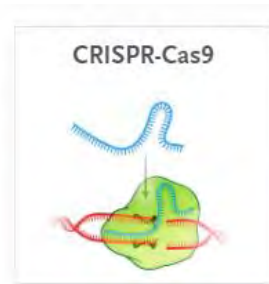
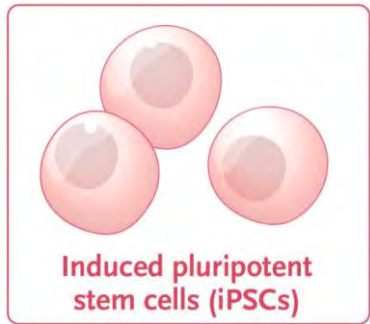
Enrichment of ASD gene TAD promoter de novo variants in ASD are pronounced when brain/stem cell TAD data are used

Experimental analysis of the effect of selected ASD gene TAD promoter DNVs on gene expression profiles

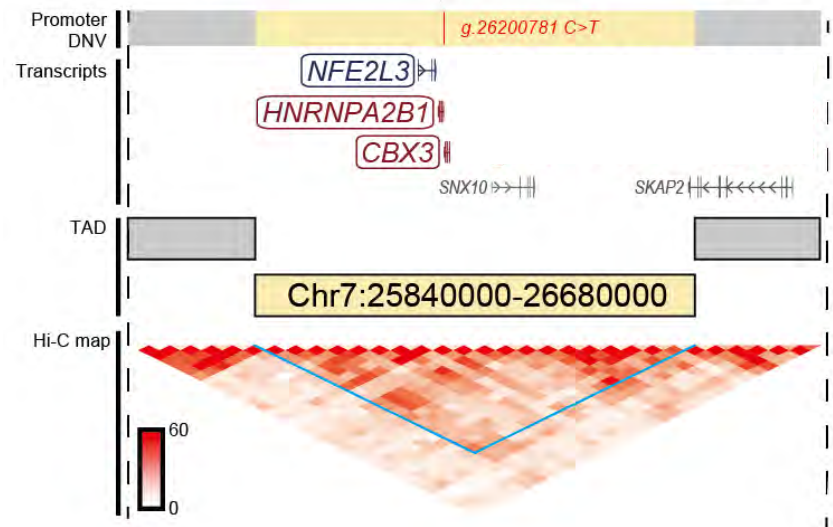
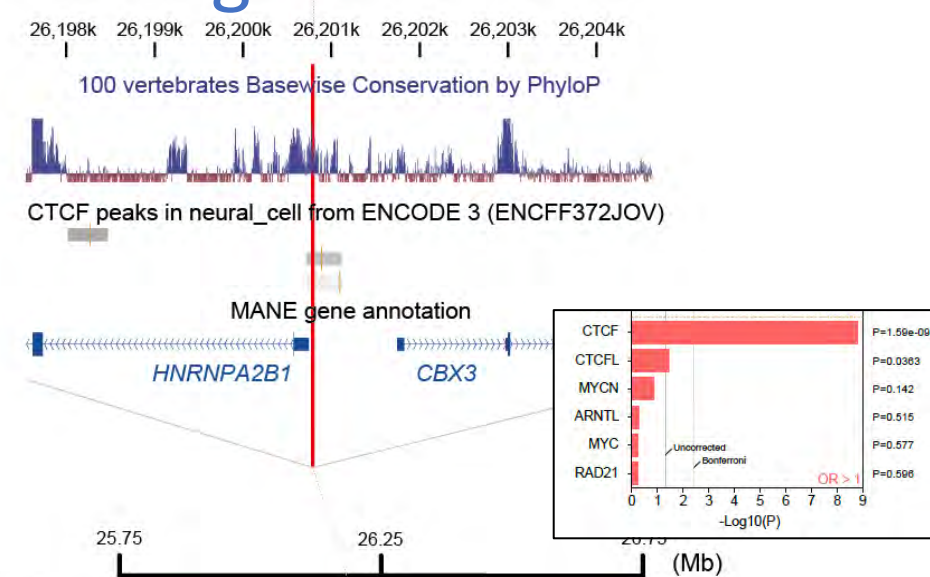


#1, chr7:g.26200781C>T

- Altering a conserved base
- Bound to CTCF in neuronal cells (ENCODE3)
- High-confidence ASD genes (SFARI gene score S, 1 or 2) in the same TAD
- High specificity gRNA can be designed



parental line 5'- AGAACCTTCCCCCGCACTAACGCGTCTTCCGCTACG -3'
Control 5'- AGAACCTTCCCCCGCACTAACGCGTCTTCCGCTACG -3'
Homo-KI 5'- AGAACCTTCCCCCGCACTAACGCGTCTTCCGCTACG -3'



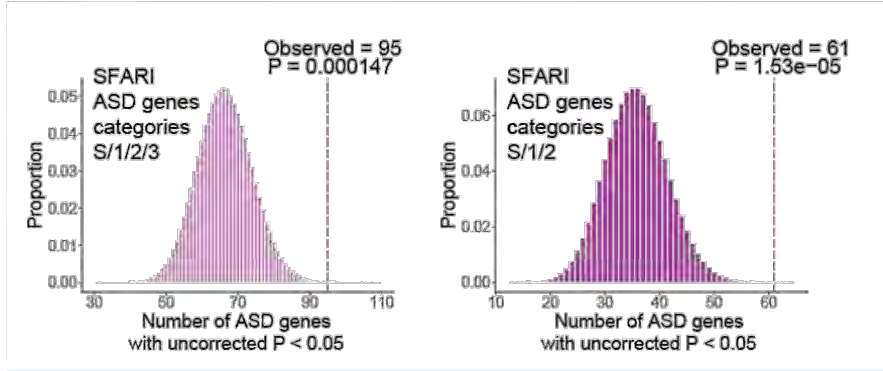
Local alteration of multiple genes in the same TAD

Effect of the chr7:g.26200781C>T variant on the transcriptomic profile

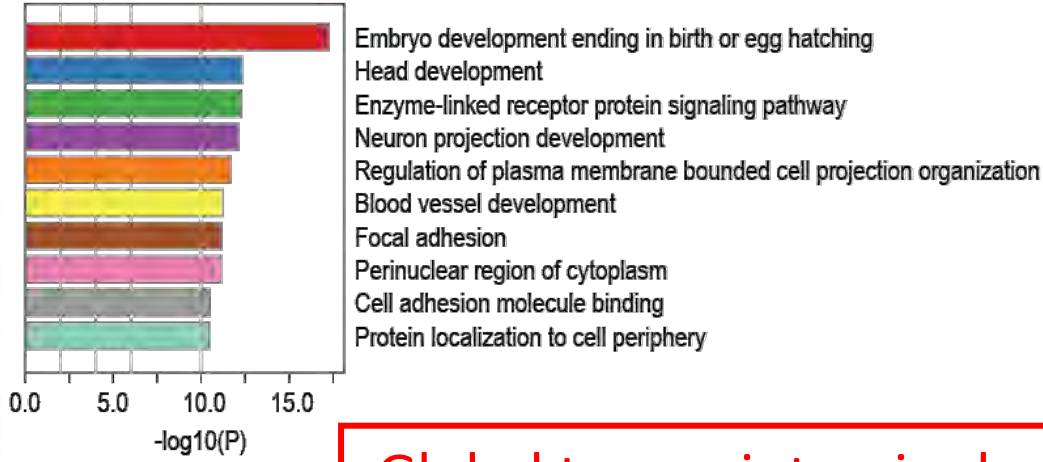
Significant enrichment of ASD genes

SFARI ASD genes categories S/1/2

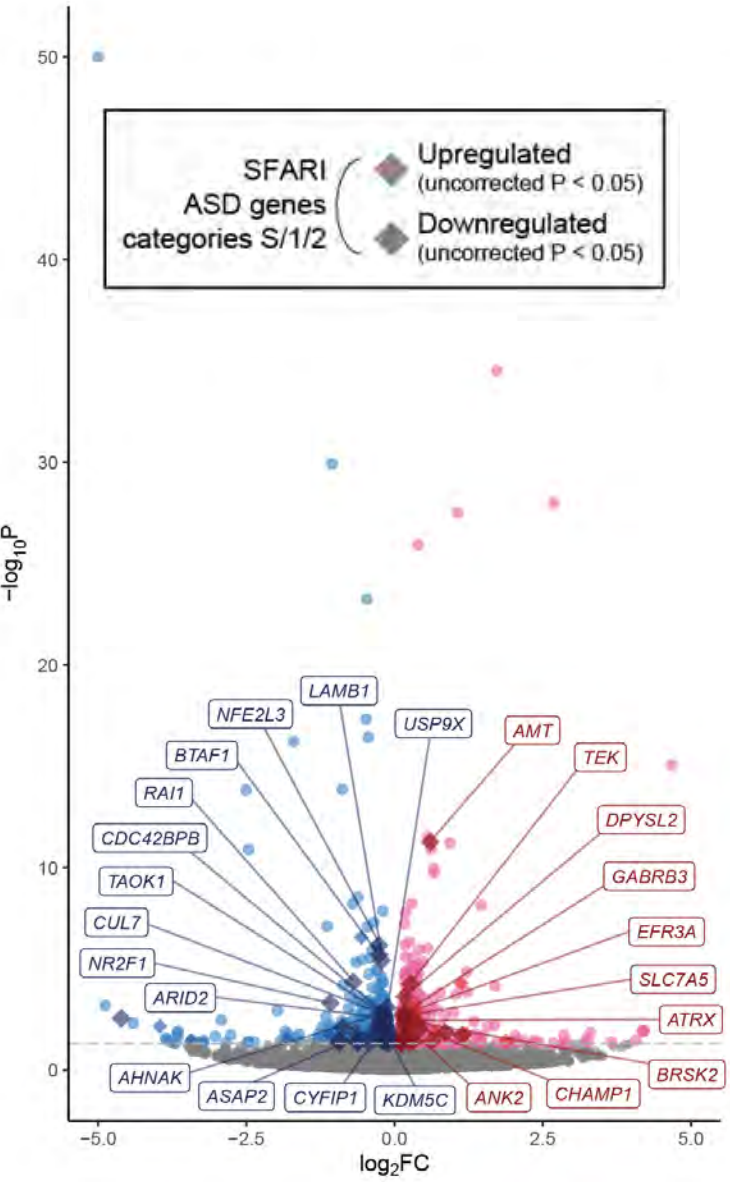
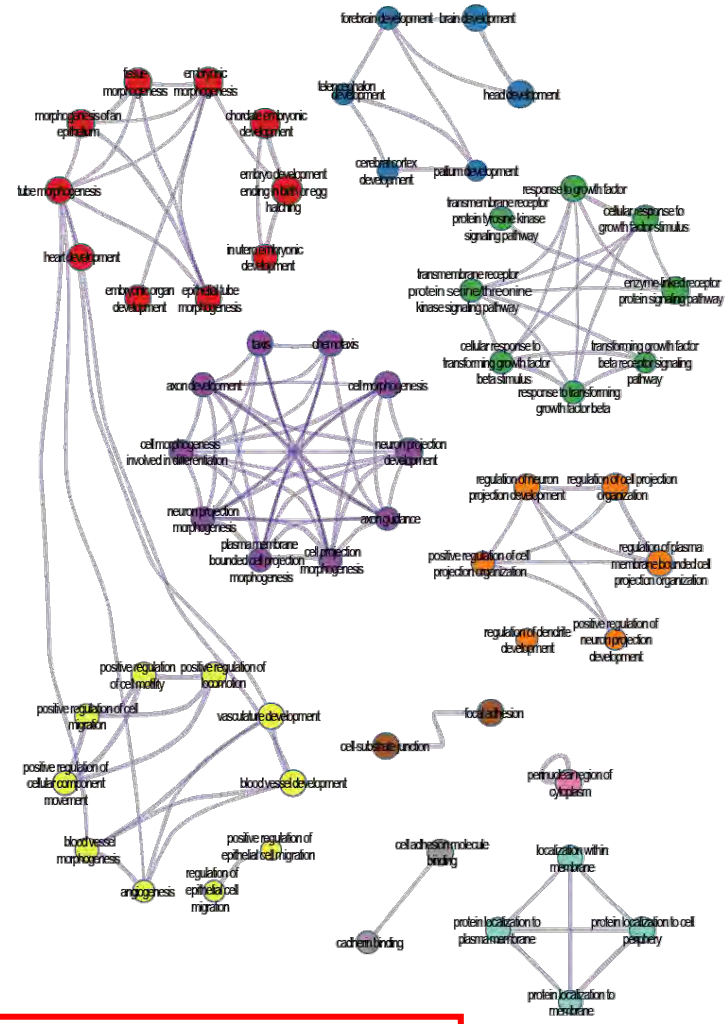
- Upregulated (uncorrected $P < 0.05$)
- Downregulated (uncorrected $P < 0.05$)



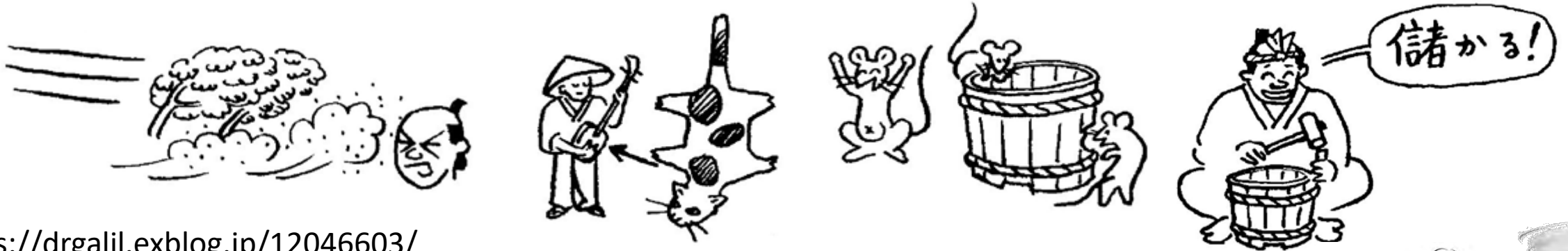
GO Enrichment Analysis



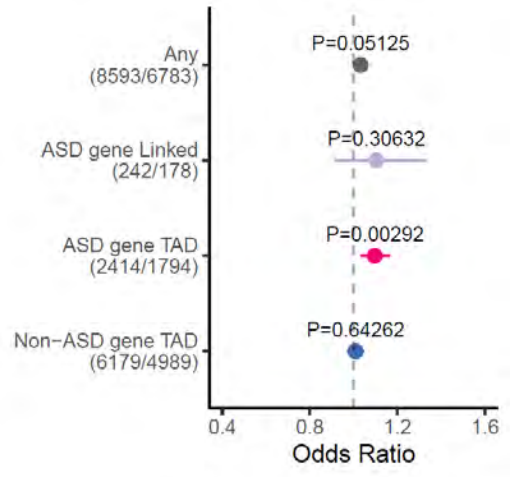
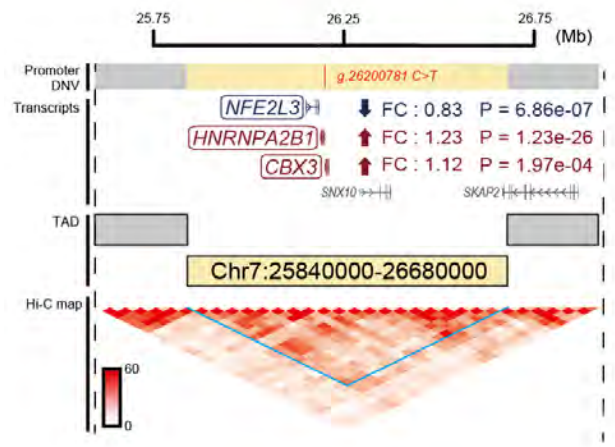
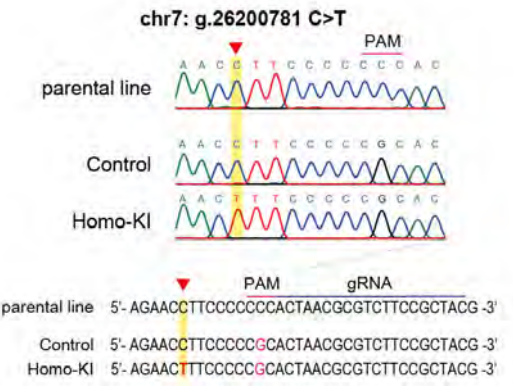
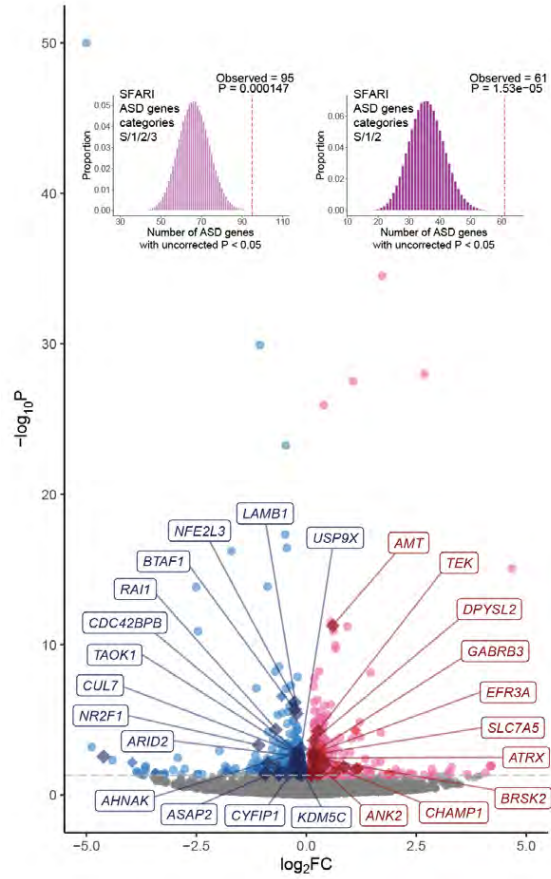
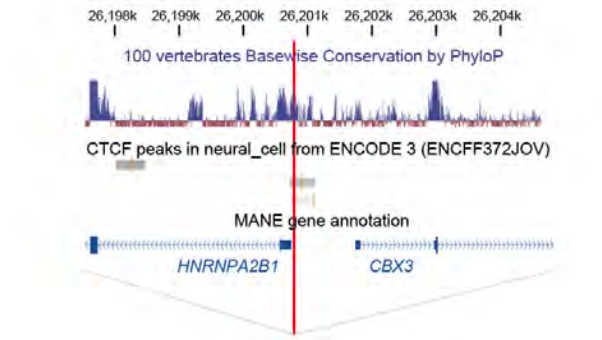
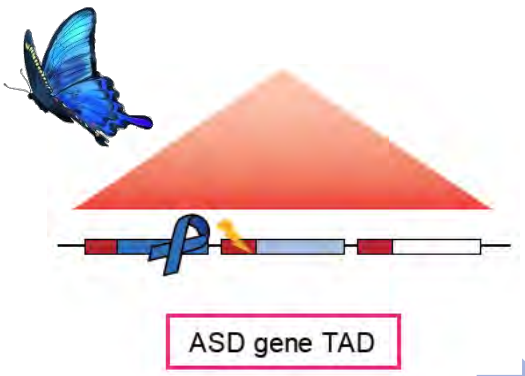
Global transcriptomic change enriched for ASD/neurodevelopment genes



風が吹けば桶屋が儲かる (as a strong wind blows, business comes to barrel makers)

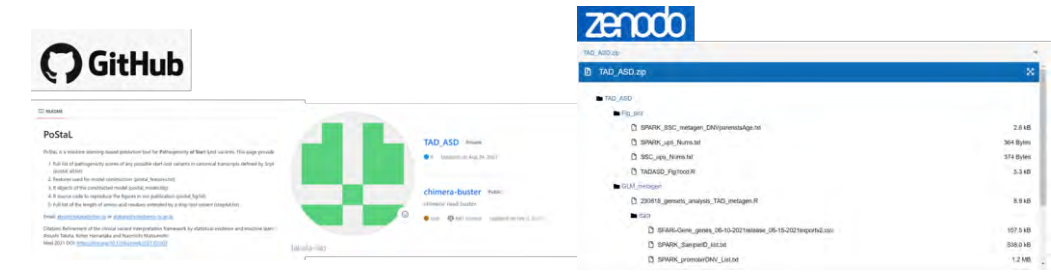


<https://drgalil.exblog.jp/12046603/>



Development and generation of publicly available resources

Utilization of open life science data



GitHub repository for TAD_ASD and Zenodo archive for TAD_ASD.zip

Human data NBDCヒトデータベース

Research ID	研究題目	公開日	データの種類	研究方法	手法	研究者 (対象集団)	提供者	アクセス制限
hum0421 v1	腸管神経系欠損における免疫反応	v1:2024/01/25	NGS (RNA-seq)	発現	Illumina (NovaSeq 5000)	自費スークラム 産業界共同で得られた変異をセレクトした遺伝子変異ヒトIPS細胞株: 5株 コントロール遺伝子変異ヒトIPS細胞株: 4株 (細胞株)	高野 (Type 1)	

CBS Data Sharing Platform

Transcriptomic dysregulation and autistic-like behaviors in Kmt2c haploinsufficient mice rescued by an LSD1 inhibitor

Takumi Nakamura¹, Toru Yoshihara², Chiharu Tanegashima³, Mitsutaka Kadota³, Yuki Kobayashi¹, Kûrara Honda¹, Mizuho Ishiwata¹, Junko Ueda¹, Tomonori Hara¹, Moe Nakanishi¹, Toru Takumi⁵, Shigeyoshi Itohara¹, Shigehiro Kuraku⁶, Masahide Asano², Takaoki Kasahara⁷, Kazuo Nakajima⁸, Takashi Tsuboi⁷, Atsushi Takata¹, Tadafumi Kato⁹




gnomAD Genome Aggregation Database

ClinVar Clinically relevant variation

HGMD[®] Human Gene Mutation Database

SIM NS FOUNDATION

SPARK Simons Powering Autism Research

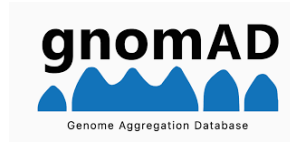
PROVEAN (Protein Variation Effect Analyzer)

ENCODE

Simons Simplex Collection

PsychENCODE Knowledge Portal

Summary



- By analyzing large open data of genomic variants from both the general population and affected individuals with statistical approaches and machine learning techniques, we proposed a method for refining standard genetic diagnosis guidelines.



Human Gene Mutation Database

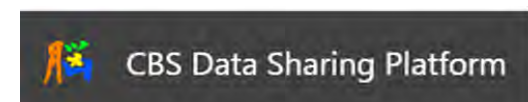


- An analysis of publicly available disease genome sequencing data from a unique viewpoint, utilizing open data for annotation, has led to the discovery of the role of TADs in the impact of promoter variants on ASD risk.



PsychENCODE
Knowledge Portal

- The formation of a virtuous cycle of utilizing open data and developing new open resources from it will be key to the success of open life sciences.



Acknowledgments

RIKEN CBS

Takumi Nakamura

Junko Ueda

Shota Mizuno

Naoki Hirose

An-a Kazuno

Kurara Honda

Emiko Koyama

Hirona Yamamoto

Tomonori Hara

Ayumu Kawasaki

Li Sin Yi

Risitha Subasinghe

Yuki Niwa

Sumina Atarashi

Hiroyo Yamaguchi

Tomoko Toyota

RRD members

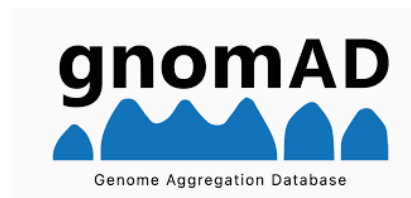
Yokohama City Univ

Kohei Hamanaka

Naomichi Matsumoto



and many others including
all study participants



and the providers/developers of
many other resources

