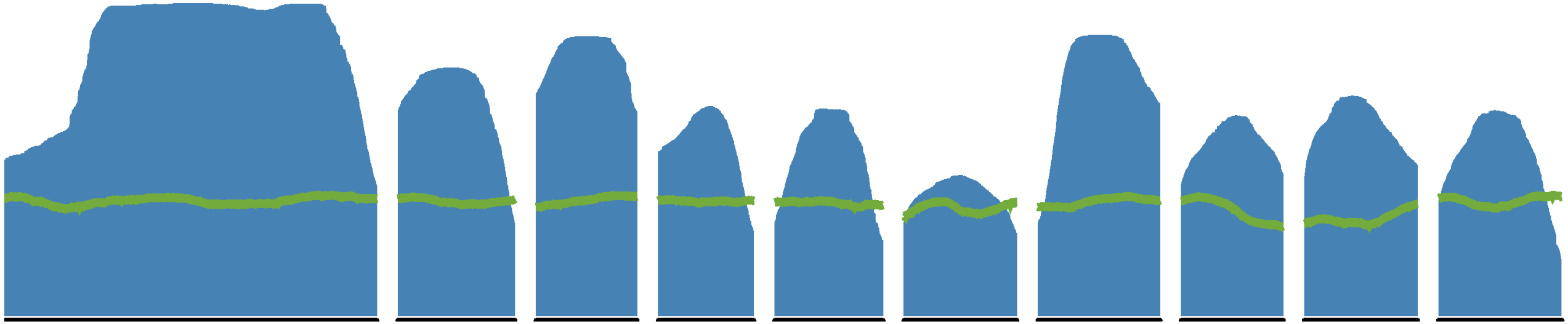


Use of the Genome Aggregation Database (gnomAD)

Anne O'Donnell-Luria, MD, PhD

Broad Institute of MIT and Harvard

Boston Children's Hospital

H3Africa Rare Disease Workshop February 2021

Twitter: @AnneOtation

Every genome contains many rare, potentially functional variants

- ~2 million high quality variants in a variant call file (vcf)
- ~20,000 variants within coding (exome) regions
- ~500 rare missense variants (1/3 of which are predicted
- [illegible]
- [illegible]
- ~50 reported disease-causing mutations (!)
- 1-2 *de novo* coding mutations
- Unknown number of sequencing errors

In Mendelian disease analysis, how can we identify the pathogenic genetic variant(s) in the sea of benign variation?

The power of 7.7 billion people

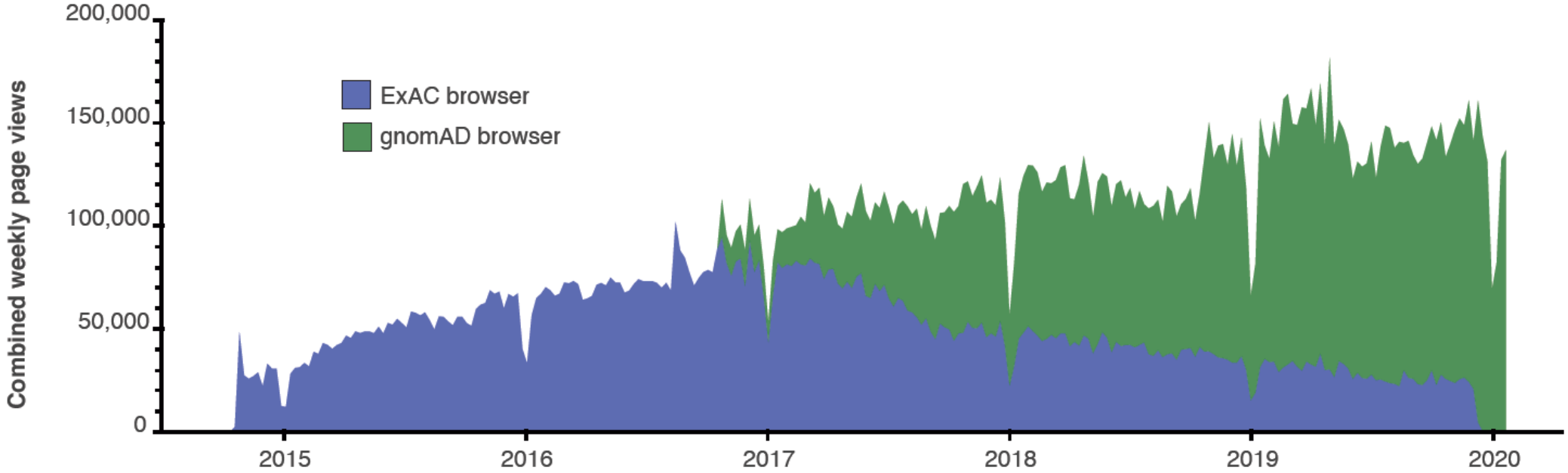
Given known mutation rates, it is almost certain that **every possible single base change compatible with life exists in a living human**

Variant aggregation at Broad

Exome Aggregation Consortium (ExAC) – v1
60,076 exomes
Genome Build 37
released October 2014

Genome Aggregation Database (gnomAD) – v2
125,748 exomes + 15,708 genomes
Genome Build 37
released October 2016

Genome Aggregation Database (gnomAD) – v3
71,702 genomes
Genome Build 38
released October 2019

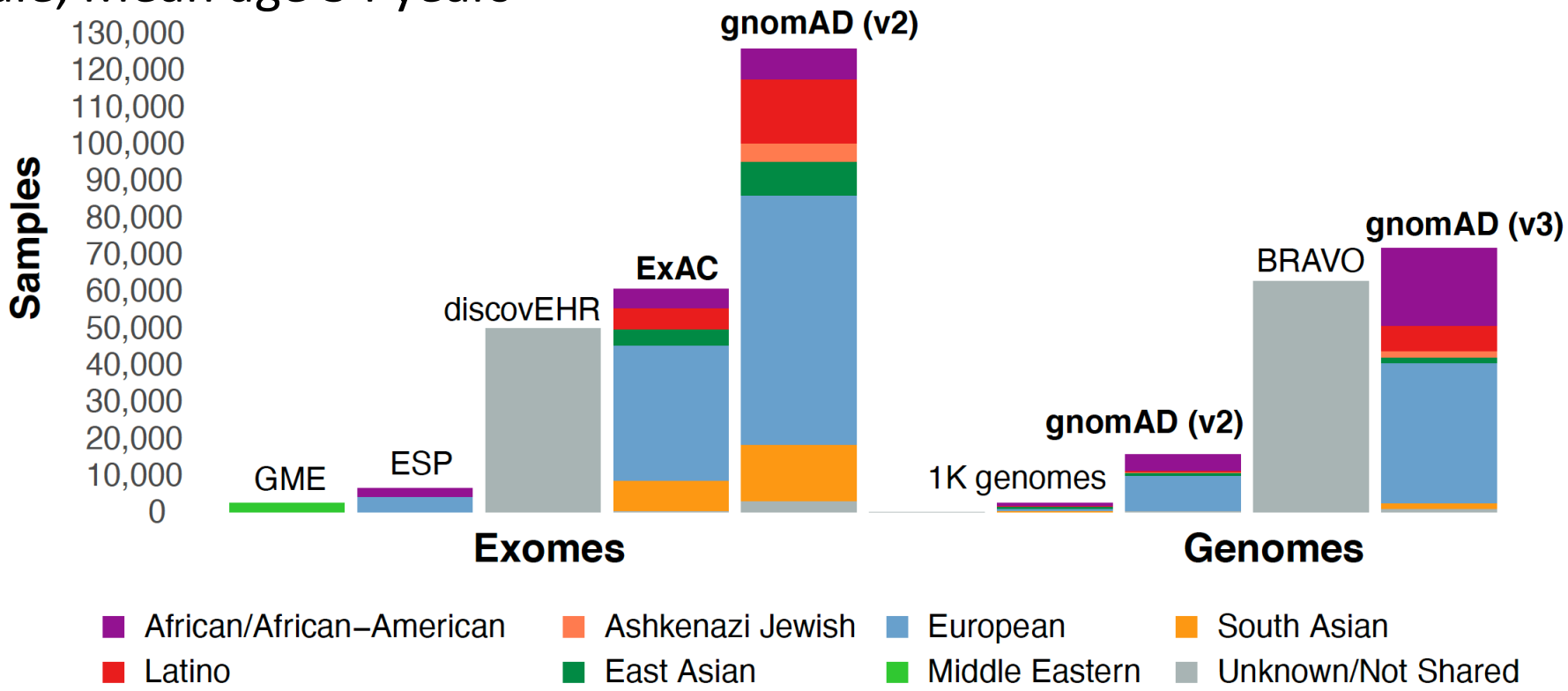


- 20+ M pageviews from 184 countries
- Aided in the diagnosis of over 200,000 patients with rare disease

Matt Solomonson

Who's in the gnomAD database?

- Case-control studies of complex adult-onset diseases (e.g. type 2 diabetes, heart attack, migraine, bipolar)
- Depleted as much as possible of people **known** to have severe pediatric disease, and their first-degree relatives
- 55% Male, Mean age 54 years



Grace Tiao

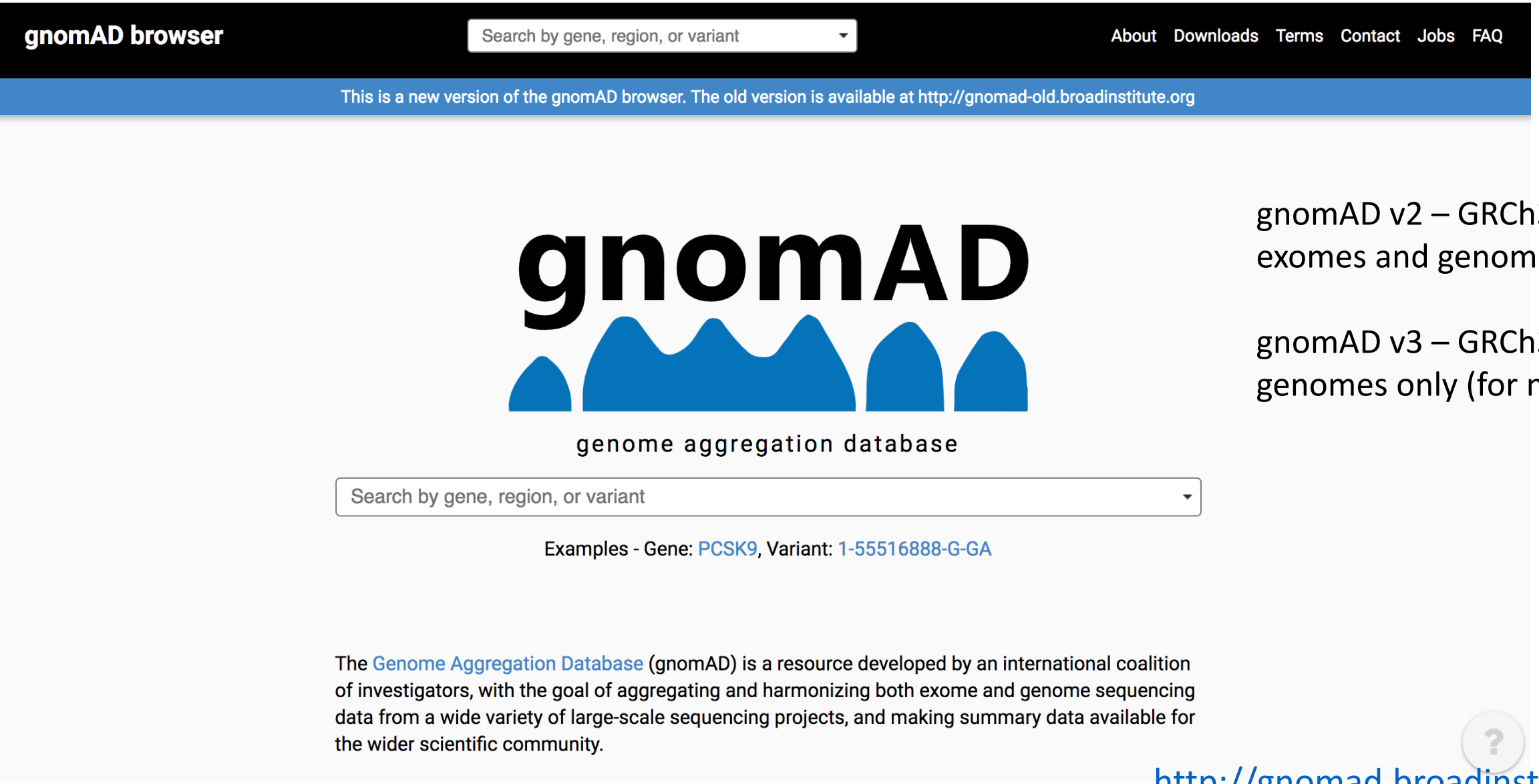
Thank you to the PIs who share data in gnomAD

Data processed through same pipeline and called jointly

Maria Abreu
Carlos A Aguilar Salinas
Tariq Ahmad
Christine M. Albert
Nicholette Allred
David Altshuler
Diego Ardissino
Gil Atzmon
John Barnard
Laurent Beaugerie
Gary Beecham
Emelia J. Benjamin
Michael Boehnke
Lori Bonnycastle
Erwin Bottinger
Donald Bowden
Matthew Bown
Steven Brant
Hannia Campos
John Chambers
Juliana Chan
Daniel Chasman
Rex Chisholm
Judy Cho
Rajiv Chowdhury
Mina Chung
Wendy Chung
Bruce Cohen
Adolfo Correa
Dana Dabelea
Mark Daly
John Danesh
Dawood Darbar
Joshua Denny
Ravindranath Duggirala
Josée Dupuis
Patrick T. Ellinor
Roberto Elosua
Jeanette Erdmann
Tõnu Esko
Martt Färkkilä
Diane Fatkin
Jose Florez
Andre Franke
Gad Getz
David Glahn
Ben Glaser
Stephen Glatt
David Goldstein
Clicerio Gonzalez
Leif Groop
Christopher Haiman
Ira Hall
Craig Hanis
Matthew Harms
Mikko Hiltunen
Matti Holi
Christina Hultman
Chaim Jalas
Mikko Kallela
Jaakko Kaprio
Sekar Kathiresan
Eimear Kenny
Bong-Jo Kim
Young Jin Kim
George Kirov
Jaspal Kooner
Seppo Koskinen
Harlan M. Krumholz
Subra Kugathasan
Soo Heon Kwak
Markku Laakso
Terho Lehtimäki
Ruth Loos
Steven A. Lubitz
Ronald Ma
Daniel MacArthur
Gregory M. Marcus
Jaume Marrugat
Kari Mattila
Steve McCarroll
Mark McCarthy
Jacob McCauley
Dermot McGovern
Ruth McPherson
James Meigs
Olle Melander
Deborah Meyers
Lili Milani
Braxton Mitchell
Aliya Naheed
Saman Nazarian
Ben Neale
Peter Nilsson
Michael O'Donovan
Dost Ongur
Lorena Orozco
Michael Owen
Colin Palmer
Aarno Palotie
Kyong Soo Park
Carlos Pato
Ann Pulver
Dan Rader
Nazneen Rahman
Alex Reiner
Anne Remes
Stephen Rich
John D. Rioux
Samuli Ripatti
Dan Roden
Jerome I. Rotter
Danish Saleheen
Veikko Salomaa
Nilesh Samani
Jeremiah Scharf
Heribert Schunkert
Svati Shah
Moore Shoemaker
Tai Shyong
Edwin K. Silverman
Pamela Sklar
Gustav Smith
Hilkka Soininen
Harry Sokol
Tim Spector
Nathan Stitziel
Patrick Sullivan
Jaana Suvisaari
Kent Taylor
Yik Ying Teo
Tuomi Tiinamaija
Ming Tsuang
Dan Turner
Teresa Tusie Luna
Erkki Vartiainen
James Ware
Hugh Watkins
Rinse Weersma
Maija Wessman
James Wilson
Ramnik Xavier

Broad Genomics Platform

Broad Data Sciences Platform



gnomAD v2 – GRCh37
exomes and genomes

gnomAD v3 – GRCh38
genomes only (for now)

http://gnomad.broadinstitute.org

MPG Primers on using gnomAD

https://www.youtube.com/watch?v=J29_Pf8Ti1s
https://www.youtube.com/watch?v=Bh8AKkI-DhY

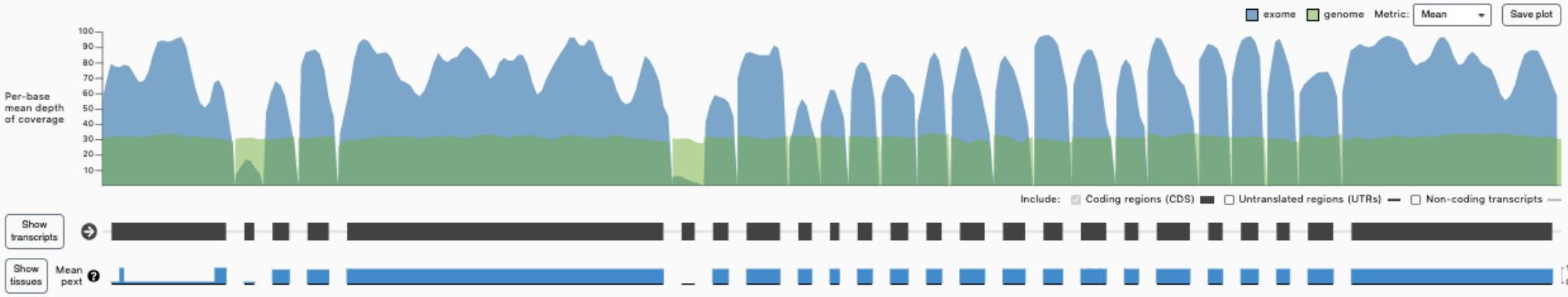
NSD1 nuclear receptor binding SET domain protein 1

Dataset gnomAD v2.1.1 gnomAD SVs v2.1

Genome build GRCh37 / hg19
Ensembl gene ID ENSG00000165671.14
Ensembl canonical transcript ENST00000439151.2
Other transcripts ENST00000347982.4, ENST00000503056.1, and 13 more
Region 5:176560026-176727216
References Ensembl, UCSC Browser, and more

Constraint

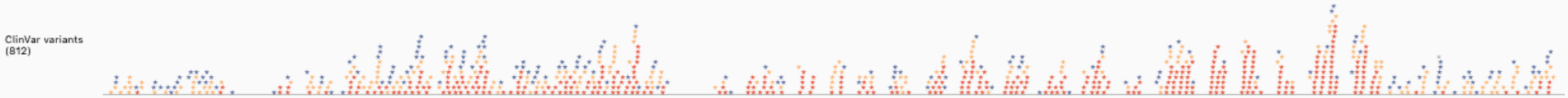
Category	Exp. SNVs	Obs. SNVs	Constraint metrics
Synonymous	525.3	523	Z = 0.08 o/e = 1 (0.93 - 1.07)
Missense	1427.7	1065	Z = 3.41 o/e = 0.75 (0.71 - 0.79)
pLoF	110.6	5	pLI = 1 o/e = 0.05 (0.02 - 0.1)



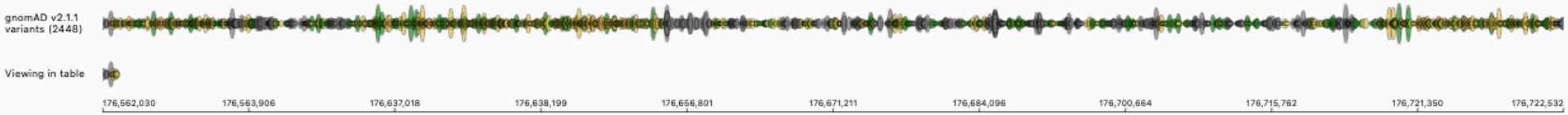
ClinVar variants

Pathogenic / likely pathogenic only | Uncertain significance / conflicting only | Benign / likely benign only | Other only

Collapse to bins



gnomAD variants



pLoF only | Missense only | Synonymous only | Other only

Exomes Genomes | SNVs Indels | Filtered variants

Search variant table

Export variants to CSV

Note Only variants located in or within 75 base pairs of a coding exon are shown here. To see variants in UTRs or introns, use the region view.

The table below shows the HGVS consequence and VEP annotation for each variant's most severe consequence across all transcripts in this gene. Cases where the most severe consequence occurs in a non-canonical transcript are denoted with †. To see consequences in a specific transcript, use the transcript view.

Variant ID	Source	HGVS Consequence	VEP Annotation	Clinical Significance	Flags	Allele Count	Allele Number	Allele Frequency	Number of Homozygote
5-176562031-AGAGTC-A	G	c.-17-51_-17-47delGAGTC	intron			2	31382	6.37e-5	0
5-176562058-C-A	G	c.-17-30C>A	intron			1	31394	3.19e-5	0
5-176562062-T-G	E	c.-17-26T>G	intron			2	250150	8.00e-6	0
5-176562065-A-G	E	c.-17-23A>G	intron			3	250392	1.20e-5	0

Single nucleotide variant: 5-176562864-C-T (GRCh37)

Dataset: gnomAD v2.1.1

	Exomes	Genomes	Total
Filter	Pass	Pass	
Allele Count	88	38	126
Allele Number	249800	31406	281206
Allele Frequency	0.0003523	0.001210	0.0004481
Popmax Filtering AF (95% confidence)	0.004016	0.003068	
Number of homozygotes	0	0	0

This variant is multiallelic. Other alt alleles are:

- 5-176562864-C-G

References

- dbSNP (rs149334244)
- UCSC
- ClinVar (159441)

Report

- Report this variant
- Request additional information

Annotations

This variant falls on 8 transcripts in 1 gene.

missense

- NSD1
  - ENST00000439151 *
    HGVSp: p.Leu254Phe
    Polyphen: possibly_damaging
    SIFT: tolerated_low_confidence
  - ENST00000361032
    HGVSp: p.Leu254Phe
    Polyphen: possibly_damaging
    SIFT: tolerated_low_confidence

intron

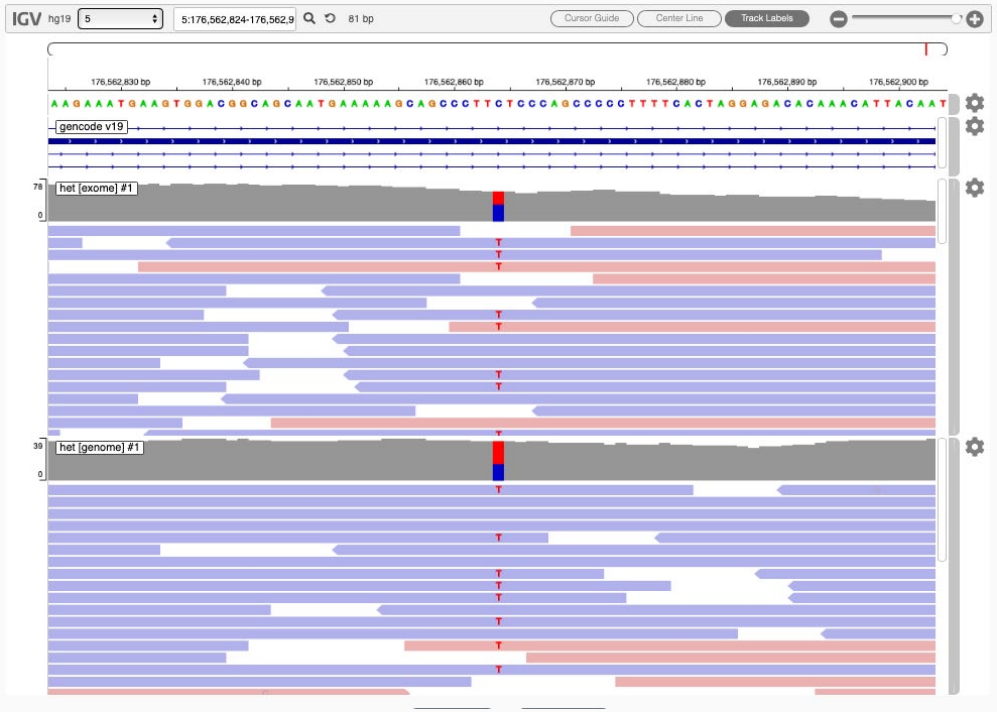
- NSD1
  - ENST00000347982
  - ENST00000354179
  - ENST00000508896
  - and 2 more

non coding transcript exon

- NSD1
  - ENST00000602285

Population Frequencies

Population	Allele Count	Allele Number	Number of Homozygote
African	115	24888	0
Other	3	7168	0
Latino	7	35024	0
European (non-Finnish)	1	128728	0
Ashkenazi Jewish	0	10264	0
East Asian	0	19910	0
European (Finnish)	0	25070	0
South Asian	0	30154	0
Female	72	128682	0
Male	54	152524	0
Total	**126**	**281206**	**0**



Population Frequencies

Population		Allele Count	Allele Number	Number of Homozygotes	Allele Frequency
African		115	24888	0	0.004621
Other		3	7168	0	0.0004185
Latino		7	35024	0	0.0001999
European (non-Finnish)	Overall	1	128728	0	0.000007768
	Other non-Finnish European	1	32984	0	0.00003032
	Bulgarian	0	2666	0	0.000
	Estonian	0	4836	0	0.000
	North-western European	0	50588	0	0.000
	Southern European	0	11578	0	0.000
	Swedish	0	26076	0	0.000
	Male	1	72028	0	0.00001388
	Female	0	56700	0	0.000
Ashkenazi Jewish		0	10264	0	0.000
East Asian	Overall	0	19910	0	0.000
	Japanese	0	150	0	0.000
	Korean	0	3818	0	0.000
	Other East Asian	0	14382	0	0.000
	Female	0	9840	0	0.000
	Male	0	10070	0	0.000
European (Finnish)		0	25070	0	0.000
South Asian		0	30154	0	0.000
emale		72	128682	0	0.0005595
Male		54	152524	0	0.0003540
Total		**126**	**281206**	**0**	**0.0004481**

Additional features in gnomAD v3 (GRCh38)

In silico predictors

- REVEL: 0.325
- CADD: 21.9
- SpliceAI: 0.00 (No consequence)
- PrimateAI: 0.517

Released last week (also in gnomAD v2)

ClinVar

ClinVar Variation ID 96079
Conditions not specified, Beckwith-Wiedemann syndrome, History of neurodevelopmental disorder
Clinical significance Benign
Review status criteria provided, multiple submitters, no conflicts (2 stars)
Last evaluated 2019-12-31

See all 5 submissions or find more information on the ClinVar website.

Population Frequencies

gnomAD | HGDP | 1KG

Population	Allele Count	Allele Number	Number of Homozygotes	Allele Frequency
South Asian	262	4832	12	0.05422
Other	52	2092	0	0.02486
African/African-American	1027	41440	14	0.02478
European (non-Finnish)	1573	68026	21	0.02312
Middle Eastern	7	316	0	0.02215
Latino/Admixed American	326	15290	3	0.02132
East Asian	89	5206	3	0.01710
European (Finnish)	124	10610	0	0.01169
Amish	10	912	0	0.01096
Ashkenazi Jewish	14	3470	0	0.004035
XX	1796	77830	26	0.02308
XY	1688	74364	27	0.02270
Total	**3484**	**152194**	**53**	**0.02289**

Additional features in gnomAD v3 (GRCh38)

Two open access datasets

Human Genome Diversity Project (HGDP)

1000Genomes (1KG)

gnomAD | HGDP | 1KG

Population	Allele Count	Allele Number	Number of Homozygotes	Allele Frequency
Central/South Asian	25	332	1	0.07530
African	6	96	0	0.06250
Middle Eastern	10	308	0	0.03247
East Asian	14	456	0	0.03070
European	5	298	0	0.01678
Native American	0	70	0	0.000
XX	16	568	0	0.02817
XY	44	992	1	0.04435
Total	**60**	**1560**	**1**	**0.03846**

Warning Because of low sample sizes for HGDP populations, allele frequencies may not be representative.

Additional features in gnomAD v3 (GRCh38)

Two open access datasets

Human Genome Diversity Project (HGDP)

1000Genomes (1KG)

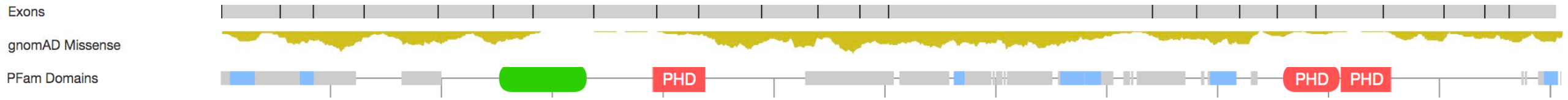
gnomAD | HGDP | 1KG

Population		Allele Count	Allele Number	Number of Homozygotes	Allele Frequency
▸ Central/South Asian		25	332	1	0.07530
▾ African	Overall	6	96	0	0.06250
	Bantu (Kenya)	3	14	0	0.2143
	Mandenka	2	42	0	0.04762
	Yoruba	1	38	0	0.02632
	Bantu (South Africa)	0	2	0	0.000
	XX	2	32	0	0.06250
	XY	4	64	0	0.06250
▸ Middle Eastern		10	308	0	0.03247
▸ East Asian		14	456	0	0.03070
▸ European		5	298	0	0.01678
▸ Native American		0	70	0	0.000
XX		16	568	0	0.02817
XY		44	992	1	0.04435
Total		**60**	**1560**	**1**	**0.03846**

Warning Because of low sample sizes for HGDP populations, allele frequencies may not be representative.

Constraint

- Where to we see depletion of variation across human populations?



Conservation is between species
Constraint is within the human species

Developed a mutational model to calculate expected number of variants per gene

Category	Exp. SNVs	Obs. SNVs
Synonymous	525.3	523
Missense	1427.7	1065
pLoF	110.6	5

Kaitlin Samocha

Konrad Karczewski

Daniel MacArthur

Mark Daly

(Samocha *et al*. 2014; Lek *et al*. 2016)

Constraint metrics

Category	Exp. SNVs	Obs. SNVs	Constraint metrics
Synonymous	525.3	523	Z = 0.08 o/e = 1 (0.93 - 1.07)
Missense	1427.7	1065	Z = 3.41 o/e = 0.75 (0.71 - 0.79)
pLoF	110.6	5	pLI = 1 o/e = 0.05 (0.02 - 0.1)

LOEUF

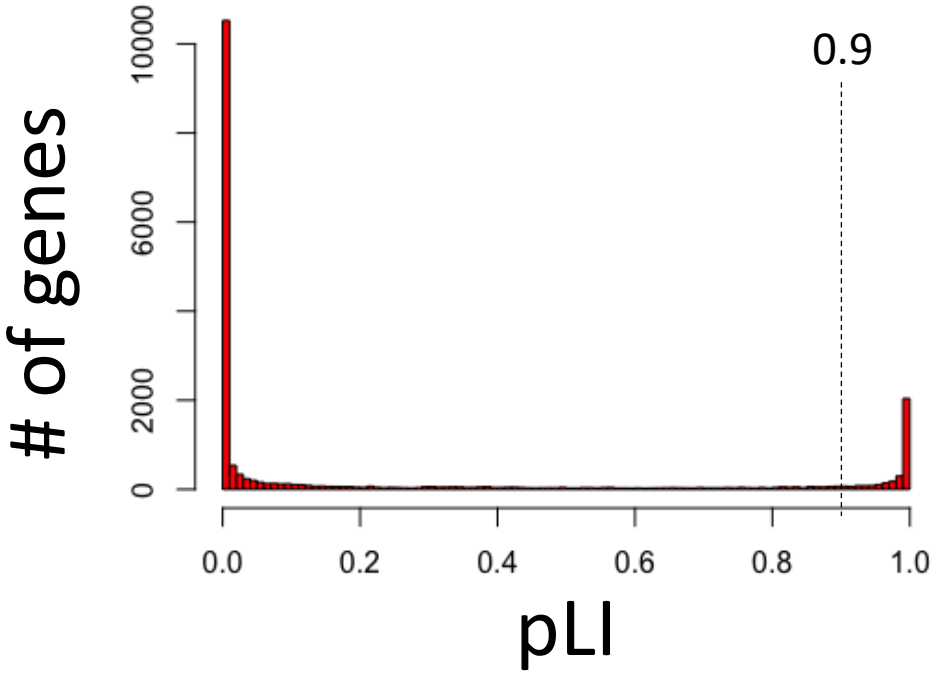
- Used z-scores to assess depletion from expectation
- For pLoF variants, correlation with gene length (short genes under powered)

Constraint metrics

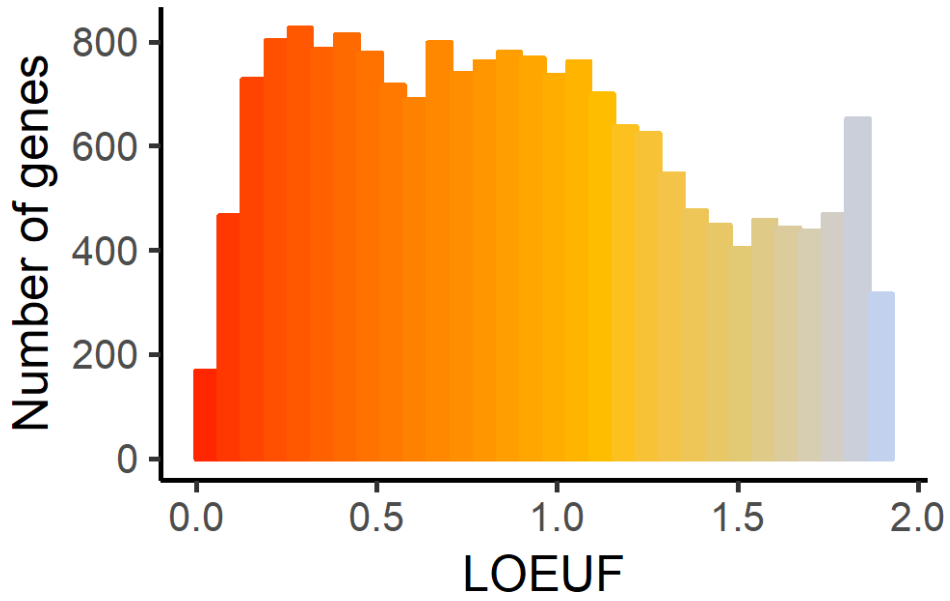
Category	Exp. SNVs	Obs. SNVs	Constraint metrics
Synonymous	525.3	523	Z = 0.08 o/e = 1 (0.93 - 1.07)
Missense	1427.7	1065	Z = 3.41 o/e = 0.75 (0.71 - 0.79)
pLoF	110.6	5	pLI = 1 o/e = 0.05 (0.02 - 0.1)

LOEUF

pLI = Probability of loss function intolerance
Dichotomous metric (>0.9 LoF constrained)



LOEUF = Loss-of-function observed/expected upper bound fraction
Continuous metric (lower scores more LoF constrained)



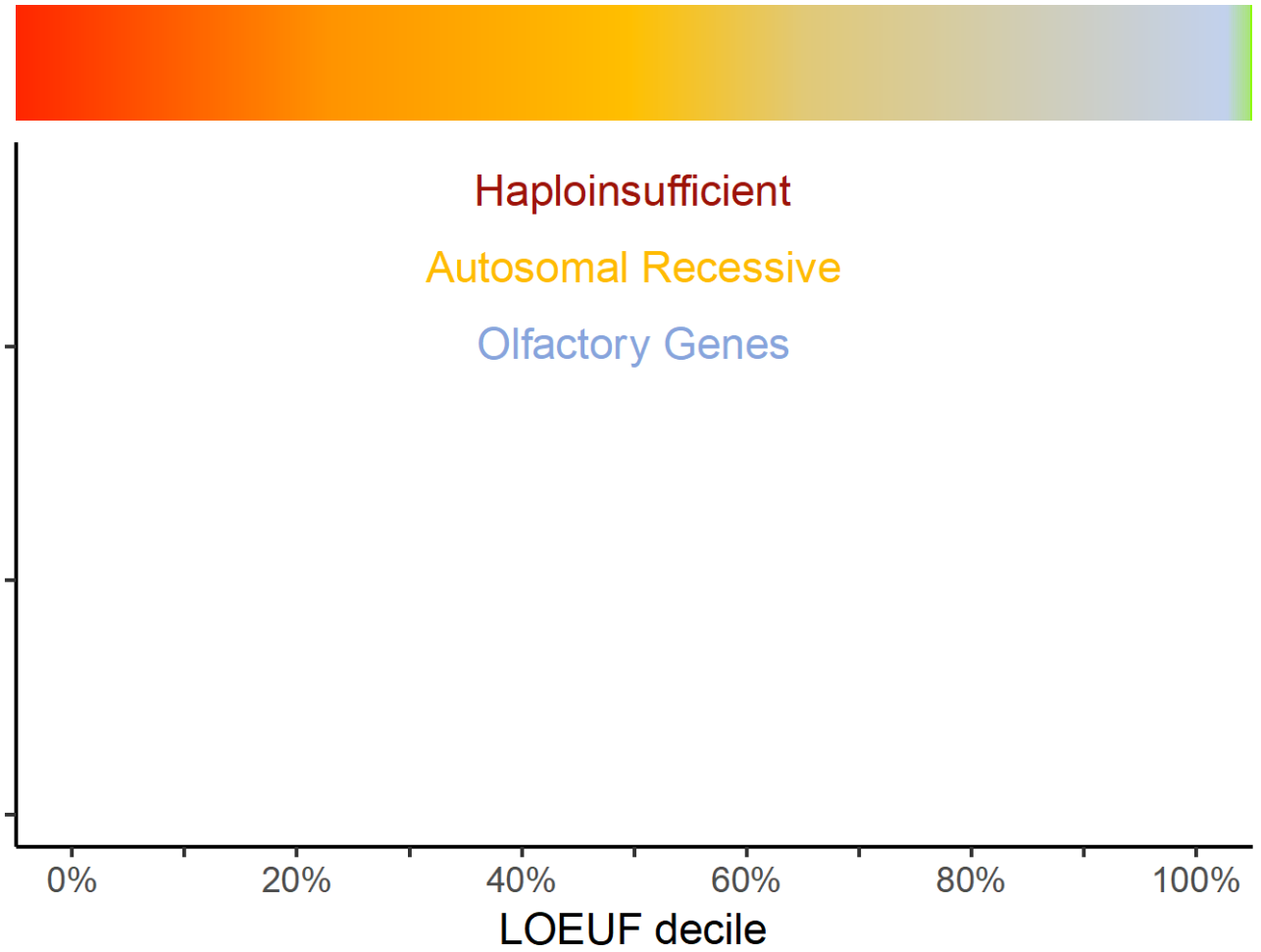
Samocha *et al*. 2014; Lek *et al*. 2016

Karczewski *et al*., Nature, 2020

Resolving the spectrum of LoF intolerance

- Binning this spectrum into deciles

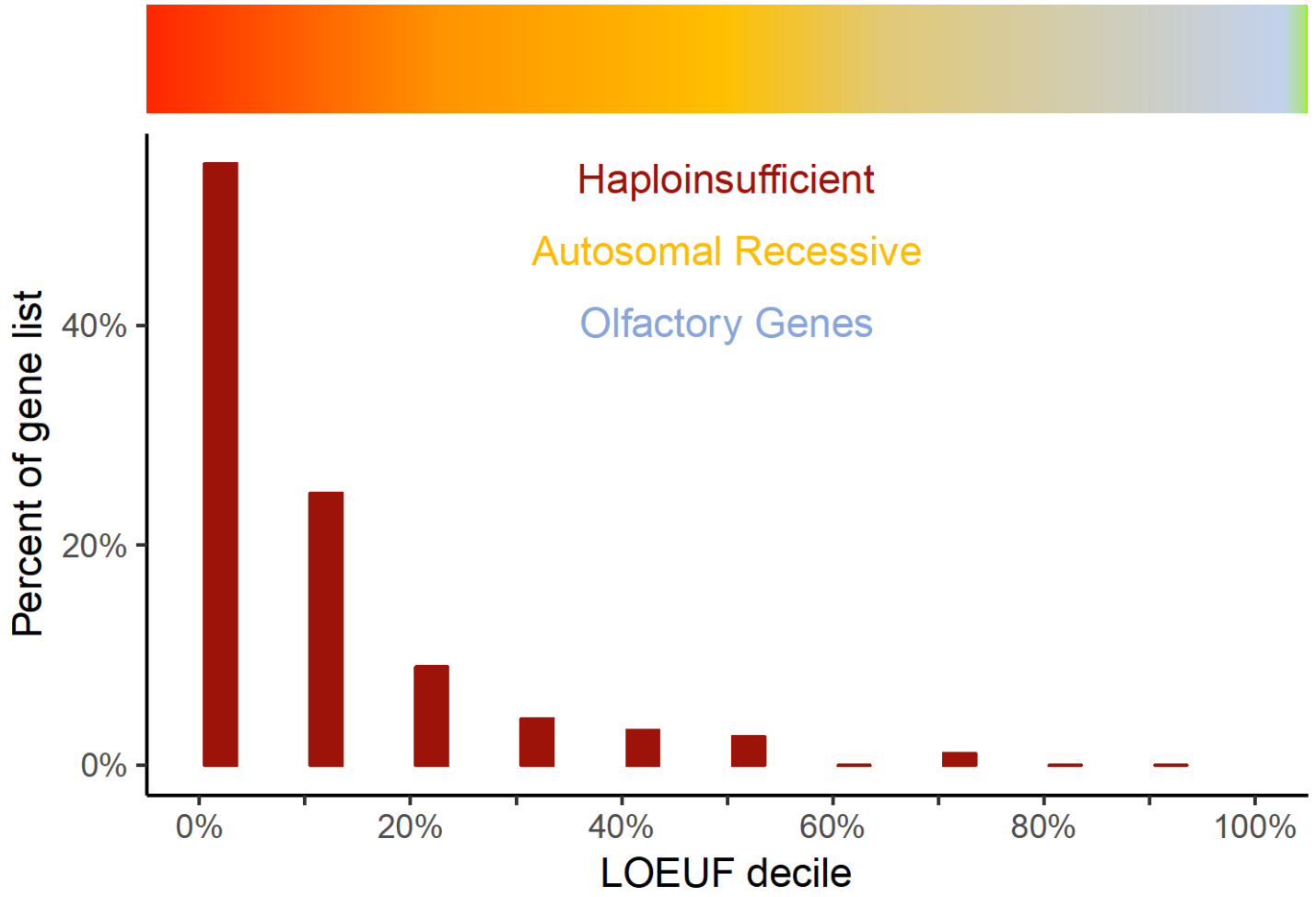
More depleted
More constrained



More tolerant
Less constrained

Resolving the spectrum of LoF intolerance

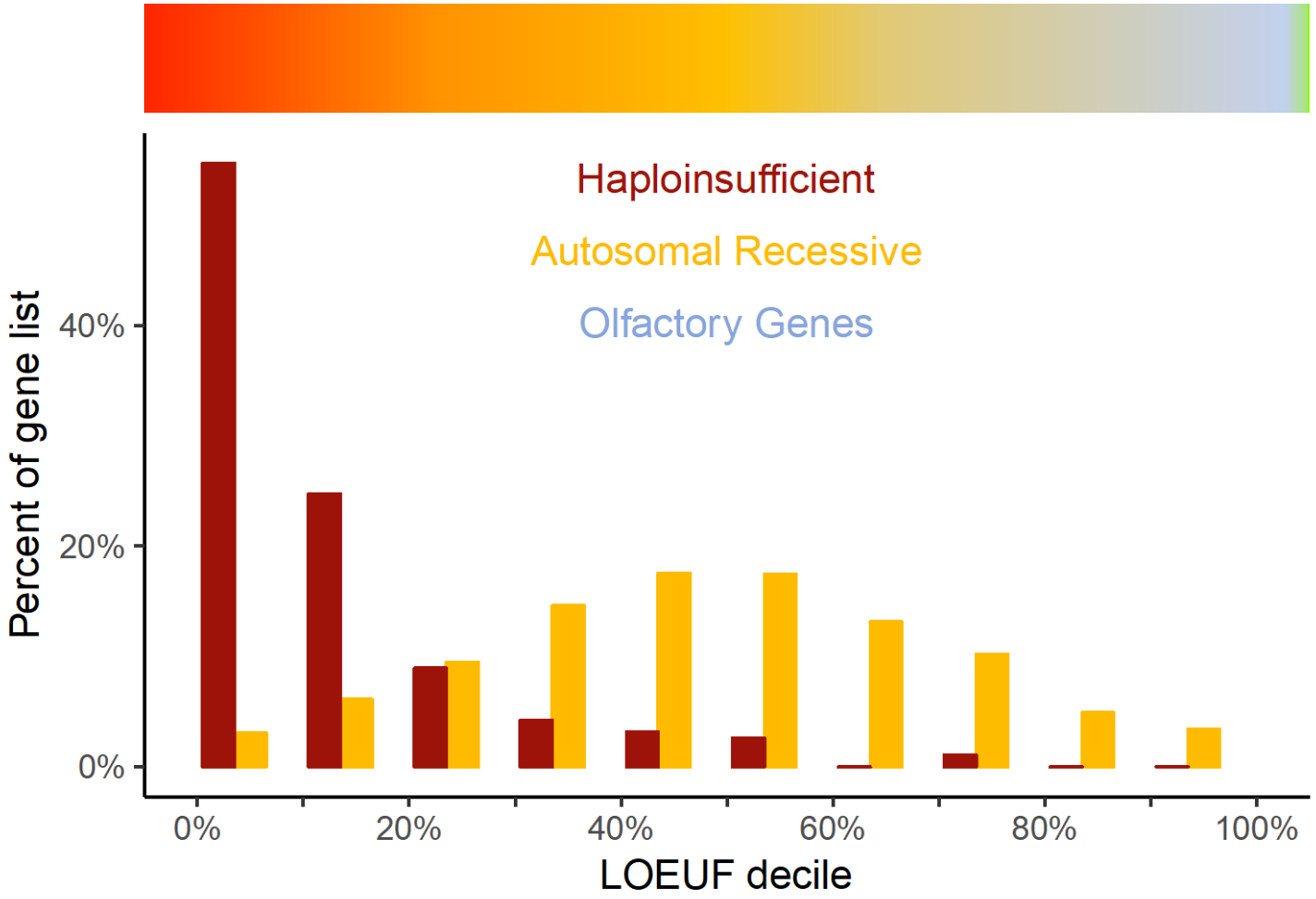
- Known haploinsufficient genes where loss of one copy results in disease



ClinGen gene dosage map

Resolving the spectrum of LoF intolerance

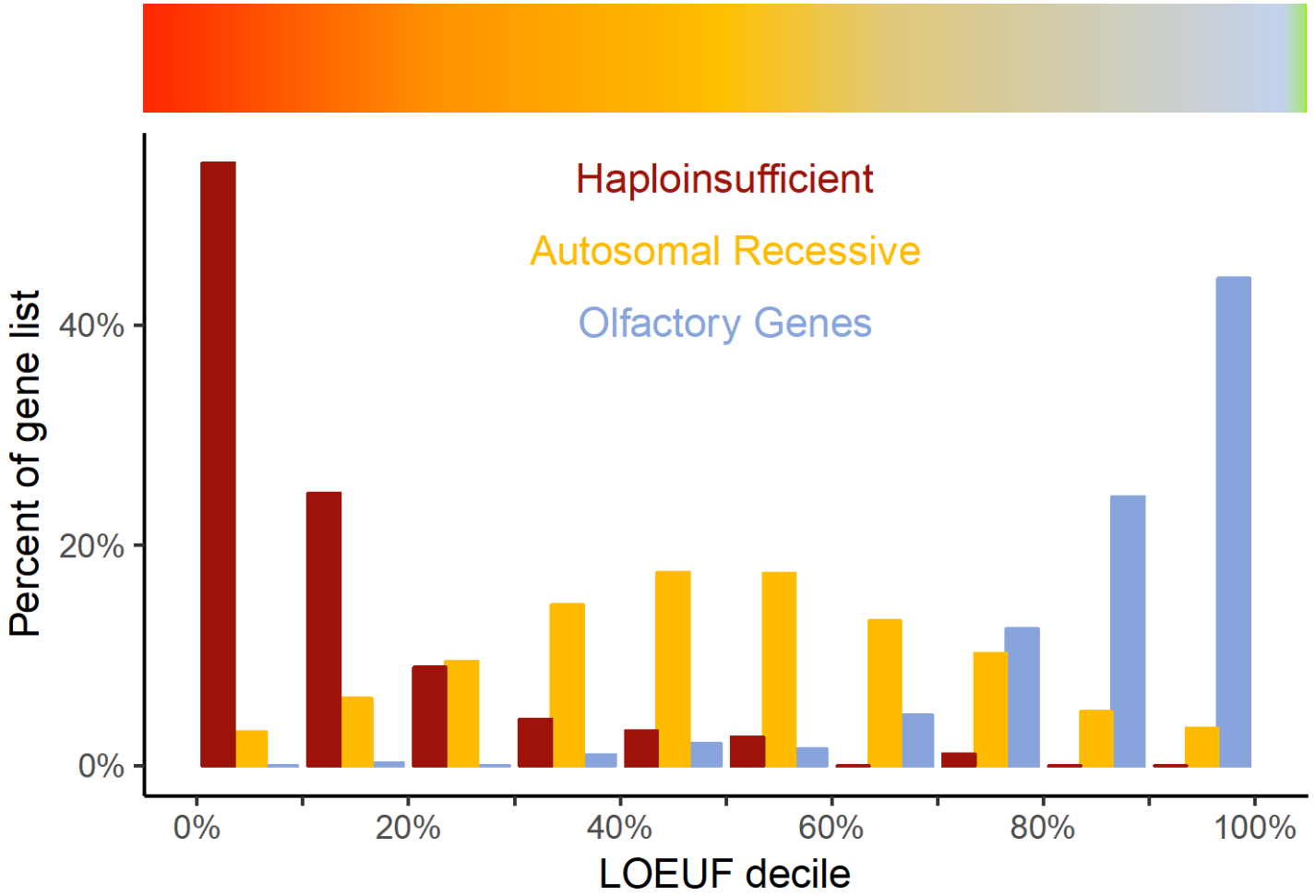
- Autosomal recessive genes are centered around 60% of expected



Gene list from:
Blekhman et al., 2008
Berg et al., 2013

Resolving the spectrum of LoF intolerance

- Some genes, e.g. olfactory receptors, are unconstrained



Important notes on constraint

- **Primarily for dominant disease genes;** do not expect to see strong constraint signal (depletion of pLoF variation) for recessive disease genes
- Not necessarily a sign of a strong phenotype – can see evidence of constraint from mild degrees of negative selection
- Selection occurs through reproduction so deleterious effect must be before/during reproductive years
  - Example: *BRCA1* does not show evidence of loss of function constraint

BRCA1

Autosomal dominant

Breast and ovarian cancer

Category	Exp. SNVs	Obs. SNVs	Constraint metrics
Synonymous	341	329	Z = 0.51 o/e = 0.96 (0.88 - 1.06)
Missense	941.3	891	Z = 0.58 o/e = 0.95 (0.9 - 1)
pLoF	75.2	55	pLI = 0 o/e = 0.73 (0.59 - 0.92)

NSD1 nuclear receptor binding SET domain protein 1

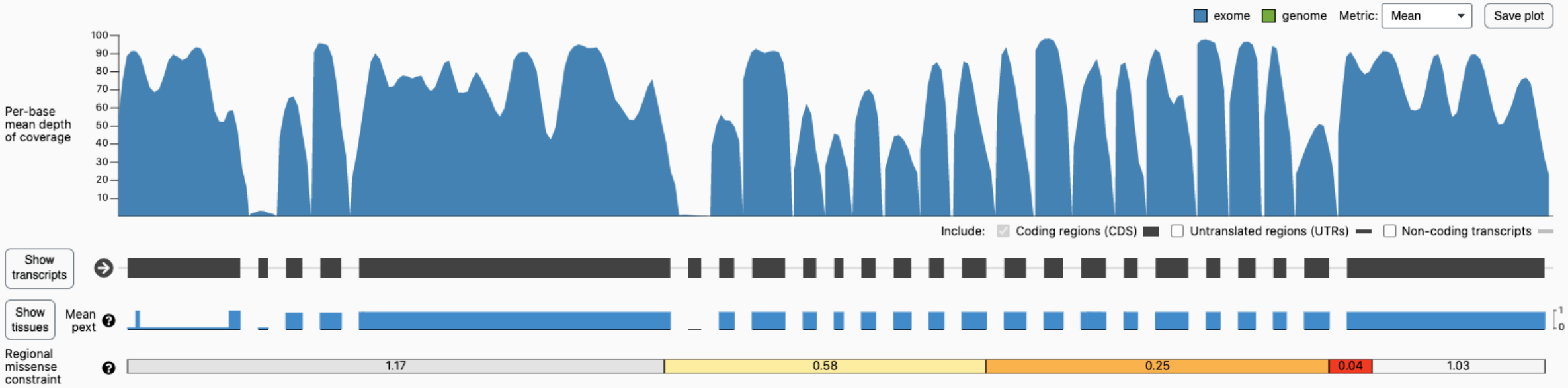
Dataset ExAC v1.0 gnomAD SVs v2.1

Regional missense constraint (Samocha et al.)
https://www.biorxiv.org/content/early/2017/06/12/148353

more

Constraint

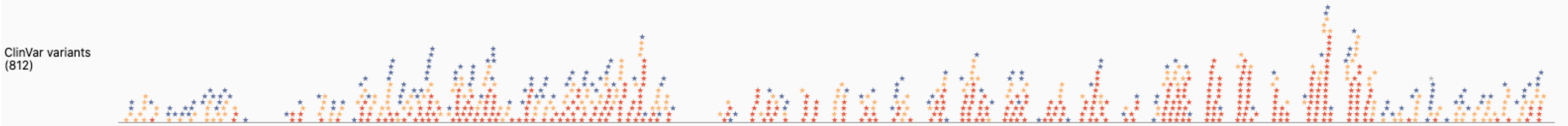
Category	Exp. SNVs	Obs. SNVs	Constraint metric
Synonymous	323.8	345	Z = -0.73
Missense	833.7	693	Z = 2.38
pLoF	74.8	4	pLI = 1



ClinVar variants

Pathogenic / likely pathogenic only | Uncertain significance / conflicting only | Benign / likely benign only | Other only

Collapse to bins



gnomAD variants

NSD1 nuclear receptor binding SET domain protein 1

Dataset ExAC v1.0 | gnomAD SVs v2.1

Genome build GRCh37 / hg19

Ensembl gene ID ENSG00000165671.14

Ensembl canonical transcript ENST00000439151.2

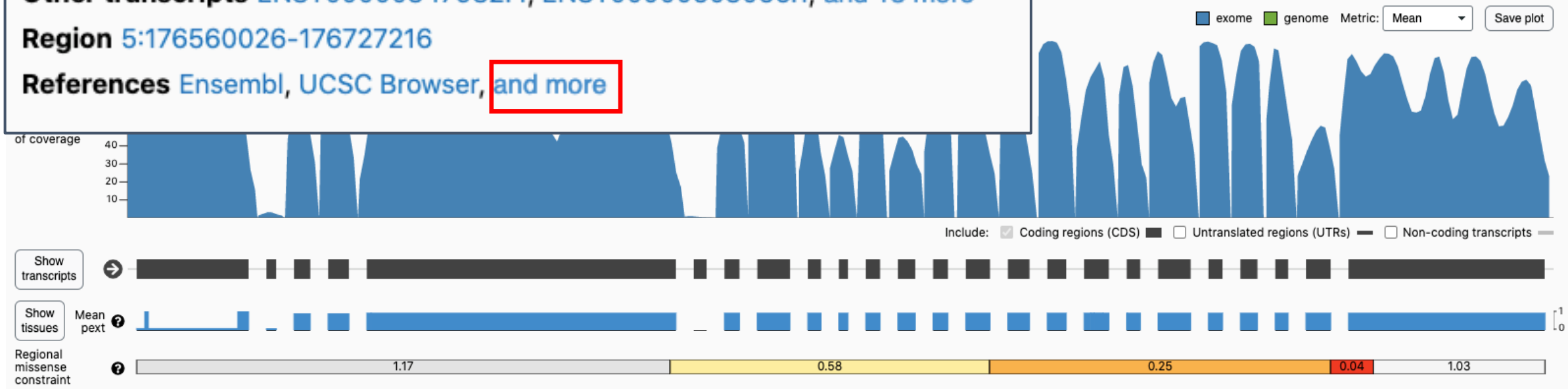
Other transcripts ENST00000347982.4, ENST00000503056.1, and 13 more

Region 5:176560026-176727216

References Ensembl, UCSC Browser, and more

Constraint

Category	Exp. SNVs	Obs. SNVs	Constraint metric
Synonymous	323.8	345	Z = -0.73
Missense	833.7	693	Z = 2.38
pLoF	74.8	4	pLI = 1



ClinVar variants

gnomAD variants

NSD1 nuclear receptor binding SET domain protein 1

Dataset ExAC v1.0 gnomAD SVs v2.1

Genome build GRCh37 / hg19

Ensembl gene ID ENSG00000165671.14

Ensembl canonical transcript ENST000004391

Other transcripts ENST00000347982.4, ENST0000

Region 5:176560026-176727216

References Ensembl, UCSC Browser, and more

of coverage

40
30
20
10

Show transcripts

Show tissues

Mean pext

Regional missense constraint

1.17

References for NSD1

- Ensembl
- UCSC Browser
- GeneCards
- OMIM
- DECIPHER
- ClinGen
- Gene Curation Coalition (GenCC)
- HGNC

Ok

ClinVar variants

Pathogenic / likely pathogenic only | Uncertain significance / conflicting only | Benign / likely benign only | Other only

Collapse to bins

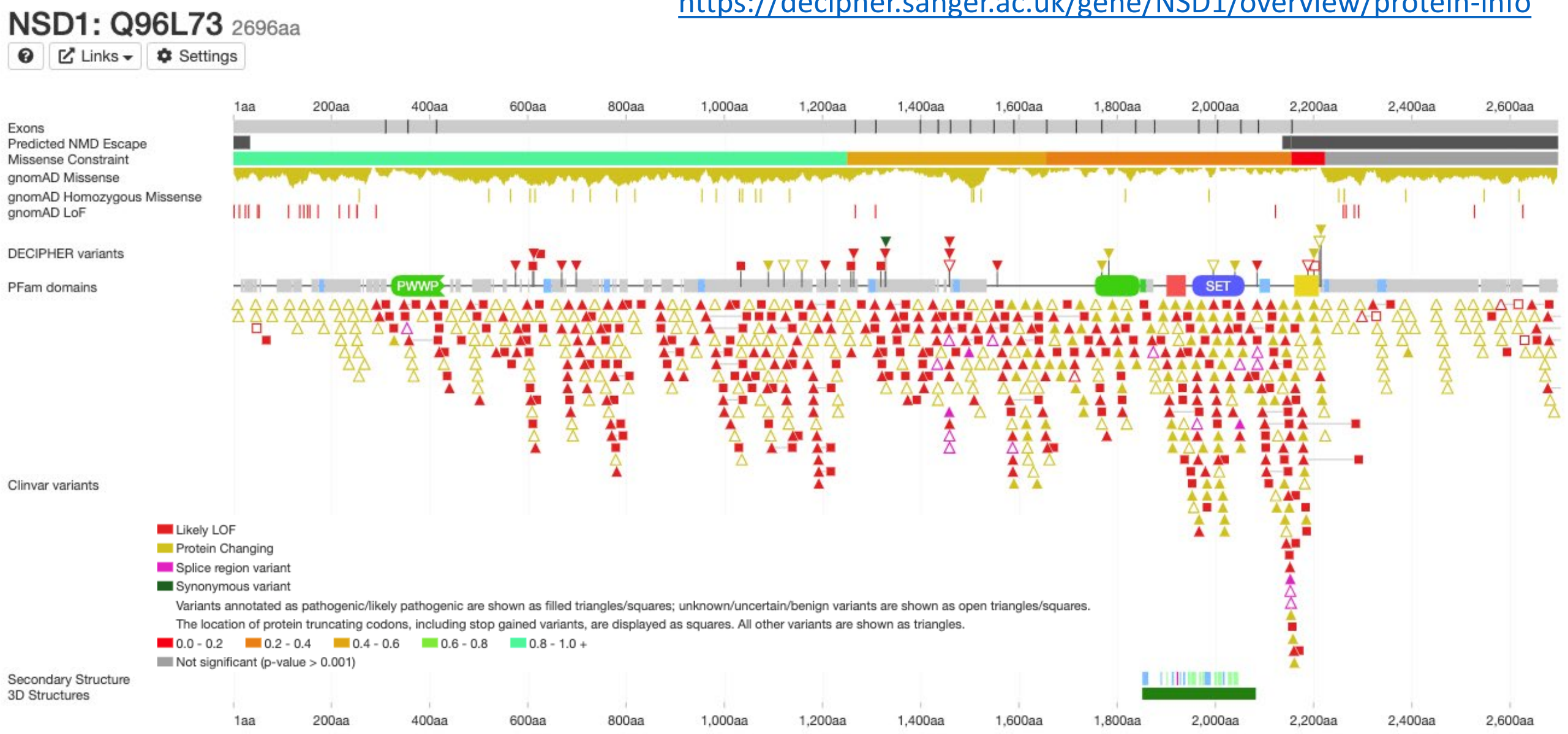
ClinVar variants (812)

gnomAD variants

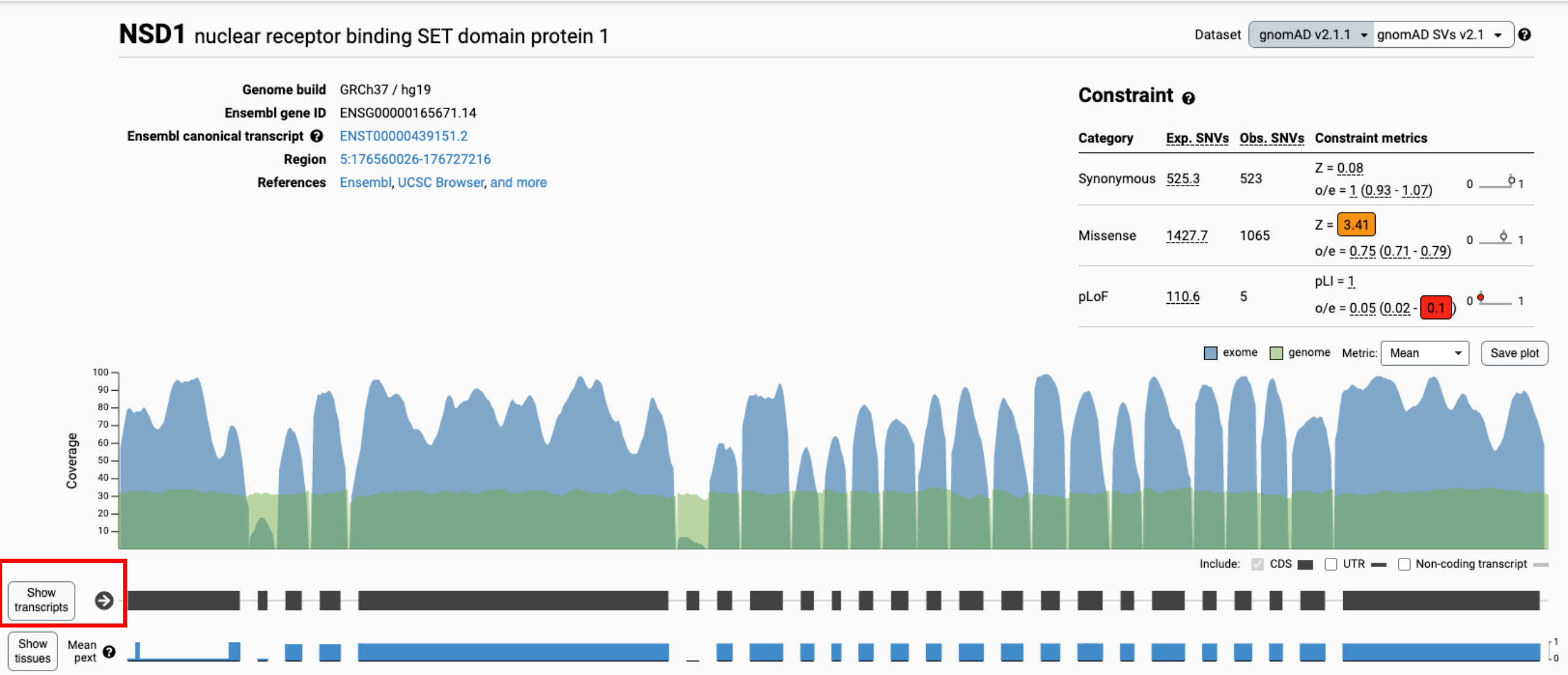
ExAC v1.0 variants

DECIPHER Protein View: *NSD1*

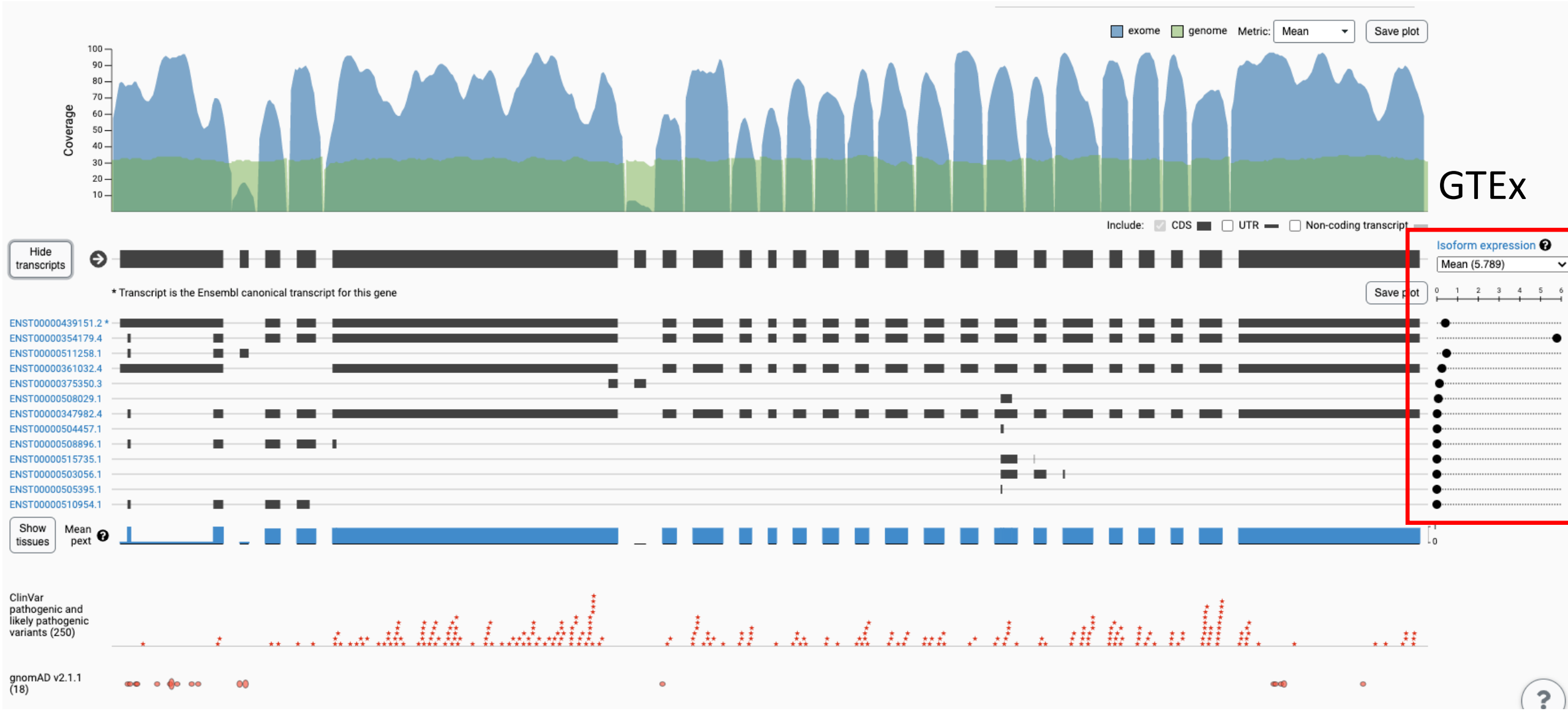
https://decipher.sanger.ac.uk/gene/NSD1/overview/protein-info



Expression data from GTEx in gnomAD

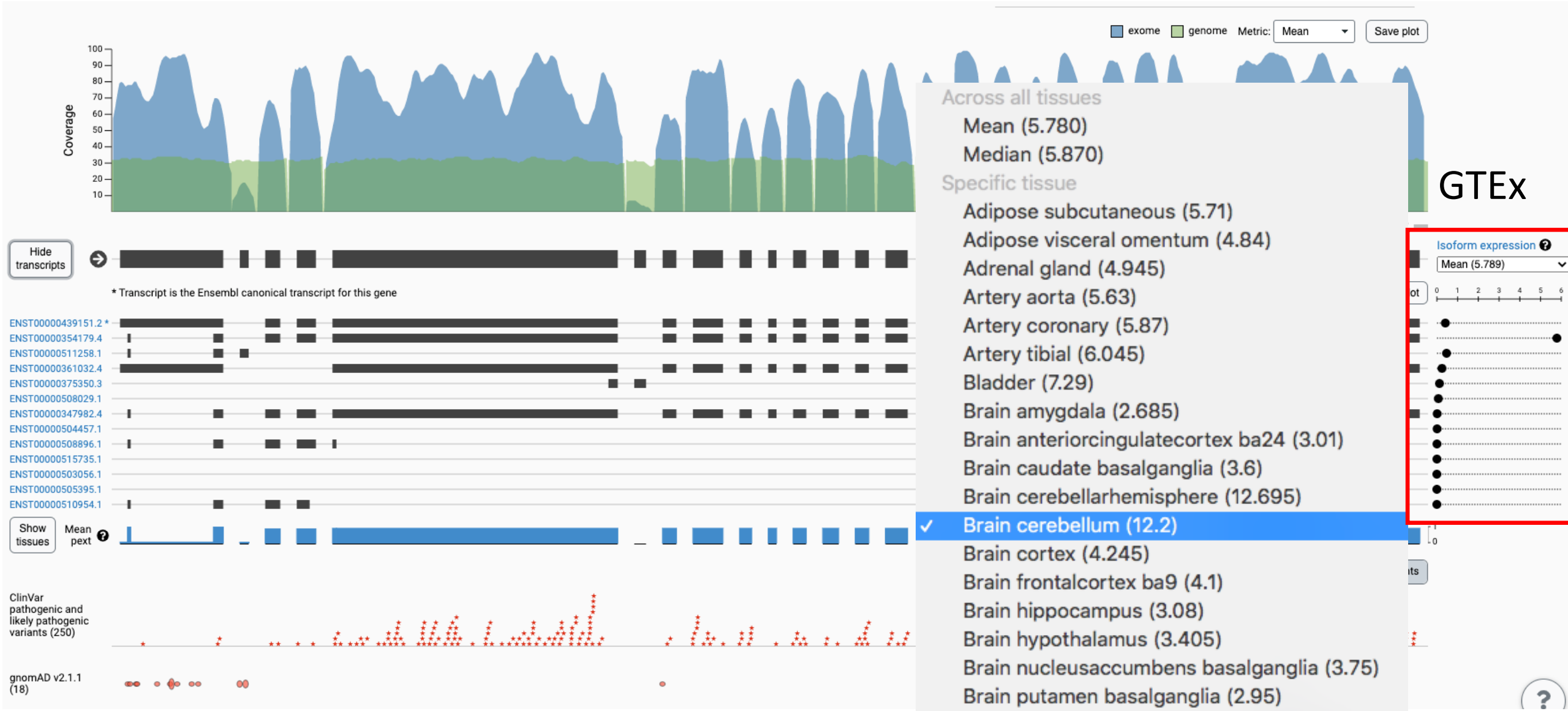


NSD1

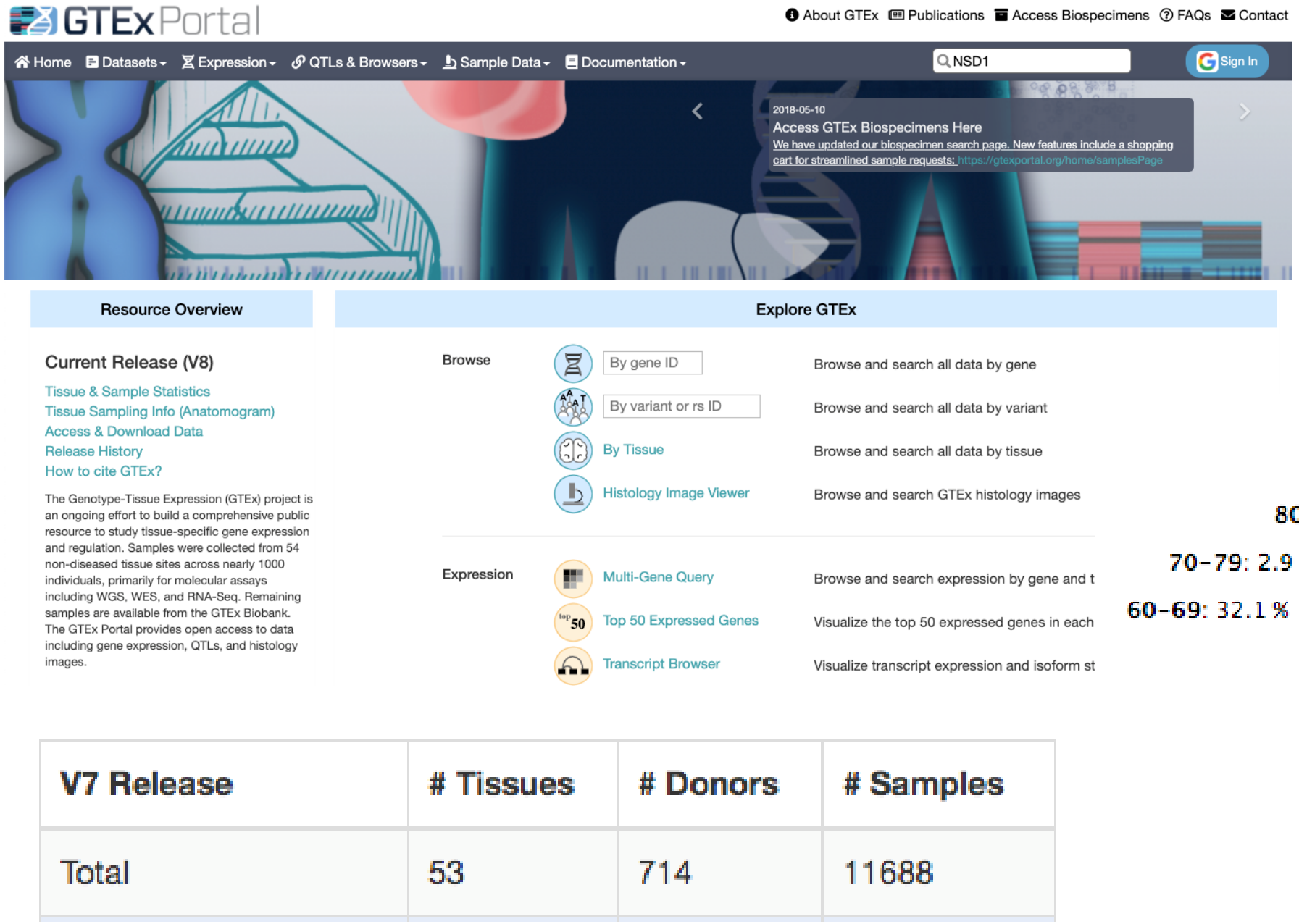


Showing only Pathogenic pLoF variants

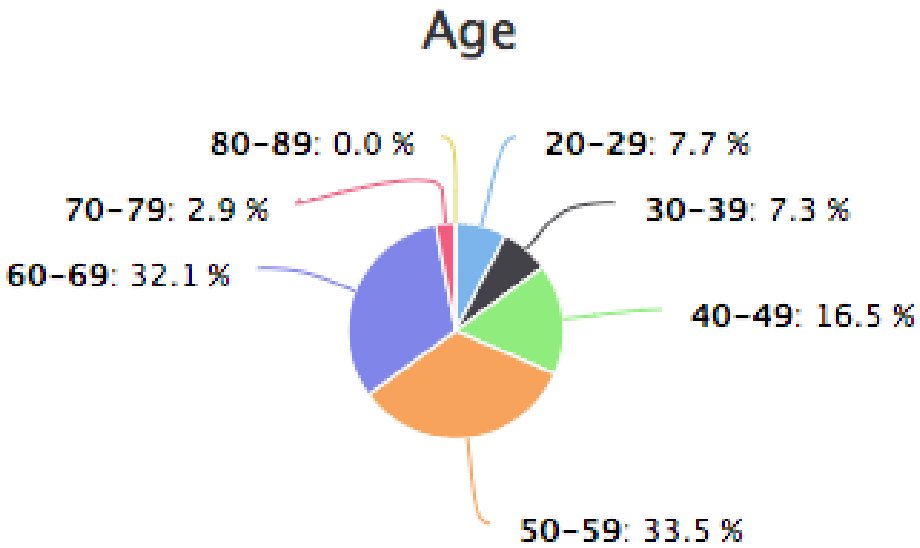
NSD1


exome
genome
Metric:
Mean
Save plot
Coverage
Hide transcripts
* Transcript is the Ensembl canonical transcript for this gene
ENST00000439151.2 *
ENST00000354179.4
ENST00000511258.1
ENST00000361032.4
ENST00000375350.3
ENST00000508029.1
ENST00000347982.4
ENST00000504457.1
ENST00000508896.1
ENST00000515735.1
ENST00000503056.1
ENST00000505395.1
ENST00000510954.1
Show tissues
Mean pext
ClinVar pathogenic and likely pathogenic variants (250)
gnomAD v2.1.1 (18)
Across all tissues
Mean (5.780)
Median (5.870)
Specific tissue
Adipose subcutaneous (5.71)
Adipose visceral omentum (4.84)
Adrenal gland (4.945)
Artery aorta (5.63)
Artery coronary (5.87)
Artery tibial (6.045)
Bladder (7.29)
Brain amygdala (2.685)
Brain anteriorcingulatecortex ba24 (3.01)
Brain caudate basalganglia (3.6)
Brain cerebellarhemisphere (12.695)
Brain cerebellum (12.2)
Brain cortex (4.245)
Brain frontalcortex ba9 (4.1)
Brain hippocampus (3.08)
Brain hypothalamus (3.405)
Brain nucleusaccumbens basalganglia (3.75)
Brain putamen basalganglia (2.95)
Brain spinalcord cervical c1 (3.73)
Isoform expression
Mean (5.789)

GTEx



Gene expression
Transcript expression
eQTLs

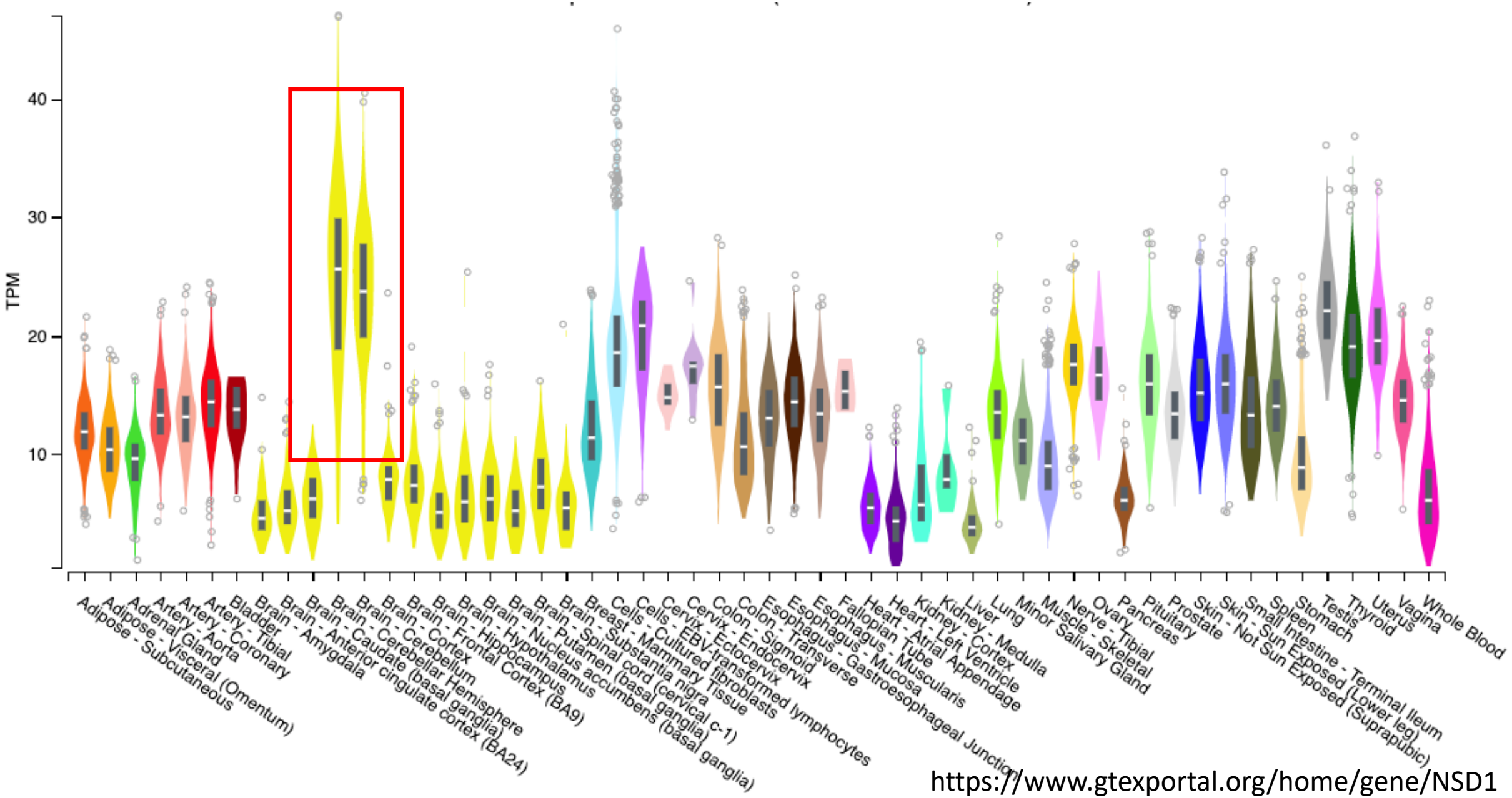


V7 Release	# Tissues	# Donors	# Samples
Total	53	714	11688

GTEx Consortium et al., Nature, 2017

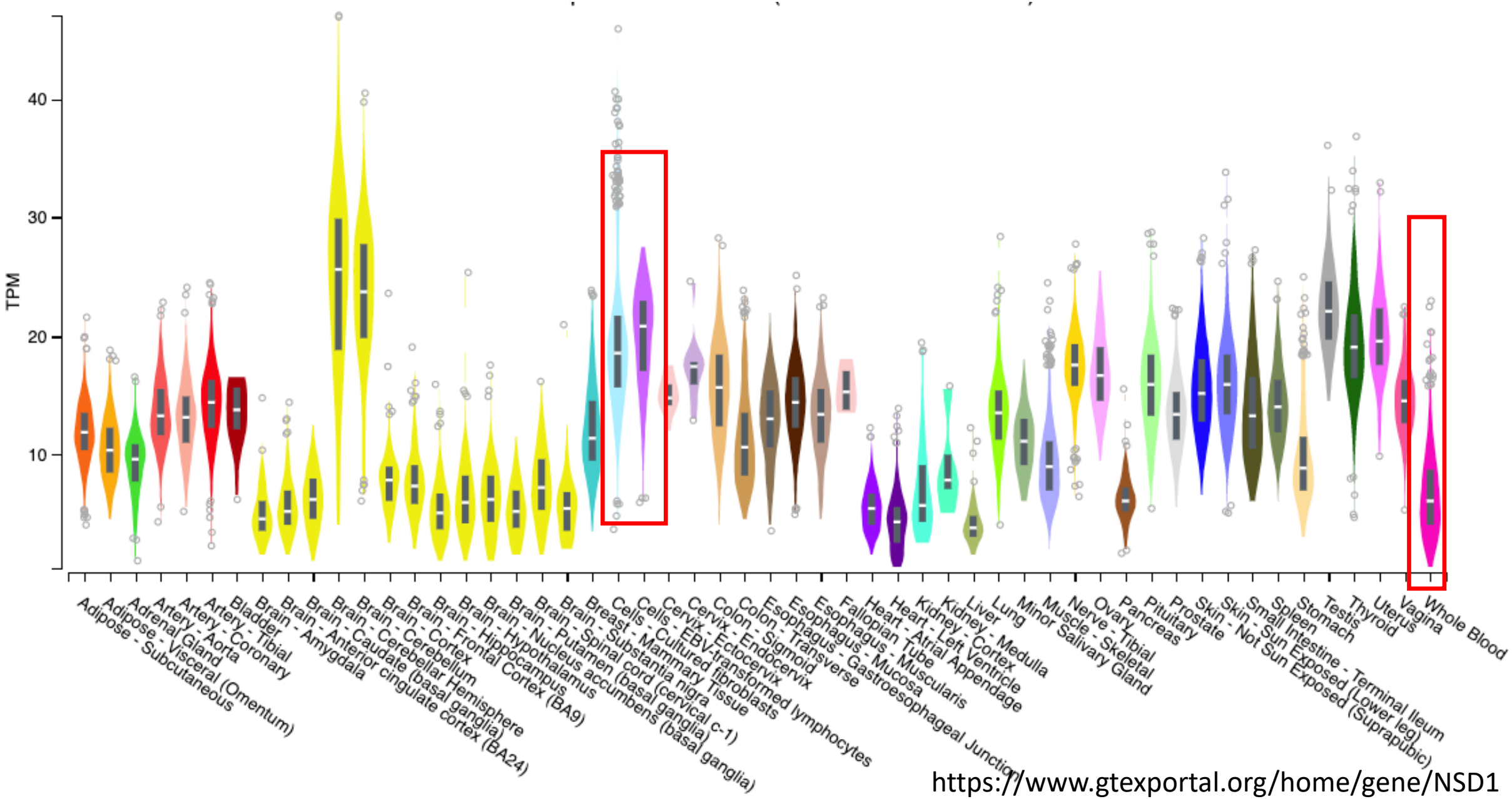
https://www.gtexportal.org/home/

Example gene expression across tissues: *NSD1*



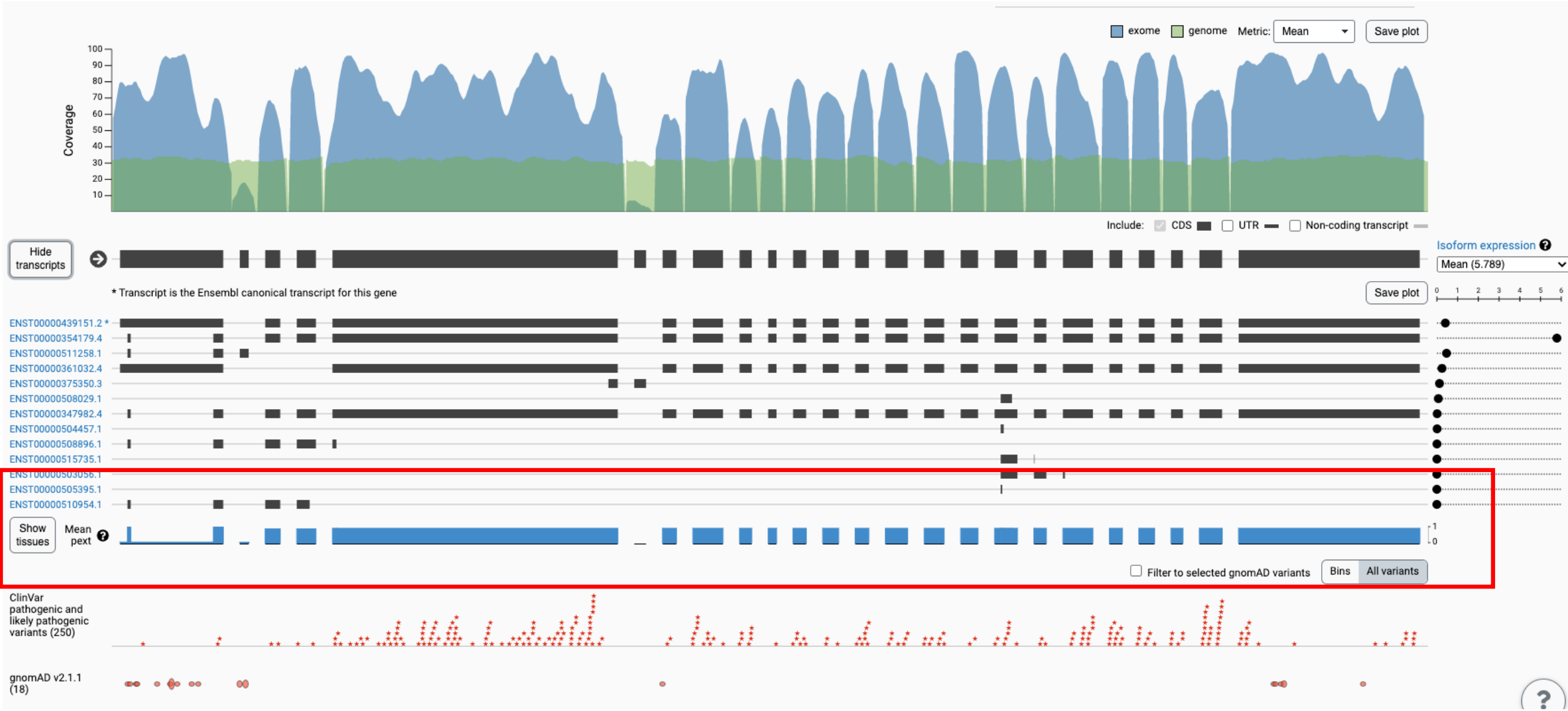
https://www.gtexportal.org/home/gene/NSD1

Example gene expression across tissues: *NSD1*



https://www.gtexportal.org/home/gene/NSD1

Transcript aware expression analysis - pext score



Cummings et al., Nature, 2020

Exploring subsets

gnomAD v2.1.1

- gnomAD v2.1.1 — 141,456 samples
- gnomAD v2.1.1 (non-TOPMed) — 135,743 samples
- gnomAD v2.1.1 (non-cancer) — 134,187 samples
- gnomAD v2.1.1 (non-neuro) — 114,704 samples
- gnomAD v2.1.1 (controls) — 60,146 samples
- ExAC v1.0 — 60,706 samples

Single nucleotide variant: 5-176631141-C-T (

	Exomes	Genomes	Total
Filter	Pass	No variant	
Allele Count	1		1
Allele Number	251484		251484
Allele Frequency	0.000003976		0.000003976
Popmax Filtering AF (95% confidence)	—		
Number of homozygotes	0		0

missense

- NSD1
  - ENST00000439151.2

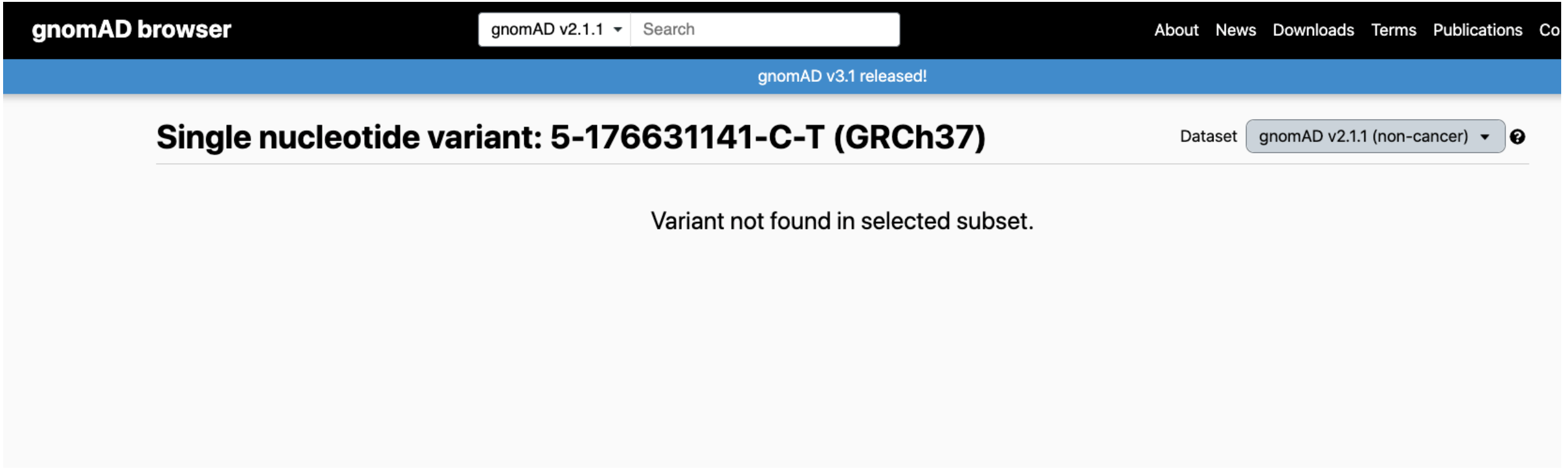
    Ensembl canonical transcript for

    HGVSp: p.Arg362Trp

Population Frequencies

Population	Allele Count	Allele Number	Number of Homozygotes	Allele Frequency
African	1	16256	0	0.00006152
Latino	0	34590	0	0.000
Ashkenazi Jewish	0	10080	0	0.000
East Asian	0	18394	0	0.000
European (Finnish)	0	21648	0	0.000

Exploring subsets

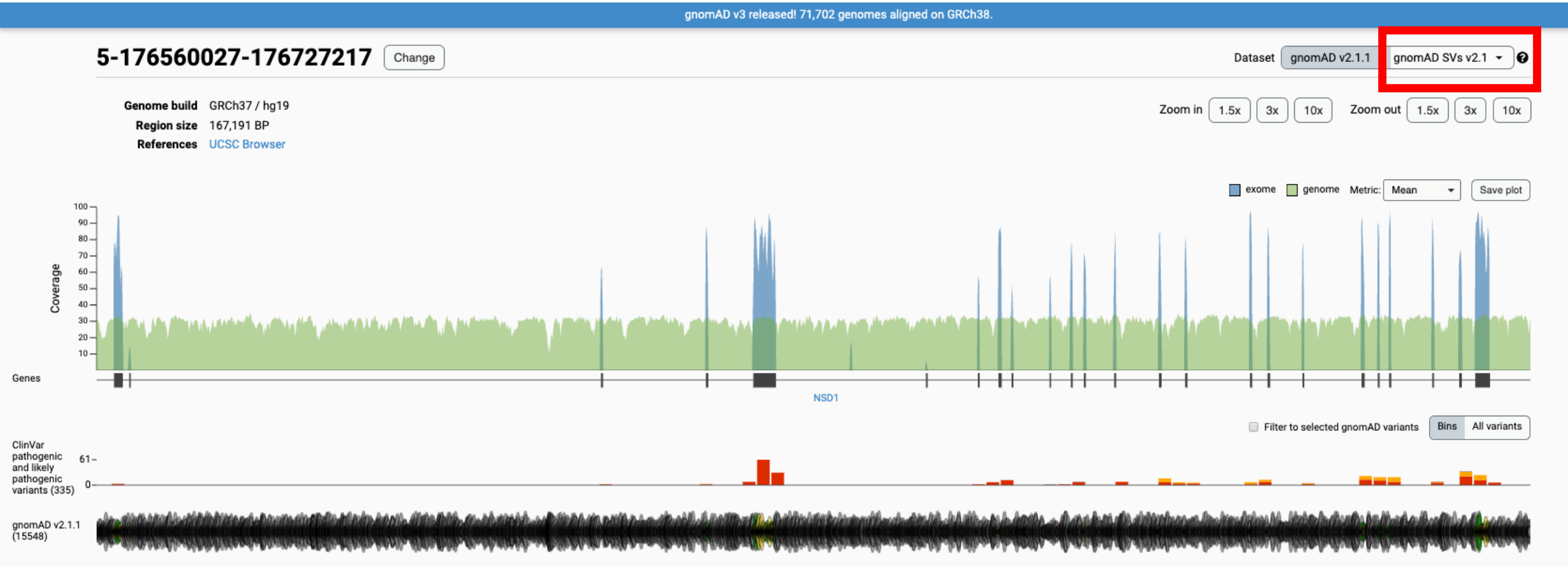


Cancer means germline sample (typically blood) for individual recruited to a study for a non-blood cancer

Gene page region view

NSD1 nuclear receptor binding SET domain protein 1

Genome build	GRCh37 / hg19
Ensembl gene ID	ENSG00000165671
Canonical transcript ID	ENST00000439151
Region	5:176560027-176727217
References	Ensembl, UCSC Browser, and more



Article | Open Access | Published: 27 May 2020

A structural variation reference for medical and population genetics

Ryan L. Collins, Harrison Brand, [...] Michael E. Talkowski

Nature **581**, 444–451(2020) | Cite this article

29k Accesses | **38** Citations | **168** Altmetric | Metrics


Ryan Collins

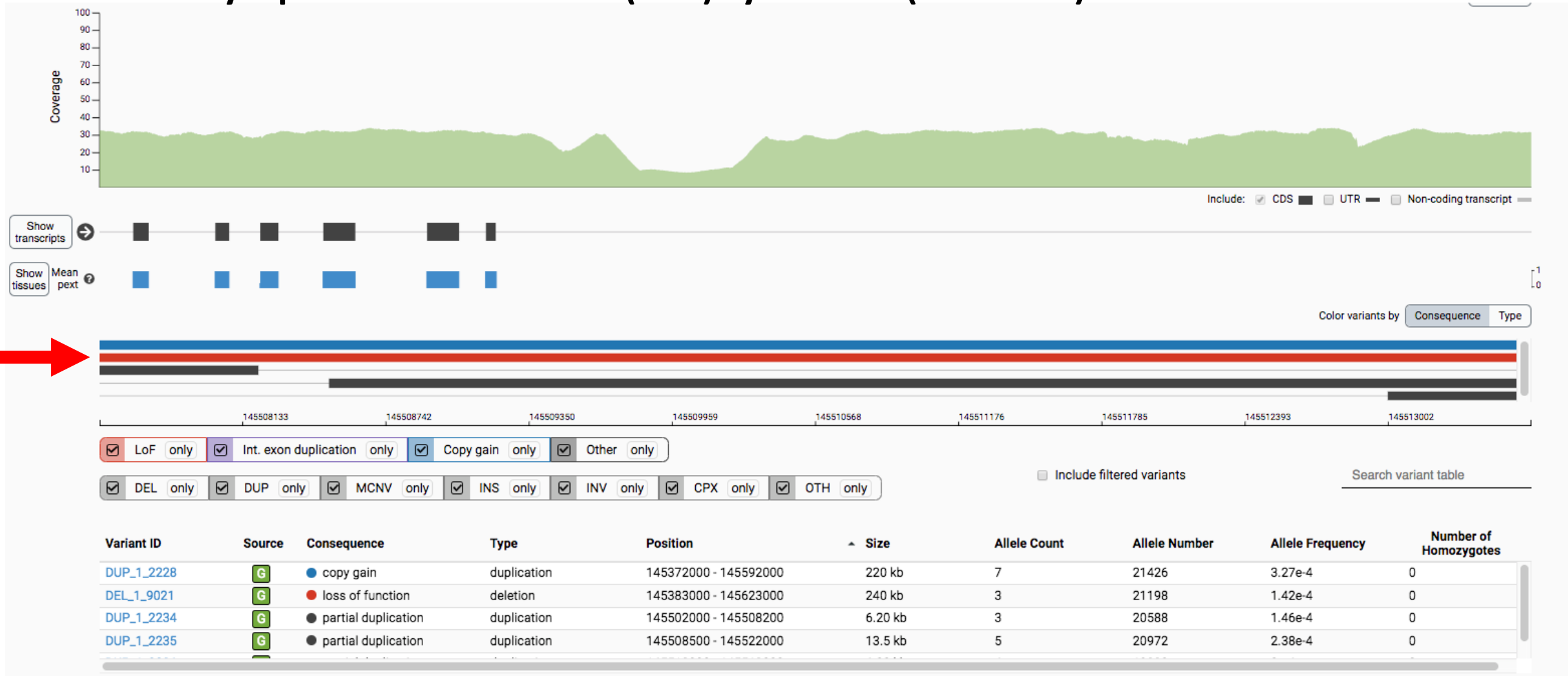

Harrison Brand


Mike Talkowski

- gnomAD-SV callset from 10,847 whole genomes
- Multiple algorithms followed by integration step
- 433,371 unique structural variants with frequency data

gnomAD Structural Variants: *RBM8A*

Thrombocytopenia absent radius (TAR) syndrome (recessive)

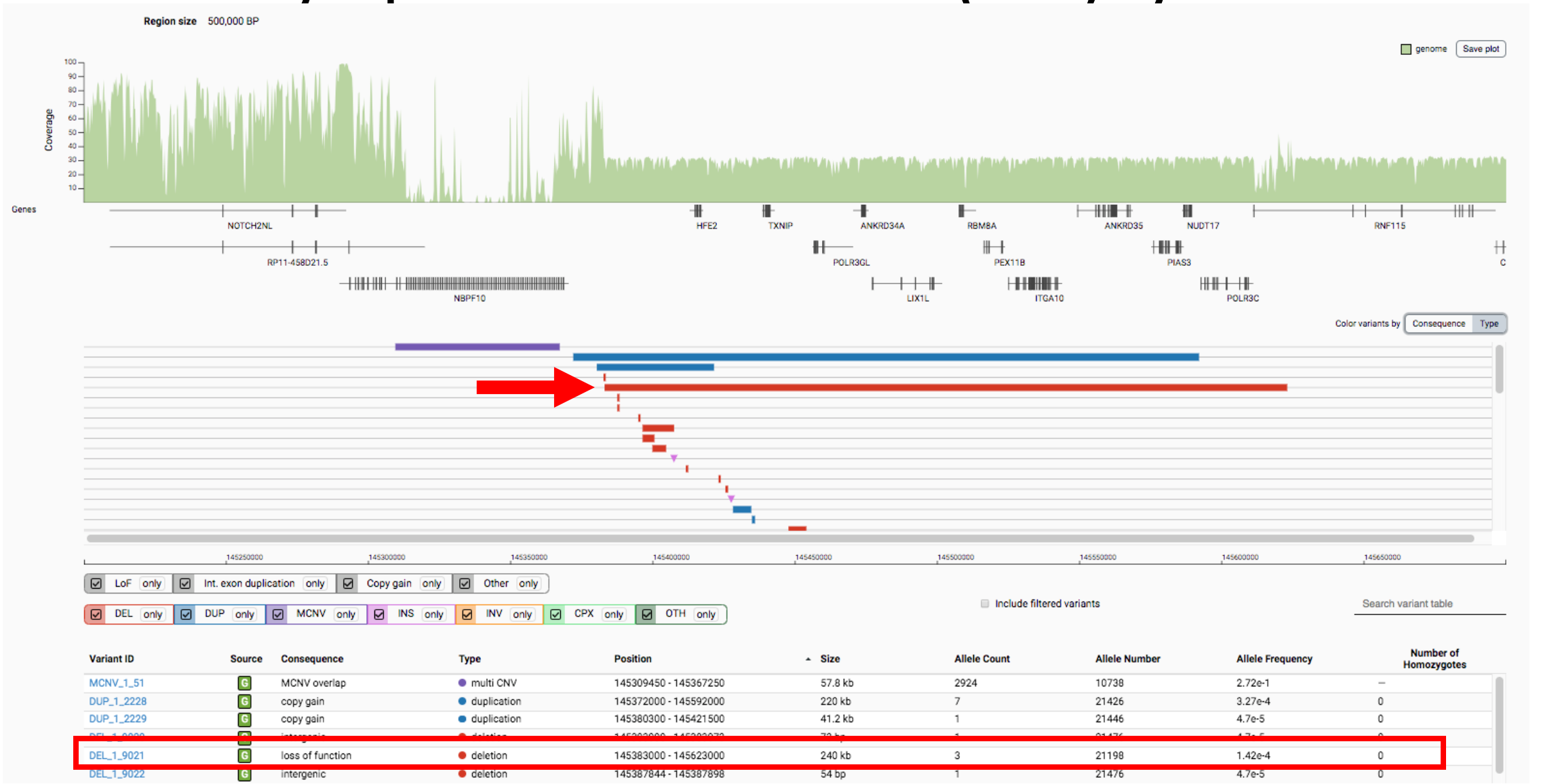


Variant ID	Source	Consequence	Type	Position	Size	Allele Count	Allele Number	Allele Frequency	Number of Homozygotes
DUP_1_2228	G	copy gain	duplication	145372000 - 145592000	220 kb	7	21426	3.27e-4	0
DEL_1_9021	G	loss of function	deletion	145383000 - 145623000	240 kb	3	21198	1.42e-4	0
DUP_1_2234	G	partial duplication	duplication	145502000 - 145508200	6.20 kb	3	20588	1.46e-4	0
DUP_1_2235	G	partial duplication	duplication	145508500 - 145522000	13.5 kb	5	20972	2.38e-4	0

Collins et al., Nature, 2020

gnomad.broadinstitute.org

Previously reported “common” deletion in Thrombocytopenia absent radius (TAR) syndrome



Variant ID	Source	Consequence	Type	Position	Size	Allele Count	Allele Number	Allele Frequency	Number of Homozygotes
MCNV_1_51	G	MCNV overlap	multi CNV	145309450 - 145367250	57.8 kb	2924	10738	2.72e-1	–
DUP_1_2228	G	copy gain	duplication	145372000 - 145592000	220 kb	7	21426	3.27e-4	0
DUP_1_2229	G	copy gain	duplication	145380300 - 145421500	41.2 kb	1	21446	4.7e-5	0
DEL_1_9021	G	loss of function	deletion	145383000 - 145623000	240 kb	3	21198	1.42e-4	0
DEL_1_9022	G	intergenic	deletion	145387844 - 145387898	54 bp	1	21476	4.7e-5	0

Structural variant: DEL_1_7663

Dataset: gnomAD SVs v2.1

- gnomAD SVs v2.1 — 10,847 samples
- gnomAD SVs v2.1 (controls) — 5,192 samples
- gnomAD SVs v2.1 (non-neuro) — 8,342 samples

Filter	Pass
Allele Count	3
Allele Number	21314
Allele Frequency	0.0001408
Quality score	999
Position	1:145383000-145833000
Size	450,000 bp
Class	deletion
Evidence	Read depth
Algorithms	Depth

References

- UCSC

Report

- Report this variant

Consequences

This variant has consequences in 16 genes.

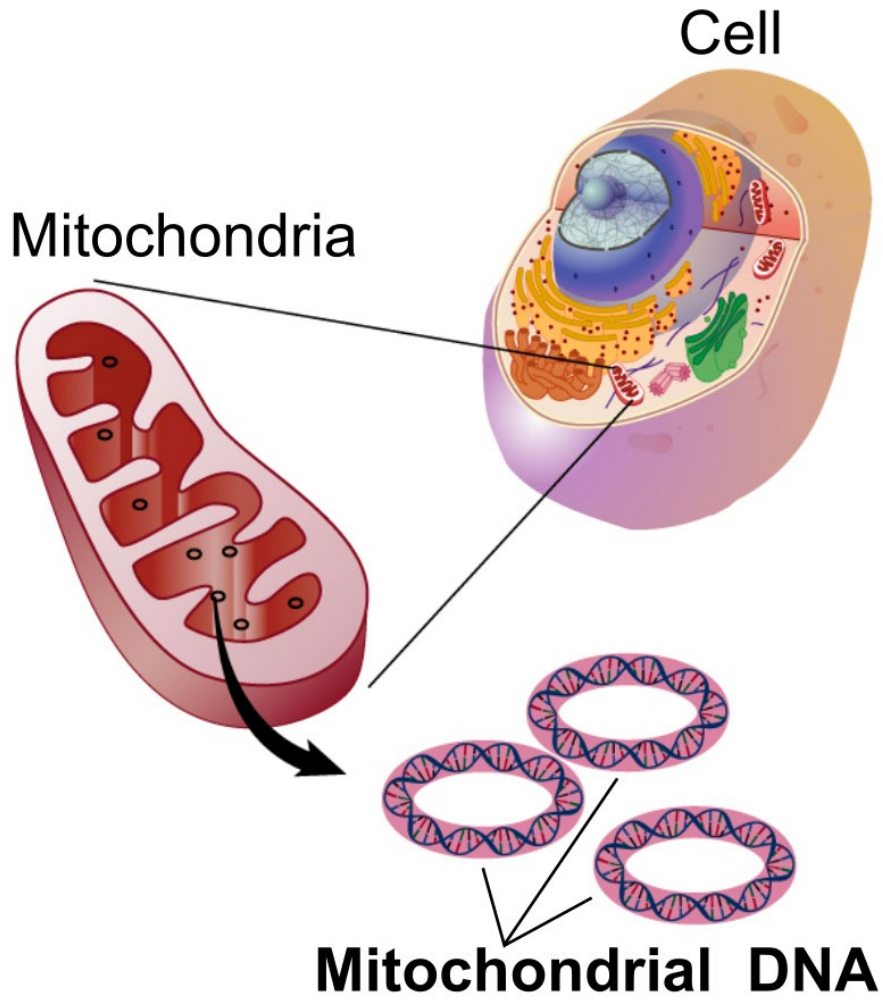
loss of function

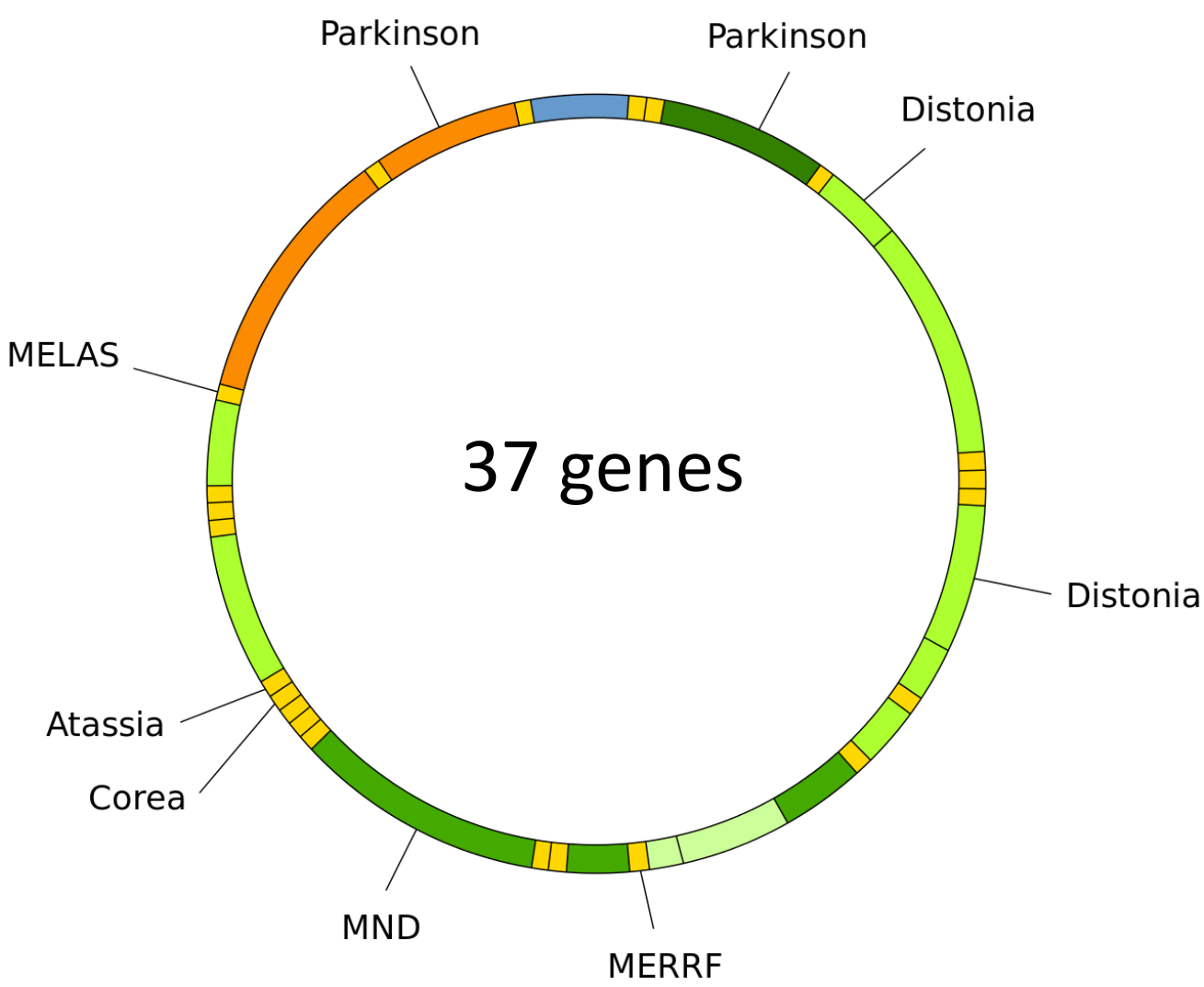
- ANKRD34A
- ANKRD35
- CD160
- and 13 more

Population Frequencies

Population	Allele Count	Allele Number	Number of Homozygotes	Allele Frequency
European	2	7534	0	0.0002655
African	1	9406	0	0.0001063
Latino	0	1928	0	0.000
East Asian	0	2258	0	0.000
Other	0	188	0	0.000
Female	2	10416	0	0.0001920
Male	1	10852	0	0.00009215
Total	**3**	**21314**	**0**	**0.0001408**

Mitochondrial variants in gnomAD v3





Siekevitz P, Scientific American (Wikipedia)

MT-ND5 mitochondrially encoded NADH:ubiquinone oxidoreductase core subunit 5

Dataset gnomAD v3.1

Genome build GRCh38 / hg38
Ensembl gene ID ENSG00000198786.2
MANE Select transcript Not available
Ensembl canonical transcript Not available
Other transcripts ENST00000361567.2
Region M:12337-14148
References Ensembl, UCSC Browser, and more

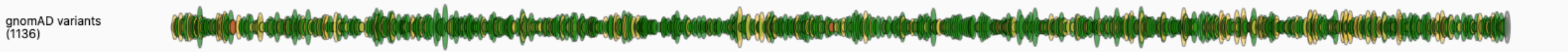
Constraint

Constraint not yet available for gnomAD v3.

ClinVar variants

Pathogenic / likely pathogenic only | Uncertain significance / conflicting only | Benign / likely benign only | Other only | Collapse to bins

gnomAD variants



Variant ID	Source	HGVS Consequence	VEP Annotation	Clinical Significance	Flags	Allele Number	Homoplasmic Allele Count	Homoplasmic Allele Frequency	Heteroplasmic Allele Count	Heteroplasmic Allele Frequency	Max observed heteroplasmy
M-12372-G-A	G	p.Leu12Leu	synonymous	Likely pathogenic	Common Low Heteroplasmy	56353	8901	1.58e-1	8	1.42e-4	1
M-12372-G-C	G	p.Leu12Leu	synonymous			56433	1	1.77e-5	0	0	0.988
M-12373-A-G	G	p.Thr13Ala	missense	Benign		56429	52	9.22e-4	1	1.77e-5	1
M-12375-T-C	G	p.Thr13Thr	synonymous			56432	20	3.54e-4	2	3.54e-5	1

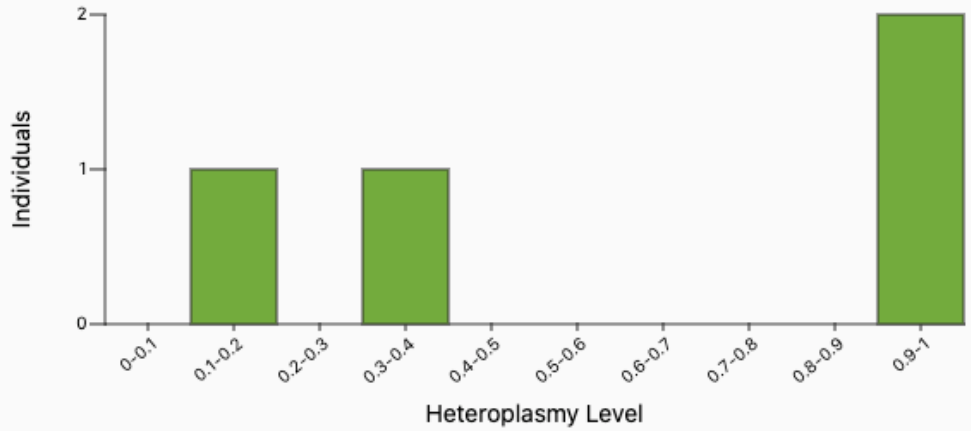
Mitochondrial variant page

Vamsi Mootha
Sarah Calvo
Kristen Laricchia
Monkol Lek
Nicole Lake
Daniel MacArthur

Single nucleotide variant: M-12451-A-G (GRCh38)

Filter Pass
Allele Number 56429
Homoplasmic AC 2
Homoplasmic AF 0.00003544
Heteroplasmic AC 2
Heteroplasmic AF 0.00003544
Max Observed Heteroplasmy 0.9940
Excluded Allele Count 5
Haplogroup Defining Yes

Heteroplasmy Distribution



Annotations

missense

- MT-ND5
  - ENST00000361567.2
    HGVSp: p.Ile39Val
    Polyphen: benign
    SIFT: tolerated_low_confidence

ClinVar

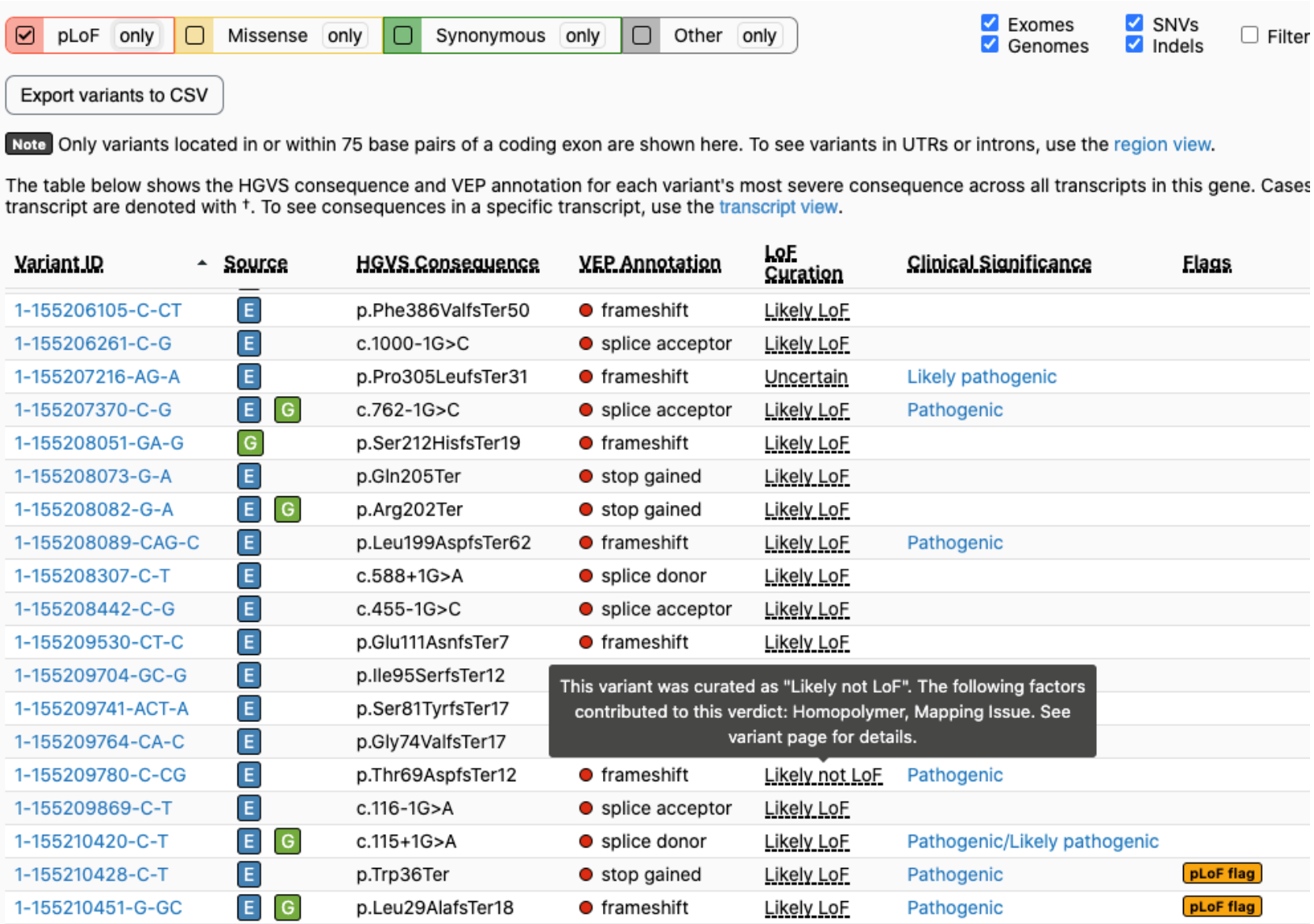
ClinVar Variation ID 693446
Conditions Leigh syndrome
Clinical significance Likely benign
Review status criteria provided, single submitter (1 star)
Last evaluated 2019-10-17

See submission or find more information on the ClinVar website.

Haplogroup Frequencies

Haplogroup	Allele Number	Homoplasmic AC	Homoplasmic AF	Heteroplasmic AC	Heteroplasmic AF
H	14783	0	0.000	1	0.00006765
L2	4724	0	0.000	1	0.0002117
M	1298	1	0.0007704	0	0.000
U	6036	1	0.0001657	0	0.000
Total	**26841**	**2**	**0.00007451**	**2**	**0.00007451**

pLoF curations: predictions not the full story



pLoF only | Missense only | Synonymous only | Other only

Exomes Genomes | SNVs Indels | Filter

Export variants to CSV

Note Only variants located in or within 75 base pairs of a coding exon are shown here. To see variants in UTRs or introns, use the region view.

The table below shows the HGVS consequence and VEP annotation for each variant's most severe consequence across all transcripts in this gene. Cases
transcript are denoted with †. To see consequences in a specific transcript, use the transcript view.

Variant ID	Source	HGVS Consequence	VEP Annotation	LoF Curation	Clinical Significance	Flags
1-155206105-C-CT	E	p.Phe386ValfsTer50	frameshift	Likely LoF		
1-155206261-C-G	E	c.1000-1G>C	splice acceptor	Likely LoF		
1-155207216-AG-A	E	p.Pro305LeufsTer31	frameshift	Uncertain	Likely pathogenic	
1-155207370-C-G	E G	c.762-1G>C	splice acceptor	Likely LoF	Pathogenic	
1-155208051-GA-G	G	p.Ser212HisfsTer19	frameshift	Likely LoF		
1-155208073-G-A	E	p.Gln205Ter	stop gained	Likely LoF		
1-155208082-G-A	E G	p.Arg202Ter	stop gained	Likely LoF		
1-155208089-CAG-C	E	p.Leu199AspfsTer62	frameshift	Likely LoF	Pathogenic	
1-155208307-C-T	E	c.588+1G>A	splice donor	Likely LoF		
1-155208442-C-G	E	c.455-1G>C	splice acceptor	Likely LoF		
1-155209530-CT-C	E	p.Glu111AsnfsTer7	frameshift	Likely LoF		
1-155209704-GC-G	E	p.Ile95SerfsTer12				
1-155209741-ACT-A	E	p.Ser81TyrfsTer17				
1-155209764-CA-C	E	p.Gly74ValfsTer17				
1-155209780-C-CG	E	p.Thr69AspfsTer12	frameshift	Likely not LoF	Pathogenic	
1-155209869-C-T	E	c.116-1G>A	splice acceptor	Likely LoF		
1-155210420-C-T	E G	c.115+1G>A	splice donor	Likely LoF	Pathogenic/Likely pathogenic	
1-155210428-C-T	E	p.Trp36Ter	stop gained	Likely LoF	Pathogenic	pLoF flag
1-155210451-G-GC	E G	p.Leu29AlafsTer18	frameshift	Likely LoF	Pathogenic	pLoF flag

This variant was curated as "Likely not LoF". The following factors contributed to this verdict: Homopolymer, Mapping Issue. See variant page for details.

Manually curated (gnomAD v2)

- All homozygous pLOF variants
  - 28% not LoF
- Heterozygous pLoF variants in some recessive genes
  - 25% not LoF
- Working on pLoF genes in haploinsufficient genes
  - 60% not LoF

LoF Curation column if the gene has variants curated by the gnomAD team

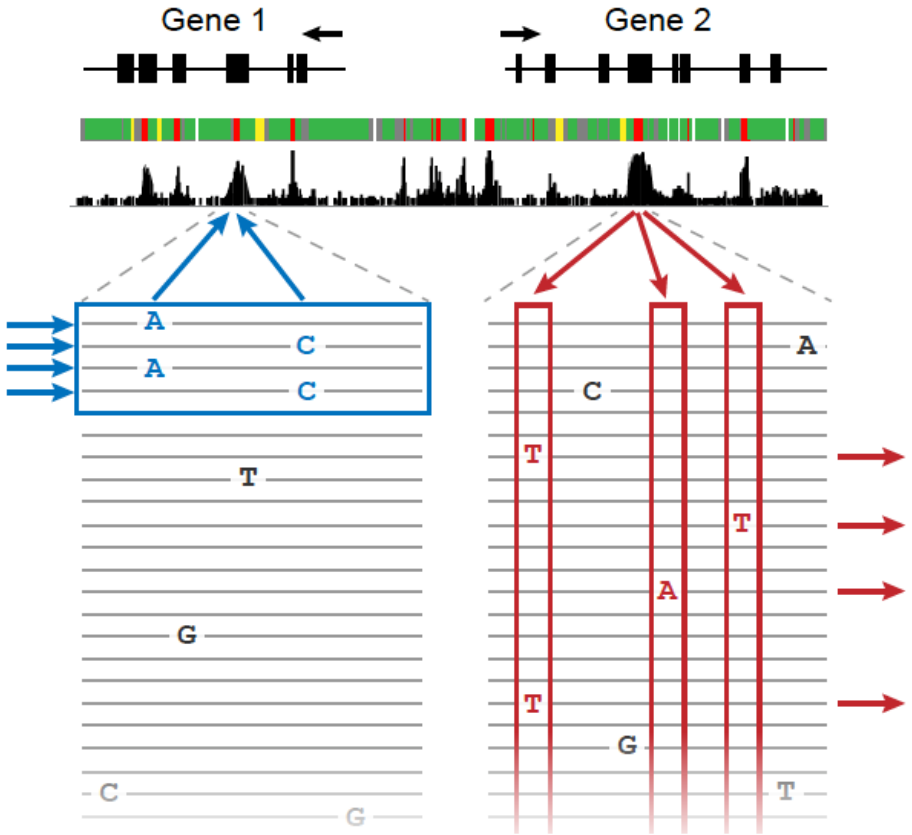
Moriel Singer-Berk

https://gnomad.broadinstitute.org/blog/2020-10-loss-of-function-curations-in-gnomad/

Genetics as a lever to understand biology

Forward genetics

Find the genetic basis of a phenotype or trait



Reverse genetics

Find phenotypes associated with a particular genetic variant

Discovery of novel disease genes

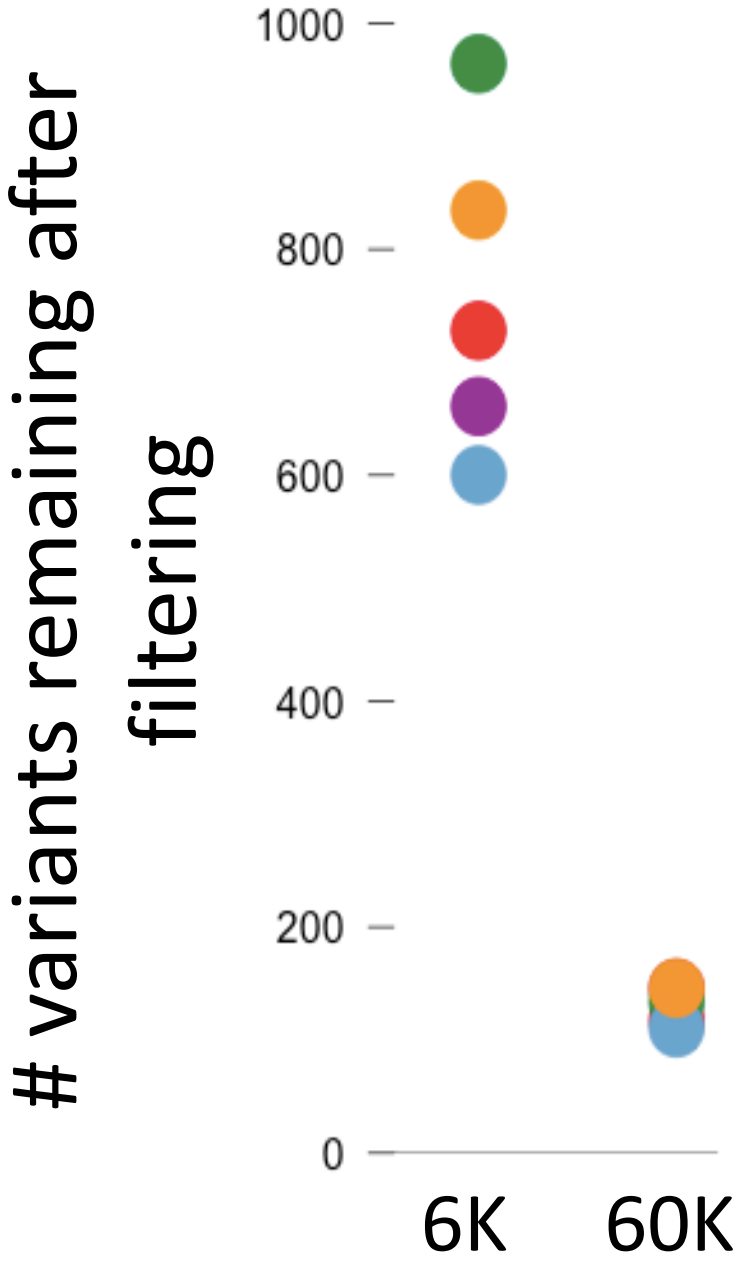
Phenotyping humans with LoF variants

Slide from Daniel MacArthur

Harnessing the power of allele frequency

vs

Making sense of one exome requires **tens of thousands** of exomes (or genomes) to reveal rare variants



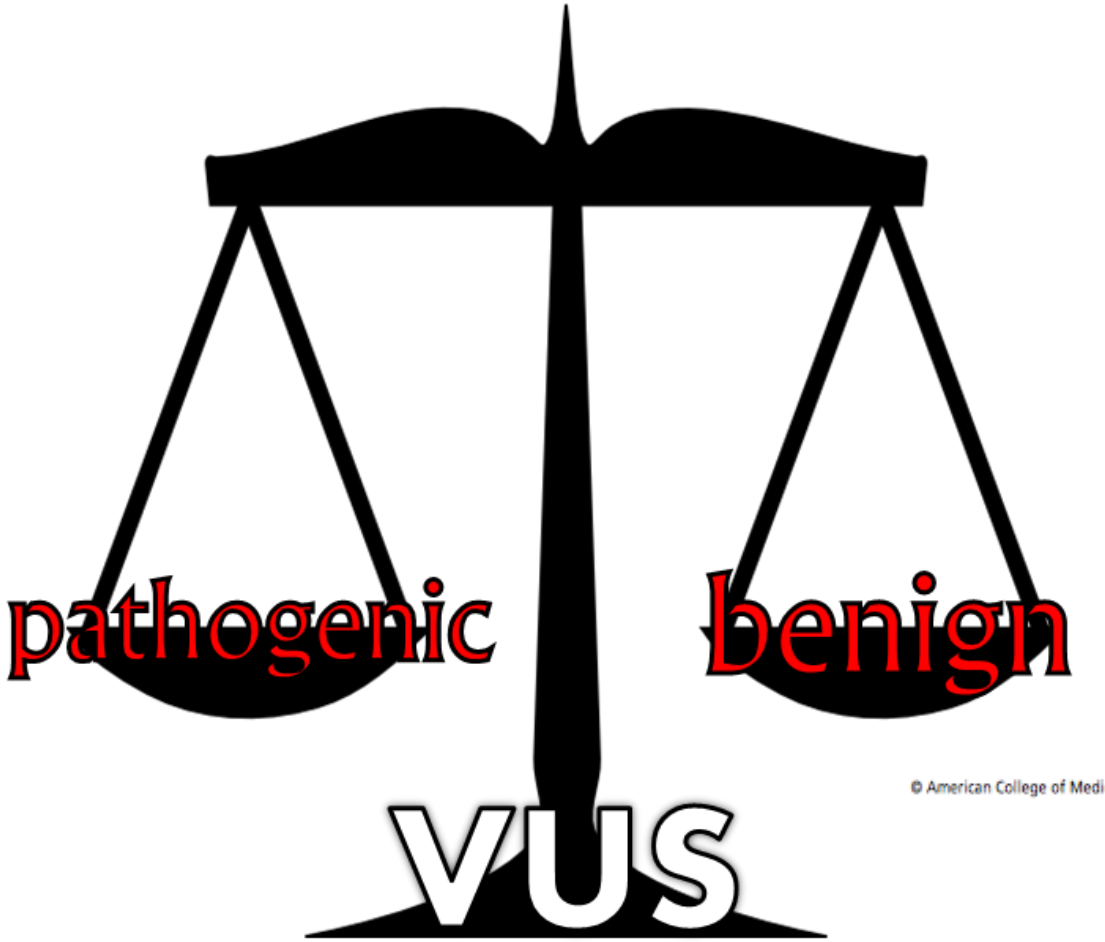
East Asian
South Asian
Latino
African
European

Five-fold reduction in number of very rare variants with large reference databases

variants remaining in an exome after applying a 0.1% filter across all populations

Both **size and ancestral diversity** increase filtering power

Lek et al., *Nature*, 2016



Evaluating rare variant pathogenicity

ACMG STANDARDS AND GUIDELINES | Genetics inMedicine
2015

Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology

Sue Richards, PhD[1], Nazneen Aziz, PhD[2,16], Sherri Bale, PhD[3], David Bick, MD[4], Soma Das, PhD[5], Julie Gastier-Foster, PhD[6,7,8], Wayne W. Grody, MD, PhD[9,10,11], Madhuri Hegde, PhD[12], Elaine Lyon, PhD[13], Elaine Spector, PhD[14], Karl Voelkerding, MD[13] and Heidi L. Rehm, PhD[15]; on behalf of the ACMG Laboratory Quality Assurance Committee

	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		

Richards et al., Genet Med, 2015

Conclusions

- **Reference population databases** are public resources that are critically important to **evaluate variant rarity**, which is necessary but not sufficient for pathogenicity for rare disease
- These datasets provide multiple tools (constraint, pext) to help prioritize variants to understand human biology
- The power of reference population datasets will increase as they grow in size and diversity
  - gnomAD v4 with >300,000 exomes on GRCh38 anticipated in 2021 or 2022

gnomAD browser

Browser | About | News | Downloads | Terms | Publications | Contact | FAQ

Available on

- Google Cloud Public Datasets
- Registry of Open Data on AWS
- Azure Open Datasets

Downloads

gnomAD v2 | gnomAD v2 liftover | gnomAD v3 | ExAC

Summary

- Variants
- Coverage
- Constraint
- Multi-nucleotide variants (MNVs)
- Proportion expressed across transcripts (pext)
- Structural variants
- Loss-of-function curation results
- Linkage disequilibrium
- Resources

Variants

Note Find out what changed in the latest release in the gnomAD v2.1.1 README.

The variant dataset files below contain all subsets (non-neuro, non-cancer, controls-only, and non-TOPMed).

Exomes

- Sites Hail Table
  Show URL for Google / Amazon / Microsoft
  Copy URL for Google / Amazon / Microsoft
- All chromosomes sites VCF
  58.81 GiB, MD5: f034173bf6e57fbb5e8ce680e95134f2
  Download from Google / Amazon / Microsoft
  Download TBI from Google / Amazon / Microsoft
- chr1 sites VCF
  5.77 GiB, MD5: 9817acdf1d9600efb3355e4cb4b7ee1f
  Download from Google / Amazon / Microsoft
  Download TBI from Google / Amazon / Microsoft

Genomes

- Sites Hail Table
  Show URL for Google / Amazon / Microsoft
  Copy URL for Google / Amazon / Microsoft
- All chromosomes sites VCF
  460.93 GiB, MD5: e6eadf5ac7b2821b40f350da6e1279a2
  Download from Google / Amazon / Microsoft
  Download TBI from Google / Amazon / Microsoft
- Exome calling intervals VCF
  9.7 GiB, MD5: e5bd69a0f89468149bc3afca78cd5acc
  Download from Google / Amazon / Microsoft
  Download TBI from Google / Amazon / Microsoft

- Exome and genome data is in separate files
- Exome data is 125K exomes
- Genome data is 15K genomes

gnomAD v3.1 Mitochondrial DNA Variants

November 17, 2020 in Announcements / Releases

Kristen Laricchia, Sarah E. Calvo

Overview

Mitochondrial DNA (mtDNA) variants for gnomAD are now available for the first time! We have called mtDNA variants for 56,434 whole genome samples in the v3.1 release. This initial release includes population frequencies for 10,850 unique mtDNA variants defined at more than half of all mtDNA bases. The vast majority of variant calls (98%) are homoplasmic or near homoplasmic, whereas 2% are heteroplasmic. Variation in mitochondrial genomes contributes to many human diseases and has had unique value in the study of human evolutionary genetics. We hope that the addition of mtDNA to gnomAD will enable researchers to better understand the role of mtDNA variation in both health and disease states.

Previous gnomAD callsets have not included mtDNA variants because their properties do not fit the assumptions that we use with our nuclear variant calling pipeline. These properties include:

Continue reading

gnomAD v3.1

October 29, 2020 in Announcements / Releases

gnomAD Production Team

Latest posts

gnomAD v3.1 Mitochondrial DNA Variants
November 17, 2020

gnomAD v3.1
October 29, 2020

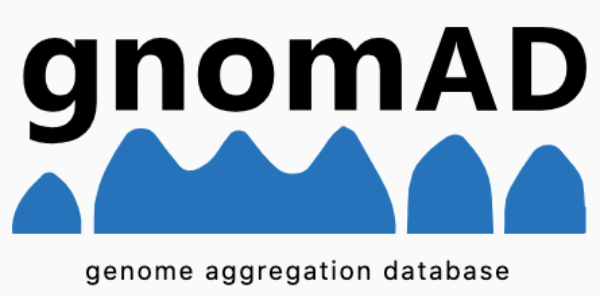
gnomAD v3.1 New Content, Methods, Annotations, and Data Availability
October 29, 2020

Loss-of-Function Curations in gnomAD
October 29, 2020

Open access to gnomAD data on multiple cloud providers
October 29, 2020

Acknowledgements

gnomAD papers at:
https://www.nature.com/collections/afbgiddede



Steering Committee
Heidi Rehm (co-PI)
Mark Daly (co-PI)
Daniel MacArthur
Ben Neale
Mike Talkowski
Anne O'Donnell-Luria
Konrad Karczewski
Grace Tiao
Matt Solomonson
Samantha Baxter

Analysis team
Konrad Karczewski
Laurent Francioli
Grace Tiao
Kristen Laricchia
Beryl Cummings
Eric Minikel
Irina Armean
James Ware
Kaitlin Samocha
Nicola Whiffin
Qingbo Wang
Ryan Collins

Ethics team
Andrea Saltzman
Molly Schleicher
Namrata Gupta
Stacey Donnelly

Production team
Grace Tiao
Julia Goodrich
Katherine Chao
Michael Wilson
William Phu
Eric Banks
Charlotte Tolonen
Christopher Llanwarne
David Roazen
Diane Kaplan
Gordon Wade
Jeff Gentry
Jose Soto
Kathleen Tibbetts
Kristian Cibulskis
Laura Gauthier
Louis Bergelson
Miguel Covarrubias
Nikelle Petrillo
Ruchi Munshi
Sam Novod
Thibault Jeandet
Valentin Ruano-Rubio
Yossi Farjoun

Structural Variation
Ryan Collins
Harrison Brand
Konrad Karczewski
Laurent Francioli
Nick Watts
Matthew Solomonson
Xuefang Zhao
Laura Gauthier
Harold Wang
Chelsea Lowther
Mark Walker
Christopher Whelan
Ted Brookings
Ted Sharpe
Jack Fu
Eric Banks
Michael Talkowski

hail

Ben Neale
Cotton Seed
Tim Poterba
Arcturus Wang
Chris Vittal

Browser team
Matthew Solomonson
Nick Watts
Ben Weisburd

Broad Genomics Platform
Stacey Gabriel
Kristen Connolly
Steven Ferriera

https://gnomad.broadinstitute.org/about
Contact for ?s: gnomad@broadinstitute.org or odonnell@broadinstitute.org