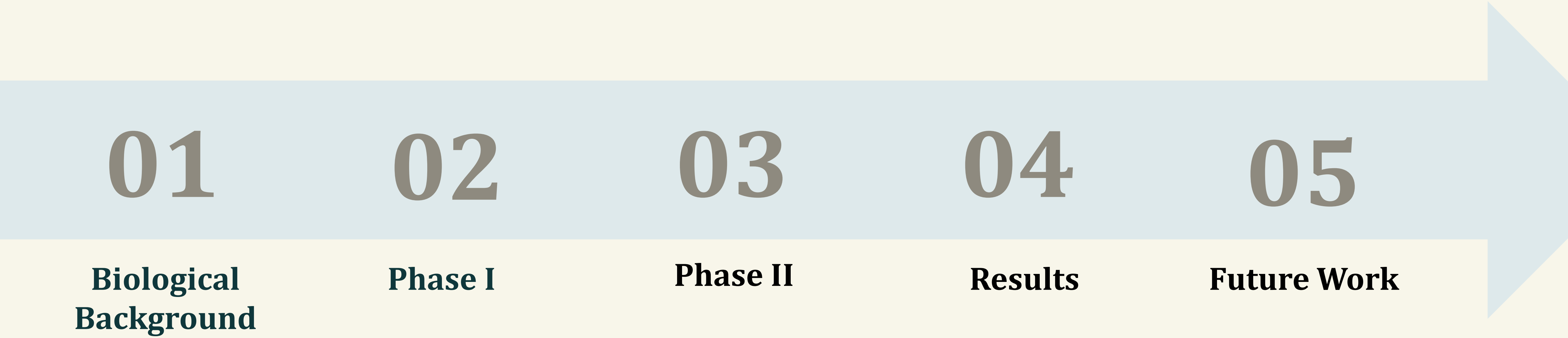
A shortcut to the genetic basis of Neurodegenerative Diseases

Lior Sivan Schapiro

Supervisors: Prof. Michal Linial and Roni Rasnic

The Timeline



01 Biological Background

02 Phase I

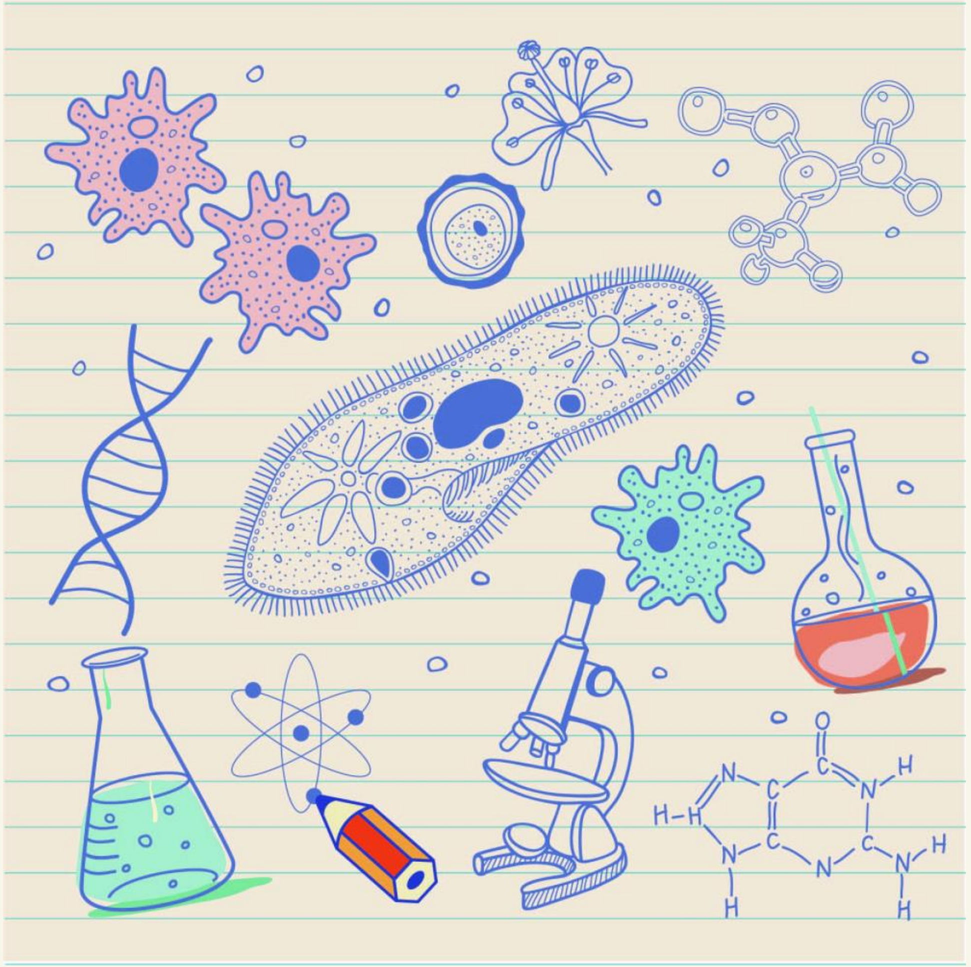
03 Phase II

04 Results

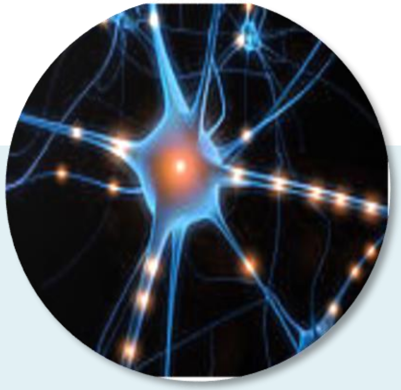
05 Future Work

01

Biological Background



Neurodegenerative Diseases



Cells in the brain lose function over time and ultimately die

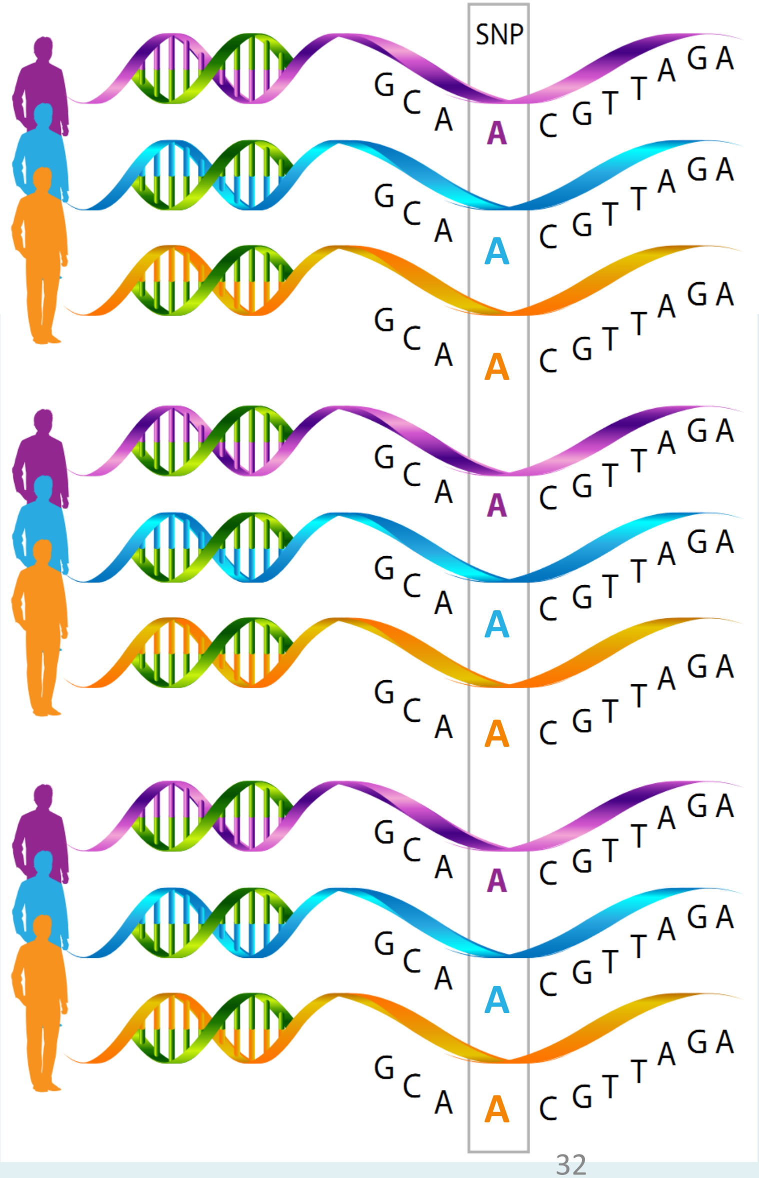
Examples- Alzheimer’s disease, Parkinson’s disease (PD), ALS

The risk of being affected increases dramatically with age

Constraint

Rank of genic functional intolerance to mutation variants using evolutionary conservation of protein sequences within species.

More than 150,000 human genomes (exomes) where tested

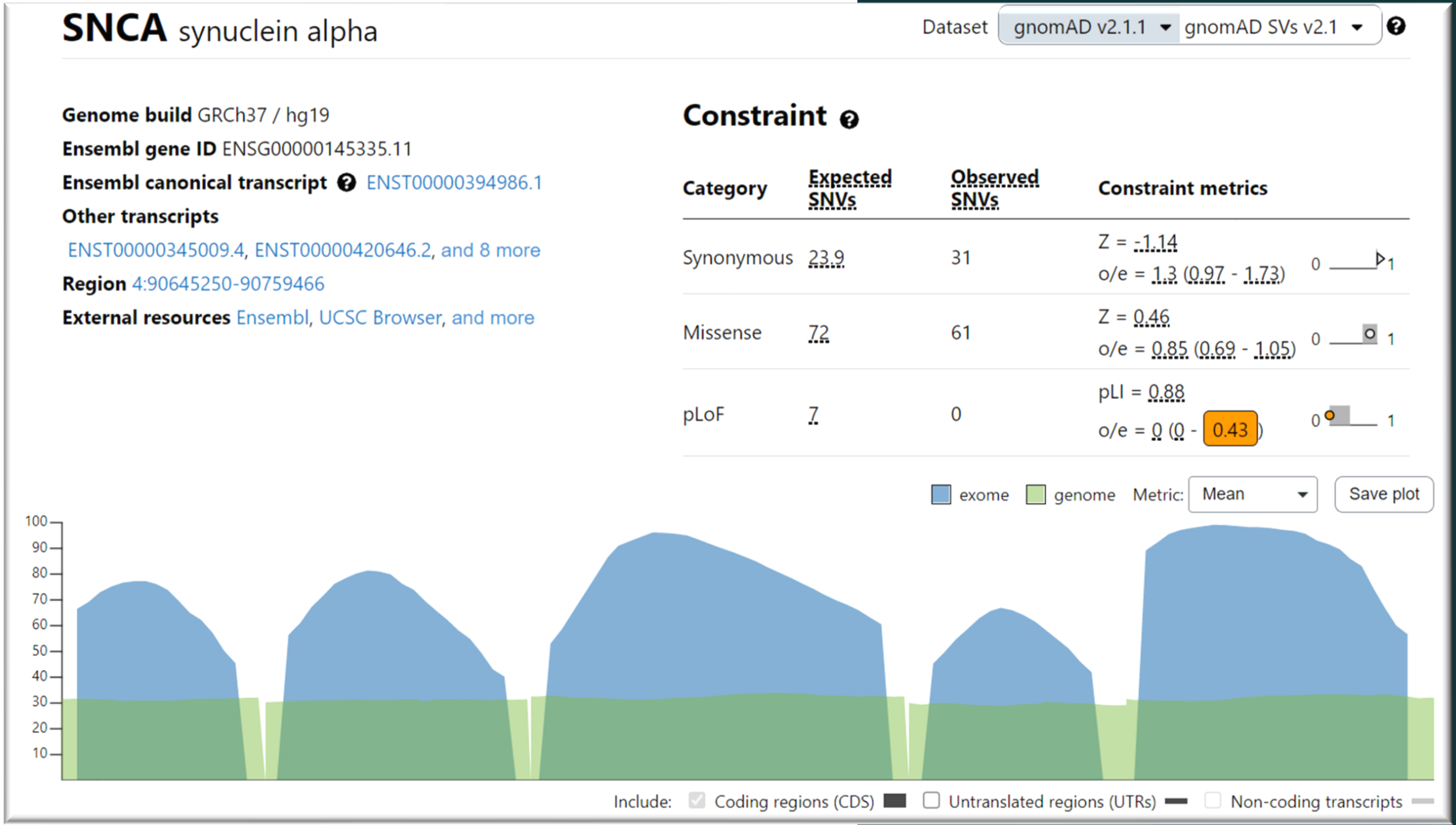


02

Phase I

Can we learn from a healthy population about the genetic basis of Neurodegenerative Diseases?

GnomAD data



SNCA synuclein alpha

Dataset gnomAD v2.1.1 ▾ gnomAD SVs v2.1 ▾

Genome build GRCh37 / hg19

Ensembl gene ID ENSG00000145335.11

Ensembl canonical transcript ENST00000394986.1

Other transcripts

ENST00000345009.4, ENST00000420646.2, and 8 more

Region 4:90645250-90759466

External resources Ensembl, UCSC Browser, and more

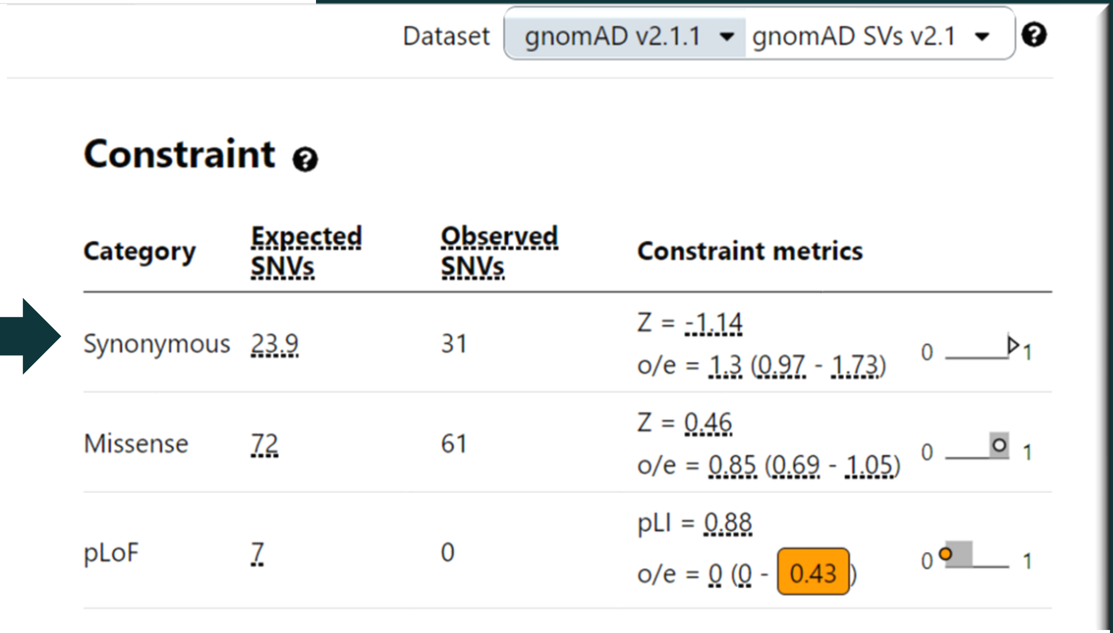
Constraint

Category	Expected SNVs	Observed SNVs	Constraint metrics
Synonymous	23.9	31	Z = -1.14 o/e = 1.3 (0.97 - 1.73)
Missense	72	61	Z = 0.46 o/e = 0.85 (0.69 - 1.05)
pLoF	7	0	pLI = 0.88 o/e = 0 (0 - 0.43)

Include: ☑ Coding regions (CDS) □ Untranslated regions (UTRs) □ Non-coding transcripts

GnomAD data

change that doesn't change the amino acid



Dataset gnomAD v2.1.1 gnomAD SVs v2.1

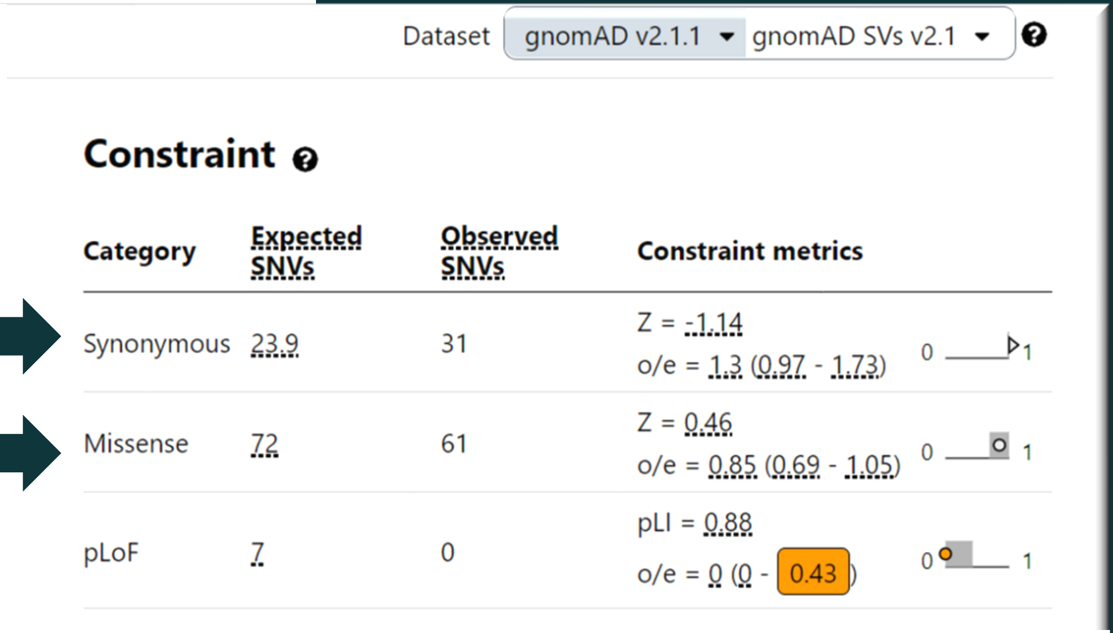
Constraint

Category	Expected SNVs	Observed SNVs	Constraint metrics
Synonymous	23.9	31	Z = -1.14 o/e = 1.3 (0.97 - 1.73)
Missense	72	61	Z = 0.46 o/e = 0.85 (0.69 - 1.05)
pLoF	7	0	pLI = 0.88 o/e = 0 (0 - 0.43)

GnomAD data

change that doesn't change the amino acid

change of a single nucleotide



Dataset gnomAD v2.1.1 gnomAD SVs v2.1

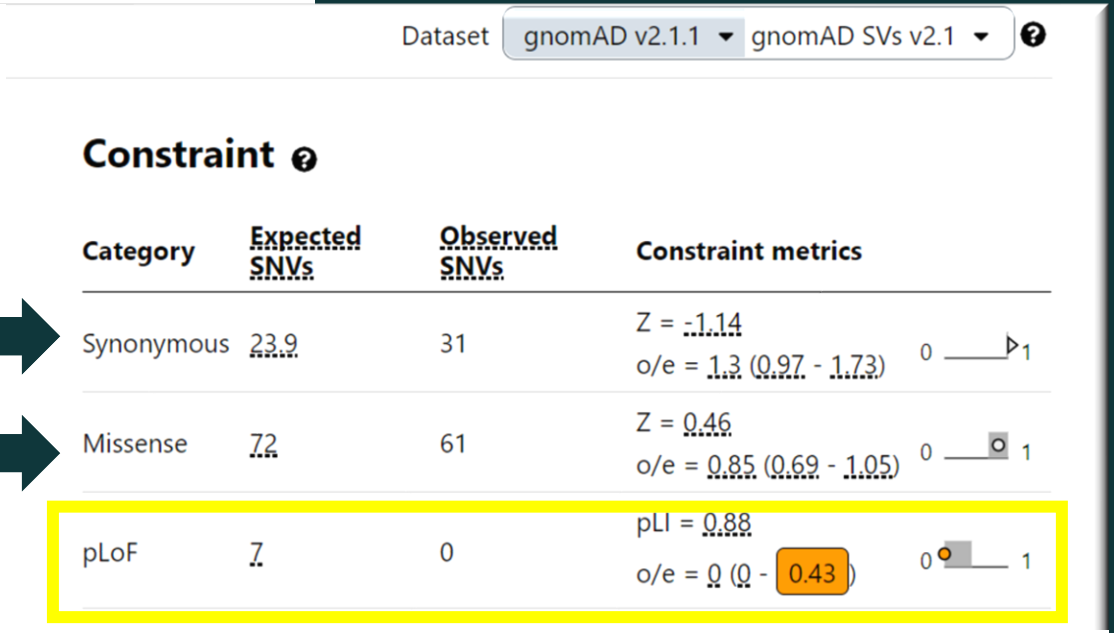
Constraint

Category	Expected SNVs	Observed SNVs	Constraint metrics
Synonymous	23.9	31	Z = -1.14 o/e = 1.3 (0.97 - 1.73)
Missense	72	61	Z = 0.46 o/e = 0.85 (0.69 - 1.05)
pLoF	7	0	pLI = 0.88 o/e = 0 (0 - 0.43)

GnomAD data

change that doesn’t change the amino acid

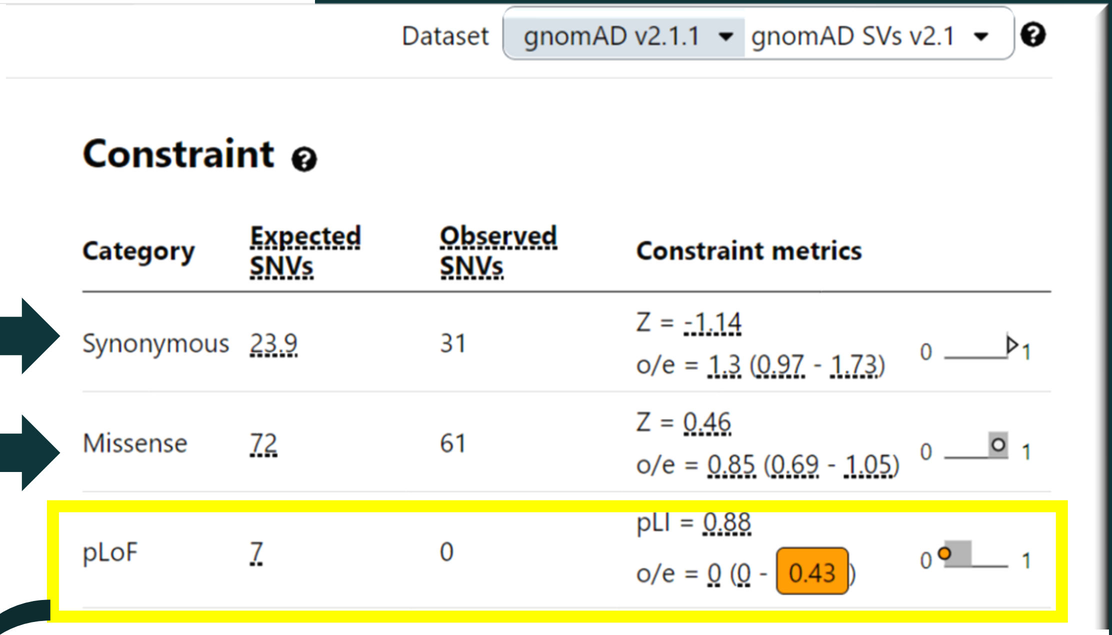
change of a single nucleotide



Constraint

Category	Expected SNVs	Observed SNVs	Constraint metrics	
Synonymous	23.9	31	Z = -1.14 o/e = 1.3 (0.97 - 1.73)	0 1
Missense	72	61	Z = 0.46 o/e = 0.85 (0.69 - 1.05)	0 1
pLoF	7	0	pLI = 0.88 o/e = 0 (0 - 0.43)	0 1

GnomAD data



Dataset gnomAD v2.1.1 ▾ gnomAD SVs v2.1 ▾

Constraint

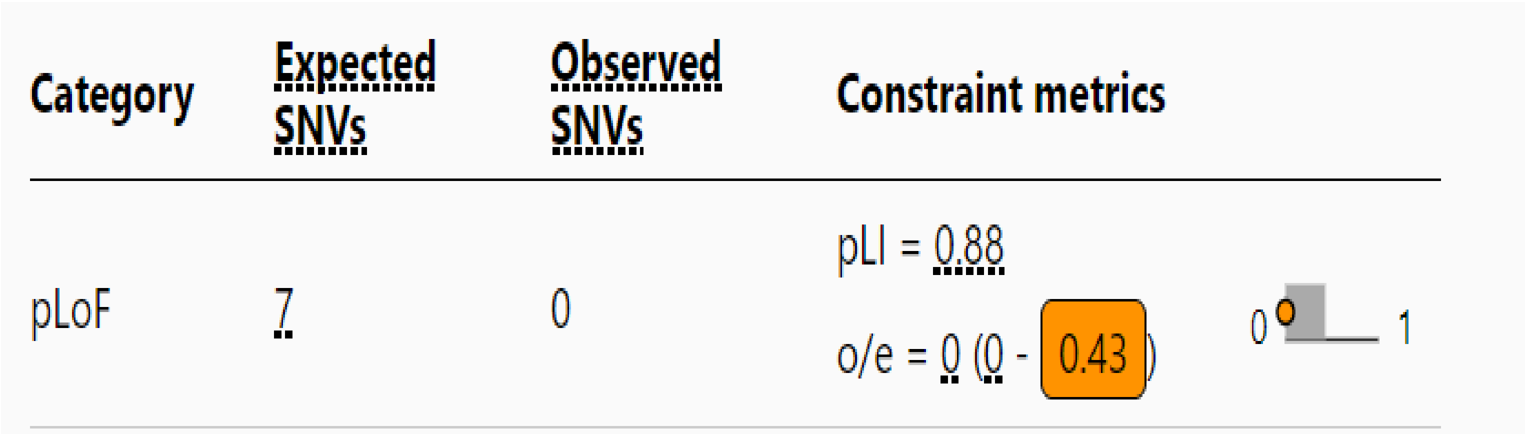
Category	Expected SNVs	Observed SNVs	Constraint metrics	
Synonymous	23.9	31	Z = -1.14 o/e = 1.3 (0.97 - 1.73)	0 1
Missense	72	61	Z = 0.46 o/e = 0.85 (0.69 - 1.05)	0 1
pLoF	7	0	pLI = 0.88 o/e = 0 (0 - 0.43)	0 1

change that doesn't change the amino acid

change of a single nucleotide

changes that interrupt the protein function

Re-calculation of the pLoF score – WHY?



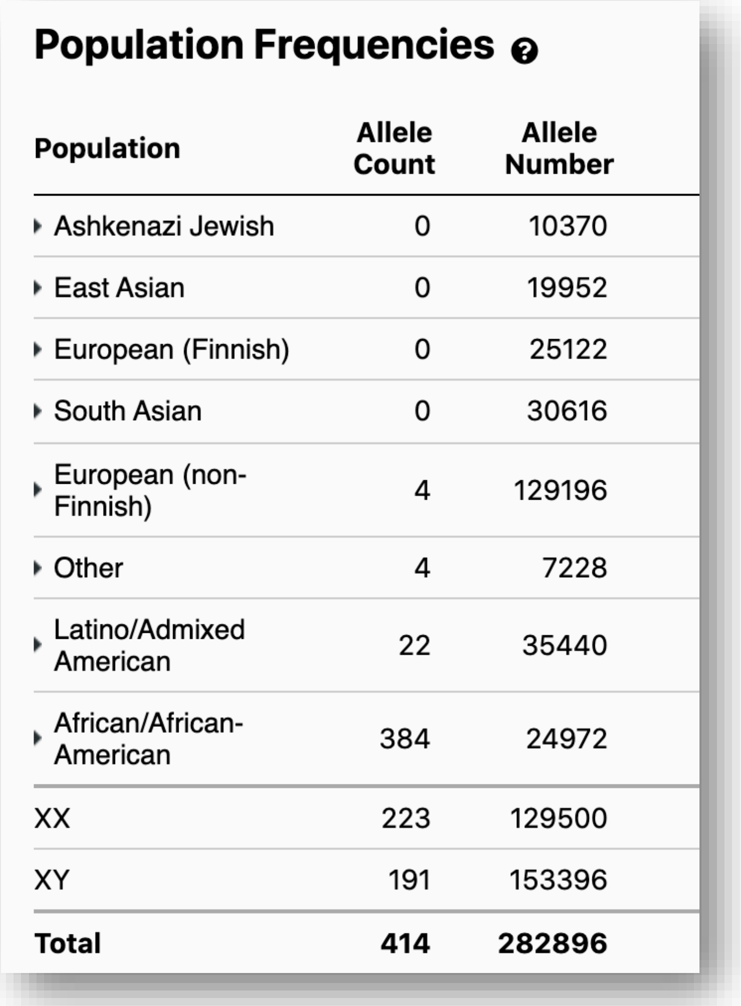
Category	Expected SNVs	Observed SNVs	Constraint metrics	
pLoF	7	0	pLI = 0.88 o/e = 0 (0 - 0.43)	0 1

Re-calculation of the pLoF score – WHY?

0 : 10370

0 : 19952

1: 32299

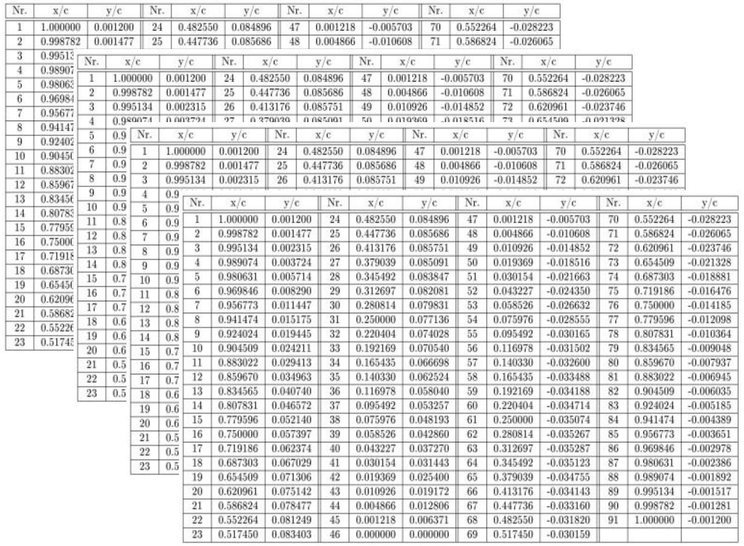


Population Frequencies

Population	Allele Count	Allele Number
Ashkenazi Jewish	0	10370
East Asian	0	19952
European (Finnish)	0	25122
South Asian	0	30616
European (non-Finnish)	4	129196
Other	4	7228
Latino/Admixed American	22	35440
African/African-American	384	24972
XX	223	129500
XY	191	153396
Total	**414**	**282896**

Re-calculation of the pLoF score

Chromosomal VCF files



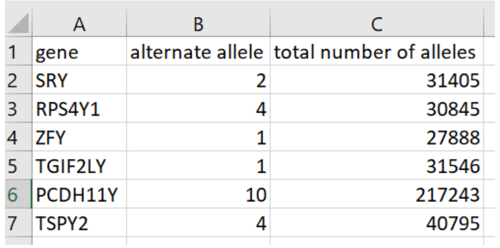
Nr.	x/c	y/c	Nr.	x/c	y/c	Nr.	x/c	y/c	Nr.	x/c	y/c
1	1.000000	0.001200	24	0.482550	0.084896	47	0.001218	-0.005703	70	0.552264	-0.028223
2	0.998782	0.001477	25	0.447736	0.085686	48	0.004866	-0.010608	71	0.586824	-0.026065
3	0.995134	0.002315	26	0.413176	0.085751	49	0.010926	-0.014852	72	0.620961	-0.023746
4	0.989074	0.003724	27	0.379039	0.085091	50	0.019369	-0.018516	73	0.654509	-0.021328
5	0.980631	0.005714	28	0.345492	0.083847	51	0.030154	-0.021663	74	0.687303	-0.018881
6	0.969846	0.008290	29	0.312697	0.082081	52	0.043227	-0.024350	75	0.719186	-0.016476
7	0.956773	0.011447	30	0.280814	0.079831	53	0.058526	-0.026632	76	0.750000	-0.014185
8	0.941474	0.015175	31	0.250000	0.077136	54	0.075976	-0.028555	77	0.779596	-0.012098
9	0.924024	0.019445	32	0.220404	0.074028	55	0.095492	-0.030165	78	0.807831	-0.010364
10	0.904509	0.024211	33	0.192169	0.070540	56	0.116978	-0.031502	79	0.834565	-0.009048
11	0.883022	0.029413	34	0.165435	0.066698	57	0.140330	-0.032600	80	0.859670	-0.007937
12	0.859670	0.034963	35	0.140330	0.062524	58	0.165435	-0.033488	81	0.883022	-0.006945
13	0.834565	0.040740	36	0.116978	0.058040	59	0.192169	-0.034188	82	0.904509	-0.006035
14	0.807831	0.046572	37	0.095492	0.053257	60	0.220404	-0.034714	83	0.924024	-0.005185
15	0.779596	0.052140	38	0.075976	0.048193	61	0.250000	-0.035074	84	0.941474	-0.004389
16	0.750000	0.057397	39	0.058526	0.042860	62	0.280814	-0.035267	85	0.956773	-0.003651
17	0.719186	0.062374	40	0.043227	0.037270	63	0.312697	-0.035287	86	0.969846	-0.002978
18	0.687303	0.067029	41	0.030154	0.031443	64	0.345492	-0.035123	87	0.980631	-0.002386
19	0.654509	0.071306	42	0.019369	0.025400	65	0.379039	-0.034755	88	0.989074	-0.001892
20	0.620961	0.075142	43	0.010926	0.019172	66	0.413176	-0.034143	89	0.995134	-0.001517
21	0.586824	0.078477	44	0.004866	0.012806	67	0.447736	-0.033160	90	0.998782	-0.001281
22	0.552264	0.081249	45	0.001218	0.006371	68	0.482550	-0.031820	91	1.000000	-0.001200
23	0.517450	0.083403	46	0.000000	0.000000	69	0.517450	-0.030159			

	Neuro candidates All - non neuro			non neuro		
	Alternate allele count	Total num of alleles	Alternate allele frequency	Alternate allele count	Total num of alleles	Alternate allele frequency
$var0_{geneX}$	2	31405	$6.37E-5$	2	30272	6.61E – 5
$var1_{geneX}$	0	30657	$1.30E-4$	1	30561	9.82E – 5
$var0_{geneY}$	1	27888	$3.59E-5$	4	17201	2.33E – 4
$vars$ …	…	…	…	…	…	…
	…	…	…	…	…	…

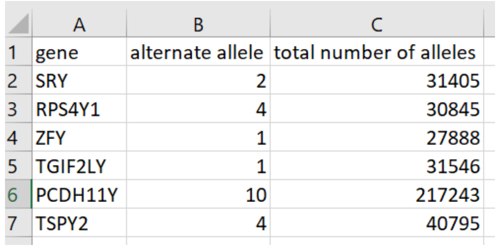
$$observedSNV_{gene} = \sum_{altarnate \in var} \sum_{lof\_var \in geneX} variant$$

$$expectedSNV_{gene} = \frac{\sum_{altarnate \in var} \sum_{lof\_var \in geneX} variant}{2}$$

2 lists with 18k gene
and their data



	A	B	C
1	gene	alternate allele	total number of alleles
2	SRY	2	31405
3	RPS4Y1	4	30845
4	ZFY	1	27888
5	TGIF2LY	1	31546
6	PCDH11Y	10	217243
7	TSPY2	4	40795



	A	B	C
1	gene	alternate allele	total number of alleles
2	SRY	2	31405
3	RPS4Y1	4	30845
4	ZFY	1	27888
5	TGIF2LY	1	31546
6	PCDH11Y	10	217243
7	TSPY2	4	40795

Re-calculation of the pLoF score

$$observedSNV_{gene} = \sum_{altarnate \in var} \sum_{lof\_var \in geneX} variant$$

Assumption

- One of the PLOF mutation:

  transcript ablation

  Splice acceptor variant

  Splice donor variant

  Stop gained

  frameshift variant

- Alternate allele > 0 → total ++
- $\frac{Total\ allele\ number}{2}$ s.t every person has 1

2 lists of 18,000 genes with their sum of alternate allele count and total number of allele count

Fisher exact and Bonferroni Correction

$$p = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{a!\,b!\,c!\,d!\,n!}$$

$$\alpha^{*} = 1 - \left(1 - \frac{\alpha}{n}\right)^{\mathrm{n}}$$

$p = p - value$
$a, b, c, d = values\ of\ table$
$n = set\ size$
$\alpha = given\ alpha$

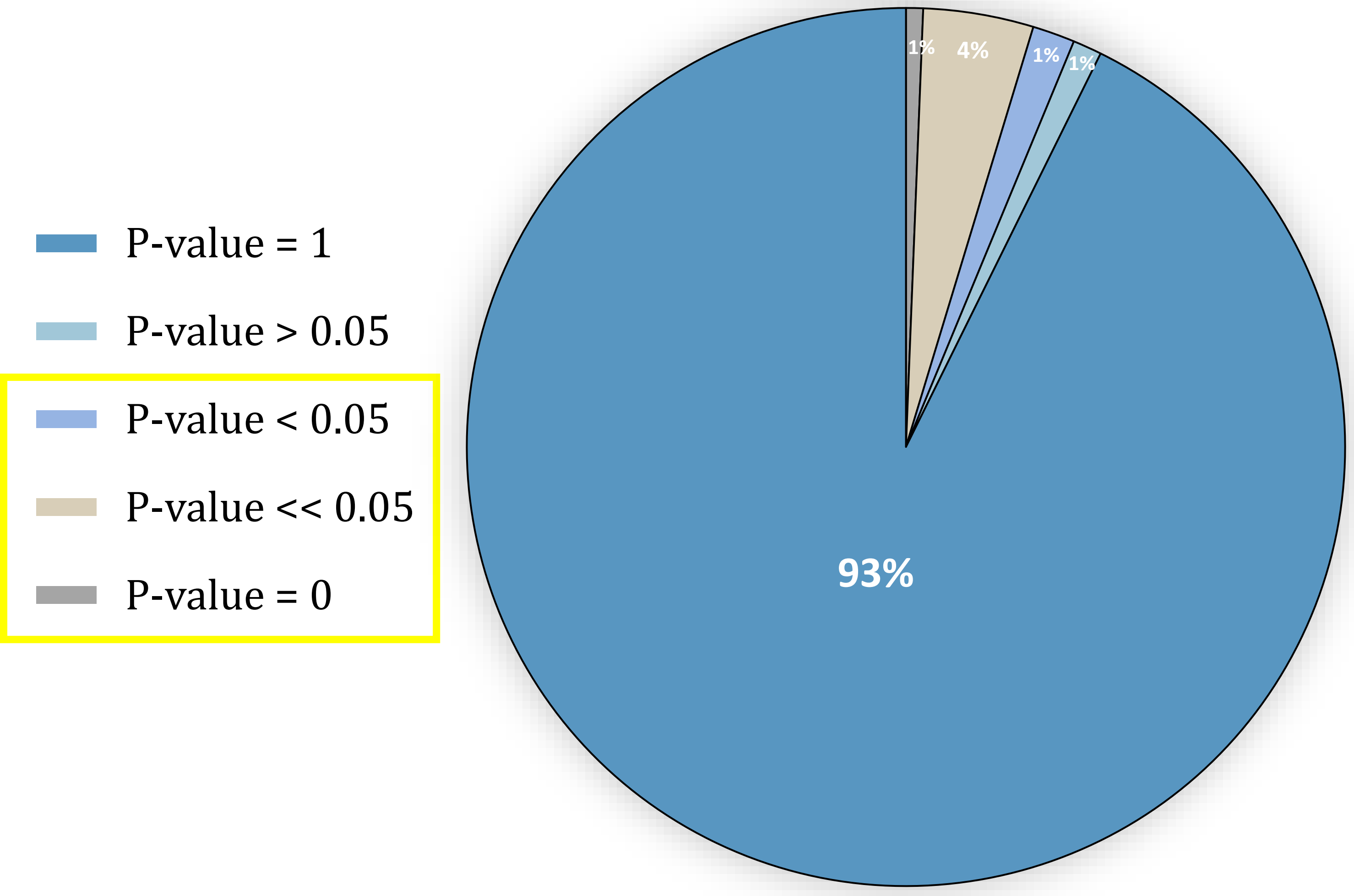
Using Fisher p-value as pLoF score

18k genes ×

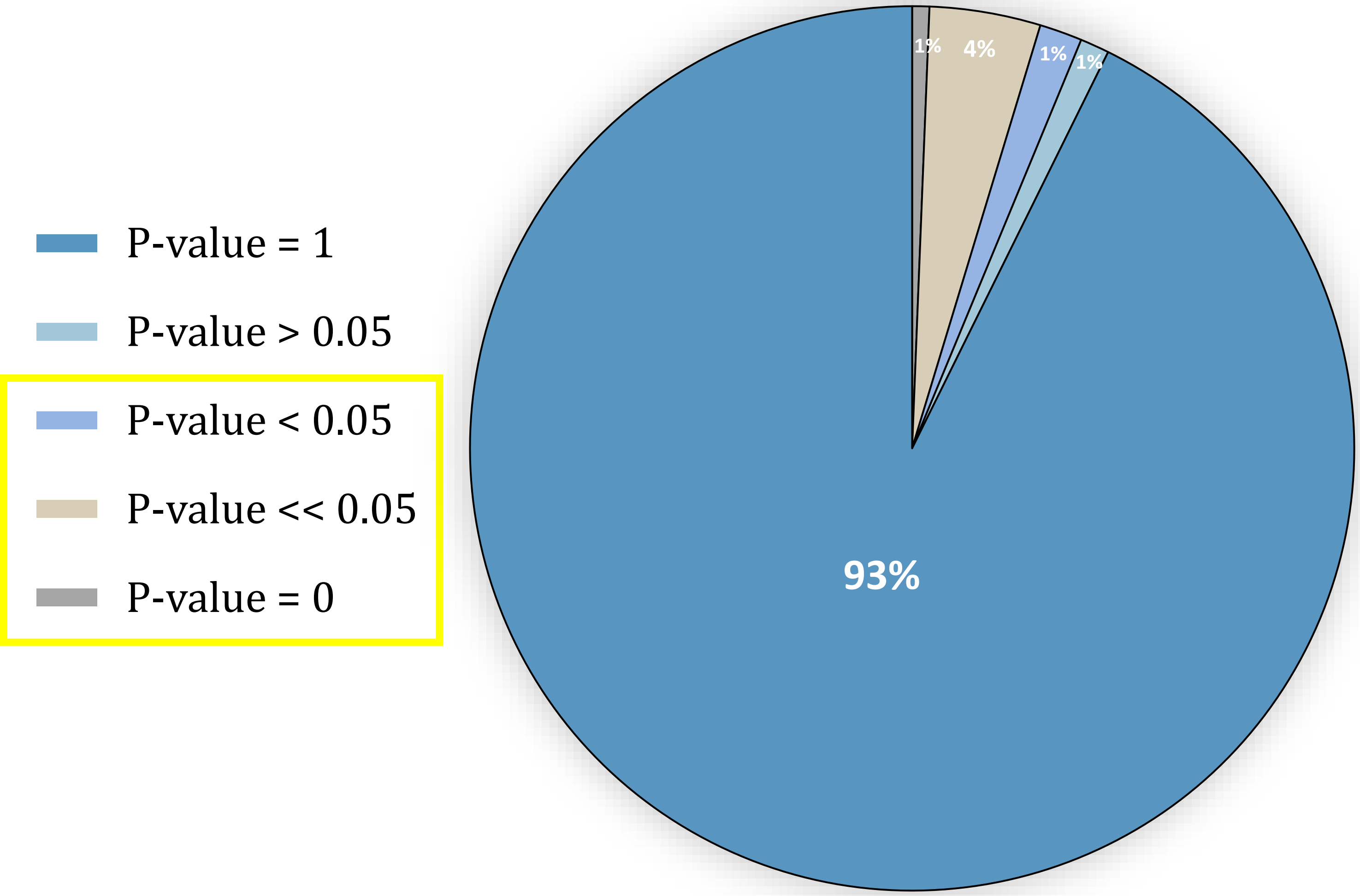
	altearante allele = num of Lof var	*total count = total vars*	
non − neuro	20	20,000	20020
candidates	30	30,000	30030
	50	50000	50,050

P-value

Gene's constraint distribution of European

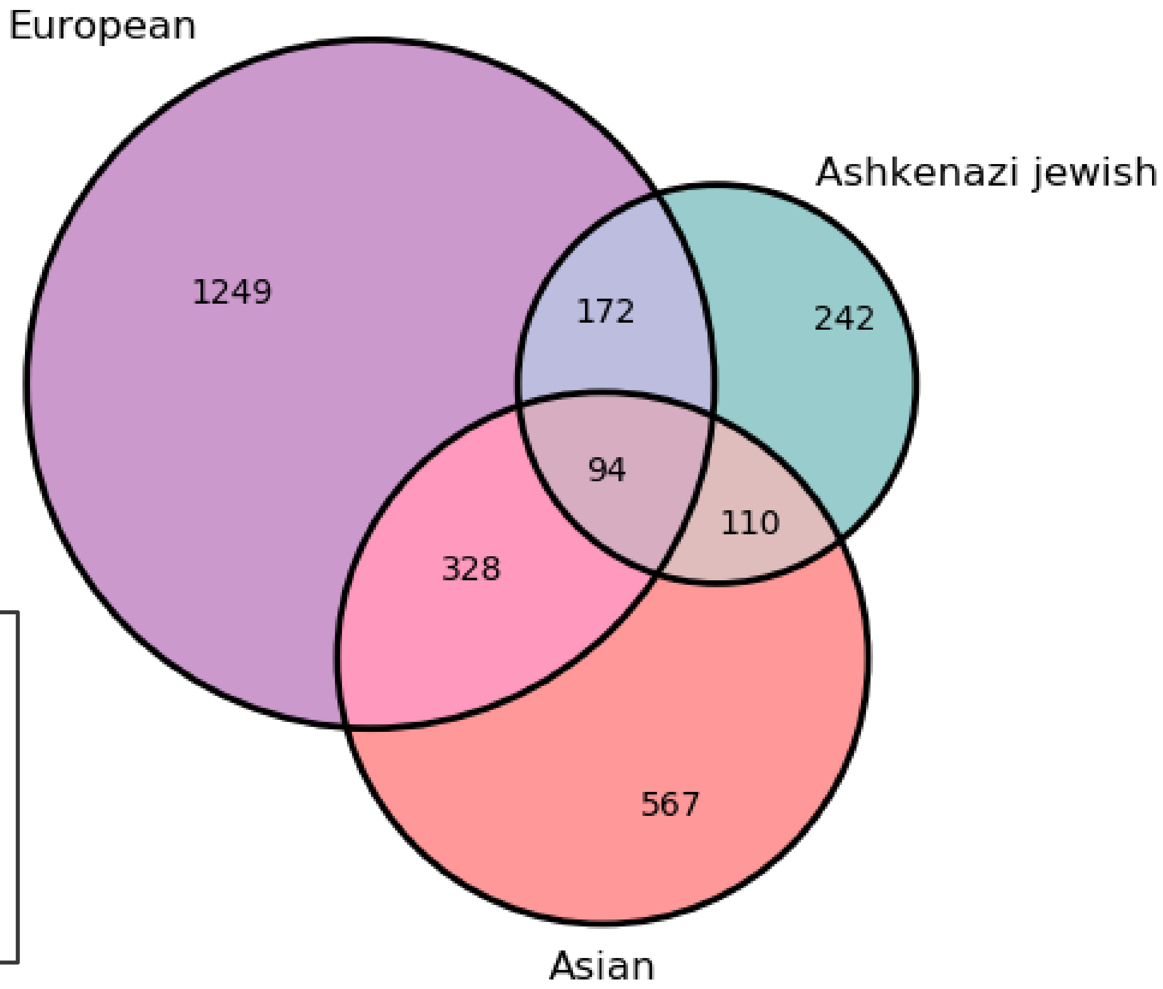


Gene's constraint distribution of European



Most of the genes have p-value = 1 → can change and have number of variants

Constraint in different populations

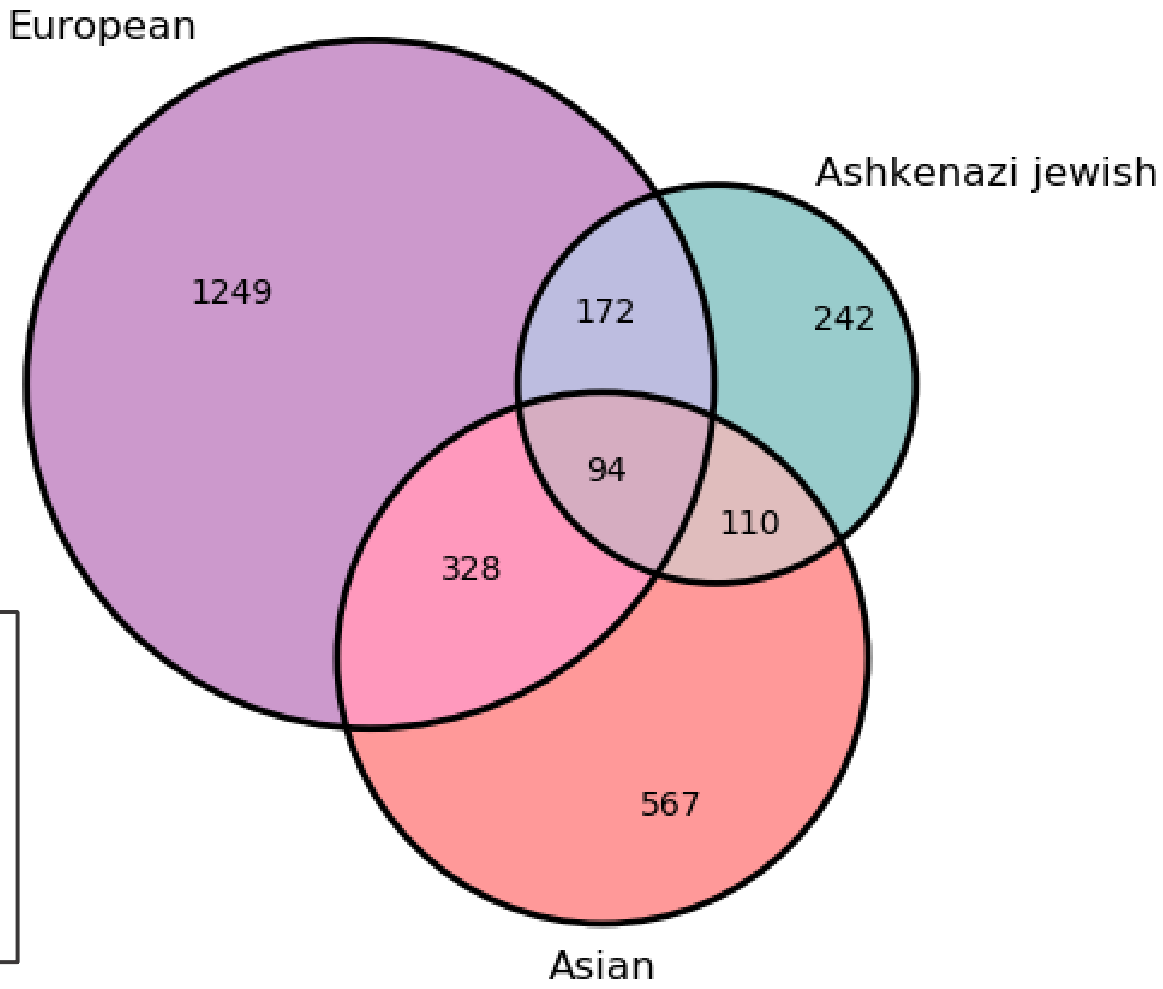


European = 1249

Ashkenazi Jewish = 242

Asian = 567

Constraint in different populations



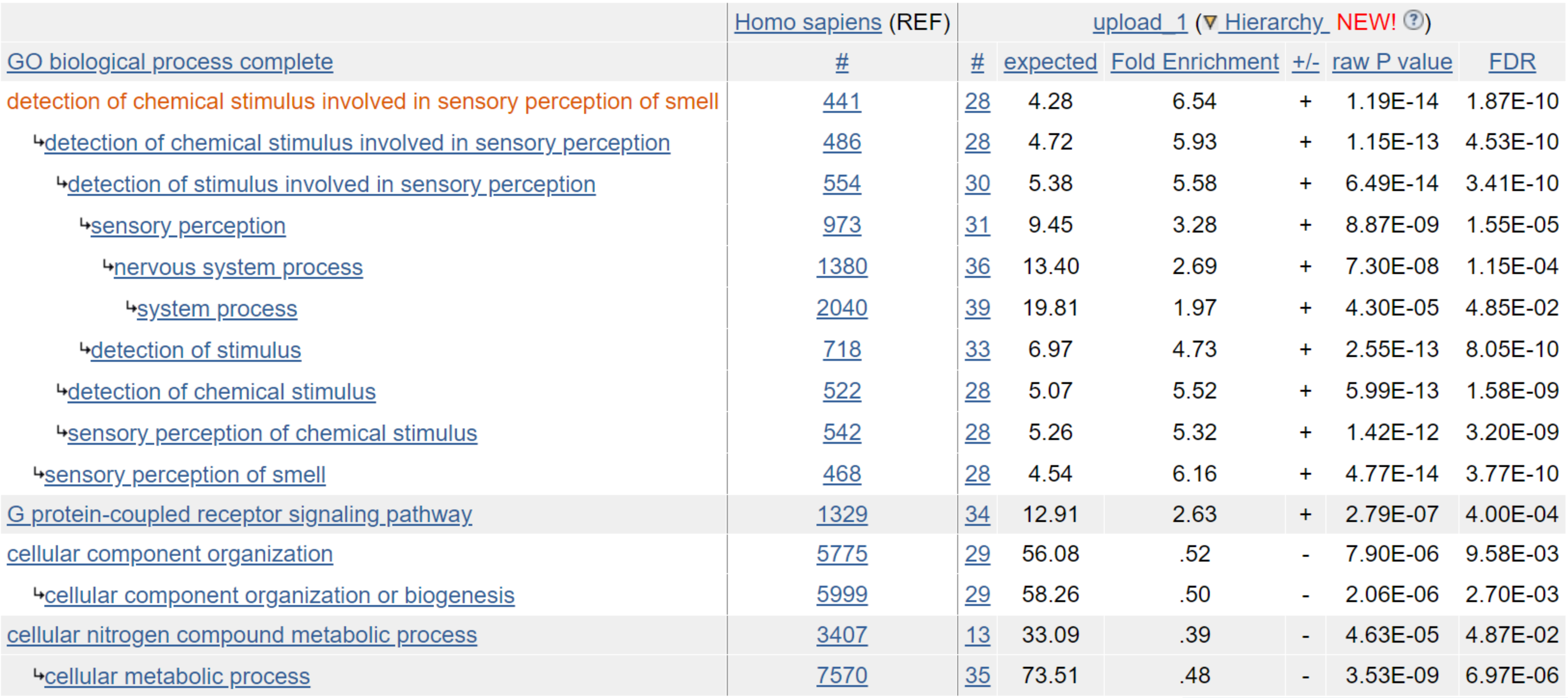
European = 1249

Ashkenazi Jewish = 242

Asian = 567

94 overlap genes

Ontology results of the most suspected genes



	Homo sapiens (REF)	upload_1 (▼ Hierarchy NEW! ?)					
GO biological process complete	#	#	expected	Fold Enrichment	+/-	raw P value	FDR
detection of chemical stimulus involved in sensory perception of smell	441	28	4.28	6.54	+	1.19E-14	1.87E-10
↳detection of chemical stimulus involved in sensory perception	486	28	4.72	5.93	+	1.15E-13	4.53E-10
↳detection of stimulus involved in sensory perception	554	30	5.38	5.58	+	6.49E-14	3.41E-10
↳sensory perception	973	31	9.45	3.28	+	8.87E-09	1.55E-05
↳nervous system process	1380	36	13.40	2.69	+	7.30E-08	1.15E-04
↳system process	2040	39	19.81	1.97	+	4.30E-05	4.85E-02
↳detection of stimulus	718	33	6.97	4.73	+	2.55E-13	8.05E-10
↳detection of chemical stimulus	522	28	5.07	5.52	+	5.99E-13	1.58E-09
↳sensory perception of chemical stimulus	542	28	5.26	5.32	+	1.42E-12	3.20E-09
↳sensory perception of smell	468	28	4.54	6.16	+	4.77E-14	3.77E-10
G protein-coupled receptor signaling pathway	1329	34	12.91	2.63	+	2.79E-07	4.00E-04
cellular component organization	5775	29	56.08	.52	-	7.90E-06	9.58E-03
↳cellular component organization or biogenesis	5999	29	58.26	.50	-	2.06E-06	2.70E-03
cellular nitrogen compound metabolic process	3407	13	33.09	.39	-	4.63E-05	4.87E-02
↳cellular metabolic process	7570	35	73.51	.48	-	3.53E-09	6.97E-06

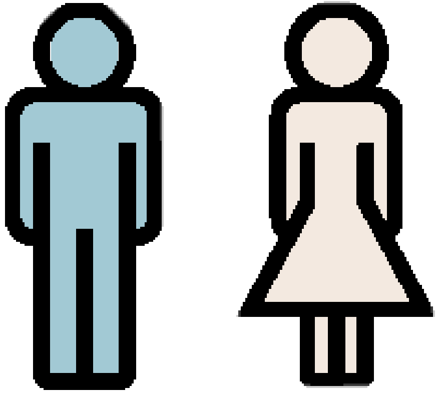
03

Phase II

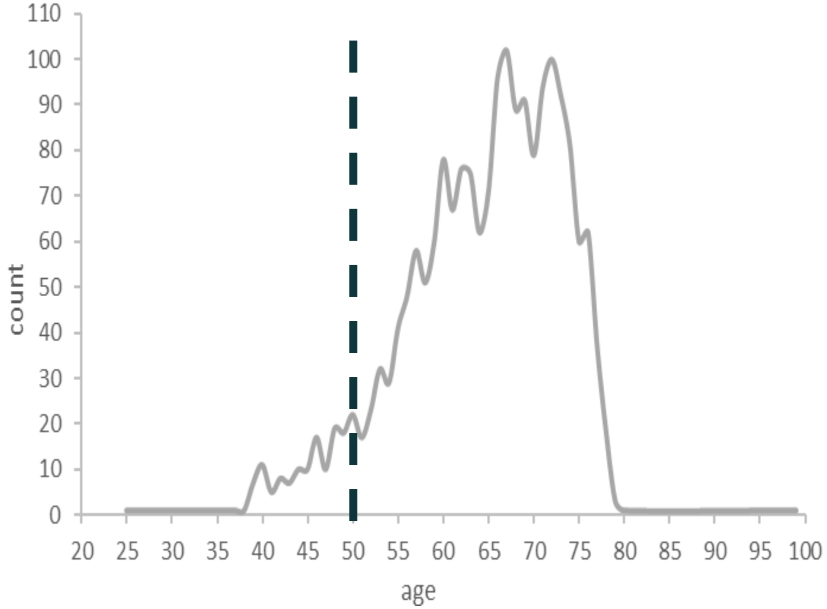
Can Early Onset of Complex Diseases

be a hint for Etiology?

Parkinson in numbers



60% 40%



Age distribution

10M

people worldwide

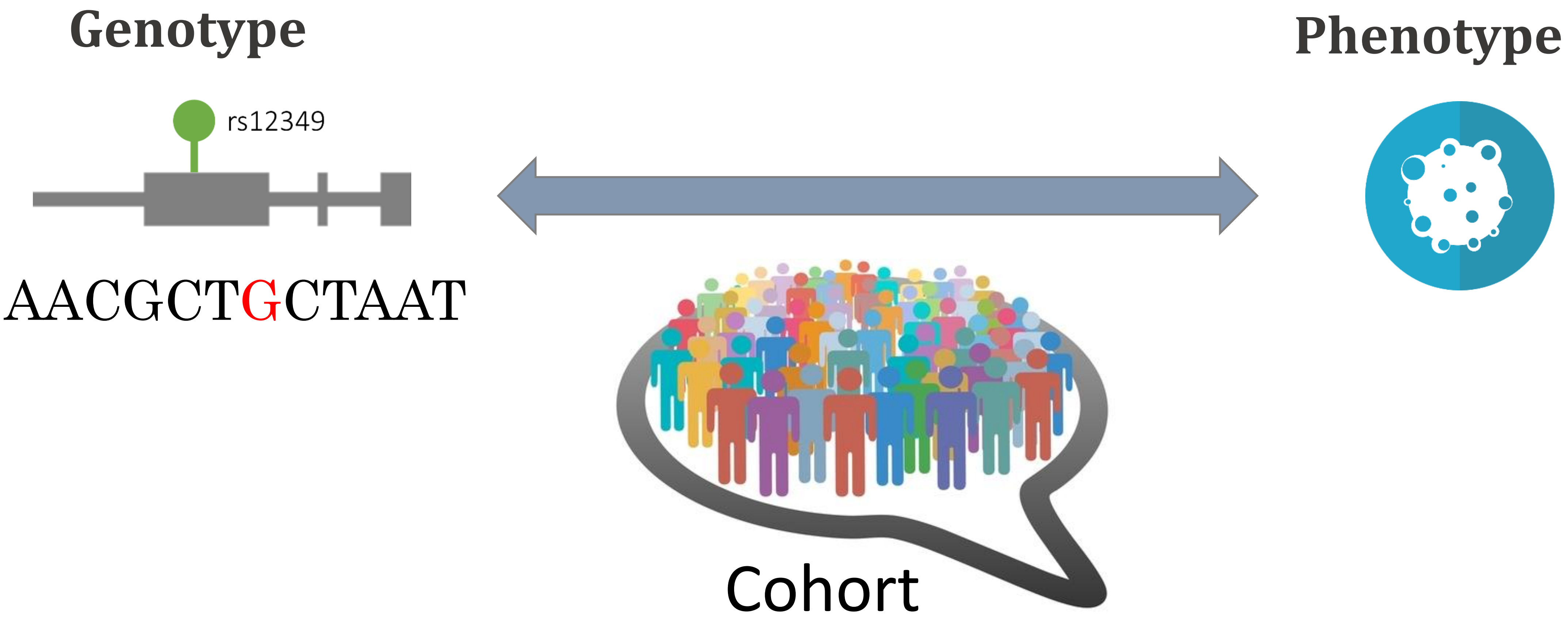
60k

new cases every year

52B$

every year

GWAS - genome wide association study



Genotype

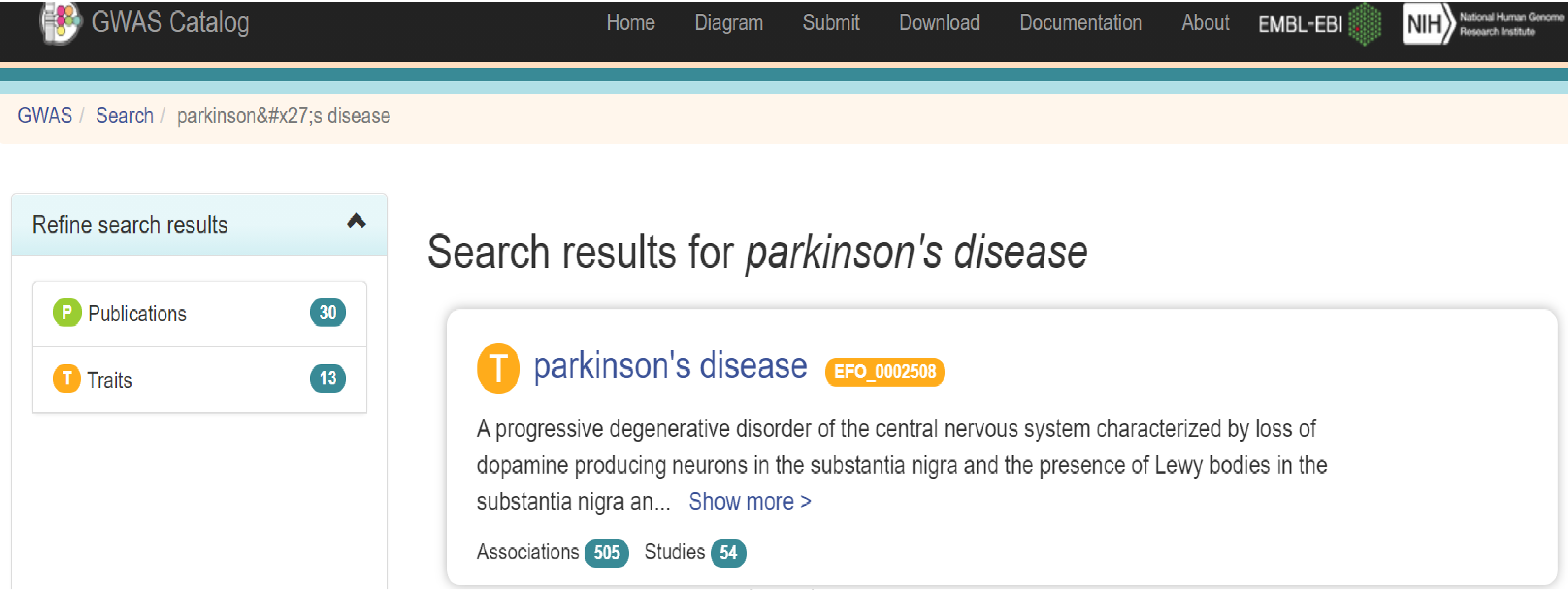
rs12349

AACGCTGCTAAT

Phenotype

Cohort

GWAS - genome wide association study



GWAS Catalog
Home
Diagram
Submit
Download
Documentation
About
EMBL-EBI
NIH National Human Genome Research Institute
GWAS / Search / parkinson&#x27;s disease
Refine search results

P Publications	30
T Traits	13

Search results for *parkinson's disease*

T parkinson's disease EFO_0002508

A progressive degenerative disorder of the central nervous system characterized by loss of dopamine producing neurons in the substantia nigra and the presence of Lewy bodies in the substantia nigra an... Show more >

Associations 505 Studies 54

What data did we have from GWAS?

Late: 1236 variants
→ 43 genes

Early: 2254 variants
→ 161 genes

4 overlap genes (all are known as related)

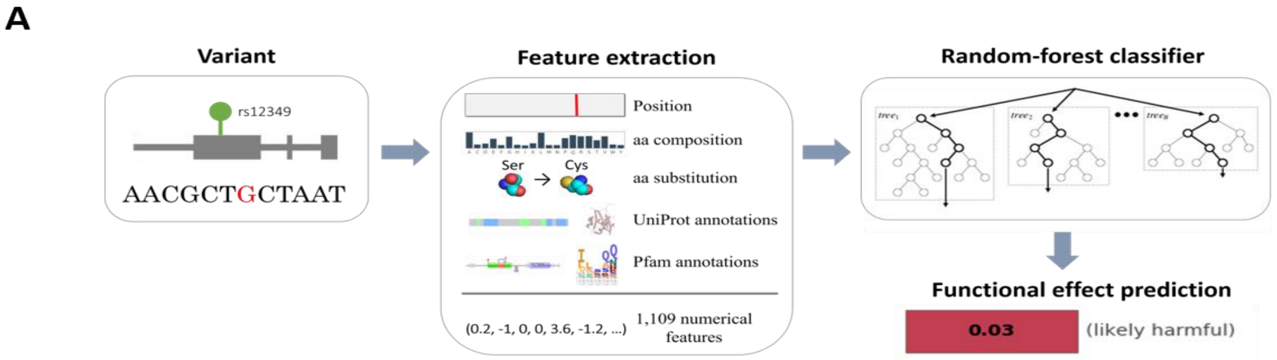
FABRIC

(**F**unctional **A**lteration **B**ias **R**ecovery **I**n **C**oding-regions)

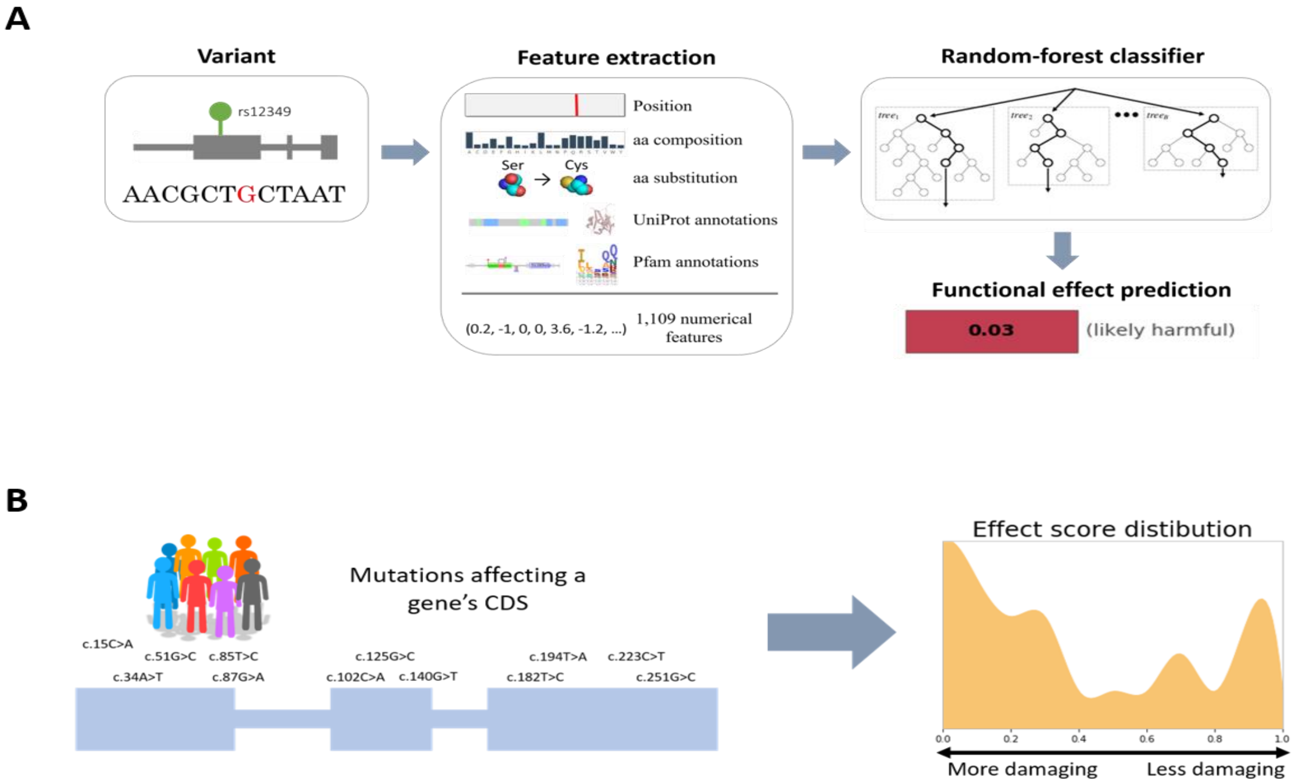
- Assess the **impact** of mutations on gene/protein **function**

- Find genes more damaged than expected
  - Given this number of *random* mutations

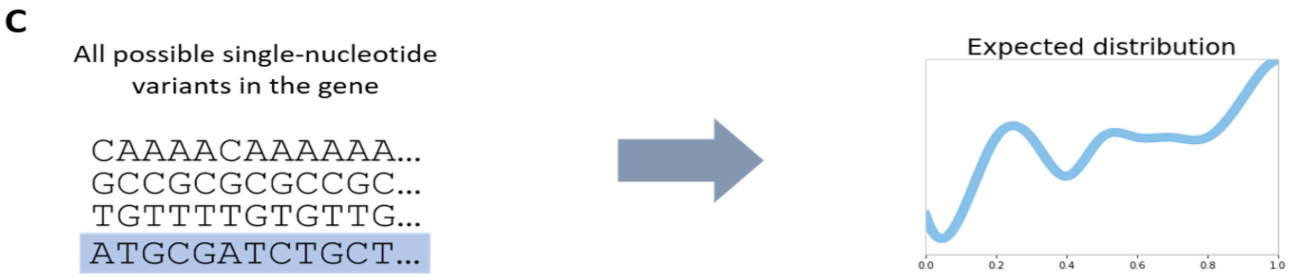
FABRIC


A
Variant
rs12349
AACGCTGCTAAT
Feature extraction
Position
aa composition
Ser
Cys
aa substitution
UniProt annotations
Pfam annotations
(0.2, -1, 0, 0, 3.6, -1.2, ...)
1,109 numerical features
Random-forest classifier
tree1
tree2
treeB
Functional effect prediction
0.03
(likely harmful)

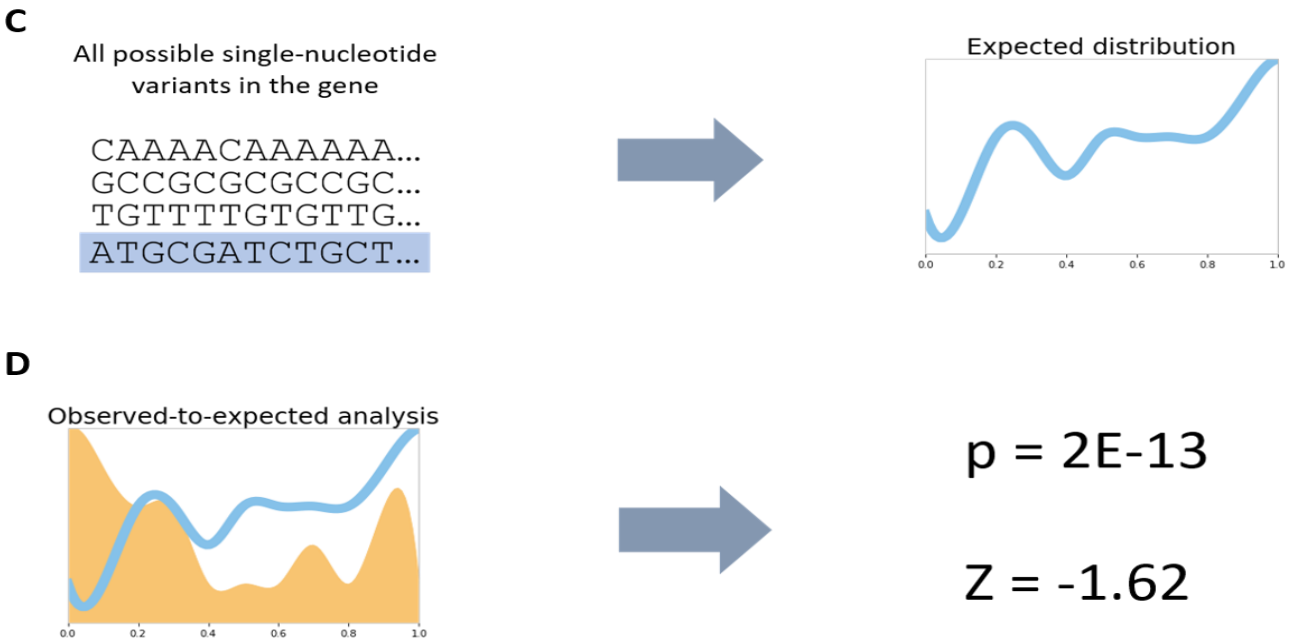
FABRIC


A
Variant
rs12349
AACGCTGCTAAT
Feature extraction
Position
aa composition
Ser
Cys
aa substitution
UniProt annotations
Pfam annotations
(0.2, -1, 0, 0, 3.6, -1.2, ...)
1,109 numerical features
Random-forest classifier
tree1
tree2
treeN
Functional effect prediction
0.03
(likely harmful)
B
Mutations affecting a gene's CDS
c.15C>A
c.51G>C
c.85T>C
c.34A>T
c.87G>A
c.125G>C
c.102C>A
c.140G>T
c.194T>A
c.223C>T
c.182T>C
c.251G>C
Effect score distibution
0.0
0.2
0.4
0.6
0.8
1.0
More damaging
Less damaging

FABRIC


C
All possible single-nucleotide variants in the gene
CAAAACAAAAAA...
GCCGCGCGCCGC...
TGTTTTGTGTTG...
ATGCGATCTGCT...
Expected distribution
0.0
0.2
0.4
0.6
0.8
1.0

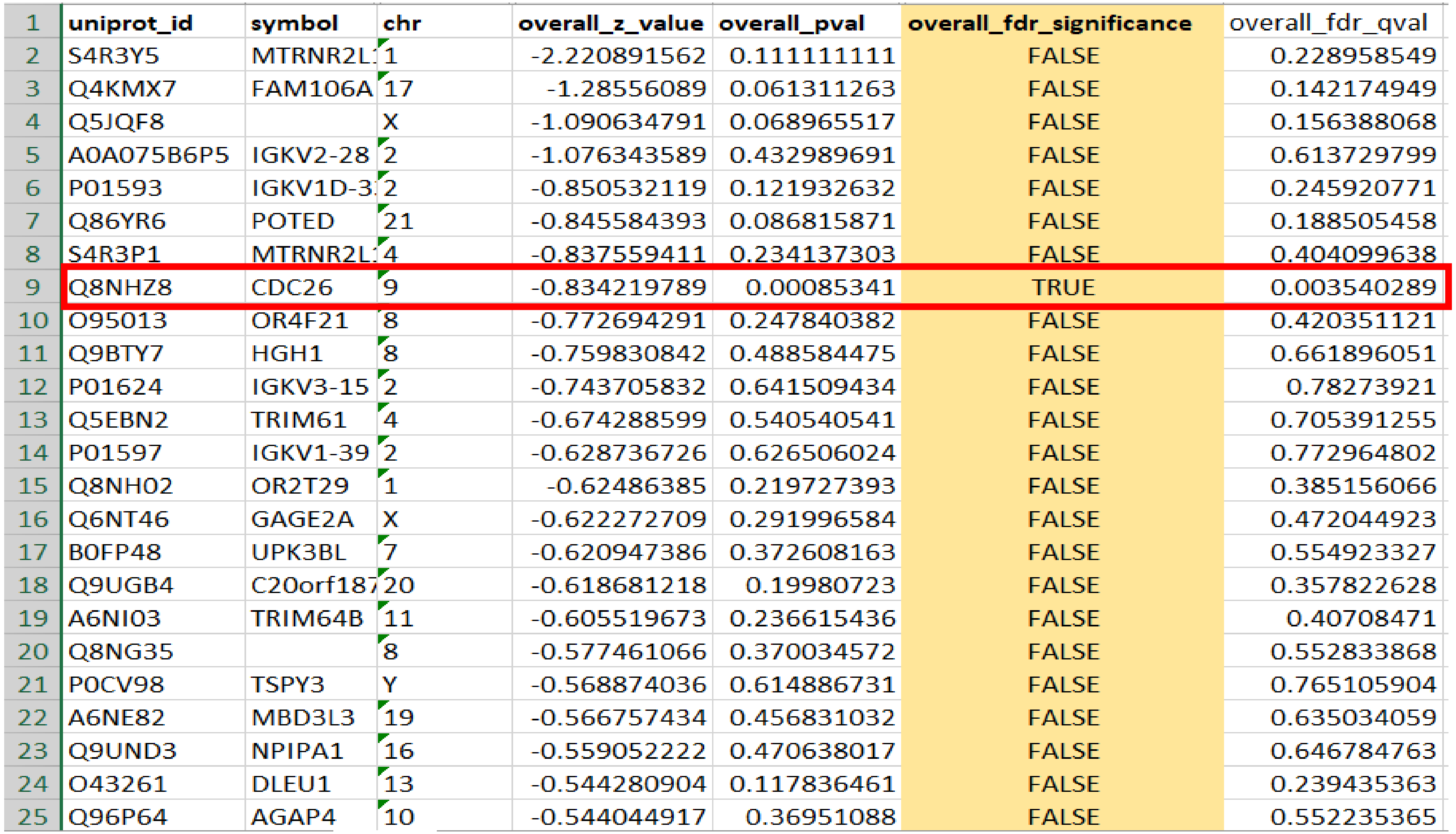
FABRIC



$p = 2E\text{-}13$

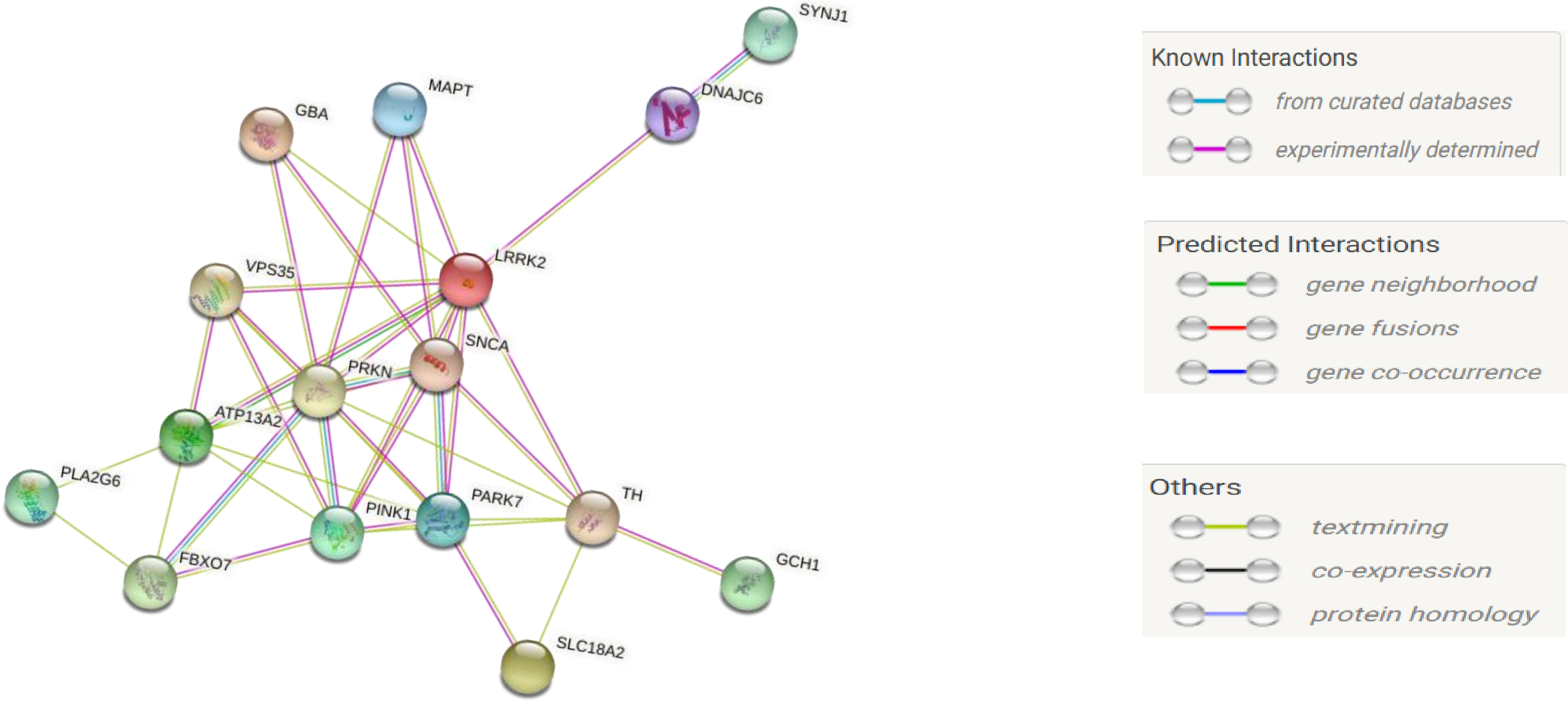
$Z = -1.62$

FABRIC output



1	uniprot_id	symbol	chr	overall_z_value	overall_pval	overall_fdr_significance	overall_fdr_qval
2	S4R3Y5	MTRNR2L1	1	-2.220891562	0.111111111	FALSE	0.228958549
3	Q4KMX7	FAM106A	17	-1.28556089	0.061311263	FALSE	0.142174949
4	Q5JQF8		X	-1.090634791	0.068965517	FALSE	0.156388068
5	A0A075B6P5	IGKV2-28	2	-1.076343589	0.432989691	FALSE	0.613729799
6	P01593	IGKV1D-3	2	-0.850532119	0.121932632	FALSE	0.245920771
7	Q86YR6	POTED	21	-0.845584393	0.086815871	FALSE	0.188505458
8	S4R3P1	MTRNR2L1	4	-0.837559411	0.234137303	FALSE	0.404099638
9	Q8NHZ8	CDC26	9	-0.834219789	0.00085341	TRUE	0.003540289
10	O95013	OR4F21	8	-0.772694291	0.247840382	FALSE	0.420351121
11	Q9BTY7	HGH1	8	-0.759830842	0.488584475	FALSE	0.661896051
12	P01624	IGKV3-15	2	-0.743705832	0.641509434	FALSE	0.78273921
13	Q5EBN2	TRIM61	4	-0.674288599	0.540540541	FALSE	0.705391255
14	P01597	IGKV1-39	2	-0.628736726	0.626506024	FALSE	0.772964802
15	Q8NH02	OR2T29	1	-0.62486385	0.219727393	FALSE	0.385156066
16	Q6NT46	GAGE2A	X	-0.622272709	0.291996584	FALSE	0.472044923
17	B0FP48	UPK3BL	7	-0.620947386	0.372608163	FALSE	0.554923327
18	Q9UGB4	C20orf187	20	-0.618681218	0.19980723	FALSE	0.357822628
19	A6NI03	TRIM64B	11	-0.605519673	0.236615436	FALSE	0.40708471
20	Q8NG35		8	-0.577461066	0.370034572	FALSE	0.552833868
21	P0CV98	TSPY3	Y	-0.568874036	0.614886731	FALSE	0.765105904
22	A6NE82	MBD3L3	19	-0.566757434	0.456831032	FALSE	0.635034059
23	Q9UND3	NPIPA1	16	-0.559052222	0.470638017	FALSE	0.646784763
24	O43261	DLEU1	13	-0.544280904	0.117836461	FALSE	0.239435363
25	Q96P64	AGAP4	10	-0.544044917	0.36951088	FALSE	0.552235365

Network of Parkinson genes identified by FABRIC



25 genes were marked as “significant” (positive selection) for early onset, none for late onset

04

What’s Next?

What’s Next?

What’s Next?

Run fabric on the 94 constraint genes in order to test how damaged, they are.

What’s Next?

Run fabric on the
94 constraint genes
in order to test how
damaged, they are.

Compare between
other population &
Run on more diseases
like Alzheimer

What’s Next?

Run fabric on the 94 constraint genes in order to test how damaged, they are.

Compare between other population & Run on more diseases like Alzheimer

Analyze exomes from UK bio-bank & build predictor for individuals

Michal, Amir and Roni

Questions?