Discovering the genetic basis of common disease using sequencing-based cohort studies

EMBO Meeting 7th October 2018

Slavé Petrovski, PhD
Honorary Assoc. Prof. Department of Medicine, University of Melbourne
Honorary Visiting Fellow, Department of Public Health and Primary Care, Cambridge University,

Currently: VP & Head of Genome Analytics & Informatics, Centre for Genomics Research, IMED Biotech Unit, AstraZeneca, Cambridge, UK

Collapsing Analyses

Common Complex Disorders: Rare Variants

Key Accommodations: Allelic and Locus Heterogeneity

Do trait-ascertained samples have more 'qualifying variants' in gene X than controls?

Collapsing Analyses

Common Complex Disorders: Rare Variants

Key Accommodations: Allelic and Locus Heterogeneity

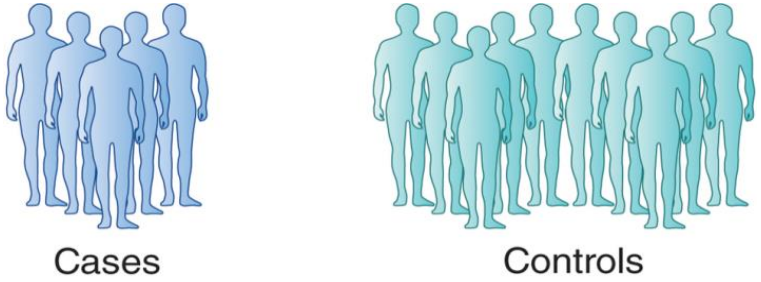
Do trait-ascertained samples have more 'qualifying variants' in gene X than controls?

Example Published Applications:

Amyotrophic Lateral Sclerosis	-	Cirulli E, Lasseigne B, Petrovski S, et al. *Science* 2015
Genetic Generalized Epilepsy*	-	EPI4K Consortium. *Lancet Neurology* 2017
Idiopathic Pulmonary Fibrosis	-	Petrovski S, Todd J, Durheim M, et al. *AJRCCM* 2017
Non-acquired Focal Epilepsy*	-	EPI4K Consortium. *Lancet Neurology* 2017
Sudden Unexplained Death	-	Bagnall R, Crompton D, Petrovski S, et al. *Annals of Neurology* 2016

Cohort Design

Step 1: Select an appropriate control sample, including “controls of convenience”



Step 2: Evaluate QC metrics of samples to ensure high quality WES samples remain. Outlier removal across various sample-level sequencing metrics.

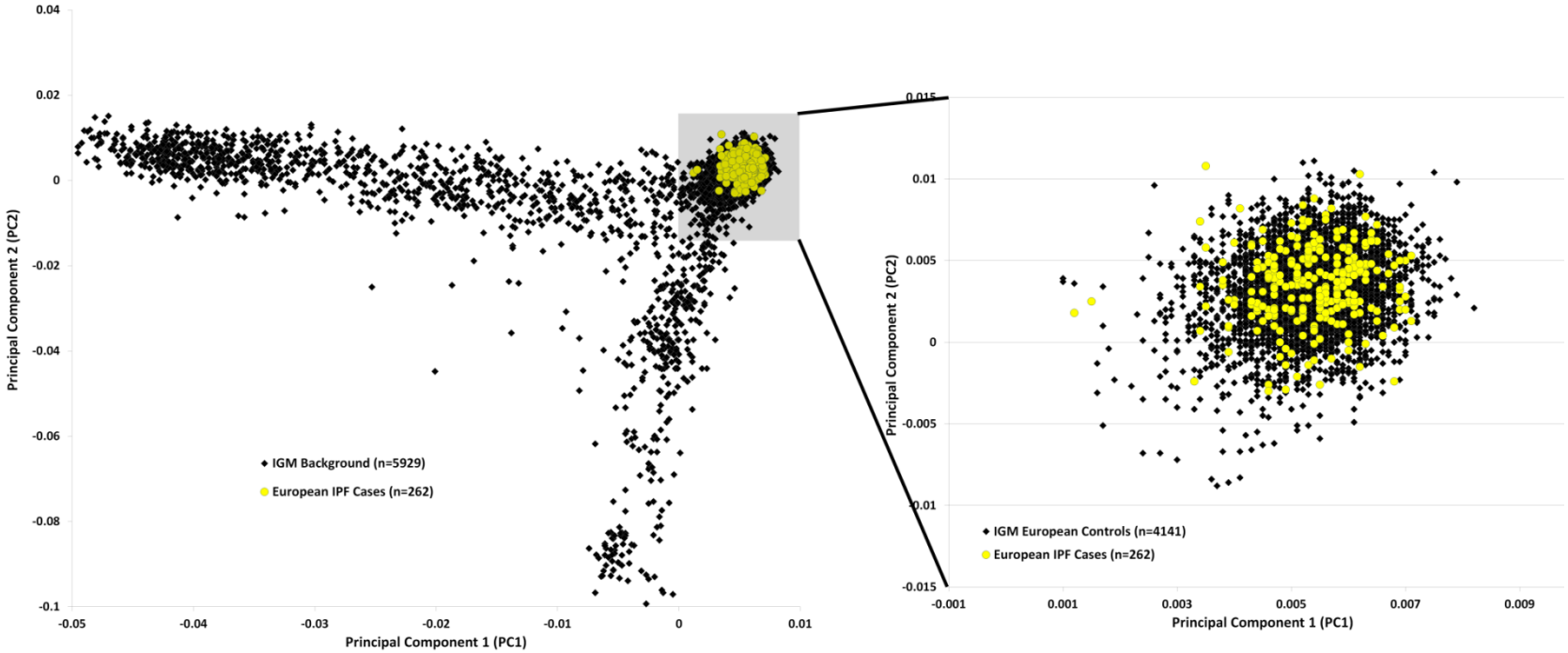
Step 3: Identify pairs with evidence for cryptic relatedness in test cohort; removing one from each pair to focus tests on unrelated index samples

Step 4: Run PCA on the common exome variation to predict genetic ancestry and identify population outliers.[1] Stringency can depend on genetic model of interest.

Petrovski S & Goldstein DB; Genome Bio 2016

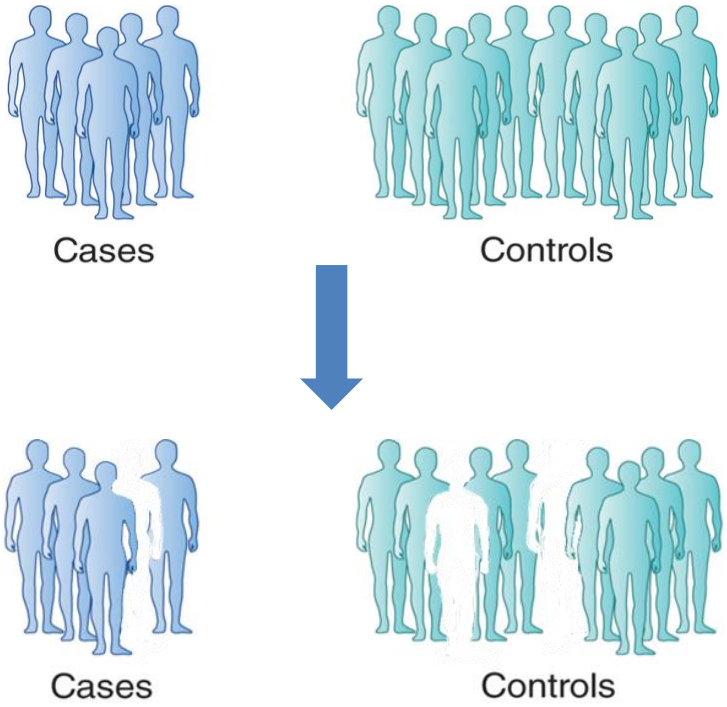
Cohort Design

Step 4 (cont.): Require the probability of being European > 0.95. Furthermore, samples required to be within 4SD of the Pr(European)>0.95 sample centroid.



Petrovski S, et al. AJRCCM (2017)

Cohort Design

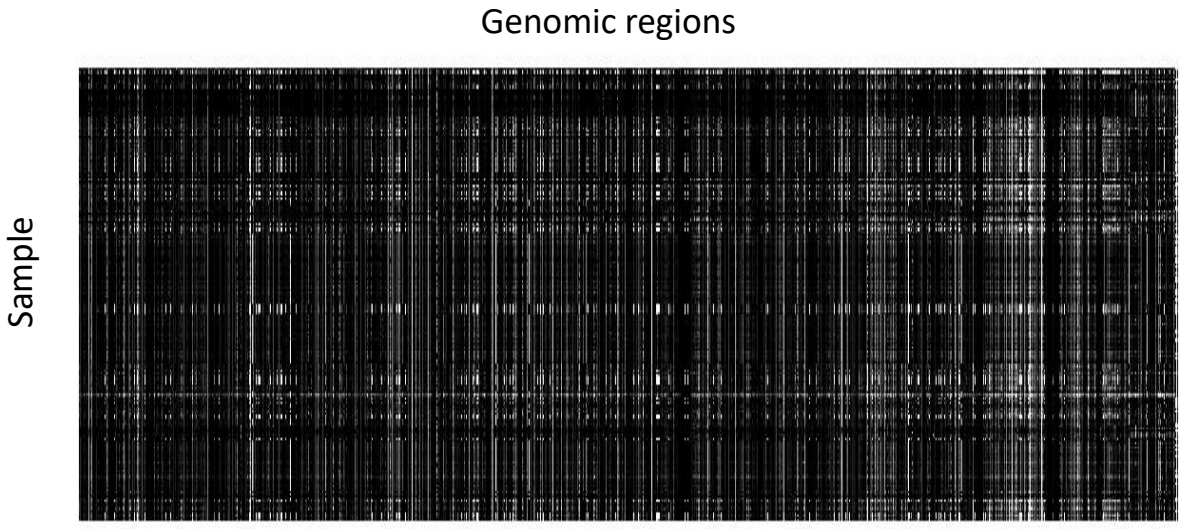


Status	Test Cohort	% Test Cohort
Initial Cohort	6331	100%
Contamination >2% based on VerifyBamID	6318	99.8%
Gender discordance between clinically-reported and X:Y coverage ratios	6315	99.7%
Autosomal Average Coverage <40-fold	6271	99.1%
<84.7% of CCDS 33.27M bases covered with ≥10-fold coverage	6250	98.7%
Cryptic Relatedness (KING and PLINK v1.07)	6218	98.2%
Self-declared Non-European	5114	80.8%
EIGENSTRAT multinomial Pr(European ancestry) <0.95	4486	70.9%
±4SD outside of PC 1 – 5 Pr(European ancestry) centroid	4403	69.5%
Final Test Cohort	4403	69.5%

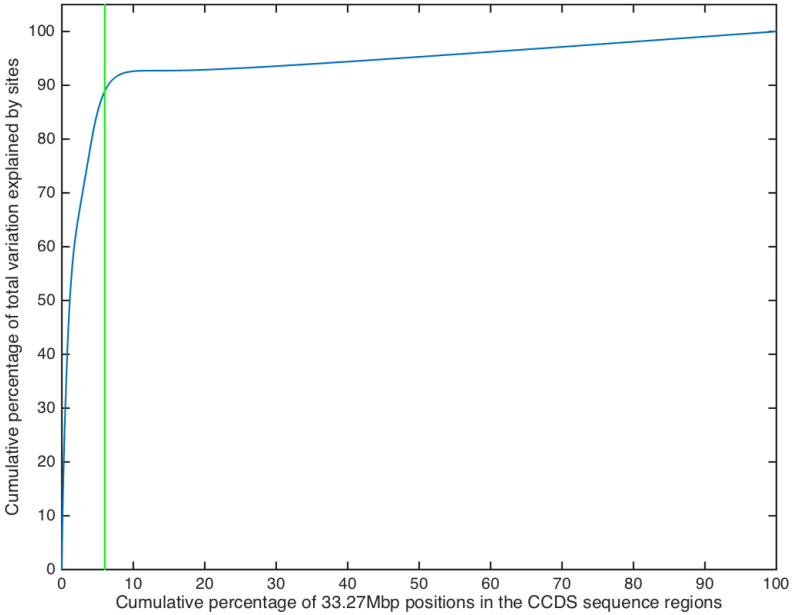
Petrovski S, et al. AJRCCM (2017)

Opportunity Bias

- Underlying issue: for a given gene, cases and controls can be imbalanced for their sequencing coverage ability to have called a variant. This can cause enrichment bias in one group.



Opportunity Bias

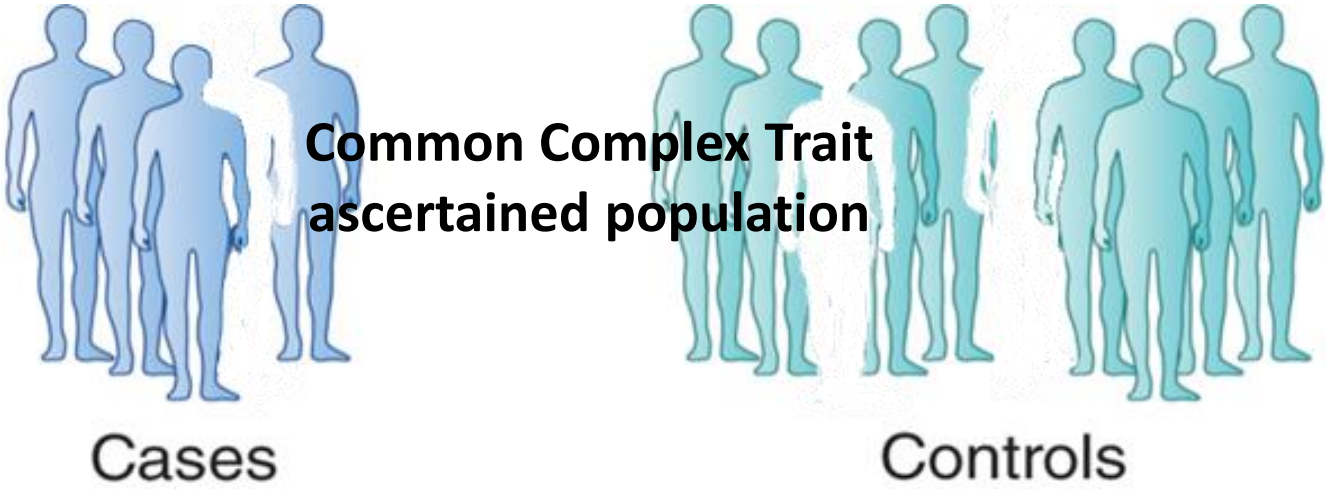


Y-axis = Cumulative sum of variation explained. Green line = point at which we maximize the amount of study-wide variation explained (here 89.1%) while minimizing the % of the exome that is pruned out (here 7.8%).

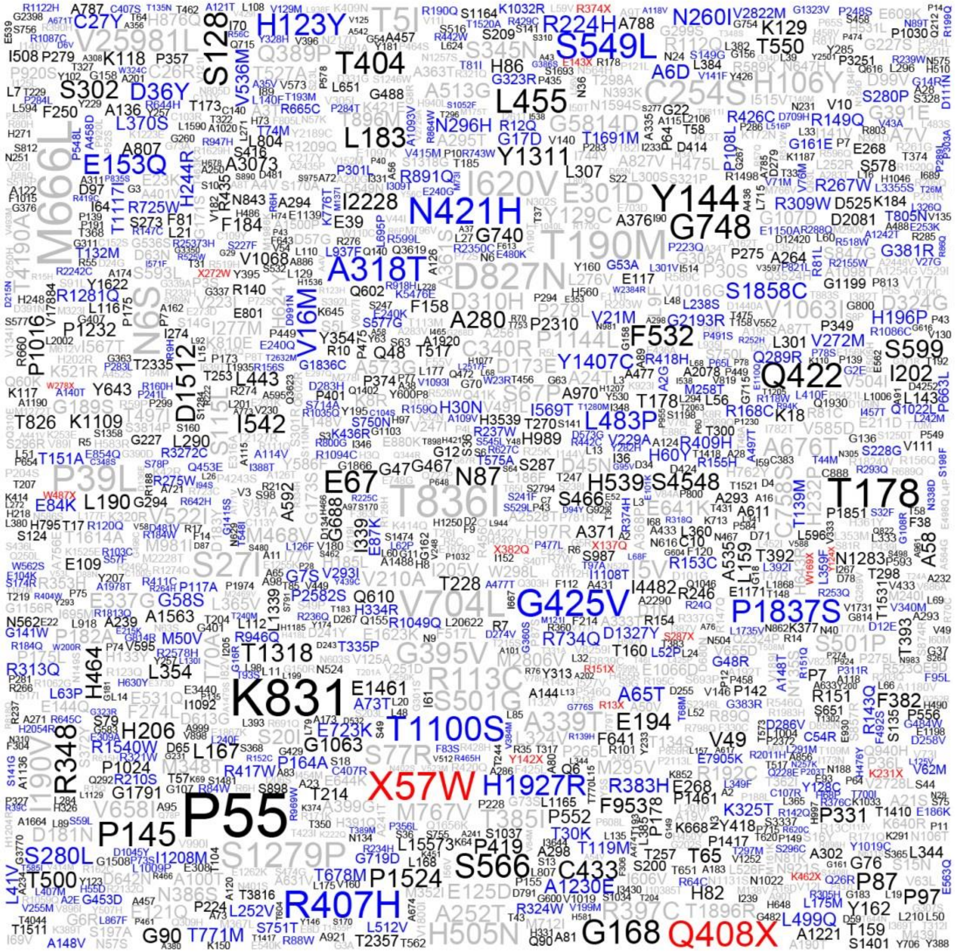
Post-pruned CCDS ≥10-fold coverage: Cases = 98.1%±0.3%. Controls = 97.9%±0.8% of sites.

Petrovski S, et al. AJRCCM (2017)

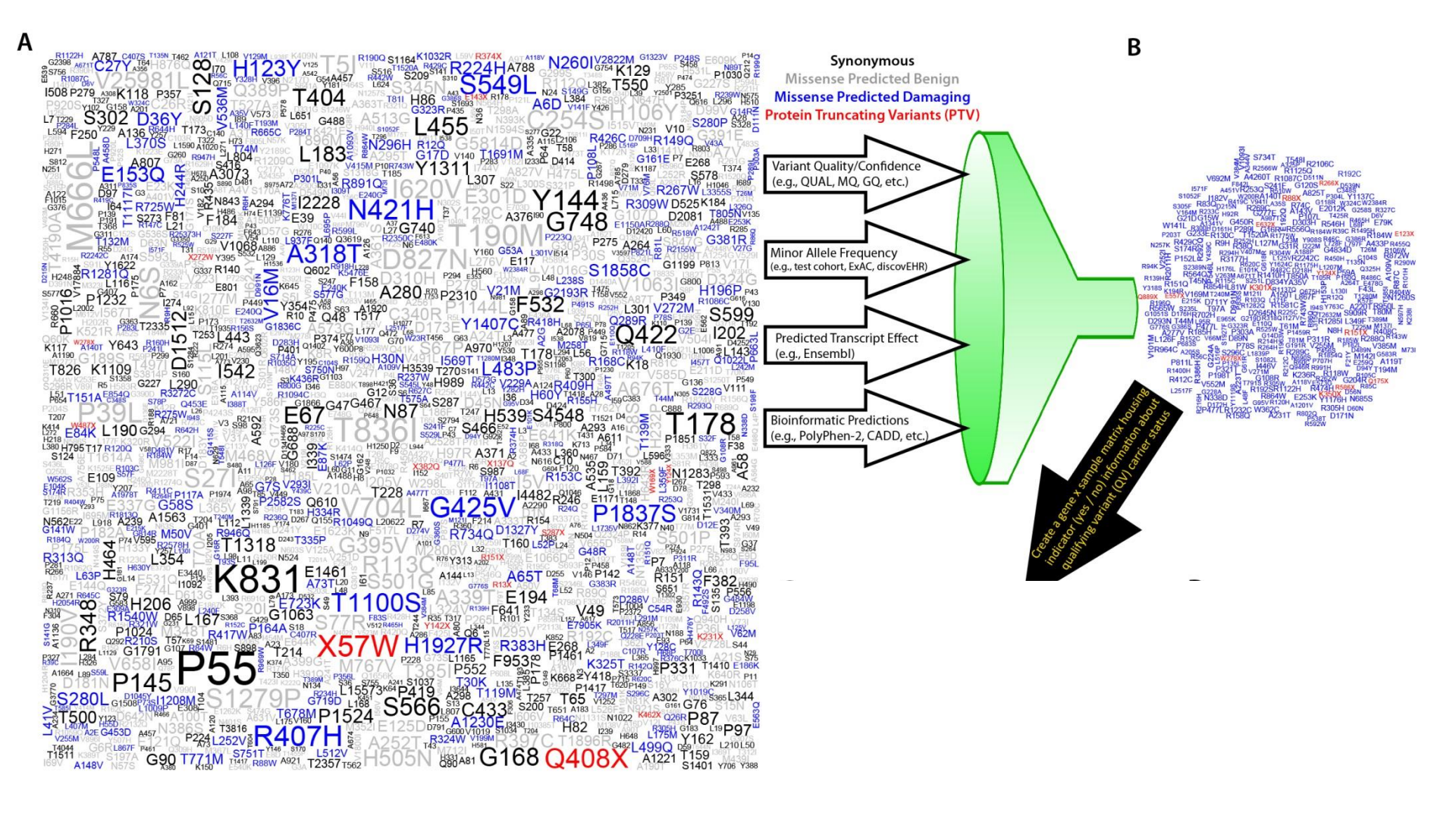
Collapsing Analyses

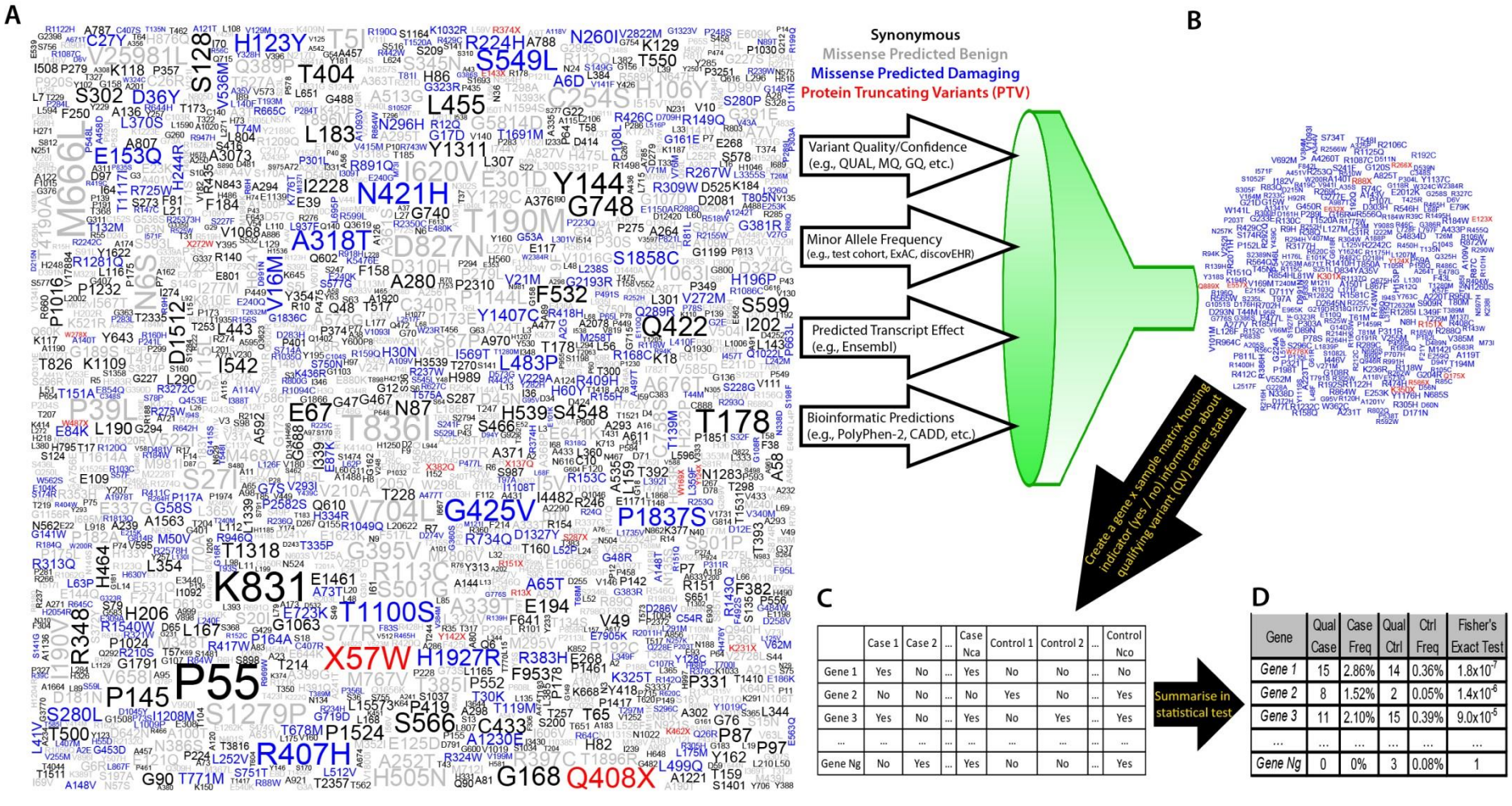


A



P55
K831
X57W
Q408X
S549L
N421H
A318T
T1100S
R407H
G425V
P1837S
H123Y
H1927R
E153Q
S566
T178
Q422
Y144
G748
L455
T404
L183
E67
F532
G168
P145
R348
S128
D36Y
S302
M666L
T190M
D827N
T836I
V704L
I620V
S1279P
H505N


A
B
Synonymous
Missense Predicted Benign
Missense Predicted Damaging
Protein Truncating Variants (PTV)
Variant Quality/Confidence
(e.g., QUAL, MQ, GQ, etc.)
Minor Allele Frequency
(e.g., test cohort, ExAC, discovEHR)
Predicted Transcript Effect
(e.g., Ensembl)
Bioinformatic Predictions
(e.g., PolyPhen-2, CADD, etc.)
Create a gene x sample matrix housing indicator (yes / no) information about qualifying variant (QV) carrier status



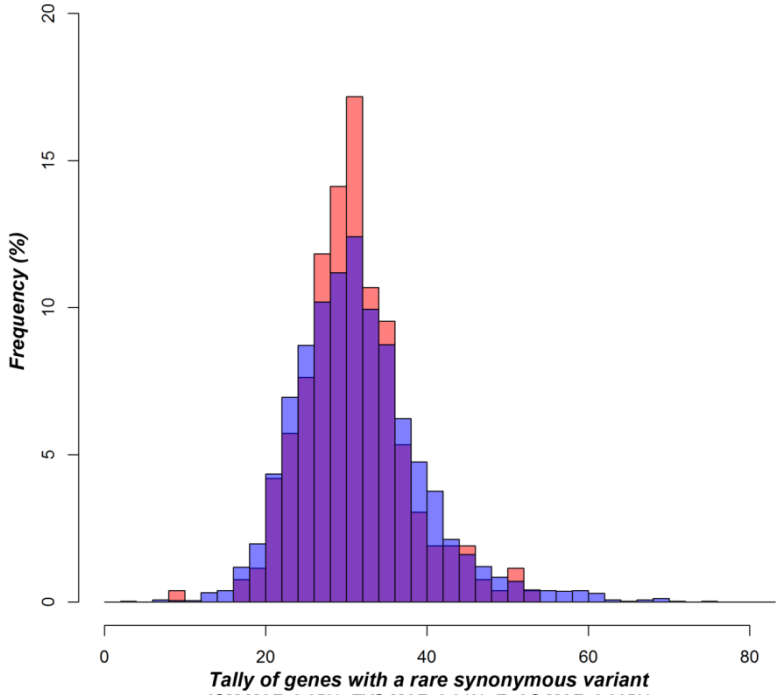
	Case 1	Case 2	...	Case Nca	Control 1	Control 2	...	Control Nco
Gene 1	Yes	No	...	Yes	No	No	...	No
Gene 2	No	No	...	No	Yes	No	...	Yes
Gene 3	Yes	No	...	Yes	No	Yes	...	Yes
...	...	...	...	...	...	...	...	...
Gene Ng	No	Yes	...	Yes	No	No	...	Yes

Gene	Qual Case	Case Freq	Qual Ctrl	Ctrl Freq	Fisher's Exact Test
Gene 1	15	2.86%	14	0.36%	1.8×10^{-7}
Gene 2	8	1.52%	2	0.05%	1.4×10^{-6}
Gene 3	11	2.10%	15	0.39%	9.0×10^{-5}
...	...	...	...	...	...
Gene Ng	0	0%	3	0.08%	1

Example “Qualifying Variants” Classes

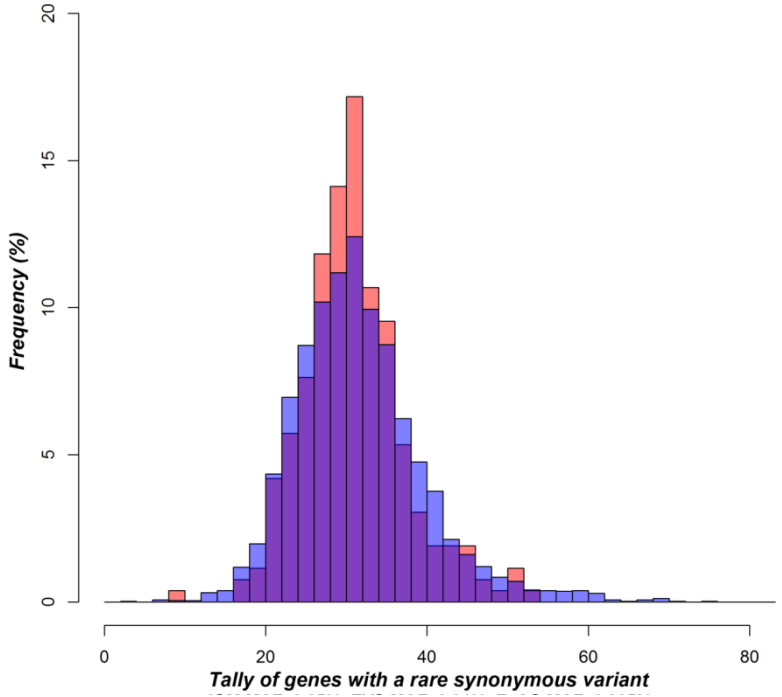
Model	Internal MAF(%)	External MAF(%)	Variant Effects
Ultra-rare (Primary)	0.05%	0%	PTV and PolyPhen-2 “probably”
PTV (LoF)	0.1%	0.1%	PTV (LoF)
Rare non-syn (MAF<0.1%)	0.1%	0.1%	PTV and missense
Neutral (Ultra-rare)	0.05%	0%	Synonymous


No Go
Go

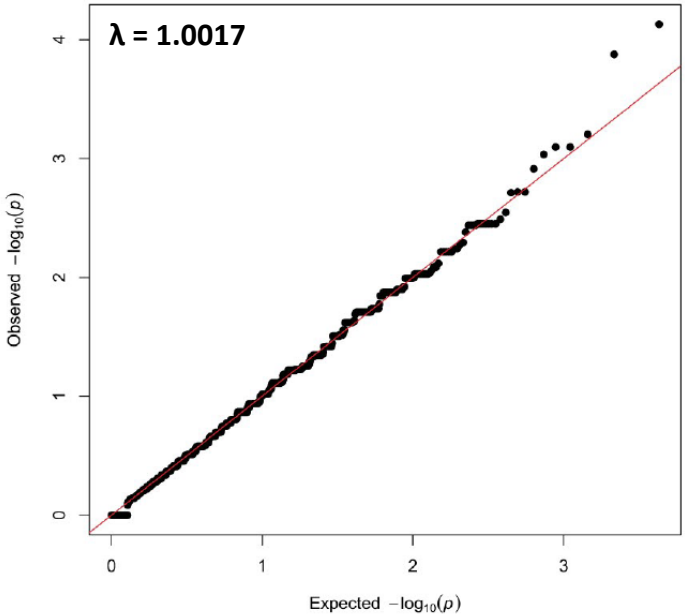


*PF case cohort (red; average 30.7±6.5 qualifying genes) to the control cohort (blue; average 31.2±7.9 qualifying genes), (Mann-Whitney U test, **p=0.68**).*

Petrovski S, et al. AJRCCM (2017)



*PF case cohort (red; average 30.7±6.5 qualifying genes) to the control cohort (blue; average 31.2±7.9 qualifying genes), (Mann-Whitney U test, **p=0.68**).*



S4K: Pulmonary Fibrosis Rare Synonymous (Neutral) Model QQ-Plot

***QQPerm**: https://cran.r-project.org/package=QQperm*
Permutation QQ plots reflecting the empirical NULL distribution

Petrovski S, et al. AJRCCM (2017)

Collapsing Analyses

Some Published Applications (to date):

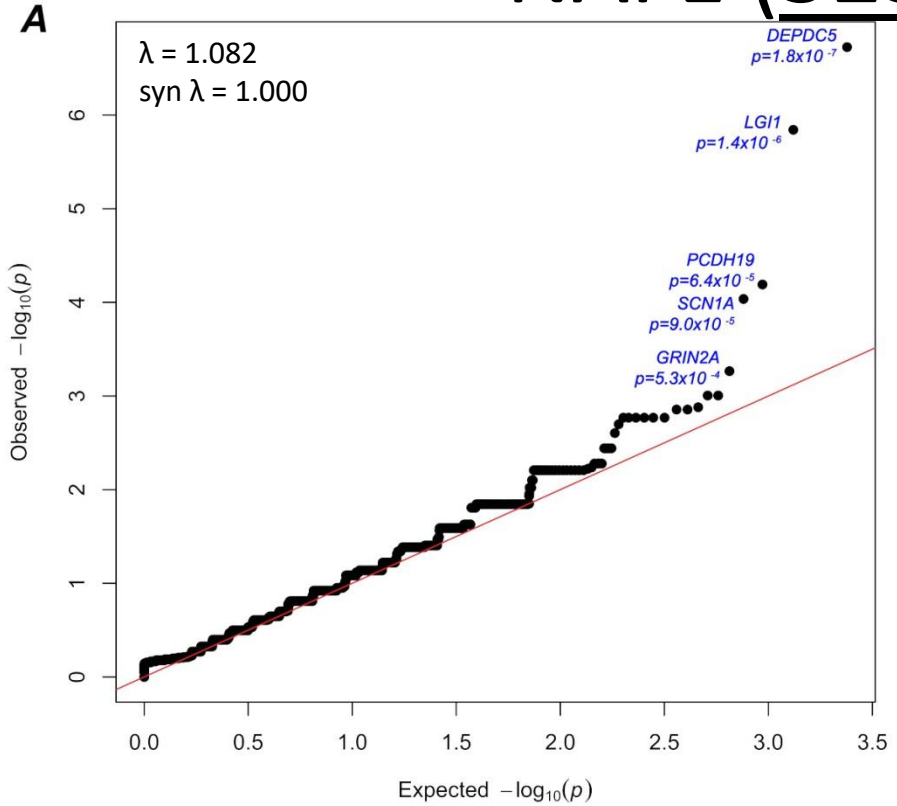
Amyotrophic Lateral Sclerosis	-	Cirulli E, Lasseigne B, Petrovski S, et al. *Science* 2015
Genetic Generalized Epilepsy*	-	EPI4K Consortium. *Lancet Neurology* 2017
Idiopathic Pulmonary Fibrosis	-	Petrovski S, Todd J, Durheim M, et al. *AJRCCM* 2017
Non-acquired Focal Epilepsy*	-	EPI4K Consortium. *Lancet Neurology* 2017
Sudden Unexplained Death	-	Bagnall R, Crompton D, Petrovski S, et al. *Annals of Neurology* 2016

Collapsing analyses of the common complex epilepsies (*familial ascertainment*)

Publication:

Ultra-rare genetic variation in common epilepsies: a case-control sequencing study
Epi4K Consortium. The Lancet Neurology (2017); 16 (2), 135-143

NAFE (525 vs 3,877)



HGNC	RVIS%	Qual Case	Case Freq	Qual Ctrl	Ctrl Freq	FET p- value
DEPDC5	6.7%	15	2.86%	14	0.36%	**1.82E-07**
LGI1	8.8%	8	1.52%	2	0.05%	**1.41E-06**
PCDH19	5.3%	6	1.14%	2	0.05%	**6.35E-05**
SCN1A	2.4%	11	2.10%	15	0.39%	**8.99E-05**
GRIN2A	1.2%	7	1.33%	7	0.18%	**5.33E-04**
TYRO3	10.6%	5	0.95%	3	0.08%	**9.74E-04**
LMAN1L	78.1%	5	0.95%	3	0.08%	**9.74E-04**
PKHD1	67.4%	10	1.90%	19	0.49%	**0.0013**
ATP8B1	39.3%	6	1.14%	6	0.15%	**0.0014**
PCDHB6	98.5%	6	1.14%	6	0.15%	**0.0014**

Summary:

Likelihood of getting five of 43 known genes occupy genome-wide ranks [1-5] of ~18K tested genes, ***$p=5.7x10^{-14}$***

QV in one of these five epilepsy genes contributes to disease risk in ~8% of cases with OR 13.2 [95%CI 8.0 – 22.1].

Epi4K Consortium. Lancet Neurology (2017)

Population Reference cohort resolution:
What minor allele frequencies (MAF) are we able to estimate?

Cohort:	EVS
Sample:	6,503
MAF res.:	<0.008%

gnomAD – 141,352 population reference cohort

http://gnomad.broadinstitute.org/

Population Reference cohort resolution:

What minor allele frequencies (MAF) are we able to estimate?

Cohort:	EVS	->	ExAC
Sample:	6,503	->	60,706
MAF res.:	<0.008%	->	<0.0008%

gnomAD – 141,352 population reference cohort

http://gnomad.broadinstitute.org/

Population Reference cohort resolution:

What minor allele frequencies (MAF) are we able to estimate?

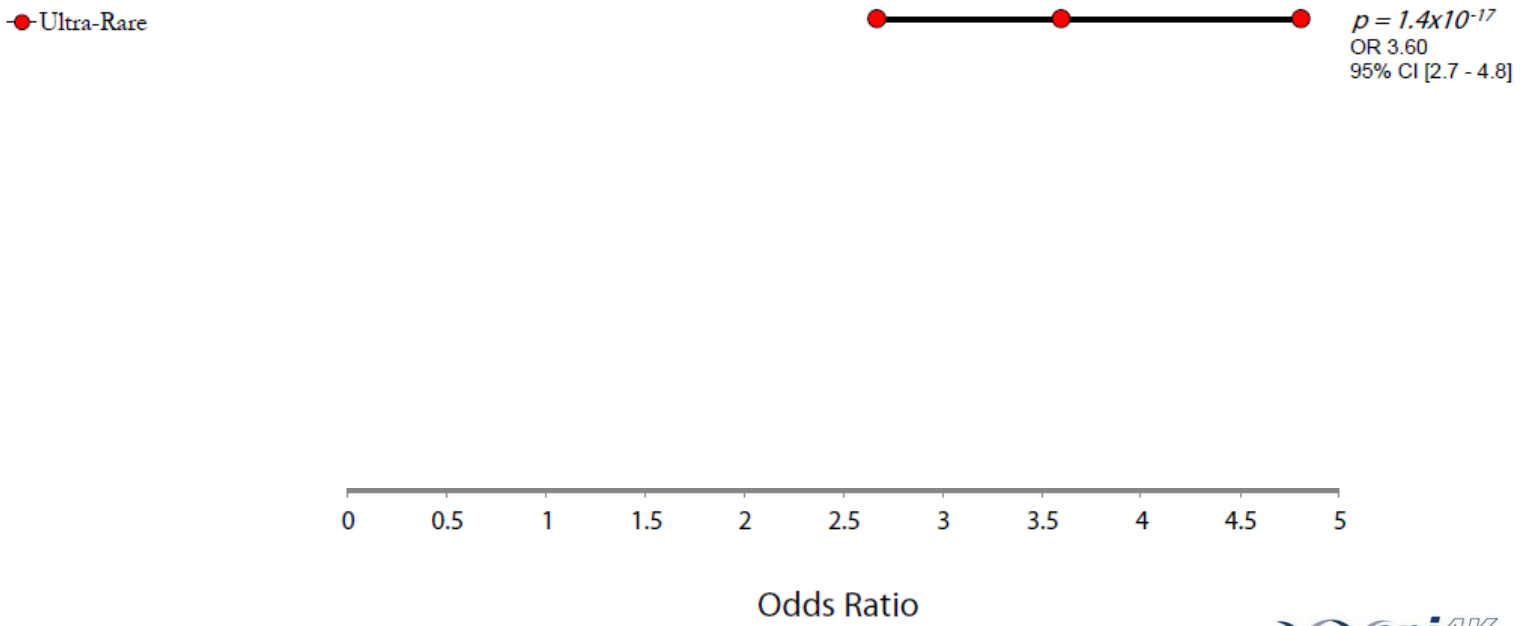
Cohort:	EVS	->	ExAC	->	gnomAD
Sample:	6,503	->	60,706	->	141,352
MAF res.:	<0.008%	->	<0.0008%	->	<0.0004%

gnomAD – 141,352 population reference cohort

http://gnomad.broadinstitute.org/

NAFE Architecture: Comparing relative contribution of rare allele frequencies

Based on the enrichment of variants among dominant epilepsy genes



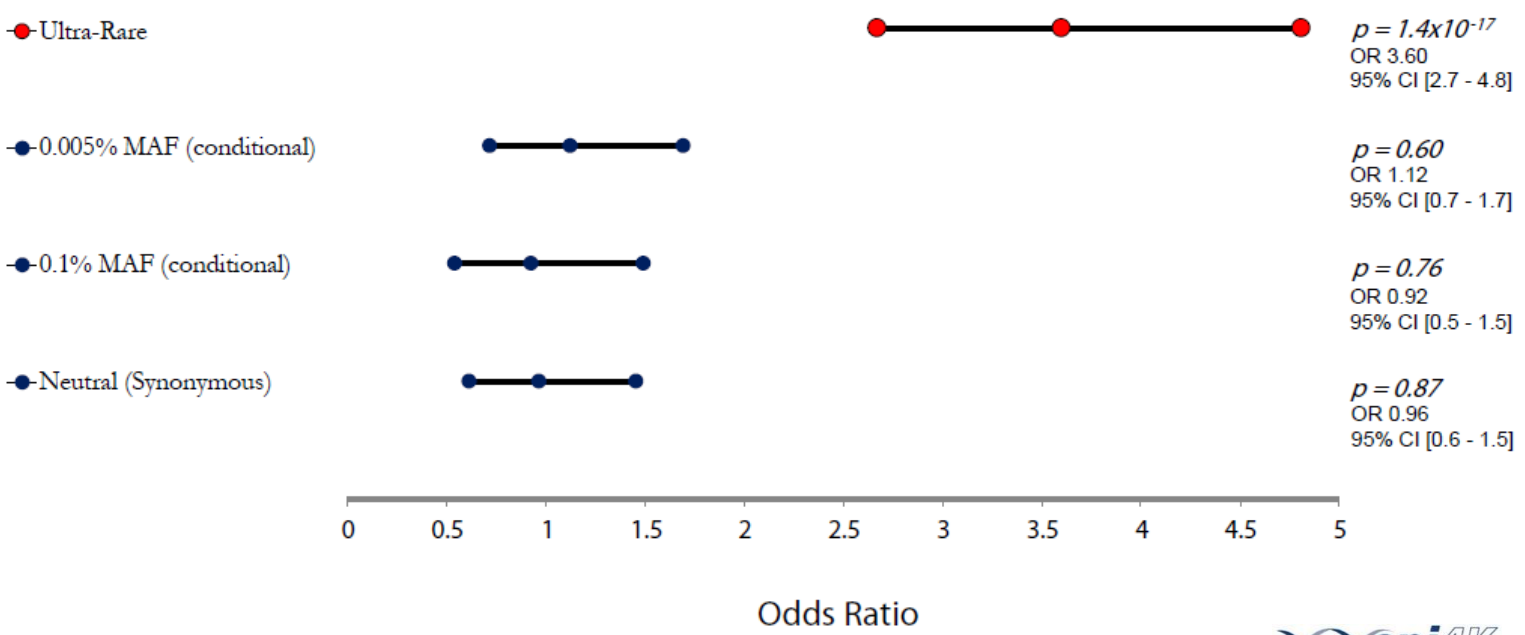
Ultra-rare* : MAF ≤0.05% among combined test population, while absent (MAF=0**) in both EVS and ExAC reference cohorts.*

epi4K GENE DISCOVERY IN EPILEPSY

Epi4K Consortium. Lancet Neurology (2017)

NAFE Architecture: Comparing relative contribution of rare allele frequencies

Based on the enrichment of variants among dominant epilepsy genes



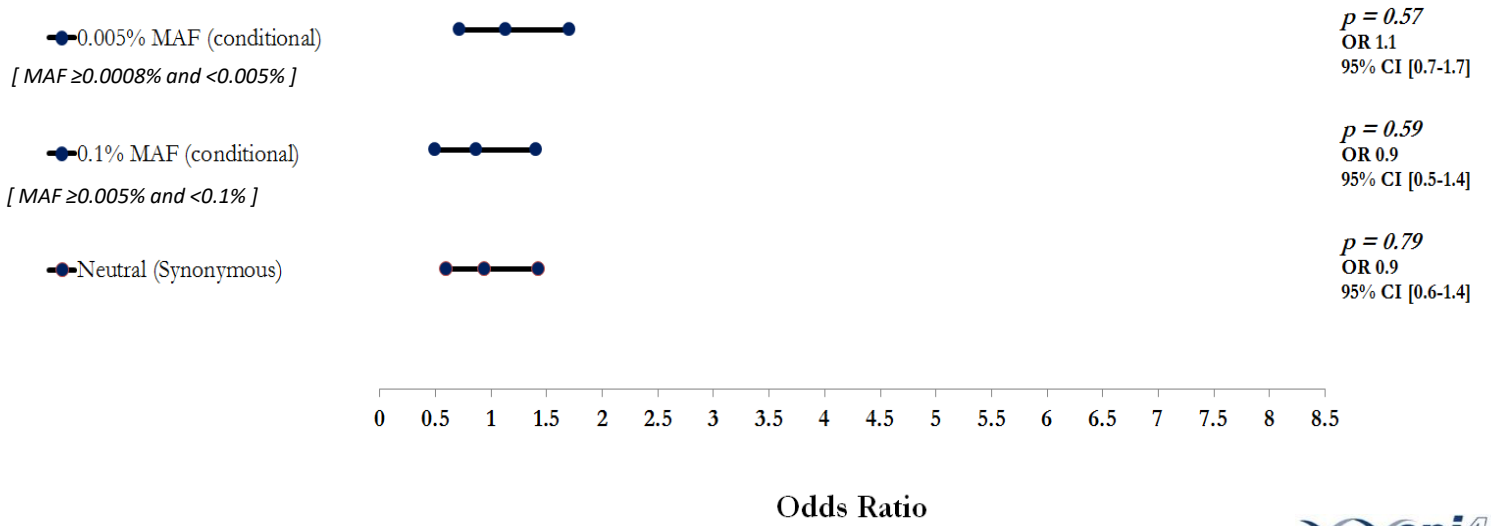
Ultra-rare* : MAF ≤0.05% among combined test population, while absent (MAF=0**) in both EVS and ExAC reference cohorts.*

epi4K GENE DISCOVERY IN EPILEPSY

Epi4K Consortium. Lancet Neurology (2017)

NAFE Architecture: Comparing relative contribution of rare allele frequencies

Based on the enrichment of variants among dominant epilepsy genes



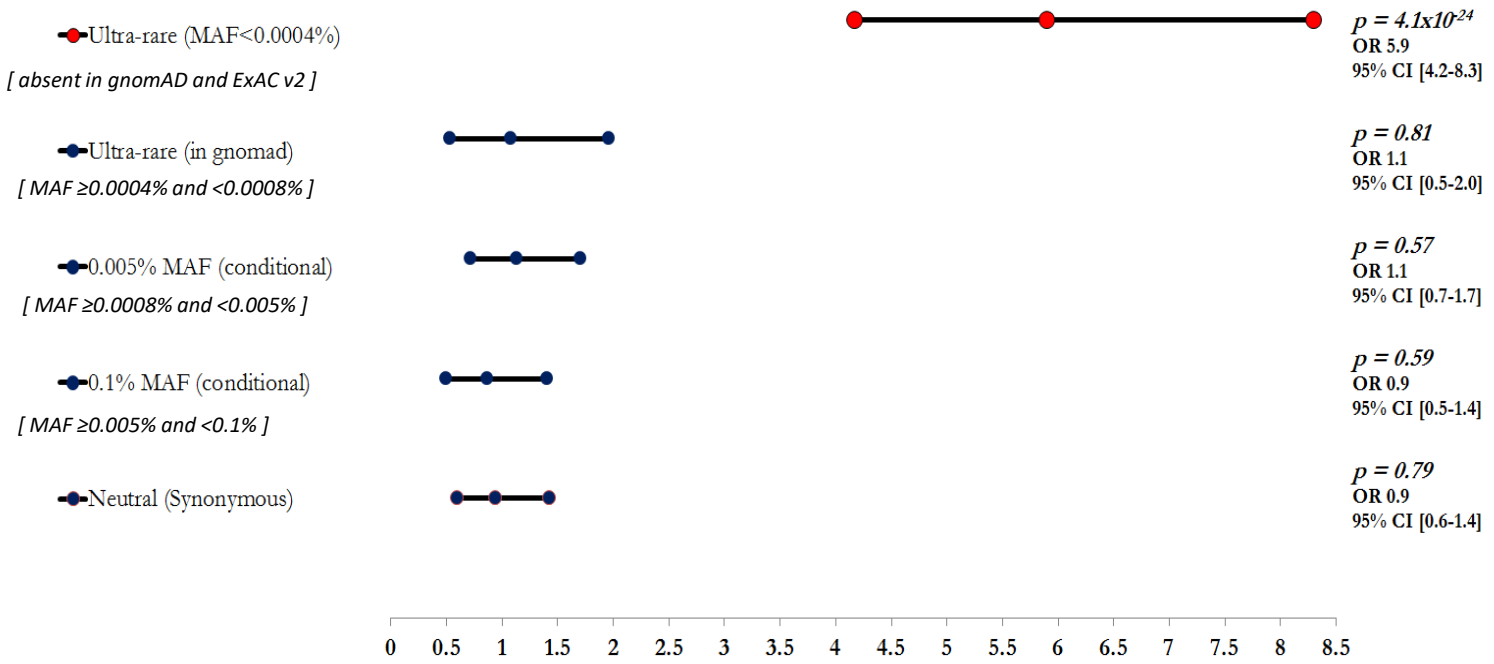
Ultra-rare* : MAF ≤0.05% among combined test population, while absent (*MAF=0***) in both EVS and ExAC reference cohorts.

epi4K GENE DISCOVERY IN EPILEPSY

Epi4K Consortium. Lancet Neurology (2017)

NAFE Architecture: Comparing relative contribution of rare allele frequencies

Based on the enrichment of variants among dominant epilepsy genes



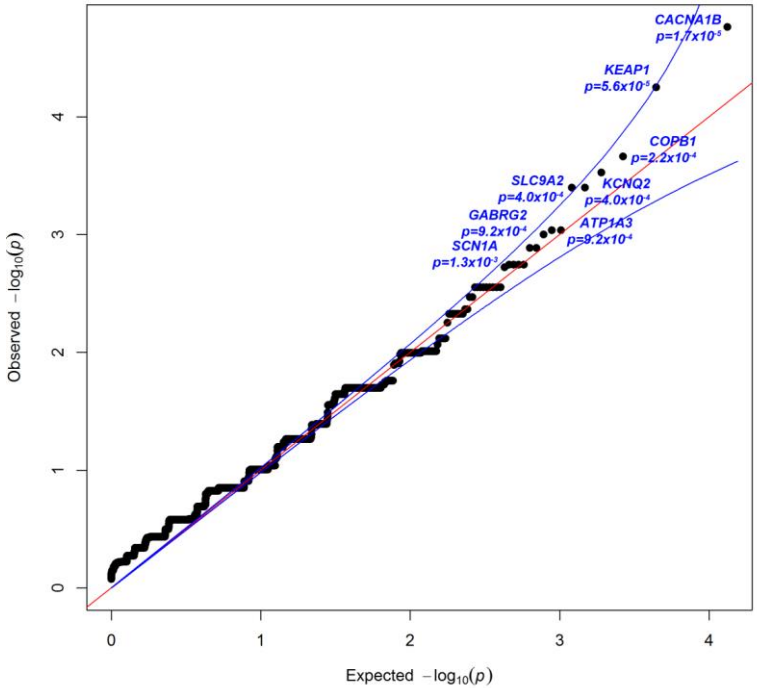
Ultra-rare : MAF ≤0.05% among combined test population, while absent (MAF=0**) in both EVS and ExAC reference cohorts.*

epi4K GENE DISCOVERY IN EPILEPSY

Epi4K Consortium. Lancet Neurology (2017)

Do patients with epilepsy have more 'qualifying variants' in gene X than general controls?

GGE (640 vs 3,877)



Qualifying variant:

High confidence variant call

LoF / Polyphen "Probably" prediction

Ultra-rare and absent among EVS and ExAC (i.e., ~ <0.0008% MAF)

HGNC	RVIS%	Qual Case	Case Freq	Qual Ctrl	Ctrl Freq	FET p-value
CACNA1B	0.8%	8	1.25%	3	0.08%	**1.73E-05**
KEAP1	8.8%	5	0.78%	0	0%	**5.63E-05**
COPB1	24.9%	7	1.09%	4	0.10%	**2.18E-04**
PHTF1	32.5%	5	0.78%	1	0.03%	**2.98E-04**
KCNQ2	5.9%	4	0.62%	0	0%	**4.00E-04**
SLC9A2	4.0%	4	0.62%	0	0%	**4.00E-04**
ATP1A3	2.2%	5	0.78%	2	0.05%	**9.22E-04**
GABRG2	10.5%	5	0.78%	2	0.05%	**9.22E-04**
ZNF100	69.2%	6	0.94%	4	0.10%	**9.99E-04**
CUX1	2.3%	9	1.41%	12	0.31%	**0.0013**
SCN1A	2.4%	10	1.56%	15	0.39%	**0.0013**
ARNT2	5.5%	4	0.62%	1	0.03%	**0.0018**

Summary:

No single gene is genome-wide significant:

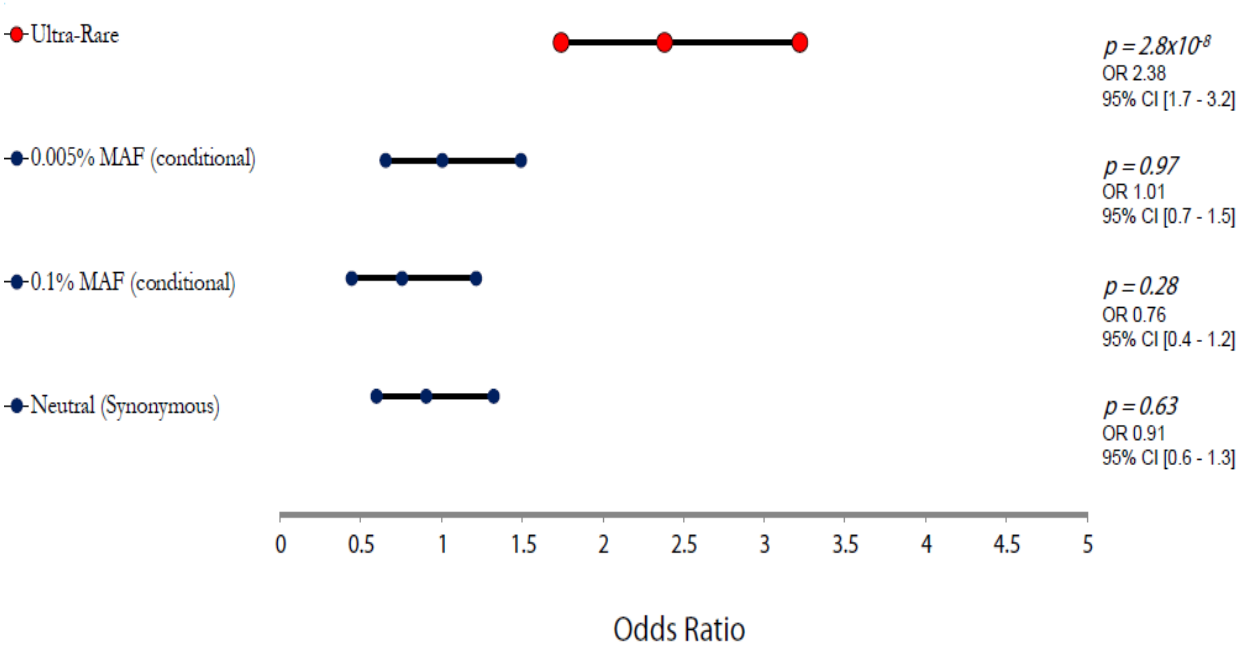
adjusted alpha $p=2x10^{-6}$

epi4K GENE DISCOVERY IN EPILEPSY

Epi4K Consortium. Lancet Neurology (2017)

GGE Architecture: Comparing relative contribution of rare allele frequencies

Based on the enrichment of variants among dominant epilepsy genes



Ultra-rare* : *MAF ≤0.05% among combined test population, while absent (MAF=0**) in both EVS and ExAC reference cohorts.*

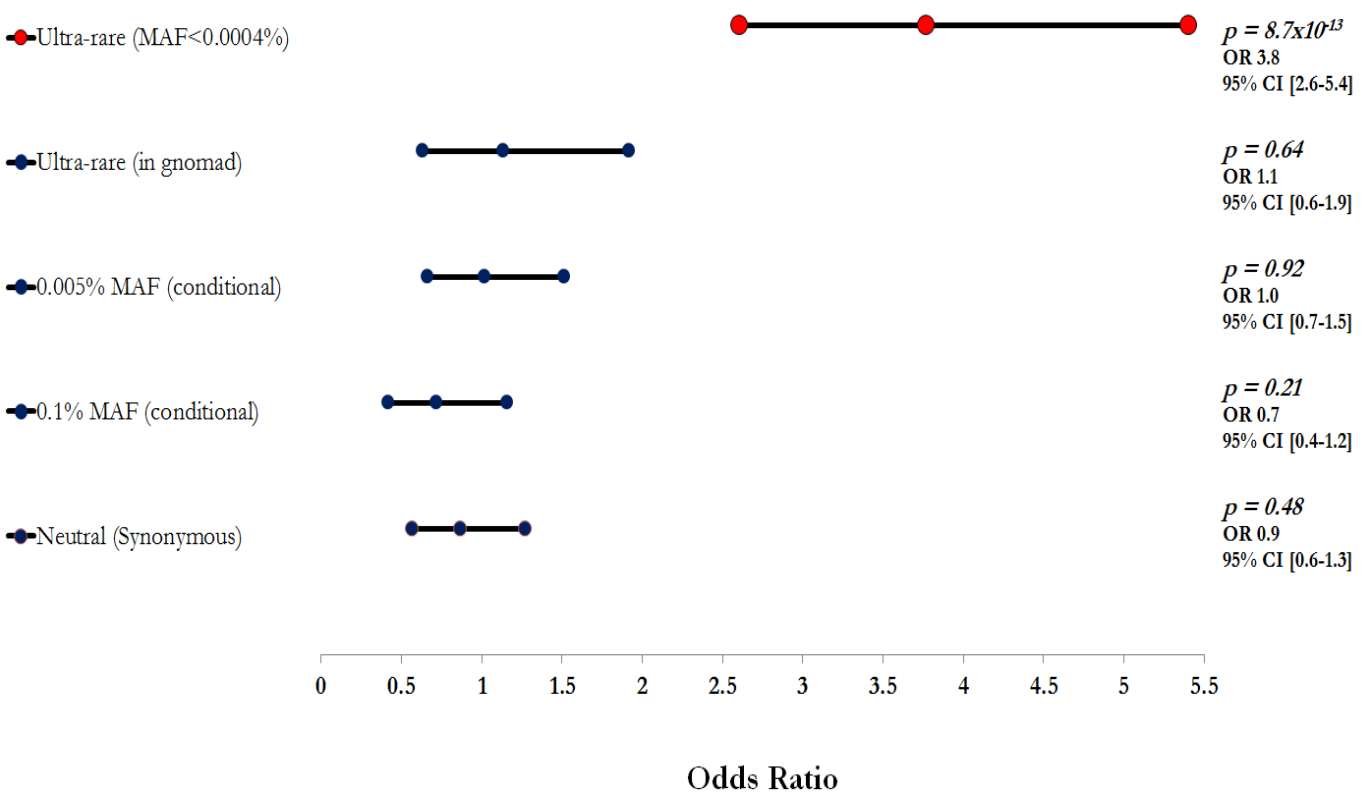
COLUMBIA UNIVERSITY
College of Physicians and Surgeons

IGM Institute for Genomic Medicine

epi4K
GENE DISCOVERY IN EPILEPSY

GGE Architecture: Comparing relative contribution of rare allele frequencies

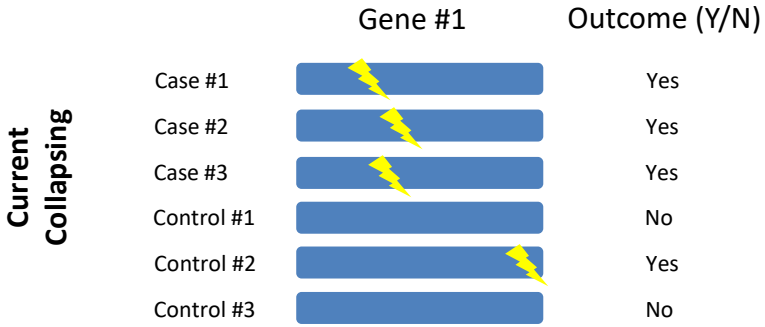
Based on the enrichment of variants among dominant epilepsy genes



Ultra-rare* : *MAF ≤0.05% among combined test population, while absent (MAF=0**) in both EVS and ExAC reference cohorts.*

COLUMBIA UNIVERSITY College of Physicians and Surgeons | IGM Institute for Genomic Medicine | epi4K GENE DISCOVERY IN EPILEPSY

Mega-Gene Burden



Regression correcting for: gender, exome-wide CCDS coverage, exome-wide average read depth and ultra-rare synonymous rate in corresponding mega-gene. Permutation-based implementation supported.

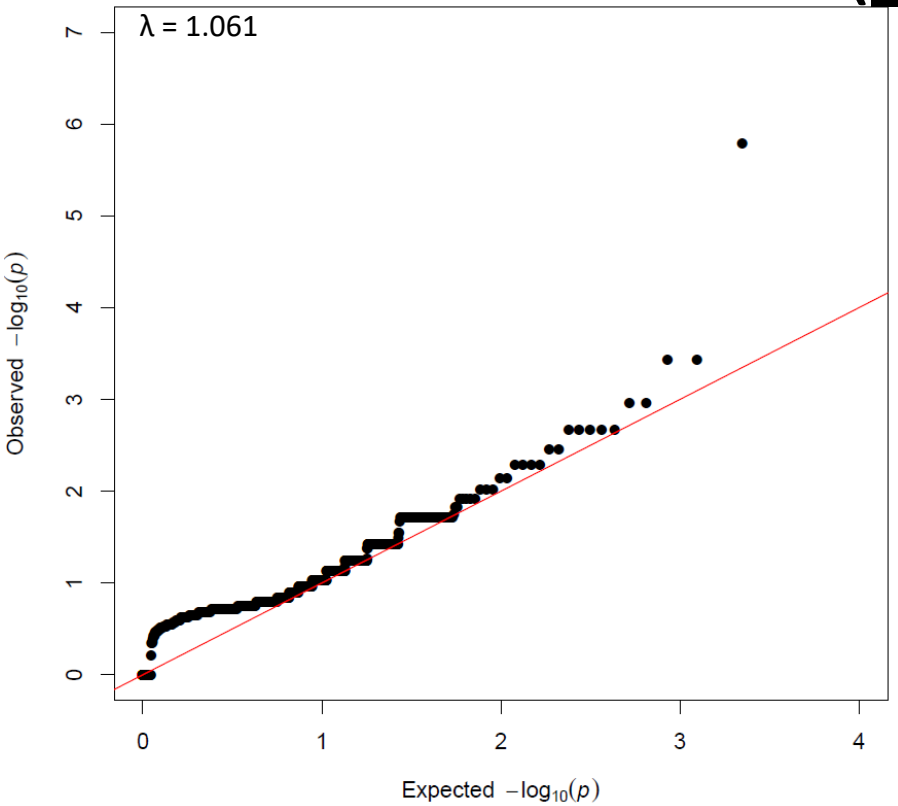
Mega-Gene Analysis

Gene set	Number of genes	Average qualifying variants[a]	Qualifying variants enrichment p-value (Odds Ratio [95% CI])	Neutral variation enrichment p-value	Enrichment after removing the 43 epilepsy genes p-value
Known	43	0.052	$p = 9.1x10^{-8}$ (OR=2.3 [95% CI 1.7 - 3.2])	p = 0.86	N/A
Known (EE)	33	0.037	$p = 2.6x10^{-7}$ (OR=2.6 [95% CI 1.8 - 3.6])	p = 0.34	N/A
Ion Channel	209	0.264	p = 0.028 (OR=1.2 [95% CI 1.0 - 1.5])	p = 0.73	p = 0.21
FMRP	823	1.481	p = 0.034 (OR=1.3 [95% CI 1.0 - 1.6])	p = 0.94	p = 0.04
NMDAR & ARC	78	0.067	p = 0.004 (OR=1.6 [95% CI 1.1 - 2.1])	p = 0.80	p = 0.007
MGI Seizure	235	0.269	p = 0.003 (OR=1.3 [95% CI 1.1 - 1.6])	p = 0.97	p = 0.17

[a]Average number of qualifying variants in the corresponding gene set, per sample in the test population

Epi4K Consortium. Lancet Neurology (2017)

SUDEP (58 vs 2,936)



Rank	HGNC	Case Freq	Ctrl Freq	FET P
1	*DEPDC5*	8.6%	0.2%	1.6x10-6
2	*RSPO2*	3.5%	0%	3.7x10-4
3	*NFE2L2*	3.5%	0%	3.7x10-4
…	…	…	…	…
15	*SCN2A*	3.5%	0.1%	0.005
…	…	…	…	…
17	*KCNH2*	3.5%	0.2%	0.007

Both *SCN2A* confirmed *de novo* mutations through trio-based Sanger validation.

Both *KCNH2* variants previously reported as pathogenic in unrelated samples for Long QT syndrome.

Bagnall R, et al. Annals of Neurology (2016)

Missense Tolerance Ratio (MTR)

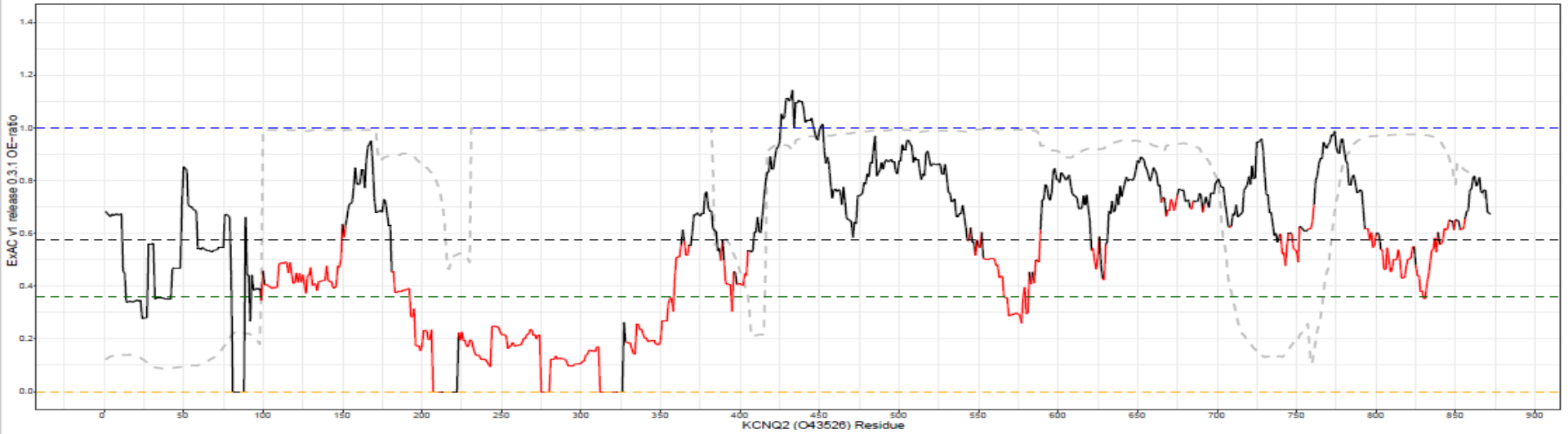
- Use sequence context to estimate a regions **expected proportion** of non-synonymous variation, taking into account the underlying mutation rate.

- Using gnomAD reference cohort extract the **observed proportion** of non-synonymous variation.

- Take the ratio of Observed over Expected proportions (MTR) as a metric to quantify the departure of the observed from the expected proportion of non-synonymous variants in a given coding region.

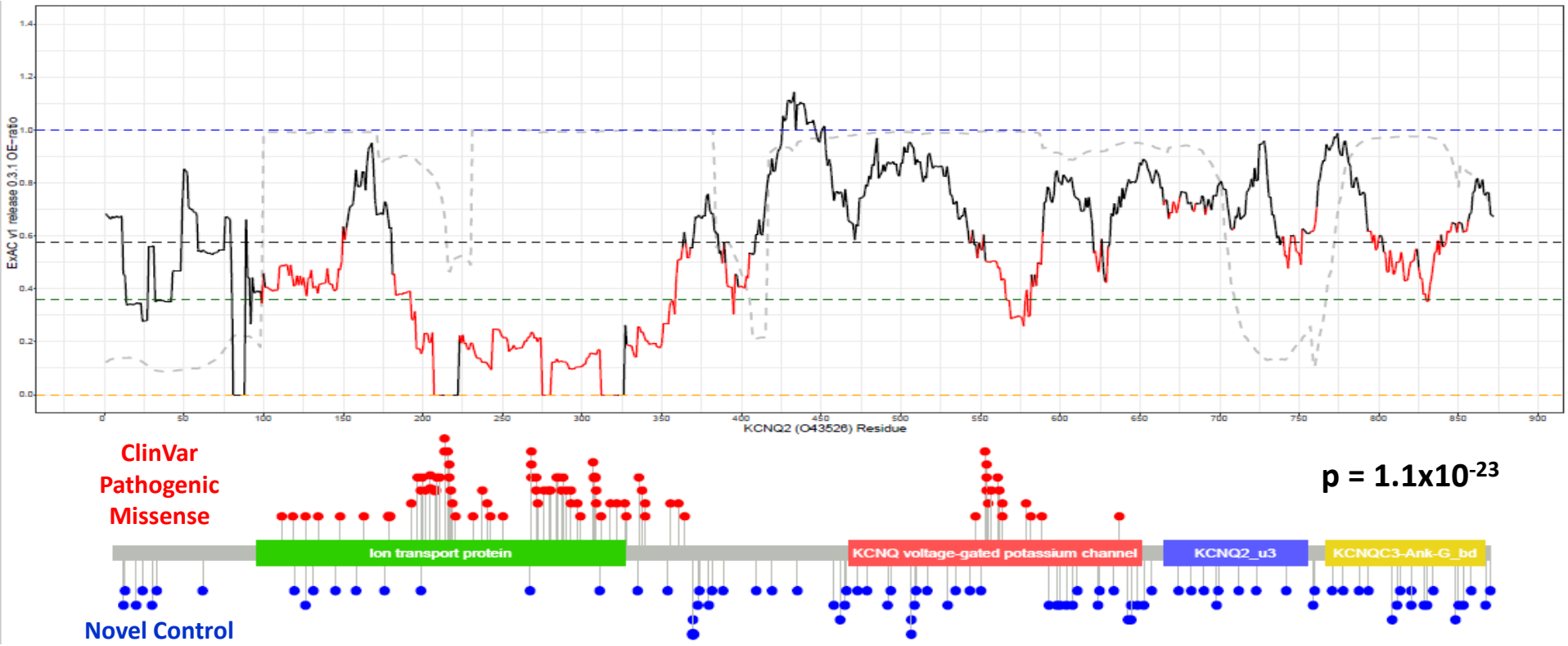
Traynelis et al. Genome Research (2017)

KCNQ2 example...



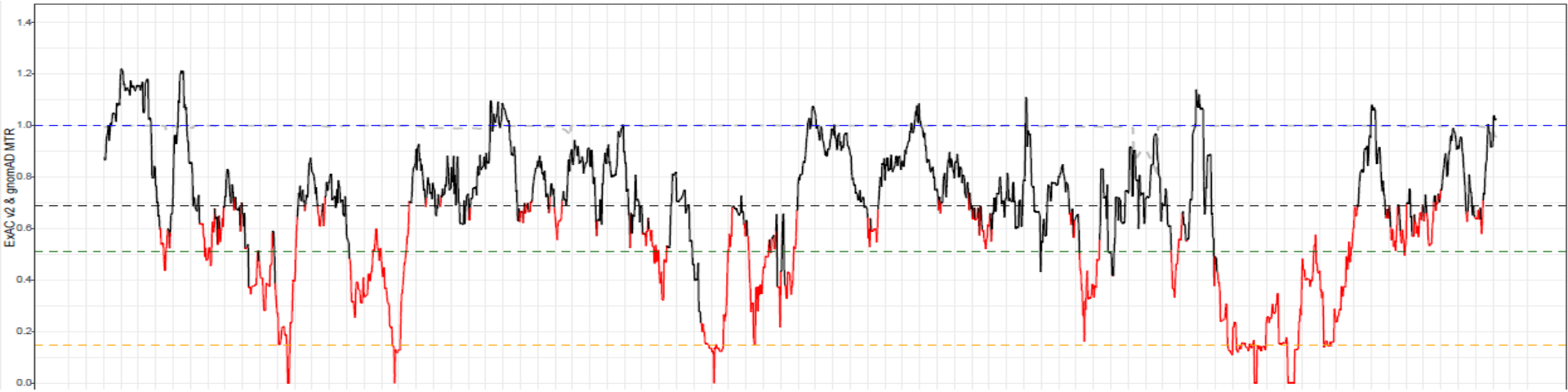
Traynelis et al. Genome Research (2017)

KCNQ2 example...



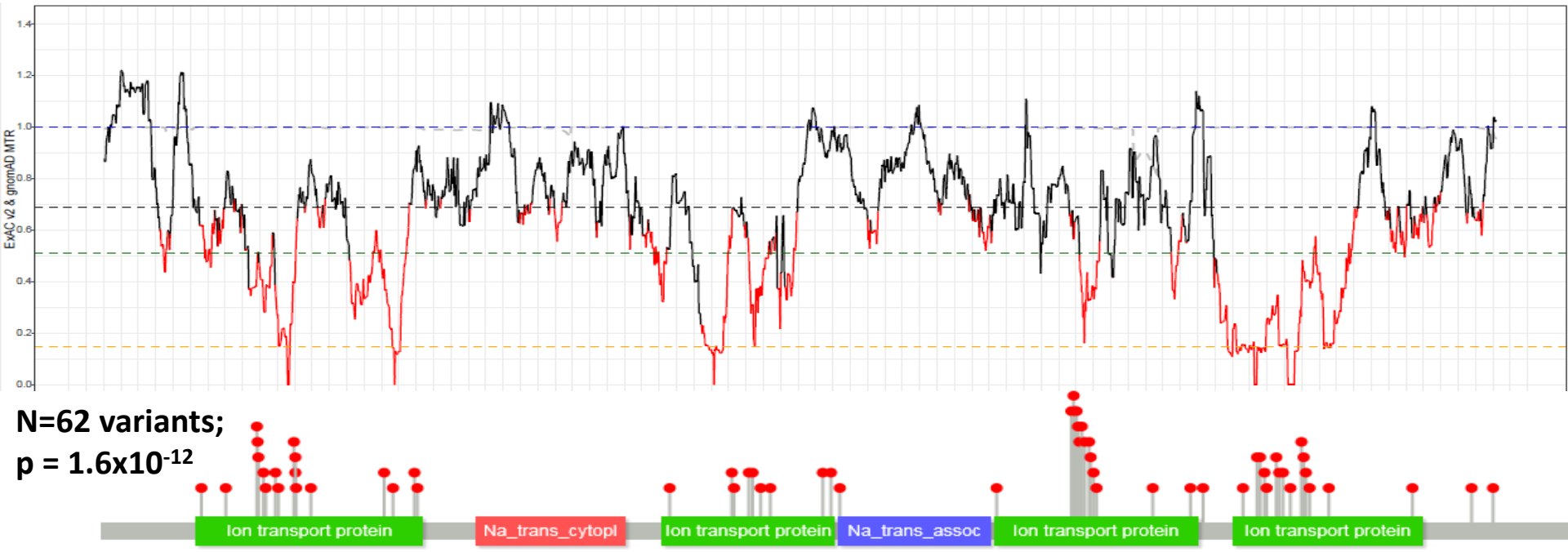
Traynelis et al. Genome Research (2017)

SCN2A Missense Tolerance Ratio (MTR)



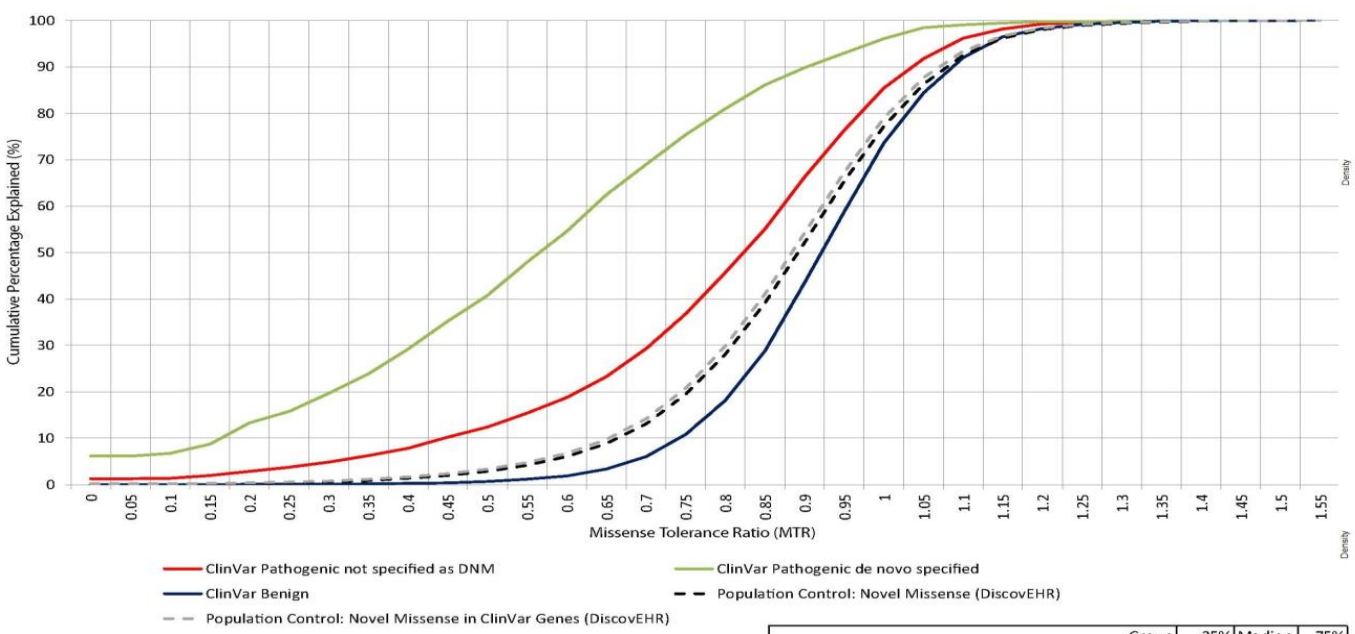
Traynelis et al. Genome Research (2017)

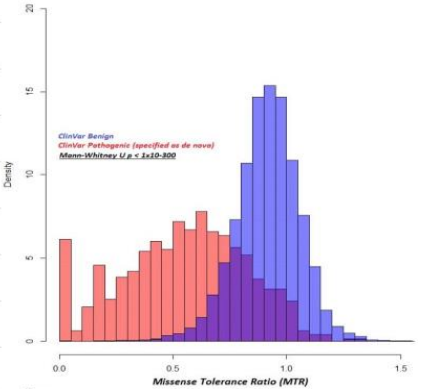
SCN2A Missense Tolerance Ratio (MTR)

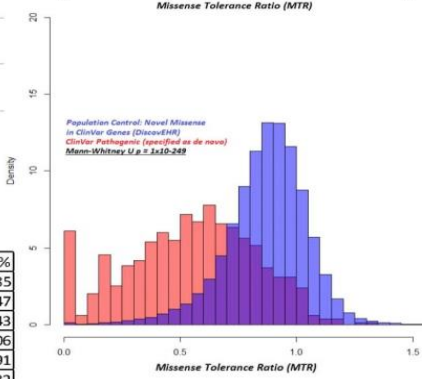


Traynelis et al. Genome Research (2017)

Genic regional intolerance: MTR





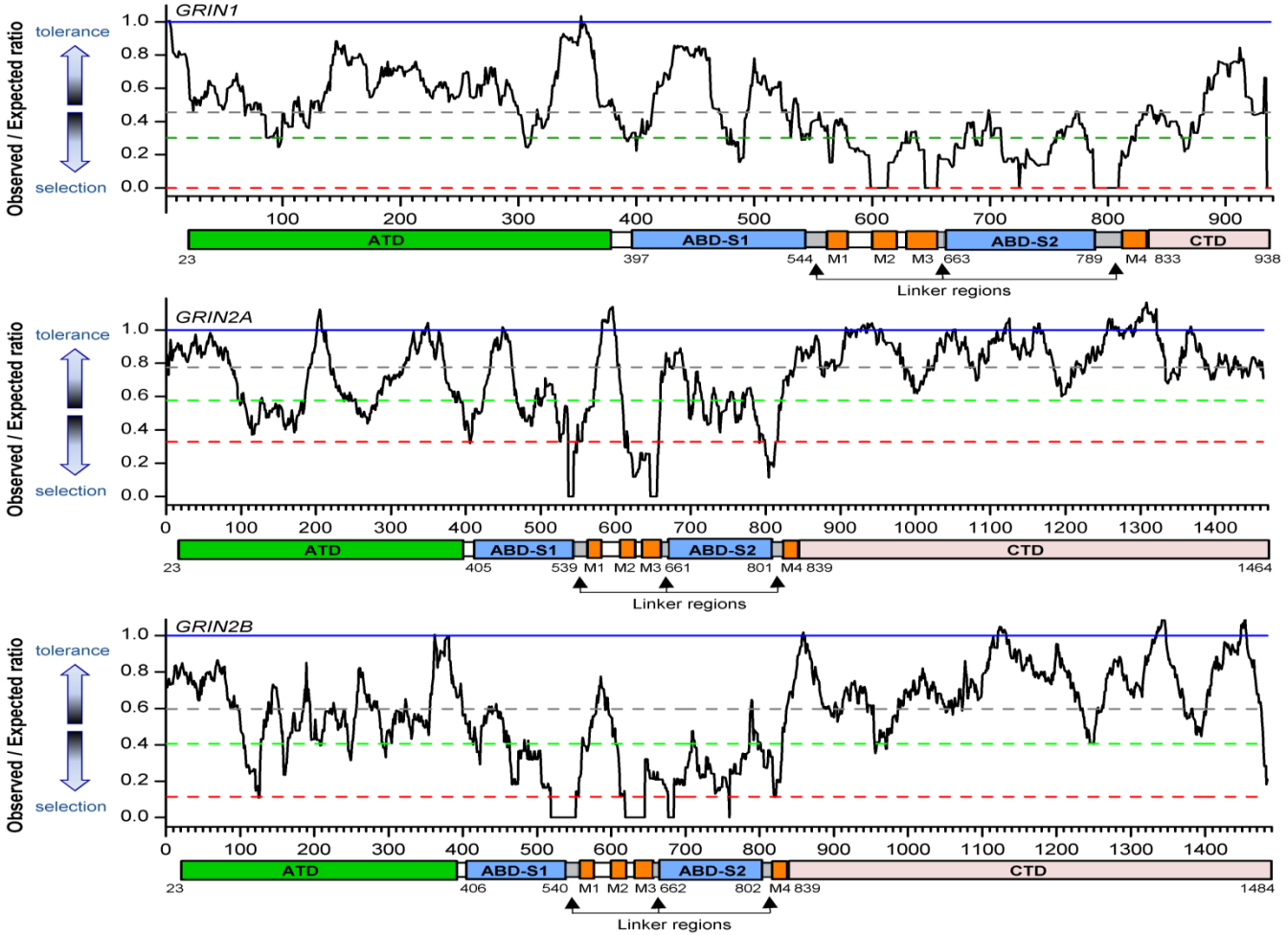


Mann-Whitney U test comparing to Novel Population Missense Variants (DiscovEHR)
835 ClinVar Pathogenic variants specified as DNM ($p<10^{-300}$)
24,900 ClinVar Pathogenic variants not specified as DNM ($p<10^{-300}$)

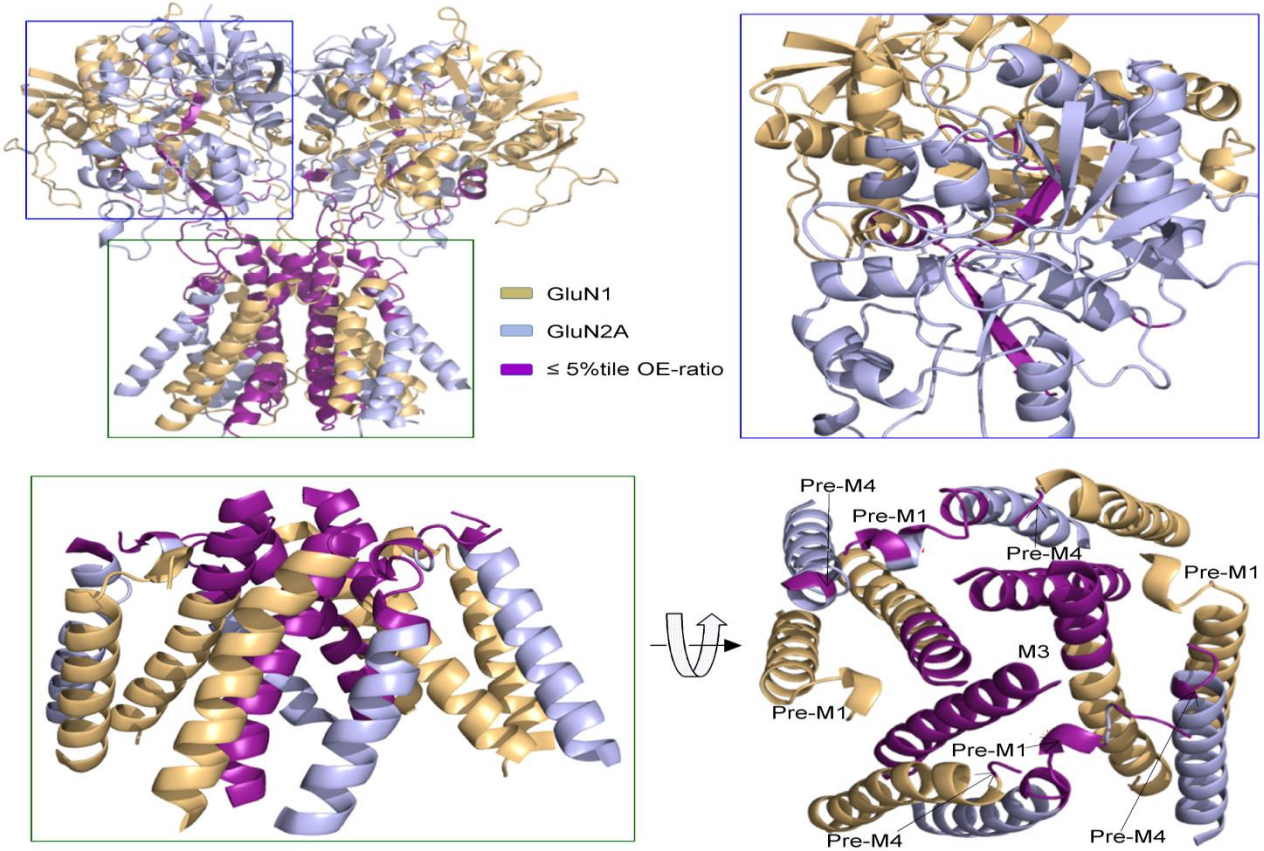
Group	25%	Median	75%
Exome-Wide	0.776	0.887	0.985
ClinVar Pathogenic DNM	0.366	0.565	0.747
ClinVar Pathogenic Other	0.664	0.825	0.943
ClinVar Benign	0.835	0.922	1.006
Novel Control Missense (DiscovEHR)	0.785	0.892	0.991
Novel Control Missense in ClinVar Genes (DiscovEHR)	0.776	0.885	0.982

exome-wide percentile	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%	55%	60%	65%	70%	75%	80%	85%	90%	95%	100%
MTR value	0.5462	0.6477	0.7041	0.744	0.7757	0.8024	0.8262	0.8476	0.8678	0.8872	0.9061	0.9249	0.9441	0.9641	0.9854	1.009	1.0363	1.0707	1.1228	1.6099

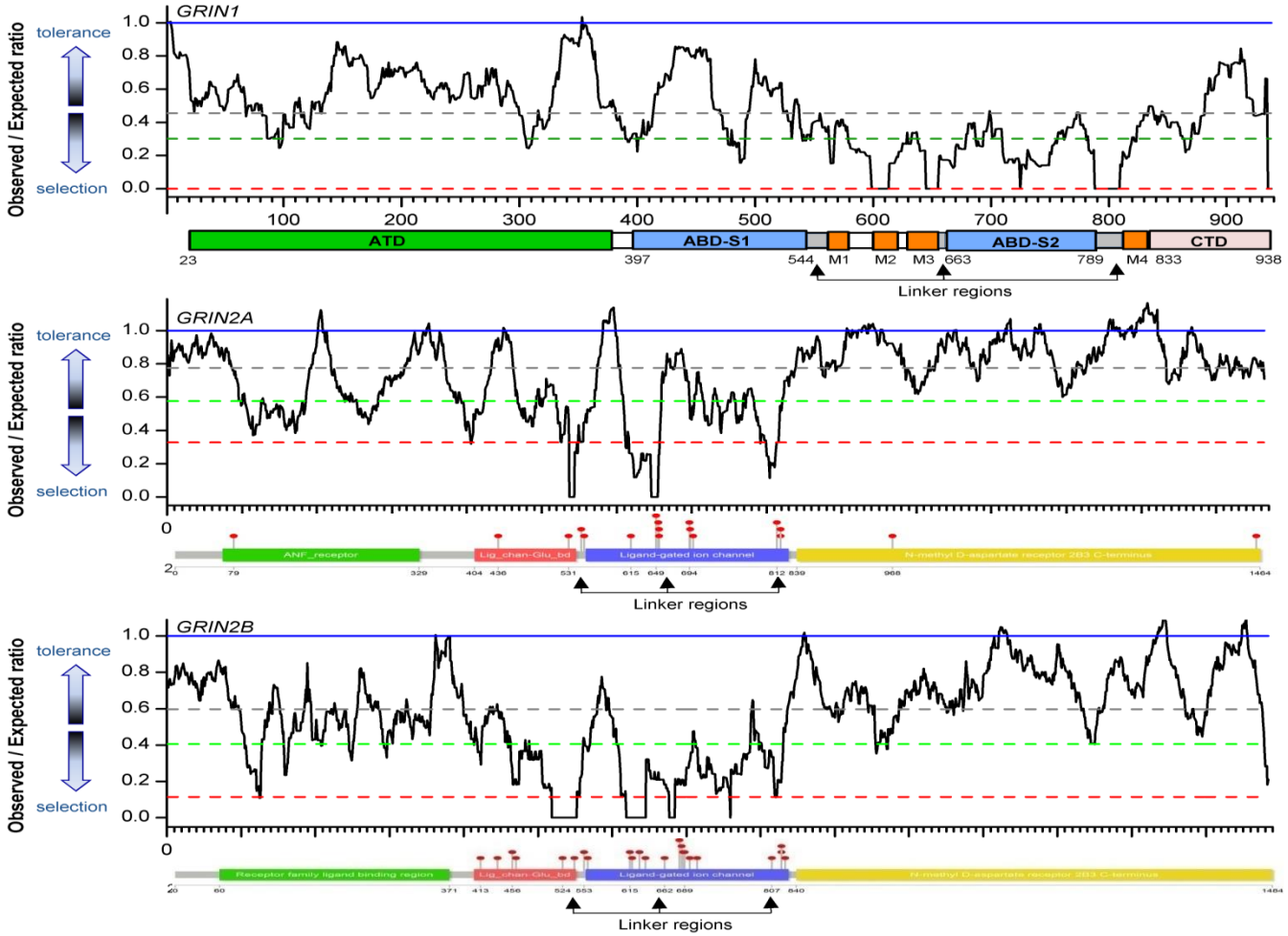
http://mtr-viewer.mdhs.unimelb.edu.au/



Ogden et al. PLOS Genetics (2017)



Ogden et al. PLOS Genetics (2017)



Swanger et al. AJHG (2016)

Leveraging MTR in Collapsing

Despite all case and control missense variants going through precisely same filtering (including absent in ExAC and predicted to be probably damaging by PolyPhen-2), their MTR distributions significantly differ (median MTR of 33.4% [20 case variants] and 70.4% [14 control variants]; Mann-Whitney U p = 0.004). Alternatively, using a *SCN1A* MTR 50th percentile threshold (MTR<0.746 for *SCN1A*) finds case variants preferentially residing among intolerant sequence (16/20 case vs. 4/14 control missense variants; Fisher's exact test p=0.005).