SUPPLEMENTARY INFORMATION

Enrichment of rare protein truncating variants in amyotrophic lateral sclerosis patients

Supplementary Tables

Supplementary Figures

Table S1. Dataset quality control

Quality control criteria applied	Total	Cases	Controls
Total samples in initial dataset	15,722	5,073	10,649
Samples successfully incorporated into VCF	15,428	4,851	10,577
Initial sample QC: (1) Removed call rate <90 (2) Removed dp Mean <10 (3) Removed gq Mean <65	14,143	4,569	10,575
Sex imputation: (1) Removed samples with ambiguous sex	15,128	4,558	10,570
Principal component analysis: (1) Case vs controls (2) Case, controls, and 1000 Genomes (3) Isolated overlapping case and control samples	12,317	4,429	7,888
Identity by descent: (1) Removed related or duplicated samples, IBD > 0.2	11,703	3,864	7,839

Table S2. Neurodegenerative disease genes

ABCA7	*BST1*	*CSF1R*	*ECHDC3*	*GCH1*	*ITM2B*	*NR4A2*	*PLD3*	*SCARB2*	*SPPL2A*	*TYROBP*
ACMSD	*C9orf72*	*DAO*	*EIF4G1*	*GIGYF2*	*LRRK2*	*NUCKS1*	*PM20D1*	*SCIMP*	*SQSTM1*	*UBA1*
ALS2	*CASS4*	*DCTN1*	*EPHA1*	*GPNMB*	*MAPT*	*OPTN*	*PNPLA6*	*SETX*	*STK39*	*UBQLN2*
ANG	*CCDC62*	*DDRGK1*	*FAM47E*	*GRN*	*MC1R*	*PANK2*	*PRNP*	*SIGMAR1*	*STX1B*	*UCHL1*
APOE	*CD2AP*	*DGKQ*	*FBXO7*	*HBEGF*	*MCCC1*	*PARK2*	*PRPH*	*SIPA1L2*	*TAF15*	*UNC13A*
APP	*CD33*	*DLG2*	*FERMT2*	*HLA-DRB5*	*MEF2C*	*PARK7*	*PSEN1*	*SLC24A4*	*TARDBP*	*VAPB*
AR	*CELF1*	*DNAJC13*	*FGF20*	*HNRNPA1*	*MS4A4E*	*PARL*	*PSEN2*	*SMN1*	*TBK1*	*VCP*
ARPP21	*CENPV*	*DNAJC6*	*FIG4*	*HNRNPA2B1*	*MS4A6A*	*PFN1*	*PTGER2*	*SMN2*	*TMEM163*	*VPS13C*
ATP13A2	*CHMP2B*	*DNMT1*	*FUS*	*HTRA2*	*NEFH*	*PICALM*	*PTK2B*	*SNCA*	*TMEM175*	*VPS35*
ATXN2	*CLU*	*DSG2*	*GAK*	*INPP5D*	*NEK1*	*PINK1*	*RAB7L1*	*SOD1*	*TMEM229B*	*ZCWPW1*
BIN1	*CR1*	*DYNC1H1*	*GBA*	*INPP5F*	*NME8*	*PLA2G6*	*RIT2*	*SORL1*	*TREM2*	

Table S3. Annotated variants across allele frequency categories

Number of variants	**Singletons (AC=1)**	**Doubletons (AC=2)**	**Ultra-rare singletons (AC=1)**	**Rare variants (MAF <0.01%)**
Total	536,415	232,452	309,281	2,802,758
Per case (3,864)	14-216	5-114	4-100	164-738
Per control (7,839)	7-199	3-92	5-95	173-583

Abbreviations are as follows: AC, allele count; MAF, minor allele frequency. The different frequency thresholds are as follows: (1) singletons (AC=1); (2) doubletons (AC=2); (3) ultra-rare singletons (AC=1, DiscovEHR=0); and (4) rare variants (MAF <0.01 in our dataset; ExAC (non-psychiatric studies, >45,000 samples); and in DiscovEHR (>50,000 samples).

Table S4. Proposed ALS genes and their signals under the protein-truncating variant model

Gene	Qualifying variants in cases (3,864)	Qualifying variants in controls (7,839)	Qualifying variants in total controls (28,910)	OR N_{cases}: 3,864 $N_{controls}$: 7,839	P-value N_{cases}: 3,864 $N_{controls}$: 7,839	OR N_{cases}: 3,864 $N_{controls}$: 28,910	P-value N_{cases}: 3,864 $N_{controls}$: 28,910
ALS2	2	5	17	0.81	1.00	0.88	1.00
ANG	NA	NA	NA	NA	NA	NA	NA
ARHGEF28	5	4	16	2.54	0.17	2.34	0.09
ARPP21	0	1	10	0.68	1.00	0.36	0.62
ATXN2	0	1	45	0.68	1.00	0.08	8.38×10^{-3} #
C21orf2	1	2	9	1.01	0.02	0.84	1.00
C9orf72	1	2	13	1.01	0.02	0.58	1.00
CENPV	NA	NA	NA	NA	NA	NA	NA
CHMP2B	4	1	31	8.12	0.04	0.97	1.00
DAO	0	2	27	0.41	1.00	0.14	0.07
DCTN1	1	1	9	2.03	0.55	0.83	1.00
FIG4	9	13	51	1.41	0.50	1.32	0.42
FUS	6	0	0	26.41	1.29×10^{-3}	97.40	2.68×10^{-6} #
GRN	1	1	8	2.03	0.55	0.94	1.00
HNRNPA1	0	1	18	0.68	1.00	0.20	0.26
HNRNPA2B1	NA	NA	NA	NA	NA	NA	NA
KIF5A	3	0	2	14.21	0.04	11.23	0.01 #
MAPT	0	2	46	0.41	1.00	0.08	5.067×10^{-3} #

MATR3	NA	NA	NA	NA	NA	NA	NA
MOBP	NA	NA	NA	NA	NA	NA	NA
NEFH	0	1	8	0.68	1.00	0.44	0.61
NEK1	25	4	29	12.80	$4.59x10^{-9}$*	6.48	$3.03x10^{-10}$#
OPTN	13	4	38	6.60	$3.04x10^{-4}$	2.56	$6.90x10^{-3}$
PFN1	NA	NA	NA	NA	NA	NA	NA
PNPLA6	3	8	44	0.76	1.00	0.51	0.36
PRPH	1	2	20	1.01	1.00	0.37	0.50
SCFD1	0	1	12	0.68	1.00	0.30	0.38
SETX	1	4	28	0.51	1.00	0.27	0.25
SIGMAR1	1	0	3	6.09	0.33	2.49	0.39
SOD1	1	1	2	2.03	0.55	3.74	0.31
SQSTM1	0	3	9	0.11	0.56	0.39	0.61
TAF15	NA	NA	NA	NA	NA	NA	NA
TARDBP	NA	NA	NA	NA	NA	NA	NA
TBK1	5	0	3	22.34	$3.92x10^{-3}$	12.48	$9.35x10^{-4}$#
UBQLN2	NA	NA	NA	NA	NA	NA	NA
UNC13A	NA	NA	NA	NA	NA	NA	NA
VAPB	NA	NA	NA	NA	NA	NA	NA
VCP	NA	NA	NA	NA	NA	NA	NA

*Passed exome-wide significance (P-value <2.5x10-6) in first analysis (3,864 cases and 7,839 controls.
#OR direction is maintained in secondary analysis (3,864 cases and 28,910 controls) and P-value is lower.
NA values in cells implies there were no qualifying variants in this gene under this model.

Table S5. Proposed ALS genes and their signals under the damaging missense model

Gene	Qualifying variants in cases (3,864)	Qualifying variants in controls (7,839)	Qualifying variants in total controls (28,910)	OR N_{cases}: 3,864 $N_{controls}$: 7,839	P-value N_{cases}: 3,864 $N_{controls}$: 7,839	OR N_{cases}: 3,864 $N_{controls}$: 28,910	P-value N_{cases}: 3,864 $N_{controls}$: 28,910
ALS2	7	17	73	0.84	0.83	0.72	0.49
ANG	0	2	8	0.41	1.00	0.44	0.61
ARHGEF28	12	14	128	1.74	0.21	0.70	0.29
ARPP21	3	5	55	1.22	0.72	0.41	0.15
ATXN2	NA	NA	NA	NA	NA	NA	NA
C21orf2	1	3	33	0.68	1.00	0.23	0.18
C9orf72	1	5	25	0.41	0.67	0.30	0.36
CENPV	1	2	13	1.01	1.00	0.58	1.00
CHMP2B	3	1	9	6.09	0.11	2.49	0.16
DAO	3	20	114	0.30	0.05	0.20	7.52×10^{-4} #
DCTN1	9	15	70	1.22	0.67	0.96	1.00
FIG4	7	11	51	1.29	0.62	1.03	0.84
FUS	0	2	7	0.41	1.00	0.50	1.00
GRN	11	13	59	1.72	0.20	1.40	0.35
HNRNPA1	NA	NA	NA	NA	NA	NA	NA
HNRNPA2B1	0	1	2	0.68	1.00	1.50	1.00
KIF5A	3	8	31	0.76	1.00	0.72	0.79
MAPT	7	8	49	1.78	0.28	1.07	0.84
MATR3	6	2	7	6.09	0.02	6.42	2.19×10^{-3} #

MOBP	NA	NA	NA	NA	NA	NA	NA
NEFH	NA	NA	NA	NA	NA	NA	NA
NEK1	10	22	93	0.92	1.00	0.80	0.65
OPTN	12	5	34	4.88	2.79x10^{-3}	2.65	8.99x10^{-3}
PFN1	0	1	65	0.68	1.00	0.06	6.70x10^{-4}[#]
PNPLA6	7	11	67	1.29	0.62	0.78	0.72
PRPH	8	13	69	1.25	0.65	0.87	0.86
SCFD1	3	8	35	0.76	1.00	0.64	0.62
SETX	32	48	265	1.36	0.19	0.90	0.65
SIGMAR1	1	2	18	1.01	1.00	0.42	0.72
SOD1	21	0	2	87.70	7.55x10^{-11}*	78.98	6.04x10^{-18}[#]
SQSTM1	5	11	38	0.92	1.00	0.98	1.00
TAF15	NA	NA	NA	NA	NA	NA	NA
TARDBP	NA	NA	NA	NA	NA	NA	NA
TBK1	6	5	24	2.44	0.20	1.87	0.16
UBQLN2	0	3	15	0.29	0.56	0.24	0.24
UNC13A	7	6	49	2.37	0.14	1.07	0.84
VAPB	2	1	4	4.06	0.26	3.74	0.15
VCP	6	0	18	26.41	1.29x10^{-3}	2.50	0.06

*Passed exome-wide significance (P-value <2.5x10-6) in first analysis (3,864 cases and 7,839 controls.
[#]OR direction is maintained in secondary analysis (3,864 cases and 28,910 controls) and P-value is lower.
NA values in cells implies there were no qualifying variants in this gene under this model.

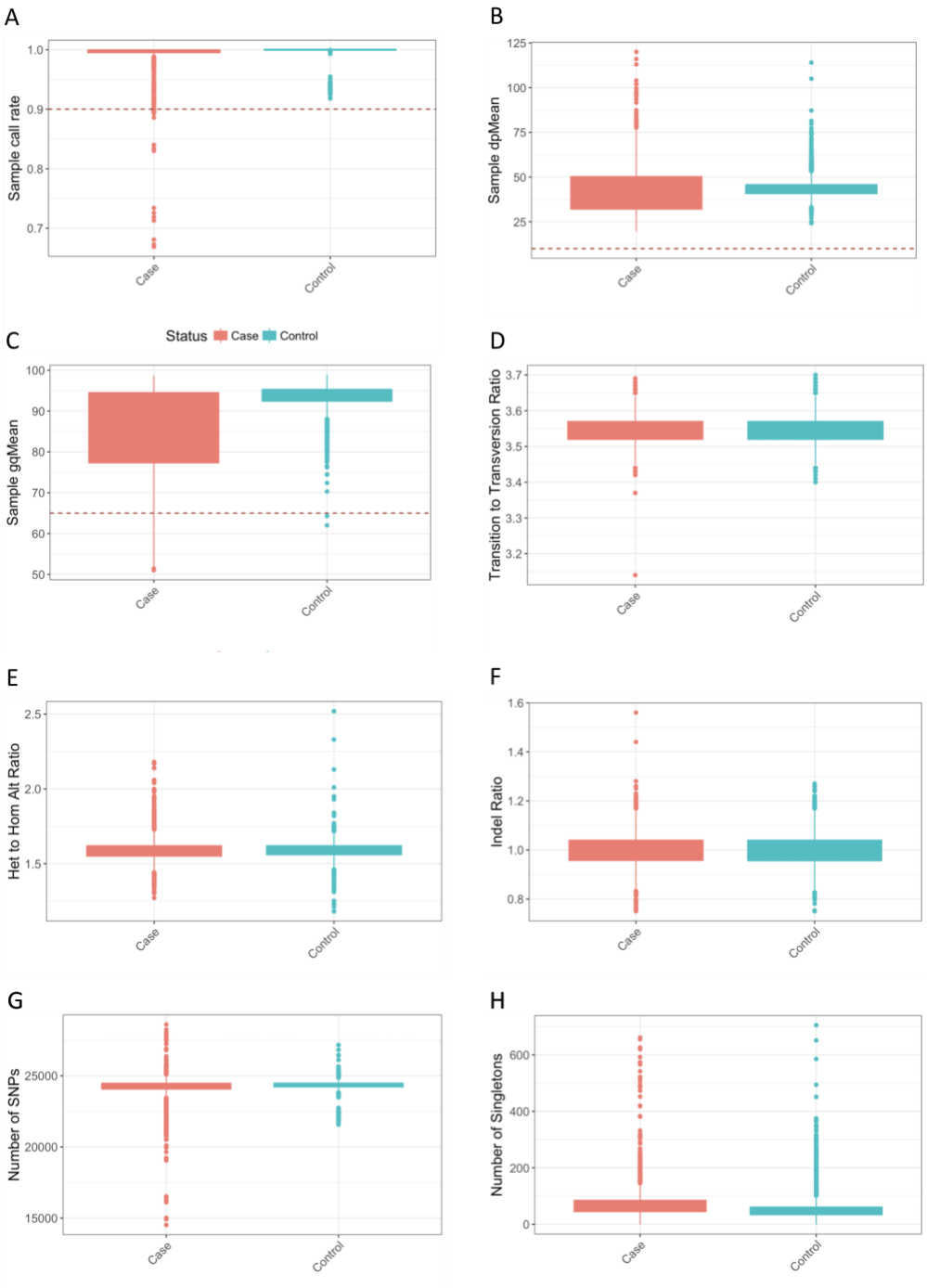


Figure S1. Initial sample quality control analysis

(A) Sample call rate. (B) Sample mean depth. (C) Sample mean genotype quality. (D) Sample transition to transversion ratio. (E) Sample heterozygous to homozygous ratio. (F) Sample insertion to deletion ratio. (G) Number of SNPs in each sample. (H) Number of singletons in each sample.

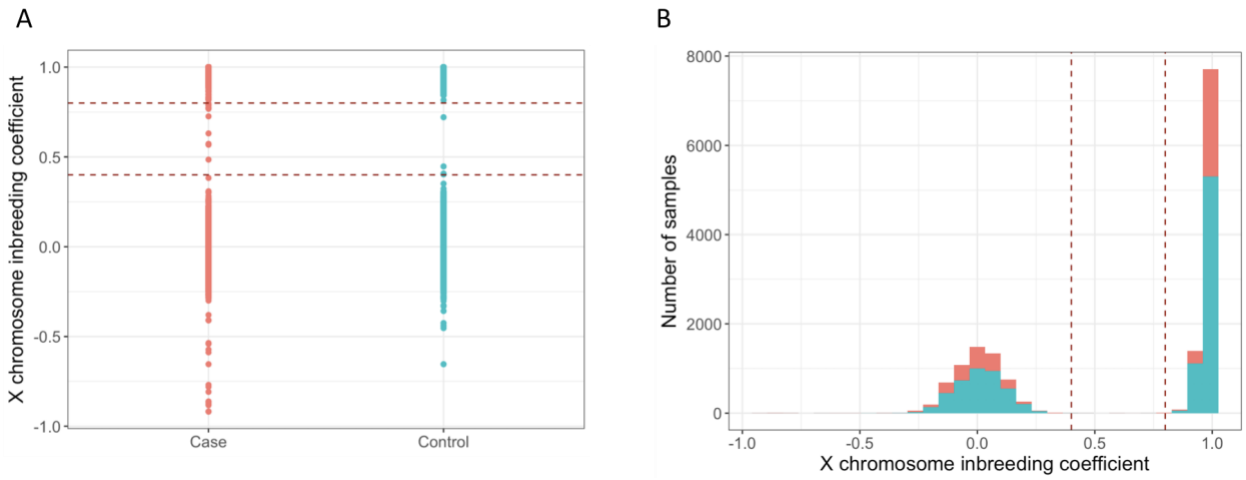


Figure S2. Sex imputation analysis

We excluded samples with ambiguous sex status. Samples with an X chromosome inbreeding coefficient >0.8 were classified as males and samples with an X chromosome inbreeding coefficient <0.4 were classified as females. Samples <0.8 and >0.4 were classified as samples with ambiguous sex status. (A) and (B) are depicting the same data of sex status between cases and controls differently.

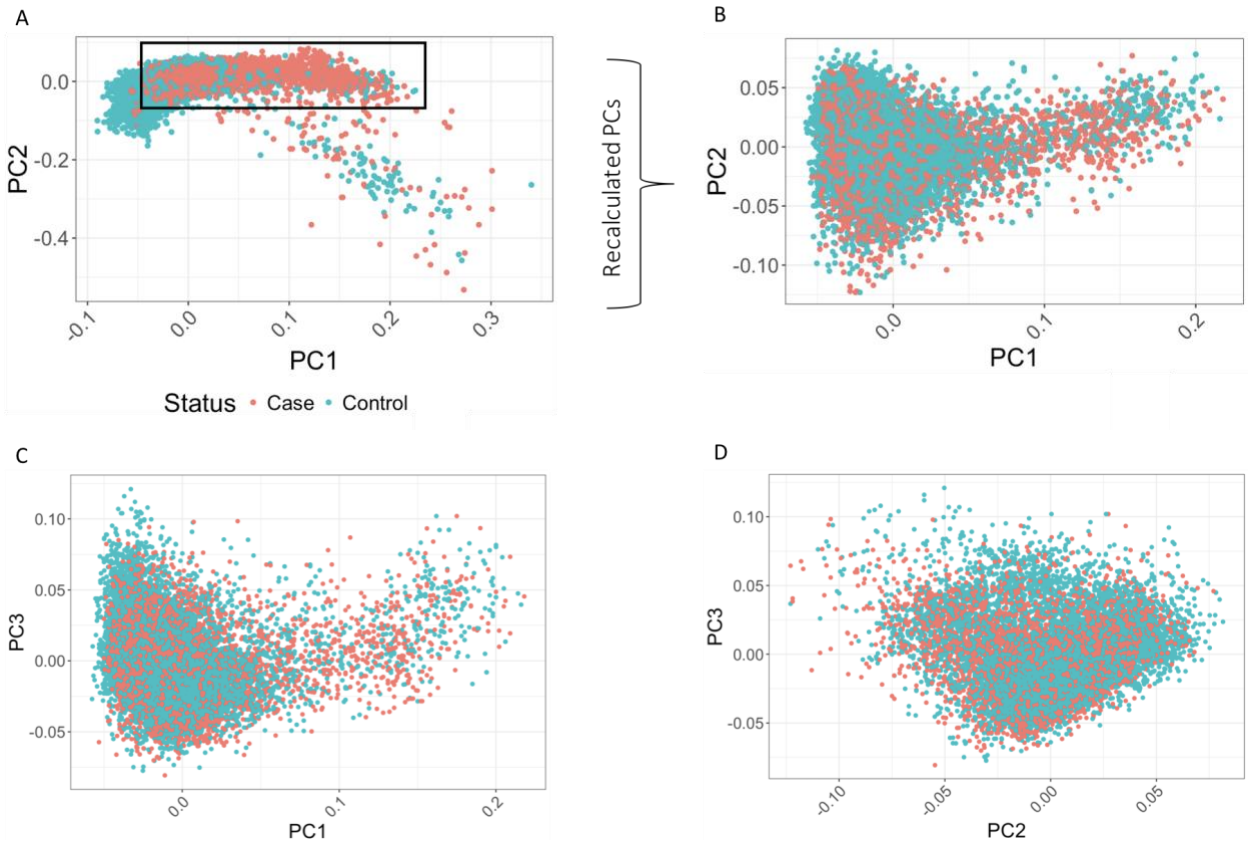


Figure S3. Principal component analysis of ALS dataset

(A) PC1 and PC2 of ALS dataset, cases are in red, and controls are in blue. All samples outside of the black rectangle were excluded from the dataset to prevent any false signals due to ancestry.

(B) PC1 and PC2 of ALS dataset after removing non-overlapping samples and recalculating the PCs.

(C) PC1 and PC3 of ALS dataset after removing non-overlapping samples and recalculating the PCs.

(D) PC2 and PC3 of ALS dataset after removing non-overlapping samples and recalculating the PCs.

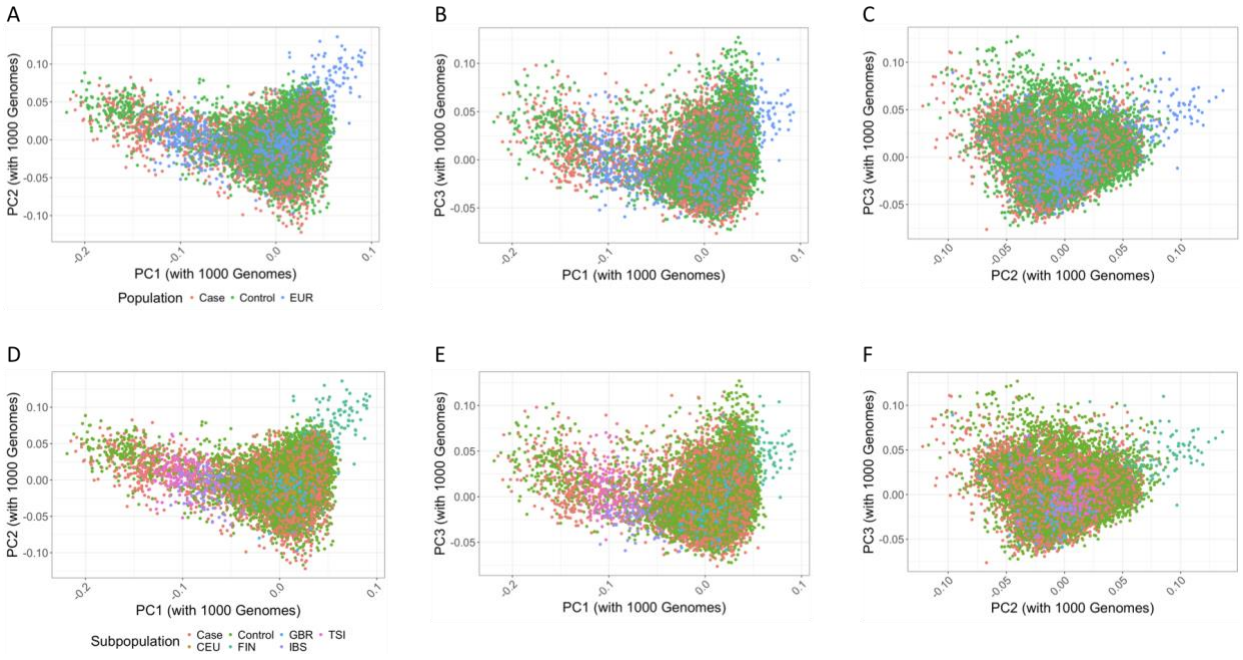


Figure S4. Principal component analysis of ALS dataset with 1000 Genomes

(A) PC1 and PC2 of ALS dataset with 1000 Genomes. Cases, controls, and the European population is shown.

(B) PC1 and PC3 of ALS dataset with 1000 Genomes. Cases, controls, and the European population is shown.

(C) PC2 and PC3 of ALS dataset with 1000 Genomes. Cases, controls, and the European population is shown.

(D) PC1 and PC2 of ALS dataset with 1000 Genomes. Cases, controls, and the European subpopulations are shown.

(E) PC1 and PC3 of ALS dataset with 1000 Genomes. Cases, controls, and the European subpopulations are shown.

(F) PC2 and PC3 of ALS dataset with 1000 Genomes. Cases, controls, and the European subpopulations are shown.

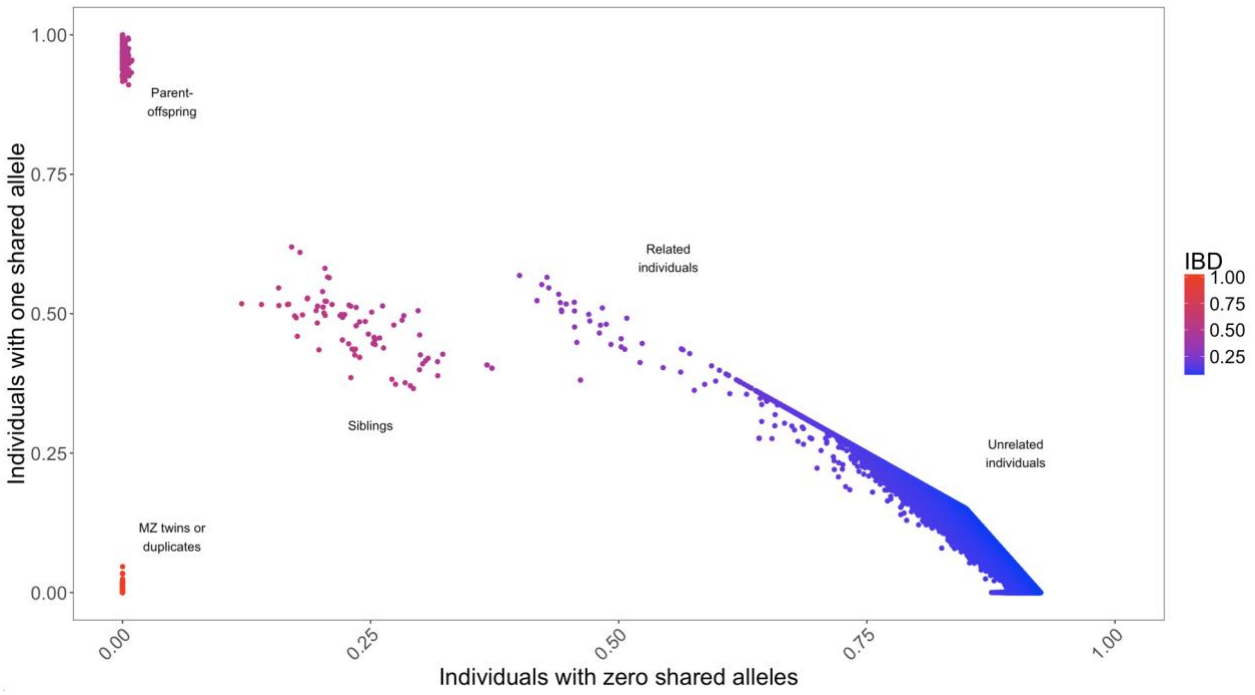


Figure S5. Identity-by-descent analysis to remove related individuals. Only unrelated individuals (dark blue) were included in our dataset

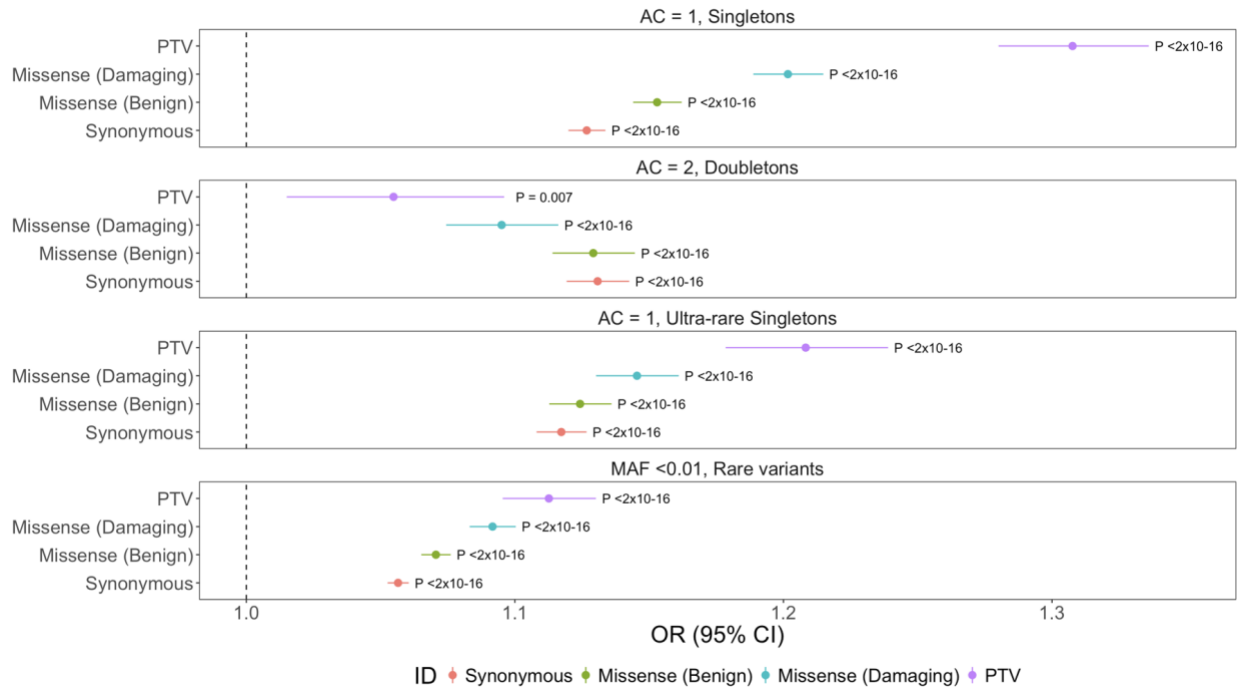


Figure S6. Exome wide analysis of variants in cases and controls within model 1 (sample variation only)

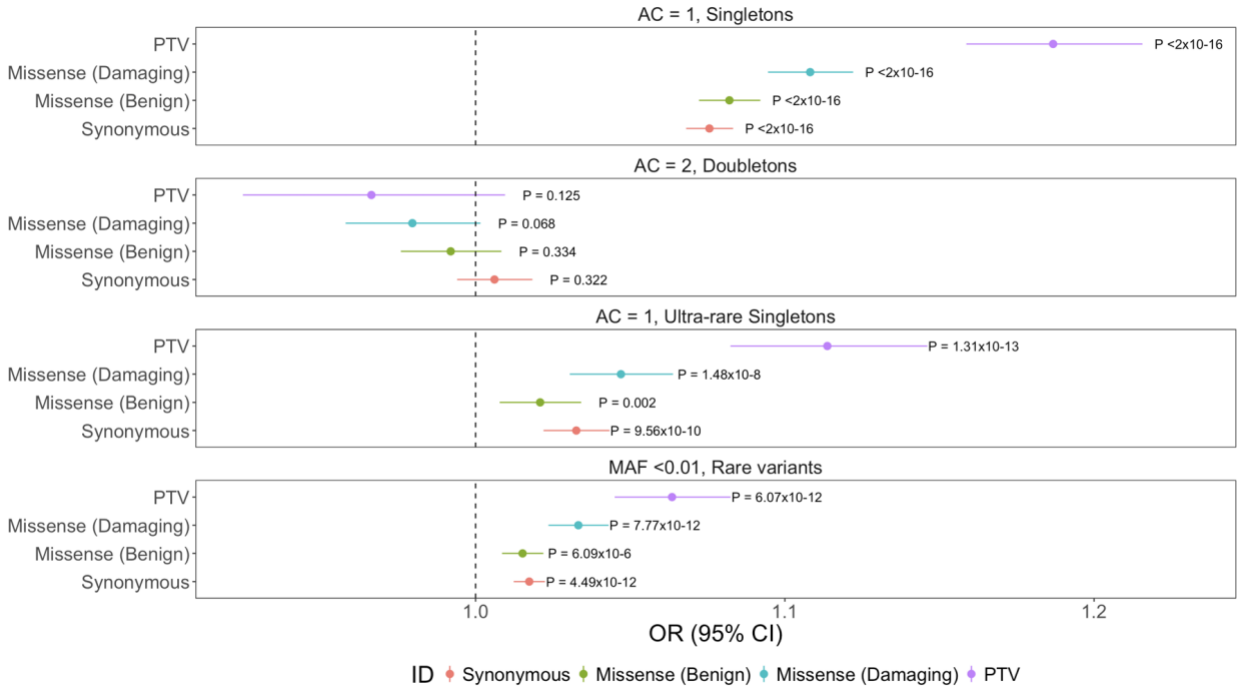


Figure S7. Exome wide analysis of variants in cases and controls within model 2 (sample variation, sample sex, and PC1-PC10)

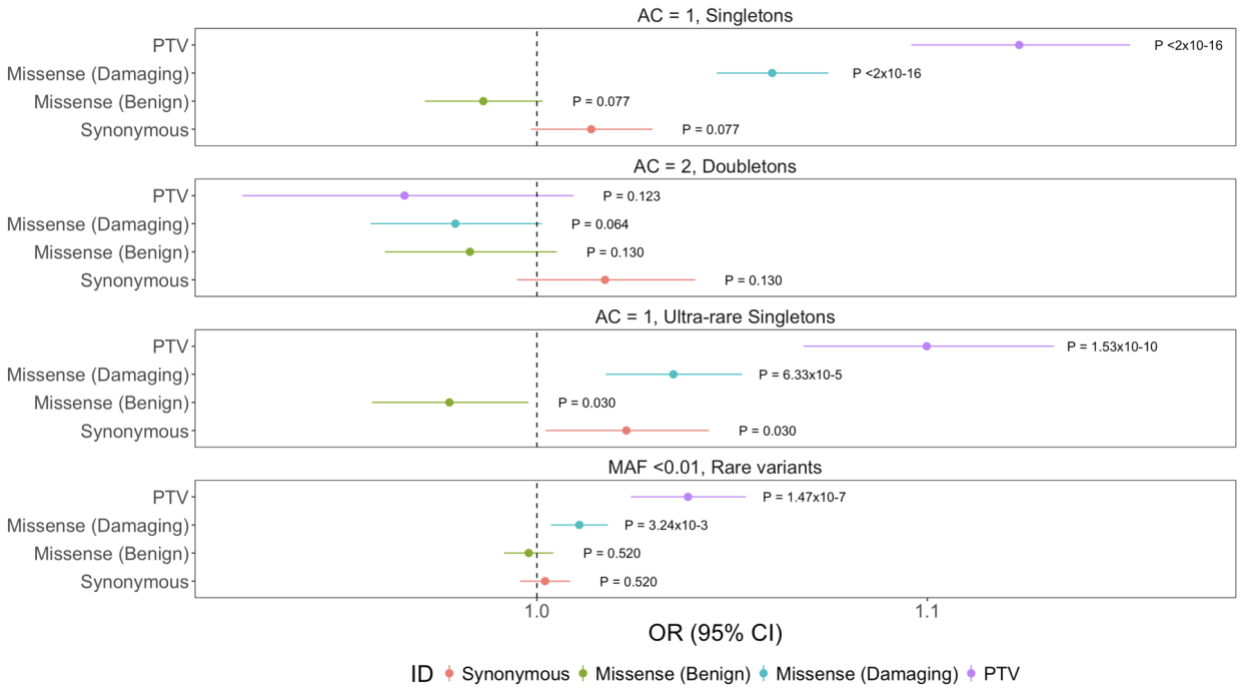


Figure S8. Exome wide analysis of variants in cases and controls within model 4 (sample variation, sample sex, PC1-PC10, and benign variation count, summation of synonymous and benign missense variation)

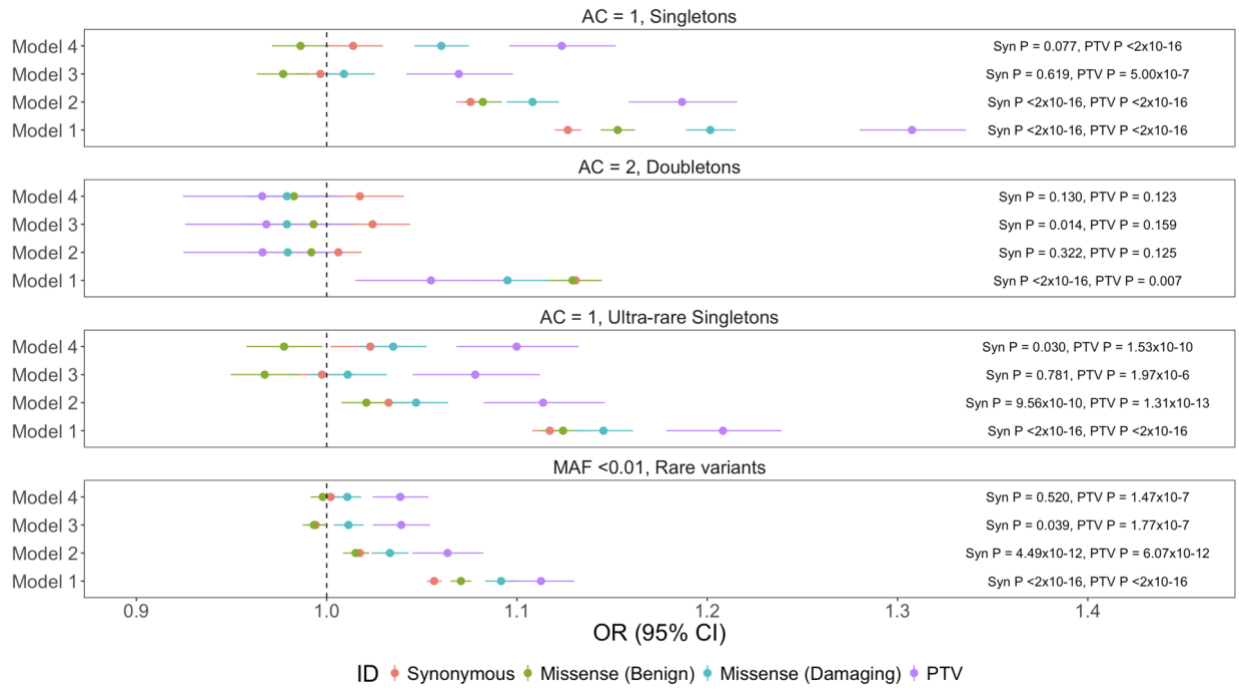


Figure S9. All models together

Model 1: Sample variation.
Model 2: Sample variation, sample sex, PC1-PC10.
Model 3: Sample variation, sample sex, PC1-PC10, and total exome count (summation of synonymous, benign missense, damaging missense, and PTV).
Model 4: Sample variation, sample sex, PC1-PC10, and benign variation count (summation of synonymous and benign missense variation).

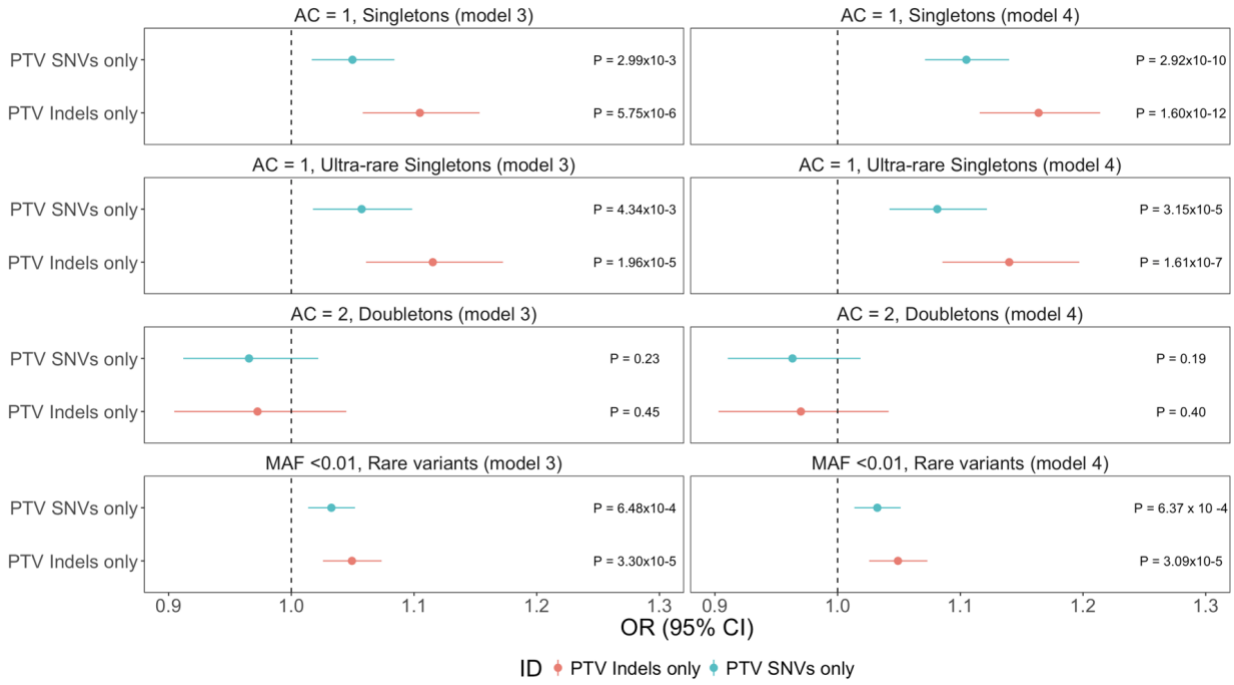


Figure S10. Exome wide enrichment of SNV-based PTVs and indel-based PTVs in ALS cases

Extension of Figure 1: Evaluating the effects of SNV-based and indel-based PTVs within singletons (AC=1), doubletons (AC=2), ultra-rare singletons (AC=1, 0 in DiscovEHR), and rare variants (MAF<0.01 in our dataset, DiscovEHR and ExAC). Odds ratios and 95% confidence intervals for each class of variation are depicted by different colors. P-values are also displayed. Model 3 evaluates sample variation with the covariates, sample sex, PC1-10, and total exome count (summation of synonymous variation, benign missense variation, damaging missense variation, and PTV SNV or PTV indel). Model 4 evaluates sample variation with the covariates, sample sex, PC1-10, and benign variation (summation of synonymous and benign missense variation).

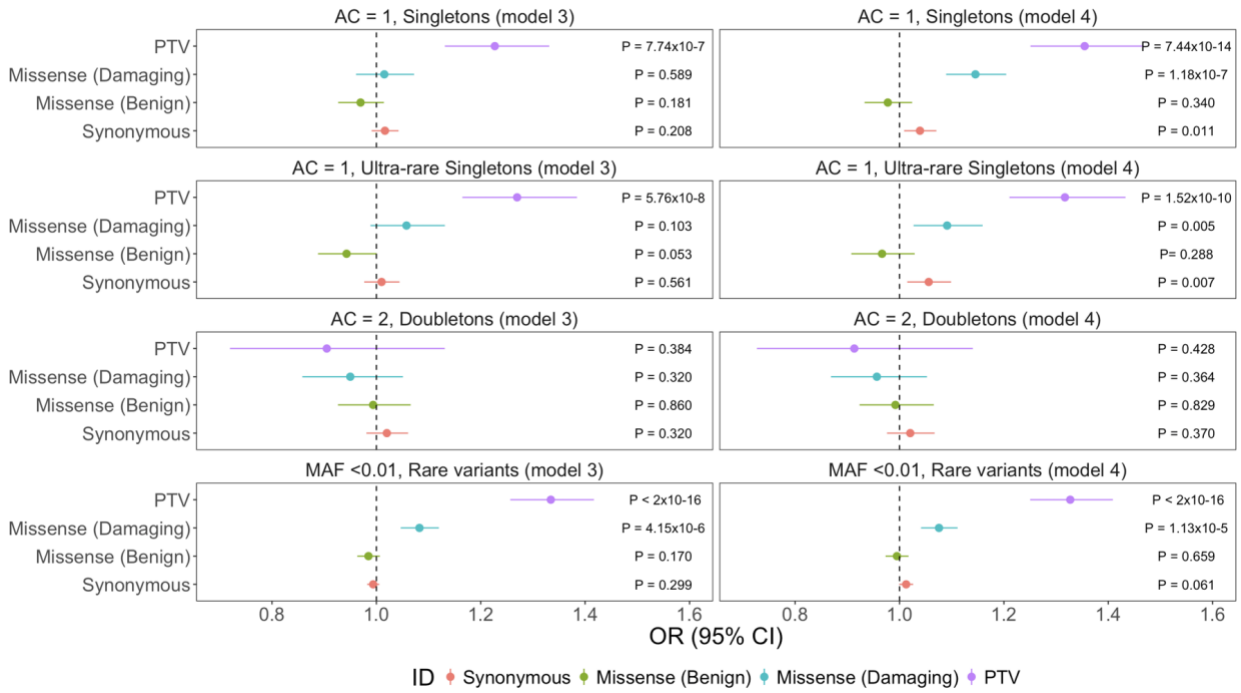


Figure S11. Enrichment of variants in constrained genes in ALS cases

Extension of Figure 2A: Evaluating the effects of constrained genes in synonymous variants, benign missense variants, damaging missense variants, and PTVs within singletons (AC=1), doubletons (AC=2), ultra-rare singletons (AC=1, 0 in DiscovEHR), and rare variants (MAF<0.01 in our dataset, DiscovEHR and ExAC). Odds ratios and 95% confidence intervals for each class of variation are depicted by different colors. P-values are also displayed. Model 3 evaluates sample variation with the covariates, sample sex, PC1-10, and total exome count (summation of synonymous variation, benign missense variation, damaging missense variation, and PTV). Model 4 evaluates sample variation with the covariates, sample sex, PC1-10, and benign variation (summation of synonymous and benign missense variation).

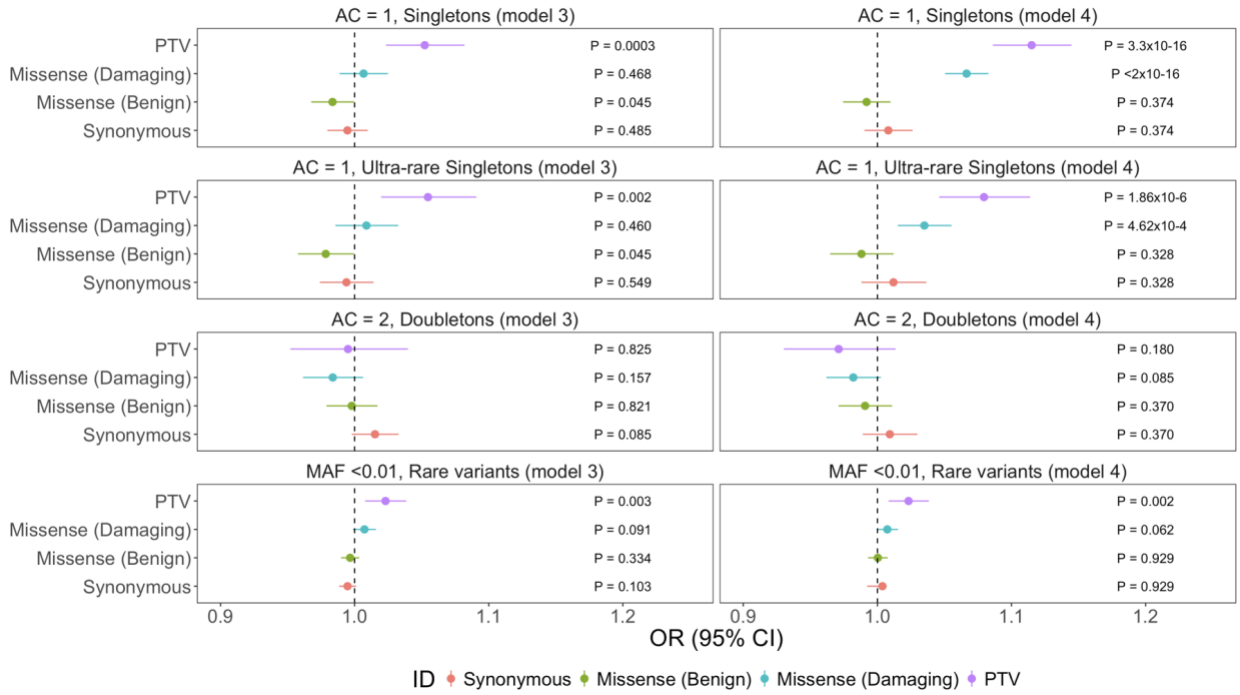


Figure S12. Exome-wide analysis with constrained genes removed

Extension of Figure 2B: Evaluating the residual effects with constrained genes removed in synonymous variants, benign missense variants, damaging missense variants, and PTVs within singletons (AC=1), doubletons (AC=2), ultra-rare singletons (AC=1, 0 in DiscovEHR), and rare variants (MAF<0.01 in our dataset, DiscovEHR and ExAC). Odds ratios and 95% confidence intervals for each class of variation are depicted by different colors. P-values are also displayed. Model 3 evaluates sample variation with the covariates, sample sex, PC1-10, and total exome count (summation of synonymous variation, benign missense variation, damaging missense variation, and PTV). Model 4 evaluates sample variation with the covariates, sample sex, PC1-10, and benign variation (summation of synonymous and benign missense variation).

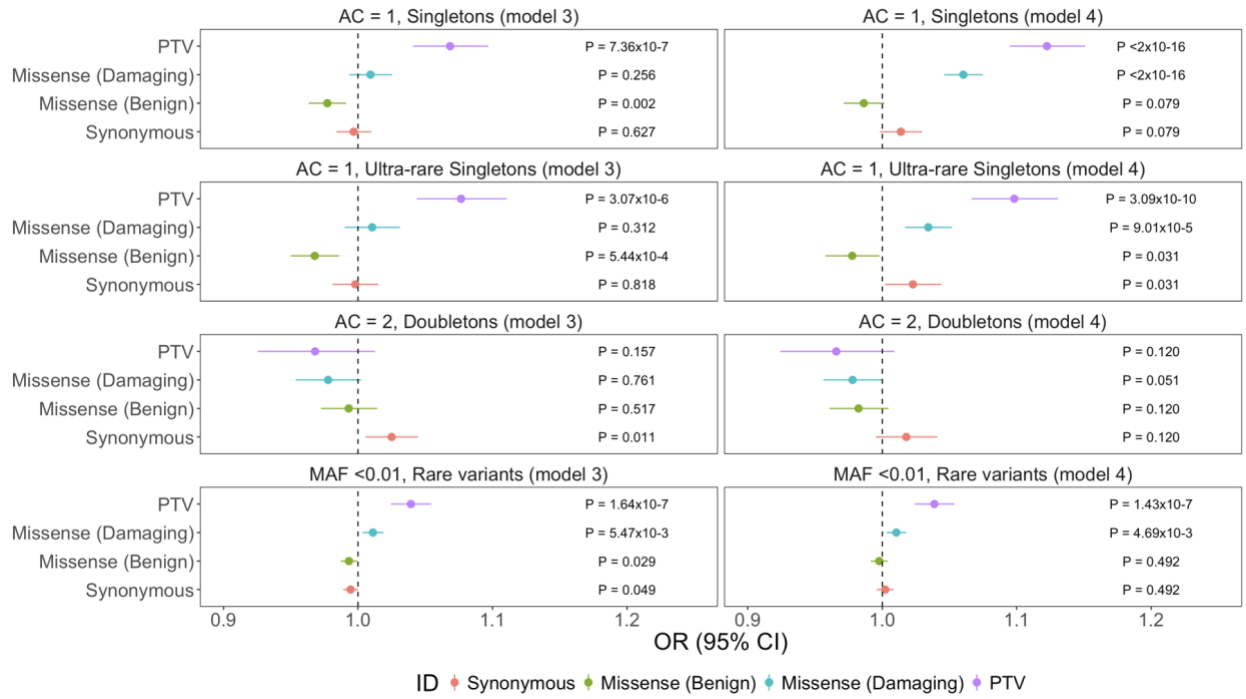


Figure S13. Exome-wide analysis with known ALS genes removed

Extension of Figure 3A: Evaluating the effects of known ALS genes removed in synonymous variants, benign missense variants, damaging missense variants, and PTVs within singletons (AC=1), doubletons (AC=2), ultra-rare singletons (AC=1, 0 in DiscovEHR), and rare variants (MAF<0.01 in our dataset, DiscovEHR and ExAC). Odds ratios and 95% confidence intervals for each class of variation are depicted by different colors. P-values are also displayed. Model 3 evaluates sample variation with the covariates, sample sex, PC1-10, and total exome count (summation of synonymous variation, benign missense variation, damaging missense variation, and PTV). Model 4 evaluates sample variation with the covariates, sample sex, PC1-10, and benign variation (summation of synonymous and benign missense variation).

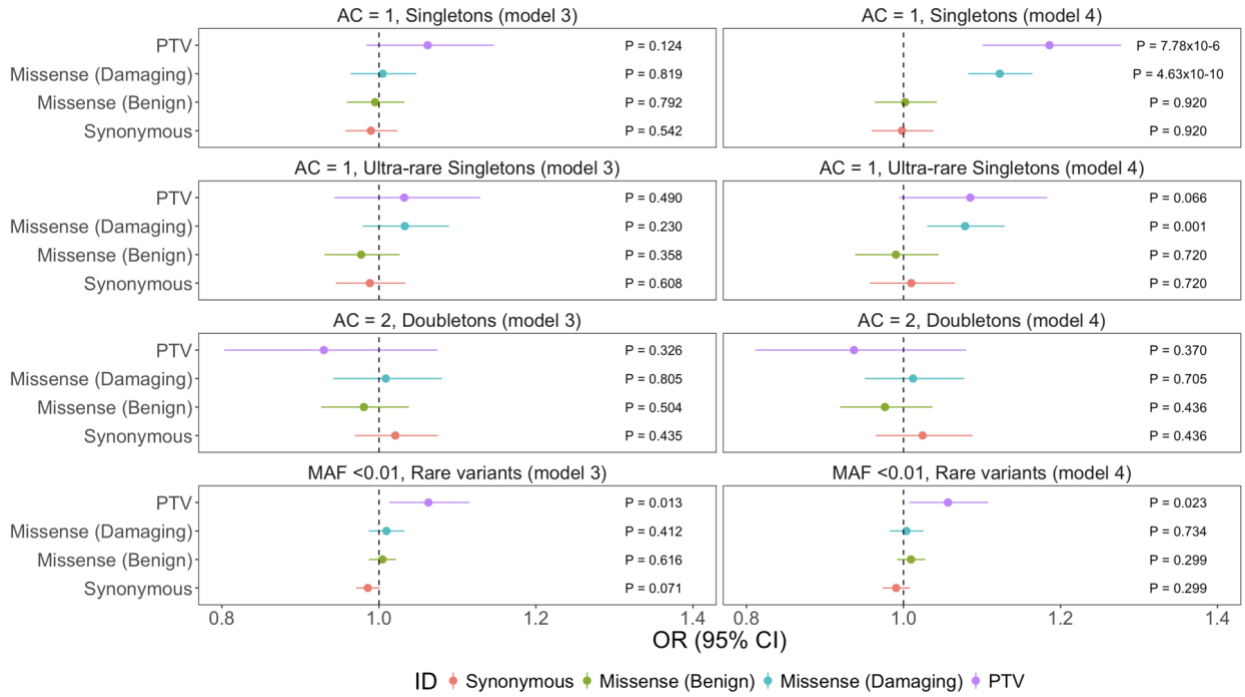


Figure S14. Analysis of other neurodegenerative disease genes

Extension of Figure 3B: Evaluating the effects of genes associated with other neurodegenerative disease (motor neuron diseases: primary lateral sclerosis, progressive muscular atrophy, progressive bulbar palsy, and spinal muscular atrophy; diseases with overlapping phenotypes: frontotemporal dementia, Parkinson's disease, Pick's disease, and Alzheimer's disease) in synonymous variants, benign missense variants, damaging missense variants, and PTVs within singletons (AC=1), doubletons (AC=2), ultra-rare singletons (AC=1, 0 in DiscovEHR), and rare variants (MAF<0.01 in our dataset, DiscovEHR and ExAC). Odds ratios and 95% confidence intervals for each class of variation are depicted by different colors. P-values are also displayed. Model 3 evaluates sample variation with the covariates, sample sex, PC1-10, and total exome count (summation of synonymous variation, benign missense variation, damaging missense variation, and PTV). Model 4 evaluates sample variation with the covariates, sample sex, PC1-10, and benign variation (summation of synonymous and benign missense variation).

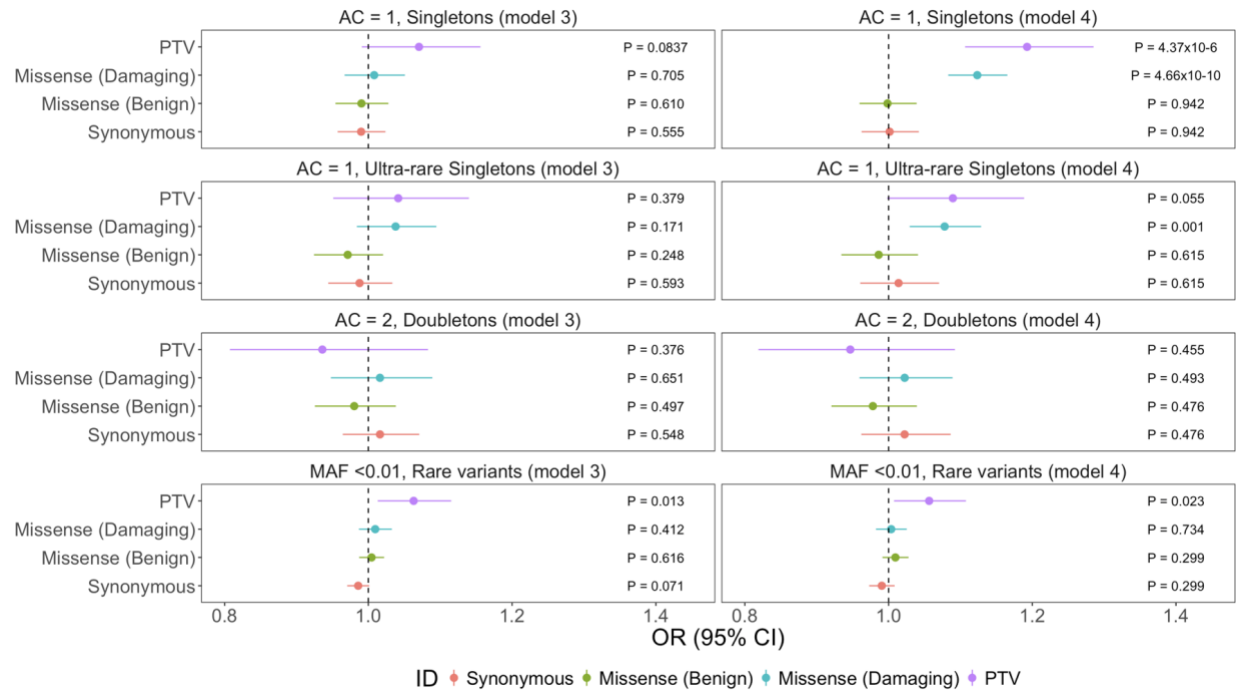
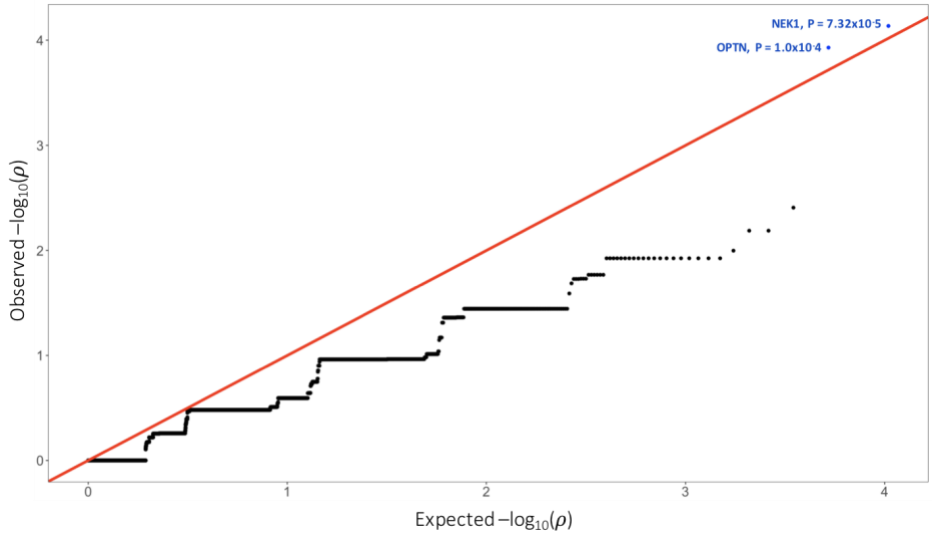


Figure S15. Analysis of brain specific genes

Extension of Figure 3C: Analysis of brain specific genes in synonymous variants, benign missense variants, damaging missense variants, and PTVs within singletons (AC=1), doubletons (AC=2), ultra-rare singletons (AC=1, 0 in DiscovEHR), and rare variants (MAF<0.01 in our dataset, DiscovEHR and ExAC). Odds ratios and 95% confidence intervals for each class of variation are depicted by different colors. P-values are also displayed. Model 3 evaluates sample variation with the covariates, sample sex, PC1-10, and total exome count (summation of synonymous variation, benign missense variation, damaging missense variation, and PTV). Model 4 evaluates sample variation with the covariates, sample sex, PC1-10, and benign variation (summation of synonymous and benign missense variation).

A



B

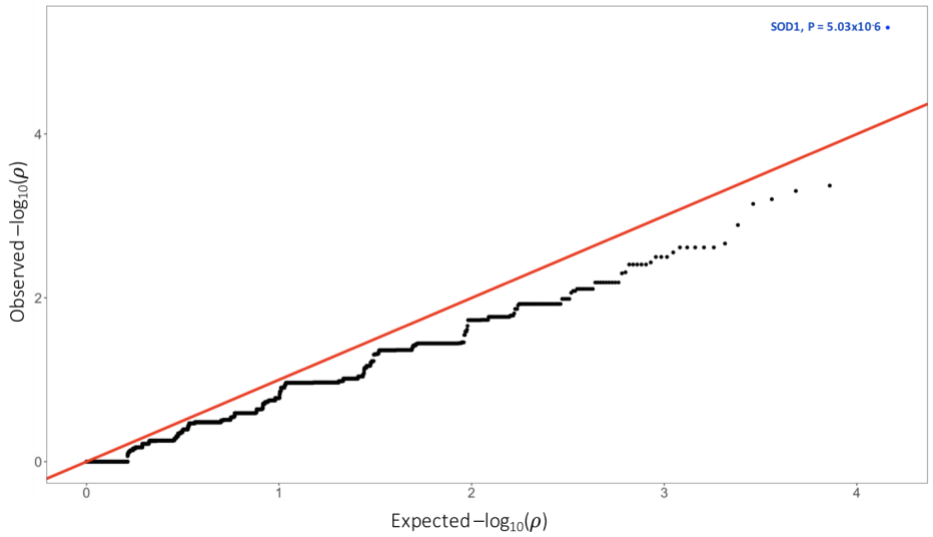
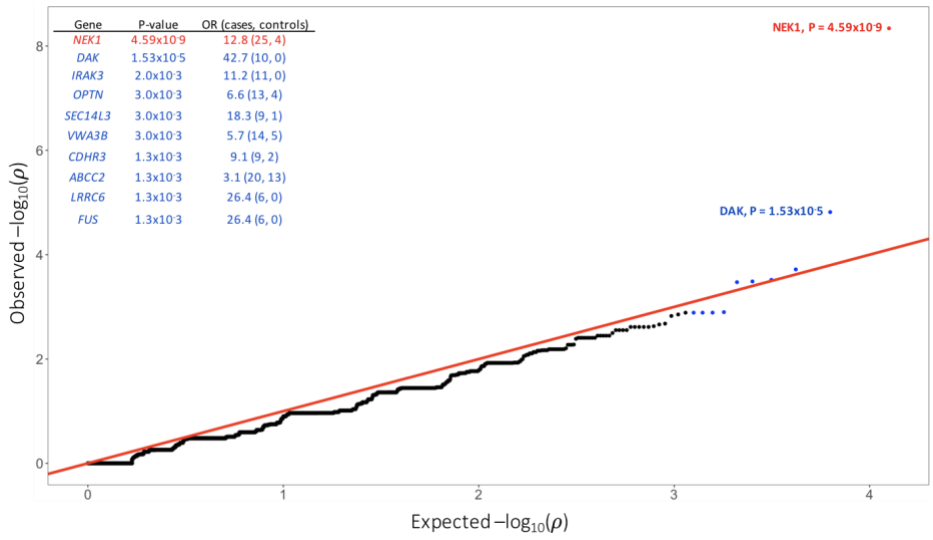


Figure S16. Quantile-quantile plot of ultra-rare singletons

(A) Ultra-rare singleton (AC=1, 0 in DiscovEHR database) for PTV. PTVs in *NEK1* and *OPTN*, which are known ALS genes, are enriched in ALS cases. *NEK1* and *OPTN* P-values are displayed.

(B) Ultra-rare singleton (AC=1, 0 in DiscovEHR database) for damaging missense variants. Damaging missense variants in *SOD1are* enriched in ALS cases. *SOD1* P-value is displayed.

A



B

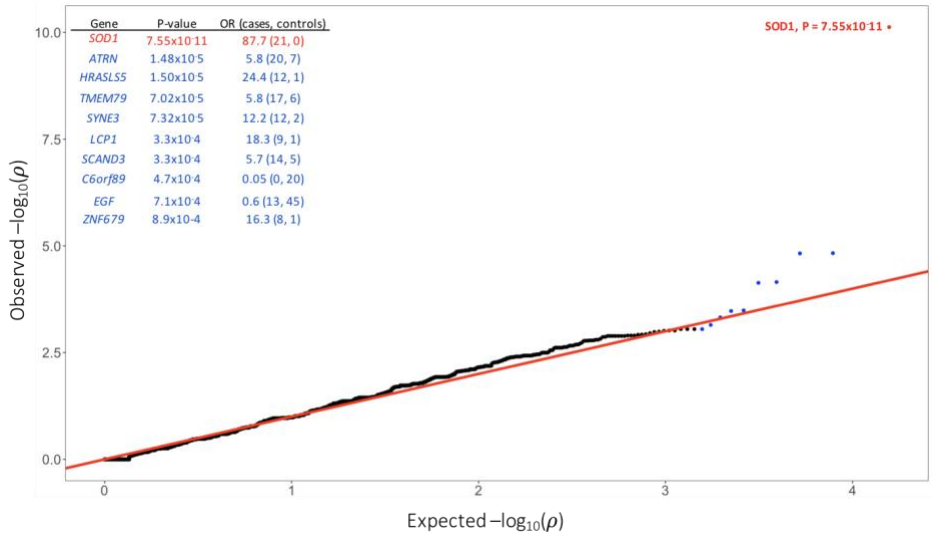
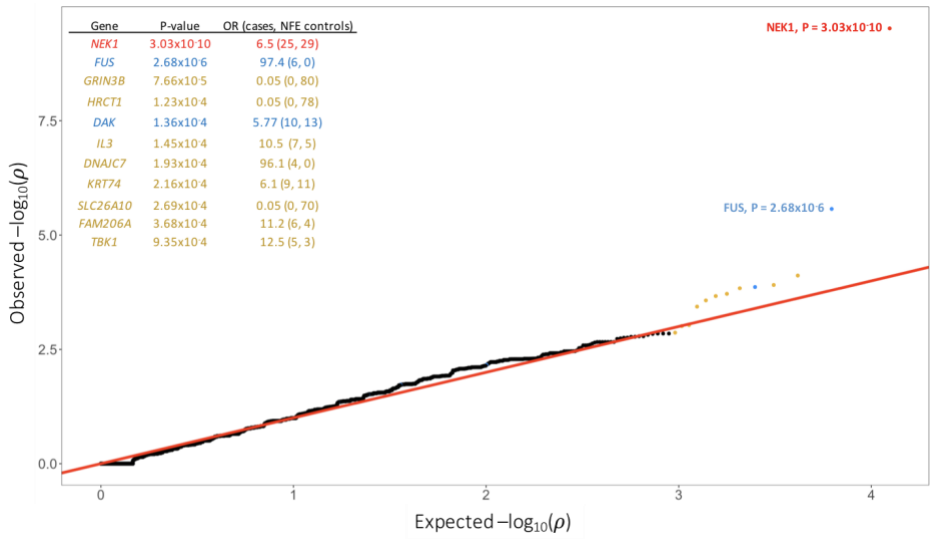


Figure S17. Quantile-quantile plot of rare variants (MAF <0.001)

(A) Rare PTVs (MAF<0.001% in dataset, DiscovEHR, and ExAC) in ALS dataset. PTVs in *NEK1* and *DAK* are enriched in ALS cases. The top 10 genes with their P-values are displayed. *NEK1* passes exome-wide significance.

(B) Rare damaging missense variants (MAF<0.001% in dataset, DiscovEHR, and ExAC) in ALS dataset. Damaging missense variants in *SOD1* are enriched in ALS cases. The top 10 genes with their P-values are displayed. *SOD1* passes exome-wide significance.

A



B

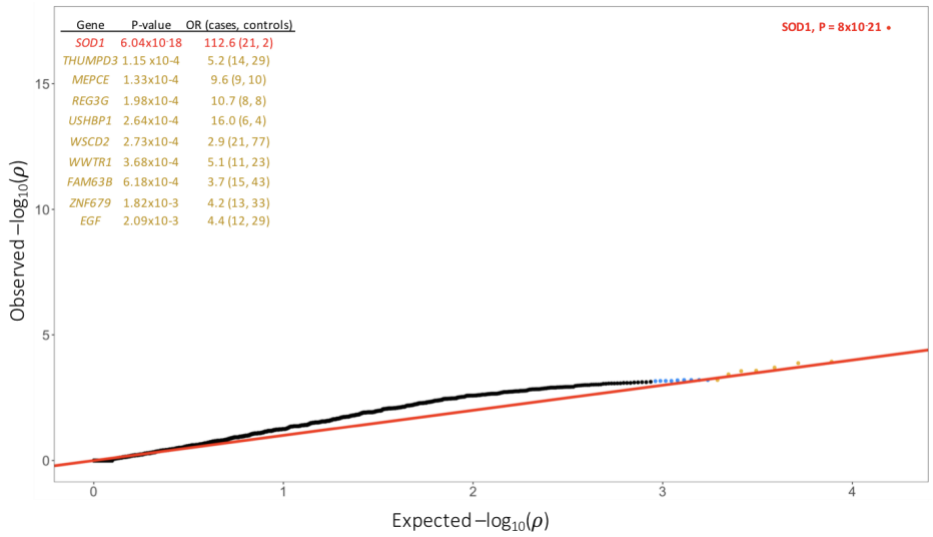


Figure S18. Quantile-quantile plot of rare variants (MAF<0.001%)

(A) Rare PTVs (MAF<0.001% in dataset, DiscovEHR, and ExAC) in ALS cases with an additional 21,071 non-Finnish European controls from ExAC for a total of 28,910 controls. The top 10 genes with their P-values are displayed. *NEK1* is shown in red and passes exome-wide significance. Genes in blue were the most significant genes in the discovery analysis. *FUS* is shown in blue and approaches exome-wide significance. Genes in yellow were the most significant genes in the secondary analysis.

(B) Rare damaging missense variants (MAF<0.001% in dataset, DiscovEHR, and ExAC) in ALS cases with an additional 21,071 non-Finnish European controls from ExAC for a total of 28,910 controls. The top 10 genes with their P-values are displayed. *SOD1* is shown in red and passes

exome-wide significance. Genes in blue were the most significant genes in the discovery analysis. Genes in yellow were the most significant genes in the secondary analysis.