SUPPLEMENTARY DATA

Human-lineage-specific genomic elements: relevance to neurodegenerative disease and *APOE* transcript usage

Zhongbo Chen[1], David Zhang[1], Regina H. Reynolds[1], Emil K. Gustavsson[1], Sonia García Ruiz[1], Karishma D'Sa[1], Aine Fairbrother-Browne[1], Jana Vandrovcova[1], International Parkinson's Disease Genomics Consortium (IPDGC)[#], John Hardy[1,2,3,4,5], Henry Houlden[6], Sarah A. Gagliano Taliun[7], Juan Botía[8], Mina Ryten[1]

[1.] *Department of Neurodegenerative Disease, Queen Square Institute of Neurology, University College London (UCL), London, UK*

[2.] *Reta Lila Weston Institute, Queen Square Institute of Neurology, UCL, London, UK*

[3.] *UK Dementia Research Institute at UCL, Queen Square Institute of Neurology, UCL, London, UK*

[4.] *NIHR University College London Hospitals Biomedical Research Centre, London, UK*

[5.] *Institute for Advanced Study, The Hong Kong University of Science and Technology, Hong Kong SAR, China*

[6.] *Department of Neuromuscular Disease, Queen Square Institute of Neurology, UCL, London, UK*

[7.] *Center for Statistical Genetics and Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, USA*

[8.] *Departamento de Ingeniería de la Información y las Comunicaciones, Universidad de Murcia, Murcia, Spain*

[#] *Full list of consortium members appended*

Correspondence to Dr Mina Ryten (mina.ryten@ucl.ac.uk)

SUPPLEMENTAL DATA

Supplementary Data Figure Titles and Legends

Supplementary Figure 1. Kernal density plots of annotation metrics. Panel **a** depicts density plot of constraint (context dependent tolerance score (CDTS): a lower CDTS represents more constrained data). Panel **b** shows the density distribution of the mean phastCons20 scores per 10bp bin. Panel **c** shows the distribution of log2 ratio (CNC score), of the reverse ranked CDTS (so a higher rank pertains to higher constraint but lower CDTS) and ranked phastCons20 scores, partitioned by regions of exon, intron and intergenic as defined by Ensembl v.92.

Supplementary Figure 2. Proportion of enriched neurologically-related GO terms in the gene set analysis compared between the annotation of interest (CNCRs) and the comparator annotation sets (a). Proportion of neurologically-related GO terms at CNCR density of 0.3 and above (b).

Supplementary Figure 3. Sanger sequencing of human hippocampus cDNA using targeted primers within *APOE*, aligned to hg38. Primers as listed in Supplementary Table 3.

Supplementary Tables

Supplementary Table 1. Annotation priority order for genomic feature. Genomic features are based on both Gencode and Ensembl. A priority order for annotation with a genomic feature is assigned to avoid conflict with overlapping features. The number of 10bp bins across the genome is also shown in the table.

Supplementary Table 2. Genome-wide association studies used in the stratified LDSC analysis. The GWAS for Parkinson's disease and major depressive disorder do not incorporate 23&Me data.

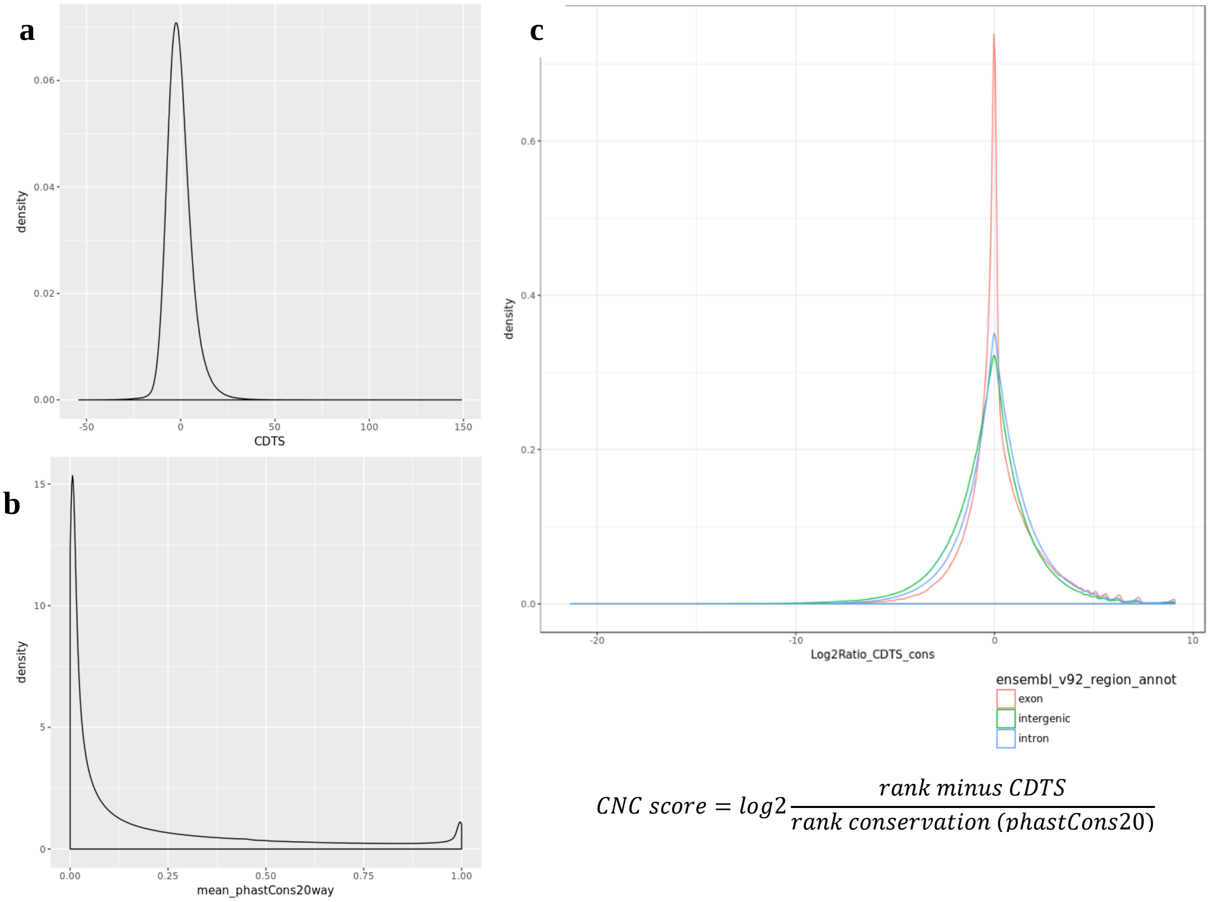
Supplementary Table 3. Primer positions and sequences used to validate the *APOE* intron-3 retention event.

Supplementary Table 4. Results for heritability, enrichment, and regression coefficient from stratified LDSC analysis. The coefficient p-values are one-sided p-values calculated from the coefficient Z-score.

Supplementary Table 5. Significantly enriched nervous system-related GO terms for CNCRs at density of 0.3. P-value relates to the p-value for enrichment calculated using g:Profiler and its own g:SCS correction method[28].

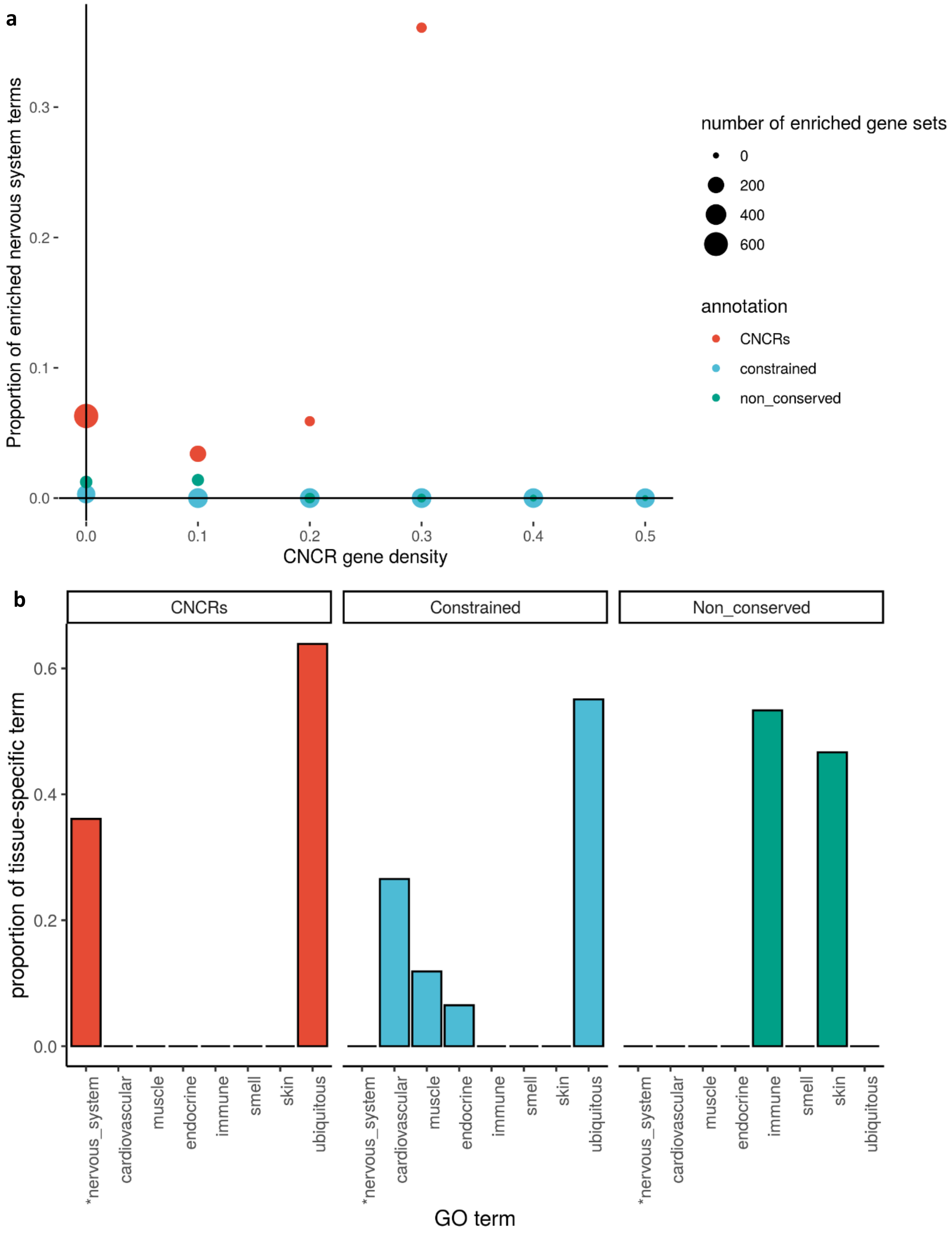
SUPPLEMENTAL DATA

Supplementary Figures

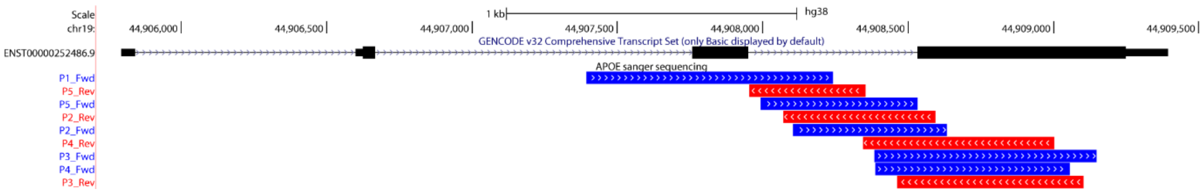


$$CNC\ score = log2\frac{rank\ minus\ CDTS}{rank\ conservation\ (phastCons20)}$$

Supplementary Figure 1. Kernal density plots of annotation metrics.



Supplementary Figure 2. Proportion of enriched neurologically-related GO terms in the gene set analysis compared between the annotation of interest (CNCRs) and the comparator annotation sets (a). Proportion of neurologically-related GO terms at CNCR density of 0.3 and above (b).



Supplementary Figure 3. Sanger sequencing of human hippocampus cDNA using targeted primers within *APOE*, aligned to hg38.

Supplementary Tables

Annotation priority order	Genomic feature	Number of 10bp bins	Description
1	Exon PCCDS	1,453,269	Exon, protein-coding sequence
2	Exon NCRNA	1,156,726	Exon, non-coding RNA, e.g. lincRNA
3	Exon PCUTR	892,210	Exon, protein-coding UTR
4	Promoter	820,321	Promoter
5	Promoter Flanking	1,074,641	Cluster with promoters or distal cis-regulatory elements
6	Enhancer	251,636	Enhancer
7	Intron, cis	108,670	Introns located in genes <10bp from splice-site
8	Intron, trans	15,204,447	Introns located in genes >10bp from splice-site
9	Intergenic	689,419	Not annotated in GenCode/ Ensembl
10	H3K9me3	2,082,553	Only overlap with H3K9me3
11	H3K27me3	777,409	Only overlap with H3K27me3
12	Multiple histones	5,199,455	Overlap with a combination of histone marks
13	Other	1,404,860	Includes open chromatin and unannotated features

Supplementary Table 1. Annotation priority order for genomic feature.

Disease	Author, Year, Reference	n case
Intelligence	Savage, 2018[23]	269,858
Alzheimer’s disease (AD)	Jansen, 2018[24]	71,880
Parkinson’s disease (PD)	Nalls, 2019 (excluding 23&Me data)[25]	33,674
Major depressive disorder (MDD)	Wray, 2018 (excluding 23&Me data)[27]	59,851
Schizophrenia (SCZ)	Pardiñas, 2018[26]	40,675

Supplementary Table 2. Genome-wide association studies used in the stratified LDSC analysis.

Primer name	5’ – 3’ sequence	Strand	Chr: Start-End (hg38)
P1_Fwd	ACAAGGACACTCAATACATGC	+	19:44907289-44907309
P1_Rev	CAGAGACGAAGAAGGAGCTAG	-	19:44908338-44908358
P2_Fwd	GGTTCTAGCTTCCTCTTCCC	+	19:44908064-44908083
P2_Rev	CGCCTGCAGCTCCTTGGACAG	-	19:44908627-44908647
P3_Fwd	CCTAGCTCCTTCTTCGTCTC	+	19:44908337-44908356
P3_Rev	CTCGAACCAGCTCTTGAGG	-	19:44909130-44909148
P4_Fwd	CCTTCTTCGTCTCTGCCTC	+	19:44908344-44908362
P4_Rev	CTGCTCCTTCACCTCGTC	-	19:44909037-44909055
P5_Fwd	GTGAGTGTCCCCATCCTGG	+	19:44907953-4490771
P5_Rev	CTGCGGCCGAGAGGGCGGGAG	-	19:44908512-44908532

Supplementary Table 3. Primer positions and sequences used to validate the *APOE* intron-3 retention event.

Annotation	GWAS	Proportion SNPs	Proportion heritability	Enrichment	Enrichment p-value	Regression Coefficient	Coefficient Z-score	Z- score -log P-value
CNCR	**Intelligence 2018**	0.03071	0.33916	11.04414	5.12E-20	2.96E-07	10.05909	23.37797
Constrained		0.0547	0.441239	8.06649	3.20E-21	1.85E-07	9.413106	20.61827
Non-conserved		0.12551	0.329821	2.627846	1.32E-05	6.28E-08	5.125337	6.828264
CNCR	**AD 2019**	0.03071	0.398428	12.9741	0.009868	1.89E-08	1.960767	1.602875
Constrained		0.0547	0.532373	9.732543	0.001961	1.12E-08	1.964995	1.607173
Non-conserved		0.12551	-0.34052	-2.71312	0.216138	-8.51E-09	-1.58533	0.025233
CNCR	**PD 2019 (ex.23&Me)**	0.03071	0.334257	10.88446	0.001934	2.57E-08	2.76684	2.548194
Constrained		0.0547	0.367301	6.714792	0.008806	1.32E-08	2.080212	1.726928
Non-conserved		0.12551	0.149455	1.190777	0.856765	1.28E-10	0.036813	0.313975
CNCR	**MDD 2018 (ex.23&Me)**	0.03071	0.330293	10.7554	1.39E-07	1.13E-07	5.421715	7.529959
Constrained		0.0547	0.403657	7.379441	1.51E-08	6.29E-08	4.940762	6.409951
Non-conserved		0.12551	0.432541	3.446263	5.02E-04	3.84E-08	3.908254	4.332707
CNCR	**SCZ 2018**	0.03071	0.33881	11.03275	2.50E-16	6.53E-07	8.829352	18.27867
Constrained		0.0547	0.425132	7.772029	2.19E-17	4.04E-07	8.456392	16.86047
Non-conserved		0.12551	0.308866	2.460883	8.75E-04	1.18E-07	3.576297	3.758833

Supplementary Table 4. Results for heritability enrichment, and regression coefficient from stratified LDSC analysis.

GO ID	GO term description	P-value
GO:0048663	neuron fate commitment	5.46E-07
GO:0048665	neuron fate specification	0.0012
GO:0021510	spinal cord development	0.00129
GO:0021517	ventral spinal cord development	0.00175
GO:0021515	cell differentiation in spinal cord	3.64E-07
GO:0021953	central nervous system neuron differentiation	7.44E-05
GO:0021522	spinal cord motor neuron differentiation	3.48E-04
GO:0021520	spinal cord motor neuron cell fate specification	0.0479
GO:0021527	spinal cord association neuron differentiation	0.00533
GO:0021871	forebrain regionalization	7.91E-05
GO:0021978	telencephalon regionalization	0.00313
GO:0030902	hindbrain development	0.0337
GO:0021536	diencephalon development	0.045

Supplementary Table 5. Significantly enriched nervous system-related GO terms for CNCRs at density of 0.3. P-value relates to the p-value for enrichment calculated using g:Profiler and its own g:SCS correction method[28].