Science Advances
AAAS

advances.sciencemag.org/cgi/content/full/6/24/eaay8299/DC1

Supplementary Materials for

Incomplete annotation has a disproportionate impact on our understanding of Mendelian and complex neurogenetic disorders

David Zhang, Sebastian Guelfi, Sonia Garcia-Ruiz, Beatrice Costa, Regina H. Reynolds, Karishma D'Sa, Wenfei Liu, Thomas Courtin, Amy Peterson, Andrew E. Jaffe, John Hardy, Juan A. Botía, Leonardo Collado-Torres, Mina Ryten*

*Corresponding author. Email: mina.ryten@ucl.ac.uk

Published 10 June 2020, *Sci. Adv.* **6**, eaay8299 (2020)
DOI: 10.1126/sciadv.aay8299

The PDF file includes:

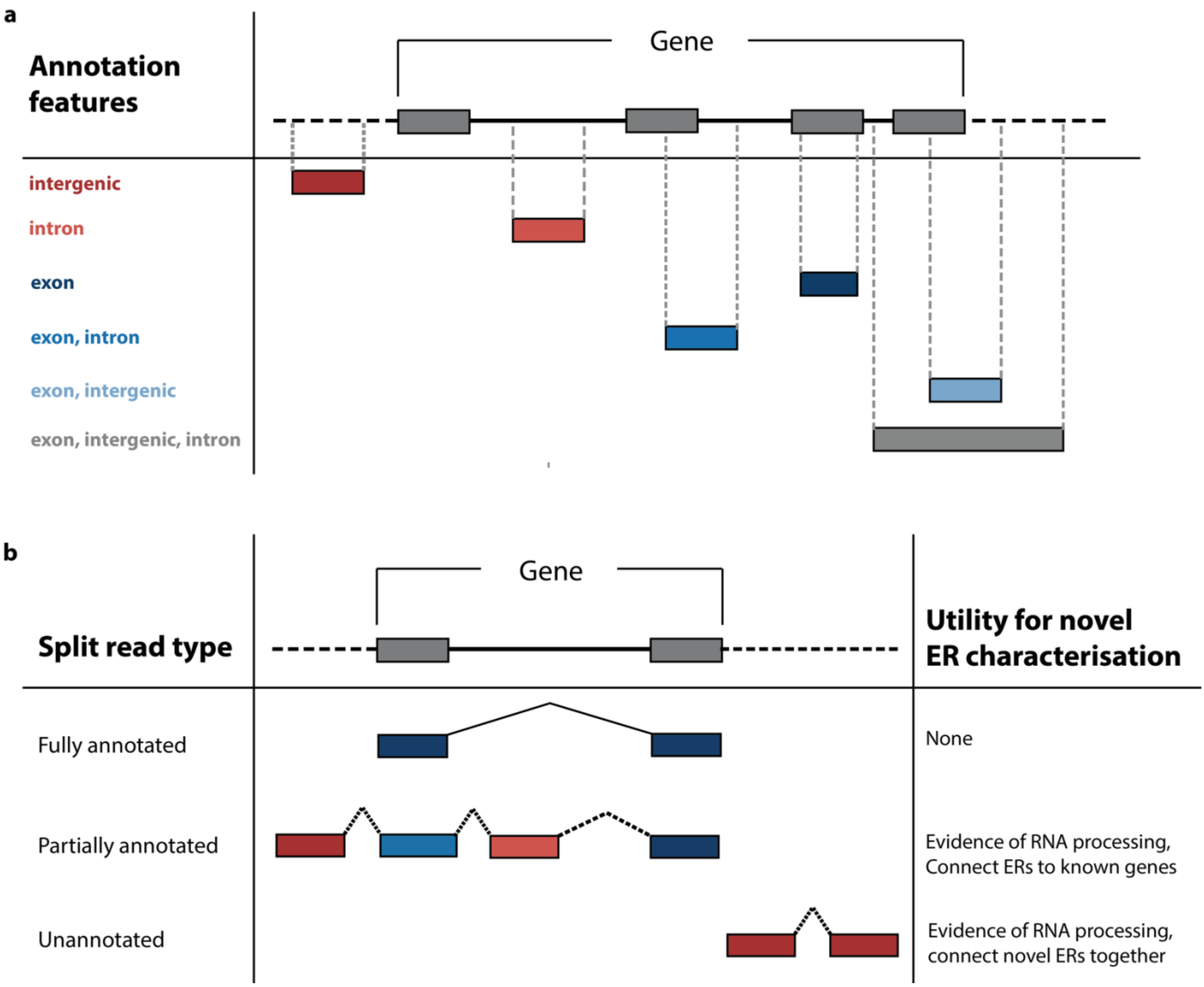
Figs. S1 to S6
Tables S1 and S3 to S5

Other Supplementary Material for this manuscript includes the following:

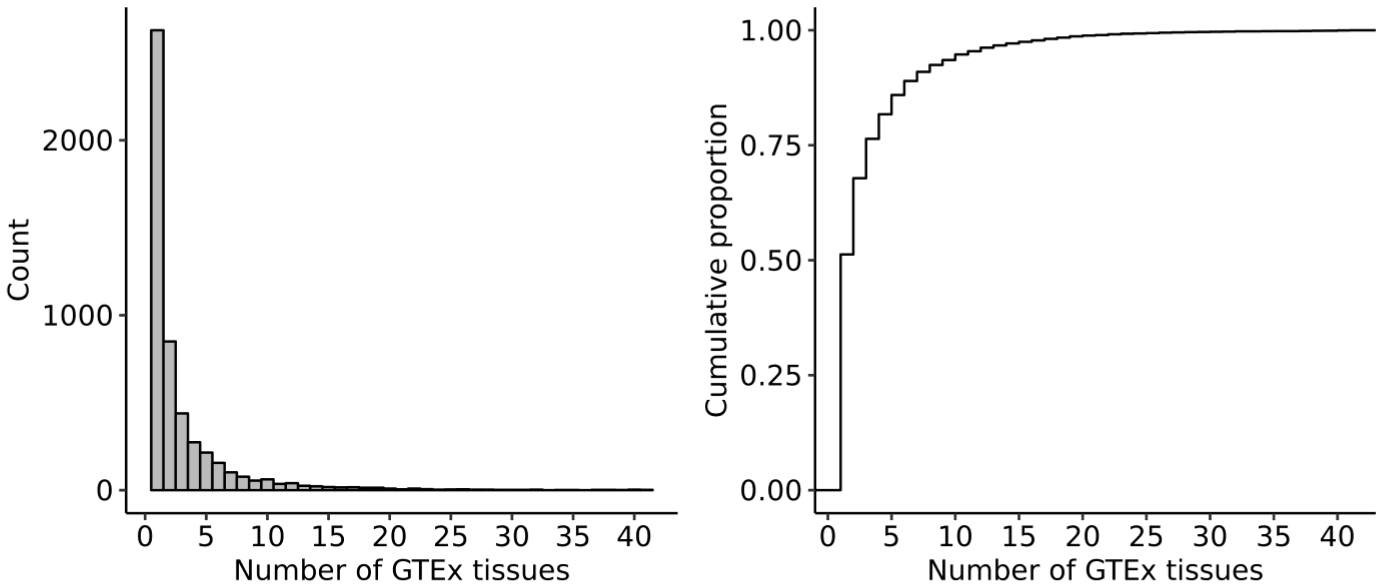
(available at advances.sciencemag.org/cgi/content/full/6/24/eaay8299/DC1)

Tables S2, S6, and S7

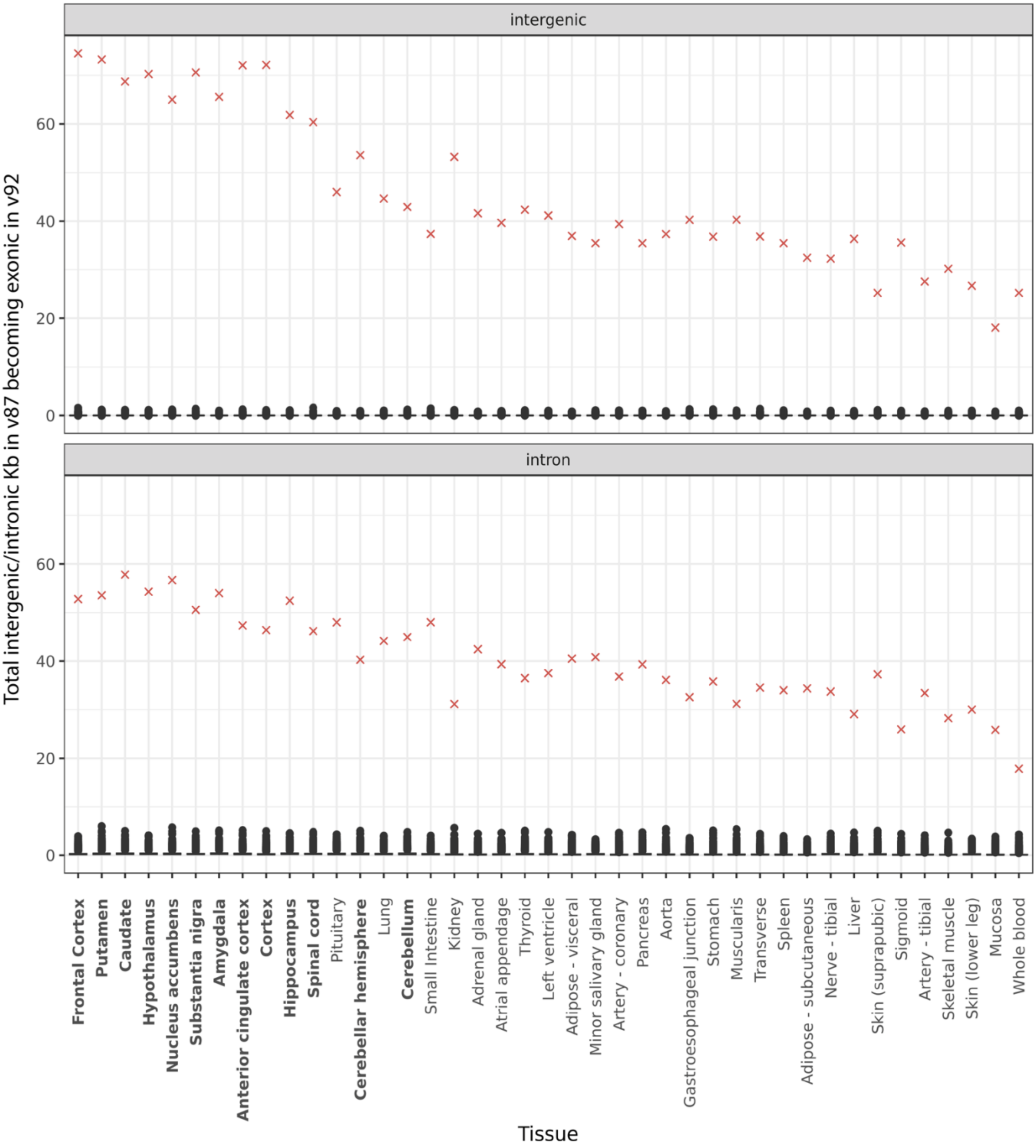
Supplementary figures



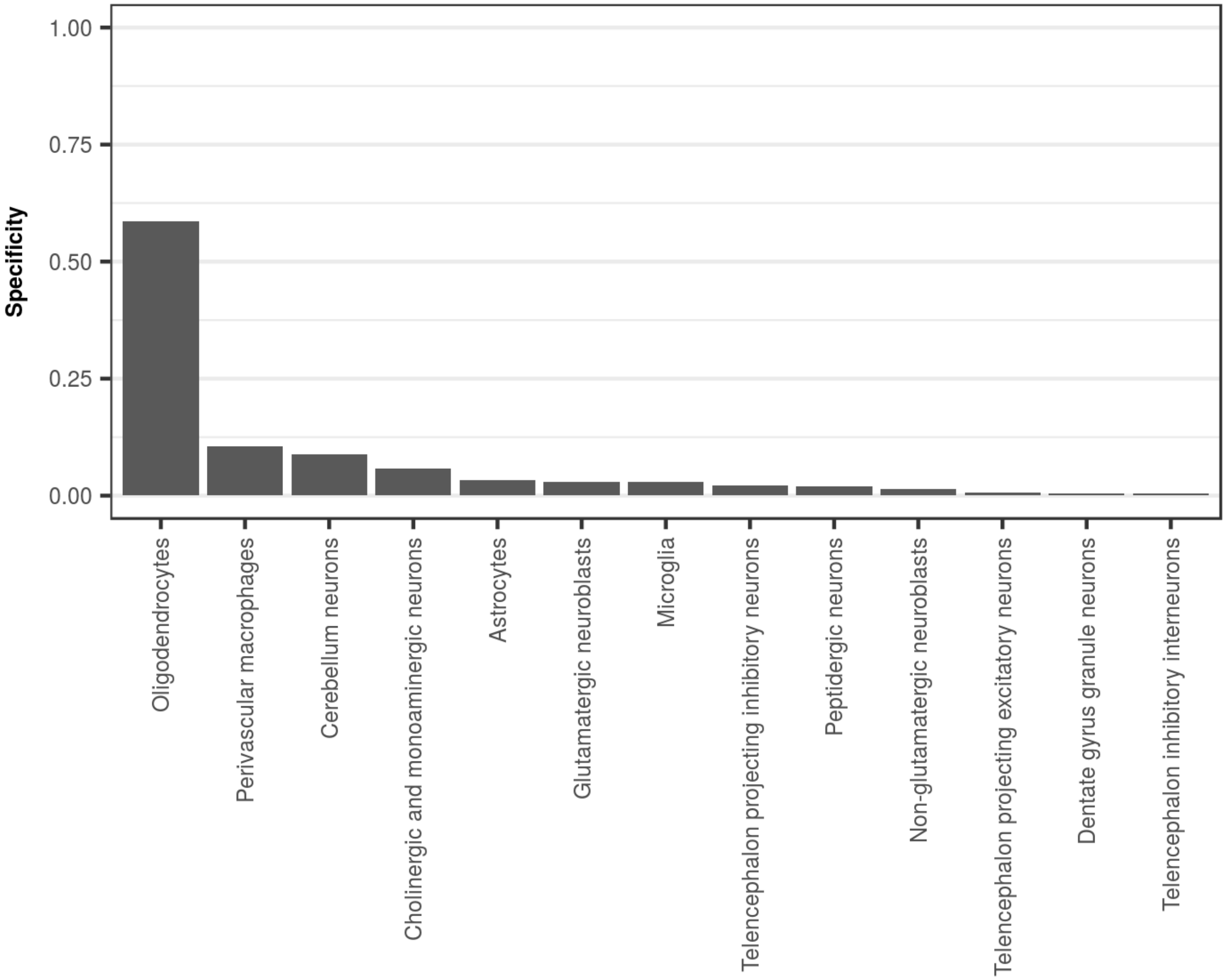
*Supplementary figure 1 – **Characterising ERs using Ensembl annotation features and split reads.** a) Illustration of the ER categorisation dependent on overlap with existing gene annotation. ERs in red are considered novel transcription. Blue ERs are those that overlap existing exons and are considered part of existing annotation. Grey ERs were uninformative and likely an artefact generated from genomic regions with high amounts of noise, pre-mRNA or overlapping genes, therefore were removed from all downstream analysis. b) Diagram showing the use of split reads (reads with a gapped alignment to the genome) to characterise novel ERs. Split reads were classified as annotated, partially annotated or unannotated dependent on whether the acceptor or donor sites both overlapped, only 1 of the acceptor or donor sites overlapped or neither overlapped known Ensembl v92 exon boundaries respectively. Partially annotated split reads were used to connect novel ERs to known genes. Partially annotated and unannotated split reads were used to provide evidence of RNA processing for novel ERs.*



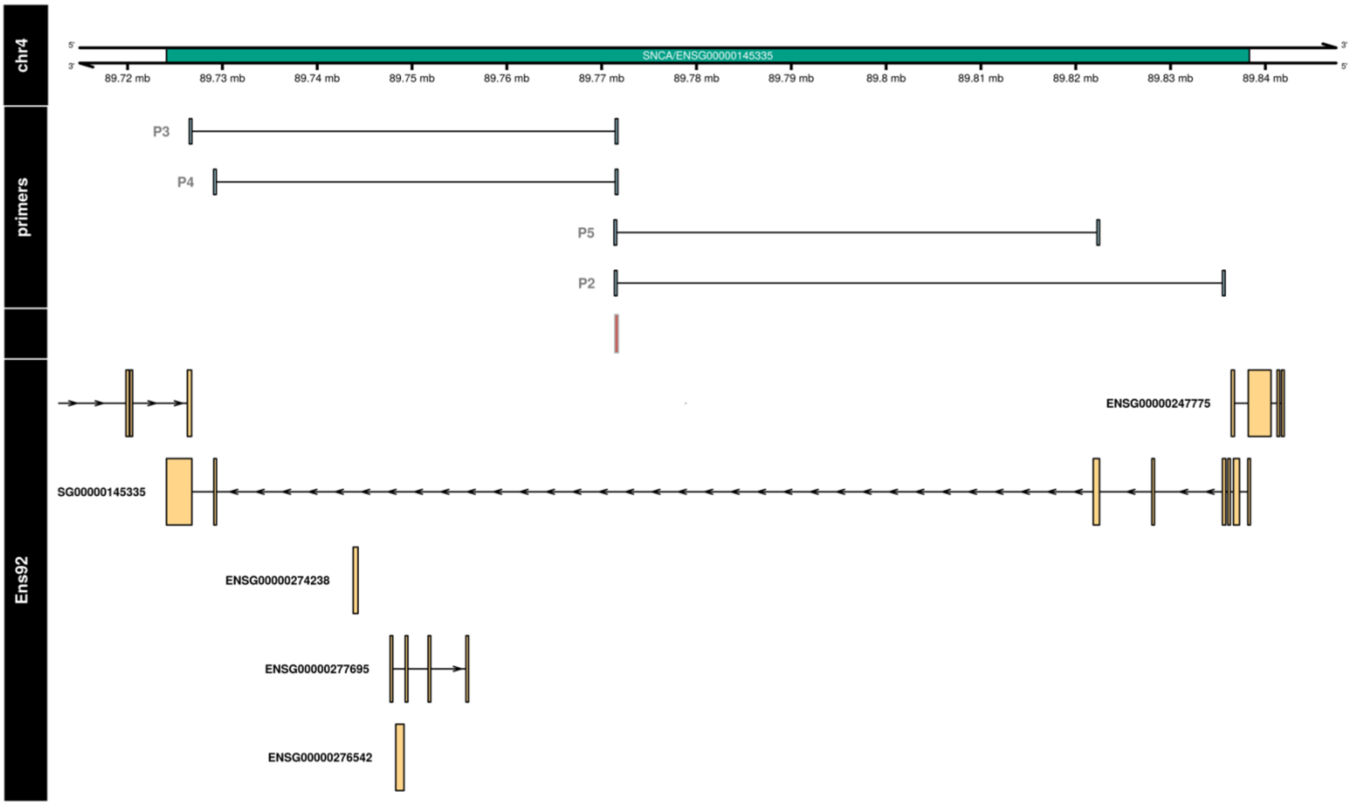
Supplementary figure 2 – ***Tissue specificity of novel ERs.*** *Taking all intronic and intergenic ERs that were intersected by two non-overlapping split reads, we inferred the precise boundaries of this set of 5,129 unique novel ERs. We then counted the number of tissues in which these ERs were detected. The majority (51.3%) of ERs were detected in only 1 tissue and 85.9% were detected in less than 5 tissues.*



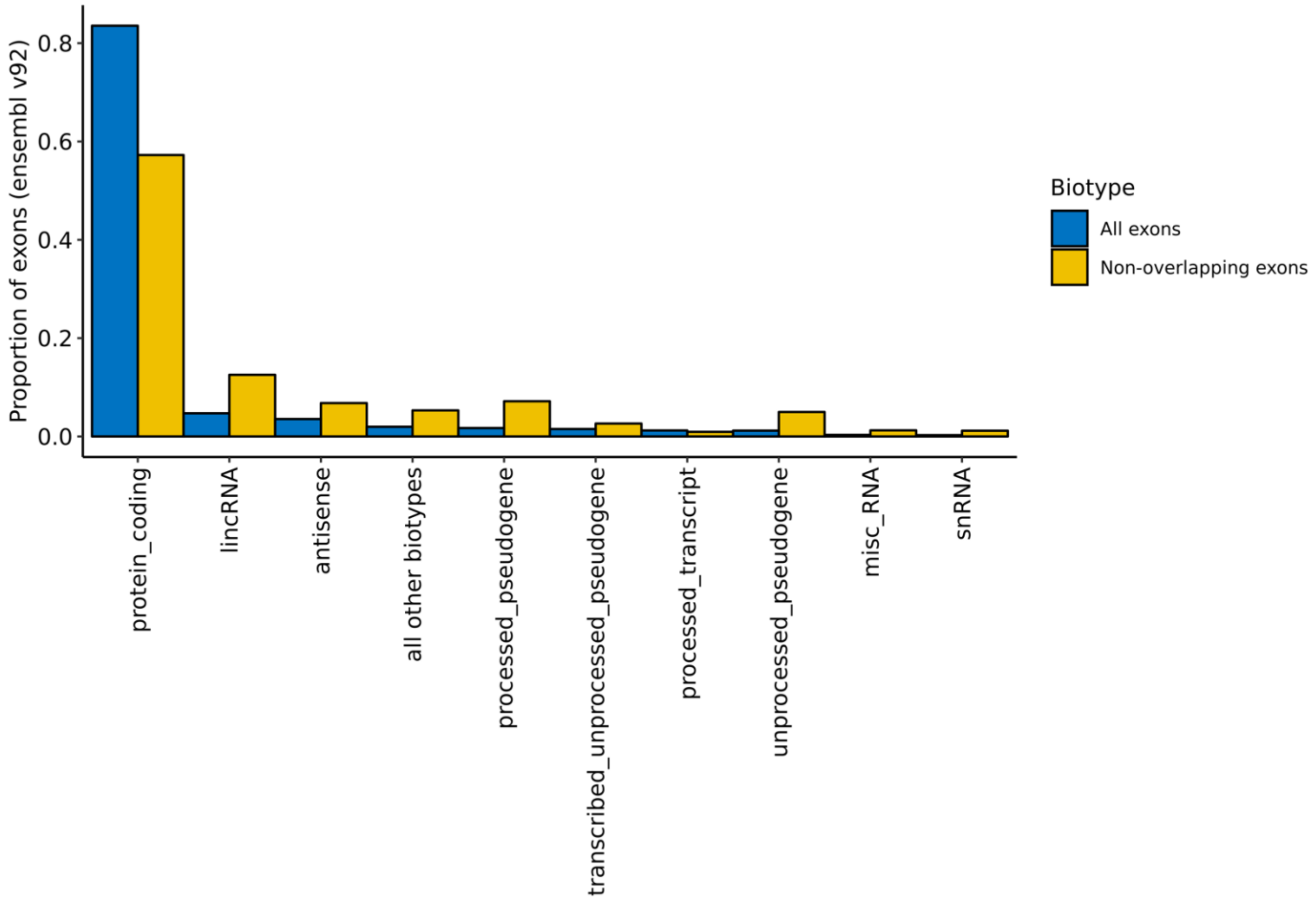
Supplementary figure 3 – ***Total Kb of novel ER entering Ensembl v92 annotation compared to random, length-matched intron and intergenic regions*** *For each of the 41 tissues, 10,000 random sets of intron and intergenic (with respect to Ensembl v87) regions were generated and length matched to the intron and intergenic ERs derived from that tissue. For all 10,000 sets, we counted the total Kb of regions that were now exonic in Ensembl v92, shown by distributions of black dots on the graph. Red "X"'s mark the actual total Kb of novel ERs for each tissue that were validated and one-sample Wilcoxon rank sum tests were used to test whether this quantity was significantly different from the randomised sets (all p-values < 2e-16).*



*Supplementary figure 4 – **MBP is specifically highly expressed in oligodendrocytes compared to other brain-specific cell-types** The specificity of MBP expression was calculated across various brain-specific cell-types. Oligodendrocytes had a 5x increased specificity of MBP expression compared to perivascular macrophages.*



Supplementary figure 5 – ***Primer locations for sanger sequence validation of SNCA novel exon*** *The genomic locations of the of the primers used for sanger sequence validation are displayed in relation to the SNCA gene structure and the novel exon (in red). P4 and P5 sequenced from the novel exon to flanking exons of SNCA, whilst P2 and P3 sequenced from the novel exon to the first and last coding exons of SNCA. Full details of primer sequences are found in supplementary table 4.*



Supplementary figure 6 – ***Proportion of exons that fall into different gene biotypes.*** *Comparison of the proportion of exons that are classified within the different gene biotypes between all exons from Ensembl v92 and the non-overlapping set of exons used to optimise the detection of transcription.*

Supplementary tables

Gene property	Estimate	P-value
Brain-specific	0.093	***
Transcript count	0.016	***
Gene length	4.18E-07	***
Gene biotype - protein coding	0.218	***
Gene biotype – lincRNA	-0.039	***
Gene biotype - processed pseudogene	-0.154	***
Gene biotype - unprocessed pseudogene	-0.093	***
Gene biotype – other	-0.113	***
Gene TPM	-2.62E-06	0.4
Overlapping gene	1	0.83

*** p <= 2e-16

Supplementary table 1 – ***Gene properties influencing re-annotation.*** *Gene characteristics such as brain specificity, transcript count, gene length, mean TPM and whether the gene overlapped with another were used to assess which genes were the most likely to be identified as re-annotated. Brain-specific, longer, protein-coding genes of high transcript complexity were the most likely to be re-annotated. Blue and red highlights positive and negative significant estimates, respectively.*

subgroup	pvalue	sign_change
other	0.671554608	-
neuropsychiatric	0.748582162	-
neurodegenerative	0.00405871	+
other neurological conditions	0.748582162	-

*Supplementary Table 3 - **Enrichment of re-annotated genes amongst neurological GWAS-associated genes.** Re-annotated genes were tested for enrichment amongst genes associated with neurological GWASs. Fisher's exact test was used for each comparison and the Bejamini-Hochberg (FDR) method was used to correct for multiple testing.*

neurodegenerative_disease	reannotated_gene_names
alzheimer disease	ABCA7;ACE;ACKR2;ANKRD55;APOC1;APOE;ARL17B;BCAM;BCAS3;BCHE;BCL3;BIN1;BLOC1S3;BZW2;CASS4;CD2AP;CD33;CEACAM19;CELF1;CELF2;CLASRP;CLPTM1;CLU;CR1;CRADD;CTNNA2;EED;EPHA1;ETS1;F5;FBXL7;FERMT2;GLIS3;IL1RAP;IL6R;INPP5D;KLK4;ME3;MEF2C;MIDN;MS4A4E;MTHFD1L;MYH7B;MYO16;NME8;NYAP1;PDE7B;PICALM;PMS2;PREX2;PTK2B;RIN3;RRAS2;SAP30L;SCARA3;SH3RF3;SORL1;SPATA20;SUCLG2;TNRC6A
amyotrophic lateral sclerosis	ADGRD1;ALCAM;C9orf72;CAMK1G;CTNND2;ITGA9;KIFAP3;MOB3B;OPCML;PFKP;TIAM1;UNC13A;WNT9A;ZNF783
creutzfeldt-jakob syndrome	NBPF3;NPAS2;PRNP;SERPINF2;ZDHHC2
frontotemporal lobar degeneration	NA
parkinson disease	ACMSD;ADAMTSL1;ARHGAP27;ATXN7L3;BST1;CD38;CNTN1;COL3A1;CPNE3;CTIF;DDRGK1;DSC2;FGD4;GAK;GCH1;HDAC5;HIP1R;HOOK3;IDUA;IGSF9B;INPP5F;KANSL1;KCNN3;LRRK2;MCCC1;PCGF3;PLEKHM1;PXDNL;RAB25;RIMS2;RIT2;SIPA1L2;SLC45A3;SNCA;SND1;SPPL3;STK39;TMEM229B;TPTE2;WNT3
prion diseases	NA
supranuclear palsy, progressive	ARHGAP27;CD8B;MAPT;MOBP;RPIA;STX6

*Supplementary Table 4 - **Neurodegenerative degenerative genes that are re-annotated split by GWAS.** The genes that were found to be reannotated and also a significant hit according to a neurodegenerative GWAS.*

primer_name	sequence	chr	strand	start	end	group
SNCA_PF3_ER_3	TCTCCTCTTACTTTGGCACTGG	chr4	-	89771561	89771582	P3
SNCA_PR3_3_ER	CTTCAGGTTCGTAGTCTTGATACCC	chr4	+	89726633	89726657	P3
SNCA_PF2_5_ER	GTGGCTGCTGCTGAGAAAACC	chr4	+	89771473	89771494	P2
SNCA_PR2_ER_5	CTCCAATTCTCGCCACTTCTGG	chr4	-	89835602	89835622	P2
SNCA_PF4	GGCATTTCATAAGCCTCATTGTC	chr4	+	89729201	89729223	P4
SNCA_PR4	ATCTCCTCTTACTTTGGCACTGG	chr4	-	89771561	89771583	P4
SNCA_PF5	AACATCAAAGGCGCTGGTTC	chr4	+	89771433	89771452	P5
SNCA_PR5	GCTGAGAAGACCAAAGAGCAAG	chr4	-	89822365	89822386	P5

Supplementary Table 5 - ***Primer positions and sequences used to experimentally validate the novel ER of SNCA.*** *The primers used for experimental validation of the novel expressed region found in SNCA and detailed in figure 6b.*