

# 1 Incomplete annotation of disease-associated genes is limiting our 2 understanding of Mendelian and complex neurogenetic disorders

3 David Zhang<sup>1\*</sup>, Sebastian Guelfi<sup>1\*</sup>, Sonia Garcia Ruiz<sup>1</sup>, Beatrice Costa<sup>1</sup>, Regina H. Reynolds<sup>1</sup>, Karishma D'Sa<sup>1</sup>,  
4 Wenfei Liu<sup>1</sup>, Thomas Courtin<sup>2</sup>, Amy Peterson<sup>3</sup>, Andrew E. Jaffe<sup>3,4,5,6,7,8</sup>, John Hardy<sup>1</sup>, Juan Botia<sup>1,9</sup>, Leonardo  
5 Collado-Torres<sup>3,4</sup> & Mina Ryten<sup>1</sup>

6  
7 1. Institute of Neurology, University College London (UCL), London, UK

8 2. Sorbonne Universités, UPMC Université Paris 06, UMR S 1127, Inserm U 1127, CNRS UMR 7225, ICM,  
9 Paris, France

10 3. Lieber Institute for Brain Development, Baltimore, Maryland, USA

11 4. Center for Computational Biology, Johns Hopkins University, Baltimore, Maryland, USA

12 5. Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland,  
13 USA

14 6. Department of Psychiatry and Behavioral Sciences, Johns Hopkins School of Medicine, Baltimore, MD,  
15 USA

16 7. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine,  
17 Baltimore, MD, USA

18 8. Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

19 9. Departamento de Ingeniería de la Información y las Comunicaciones, Universidad de Murcia, 30100,  
20 Murcia, Spain

21 \*These authors contributed equally.

## 2 Abstract

3 There is growing evidence to suggest that human gene annotation remains incomplete, with a disproportionate  
4 impact on the brain transcriptome. We used RNA-sequencing data from GTEx to detect novel transcription in  
5 an annotation-agnostic manner across 13 human brain regions and 28 human tissues. We found that genes  
6 highly expressed in brain are significantly more likely to be re-annotated, as are genes associated with  
7 Mendelian and complex neurodegenerative disorders. We improved the annotation of 63% of known OMIM-  
8 morbid genes and 65% of those with a neurological phenotype. We determined that novel transcribed regions,  
9 particularly those identified in brain, tend to be poorly conserved across mammals but are significantly  
10 depleted for genetic variation within humans. As exemplified by *SNCA*, we explored the implications of re-  
11 annotation for Mendelian and complex Parkinson's disease. We validated in silico and experimentally a novel,  
12 brain-specific, potentially protein-coding exon of *SNCA*. We release our findings as tissue-specific  
13 transcriptomes in BED format and via vizER: <http://rytenlab.com/browser/app/vizER>. Together these  
14 resources will facilitate basic genomics research with the greatest impact on neurogenetics.

## 35 Introduction

36

37 Genetic and transcriptomic studies are fundamentally reliant on accurate and complete human gene  
38 annotation. Gene definitions (namely genic coordinates and the isoforms/exons of which they are composed)  
39 are required for the quantification of expression or splicing from RNA-sequencing experiments, interpretation  
40 of significant genome-wide association studies (GWAS) signals and variant interpretation from genetic tests. As  
41 our understanding of transcriptomic complexity improves it is apparent that existing annotation remains  
42 incomplete even amongst known genes. Comparison of different gene annotation databases reveals that over  
43 17,000 Ensembl genes fall into intronic or intergenic regions according to the AceView database and  
44 predictably, the choice of reference annotation greatly influences the output of variant interpretation software  
45 such as VEP and ANNOVAR<sup>1,2</sup>. Thus, incomplete annotation may cause pathogenic variants to be overlooked  
46 within exonic regions that are yet to be annotated and limit our understanding of risk loci.

47 Importantly, the impact of incomplete annotation of the transcriptome may not be evenly distributed  
48 across all types of tissues or cells and there is reason to believe that improvements to gene annotation may  
49 have a disproportionate impact on the understanding of Mendelian and complex neurological diseases. This  
50 view is supported by several analyses of bulk RNA-sequencing data derived from human brain tissues, which  
51 have discovered transcription originating from intronic or intergenic regions (henceforth termed novel)<sup>3-5</sup>. In  
52 particular, Jaffe and colleagues found that as much as 41% of transcription in the human frontal cortex was  
53 novel. Furthermore, it is becoming increasingly clear that RNA processing is highly complex in the human  
54 central nervous system due to the expression of long genes and the large diversity of cell types present<sup>6,7</sup>. In  
55 combination these factors could result in rare yet important transcripts being overlooked.

56 In this study, we address this issue by leveraging publicly available transcriptomic data available through  
57 the Genotype-Tissue Expression Consortium (GTEx) to improve the annotation of genes across the genome. We  
58 define transcription in an annotation-agnostic manner using RNA-sequencing data from 13 regions of the  
59 human central nervous system and compare this to definitions generated from a further 28 GTEx non-brain  
60 tissues. While we discover novel transcription to be widespread across all tissues, it is most prevalent in human  
61 brain. We provide evidence to suggest that the additional annotations we generate are likely to be functionally  
62 important on the basis of the tissue and cell-type specificity of novel expressed regions (ERs), the significant  
63 depletion of genetic variation amongst humans within ERs and their protein coding potential. Finally, by  
64 combining novel expressed regions (ERs) with split read data, defined as reads that have a gapped alignment to  
65 the genome, we link these regions to known genes associated with Mendelian and complex neurodegenerative  
66 and neuropsychiatric disorders. We release our findings as tissue-specific transcriptomes in a BED format and  
67 in an online platform vizER (<http://www.rytenlab.com/browser/app/vizER>), which allows individual genes to  
68 be queried and visualised. Overall, we improve the annotation of 1929 (63%) OMIM genes and a further 317  
69 genes associated with complex neurodegenerative and neuropsychiatric disease. We anticipate that this will  
70 lead to improvements in diagnostic yield from whole genome sequencing (WGS) and the understanding of  
71 neurogenetic disorders.

## 72 Results

### 73 Optimising the annotation-agnostic detection of transcription using known exons

74 Pervasive transcription of the human genome, the presence of pre-mRNA even within polyA-selected  
75 RNA-sequencing libraries and variability in read depth complicates the identification of novel exons and  
76 transcripts using RNA-sequencing data<sup>8,9</sup>. With this in mind, we identified a set of exons with the most reliable  
77 and accurate boundaries (namely all exons from Ensembl v92 that did not overlap with any other exon<sup>10</sup>) and  
78 then used this exon set to calibrate the detection of transcription from 41 GTEx tissues<sup>11</sup>. We used the well-  
79 established tool, *derfinder*, to perform this analysis<sup>12</sup>. However, we noted that while *derfinder* enables the  
80 detection of continuous blocks of transcribed bases termed expressed regions (ERs) in an annotation-agnostic  
81 manner, the mean coverage cut-off (MCC) applied to determine transcribed bases is difficult to define and  
82 variability in read depth even across an individual exon can result in false segmentation of blocks of expressed  
83 sequence. Therefore, in order to improve our analysis and define ERs more accurately, we applied *derfinder*,  
84 but with the inclusion of an additional parameter we term the max region gap (MRG), which merges adjacent  
85 ERs (see detailed Methods). Next, we sought to identify the optimal values for MCC and MRG using our learning  
86 set of known, non-overlapping exons.

87 This process involved generating 506 transcriptome definitions for each tissue using unique pairs of  
88 MRCs and MRGs, resulting in a total of 20,746 transcriptome definitions across all 41 tissues. For each of the  
89 20,746 transcriptome definitions, all ERs that intersected non-overlapping exons were extracted and the  
90 absolute difference between the ER definition and the corresponding exon boundaries, termed the exon delta,  
91 was calculated (**Figure 1a**). We summarised the exon delta for each transcriptome using two metrics, the  
92 median exon delta and the number of ERs with exon delta equal to 0. The median exon delta represents the  
93 overall accuracy of all ER definitions, whereas, the number of ERs with exon delta equal to 0 indicates the  
94 extent to which ER definitions precisely match overlapping exon boundaries. The MCC and MRG pair that  
95 generated the transcriptome with the lowest median exon delta and highest number of ERs with exon delta  
96 equal to 0 was chosen as the most accurate transcriptome definition for each tissue. Across all tissues, 50-54%  
97 of the ERs tested had an exon delta = 0, suggesting we had defined the majority of ERs accurately. Taking the  
98 cerebellum as an example and comparing ER definitions to those which would have been generated applying  
99 the default *derfinder* parameters used in the existing literature (MCC: 0.5, MRG: None equivalent to 0), we noted  
00 an 96bp refinement in ER size, equating to 67% of median exon size (**Figure 1b &1c**). In summary, by using  
01 known exons to calibrate the detection of transcription, we generated more accurate annotation-agnostic  
02 transcriptome definitions for 13 regions of the CNS and a further 28 human tissues.

### 03

### 04 Novel transcription is most commonly observed in the central nervous system

05 To assess how much of the detected transcription was novel, we calculated the total size in base pairs of  
06 ERs that did not overlap known annotation. ERs were then categorised with respect to the genomic features  
07 with which they overlapped as defined by the Ensembl v92 reference annotation (exons, introns, intergenic;  
08 **Supplementary Figure 1a**). Those that solely overlapped intronic or intergenic regions were classified as  
09 novel. We discovered 8.4 to 22Mb of potentially novel transcription across all tissues, consistent with previous

10 reports that annotation remains incomplete<sup>13,14</sup>. Novel ERs predominantly fell into intragenic regions  
11 suggesting that we were preferentially improving the annotation of known genes, rather than identifying new  
12 genes (**Figure 2a**). Although novel transcription was found to be ubiquitous across tissues, the abundance  
13 varied greatly between tissues (**Figure 2b, 2d, 2e**). To investigate this further, we calculated the coefficient of  
14 variation for exonic, intronic and intergenic ERs. We found that the levels of novel transcription varied 3.4-7.7x  
15 more between tissues than the expression of exonic ERs (coefficient of variation of exonic ERs: 0.066Mb,  
16 intronic ERs: 0.222Mb, intergenic ERs: 0.481Mb). Furthermore, focusing on a subset of novel ERs for which we  
17 could infer the precise boundaries of the presumed novel exon (using intersecting split reads), we found that  
18 more than half of these ERs were detected in only 1 tissue and that 86.3% were found in less than 5 tissues  
19 (**Supplementary Figure 2a**). Even when restricting to ERs derived from only the 13 CNS tissues, 34.3% were  
20 specific to 1 CNS region (**Supplementary Figure 2b**). This suggests that novel ERs are largely derived from  
21 tissue-specific transcription, potentially explaining why they had not already been discovered.

22 This finding lead us to hypothesise that genes highly expressed in brain would be amongst the most  
23 likely to be re-annotated due to the difficulty of sampling human brain tissue, the cellular heterogeneity of this  
24 tissue and the particularly high prevalence of alternative splicing<sup>3</sup>. As we predicted, the quantity of novel  
25 transcription found within brain was significantly higher than non-brain tissues (p-value: 2.35e-10) (**Figure 2e**  
26 & **2f**). In fact, ranking the tissues by descending Mb of novel transcription demonstrated that tissues of the CNS  
27 constituted 13 of the top 14 tissues. Interestingly, the importance of improving annotation in the human brain  
28 tissue was most apparent when considering purely intergenic ERs and ERs that overlapped exons and extended  
29 into intergenic regions (**Figure 2d & 2e**).

30 This observation raised the question of whether there were specific genic features, which could be used  
31 to predict which genes were most likely to be re-annotated (connected to a novel ER). We used logistic  
32 regression to determine whether specific properties, including measures of structural gene complexity and  
33 specificity of expression to brain increased the likelihood of re-annotation. We also accounted for factors which  
34 might be expected to contribute to errors in ER identification, including whether the gene overlapped with  
35 another known gene making attribution of reads more complex. We found that the annotation of brain-specific  
36 genes and those with higher transcript complexity were more likely to have evidence for incomplete annotation  
37 (**Table 1**). Importantly, overlapping genes were not significantly more likely to be re-annotated (taking into  
38 account gene length), demonstrating that novel transcription is not merely a product of noise from intersecting  
39 genes. Taken together these findings demonstrated that widespread novel transcription is found across all  
40 human tissues, the quantity of which varies extensively between tissues. CNS tissues displayed the greatest  
41 quantity of novel transcription and accordingly, genes highly expressed in the human brain are most likely to  
42 be re-annotated.

43

#### 44 *Validation of novel transcription across Ensembl versions and within an independent dataset*

45 We recognise that a proportion of novel transcription may originate from technical variability or pre-  
46 mRNA contamination. Therefore, we assessed the reliability of novel ERs by classifying ERs using different  
47 versions of Ensembl and through an independent dataset. Firstly, we measured how many Kb of the  
48 transcription we detected would have been classified as novel with respect to Ensembl v87, but was now

19 annotated in Ensembl v92 and found that across all tissues an average of 68Kb (43-127Kb) had changed status.  
20 This value was 5.3x (3.2-10.1x) greater in every tissue compared to the Kb of ERs overlapping exons in  
21 Ensembl v87 that had become purely intronic or intergenic in Ensembl v92 (**Figure 3a**). To further assess  
22 whether this was greater than what would be expected by chance, we compared the total Kb of novel ERs  
23 entering v92 annotation for each tissue to 10,000 sets of random length-matched intronic and intergenic  
24 regions. For all tissues, the total Kb of both intronic and intergenic ERs that were now annotated in Ensembl  
25 v92 was significantly higher than the total Kb distribution of the randomised negative control regions, implying  
26 a high validation rate of novel ERs (**Supplementary figure 3**). Notably, brain regions had significantly higher  
27 Kb of ERs entering Ensembl v92 annotation from Ensembl v87 than non-brain tissues, even when subtracting  
28 the Kb of ERs leaving Ensembl v87 (p-value: 7.6e-9), suggesting the greater abundance of brain-specific novel  
29 transcription was not purely attributed to increased transcriptional noise.

30 While our analysis of novel ERs across different Ensembl versions provided a high level of confidence in  
31 the quality of ER calling, it was limited to ERs which had already been incorporated into annotation and did not  
32 provide an overall indication of the rate of validation across all ERs. Therefore, we investigated whether our  
33 GTEx frontal cortex derived ERs could also be discovered in an independent frontal cortex dataset reported by  
34 Labadord and colleagues<sup>15</sup>. As expected, ERs which overlapped with annotated exons had near complete  
35 validation (>=89%), but importantly 62% of intergenic and 70% of intronic ERs respectively were also  
36 detected in the second independent frontal cortex dataset (**Figure 3b**). While this high validation rate implied  
37 the majority of all ERs were reliably detected, we investigated whether a subset of ERs supported with  
38 evidence of RNA splicing as well as transcription would have even better rates of validation. Evidence of  
39 transcription is provided by the coverage data derived using derfinder, whilst split reads, which are reads with  
40 a gapped alignment to the genome provide evidence of the splicing out of an intron (**Supplementary figure**  
41 **1b**). With this in mind, we focused our attention on the putative spliced ERs as indicated by the presence of an  
42 overlapping split read. Consistent with expectation, we found that ERs with split read support had higher  
43 validation rates than ERs lacking this additional feature. This increase in validation rate for ERs with split read  
44 support was greatest for intergenic and intronic ERs with the validation rate rising to 87% for intergenic ERs  
45 and 88% for intronic ERs (as compared to 99% for ERs overlapping exons, **Figure 3b**). Even when considering  
46 this set of highly validated ERs with split read support, 1.7-3.8Mb of intronic and 0.5-2.2Mb of intergenic  
47 transcription was detected across all 41 tissues. Thus, in summary, the majority of novel ERs were reliably  
48 detected and validated in an independent dataset.  
49

50 *Unannotated expressed regions are depleted for genetic variation and some have the potential to  
51 be protein coding suggesting they are functionally significant*

52 Given recent reports suggesting widespread transcriptional noise and acknowledging that  
53 transcription, even when tissue-specific, does not necessarily translate to function we investigated whether  
54 novel ERs were likely to be of functional significance using measures of both conservation and genetic  
55 constraint<sup>14,16</sup>. The degree to which a base is evolutionarily conserved across species is strongly dependent on  
56 its functional importance and accordingly, conservation scores have been used to aid exon identification<sup>17</sup>.  
57 However, this measure is unable to capture genomic regions of human-specific importance. Thus, we

38 investigated novel ERs not only in terms of conservation but also genetic constraint. Constraint scores,  
39 measured here as a context-dependent tolerance score (CDTS), represent the likelihood a base is mutated  
40 within humans<sup>18</sup>. By comparing our detected novel ERs to 10,000 randomised sets of length-matched intronic  
41 and intergenic regions, we found that both intronic and intergenic ERs were significantly less conserved, but  
42 more constrained than expected by chance (p-value < 2e-16, **Figure 4a**). This would suggest that they have an  
43 important functional role specifically in humans. Furthermore, considering the importance of higher-order  
44 cognitive functions in differentiating humans from other species, we measured the constraint of brain-specific  
45 novel ERs separately on the basis that these ERs may be the most genetically constrained of all novel ERs  
46 identified. Indeed, we found that brain-specific novel ERs were even more constrained than other novel ERs,  
47 supporting the view that improvements in gene annotation are likely to have a disproportionate impact on our  
48 understanding of human brain diseases.

49 Another metric of functional importance is whether a region of the genome is translated into protein  
50 and notably the vast majority of all known Mendelian disease mutations fall within protein-coding regions. For  
51 this reason, we investigated whether novel ERs could potentially encode for proteins. Here, we focused on the  
52 subset of novel ERs which had evidence of splicing, since the overlapping split reads can be used to assign the  
53 precise boundaries of ERs, allowing us to confidently retrieve the DNA sequence and corresponding amino acid  
54 sequence for each novel ER. A total of 2,961 ERs covering 274Kb was found to be potentially protein coding,  
55 which represented 57% of the ERs analysed (**Figure 4b**). Amongst this set of ERs with protein coding potential,  
56 758 ERs also fell within the top 20% of most constrained regions of the genome. These ERs connect to 694  
57 genes, 30% of which are expressed specifically in the CNS (**Supplementary table 1**). Overall, we discovered  
58 that novel ERs broadly are likely to have a human-specific function. We also identified an important subset of  
59 novel ERs that have protein coding potential and are highly depleted for genetic variation in humans. Together,  
60 this suggested that at least a proportion of novel ERs are functionally significant.

61

## 62 *Incomplete annotation of brain-specific genes may be limiting our understanding of specific 63 CNS-relevant cell types and complex diseases*

64 Given that we discovered the greatest abundance of novel transcription amongst brain tissues, we  
65 investigated whether this may be impacting on our understanding of certain cell types within the brain more  
66 than others. We tested this by calculating whether our set of 2962 re-annotated brain-specific genes were  
67 significantly enriched for cell-type specific genes, when compared to the background list of 2422 brain-specific  
68 genes without re-annotations. Of the 13 brain-specific cell types considered, genes specifically expressed by  
69 oligodendrocytes had the largest difference in enrichment (p-value of re-annotated: <2e-16; not re-annotated:  
70 0.169), suggesting incomplete annotation was disproportionately limiting our understanding of this cell type  
71 (**Figure 5a**). For example, we found that *MBP*, which encodes for myelin basic protein, was amongst those  
72 genes re-annotated and with an oligodendrocyte-specific expression profile (**Supplementary figure 4**). In fact,  
73 we detected a 48bp ER specific to cortex and striatal tissues (anterior cingulate cortex, cortex, frontal cortex,  
74 nucleus accumbens, putamen), which was connected to two flanking protein-coding exons of *MBP*. The ER itself  
75 had protein-coding potential and evidence of functional importance specifically in humans, as demonstrated by  
76 low mammalian sequence conservation but depletion of genetic variation within humans (phasCons7: 0.03, top

20% CDTs) (**Figure 5b**). This finding is interesting because *MBP* specifically and oligodendroglial dysfunction more generally, have been implicated in a number of neurodegenerative disorders, including multiple system atrophy, which is characterised by myelin loss and degeneration of striatum and cortical regions<sup>19</sup>, as well as schizophrenia and Parkinson's disease<sup>19-21</sup>.

These observations led us to postulate whether incomplete annotation could also be hindering our understanding of complex neurodegenerative and neuropsychiatric disorders. Therefore, we assessed whether our list of re-annotated genes was enriched for genes associated with complex forms of neurodegenerative, neuropsychiatric or other neurological conditions. This analysis was performed by using the Systematic Target Opportunity assessment by Genetic Association Predictions (STOPGAP) database, which provides an extensive catalogue of human genetic associations mapped to effector gene candidates (see detailed methods)<sup>22</sup>. Interestingly, we found that genes associated with neurodegenerative disorders were significantly over-represented within our re-annotated set (p-value: 0.004, **Supplementary Table 2**). In particular, important neurodegenerative disease genes such as *SNCA*, *APOE* and *CLU* were amongst those re-annotated, suggesting that despite being extensively studied the annotation of these genes remains incomplete (complete list found in **Supplementary Table 3**). Thus, we demonstrate that incomplete annotation of brain-specific genes may be hindering our understanding of specific cell types and complex neurodegenerative disorders.

13

#### 14 *Incomplete annotation of OMIM genes may limit genetic diagnosis, particularly for* 15 *neurogenetic disorders*

Since re-annotation of genes already known to cause Mendelian disease would have a direct impact on clinical diagnostic pipelines, we specifically assessed this gene set. Novel ERs were first connected to known genes using split reads (**Supplementary figure 1b**). Next, we filtered for OMIM-morbid genes and found that 63% of this set of OMIM-morbid genes were re-annotated and 14% were connected to a potentially protein-coding ER, suggesting that despite many of these genes having been extensively studied, the annotation of many OMIM-morbid genes remains incomplete (**Figure 6a**). Given that OMIM-morbid genes often produce abnormalities specific to a given set of organs or systems, we investigated the relevance of novel transcription to disease by matching the human phenotype ontology (HPO) terms obtained from the disease corresponding to the OMIM-morbid gene, to the GTEx tissue from which ERs connected to that gene were derived. We discovered that 72% of re-annotated OMIM-morbid genes had an associated novel ER originating from a phenotypically relevant tissue (**Figure 6b**). This phenomenon was exemplified by the OMIM-morbid gene *ERLIN1*, which when disrupted is known to cause spastic paraplegia 62 (SPG62), an autosomal recessive form of spastic paraplegia, which has been reported in some families to cause not only lower limb spasticity, but also cerebellar abnormalities<sup>23</sup>. We detected a cerebellar-specific novel ER that was intronic with respect to *ERLIN1*. The novel ER had the potential to code for a non-truncated protein and connected through intersecting split reads to two flanking, protein-coding exons of *ERLIN1*, supporting the possibility of this ER being a novel protein-coding exon. Furthermore, the putative novel exon was highly conserved (phastcons7 score: 1) and was amongst the top 30% most constrained regions in the genome, suggesting it is functionally important both across mammals and within humans (**Figure 6c**).

55 Similarly, we detected a brain-specific novel ER in the long intron of the gene *SNCA*, which encodes alpha-  
56 synuclein protein implicated in the pathogenesis of Mendelian and complex Parkinson's disease. This ER  
57 connected to two flanking protein-coding exons through split reads (**Figure 6d**) and appeared to also have  
58 coding potential. Interestingly, while the ER sequence is not conserved within mammals (phastcons7 score:  
59 0.09) or primates (phastcons20 score: 0.21), it is in the top 19% of most constrained regions in the genome  
60 suggesting it is of functional importance specifically in humans. We validated the existence of this ER both in  
61 silico and experimentally. The expression of this ER was confirmed in silico using an independent frontal cortex  
62 dataset reported by Labadord and colleagues<sup>15</sup>. Using Sanger sequencing, we validated the junctions  
63 intersecting the ER and the flanking exons in RNA samples originating from pooled human frontal cortex  
64 samples (**Supplementary Figure 5**). In order to gain more information about the transcript structure in which  
65 the novel ER was contained, we also performed Sanger sequencing from the first (ENSE00000970013) and last  
66 coding exons (ENSE00000970014) of *SNCA* to the novel ER. This implied a full transcript structure containing a  
67 minimum of 609bp with the novel ER predicted to add an additional 63 amino acids (45% of existing transcript  
68 size). This example highlights the potential of incomplete annotation to both hinder genetic diagnosis and limit  
69 our understanding of a common complex neurological disease. Variants located in the novel ER linked to *SNCA*  
70 would not be captured using whole exome sequencing (WES) and if identified in WGS or through GWAS  
71 would be misassigned as non-coding variants.

72

73

## 33 Discussion

34

35 In this study, we demonstrate that novel transcription though commonly detected across all tissues,  
36 disproportionately affects genes highly expressed in the brain and some brain-specific cell types, namely  
37 oligodendrocytes. We provide evidence to suggest that novel ERs are functionally important, since they are  
38 more depleted for genetic variation within humans than would be expected by chance and some have the  
39 potential to code for protein. Furthermore, we find that genes known to cause Mendelian and complex  
40 neurodegenerative disorders are enriched amongst the set of genes we reannotate. In order to illustrate the  
41 potential impact of incomplete annotation on such disorders, we highlight the specific example of *SNCA*, a gene  
42 implicated in Mendelian and complex Parkinson's disease. We experimentally validate the existence of a novel  
43 transcript of *SNCA* containing a potentially protein coding novel ER. Together, this suggests that incomplete  
44 annotation may be limiting our understanding of both Mendelian and common complex diseases.

45 We find that the majority of probable novel exons we detect have a restricted expression pattern across  
46 tissues. The practical difficulty of accessing the brain reduces the number of available brain-specific datasets  
47 and its regional and cellular heterogeneity is one of the factors driving the high number of brain-specific  
48 transcripts. Furthermore, since our approach does not depend on conservation across species to annotate  
49 novel exons, we are able to identify ERs which are likely to be of human-specific importance<sup>18</sup>. In fact, we find  
50 that brain-specific ERs have the highest constraint scores, emphasising their specific importance in humans.  
51 Together these factors suggest that the resource we have generated will have the greatest impact on  
52 neurogenetic disorders.

53 Finally, we release our results through a dedicated web resource, vizER  
54 (<http://rytenlab.com/browser/app/vizER>), which enables individual genes to be queried for incomplete  
55 annotation as well as the download of all novel ER definitions. We believe that vizER will be an important  
56 resource for clinical scientists in the diagnosis of Mendelian disorders, neuroscientists studying individual gene  
57 structures and functions, and with the emergence of larger long read sequencing data sets will accelerate novel  
58 transcript discovery particularly in human brain.

## 9 Online Methods

### 10 *OMIM data*

11 Phenotype relationships and clinical synopses of all Online Mendelian Inheritance in Man (OMIM) genes were  
12 downloaded using <http://api.omim.org> on the 29<sup>th</sup> of May 2018<sup>24</sup>. OMIM genes were filtered to exclude  
13 provisional, non-disease and susceptibility phenotypes retaining 2,898 unique genes that were confidently  
14 associated to 4,034 Mendelian diseases. Phenotypic abnormality groups were linked to corresponding affected  
15 Genotype-Tissue Expression (GTEx) tissues through manual inspection of the HPO terms within each group by  
16 a medical specialist<sup>11</sup>.

17

### 18 *GTEx data*

19 RNA-seq data in base-level coverage format for 7,595 samples originating from 41 different GTEx tissues was  
20 downloaded using the R package recount version 1.4.6<sup>4</sup>. Cell lines, sex-specific tissues and tissues with 10  
21 samples or below were removed. Samples with large chromosomal deletions and duplications or large CNVs  
22 previously associated with disease were filtered out (smafrze = "USE ME"). Coverage for all remaining samples  
23 was normalised to a target library size of 40 million 100bp reads using the area under coverage value provided  
24 by recount2. For each tissue, base-level coverage was averaged across all samples to calculate the mean base-  
25 level coverage. GTEx split read data, defined as reads with a non-contiguous gapped alignment to the genome,  
26 was downloaded using the recount2 resource and filtered to include only split reads detected in at least 5% of  
27 samples for a given tissue and those that had available donor and acceptor splice sequences.

28

### 29 *Optimising the detection of transcription*

30 Transcription was detected across 41 GTEx tissues using the package derfinder version 1.14.0<sup>12</sup>. The mean  
31 coverage cut-off (MCC), defined as the number of reads supporting each base above which bases were  
32 considered to be transcribed, and max region gap (MRG), defined as the maximum number of bases between  
33 expressed regions (ERs) below which adjacent ERs will be merged, were optimised. Optimisation was  
34 performed using 156,674 non-overlapping exons (defined by Ensembl v92) as the gold standard<sup>10</sup>. Exon  
35 biotypes of all Ensembl v92 exons were compared to this set of non-overlapping exons to ensure we were not  
36 preferentially optimising for one particular biotype (**Supplementary figure 6**). Non-overlapping exons were  
37 selected as these definitions would be least likely to be influenced by ambiguous reads. For each tissue, we  
38 generated ERs using mean coverage cut-offs increasing from 1 to 10 in steps of 0.2 (46 cut-offs) and max gaps  
39 increasing from 0 to 100 in steps of 10 (11 max region gaps) to produce a total of 506 unique transcriptomes.  
40 For each set of ERs, we found all ERs that intersected with non-overlapping exons, then calculated the exon  
41 delta by summing the absolute difference between the start/stop positions of each ER and the overlapping exon  
42 (Figure 1a). Situations in which a single ER overlapped with multiple exons were removed to avoid assigning  
43 the ER to an incorrect exon when calculating downstream optimisation metrics. For each tissue, we selected the  
44 mean coverage cut-off and max region gap, which minimised the difference between ER and "gold standard"  
45 exon definitions (median exon delta) and maximised the number of ERs that precisely matched the boundaries

16 of exons (number of ERs with an exon delta equal to 0). All ERs that were <3bp in width were removed as these  
17 were below the minimum size of a microexon<sup>25</sup>.  
18

#### 19 *Calculating the transcriptome size per annotation feature*

20 ERs were classified with respect to the annotation feature (exon, intron, intergenic) with which they  
21 overlapped. A minimum of 1bp overlap was required for an ER to be categorised as belonging to a given  
22 annotation feature. ERs overlapping multiple annotation features were labelled with a combination of each.  
23 This generated 6 distinct categories – “exon”, “exon, intron”, “exon, intergenic”, “exon, intergenic, intron”,  
24 “intergenic” and “intron” (**Supplementary figure 1a**). ERs classified as “exon, intergenic, intron” were  
25 removed from all downstream analysis as these formed only 0.54% of all ERs and were presumed to be  
26 technical artefacts generated from regions of dense, overlapping gene expression. For each tissue, the length of  
27 all ERs within each annotation feature was summed generating the total Mb of ERs per annotation feature.  
28 Normalised variance of exonic, intronic and intergenic ERs was calculated by dividing the standard deviation of  
29 the total Mb of ERs across tissues by the mean total Mb of ERs for each annotation feature. To compare between  
30 brain and non-brain tissues, the total Mb of intronic and intergenic ERs were first summed together to generate  
31 an overall measure of novel transcription abundance across brain and non-brain tissues, then a two-sided  
32 Wilcoxon rank sum test was applied.  
33

#### 34 *Annotating ERs with split read data*

35 Intronic and intergenic ERs were connected to known genes using reads, which we term split reads, with a  
36 gapped alignment to the genome, presumed to be reads spanning exon-exon junctions (Supplementary figure  
37 2b). Such exon-exon junctions are defined as non-contiguous reads which fall on the boundary between two  
38 exons of the same mRNA molecule, therefore when aligned to the genome these reads have a break in the  
39 middle indicating the splicing out of an intron. Split read data was categorised into three groups: annotated  
40 split reads, with both ends falling within known exons; partially annotated split reads, with only one end falling  
41 within a known exon; and unannotated split reads, with both ends within intron or intergenic regions. In this  
42 way, intron and intergenic ERs that overlapped with partially annotated split reads were connected to known  
43 genes.  
44

#### 75 *Validation of detected transcription*

76 Transcription was validated across different versions of Ensembl and within an independent dataset. ERs that  
77 overlapped purely intronic or intergenic regions according to Ensembl v87, but fell within exons according to  
78 v92, were counted as novel transcription that was validated in later versions of Ensembl. Furthermore, ERs  
79 overlapping exonic regions in Ensembl v87 now classified as intronic or intergenic in v92 were measured to  
80 control for expected corrections in gene definitions. To assess whether the total Kb of validated novel ERs  
81 entering v92 annotation was greater than what would be expected by chance, we generated 10,000 random  
82 sets of length-matched regions for each tissue that were intronic or intergenic with respect to Ensembl. Using a  
83 one sample Wilcoxon test, we compared the total Kb of intronic and intergenic ERs entering annotation to the  
84 total Kb distribution of the randomised intronic and intergenic regions, respectively.

35 Validation within an independent dataset was performed using RNA-seq coverage data from 49 control  
36 frontal cortex (BA9) samples originally reported by Labadof and colleagues (2015) and available via the  
37 recount R package version 1.4.6<sup>4,15</sup>. ERs derived from the GTEx frontal cortex (BA9) data were re-quantified  
38 using this independent frontal cortex dataset and those that had a mean coverage of at least 1.4 (the optimised  
39 MCC for the GTEx frontal cortex data), were counted as novel transcription that was validated.  
40

### 41 *Analysing the conservation and constraint of novel ERs*

42 Conservation scores in the form of phastCons7 (derived from genome-wide alignments of 7 mammalian  
43 species) were downloaded from UCSC<sup>26,27</sup>. Constraint scores generated from the genome-wide alignment of  
44 7,794 unrelated human genomes were downloaded as context dependent tolerance scores (CDTS)<sup>18</sup>. The raw  
45 phastCons7 and CDTs were in bins of 1bp and 10bp, respectively, therefore when annotating the  
46 corresponding positions of ERs, we aggregated each score as a mean across the entire genomic region of  
47 interest. To account for missing CDTs values, we calculated the coverage of each ER by dividing the number of  
48 bases annotated by the CDTs by the total length of the ER. For all downstream analysis, we filtered out ERs for  
49 which CDTs coverage was less than 80%.

50 To assess whether our novel ERs were more constrained or conserved than by expected by chance, we  
51 compared the phastCons7 and CDTs of novel ERs to 10,000 randomised length-matched sets of intronic and  
52 intergenic ERs for each tissue. For each of the 10,000 iterations, we first selected a random intronic or  
53 intergenic region that was larger than the respective ER, then selected a random segment along the randomised  
54 region which matched the length of the corresponding ER. The randomised regions were annotated with  
55 constraint scores and CDTs using the aforementioned method. The mean CDTs and phastCons7 of the novel  
56 ERs (split by annotation feature) were compared to the corresponding distribution of CDTs and phastCons7 of  
57 the randomised regions using a one sample, two-tailed t-test. For easier interpretation when plotting, CDTs  
58 scores have been converted to their opposite sign, therefore for both phastCons and CDTs, the higher the value  
59 the greater the magnitude of conservation or constraint as shown in Figure 4a.  
60

### 61 *Checking ER protein coding potential*

62 Intronic and intergenic ERs that were intersected by 2 split reads were extracted. The split reads were used to  
63 determine the precise boundaries of the ER. The R package Biostrings version 2.46.0 was used to extract the  
64 DNA sequence corresponding to the ER genetic co-ordinates from the genome build hg38<sup>28</sup>. Since the  
65 translation frame was ambiguous without knowledge of the other exons that are part of the transcript that  
66 included the novel ER, we converted the DNA sequence to amino acid sequence for all three possible frames  
67 starting from the first, second or third base. Any ER that had at least 1 frame that did not include a stop codon  
68 was considered to be potentially protein coding.  
69

### 70 *Gene properties influencing re-annotation*

71 All Ensembl v92 genes were marked with a 1 or a 0 depending on whether we detected a re-annotation for that  
72 gene in the form of an ER connected to the gene using a split read, with 1 representing a detected re-annotation  
73 event. Details of gene length, biotype, transcript count and whether the gene overlapped another gene were

?4 retrieved from the Ensembl v92 database. Brain-specificity was assigned using the Finucane dataset and  
?5 selecting the top 10% of brain-specific genes when compared to non-brain tissues<sup>29</sup>. Mean gene TPM was  
?6 calculated by downloading tissue-specific TPM values from the GTEx portal and summarised by calculating the  
?7 mean across all tissues. The list of OMIM genes (May 2018) was used to assign whether a gene was known to  
?8 cause disease or not. We used a logistic regression to test whether different gene properties significantly  
?9 influenced the variability of re-annotation (formula = re-annotated ~ brain specific + mean TPM + overlapping  
30 gene + transcript count + gene biotype + gene length).

31

### 32 *Sanger sequencing of novel junctions*

33 Commercially purchased (Takara) frontal cortex and cerebellum RNA samples, isolated from individuals of  
34 European descent, were used for validation of novel junctions detected in *SNCA* and *ERLIN1* respectively.  
35 Tissues were chosen to match the tissue in which the re-annotation for each gene was detected. Reverse  
36 transcription was performed using 1ug of RNA from each tissue, then converted to cDNA using the High-  
37 Capacity cDNA Reverse Transcription Kit with RNase Inhibitor (Applied Biosystems) and random primers as  
38 per manufacturer's instructions. Primers were designed to span predicted exon-exon junctions using Primer-  
39 BLAST (NCBI) and ordered from Sigma (**Supplementary Table 4**). PCR was performed using FastStart PCR  
40 Master (Roche) and enzymatic clean-up of PCR products was performed using Exonuclease I (Thermo  
41 Scientific) and FastAP Thermosensitive Alkaline Phosphatase (Thermo Scientific). Sanger sequencing was  
42 performed using the BigDye terminator kit (Applied Biosystems) and sequences were viewed and exported  
43 using CodonCode Aligner (V. 8.0.2). Sequences were blatted against the human genome (hg38) and alignment  
44 visually inspected for confirmation of validation.

45

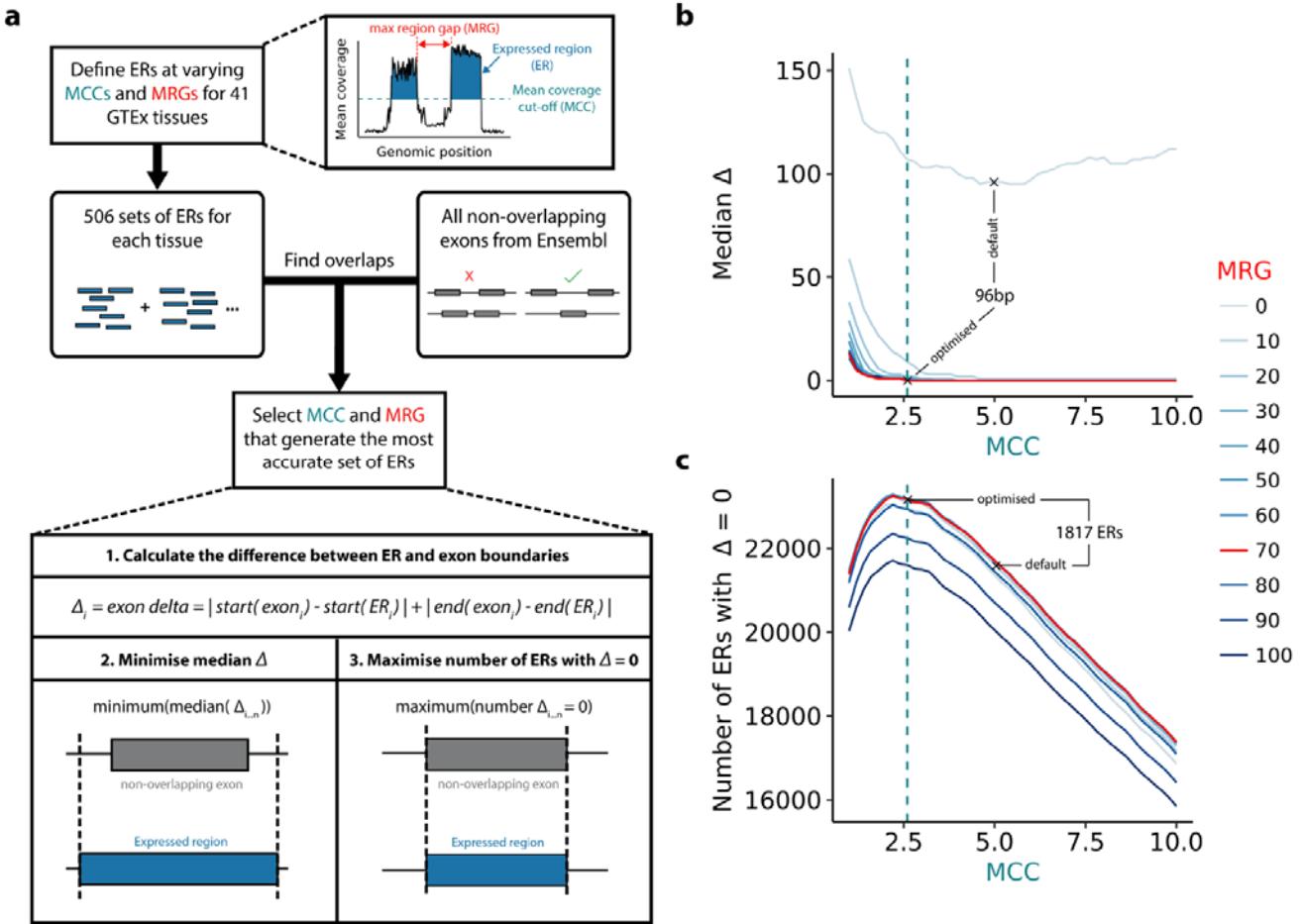
### 46 *Expression-weighted cell-type enrichment (EWCE): evaluating enrichment of theta-correlated* 47 *genes*

48 EWCE was used to determine whether brain-specific genes (both re-annotated and not re-annotated) have  
49 higher expression within particular cell types than expected by chance<sup>30</sup>. As our input, we used 1) neuronal and  
50 glial clusters of the central nervous system (CNS) identified in the Linnarsson single-cell RNA sequencing  
51 dataset (amounting to a subset of 114 of the original 265 clusters identified) and 2) lists of genes split by  
52 whether or not they were re-annotated, and if re-annotated, by their overlap with Ensembl v92 annotation  
53 features (see **Supplementary Table 5** for full list of CNS neuronal clusters and genes used)<sup>31</sup>. For each gene in  
54 the Linnarsson dataset, we estimated its cell-type specificity (the proportion of a gene's total expression in one  
55 cell type compared to all cell types) using the 'generate.celltype.data' function of the EWCE package. EWCE with  
56 the target list was run with 100,000 bootstrap replicates, which were sampled from a background list of genes  
57 that excluded all genes without a 1:1 mouse:human ortholog. We additionally controlled for transcript length  
58 and GC-content biases by selecting bootstrap lists with comparable properties to the target list. We performed  
59 the analysis with major cell-type classes (e.g. "astrocyte", "microglia", etc.). Data are displayed as  
60 standard deviations from the mean, and any values < 0, which reflect a depletion of expression, are displayed as  
61 0. P-values were corrected for multiple testing using the Benjamini-Hochberg (FDR) method over all cell types  
62 and gene lists displayed.

53 *Enrichment of re-annotated genes for neurological disorder associated genes*

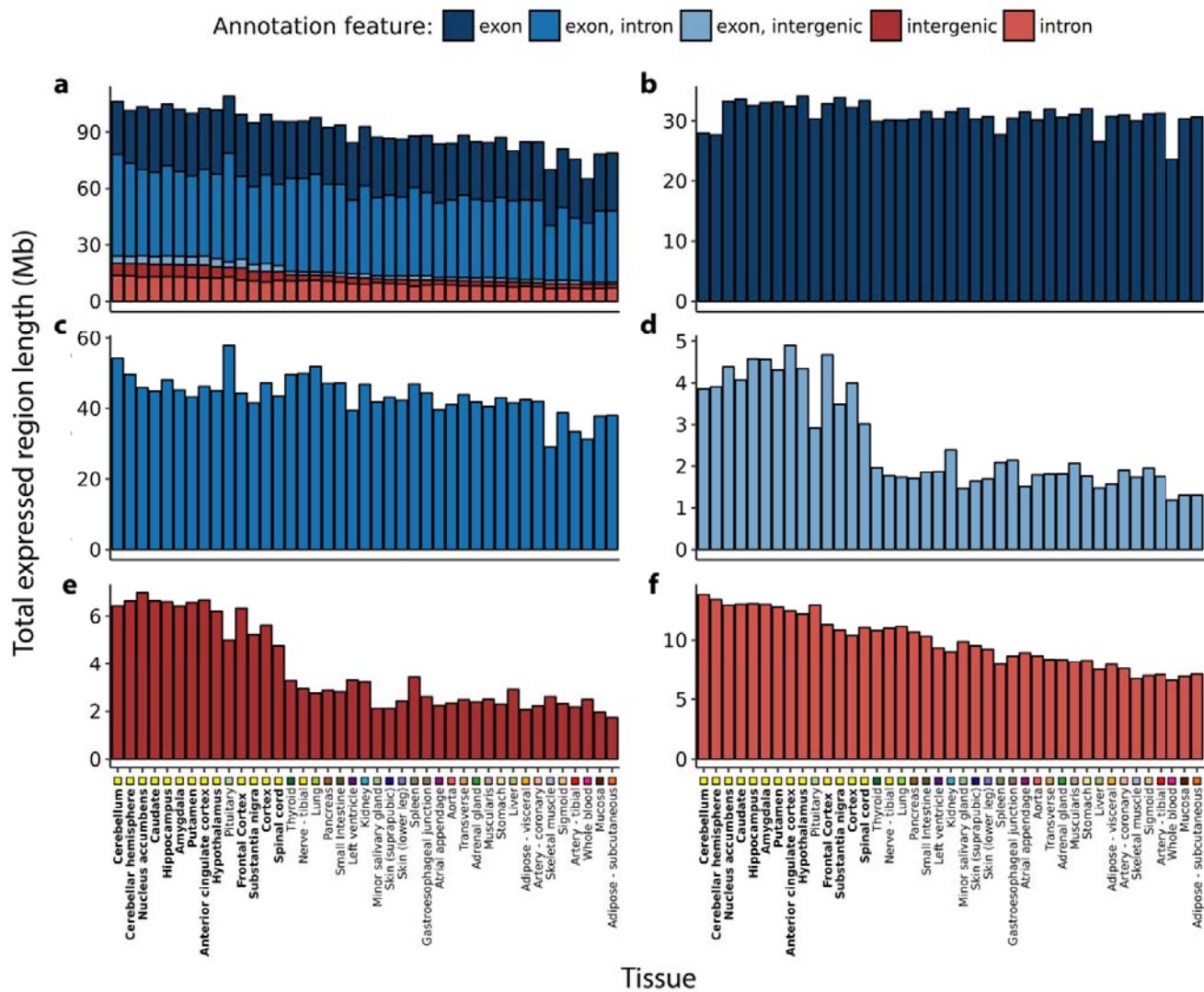
54 The STOPGAP database detailing all genes associated with 4684 GWASs was downloaded from  
55 [https://github.com/StatGenPRD/STOPGAP/blob/master/STOPGAP\\_data/stopgap.bestId.RData](https://github.com/StatGenPRD/STOPGAP/blob/master/STOPGAP_data/stopgap.bestId.RData). To select  
56 which genes were associated to a GWAS, the “best gene” as determined by STOPGAP using functional evidence  
57 was used<sup>22</sup>. The medical subject heading for each disease was used to further subgroup GWASs into 4  
58 categories; neurodegenerative, neuropsychiatric, other neurological conditions and the remaining as other  
59 (**Supplementary table 6**). For each of the subgroups, we generated a contingency table, counting the number  
70 of genes that were re-annotated or not in relation to whether they fell into that particular subgroup. For genes  
71 that were overlapping between GWASs, we classified a gene to be part of a subgroup if it was associated with at  
72 least 1 GWAS contained in that subgroup. A Fisher’s Exact test was used to examine whether our re-annotated  
73 gene list was significantly enriched for genes from any of the subgroups. Benjamini-Hochberg (FDR) method  
74 was used for to correct for multiple testing.

75 **Figures**



76

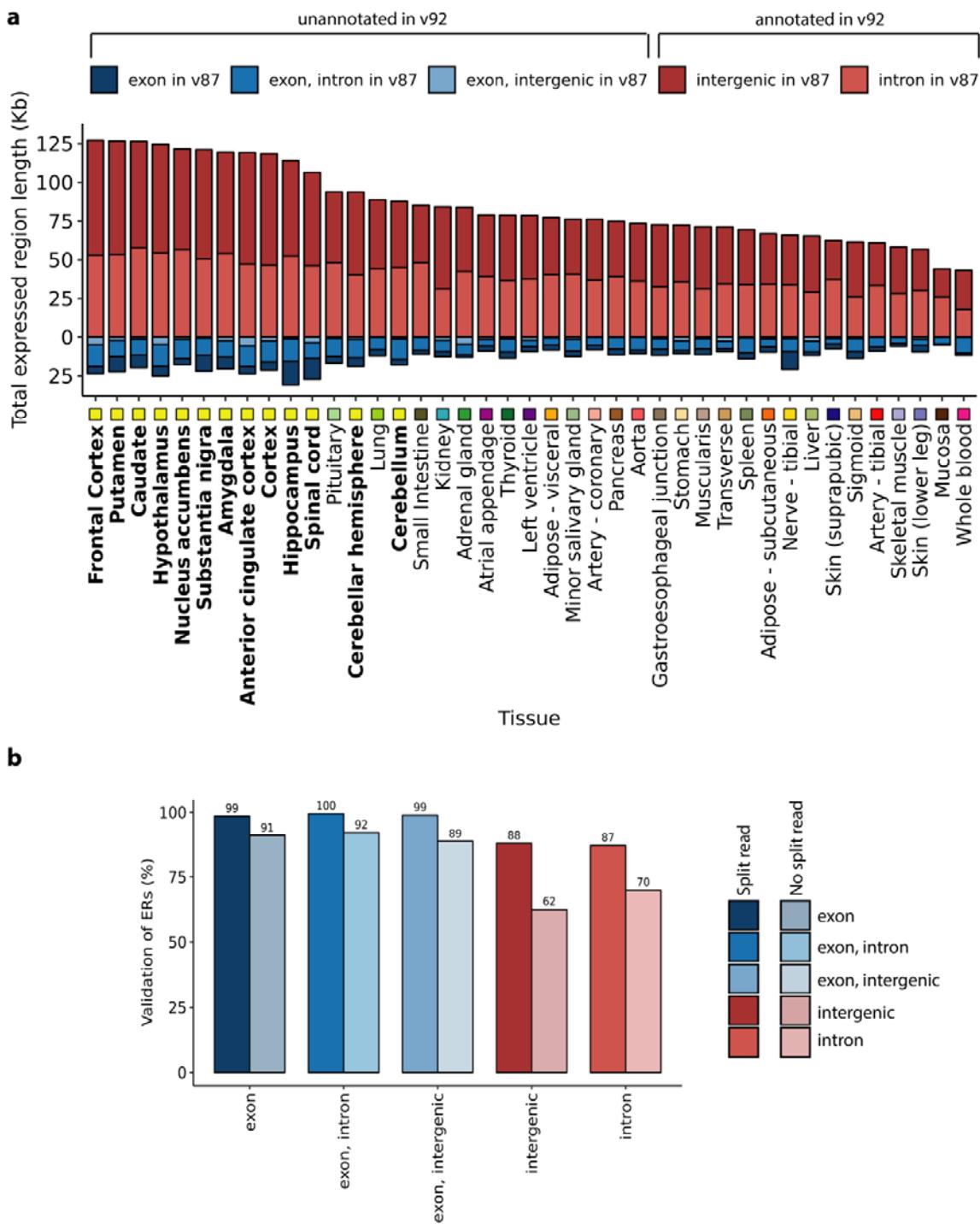
77 **Figure 1 – Optimisation of the detection of transcription.** a) Transcription in the form expressed regions (ERs) was detected in an  
 78 annotation agnostic manner across 41 human tissues. The mean coverage cut-off (MCC) is the number of reads supporting each base above  
 79 which that base would be considered transcribed and the max region gap (MRG) is the maximum number of bases between ERs below which  
 80 adjacent ERs would be merged. MCC and MRG parameters were optimised for each tissue using the non-overlapping exons from Ensembl  
 81 v92 reference annotation. b) Line plot illustrating the selection of the MCC and MRG that minimised the difference between ER and exon  
 82 definitions (median exon delta). c) Line plot illustrating the selection of the MCC and MRG that maximised the number of ERs that precisely  
 83 matched exon definitions (exon delta = 0). The cerebellum tissue is plotted for (b) and (c), which is representative of the other GTEx tissues.  
 84 Green and red lines indicate the optimal MCC (2.6) and MRG (70), respectively.



35

36 *Figure 2 – Transcription detected across 41 GTEx tissues categorised by annotation feature.* Within each tissue the length of the ERs  
 37 Mb overlapping **a)** all annotation features **b)** purely exons **c)** exons and introns **d)** exons and intergenic regions **e)** purely intergenic regions  
 38 **f)** purely introns according to Ensembl v92 was computed. Tissues are plotted in descending order based on the respective total size of  
 39 intronic and intergenic regions. Tissues are colour-coded as indicated in the x-axis, with GTEx brain regions highlighted with bold font. At  
 40 least 8.4Mb of novel transcription was discovered in each tissue, with the greatest quantity found within brain tissues (mean across brain  
 41 tissues: 18.6Mb, non-brain: 11.2Mb, two-sided Wilcoxon rank sum test p-value: 2.35e-10

12



13

14 *Figure 3 – Validation of novel transcription. a)* The classification of ERs based on v87 and v92 of Ensembl was compared. Across all  
 15 tissues, the number of intron or intergenic ERs with respect to v87 that were known to be exonic in Ensembl v92 was greater than the  
 16 number of ERs overlapping exons according to v87 that were now unannotated in v92. Tissues are plotted in descending order based on the  
 17 total Mb of novel ERs with respect to Ensembl v87 that were validated (classified as exonic in the Ensembl v92). Tissues are colour-coded as  
 18 indicated in the x-axis, with GTEx brain regions highlighted with bold font. **b)** Barplot represents the percentage of ERs seeding from the  
 19 GTEx frontal cortex that validated in an independent frontal cortex RNA-seq dataset. ERs defined in the seed tissue were re-quantified using  
 20 coverage from the validation dataset, after which the optimised mean coverage cut off was applied to determine validated ERs. Colours  
 21 represent the different annotation features that the ERs overlapped and the shade indicates whether the ER was supported by split read(s).

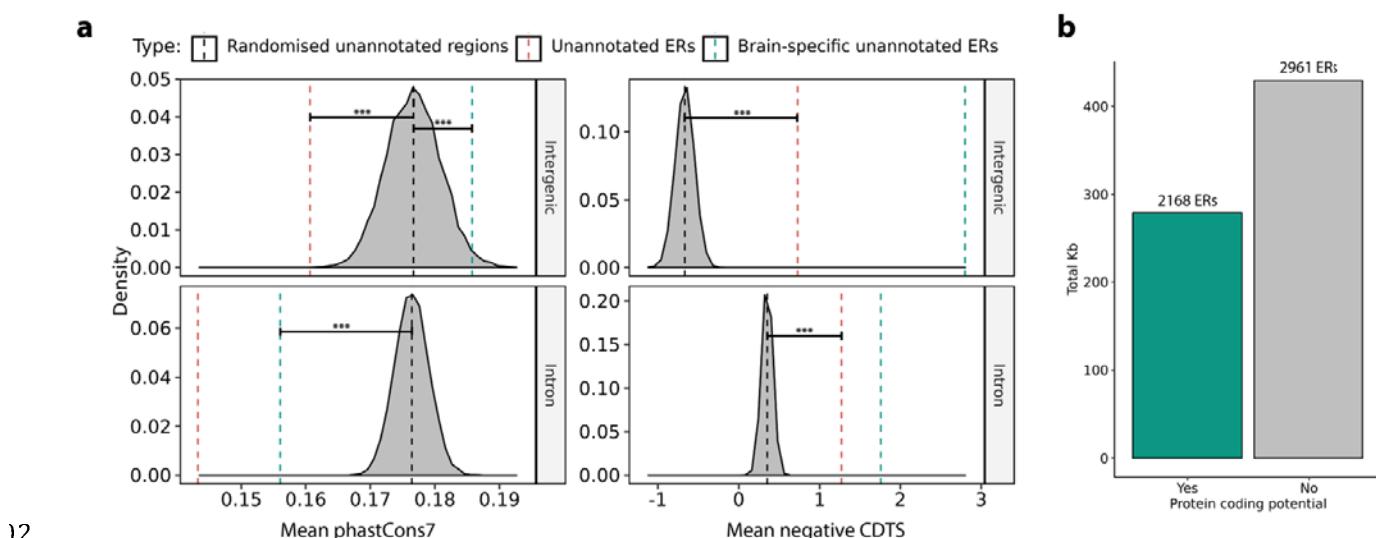
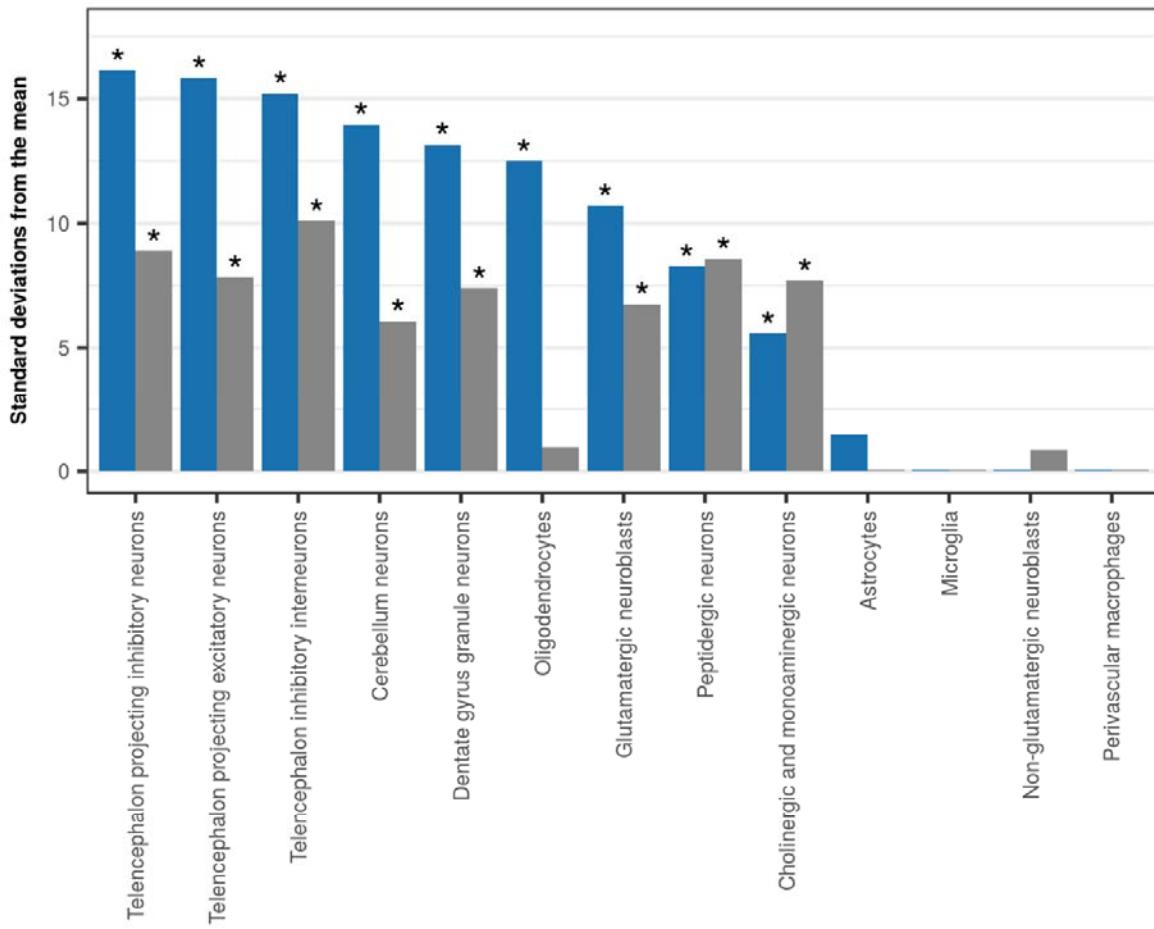
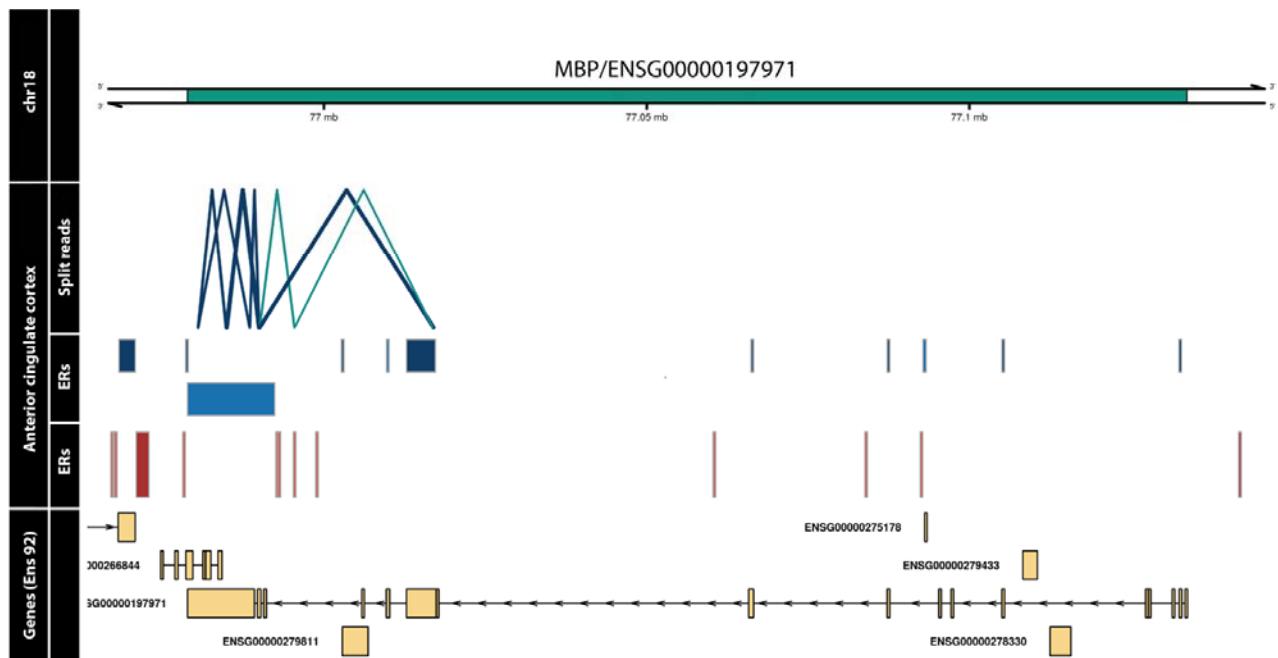


Figure 4 – **Novel ERs collectively serve an important function for humans and a proportion can form potentially protein coding transcripts.** **a)** Comparison of conservation (phastCons7) and constraint (CDTS) of intronic and intergenic ERs to 10,000 sets of random, length-matched intronic and intergenic regions. Novel ERs marked by the red, dashed line are less conserved than expected by chance, but are more constrained. Brain-specific ERs marked by the green, dashed lines are amongst the most constrained. Data for the cerebellum shown and is representative of other GTEx tissues. **b)** The DNA sequence for ERs overlapping 2 split reads was obtained and converted to amino acid sequence for all 3 possible frames. 2,168 ERs (57%) lacked a stop codon in at least 1 frame and were considered potentially protein-coding.

a

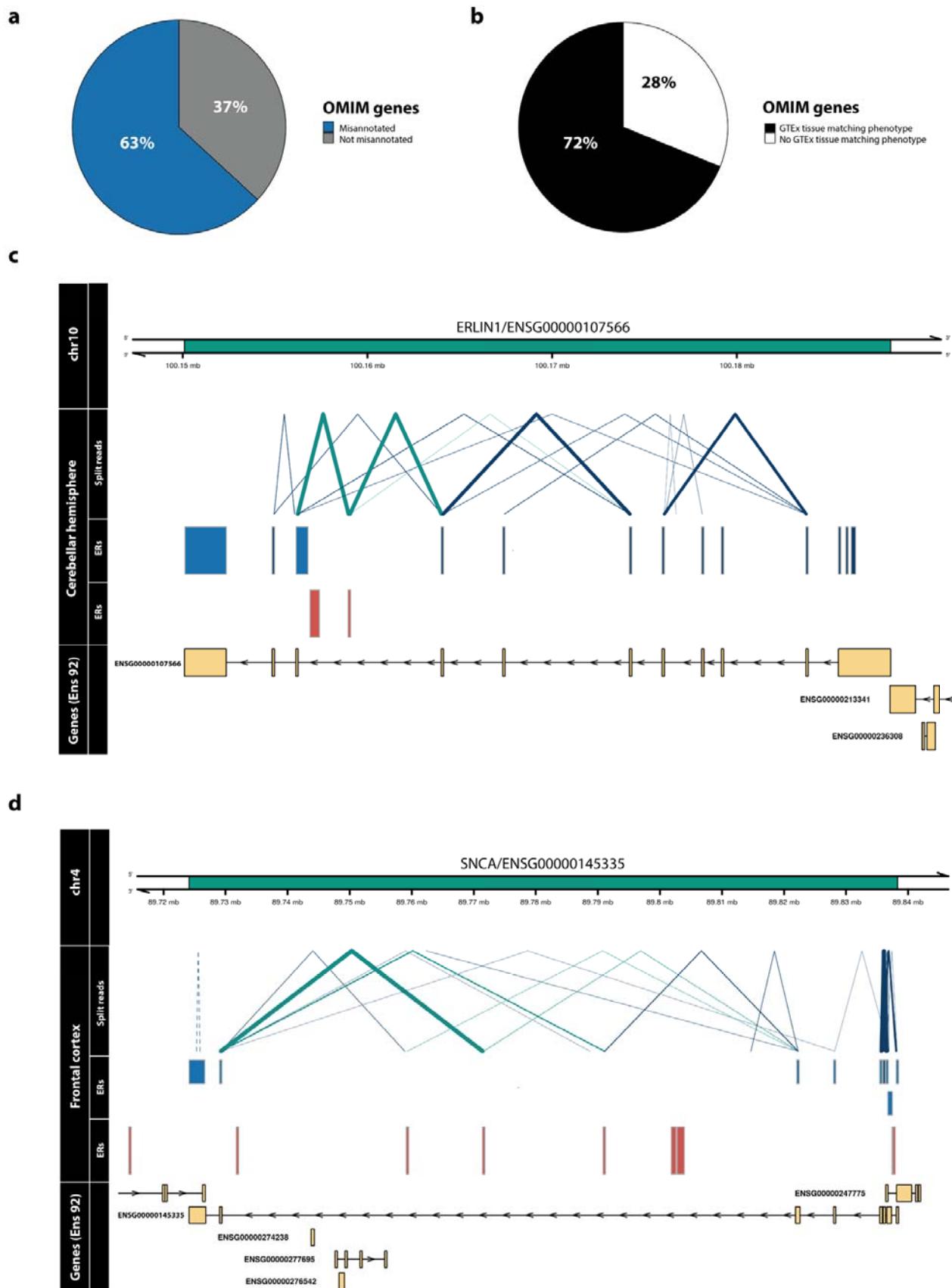


b



10

11 *Figure 5 – Incomplete annotation of genes disproportionately affects oligodendrocytes* a) Bar plot displaying the enrichment of re-  
 12 annotated and not re-annotated genes within brain cell-type specific gene sets. Blue bars represent the re-annotated genes and grey are  
 13 those without re-annotations. Of all analysed cell-types, the greatest difference between enrichment of re-annotated and not re-annotated  
 14 was observed in oligodendrocytes. b) Novel potentially protein coding ER discovered in MBP, with an oligodendrocyte specific expression



16

17 *Figure 6 – Re-annotation of OMIM genes.* a) A novel ER connected through a split read was discovered for 63% of OMIM-morbid genes. b) 18 Comparison of the phenotype (HPO terms) associated with each re-annotated OMIM-morbid gene and the GTEx tissue from which novel ERs 19 were derived. Through manual inspection, HPO terms were matched to disease-relevant GTEx tissues and for 72% of re-annotated OMIM

12 the split reads and ERs overlapping the genomic region derived from the labelled tissue. Blue ERs overlap known exonic regions and red ERs  
13 fall within intronic or intergenic regions. Blue split reads overlap blue ERs, while green split reads overlap both red and blue ERs, connecting  
14 novel ERs to OMIM-morbid genes. Thickness of split reads represents the proportion of samples of that tissue in which the split read was  
15 detected. Only partially annotated split reads (solid lines) and unannotated split reads (dashed lines) are plotted. The last track displays the  
16 genes within the region according to Ensembl v92, with all known exons of the gene collapsed into one "meta" transcript.

17

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

\*\*\* p <= 2e-16

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.

17  
18  
19  
20  
21

## 52 Acknowledgements

53 S.G. was supported through the award of an Alzheimer's Research UK PhD fellowship. R.H.R. was supported  
54 through the award of a Leonard Wolfson Doctoral Training Fellowship in Neurodegeneration. J.H. and M.R.  
55 were supported by the UK Medical Research Council (MRC), with J.H. supported by a grant (MR/N026004/)  
56 and M.R. through the award of a Tenure-track Clinician Scientist Fellowship (MR/N008324/1). J.H. was also  
57 supported by the UK Dementia Research Institute.

## 58 Author contributions

59 D.Z., S.G. and M.R. conceived and designed the study. D.Z. analysed the data, generated figures and together with  
60 M.R. wrote the first draft of the manuscript. R.H.R. performed analysis and generated figures for the cell-type  
61 specific section. S.G.R. and J.B. developed and deployed the vizER online platform. Sanger sequence validation  
62 was performed by B.C. and W.L. T.C. helped manually associate OMIM phenotypes to GTEx tissues. S.G, L.C.T.,  
63 J.B., K.D., A.P. and M.R. helped guide and troubleshoot analyses. L.C.T and A.E.F helped with the use of the  
64 recount2 data. D.Z., S.G., R.H.R., J.H., L.C.T. and M.R. contributed to the critical analysis of the manuscript.

## 65 Competing Interests

66 No competing interests to declare.

## 57 Bibliography

- 58 1. Chen, G. *et al.* Incorporating the human gene annotations in different databases significantly improved  
59 transcriptomic and genetic analyses. *RNA* **19**, 479–89 (2013).
- 60 2. McCarthy, D. J. *et al.* Choice of transcripts and software has a large effect on variant annotation. *Genome*  
61 *Med.* **6**, (2014).
- 62 3. Yeo, G., Holste, D., Kreiman, G. & Burge, C. B. Variation in alternative splicing across human tissues.  
63 *Genome Biol.* **5**, R74 (2004).
- 64 4. Collado-Torres, L. *et al.* Reproducible RNA-seq analysis using recount2. *Nat. Biotechnol.* **35**, 319–321  
65 (2017).
- 66 5. Zhang, Y. E., Landback, P., Vibranovski, M. & Long, M. New genes expressed in human brains:  
67 Implications for annotating evolving genomes. *BioEssays* **34**, 982–991 (2012).
- 68 6. Sibley, C. R. *et al.* Recursive splicing in long vertebrate genes. *Nature* (2015). doi:10.1038/nature14466
- 69 7. Sibley, C. R., Blazquez, L. & Ule, J. Lessons from non-canonical splicing. *Nature Reviews Genetics* (2016).  
70 doi:10.1038/nrg.2016.46
- 71 8. Encode, T. & Consortium, P. Identification and analysis of functional elements in 1 % of the human  
72 genome by the ENCODE pilot project. **447**, (2007).
- 73 9. Zhao, S., Zhang, Y., Gamini, R., Zhang, B. & Schack, D. Von. Evaluation of two main RNA-seq approaches for  
74 gene quantification in clinical RNA sequencing: polyA + selection versus rRNA depletion. *Sci. Rep.* 1–12  
75 (2018). doi:10.1038/s41598-018-23226-4
- 76 10. Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res.* **46**, D754–D761 (2018).
- 77 11. The GTExArd Consortium *et al.* The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene  
78 regulation in humans. *Science (80-. ).* **348**, 648–60 (2015).
- 79 12. Collado-Torres, L. *et al.* Flexible expressed region analysis for RNA-seq with derfinder. *Nucleic Acids Res.*  
80 **45**, e9 (2017).
- 81 13. Jaffe, A. E. *et al.* Developmental regulation of human cortex transcription and its clinical relevance at  
82 single base resolution. *Nat. Neurosci.* **18**, 154–161 (2015).
- 83 14. Pertea, M. *et al.* Thousands of large-scale RNA sequencing experiments yield a comprehensive new  
84 human gene list and reveal extensive transcriptional noise. *Genome Biol.* 332825 (2018).  
85 doi:10.1101/332825
- 86 15. Labadof, A. *et al.* RNA sequence analysis of human huntington disease brain reveals an extensive  
87 increase in inflammatory and developmental gene expression. *PLoS One* **10**, 1–21 (2015).
- 88 16. Doolittle, W. F. We simply cannot go on being so vague about ‘function’. *Genome Biol.* 18–20 (2018).
- 89 17. Harrow, J. *et al.* GENCODE: The Reference Human Genome Annotation for The ENCODE Project. *Genome*  
90 *Res.* **22**, 1760–1774 (2012).
- 91 18. Di Iulio, J. *et al.* The human noncoding genome defined by genetic diversity. *Nat. Genet.* **50**, 333–337  
92 (2018).
- 93 19. Ettle, B., Schlachetzki, J. C. M. & Winkler, J. Oligodendroglia and Myelin in Neurodegenerative Diseases:  
94 More Than Just Bystanders? *Mol. Neurobiol.* 3046–3062 (2016). doi:10.1007/s12035-015-9205-3

5 20. Calderon, D. *et al.* Inferring Relevant Cell Types for Complex Traits by Using Single-Cell Gene Expression.  
6 *Am. J. Hum. Genet.* **101**, 686–699 (2017).

7 21. Julien Bryois, Nathan G. Skene, Thomas Folkmann Hansen, Lisette J.A. Kogelman, Hunna J. Watson, Eating  
8 Disorders Working Group of the Psychiatric Genomics Consortium, International Headache Genetics  
9 Consortium, The 23andMe Research Team, Leo Brueggeman, G. P. F. S. *et al.* Genetic Identification of Cell  
10 Types Underlying Brain Complex Traits Yields Novel Insights Into the Etiology of Parkinson's Disease.  
11 *bioRxiv* 528463 (2019). doi:10.1101/528463

12 22. Shen, J., Song, K., Slater, A. J., Ferrero, E. & Nelson, M. R. STOPGAP: a database for systematic target  
13 opportunity assessment by genetic association predictions. *Bioinformatics* **33**, 2784–2786 (2017).

14 23. Novarino, G. *et al.* Exome Sequencing Links Corticospinal Motor Neuron Disease to Common  
15 Neurodegenerative Disorders. *Science (80-. ).* **343**, 506–511 (2014).

16 24. Hamosh, A., Scott, A. F., Amberger, J., Valle, D. & McKusick, V. A. Online Mendelian Inheritance in Man  
17 (OMIM). *Hum. Mutat.* **15**, 57–61 (2000).

18 25. Irimia, M. *et al.* A highly conserved program of neuronal microexons is misregulated in autistic brains.  
19 *Cell* **159**, 1511–1523 (2014).

20 26. Siepel, A. & Haussler, D. Phylogenetic Hidden Markov Models. **26** (2005).

21 27. Wainberg, M., Alipanahi, B. & Frey, B. Does conservation account for splicing patterns?ccc. *BMC Genomics*  
22 **17**, 1–10 (2016).

23 28. Pagès, H., Aboyoun, P., Gentleman, R. & DebRoy, S. Biostrings: Efficient manipulation of biological strings.  
24 *R package version 2.46.0* (2017). doi:10.1021/jm900485a

25 29. Finucane, H. K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant  
26 tissues and cell types. *Nat. Genet.* **50**, 621–629 (2018).

27 30. Skene, N. G. & Grant, S. G. N. Identification of vulnerable cell types in major brain disorders using single  
28 cell transcriptomes and expression weighted cell type enrichment. *Front. Neurosci.* (2016).  
29 doi:10.3389/fnins.2016.00016

30 31. Zeisel, A. *et al.* Molecular Architecture of the Mouse Nervous System. *Cell* (2018).  
31 doi:10.1016/j.cell.2018.06.021

32

33