Incomplete annotation of OMIM genes is likely to be limiting the diagnostic yield of genetic testing, particularly for neurogenetic disorders

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

1. Institute of Neurology, University College London (UCL), London, UK
2. Sorbonne Universités, UPMC Université Paris 06, UMR S 1127, Inserm U 1127, CNRS UMR 7225, ICM, Paris, France
3. Lieber Institute for Brain Development, Baltimore, Maryland, USA
4. Center for Computational Biology, Johns Hopkins University, Baltimore, Maryland, USA
5. Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA
6. Department of Psychiatry and Behavioral Sciences, Johns Hopkins School of Medicine, Baltimore, MD, USA
7. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA
8. Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
9. Departamento de Ingeniería de la Información y las Comunicaciones, Universidad de Murcia, 30100, Murcia, Spain

*These authors contributed equally.

Abstract

Although the increasing use of whole-exome and whole-genome sequencing have improved the yield of genetic testing for Mendelian disorders, an estimated 50% of patients still leave the clinic without a genetic diagnosis. This can be attributed in part to our lack of ability to accurately interpret the genetic variation detected through next-generation sequencing. Variant interpretation is fundamentally reliant on accurate and complete gene annotation, however numerous reports and discrepancies between gene annotation databases reveals that the knowledge of gene annotation remains far from comprehensive. Here, we detect and validate transcription in an annotation-agnostic manner across all 41 different GTEx tissues, then connect novel transcription to known genes, ultimately improving the annotation of 63% of the known OMIM-morbid genes. We find the majority of novel transcription to be tissue-specific in origin, with brain tissues being most susceptible to misannotation. Furthermore, we find that novel transcribed regions tend to be poorly conserved, but are significantly depleted for genetic variation within humans, suggesting they are functionally significant and potentially have human-specific functions. We present our findings through an online platform vizER, which enables individual genes to be visualised and queried for evidence of misannotation. We also release all tissue-specific transcriptomes in a BED format for ease of integration with whole-genome sequencing data. We anticipate that these resources will improve the diagnostic yield for a wide range of Mendelian disorders.

Introduction

Genetic and transcriptomic studies are fundamentally reliant on accurate and complete human gene annotation. Gene definitions are required for the quantification of expression and splicing from RNA-sequencing experiments, interpretation of significant GWAS signals and variant interpretation from genetic tests. The latter is a crucial step in molecular diagnosis, which involves ascertaining the genetic cause of disease for a patient with a suspected Mendelian disorder[1]. Importantly, successful molecular diagnosis can improve the management of symptoms, inform genetic counselling and provide therapeutic opportunities for diagnosis and prevention. With the advancement in next-generation sequencing technology and concomitant reduction in associated costs, genetic diagnosis has progressed from traditional targeted sequencing of mutation hotspots to in-silico panels using whole exome sequencing (WES) and, more recently, whole genome sequencing (WGS)[2–4]. WES and WGS have improved the ability to identify variants associated with disease and increasingly, are an integral part of the diagnostic journey. However, despite these advances, our understanding of the genetic aetiology of Mendelian disorders remains incomplete and consequently, the current rate of genetic diagnosis remains only 25-50%[5,6].

A key component of the molecular diagnosis of Mendelian disorders is the ability to distinguish pathogenic variants from the many rare, functional yet non-pathogenic variants present in any human genome. Given that the vast majority of currently known pathogenic variants fall within exonic regions, variants located within intronic regions or intergenic regions are unsurprisingly downgraded in importance[7]. However, as our understanding of transcriptomic complexity improves it is apparent that existing annotation remains incomplete even amongst known genes. Comparison of different gene annotation databases reveals that over 17,000 Ensembl genes fall into intronic or intergenic regions according to the AceView database and predictably, the choice of reference annotation greatly influences the output of variant interpretation software such as VEP and ANNOVAR[8,9]. Thus, incomplete annotation may cause pathogenic variants to be overlooked within coding regions that are yet to be annotated, despite them having been sequenced. When taken together these findings suggest that improvements to gene annotation will increase the diagnostic yield from genetic tests[10].

There is evidence to suggest that improvements to gene annotation may be most important for the diagnosis of neurogenetic disorders. While the large phenotypic overlap and variability of neurogenetic disorders has meant that unbiased WES or WGS approaches to genetic testing have greater diagnostic utility, these disorders remain amongst those with the lowest diagnostic rate. A recent report estimates that as few as 26% of patients in this category are successfully diagnosed[11]. Given that the human brain is the tissue with the highest prevalence of alternative splicing, genes important for brain function may be predicted to have disproportionately high levels of misannotation, which we define as the number of genes for which annotation remains inaccurate or incomplete[12]. In fact, several studies analysing RNA-sequencing derived from human brain tissue have discovered transcription originating from intronic or intergenic regions (henceforth termed novel)[13,14]. In particular, Jaffe and colleagues found that as much as 41% of human frontal cortex transcription was novel. Therefore, taking into account both the genetic heterogeneity of neurogenetic disorders and the

human brain's susceptibility to misannotation, improvements to gene annotation may be of greatest benefit to the diagnosis of this important set of conditions.

To address this issue, we used publicly available transcriptomic data to improve the annotation of genes across the genome, with a focus on genes known to cause Mendelian disease as reported in the Online Mendelian Inheritance in Man (OMIM) catalogue. We define transcription in an annotation-agnostic manner from 41 GTEx tissues. We find that while novel transcription is widespread across all tissues, it is most prevalent in human brain and, collectively, is relatively depleted for genetic variation, suggesting it is functionally important. By combining novel expressed regions (ERs) with split read data, defined as reads that have a gapped alignment to the genome, we link these regions to known OMIM genes. To aid the ease of integration with WGS data, we have released all 41 tissue-specific transcriptomes in a BED format and built an online platform vizER, which allows individual genes to be queried and visualised. Overall, we improve the annotation of 1929 (63%) OMIM genes, a vital step for the accurate assignment of variant pathogenicity, and anticipate that this will lead to improvements in diagnostic yield from WGS, particularly for neurogenetic disorders.

Methods

OMIM data

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

GTEx data

RNA-seq data in base-level coverage format for 7,595 samples originating from 41 different GTEx tissues was downloaded using the R package recount version 1.4.6[13]. Cell lines, sex-specific tissues and tissues with 10 samples or below were removed. Samples with large chromosomal deletions and duplications or large CNVs previously associated with disease were filtered out (smafrze = "USE ME"). Coverage for all remaining samples was normalised to a target library size of 40 million 100bp reads using the area under coverage value provided by recount2. For each tissue, base-level coverage was averaged across all samples to calculate the mean base-level coverage. GTEx split read data, defined as reads with a non-contiguous gapped alignment to the genome, was downloaded using the recount2 resource and filtered to include only split reads detected in at least 5% of samples for a given tissue and those that had available donor and acceptor splice sequences.

Optimising the detection of transcription

Transcription was detected across 41 GTEx tissues using the package derfinder version 1.14.0[17]. The mean coverage cut-off (MCC), defined as the number of reads supporting each base above which bases were considered to be transcribed, and max region gap (MRG), defined as the maximum number of bases between expressed regions (ERs) below which adjacent ERs will be merged, were optimised. Optimisation was performed using 156,674 non-overlapping exons (defined by Ensembl v92) as the gold standard[18]. Exon biotypes of all Ensembl v92 exons were compared to this set of non-overlapping exons to ensure we were not preferentially optimising for one particular biotype (Supplementary figure 1). Non-overlapping exons were selected as these definitions would be least likely to be influenced by ambiguous reads. For each tissue, we generated ERs using mean coverage cut-offs increasing from 1 to 10 in steps of 0.2 (46 cut-offs) and max gaps increasing from 0 to 100 in steps of 10 (11 max region gaps) to produce a total of 506 unique transcriptomes. For each set of ERs, we found all ERs that intersected with non-overlapping exons, then calculated the exon delta by summing the absolute difference between the start/stop positions of each ER and the overlapping exon (Figure 1a). Situations in which a single ER overlapped with multiple exons were removed to avoid assigning the ER to an incorrect exon when calculating downstream optimisation metrics. For each tissue, we selected the mean coverage cut-off and max region gap, which minimised the difference between ER and "gold standard" exon definitions (median exon delta) and maximised the number of ERs that precisely matched the boundaries

of exons (number of ERs with an exon delta equal to 0). All ERs that were <3bp in width were removed as these were below the minimum size of a microexon[19].

Calculating the transcriptome size per annotation feature

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

Annotating ERs with split read data

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

Validation of detected transcription

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

Validation within an independent dataset was performed using RNA-seq coverage data from 49 control frontal cortex (BA9) samples originally reported by Labadorf and colleagues (2015) and available via the recount R package version 1.4.6[13,20]. ERs derived from the GTEx frontal cortex (BA9) data were re-quantified using this independent frontal cortex dataset and those that had a mean coverage of at least 1.4 (the optimised MCC for the GTEx frontal cortex data), were counted as novel transcription that was validated.

Analysing the conservation and constraint of novel ERs

Conservation scores in the form of phastCons7 (derived from genome-wide alignments of 7 mammalian species) were downloaded from UCSC[21,22]. Constraint scores generated from the genome-wide alignment of 7,794 unrelated human genomes were downloaded as context dependent tolerance scores (CDTS)[23]. The raw phastCons7 and CDTS were in bins of 1bp and 10bp, respectively, therefore when annotating the corresponding positions of ERs, we aggregated each score as a mean across the entire genomic region of interest. To account for missing CDTS values, we calculated the coverage of each ER by dividing the number of bases annotated by the CDTS by the total length of the ER. For all downstream analysis, we filtered out ERs for which CDTS coverage was less than 80%.

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

Checking ER protein coding potential

Intronic and intergenic ERs that were intersected by 2 split reads were extracted. The split reads were used to determine the precise boundaries of the ER. The R package Biostrings version 2.46.0 was used to extract the DNA sequence corresponding to the ER genetic co-ordinates from the genome build hg38[24]. Since the translation frame was ambiguous without knowledge of the other exons that are part of the transcript that included the novel ER, we converted the DNA sequence to amino acid sequence for all three possible frames starting from the first, second or third base. Any ER that had at least 1 frame that did not include a stop codon was considered to be potentially protein coding.

Gene properties influencing misannotation

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

retrieved from the Ensembl v92 database. Brain-specificity was assigned using the Finucane dataset and selecting the top 10% of brain-specific genes when compared to non-brain tissues[25]. Mean gene TPM was calculated by downloading tissue-specific TPM values from the GTEx portal and summarised by calculating the mean across all tissues. The list of OMIM genes (May 2018) was used to assign whether a gene was known to cause disease or not. We used a logistic regression to test whether different gene properties significantly influenced the variability of misannotation (formula = misannotation ~ brain specific + mean TPM + overlapping gene + transcript count + gene biotype + gene length).

Results

Accurately detecting transcription in an annotation-agnostic manner

We applied the R package derfinder to discover genome-wide transcription from 41 GTEx tissues[16]. By default, derfinder utilises RNA-sequencing base-level coverage data to detect continuous blocks of transcribed bases termed expressed regions (ERs) in an annotation-agnostic manner using the mean coverage cut-off (MCC)[17]. In order to define ERs more accurately, we improved upon the original derfinder methodology by including an additional parameter we call the max region gap (MRG), which merges adjacent ERs that have been segmented due to the variability in read depth even across an individual exon (see detailed Methods). Both the MCC and MRG were optimised using a set of exons with the most reliable and accurate boundaries, namely all exons from Ensembl v92 that did not overlap with any other exon[18]. For each tissue, we first generated 506 possible transcriptomes using unique pairs of MRCs and MRGs to produce a total of 20,746 sets of ERs across all 41 tissues. Then for each transcriptome, all ERs that intersected non-overlapping exons were extracted and the absolute difference between the ER definition and the corresponding exon boundaries, termed the exon delta, was calculated (Figure 1a). We summarised the exon delta for each transcriptome using two metrics, the median exon delta and the number of ERs with exon delta equal to 0. The median exon delta represents the overall accuracy of all ER definitions, whereas, the number of ERs with exon delta equal to 0 indicates the extent to which ER definitions precisely match overlapping exon boundaries. The MCC and MRG pair that generated the transcriptome with the lowest median exon delta and highest number of ERs with exon delta equal to 0 was chosen as the most accurate transcriptome definition for each tissue. Across all tissues, 50-54% of the ERs tested had an exon delta = 0, suggesting we had defined the majority of ERs accurately. Taking the cerebellum as an example, our optimised ERs were on average 96bps (67% of the median exon size) more accurate than would be generated if we had applied the derfinder parameters used in the existing literature (MCC: 0.5, MRG: None equivalent to 0) (Figure 1a &1b). In summary, we improved upon and optimised existing methodology to detect genome-wide transcription without reliance on existing annotation and as a result, defined 41 tissue-specific transcriptomes with increased accuracy.

Novel transcription is widespread across all human tissues and most commonly observed in brain

To assess how much of the detected transcription was novel, we calculated the total size in base pairs of ERs that did not overlap known annotation. ERs were first categorised with respect to the genomic features (exons, introns, intergenic) with which they overlapped as defined by the Ensembl v92 reference annotation (Supplementary Figure 2a). Those that solely overlapped intronic or intergenic regions were classified as novel. We discovered 8.4 to 22Mb of novel transcription across all tissues, consistent with previous reports that annotation remains incomplete[26,27]. Novel ERs predominantly fell into intragenic regions suggesting we were improving the annotation of known genes rather than discovering new genes (Figure 2a).

Although novel transcription was found to be ubiquitous across tissues, the abundance varied greatly between tissues (Figure 2e, 2f). To investigate this further, we compared the variance in total Mb of exonic ERs to intergenic and intronic ERs, normalised to the mean total Mb of each respective annotation feature. We found that the levels of novel transcription varied 3.4-7.7x more between tissues than the expression of exonic

ERs (normalised variance of exonic ERs: 0.066Mb, intronic ERs: 0.222Mb, intergenic ERs: 0.481Mb). Furthermore, focusing on a subset of novel ERs for which we could infer the precise boundaries of the presumed novel exon (using intersecting split reads), we found that more than half of these ERs were detected in only 1 tissue and that 85.9% were found in less than 5 tissues (Supplementary figure 3). This suggests that novel ERs are largely derived from tissue-specific transcription, potentially explaining why they have not already been discovered.

This finding lead us to hypothesise that genes highly expressed in brain would be amongst the most prone to misannotation, due to the difficulty of sampling human brain tissue, the cellular heterogeneity of this tissue and the particularly high prevalence of alternative splicing[12]. As we predicted, the quantity of novel transcription found within brain was significantly higher than non-brain tissues (p-value: 2.35e-10) (Figure 2e & 2f). In fact, ranking the tissues by descending Mb of novel transcription demonstrated that brain tissues constituted 13 of the top 14 most misannotated tissues. Interestingly, the importance of improving annotation in the human brain tissue was most apparent when considering purely intergenic ERs and ERs that overlapped exons and extended into intergenic regions (Figure 2d & 2e). This observation lead us to question whether there were specific features of a gene, which could be used to predict, which genes were most likely to be misannotated. We ran a logistic regression testing whether gene properties including measures of structural gene complexity and specificity of expression in human brain increased its likelihood of being misannotated. We also accounted for factors which might be expected to contribute to errors in ER identification, including whether the gene overlapped with another known gene making attribution of reads more complex. We found that the annotation of brain-specific genes and those with higher transcript complexity were more likely to have evidence for incomplete annotation (Table 1). Overlapping genes, with gene length accounted for, were not significantly more misannotated, demonstrating that novel transcription is not merely a product of noise from intersecting genes. Together these findings demonstrate that widespread novel transcription is found in all human tissues, but the quantity varies extensively between tissues, with the genes with brain-specific expression being most prominently misannotated.

Novel transcription validates across different versions of Ensembl and within an independent dataset

We recognised that a proportion of novel transcription may originate from technical variability or pre-mRNA contamination, thus we aimed to assess the reliability of novel ERs through validation across different versions of Ensembl and within an independent dataset. Firstly, we measured how many Kb of novel transcription would now be considered annotated in Ensembl v92 if we had performed this categorisation using Ensembl v87. Across all tissues, an average of 68Kb (43-127Kb) of transcription that was novel with respect to Ensembl v87 was now annotated in Ensembl v92. This was 5.3x (3.2-10.1x) greater in every tissue compared to the Kb of ERs overlapping exons in Ensembl v87 that had become purely intronic or intergenic in Ensembl v92, suggesting that the quantity of validated novel ERs was over and above what would be estimated to be detected solely through refinements to the gene annotation across Ensembl versions (Figure 3a). To further assess whether this was greater than what would be expected by chance, we compared the total Kb of novel ERs entering v92 annotation for each tissue to 10,000 sets of random length-matched intronic and intergenic

regions. For all tissues, the total Kb of both intronic and intergenic ERs that were now annotated in Ensembl v92 was significantly higher than the total Kb distribution of the randomised negative control regions, implying a high validation rate of novel ERs (Supplementary figure 4). Notably, brain regions had significantly higher Kb of ERs entering Ensembl v92 annotation from Ensembl v87 than other non-brain tissues, even when subtracting the Kb of ERs leaving Ensembl v87 (p-value: 7.6e-9), suggesting the greater abundance of brain-specific misannotation was not purely attributed to increased level of noise.

Since comparison of different Ensembl versions was limited to confirming a small subset of novel ERs, in order to gain an overall indication of the rate of validation across all ERs, we investigated whether our GTEx frontal cortex derived ERs could also be discovered in an independent frontal cortex dataset reported by Labadord and colleagues[20]. As expected, ERs which overlapped with annotated exons had near complete validation (>= 89%), but importantly 62% of intergenic and 70% of intronic ERs respectively were also detected in the second independent frontal cortex dataset (Figure 3b). While this high validation rate implied the majority of all ERs were reliably detected, we investigated whether a subset of ERs that had evidence of RNA splicing as well as transcription would have even better rates of validation. Evidence of transcription is provided by the coverage data derived using derfinder, whilst split reads, which are reads with a gapped alignment to the genome, provide evidence of intron splicing (Supplementary figure 2b). With this in mind, we focused our attention on the putative spliced ERs as indicated by the presence of an overlapping split read. Consistent with expectation, we found that ERs with split read support had higher validation rates than ERs lacking this additional feature. This increase in validation rate for ERs with split read support was greatest for intergenic and intronic ERs with the validation rate rising to 87% for intergenic ERs and 88% for intronic ERs (as compared to 99% for ERs overlapping exons, Figure 3b). Even when considering this set of highly validated ERs with split read support, 1.7-3.8Mb of intronic and 0.5-2.2Mb of intergenic transcription was detected across all 41 tissues. In summary, the majority of novel ERs were reliably detected and validated in an independent dataset.

Unannotated expressed regions are functionally important and some have the potential to be protein coding

We investigated whether novel ERs were likely to be of functional significance using measures of both conservation and genetic constraint. The degree to which a base is evolutionarily conserved across species is strongly dependent on its functional importance and accordingly, conservation scores have been used to aid exon identification[28]. However, this measure is unable to capture genomic regions of human-specific importance. Thus, we investigated novel ERs not only in terms of conservation but also genetic constraint. Constraint scores, measured here as a context-dependent tolerance score (CDTS), represent the likelihood of a base to be mutated within humans[23]. By comparing our detected novel ERs to 10,000 randomised sets of length-matched intronic and intergenic regions, we found that novel ERs were less conserved than expected, but significantly more constrained than expected by chance (p-value < 2e-16, Figure 4a). This would suggest that they have an important functional role specifically in humans. Furthermore, considering the importance of higher-order cognitive functions in differentiating humans from other species, we measured the constraint of brain-specific novel ERs separately, on the basis that these ERs may be the most genetically constrained of all

novel ERs identified. Indeed, we found that brain-specific novel ERs were significantly more constrained than other novel ERs, supporting the view that improvements in gene annotation are likely to have a disproportionate impact on our understanding of human brain diseases (Figure 4b).

Another metric of functional importance is whether a region of the genome is translated into protein and notably most known Mendelian disease mutations fall within protein-coding regions. For this reason, we investigated whether novel ERs could potentially encode for proteins. Here, we focused on the subset of novel ERs which had evidence of splicing, since the overlapping split reads could be used to assign the precise boundaries of ERs, allowing us to confidently retrieve the DNA sequence and corresponding amino acid sequence for each novel ER. A total of 2,961 ERs covering 274Kb were found to be potentially protein coding, which represented 57% of the ERs analysed, highlighting the possibility of pathogenic variants disrupting protein function having been overlooked within these regions due to incomplete annotation.

Misannotation of OMIM genes may limit genetic diagnosis

We assessed the completeness of annotation for genes already known to cause Mendelian disease, since misannotation of this set would likely have the greatest impact on genetic diagnosis. Novel ERs were first connected to known genes using split reads (Supplementary figure 2b, see detailed Methods) to provide a conservative estimate of novel annotation. Next, we filtered for OMIM-morbid genes and found that 63% of this set of OMIM-morbid genes were misannotated, suggesting that despite many of these genes having been extensively studied, the annotation of most OMIM-morbid genes remains incomplete (Figure 5a). 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 misannotated OMIM-morbid genes had an associated novel ER originating from a phenotypically relevant tissue (Figure 5b). This phenomenon was exemplified by *MYH3*, which encodes the embryonic myosin heavy chain 3 protein and when mutated can cause distal arthrogryposis types 2A (Freeman-Sheldon syndrome), 2B (Sheldon-Hall syndrome) and type 8 (multiple pterygium syndrome)[29,30]. We detect a 117bp intronic ER in *MYH3* in skeletal muscle, which matches the affected disease tissue (Figure 5d). Interestingly, this ER which connects to two protein-coding exons of *MYH3* through split reads, is not conserved within mammals (phastCons7: 0) but is amongst the top 11% of most constrained regions of the genome suggesting a human-specific function. We postulate that the poor conservation of the ER and the complex pattern of *MYH3* gene expression with post-natal expression being limited to specialised muscle tissues (such as extraocular, jaw-closing and regenerating muscle)[31], would explain why this probable novel exon was not previously reported. Similarly, we detected a cerebellar-specific 72bp intronic ER with respect to *ERLIN1.* When disrupted this gene 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[32]. The novel ER we detected in cerebellum 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 5c). Currently, variants located in the novel ERs detected in both *MYH3* and *ERLIN1* would not be captured using WES and if identified in WGS would be misassigned as non-coding variants. This would mean that identification of pathogenicity would be highly driven by statistical evidence of the variant associating with cases rather than controls which would be very challenging for such rare disorders (–100 Freeman Sheldon syndrome: ~100 known cases, Sheldon-Hall syndrome: <100 known cases, Multiple Pterygium syndrome: ~50 known cases, Spastic Paraplegia 62: prevalence <1/1,000,000).

Discussion

In recent years, the use of next-generation sequencing has changed the landscape of clinical genetics. WES and, to a lesser extent, WGS are becoming key components of diagnostic testing and have dramatically accelerated the discovery of new disease-causing genes. However, recent analyses predict that there is a finite pool of disease-causing genes, which will be exhausted by 2020 or earlier[33]. With the reducing number of potential disease genes left to discover, we believe genetic diagnosis will become ever more reliant on the accuracy and completeness of the annotation of known disease-related genes. Foreseeing this, we build on existing resources to develop a method to accurately detect novel transcription in an annotation-agnostic manner, connect novel ERs to known genes and ultimately, improve the annotation of 63% of all OMIM-morbid genes.

We find that most probable novel exons we detect have a restricted expression pattern, which is often disease-relevant and significantly more abundant in brain. Furthermore, since our approach does not rely on conservation across species to annotate novel exons, we are able to identify ERs which are likely to be of human-specific importance. Using constraint scores generated from aligning 7,794 human genomes and PhastCons conservation scores we find that collectively our probable novel exons, while not necessarily conserved are depleted for genetic variation within humans suggesting that they are potential sites for pathogenic variation[23]. Interestingly, the putative tissue-specific origin and human-specific functions of the novel transcription we detect also provides a reasonable explanation for their omission from existing annotation databases and the abundance of novel transcription in human brain. The practical difficulty of accessing the brain reduces the number of available brain-specific datasets and its higher transcriptomic diversity is known to generate a higher number of brain-specific transcripts. In addition, we find that brain-specific ERs have the highest constraint scores, emphasising their specific importance in humans. Together these factors suggest that the resource we have generated will have the greatest impact on the diagnosis of neurogenetic disorders.

Since the underlying aim of our analyses is the improvement of diagnostic yield for genetic testing, we provide the vizER web interface and BED-formatted descriptions of novel ERs, which will serve as an important resource for clinical scientists and clinicians in the diagnosis of Mendelian disorders. Finally, we note that as the availability of cell-specific and cell state-specific RNA-seq data increases, novel exon discovery is likely to accelerate and will provide additional insights into the molecular processes underpinning rare genetic disorders.

Figures

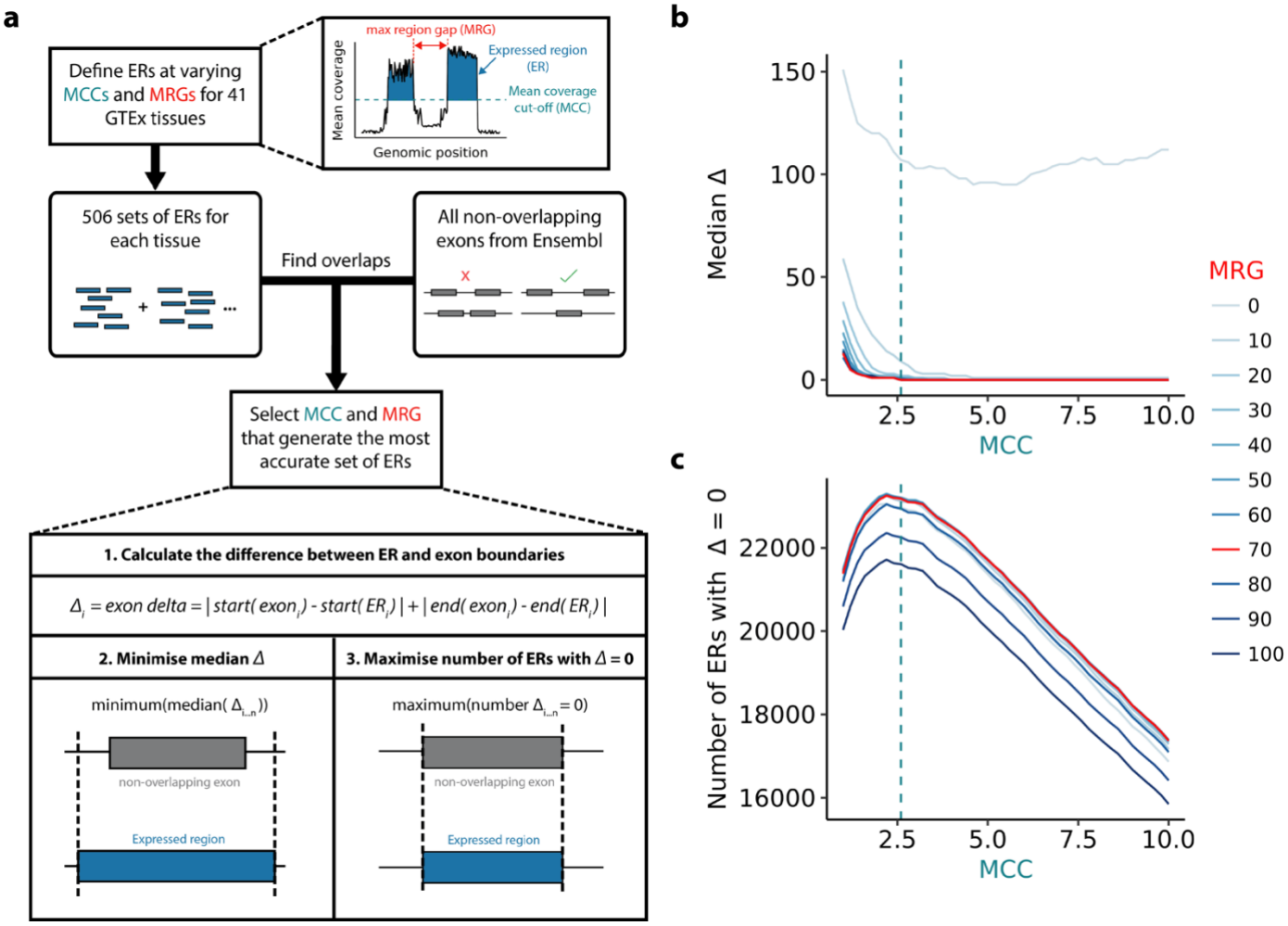


Figure 1 – ***Optimisation of the detection of transcription. a)*** *Transcription in the form expressed regions (ERs) was detected in an annotation agnostic manner across 41 human tissues. The mean coverage cut-off (MCC) is the number of reads supporting each base above which that base would be considered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

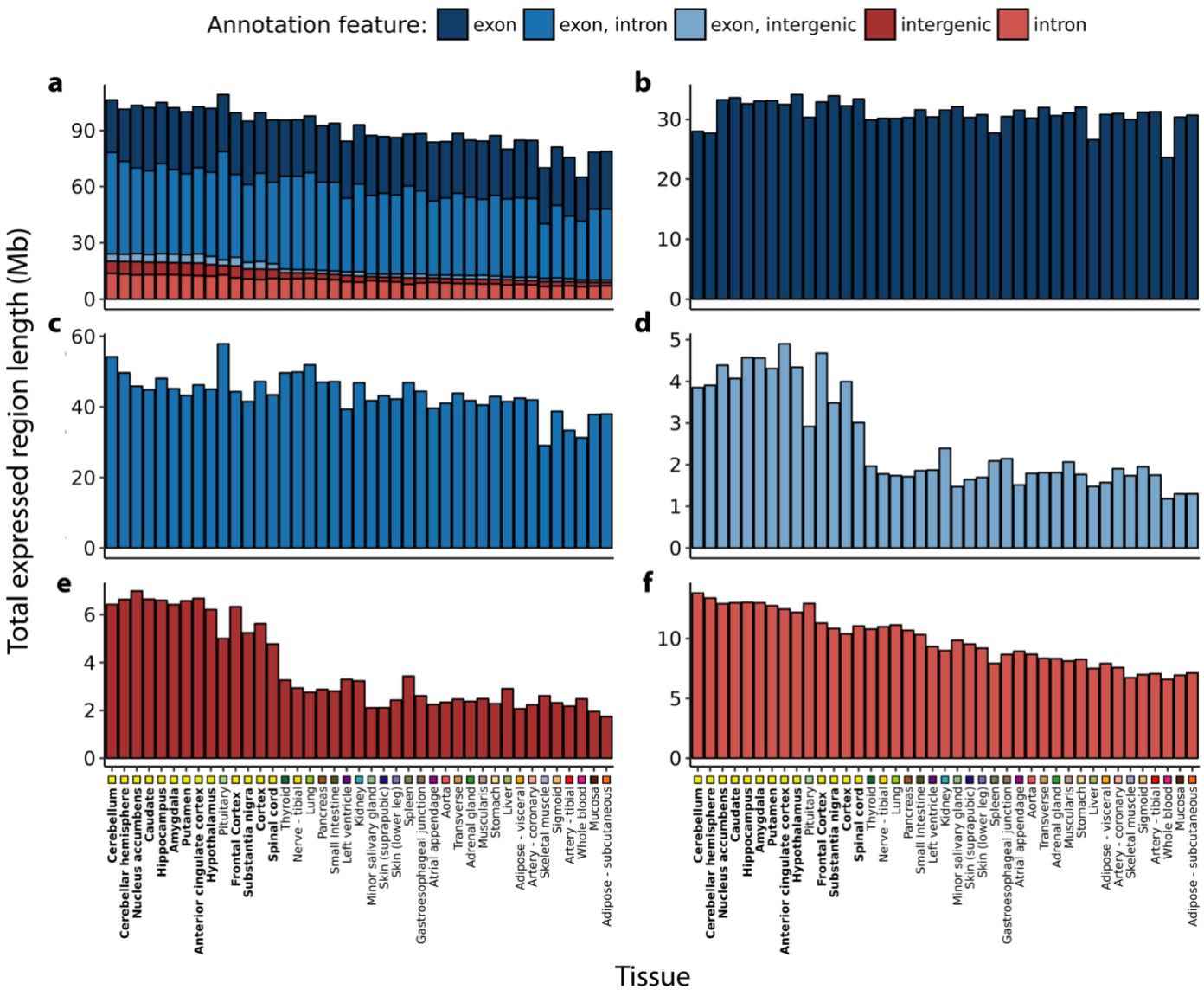


Figure 2 – ***Transcription detected across 41 GTEx tissues categorised by annotation feature.*** *Within each tissue the length of the ERs Mb overlapping* ***a)*** *all annotation features* ***b)*** *purely exons* ***c)*** *exons and introns* ***d)*** *exons and intergenic regions* ***e)*** *purely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

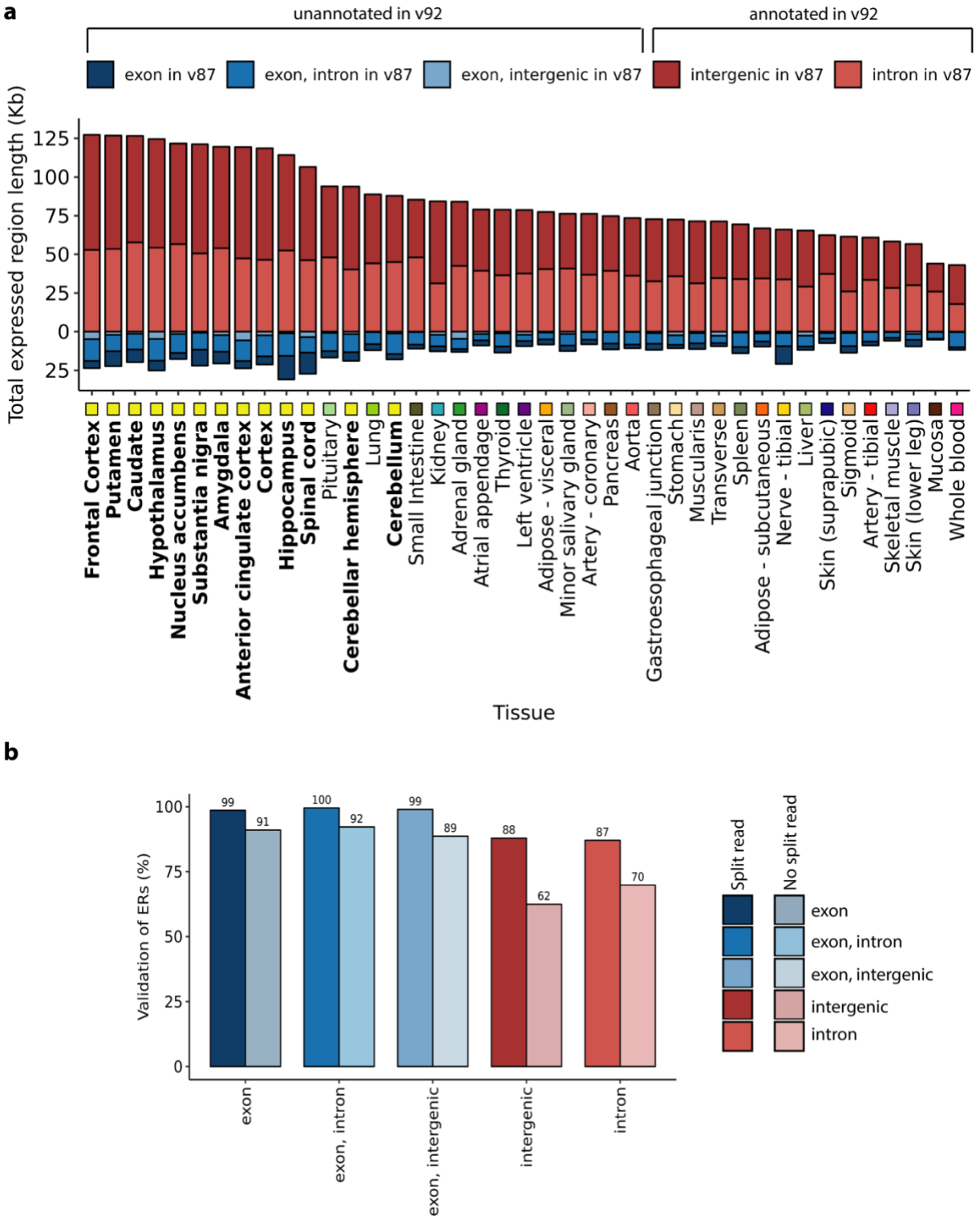


Figure 3 – ***Validation of novel transcription. a)*** *The classification of ERs based on v87 and v92 of Ensembl was compared. Across all tissues, the number of intron or intergenic ERs with respect to v87 that were known to be exonic in the newest version of Ensembl (minimum 250) was greater than the number of ERs overlapping exons according to v87 that were now unannotated in v92 (maximum 87). Tissues are plotted in descending order based on their respective total size of intergenic and intronic ERs. Tissues are colour-coded as indicated in the x-axis, with GTEx brain regions highlighted with bold font.* ***b)*** *Barplot represents the percentage of ERs seeding from the GTEx frontal cortex that validated in an independent frontal cortex RNA-seq dataset reported by Labadord and colleagues. ERs defined in the seed tissue were re-quantified using coverage from the validation dataset, after which the optimised mean coverage cut off was applied to determine validated ERs. Colours represent the different annotation features that the ERs overlapped and the shade indicates whether the ER was supported by split read(s).*

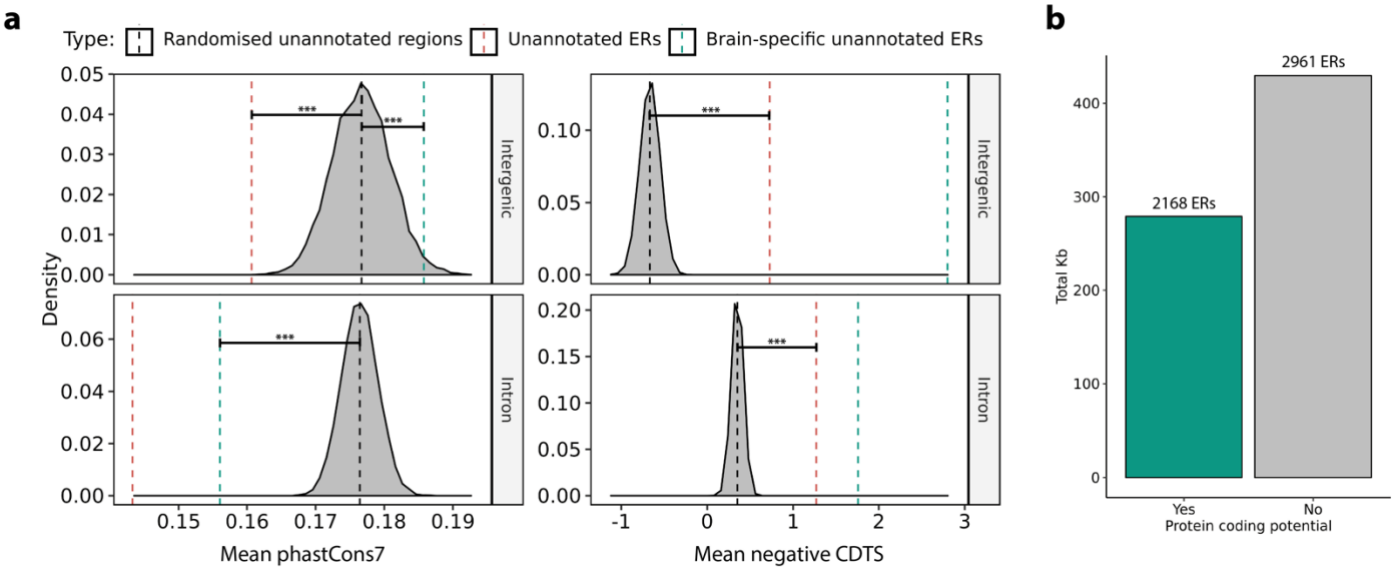


Figure 4 – ***Novel ERs collectively serve an important function for humans and a proportion can form potentially protein coding transcripts. a)*** *Comparison of conservation (phastCons7) and constraint (CDTS) of intronic and intergenic ERs to 10,000 sets of random, length-matched intronic and intergenic 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.*

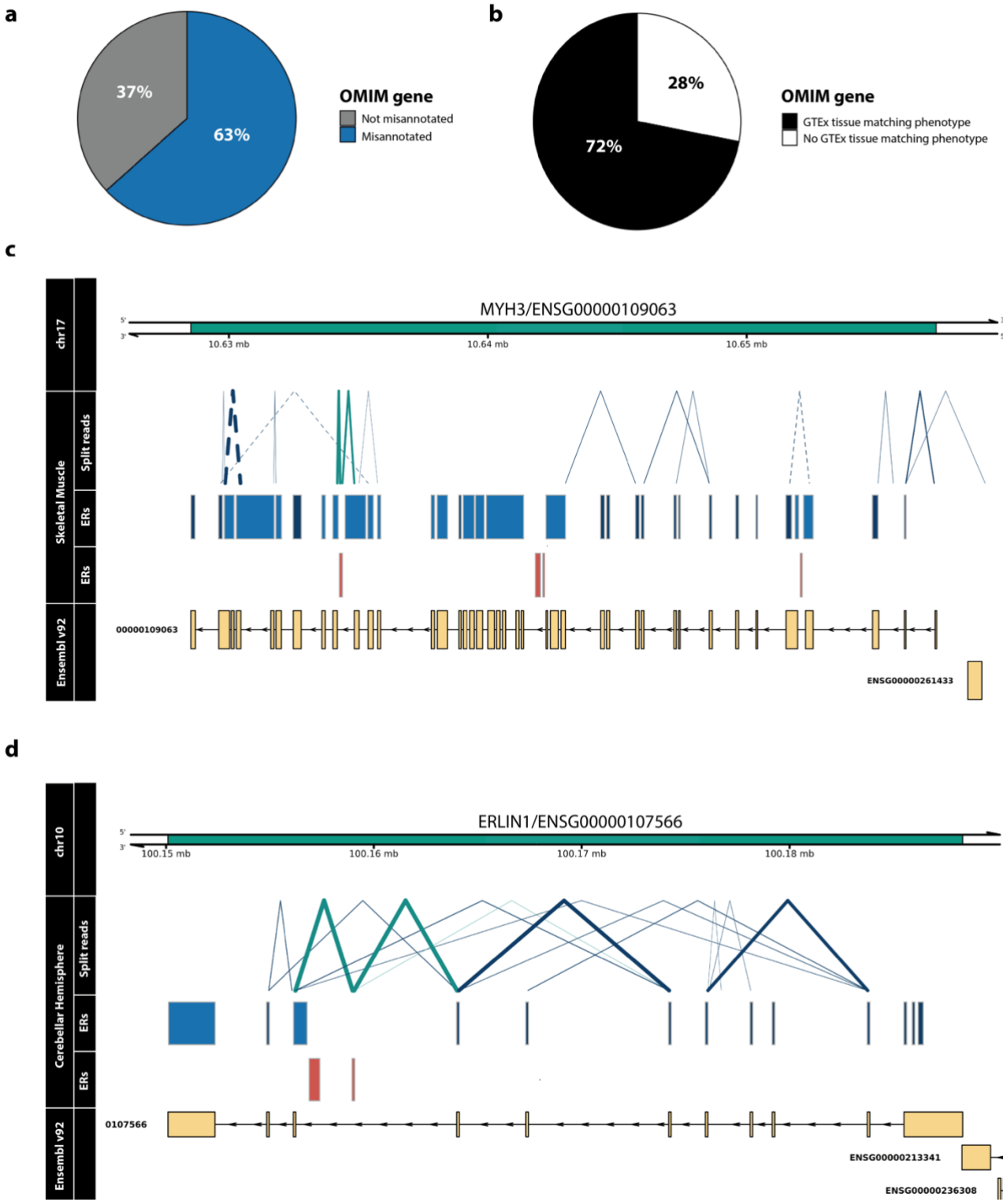


Figure 5 – ***Misannotation of OMIM genes. a)*** *A novel ER connected through a split read was discovered for 63% of OMIM-morbid genes.* ***b)*** *Comparison of the phenotype (HPO terms) associated with each misannotated OMIM-morbid gene and the GTEx tissue from which misannotations were derived. Through manual inspection, HPO terms were matched to disease-relevant GTEx tissues and for 72% of misannotated OMIM genes, the associated novel ER was detected in the phenotype-relevant tissue. Visualised examples of misannotated OMIM-morbid genes* ***c)*** *ERLIN1 and* ***d)*** *MYH3. Top track represents the genomic region including the gene of interest marked in green. Second group of tracks detail the split reads and ERs overlapping the genomic region derived from the labelled tissue. Blue ERs overlap known exonic regions and red ERs fall within intronic or intergenic regions. Blue split reads overlap blue ERs, while green split reads overlap both red and blue ERs, connecting novel ERs to OMIM-morbid genes. Thickness of split reads represents the proportion of samples of that tissue in which the split read was detected. Only partially annotated split reads (solid lines) and unannotated split reads (dashed lines) are*

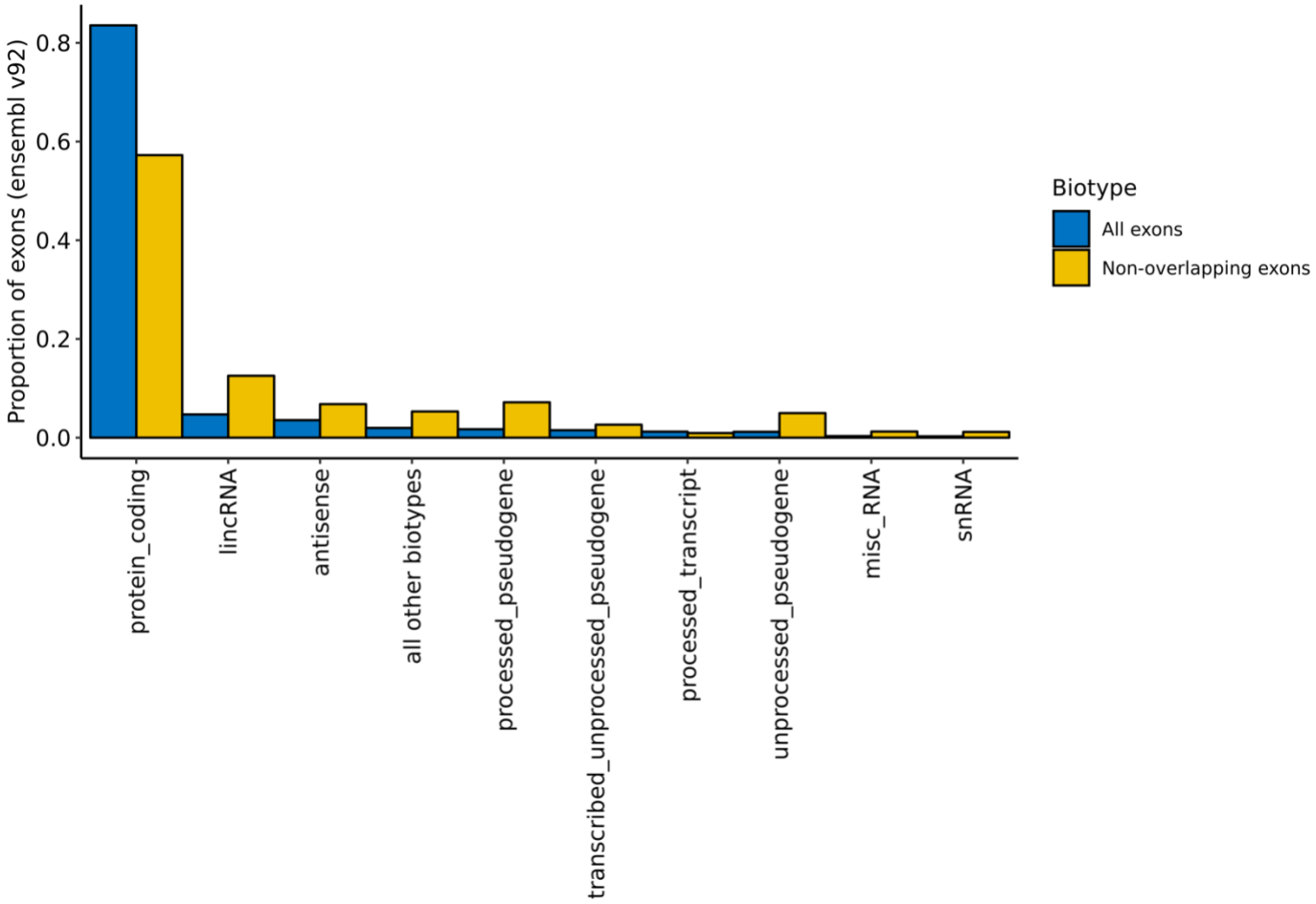
plotted. The last track displays the genes within the region according to Ensembl v92, with all known exons of the gene collapsed into one "meta" transcript.

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

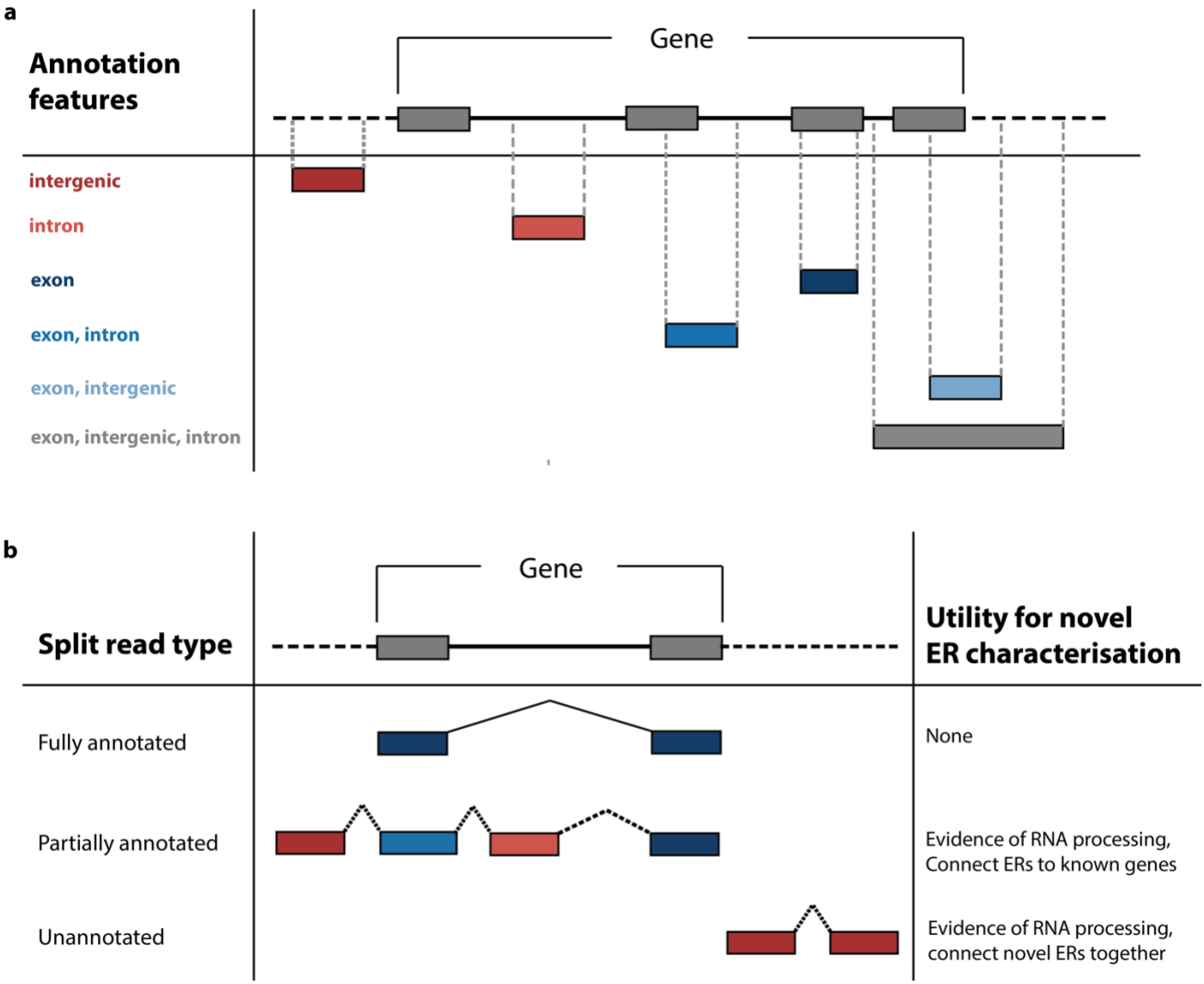
*** p <= 2e-16

Table 1 – ***Gene properties influencing misannotation*** *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 misannotated. Brain-specific, longer, protein-coding genes of high transcript complexity were the most likely to be misannotated. Blue and red highlights positive and negative significant estimates, respectively.*

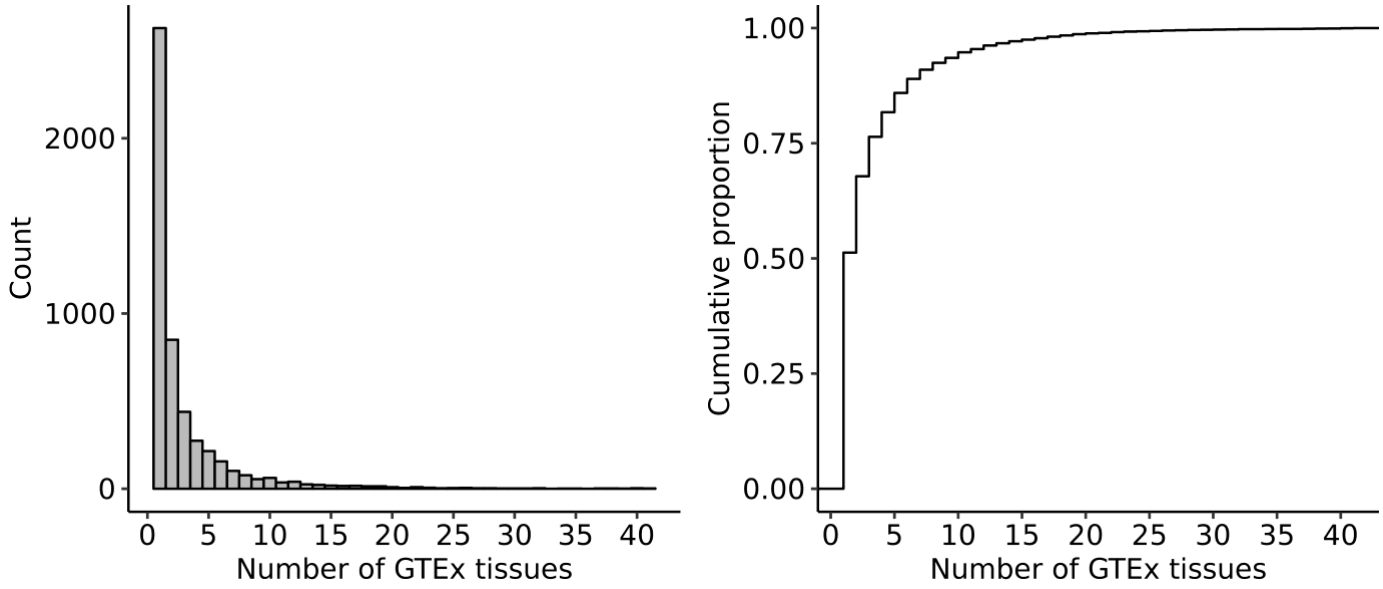
Supplementary figures



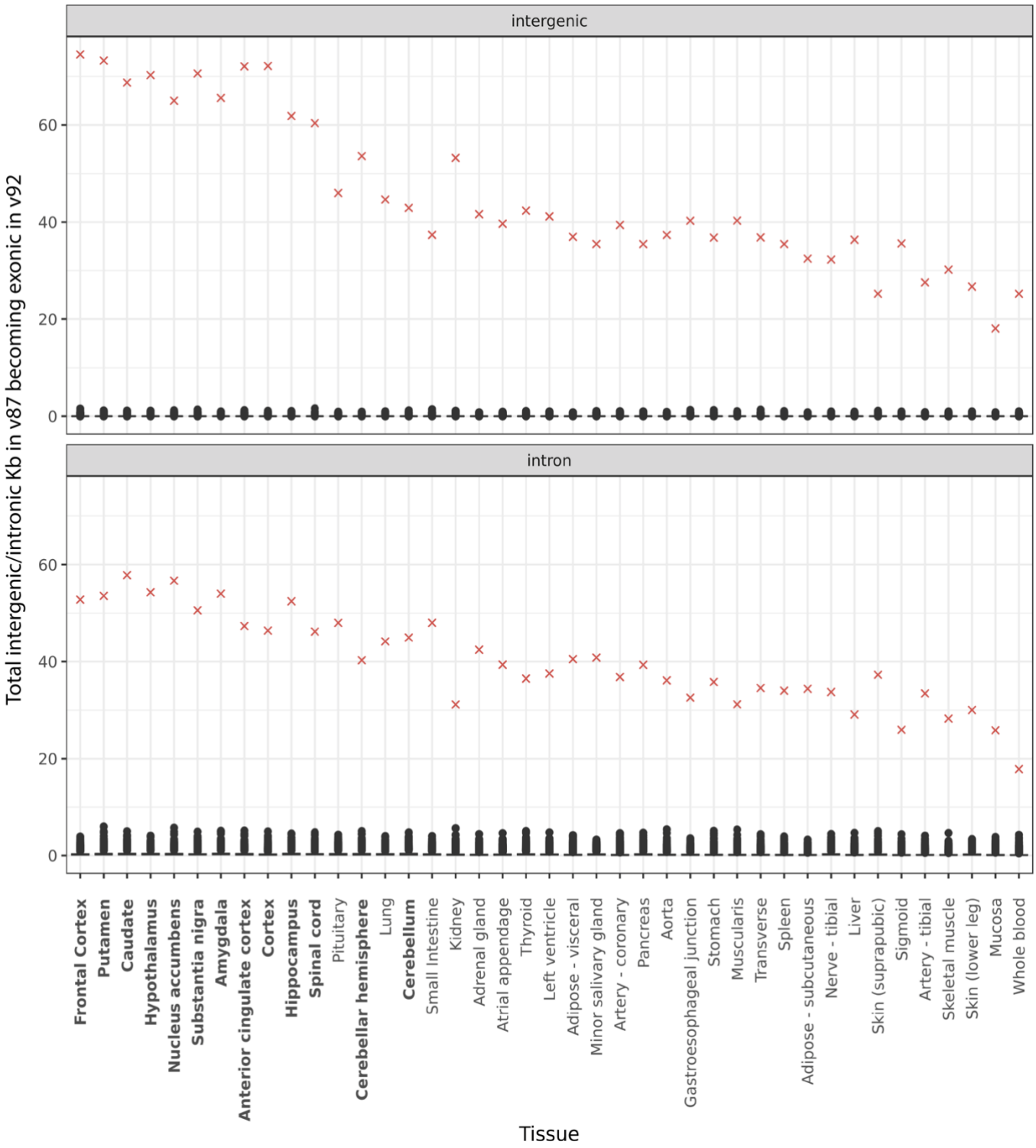
Supplementary figure 1 – ***Proportion of exons that fall into different gene biotypes.*** *Comparison of the proportion of exons that are classified within the different gene biotypes between all exons from Ensembl v92 and the non-overlapping set of exons used to optimise the detection of transcription.*



Supplementary figure 2 – ***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 3 – ***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 4 – ***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).*

Bibliography

1. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17,** 405–424 (2015).
2. Bamshad, M. J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nat. Rev. Genet.* **12,** 745–755 (2011).
3. Gilissen, C. *et al.* Genome sequencing identifies major causes of severe intellectual disability. *Nature* **511,** 344–347 (2014).
4. Belkadi, A. *et al.* Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc. Natl. Acad. Sci.* **112,** 5473–5478 (2015).
5. Taylor, J. C. *et al.* Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. *Nat. Genet.* **47,** 717–726 (2015).
6. Yang, Y. *et al.* Molecular findings among patients referred for clinical whole-exome sequencing. *JAMA - J. Am. Med. Assoc.* **312,** 1870–1879 (2014).
7. Chong, J. X. *et al.* The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *Am. J. Hum. Genet.* **97,** 199–215 (2015).
8. Chen, G. *et al.* Incorporating the human gene annotations in different databases significantly improved transcriptomic and genetic analyses. *RNA* **19,** 479–89 (2013).
9. McCarthy, D. J. *et al.* Choice of transcripts and software has a large effect on variant annotation. *Genome Med.* **6,** (2014).
10. Steward, C. A. *et al.* Genome annotation for clinical genomic diagnostics: Strengths and weaknesses. *Genome Med.* **9,** 1–19 (2017).
11. Fogel, B. L., Lee, H., Strom, S. P., Deignan, J. L. & Nelson, S. F. Clinical exome sequencing in neurogenetic and neuropsychiatric disorders. *Ann. N. Y. Acad. Sci.* **1366,** 49–60 (2016).
12. Yeo, G., Holste, D., Kreiman, G. & Burge, C. B. Variation in alternative splicing across human tissues. *Genome Biol.* **5,** R74 (2004).
13. Collado-Torres, L. *et al.* Reproducible RNA-seq analysis using recount2. *Nat. Biotechnol.* **35,** 319–321 (2017).
14. Zhang, Y. E., Landback, P., Vibranovski, M. & Long, M. New genes expressed in human brains: Implications for annotating evolving genomes. *BioEssays* **34,** 982–991 (2012).
15. Hamosh, A., Scott, A. F., Amberger, J., Valle, D. & McKusick, V. A. Online Mendelian Inheritance in Man (OMIM). *Hum. Mutat.* **15,** 57–61 (2000).
16. The GTExArd Consortium *et al.* The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science (80-.).* **348,** 648–60 (2015).
17. Collado-Torres, L. *et al.* Flexible expressed region analysis for RNA-seq with derfinder. *Nucleic Acids Res.* **45,** e9 (2017).
18. Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res.* **46,** D754–D761 (2018).
19. Irimia, M. *et al.* A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell* **159,** 1511–1523 (2014).

20. Labadorf, A. *et al.* RNA sequence analysis of human huntington disease brain reveals an extensive increase in inflammatory and developmental gene expression. *PLoS One* **10,** 1–21 (2015).
21. Siepel, A. & Haussler, D. Phylogenetic Hidden Markov Models. 26 (2005).
22. Wainberg, M., Alipanahi, B. & Frey, B. Does conservation account for splicing patterns? *BMC Genomics* **17,** 1–10 (2016).
23. Di Iulio, J. *et al.* The human noncoding genome defined by genetic diversity. *Nat. Genet.* **50,** 333–337 (2018).
24. Pagès, H., Aboyoun, P., Gentleman, R. & DebRoy, S. Biostrings: Efficient manipulation of biological strings. *R package version 2.46.0* (2017). doi:10.1021/jm900485a
25. Finucane, H. K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50,** 621–629 (2018).
26. Jaffe, A. E. *et al.* Developmental regulation of human cortex transcription and its clinical relevance at single base resolution. *Nat. Neurosci.* **18,** 154–161 (2015).
27. Pertea, M. *et al.* Thousands of large-scale RNA sequencing experiments yield a comprehensive new human gene list and reveal extensive transcriptional noise. *Genome Biol.* 332825 (2018). doi:10.1101/332825
28. Harrow, J. *et al.* GENCODE: The Reference Human Genome Annotation for The ENCODE Project. *Genome Res.* **22,** 1760–1774 (2012).
29. Toydemir, R. M. *et al.* Mutations in embryonic myosin heavy chain (MYH3) cause Freeman-Sheldon syndrome and Sheldon-Hall syndrome. *Nat. Genet.* **38,** 561–565 (2006).
30. Chong, J. X. *et al.* Autosomal-dominant multiple pterygium syndrome is caused by mutations in MYH3. *Am. J. Hum. Genet.* **96,** 841–849 (2015).
31. Schiaffino, S., Rossi, A. C., Smerdu, V., Leinwand, L. A. & Reggiani, C. Developmental myosins: Expression patterns and functional significance. *Skelet. Muscle* **5,** 1–14 (2015).
32. Novarino, G. *et al.* Exome Sequencing Links Corticospinal Motor Neuron Disease to Common Neurodegenerative Disorders. *Science (80-.).* **343,** 506–511 (2014).
33. Boycott, K. M., Vanstone, M. R., Bulman, D. E. & MacKenzie, A. E. Rare-disease genetics in the era of next-generation sequencing: Discovery to translation. *Nat. Rev. Genet.* **14,** 681–691 (2013).