Title: Long-read sequencing reveals the splicing profile of the calcium channel gene *CACNA1C* in human brain

Authors: Michael B Clark[1,2#], Tomasz Wrzesinski[3,#], Aintzane García-Bea[1], Joel E Kleinman[4], Thomas Hyde[4], Daniel R Weinberger[4], Wilfried Haerty[3,*], Elizabeth M Tunbridge[1,5,*]

[#,] *These authors contributed equally to the manuscript

1. Department of Psychiatry, University of Oxford, UK
2. Garvan Institute of Medical Research, NSW, Australia
3. The Earlham Institute, Norwich, UK
4. The Lieber Institute for Brain Development, Baltimore, US
5. Oxford Health NHS Foundation Trust, Oxford, UK

Corresponding Author: Elizabeth Tunbridge
University Department of Psychiatry
Warneford Hospital
Oxford, OX3, 7JX, UK
elizabeth.tunbridge@psych.ox.ac.uk

Abstract

RNA splicing is a key mechanism linking genetic variation and complex diseases, including schizophrenia. Splicing profiles are particularly diverse in the brain, but it is difficult to accurately identify and quantify full-length isoforms using standard approaches. *CACNA1C* is a large gene that shows robust genetic associations with several psychiatric disorders and encodes multiple, functionally-distinct voltage-gated calcium channels via alternative splicing. We combined long-range PCR with nanopore sequencing to characterise the full-length coding sequences of the *CACNA1C* gene in human brain. We show that its splice isoform profile varies between brain regions and is substantially more complex than currently appreciated: we identified 38 novel exons and 83 high confidence novel isoforms, many of which are predicted to alter protein function. Our findings demonstrate the capability of long-read amplicon sequencing to effectively characterise human splice isoform diversity, while the accurate characterisation of *CACNA1C* isoforms will facilitate the identification of disease-linked isoforms.

Introduction

In recent years, large-scale genomic studies have identified numerous common single nucleotide polymorphisms that are robustly associated with psychiatric disorders; the challenge is now to understand the pathophysiological mechanisms underlying these associations[1]. However, the majority of the variants identified thus far are non-coding[2-5] and therefore the molecular mechanisms of disease likely involve effects on RNA expression and splicing, rather than direct effects on the protein sequence. Notably, cellular studies implicate RNA splicing as a key mechanism mediating the effect of disease-associated, non-coding variants, including those linked with schizophrenia[6] (the psychiatric disorder for which the most genetic loci have been identified to date[5]). Indeed, examples of associations between schizophrenia risk loci and variants impacting both exon splicing and alternatively spliced transcript representation are beginning to emerge[7]. These observations highlight the pressing need to better characterize the diversity and functional relevance of splicing isoforms expressed in the human brain and the potentially deleterious impact of variants altering their expression.

Previous analyses have reported extensive splicing variation across human development and aging[8] and among tissues; especially in brain, which exhibits the highest levels of splicing diversity and the prominent use of tissue specific exons, microexons and splicing factors[9-13]. However, despite extensive efforts in improving the annotation of the human genome[14], the alternative splicing patterns of many human genes remain poorly described. Furthermore, the current annotations are incomplete, with novel coding exons (many enriched within brain expressed genes[15,16]) and transcript isoforms continually being discovered. In part, this is due to the limited availability of high-quality, post-mortem human tissue; however, there are also technical limitations associated with standard approaches. For example, short-read RNA-Seq identifies mainly relatively highly-expressed mRNAs within the total cellular pool, meaning that a significant proportion of gene and splice isoforms may be missed. It also relies on the reconstruction of fragmented sequences into full-length transcripts, which makes the disambiguation of full-length isoforms difficult, particularly in the case of large and complex genes. Exemplifying this issue, benchmarking studies have shown that short-read RNA-Seq methodologies are unable to accurately reconstruct and quantitate the majority of transcript isoforms[17,18].

The relative lack of knowledge about the splicing of individual human genes is typified by *CACNA1C*. *CACNA1C* encodes a voltage-gated calcium channel (VGCC) and has emerged as a leading candidate gene for psychiatric disorders from large-scale genomic studies, given associations with non-coding polymorphisms within its locus[2-5]. The *CACNA1C* gene is large and complex, with at least 50 exons and 40 predicted isoforms (arising from both splicing and transcriptional mechanisms; Figure 1A)[12]. The large transcript diversity displayed by VGCC genes has significant functional implications[13]. Alternative splice isoforms encode functionally different VGCCs, and

the transcript profile of *CACNA1C* differs between tissues[19]. For example, *CACNA1C* can be expressed from at least two alternative 'start' exons: exons 1A and 1B, which predominate in heart and brain, respectively, and encode proteins with alternate, functionally-distinct intracellular N termini[19,20]. However, many key details remain unknown. Firstly, because the *CACNA1C* gene is large and complex, accurate isoform identification and quantification by standard gene expression measures is extremely difficult, and so the full-length protein coding sequence of isoforms remains unclear. Secondly, information on the isoform diversity of *human* VGCC subunits in different tissue types is sparse as most studies have focused on rodents[13]. Finally, the extent to which human brain VGCC transcript profiles differ between individuals (and how these differences relate to genetic factors and disease risk) is unknown, since the few studies examining VGCC splicing in human brain have used pooled, or single-individual, samples[14,15]. Therefore, there is a need for a straightforward and cost-effective means of sequencing full-length, protein-coding isoforms from human tissues to begin to address some of these unknowns.

Recent advances in sequencing technologies have included the advent of "long-read" sequencing that does not require the fragmentation of RNA or DNA before sequencing and so can sequence full-length mRNAs. With the entire mRNA or coding sequence (CDS) contained within a single read, the presence and order of each exon can be unambiguiously determined without the need for computational transcript reconstruction methods. Previous work has validated that nanopore long-read sequencing on the Oxford Nanopore MinION can sequence full-length mRNAs and be used to uncover isoform complexity[21-23]. Despite the potential of nanopore sequencing for investigating gene expression, challenges remain. Individual reads have a high error rate and few software pipelines to allow the identification and quantification of alternatively spliced isoforms exist.

Here, we demonstrate that long-range RT-PCR can successfully be combined with long-read nanopore sequencing to characterise full-length *CACNA1C* CDSs in human, post-mortem brain. In doing so, we show that the transcript structure of human *CACNA1C* is substantially more complex than currently appreciated, with numerous novel isoforms containing un-annotated exons and in-frame deletions, which are predicted to alter the protein sequence and function. Furthermore, although not directly quantitative, our approach is sensitive to differences in the *CACNA1C* isoform profile between brain regions, indicating that it can be used to detect gross changes in splicing between tissues. This would allow, for example, the identification of splice isoforms showing relative enrichment in brain vs. peripherally, which may prove to be selective therapeutic targets for psychiatric and neurological disorders.

Methods

All molecular biology protocols were conducted according to manufacturer's recommendations unless specifically noted otherwise.

Post-mortem brain tissue

Postmortem human brain tissues were collected from the Office of the Chief Medical Examiner for the State of Maryland according to a protocol approved by the Institutional Review Board of the State of Maryland Department of Health and Mental Hygiene (#12-24) and the Western Institutional Review Board (#20111080). Audiotaped informed consent to study brain tissue was obtained for the legal next-of-kin on every case. Details of the donation process and specimens handling are described previously. Each subject was evaluated for possible neurological and/or psychiatric illness retrospectively by two board-certified psychiatrists, applying the criteria in the DSM-IV-R. Subjects with evidence of neuropathology, drug use (other than alcohol), or psychiatric illness were excluded. None of the donors smoked at the time of death. Demographic information is presented in Supplementary Table 1.

RNA extraction and RT-PCR

RNA was extracted from cerebellum, striatum, dorsolateral prefrontal cortex [DLPFC], cingulate cortex, occipital cortex and parietal cortex. Tissue was disrupted with the TissueLyser LT (Qiagen, UK) using Rnase-free 5mm stainless steel beads in QIAzol Lysis Reagent (Qiagen). RNA was extracted using the Qiagen RNeasy Lipid Tissue Mini Kit (Qiagen, UK) and eluted in RNase-free water. RNA concentration and RNA integrity (RNA Integrity Number equivalent; RINe) were measured using a Tapestation 2200 (Agilent, UK). All samples had a RINe of >7.0. RNA (1µg per sample) was converted into cDNA with GoScript™ Reverse Transcriptase (Promega, UK) using oligoDT priming.

Full-length CDS amplicons of *CACNA1C* were obtained by PCR using primers located 5' to the translation start site in Exon 1B and 3' to the stop codon in the 3' untranslated region. Specifically, *CACNA1C* was amplified from cDNA equating to 125ng RNA template from the eighteen human brain samples using PrimeSTAR GXL DNA Polymerase (Takara). Tailed primers were designed to amplify the ~6.5kb *CACNA1C* coding sequence and included sequence to allow further amplification and barcoding. Forward and Reverse primers were tttctgttggtgctgatattgcCATTTCTTCCTCTTCGTGGCTGC and acttgcctgtcgctctatcttcCCAGGTCACGAGAACAGTGAGG, respectively (*CACNA1C* sequence in capital letters). Amplification was conducted for 25 cycles of 98°C for 10 sec; 57°C for 15 sec; 68°C for 7 min. PCR products were separated on a 1.5% agarose gel (visualised with GelGreen, Biotium, UK). Full-length CDS (~6.5kb) products were excised and purified using the Qiagen Gel Purification Kit (Qiagen, UK). DNA was barcoded using the Oxford Nanopore Technologies (ONT) PCR Barcoding Kit (EXP-PBC001) using PrimeSTAR GXL DNA Polymerase and 12 cycles of PCR amplification, as described above. PCR products were purified with the Qiagen PCR Purification Kit (Qiagen, UK) and quantified by Qubit (ThermoFisher, UK).

Nanopore Sequencing of the full-length CACNA1C CDS

The PCR Barcoding Kit (EXP-PBC001, ONT) contains unique twelve barcodes, hence the 18 samples were split across two flowcells with 12 samples run on each. Six samples were sequenced on both flowcells to allow comparison and normalisation between sequencing runs. Samples and barcodes used in each sequencing run are described in Supplementary Table 2.

For each sequencing run, 132ng of each of the twelve samples was pooled and re-purified using 0.4x Agencourt Ampure XP beads (Beckman Coulter, UK) to concentrate the sample and remove any contaminating small DNA products. The presence of full-length DNA product in the pool was confirmed by 2200 or 4200 TapeStation (Agilent) analysis using gDNA screentape (Supplementary Figure 1) and the sample measured again by Qubit to ensure the presence of sufficient product for nanopore library preparation.

Sequencing libraries were prepared with the 2D Nanopore Sequencing Kit (SQK-LSK208, ONT) using 1µg of *CACNA1C* DNA. DNA end-repair and dA-tailing was performed using the NEBNext Ultra II End-Repair/dA-tailing Module (NEB, UK) in a total reaction volume of 60µl (50µl DNA, nuclease-free water (NFW) and DNA calibration strand (Run1 only), 7 µl Ultra II End-Prep buffer and 3 µl Ultra II End-Prep enzyme) for 5 min at 20ºC and 5 min at 65ºC. Samples were purified using 0.6x Ampure XP beads and eluted in 31µl of NFW. Sample concentrations were measured by Qubit and confirmed retention of >85% of the initial samples. The total volume of *CACNA1C* DNA was utilised for subsequent ligation of sequencing adaptors. Ligation reactions contained 50µl Blunt / TA Ligase Master Mix (NEB), 10 µl Adapter Mix 2D, 2 µl Hairpin Adaptor (HPA), 30µl of DNA and 8µl of NFW and incubations were performed at room temperature. After 10 min incubation 1 µl Hairpin Tether (HPT) was added, the sample mixed and incubated for a further 10 minutes. Adapted DNA was purified with MyOne C1 beads (ThermoFisher). 50µl of beads were washed 2x in 100µl bead binding buffer (BBB) and resuspended in 100µl BBB before use. Beads were mixed 1:1 with adapted DNA, incubated on a rotor for 5 min at room temperature to bind DNA to the beads and collected using a magnetic stand. Bead-bound DNA was washed 2x with 150µl of BBB on the magnetic stand and left briefly to dry the beads. Beads were resuspended in 15µl of elution buffer (ELB) and incubated at 37ºC for 10 min to elute DNA from the beads. After pelleting the beads on the magnetic stand, 14µl of supernatant containing the adapted DNA library in ELB was collected. Library concentrations were measured by Qubit, recovering >25% of the original starting amount.

Flowcells were primed with a mix of 480 µl Running Buffer with Fuel Mix (RBF) and 520 µl NFW as per manufacturer's instructions. A loading mix of 35 µl RBF, 25.5 µl library loading beads (LLB), 12 µl *CACNA1C* library and 2.5 µl NFW was prepared, mixed by pipetting and immediately loaded onto the primed “spot-on” flowcell. Sequencing was performed on fresh flowcells; Run 1 utilised a R9.0 (MIN105) flowcell, while Run 2 utilised a R9.4 (MIN106) flowcell. Pore occupancy was >80% indicating a good library and flowcell. Sequencing was allowed to continue until there was a high probability of >1000 high-quality (2D pass) reads from each barcoded sample. Run 1 was base-called with the Epi2Me cloud based service (2D Basecalling plus Barcoding for FLO-MIN105 250bps - v1.125). Run 2 was base-called with Albacore V1.1.0 (FLO-MIN106, SQK-LSK208, barcoding) after the withdrawal of the Epi2Me base-calling service. The absence of a base-calling model for R9.0 in the Albacore software

prevents the base-calling of Run1 with this software. Sequencing library metrics (Supplementary Table 3) were generated by Poretools[24] and PycoQC[25].

Read mapping

2D pass reads (i.e.: those with a q-score of >=9) with an identified barcode were mapped to the transcriptome (hg38, ENSEMBL v82) using a HPC version of BLAT (pblat-cluster v0.3, available from http://icebert.github.io/pblat-cluster/) and to the *CACNA1C* meta-gene containing known exonic sequences using GMAP version 2017-04-24[26]. Reads mapping uniquely to *CACNA1C*, and with at least 50% of their length mapped and covering at 75% of an annotated transcript were retained for further analysis.
To identify potential novel exons within *CACNA1C*, we mined the alignments of the reads to the transcriptome for inserts of at least 9 nucleotides in length located at the junctions between annotated exons. To further characterize candidate novel exons, we mapped the corresponding reads to the genome using LAST v840[27]. We retained candidate exons located within the expected intronic sequences, and at least 6 nt away from existing exons. Novel exonic sequences were subsequently introduced to the *CACNA1C* meta-gene (concatenation of all exons) and reads were mapped again to this new model using GMAP to enable further characterization of alternative splicing and quantification of isoform expression.

Transcript annotation

To annotate novel transcripts that include novel exons and/or novel junctions, we parsed the CIGAR string from the alignment to the metagene to identify the combination of exon junctions supported by each read. We subsequently clustered splicing patterns to annotate a unique set of isoforms for *CACNA1C* and enable their expression quantification.

Expression quantification

Transcript expression was quantified using the number of reads supporting the transcript model. Reads mapping to multiple transcripts were down-weighted according to the number of isoforms they could be mapped to. Read counts were normalized across libraries using the trimmed Mean of M-values normalization method (TMM)[28]. For visualisation, all normalized counts were $\log_{10}$-transformed. Expression heatmaps and PCA plots were generated using R statistical language[29], by heatmap3[30] and ggplot2 libraries[31], respectively. Because of the large difference in sequencing depth between samples, we downsampled all libraries to the smallest sequencing depth and recomputed transcript expression patterns.

Validation of novel exons and junctions

A subset of novel exons and novel exon junctions found in *CACNA1C* by Nanopore sequencing (Supplementary Table 4) were confirmed by PCR targeting the novel sequence followed by Sanger sequencing.

In the case of novel exons, two sets of nested PCRs were conducted: one spanning the upstream exon and the novel exon sequence (‘5’ confirmation’ in Supplementary Table 4), and a second spanning the novel exon sequence and the downstream exon (‘3’ confirmation’ in Supplementary Table 4). Novel junctions were confirmed using a single round of PCR using one primer spanning the novel junction, with the second primer located in the neighboring exon. PCR reactions were performed using illustra PuReTaq Ready-To-Go PCR beads (GE Healthcare Life Sciences, UK). PCR reactions were cycled as follows: 95°C for 5 min; 35 cycles

of 95°C for 35 sec, 30 sec at annealing temperature (see Supplementary Table 4); 72°C for 2 min, followed by a final 5 min extension at 72°C. Primers sequences and conditions are shown in Supplementary Table 4. PCR products were separated on agarose gels. Products of the predicted size were excised, cleaned using the Qiagen Gel Extraction Kit and ligated into pGEM-T Easy vector (Promega) for Sanger sequencing.

Impact of RNA quality on amplification of full-length CACNA1C CDSs

RNA samples were artificially degraded, in order to identify the minimum RIN required for reliable amplification of the full-length CDS of *CACNA1C*. Eleven brain RNA samples with an average RIN of 8 were pooled and aliquoted into eight samples of 7 µl (875ng each at 125 ng/µl). Each aliquot was heated at 72ºC for different periods of time (0, 2, 5, 10, 20, 35, 60 or 90 minutes) to degrade the RNA. The RINe for each aliquot was then assayed using the Agilent 4200 TapeStation system. *CACNA1C* CDS amplification in these artificially-degraded samples was assessed alongside that from striatal RNA samples of varying RINs from three additional adult donors (Supplementary Table 5) to test whether results obtained in artificially-degraded RNA were similar to those from RNA that had not been subject to any purposeful degradation.

Reverse transcription used 500 ng of RNA and was performed using GoScript™ Reverse Transcriptase. *CACNA1C* amplification was performed as described above, using 5µl of cDNA and 30 cycles of amplification. Forward and Reverse primers were: CATTTCTTCCTCTTCGTGGCTGC and CCAGGTCACGAGAACAGTGAGG, respectively, i.e. identical to those used in the main experiment, but lacking the barcoding tails. Successful amplification of PCR product was confirmed and visualised using both agarose gel electrophoresis and 4200 Tapestation analysis using equal proportions of each PCR. Samples from three additional adult donors were utilised specifically for this experiment to provide the required range of high quality and partially degraded samples.

Results

Successful sequencing of full-length CACNA1C CDS

To investigate the expression and splicing patterns of *CACNA1C* in human brain we amplified the full length ~6.5kb CACNA1C CDS (encompassing the full intron chain of *CACNA1C* from the first to the last exon) by long range RT-PCR (Figure 1A). We selected 6 brain regions, (4 cortical regions, striatum and cerebellum), to facilitate detection of regional expression diversity, including the region with the most divergent global transcriptional profile (cerebellum)[32]. Highly pure *CACNA1C* was successfully amplified from all regions and in all samples; the 18 samples were then barcoded, pooled and sequenced across two Oxford Nanopore MinION flowcells until we had generated at least 1000 high quality sequence reads from each sample. Run 1 produced 112024 reads, including 52994 (47%) 2D pass reads with an identified barcode. Run 2 produced 126314 reads, including 80331 (64%) 2D pass reads with an identified barcode (Supplementary Table 3). Visualisation of pass reads demonstrated most were full length, with the updated flowcell used in Run2 providing higher quality reads with a lower error rate (Figure 1, Supplementary Figure 2). All analyses were performed using the 2D pass barcoded reads, hereafter referred to simply as “reads”.

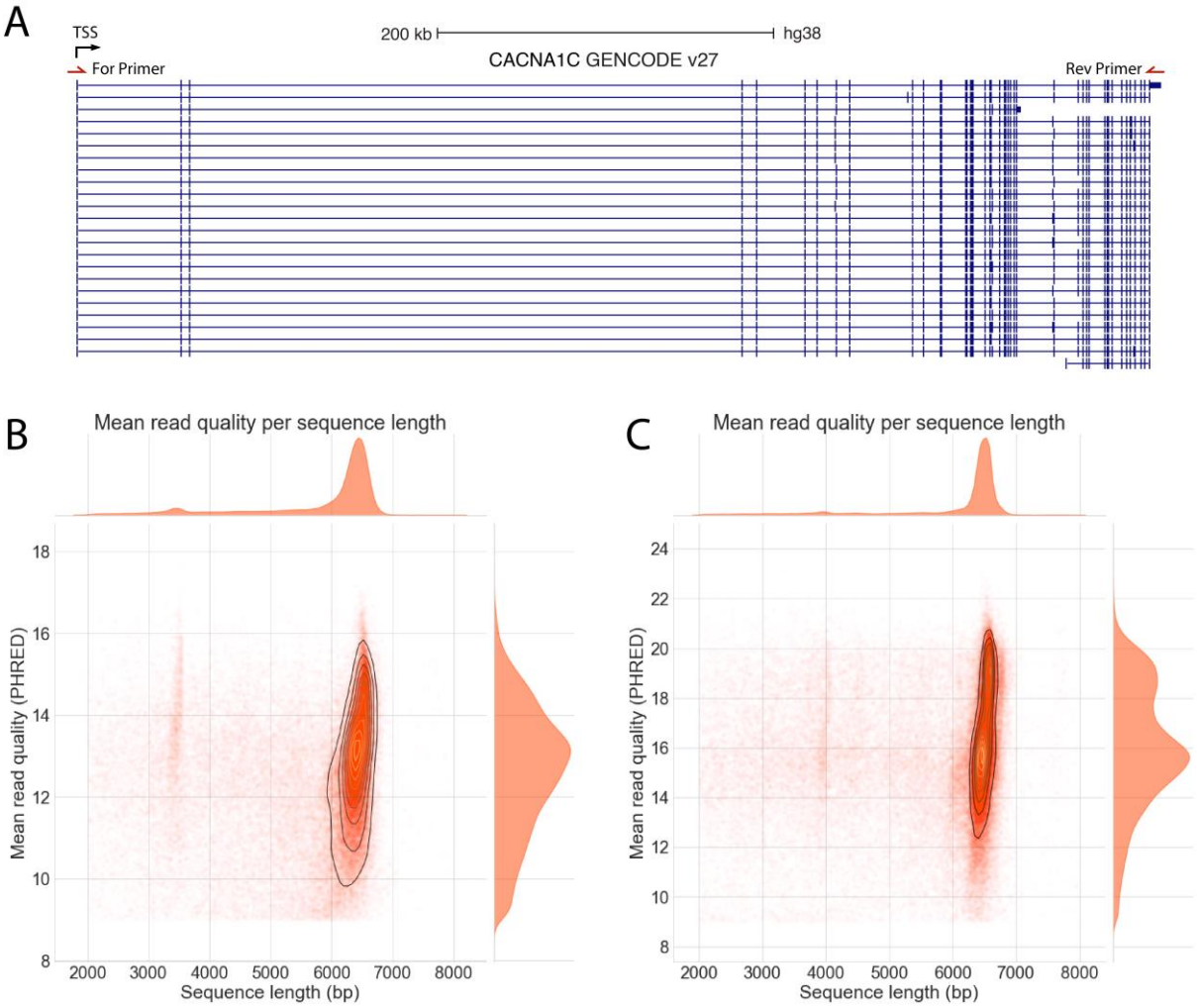

Figure1: Amplicon sequencing of *CACNA1C*. A) UCSC genome browser screenshot of *CACNA1C* isoforms annotated in GENCODE V27. All transcripts in "Basic" annotated set shown. Black arrow shows direction of transcription. TSS: Transcription start site. Position of Forward and Reverse long-range PCR primers in first exon and shared portion of final exon shown. B&C) Length vs Quality of all 2D pass reads from B) Run 1 C) Run 2. Most reads are the full-length *CACNA1C* CDS. The presence of the 3.5kb positive control CDS spike-in can be seen in Run1. Visualisation limited to reads between 2 and 8 kb, encompassing >98% of pass reads in each run.

Long-range amplicon sequencing reveals many novel CACNA1C exons and isoforms

We aimed to characterise *CACNA1C* isoforms and CDSs in adult human brain. We created a bespoke alignment and mapping pipeline to maximise the transcript information obtained from nanopore sequence reads, including the identification of novel exons and splice junctions (Supplementary Figure 3).

We initially mapped the nanopore reads to annotated *CACNA1C* transcripts, identifying events consistent with the presence of unannotated exonic sequences. Putative novel exons were

then mapped to the genome to identify their genomic location and confirm their existence. We annotated a total of 39 potential exons within the *CACNA1C* locus of which 38 were identified in at least 2 individuals or tissues and supported by at least 5 nanopore reads in each library and were included in the following analyses (Figure 2A). We selected four novel exons for validation by PCR + Sanger sequencing by confirming the splice junctions between the novel exon and its surrounding annotated exons. We confirmed 4 of 4 exons (and also discovered a fifth novel exon that was spliced in between the targeted novel exon and the nearest known exon). This successful validation of a selection of novel exons (which included some of the most and least abundant novel exons) provides high confidence that the novel exons identified by nanopore sequencing are real and actively incorporated into *CACNA1C* transcripts.

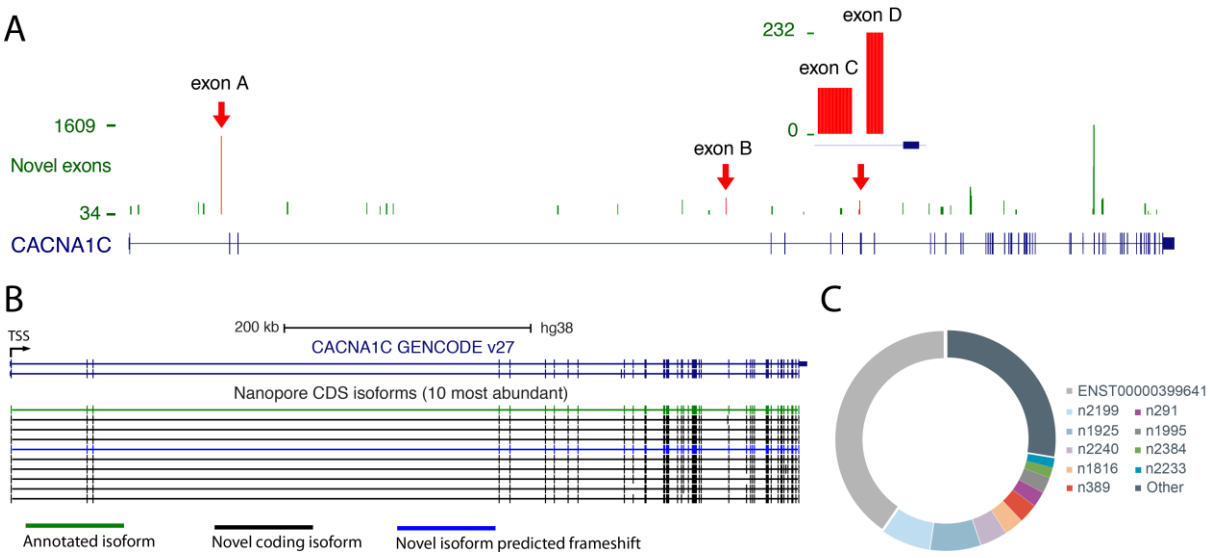

Figure 2. Novel exons and isoforms of *CACNA1C*. (**A**) Annotation and read counts (linear scale) for novel exonic sequences within *CACNA1C*. Red arrows indicate exons that have been validated (Supplementary Table 4). (**B**) Top 10 most abundant *CACNA1C* isoforms identified in brain. Colours denote transcript type. (**C**) Proportion of high confidence transcripts reads from the ten most abundant isoforms.

The integration of these 38 novel exons and the mapping of the reads to the concatenated exon sequence (metagene) enabled the identification of novel and known (annotated) isoforms. Novel isoforms included transcripts that incorporate novel exons, and/or novel junctions between annotated exons or new combinations of known junctions. We filtered the identified isoforms retaining only those found in at least 2 libraries with a minimum of 24 reads in total. We also created a high confidence set of isoforms with an increased minimum threshold of 100 reads. Unless otherwise stated all analyses were performed on the high confidence isoform set.

In total we identified 90 high confidence *CACNA1C* isoforms across the 6 brain regions, including 7 annotated isoforms (GENCODE v27) and 83 novel isoforms (Supplementary Figure 4). Seven of the novel high confidence isoforms contained novel exons (1.8% of high confidence reads in total), while the remaining 76 included previously undescribed junctions and junction combinations. The large number of transcript isoforms identified, most of which

were previously unannotated, confirms that *CACNA1C* has an extremely complex transcriptional profile and that the existing annotations of *CACNA1C* were far from complete.

We next considered whether previously annotated transcripts might be the most abundant, even if they were few in number. As expected the most highly expressed isoform is a previously annotated transcript (ENST00000399641). On average 40.2% (23.7%-50%) of reads supported this transcript. In comparison the most highly expressed novel transcript (*CACNA1C n2199*) represents on average 7.3% (2.7%-25.3%) of all reads (Figure 2B,C). However, nine of the top ten expressed isoforms were novel transcripts and eight of these are predicted to maintain the *CACNA1C* reading frame, suggesting a number of the most abundant novel isoforms could encode functionally distinct proteins (Figure 2B,C). Novel isoforms made up 56% of high confidence reads and when all reads without filtering were included 75.5% of total reads supported previously unannotated transcripts. Taken together these results suggest that novel *CACNA1C* isoforms are abundantly expressed as well as highly numerous and that current annotations are missing many of the most abundant *CACNA1C* isoforms.

Including all reads without any filtering we found evidence for the expression of 18 of the 40 annotated isoforms (GENCODE v27) in human brain. The fact that we find only 7 annotated isoforms in our high confidence set (out of 90) and that less than half show any reads supporting the transcript model, raises the possibility that a number of annotated CACNA1C isoforms may not exist and instead represent annotation errors.

Predicted impact of novel isoforms on CACNA1C *protein sequence*

CACNA1C encodes the pore-forming $Ca_V1.2$ alpha VGCC subunit. The calcium pore is made up of 24 transmembrane repeats that are clustered into 4 domains, linked by intracellular loops (Figure 3A). Disruptions to these transmembrane sequences are likely to result in a non-functional protein; therefore, we examined the predicted protein coding sequences of the novel isoforms to assess how many encode putatively functional channels. Among the 83 novel isoforms we identified, 51 are potentially coding, whilst the remainder (arising from 32 isoforms) are likely to lead to noncoding transcripts as they contain frameshifts or deletions in critical membrane-spanning regions. Notably, putatively coding isoforms represent 87.8% of total high confidence reads, demonstrating that while noncoding transcripts are quite numerous, it is coding transcripts that represent the vast majority of reads (Figure 3B). We next investigated how each of the novel, potentially coding, isoforms alter the protein sequence.

Around half of the potentially coding isoforms (25 of 51; representing 26.8% of total coding reads) consist of novel combinations of already annotated exons (including those featuring novel junctions). The remainder consists of novel isoforms containing exons and / or deletions that are not currently annotated (Figure 3B).

Many of the novel splicing events are seen across multiple novel isoforms. For example, five isoforms (representing 2.8%, on average, and 5.2%, maximally, of total coding reads) predict $Ca_V1.2$ proteins with an in-frame deletion in the intracellular linker region between domains I and II, and seven (representing 6.9%, on average, and 11.9%, maximally, of total coding reads) contain an in-frame deletion in the S4-S5 linker in domain IV (Figure 3A). Ten isoforms (representing 12.7%, on average, and 22.3%, maximally, of total coding reads) include a

microdeletion in the extracellular loop between transmembrane regions S3 and S4 in domain IV (Figure 3A), that has previously been identified using exon-to-exon PCR[33], demonstrating the utility of nanopore sequencing to reliably identify microdeletions, given sufficient reads. This group includes *CACNA1C n1925*, one of the three most abundant isoforms. Intriguingly *CACNA1C n2199*, the novel isoform with the highest maximum abundance, which consists of a novel combination of known exons, also varies in this loop from the canonical transcript (Figure 3C). An additional isoform also contains a novel exon that introduces a small, in-frame insertion in this same linker region, further highlighting this region as a 'hotspot' for alternative splicing. This finding is intriguing given previous reports of effects of splicing in this region on channel properties[33].

A number of isoforms predict alterations in the intracellular N- and C-termini, both of which are involved in coupling $Ca_V1.2$ signalling to intracellular signalling cascades[19]. At the N terminus, two isoforms (representing 0.7%, on average, and 2.1%, maximally, of total coding reads) include a novel, in-frame exon that predicts an extension, whilst two isoforms (representing 0.5%, on average, and 0.9%, maximally, of total coding reads) predict a truncated N terminus. Similarly, two isoforms (cumulatively representing 0.8%, on average, and 1.5%, maximally, of total coding reads) predict proteins with different deletions in the intracellular C terminal tail (Figure 3A)

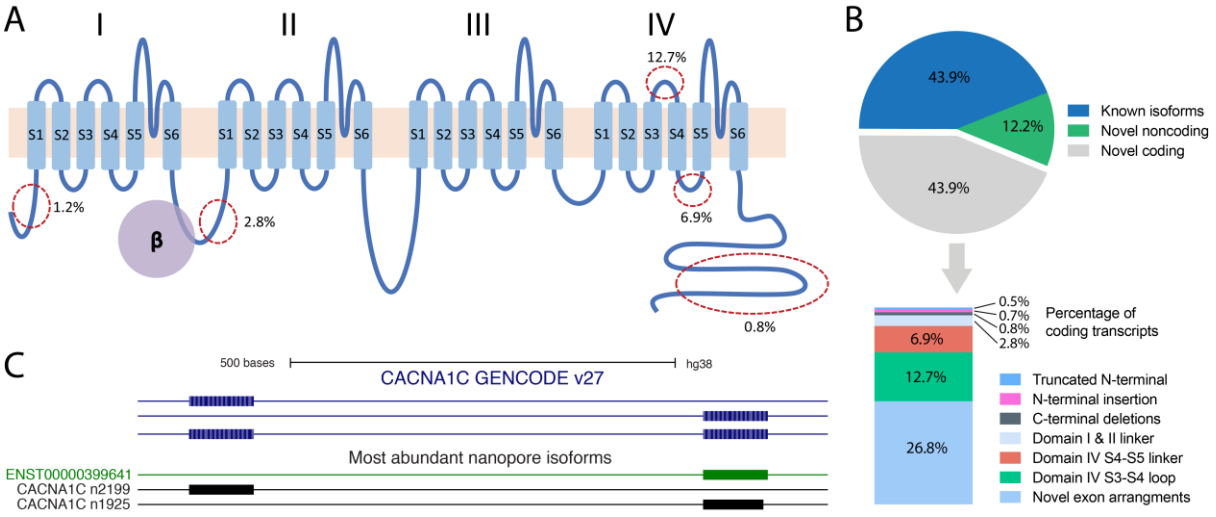

Figure 3. Impact of novel splicing isoforms on the CACNA1C protein model. (**A**) CACNA1C protein domain structure. *CACNA1C* encodes the primary pore-forming subunit of the $Ca_V1.2$ channel. $Ca_V1.2$ is formed of 4 domains (I-IV), each comprised of 6 transmembrane domains (S1-S6), which are linked by intracellular loops. The obligate beta subunit binds to the I-II intracellular loop, as shown. Red circles indicate the location of novel, in-frame insertions and deletions, discussed in the main text. Values indicate the mean proportion of coding reads containing each variant. (**B**) Percentage of reads from different isoform classes. Top: Percentage of reads from known (GENCODE v27) and novel *CACNA1C* isoforms. All known isoforms were protein coding. Below: Percentage of total coding reads in various novel coding isoform classes. (**C**) Variation of the three most highly expressed *CACNA1C* isoforms in the extracellular loop between transmembrane regions S3 and S4 in domain IV.

Differences in the expression profile of CACNA1C isoforms between brain regions

Next we examined how *CACNA1C* isoform expression varied across brain regions and between individuals. Due to the large variation in the number of reads sequenced from each of the libraries, we downsampled all libraries to 2,729 prior to normalization. Differences between tissues appears to be the main driver of the observed variation in expression between isoforms. Cerebellum and striatum were distinct from the other tissues (Figure 4A,B). In comparison we observed no major effect associated with the individual of origin. Similar observations were previously made at the transcriptome level when comparing expression across tissues and individuals[32]. The use of the more permissive filtered set of isoforms to estimate transcript expression (see Methods) further improved the separation of the tissues and also highlighted potential differences in expression between cortical regions (Supplementary Figure 5,6).

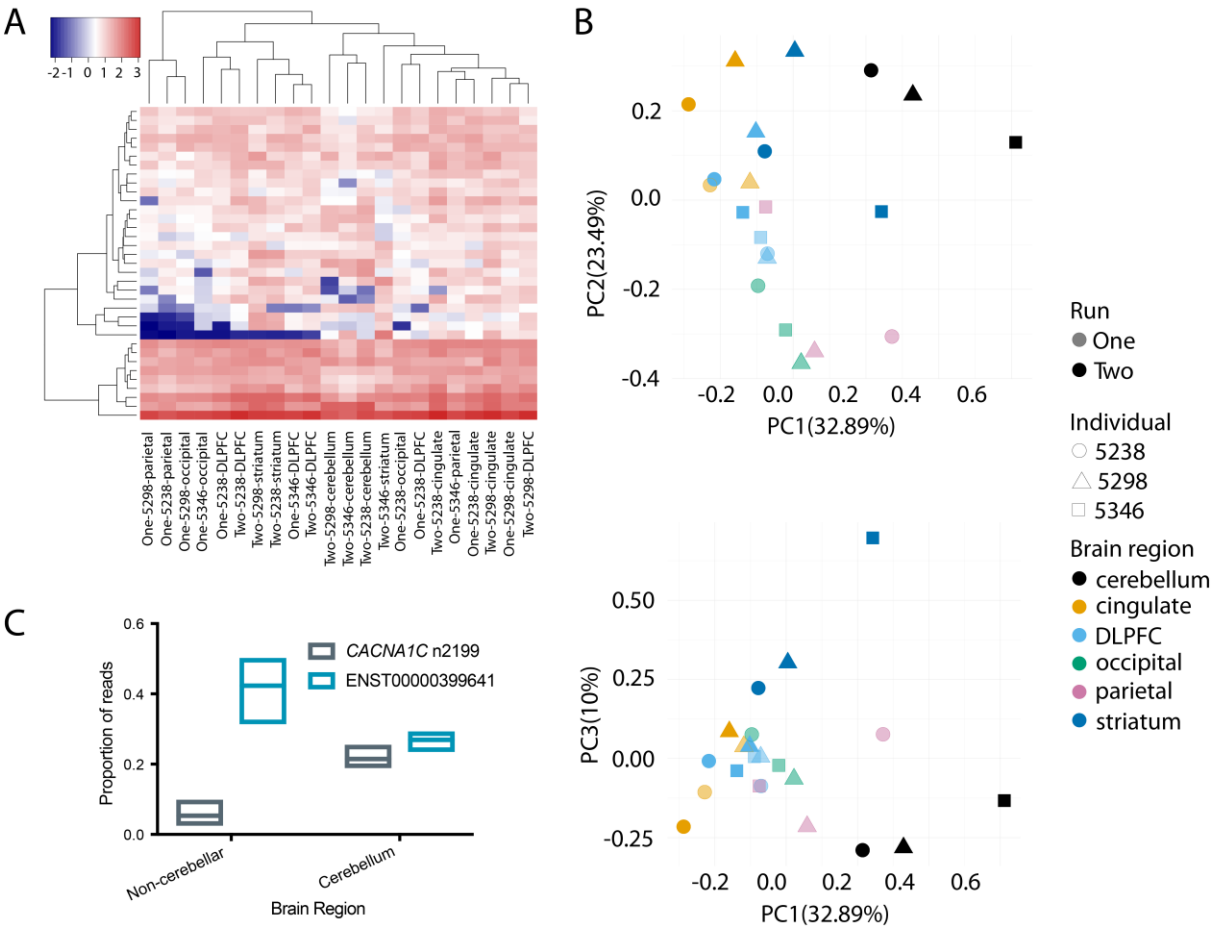

Figure 4. Comparison of *CACNA1C* isoform expression between individuals and tissues (**A**) Transcript expression levels (TPM) across tissues and individuals. “One” and “Two” denote sequencing runs. (**B**) Principal Component Analysis based on normalised transcript expression. (**C**) Isoform switching of *ENST00000399641* and *CACNA1C n2199* in cerebellum. Box plots show minimum to maximum values with line at mean value.

We next investigated isoform variation between different tissues. Although we did not find any tissue-specific isoforms amongst our set of high confidence isoforms, we did identity a pronounced isoform expression switch present solely in cerebellum. Outside of the cerebellum *ENST00000399641* made up 41.6% of high confidence isoform reads, while *CACNA1C n2199* accounted for 5.6%. Within cerebellum the proportion of expression from *CACNA1C n2199* increased dramatically to almost the same level as *ENST00000399641* at 21.5% and 24.2% of reads, respectively (Figure 4C). Cerebellum therefore appeared to be the only brain region studied with two dominant *CACNA1C* isoforms. Whether this represents two dominant isoforms expressed in the same cells or different dominant isoforms in different cerebellar cell types is unknown. Outside of cerebellum, the most highly expressed novel isoform was *CACNA1C n1925*, however this accounted for a similar proportion of *CACNA1C* reads (7.5%) in each tissue.

High quality RNA is required to reliably amplify full-length CACNA1C CDSs

The quality of RNA from post-mortem or tissues samples can be highly variable with many samples having undergone significant degradation. Conversely, RT-PCR and sequencing of long and/or full-length cDNAs requires the sample RNA to be of sufficient integrity to contain undegraded transcripts. To investigate how RNA quality impacts the feasibility of amplifying and sequencing full-length genes, and to establish a minimum recommended quality value, we artificially degraded brain RNA to create a series of RINe values and investigated the effect of RNA quality on full-length CACNA1C CDS amplification. RNA quality varied from RINe 8.2 (undegraded RNA) to 2.8 (90 min at 72ºC) (Figure 5A). *CACNA1C* amplified strongly in samples with a RINe of ~8, with lower quality samples showing decreasing amounts of *CACNA1C* and no product from samples with a RINe of ~<5 (Figure 5C). As RNA degradation from heat treatment may not accurately reflect standard sample degradation, we attempted *CACNA1C* amplification in three untreated striatal samples with RINe of 7.3, 6.5 and 5.8 with similar results (Figure 5B,C). These results suggest a minimum RINe value of 6 for generating long amplicons from post-mortem brain samples and a recommended RINe of >7 for robust amplification.

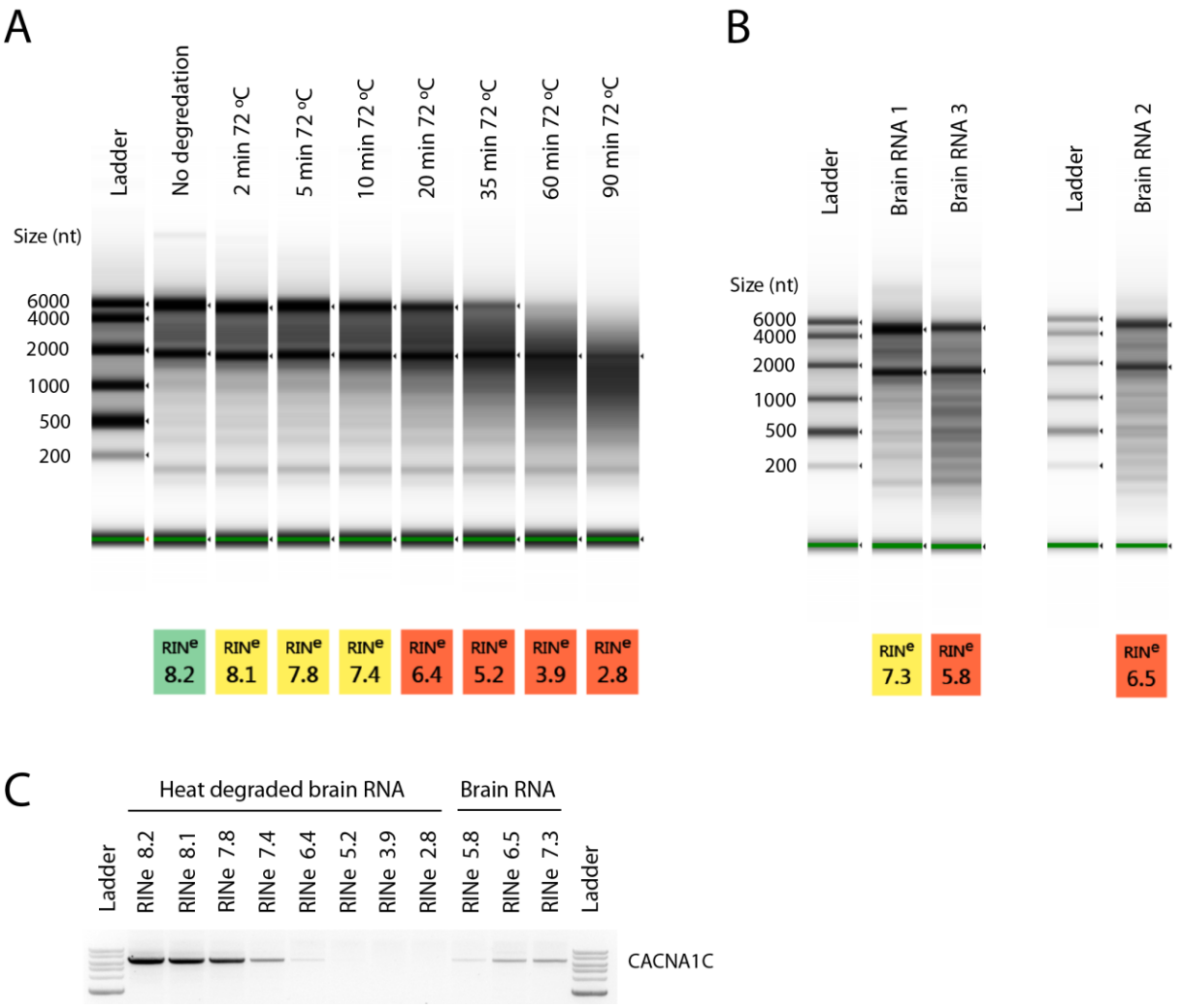

Figure 5: High-quality RNA is required for amplification of *CACNA1C* CDS. (**A**) Tapestation profile and RINe quality values of brain RNA after heat treatment at 72ºC. Untreated RNA had a RINe of 8.2. (**B**) Tapestation profiles and RINe quality values of untreated striatal brain RNA samples. Only lanes of interest from Tapestation profiles are shown. (**C**) Agarose gel showing amplification of the ~6.5kb *CACNA1C* CDS from RNA of varying quality.

Discussion

We combined long-range RT-PCR with nanopore sequencing in order to characterise the full coding sequence from isoforms of *CACNA1C*. To our knowledge, we are the first to have used this approach successfully in human postmortem tissue. The vast majority of the *CACNA1C* isoforms that we describe here are novel, and many are abundant. We demonstrate marked differences in *CACNA1C* isoform profiles between brain regions, with the cerebellum in particular showing a notable switch in isoform abundance compared with cortex. Our results demonstrate that *CACNA1C* isoforms are much more diverse than previously appreciated, and emphasise the importance of studying full-length isoforms and of access to high-quality human brain tissue. Additionally, these results are a necessary first step towards understanding *CACNA1C* gene expression and how it can be altered by the presence of a disease risk variant. We anticipate that long-read sequencing methodologies will be broadly useful in

deciphering isoform expression and splicing, especially for genes expressed in the brain, which exhibits the most prominent use of alternative splicing and tissue specific exons[9-13].

Despite annotations of very high quality and active curation[14,34], an increasing number of studies have reported novel protein coding sequences and exons in the human genome. These observations stem from the wealth of transcriptomic data generated across the human lifespan, and between tissues and individuals, allowing researchers to capture developmental stage-, tissue- and condition-specific events with unprecedented power. The rapid development of long read sequencing technologies opens the unique opportunity to gain an accurate representation of transcript diversity, as each read encompasses a full transcript. This knowledge is particularly critical for genes with complex models. This long-read sequencing approach was previously applied with success to DSCAM, identifying 7874 different transcript isoforms with a previous version of the nanopore sequencing technology[21].

Our study highlights the power of long read sequencing for the annotation and characterisation of alternatively spliced transcripts. We more than double the number of annotated transcripts for *CACNA1C*, identifying novel exons within the coding sequence as well as novel combinations of previously annotated exons. Our findings support those from transcriptome analyses that indicate that a significant proportion of gene isoforms in human brain remain unannotated[35]. Supporting their potential importance in CACNA1C's function, a number of the novel isoforms are highly abundant individually, and collectively they encompass the majority of reads. Our finding that the vast majority of reads from novel isoforms maintain the *CACNA1C* reading frame also supports the hypothesis that these are functionally relevant transcripts and not simply products of noisy splicing. Furthermore, we find that many in-frame deletions are present in a number of isoforms, cumulatively these can also be abundant and if translated could have a quite dramatic impact on the CACNA1C protein function in the cell. Notably, a number of our novel transcripts predict proteins with alterations in domains known to be important for determining channel properties and coupling to second messenger systems[20,33,36], thereby providing testable hypotheses as to their predicted functional impact. Now that the isoform structure of CACNA1C is clear, it will also be of interest to examine the *cumulative* effect of functional variation across the $Ca_V1.2$ protein on channel function. Conversely, we identified a relatively low number of annotated isoforms in our dataset. This is perhaps unsurprising, given that many of the currently annotated isoforms are likely predictions from ESTs and incomplete cDNAs. Taken together, these observations demonstrate the importance of long read sequencing for the accurate characterisation of transcript structure and alternative splicing.

Our results show that there is greater variation in *CACNA1C* isoform abundance between brain regions than between individuals. In most regions there is a single major *CACNA1C* isoform, with levels of expression almost five-fold higher than the second most highly expressed isoform. However, in the cerebellum there is a pronounced switch in isoform abundance (seen in all individuals) such that the two most abundant transcripts are expressed at similar levels to one another. The consistent nature of this switch in different individuals supports the hypothesis that this is a regulated switch in expression. Around 20% of genes are known to express multiple major isoforms simultaneously, often within the same cells, though is it currently unknown if this is the case for *CACNA1C* in cerebellum[9].

A key consideration in the use of long-read sequencing to investigate gene expression is the integrity of the starting RNA. Our results suggest that a RIN of 7 or above is required for

efficient and robust reverse transcription and amplification of a 6.5kb region of mRNA, although it is possible that shorter amplicons may be successfully sequenced at lower RINs. The average RIN value for samples in many tissue banks is often below 7; our results emphasise the improvement in the utility of samples for gene expression profiling when high RIN samples are available and the importance of high quality samples to facilitate gene expression studies.

In summary, our findings demonstrate the utility of long-range amplicon sequencing for the identification and characterisation of gene isoform profiles. More specifically, they demonstrate that the human brain *CACNA1C* isoform profile is substantially more complex than currently appreciated. Understanding the functional consequences of this isoform diversity will advance our understanding of the role of VGCCs in the human brain. This information is also of clinical relevance, given robust associations between *CACNA1C* and a number of psychiatric illnesses and related phenotypes: novel *CACNA1C* isoforms may mediate genetic associations between the *CACNA1C* locus and psychiatric phenotypes[2,3,5], and may prove novel, and more selective therapeutic targets for these disorders[37].

Acknowledgements

We are grateful to Li Chen and Arne Mould for technical assistance. The authors wish to acknowledge the following funding sources: MBC is supported by an Australian National Health and Medical Research Council (NHMRC) Early Career Fellowship [APP1072662]. EMT is supported by a Royal Society University Research Fellowship. WH and TW are supported by the BBSRC, Institute Strategic Programme Grant [BB/J004669/1], BBSRC Core Strategic Programme Grant [BB/P016774/1]. This research was supported by a Wellcome Trust [201879/Z/16/Z] award to MBC and a UK Medical Research Council [MR/P026028/1] award to EMT. This study was supported by the National Institute for Health Research Oxford Health Biomedical Research Centre. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

The contents of the published material are solely the responsibility of the administering institution, a participating institution, or individual authors and do not reflect the views of NHMRC.

Author contributions

MBC and EMT conceived the study. MBC, AGB, TW, WH and EMT designed experiments. MBC, AGB and EMT performed experiments and nanopore sequencing. AGB performed PCR validations. JEK, TH and DRW provided materials. TW and WH wrote the analysis pipeline and performed informatic analyses. MBC, TW, AGB, TW, WH and EMT wrote the manuscript.

References

1 Corvin, A. & Sullivan, P. F. What Next in Schizophrenia Genetics for the Psychiatric Genomics Consortium? *Schizophrenia Bulletin* **42**, 538-541, doi:10.1093/schbul/sbw014 (2016).

2 Green, E. K. *et al.* Replication of bipolar disorder susceptibility alleles and identification of two novel genome-wide significant associations in a new bipolar disorder case-control sample. *Molecular Psychiatry* **18**, 1302-1307, doi:10.1038/mp.2012.142 (2013).

3 Bipolar Disorder Working Group of the Psychiatric Genomics Consortium. Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nature Genetics* **43**, 977-983, doi:10.1038/ng.943 (2011).

4 Cross-Disorders Working Group of the Psychiatric Genomics Consortium. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet* **381**, 1371-1379, doi:10.1016/S0140-6736(12)62129-1 (2013).

5 Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421-427, doi:10.1038/nature13595 (2014).

6 Li, Y. I. *et al.* RNA splicing is a primary link between genetic variation and disease. *Science* **352**, 600-604, doi:10.1126/science.aad9417 (2016).

7 Li, M. *et al.* A human-specific AS3MT isoform and BORCS7 are molecular risk factors in the 10q24.32 schizophrenia-associated locus. *Nature Medicine* **22**, 649-656, doi:10.1038/nm.4096 (2016).

8 Mazin, P. *et al.* Widespread splicing changes in human brain development and aging. *Molecular Systems Biology* **9**, 633, doi:10.1038/msb.2012.67 (2013).

9 Tapial, J. *et al.* An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms. *Genome Research* **27**, 1759-1768, doi:10.1101/gr.220962.117 (2017).

10 Mele, M. *et al.* Human genomics. The human transcriptome across tissues and individuals. *Science* **348**, 660-665, doi:10.1126/science.aaa0355 (2015).

11 Jensen, K. B. *et al.* Nova-1 regulates neuron-specific alternative splicing and is essential for neuronal viability. *Neuron* **25**, 359-371 (2000).

12 Yang, Y. Y., Yin, G. L. & Darnell, R. B. The neuronal RNA-binding protein Nova-2 is implicated as the autoantigen targeted in POMA patients with dementia. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 13254-13259 (1998).

13 Raj, B. & Blencowe, B. J. Alternative Splicing in the Mammalian Nervous System: Recent Insights into Mechanisms and Functional Roles. *Neuron* **87**, 14-27, doi:10.1016/j.neuron.2015.05.004 (2015).

14 Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Research* **22**, 1760-1774, doi:10.1101/gr.135350.111 (2012).

15 Li, Y. I., Sanchez-Pulido, L., Haerty, W. & Ponting, C. P. RBFOX and PTBP1 proteins regulate the alternative splicing of micro-exons in human brain transcripts. *Genome Research* **25**, 1-13, doi:10.1101/gr.181990.114 (2015).

16 Irimia, M. *et al.* A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell* **159**, 1511-1523, doi:10.1016/j.cell.2014.11.035 (2014).

17 Steijger, T. *et al.* Assessment of transcript reconstruction methods for RNA-seq. *Nature Methods* **10**, 1177-1184, doi:10.1038/nmeth.2714 (2013).

18 Hardwick, S. A. *et al.* Spliced synthetic genes as internal controls in RNA sequencing experiments. *Nature Methods* **13**, 792-798, doi:10.1038/nmeth.3958 (2016).

19 Hofmann, F., Flockerzi, V., Kahl, S. & Wegener, J. W. L-type CaV1.2 calcium channels: from in vitro findings to in vivo function. *Physiological Reviews* **94**, 303-326, doi:10.1152/physrev.00016.2013 (2014).

20 Striessnig, J., Pinggera, A., Kaur, G., Bock, G. & Tuluc, P. L-type Ca(2+) channels in heart and brain. *Wiley Interdisciplinary Reviews of Membrane Transport and Signaling* **3**, 15-38, doi:10.1002/wmts.102 (2014).

21 Bolisetty, M. T., Rajadinakaran, G. & Graveley, B. R. Determining exon connectivity in complex mRNAs by nanopore sequencing. *Genome Biology* **16**, 204, doi:10.1186/s13059-015-0777-z (2015).

22 Weirather, J. L. *et al.* Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research* **6**, 100, doi:10.12688/f1000research.10571.2 (2017).

23 Byrne, A. *et al.* Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nature Communications* **8**, 16027, doi:10.1038/ncomms16027 (2017).

24 Loman, N. J. & Quinlan, A. R. Poretools: a toolkit for analyzing nanopore sequence data. *Bioinformatics* **30**, 3399-3401, doi:10.1093/bioinformatics/btu555 (2014).

25 Leger, A. a-slide/pycoQC: v1.1.alpha2. . doi:doi:10.5281/zenodo.1116400 (2017).

26 Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859-1875, doi:10.1093/bioinformatics/bti310 (2005).

27 Kielbasa, S. M., Wan, R., Sato, K., Horton, P. & Frith, M. C. Adaptive seeds tame genomic sequence comparison. *Genome Research* **21**, 487-493, doi:10.1101/gr.113985.110 (2011).

28 Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology* **11**, R25, doi:10.1186/gb-2010-11-3-r25 (2010).

29 R Development CoreTeam. *R: A language and environment for statistical computing.*, (2008). Vienna, Austria : the R Foundation for Statistical Computing. ISBN: 3-900051-07-0.

30 Zhao, S., Guo, Y., Sheng, Q. & Shyr, Y. Heatmap3: an improved heatmap package with more powerful and convenient features. *BMC Bioinformatics* **15**, P16, doi:10.1186/1471-2105-15-s10-p16 (2014).

31 Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag, 2009).

32 Consortium, G. T. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648-660, doi:10.1126/science.1262110 (2015).

33 Tang, Z. Z. *et al.* Transcript scanning reveals novel and extensive splice variations in human l-type voltage-gated calcium channel, Cav1.2 alpha1 subunit. *The Journal of Biological Chemistry* **279**, 44335-44343, doi:10.1074/jbc.M407023200 (2004).

34 Wright, J. C. *et al.* Improving GENCODE reference gene annotation using a high-stringency proteogenomics workflow. *Nature Communications* **7**, 11778, doi:10.1038/ncomms11778 (2016).

35 Tilgner, H. *et al.* Comprehensive transcriptome analysis using synthetic long-read sequencing reveals molecular co-association of distant splicing events. *Nature Biotechnology* **33**, 736-742, doi:10.1038/nbt.3242 (2015).

36 Kaur, G. *et al.* A Polybasic Plasma Membrane Binding Motif in the I-II Linker Stabilizes Voltage-gated CaV1.2 Calcium Channel Function. *Journal of Biological Chemistry* **290**, 21086-21100, doi:10.1074/jbc.M115.645671 (2015).

37 Cipriani, A. *et al.* A systematic review of calcium channel antagonists in bipolar disorder and some considerations for their future development. *Molecular Psychiatry* **21**, 1324-1332, doi:10.1038/mp.2016.86 (2016).