High transcriptional complexity of the retinitis pigmentosa *CERKL* gene in human and mouse

Alejandro Garanto*[1,2,3], Marina Riera*[1,2,3], Esther Pomares[1,2,3], Jon Permanyer[1,2], Marta de Castro-Miró[1], Josep F Abril[1,2], Gemma Marfany[1,2,3] and Roser Gonzàlez-Duarte[1,2,3]

From the [1]Departament de Genètica, Facultat de Biologia, [2]Institut de Biomedicina (IBUB), Universitat de Barcelona, Barcelona, Spain; [3]CIBERER, Instituto de Salud Carlos III, Barcelona, Spain.

* These two authors contributed equally to this work

Running title: *CERKL* transcriptional complexity in retina

Word count: 7,380 without references nor the title and affiliation data

Corresponding author: Roser Gonzàlez-Duarte, Departament de Genètica, Facultat de Biologia, Universitat de Barcelona, Avda.Diagonal 645, 08028 Barcelona, Spain;

Tel: +34 934021034

FAX: +34 934034420

rgonzalez@ub.edu

Grant Information: This study was supported by grants SAF2009-08079 (Ministerio de Ciencia e Innovación) and SGR2009-1427 (Generalitat de Catalunya), CIBERER (U718), Fundaluce and ONCE to R.G.-D and BFU2010-15656 to G.M.

ABSTRACT

Purpose: In order to shed light on the pathogenicity of the mutations in the retinitis pigmentosa gene *CERKL,* we aimed to characterize its transcriptional repertoire, and focused on the use of distinct promoters and alternative splicing in human and mouse tissues.

Methods: *In silico* genomic and transcriptomic computational customized analysis, combined with experimental RT-PCRs on different human and murine tissues and cell lines and immunohistochemistry, have been used to characterize the transcriptional spectrum of *CERKL*. In the mouse retina, Cerkl is mainly detected in ganglion cells and cones, but can also be observed in rods. Cerkl is mainly cytosolic: it localizes in the outer segment of photoreceptors, and also in the perinuclear region of some cells.

Results: An unexpected multiplicity of *CERKL* transcriptional start sites, 4 in each species, plus a high variety of alternative splicing events -mainly affecting the 5' half of the gene- generate more than 20 fully validated mRNA isoforms in human and 23 in mouse. Moreover, several translational start sites, compatible with a wide display of functional domains, contribute to the final protein complexity.

Conclusions: This combined approach of *in silico* and experimental characterization of the *CERKL* gene provides a comprehensive picture of the species-specific transcriptional products in the retina, underscores a highly tuned gene regulation in different tissues, and establishes a framework for the study of *CERKL* genotype-phenotype correlations.

INTRODUCTION

Spatio-temporal differential splicing, often related to developmental events or tissue differentiation processes, affects more than 95% of the human genes, as recently unveiled after massive sequencing of the human transcriptome[1, 2]. Alternative splicing, as well as the use of alternative promoters and transcriptional splice sites, are instrumental for the generation of complexity, as proteins with different functions are encoded by the transcript variants produced. Cells can thus deploy a wide array of proteins, all arising from a single genomic sequence[3, 4].

Misregulation of alternative splicing is often at the basis of human disease, given that distortions in the splicing process either directly alter the domains displayed by proteins or –more relevant to pathology– cause frameshifts, which are frequently associated to premature stop codons[5]. Therefore, prior knowledge of all the physiologically produced transcripts from a gene of interest is crucial to draw genotype-phenotype correlations in hereditary diseases, and to infer the degree of pathological severity[6-8]. This is even more relevant when considering genetic disorders of the mammalian central nervous system (CNS) and derived neurological tissues –such as the retina–, where the highest degree of alternative splicing events occurs[9-11].

Retinitis pigmentosa (RP) is a hereditary neurodegenerative disorder with extremely high genetic heterogeneity. It affects 1:4000 people worldwide and it is the major cause of non-traumatic adult blindness[12]. Although more than 45 genes have been identified as causative of RP (Retnet, http://www.sph.uth.tmc.edu/Retnet/), around 40% of the genetic cases remain unassigned, highlighting the relevance of identifying new candidates as each gene will presumably explain very few cases. The molecular diagnosis scenario becomes even more complex under the light of recent reports that reveal new mutations in known RP genes, which alter retina-specific splicing events, either by changing the number of exons included in the mature product or by modifying the relative proportion of the spliced isoforms[13-15]. These findings widen the range of molecular mechanisms underlying tissue-restricted

pathologies, decrease the number of unknown RP genes, illuminate new scenarios for tissue-specific gene function and emphasize the need for an accurate characterization of candidate splicing products, particularly since 70% of the exons in the human genome are tissue-specific[1, 16].

Our group first identified *CERKL* as an RP gene[17] by detecting a homozygous nonsense mutation (R257X) that cosegregated in consanguineous Spanish families. *CERKL* was widely expressed and the highest transcription levels were observed in the retina[17, 18]. Interestingly, the R257X mutation was embedded in an alternatively spliced exon; therefore in the patients some of the *CERKL* isoforms were *a priori* functional[19]. These results prompted us to undertake a more accurate characterization of the *CERKL* transcripts in human and mouse. Our work unveils an unexpected high complexity of the *CERKL* transcripts, particularly at the 5‘ end of the gene with: alternative first exons, inclusion/exclusion of alternatively spliced exons, intron retention and the use of additional splice sites. Overall, these results together with the bioinformatic analysis strongly support: i) the generation of many protein isoforms, ii) different roles of CERKL in retinal cells and other tissues, and iii) provide a molecular framework for genotype-phenotype correlations, as the location of the mutation in the *CERKL* gene would affect the number and type of transcripts and so, relate to the progression and severity of the disease.

MATERIAL AND METHODS

Animal handling, tissue dissection and preparation of samples

Murine tissue samples were obtained from C57BL/6J mice (Charles River Laboratories, Davis, CA). All procedures were performed according to ARVO statement for the use of animals in ophthalmic and vision research, as well as the regulations of the Animal Care facilities at the University of Barcelona. Animals were euthanized with CO_2 followed by cervical dislocation. Specific tissues and organs were dissected and immediately frozen in liquid nitrogen.

For human tissues, blood and saliva samples were collected from non-affected RP people, after informed consent and following the tenets of the Helsinki

declaration. Retina and Brain total RNA samples were supplied by Clontech Laboratories, Inc. (Mountain View, CA) and liver cDNA by BD Biosciences (San Jose, CA). Cultured cells from human origin were: Human Embryonic Kidney cells 293T (HEK293T, Bethyl Laboratories, Montgomery, TX), wild-type fibroblasts (kindly provided by D. Grinberg and L. Vilageliu), and human lung adenocarcinoma epithelial cell line (A549, Abcam, Cambridge, MA). HEK293T and fibroblasts were grown in DMEM with 4mM L-glutamine, whereas A549 was cultured in Ham's F12 L-glutamine (PAA Laboratories GmbH, Pasching, Austria). Both media were supplemented with 10% fetal bovine serum (FBS), 100 U/ml of penicillin and 100 µg/ml streptomycin (Invitrogen Life Technologies, Carlsbad, CA).

RNA extraction

Twenty-five mg of each frozen mouse tissue were homogenized using a Polytron PT 1200 E homogenizer (Kinematica AG, Lucerne, Switzerland). For total RNA extraction, High Pure RNA Tissue Kit (Roche Diagnostics, Indianapolis, IN) was used, following the manufacturer's instructions. Human and mouse blood RNA was extracted with the RiboPure-Blood Kit (Ambion/Applied Biosystems, Foster City, CA). Before proceeding, 500 µl of blood were mixed with 1.3 ml of RNA*later* (Ambion/Applied Biosystems, Foster City, CA). Saliva samples were treated as indicated in the Oragene·RNA protocol (DNA Genotek Inc., Ontario, Canada). RNA from 10^6 human cultured cells was extracted using the RNeasy kit (QIAGEN, Germantown, MD). Total RNA was quantified using the nanoquant plate in an Infinite 200 microplate reader (Tecan, Männedorf, Switzerland).

RT-PCR

Two different RT-PCRs assays were performed for human and mouse samples. The first assay was carried out with the Mint Kit (Evrogen, Moscow, Russia) according to the manufacturer's instructions and using between 300 to 750 ng of total retina RNA per reaction. A second RT-PCR with the Transcriptor High

Fidelity cDNA Synthesis Kit (Roche Diagnostics, Indianapolis, IN) was also performed following the manufacturer's protocol, using either 200 ng of mouse total RNA or 750 ng of human RNA per reaction.

For tissue expression analysis all reaction mixtures (50 µl) contained 10 µM of each primer pair, 2 µM of dNTPs, 1.5 mM $MgCl_2$ and 1 U of Go*Taq* polymerase (Promega, Madison, WI). The same pair of primers was used to amplify human and mouse *GAPDH* (Supplementary Table S1) to compare and normalize the samples. PCR conditions were as follows: 120 s at 94ºC and 30 cycles of 94ºC for 20 s and 63ºC for 120 s. Human *CERKL* reaction consisted in a three-step PCR (120 s at 94ºC followed by 40 cycles of 94ºC for 20 s, 60ºC for 30 s and 72ºC for 30 s) using primers A and B (Fig. 1 A2 and Supplementary Table S1). Mouse *Cerkl* expression was detected with primers a and b (Fig. 1 B2 and Supplementary Table S1) (120 s at 94ºC followed by 35 cycles of 94ºC for 20 s, 60ºC for 30 s and 72ºC for 20 s).

The analysis of the 5' and 3' UTRs of human and murine *CERKL* retina isoforms was performed on the cDNAs generated by the Mint Kit (Evrogen, Moscow, Russia), using either the Plug adaptor or oligo-d(T) primers (provided by the manufacturer) paired with suitable *CERKL*-specific internal primers (Fig. 1 A2 and 1 B2 and Supplementary Table S1) under the indicated PCR conditions (120 s at 94ºC, followed by 40 cycles of 94ºC for 20 s, 60ºC for 30 s and 72ºC for a variable time, depending on the amplicon size). The characterization of alternatively spliced variants was performed on cDNAs obtained by the two protocols indicated above, using a combination of the internal primers listed in Supplementary Table S1, which were located in different exons (Fig. 1A and 1B). The primers were designed to share the melting temperatures and optimized for the same amplification conditions: 120 s at 94ºC followed by 40 cycles of 94ºC for 20 s, 58ºC for 30 s and 72ºC for 90 s. All sequences have been submitted to GenBank and their accession number is quoted in Supplementary Table S2A and S2B.

Immunohistochemistry on mouse retina cryosections

Eyecups from 8 weeks old C57BL/6J mice were fixed in 4% paraformaldehyde (PFA) and 0.5% glutaraldehyde for 2 h at room temperature (RT), washed, and cryoprotected for 12 h in 8.5% acrylamide:bis-acrylamide (600:1) at 4°C. Eyecups were then embedded in O.C.T (Tissue-Tek, Sakura Finetech, Torrance, CA) and sectioned at -17°C. Sixteen micrometer sections on poly-lysine covered slides were used for immunohistochemistry.

Sections were dried (30 min), washed in PBS (10 min), and blocked in blocking solution (PBS containing 1% BSA, 3% sheep serum and 0,3% Triton X-100) at RT (1 h). Incubation with primary antibodies and peanut agglutinin (PNA) conjugated to AlexaFluor 647 (40 mg/ml, Invitrogen Life Technologies, Carlsbad, CA) was performed overnight at 4ºC in a solution of 1% sheep serum, 1% BSA, 0.05% Tween20 and 0.05% Triton X-100 in PBS. Sections were rinsed three times, followed by incubation with the corresponding secondary antibodies conjugated to either AlexaFluor 488, 546 or 568 (Invitrogen Life Technologies, Carlsbad, CA) (1:300) at RT (1 h). After 3 washes, sections were fixed in 4% PFA, rinsed twice, incubated with DAPI (1:5000, Sigma-Aldrich, Sant Louis, MO) for 15 min and washed in PBS 3 times (5 min each). Sections were cover-slipped with Fluoprep (Biomérieux, Marcy l'Etoile, France) and photographed with a Leica SP2 confocal microscope (Leica microsystems, Wetzlar, Germany).

The primary antibodies used were: mouse anti-Rhodopsin 1:500 (Abcam, Cambridge, United Kingdom), mouse anti-PKCα 1:500 (Santa Cruz Biotechnology, Inc., Santa Cruz, CA), and preabsorbed rabbit anti-CERKL 1:50.

Bioinformatic analysis of the genomic human CERKL locus

The genomic sequence of the human *CERKL* locus at chromosome 2 (March 2006 assembly version (NCBI36/hg18)) along with the upstream *ITGA4* and downstream *NEUROD1* loci, within the interval 182,029,864bp-182,259,440bp, was retrieved from the UCSC human genome browser[20] and used for most of the computational analyses conducted in this work. However, for comparative genomics purposes and in order to determine the conservation among human and other vertebrates (such as *Macaca mulatta*, *Mus musculus*, *Gallus gallus*

and *Takifugu rubripes*), pre-computed whole genome alignments were analyzed through the VISTA UCSC browser mirror, which provides the VISTA track feature[21]. The syntenic region of the mouse genome was also retrieved. BLASTN and TBLASTX alignments were performed on the syntenic sequences using the NCBI bl2seq algorithm[22] for a more in depth comparison between human and mouse.

Previously described *CERKL* isoforms were retrieved from several databases: RefSeq[23], GenBank[24], dbTSS[25], and VEGA[26]. Some of the dbTSS transcripts were already mapped on the human *CERKL* genomic region at the VEGA web site. These sequences, as well as experimentally validated *CERKL* cDNAs (this work), were mapped onto the analyzed sequence interval using Exonerate[27], following the est2genome model algorithm, in order to easily compare all the exonic structures from both sources, databases and experimental evidences.

Although a track for First-Exon-Finder program[28] on the UCSC genome browser was already available, an additional attempt to predict the presence of more CpG islands, promoters and first exons on the *CERKL* genomic region was performed (cutoff value for the first-exon a-posteriori probability (APP)=0.5, cutoff value for the promoter APP=0.4 and cutoff value for the promoter APP=0.4).

The genomic sequences for a set of 49 genes related to RP, classified into 10 distinct functional classes, were downloaded from GenBank: *RHO, PDE6A, PDE6B, CNGA1, CNGB1, SAG, GUCA1B* and *GUCY2D* (phototransduction)*; ABCA4, LRAT, RPE65, RLBP1, RGR, RDH12* and *RBP3* (retinol metabolism); *PRPH2, PROM1, EYS* and *ROM1* (photoreceptors structure); *CRX, NR2E3, NRL* and *OTX2* (transcription factors); *SEMA4A, MERTK, CRB1* and *USH2A* (cellular interaction); *PRPF3, PRPF8, PRPF31, RP9, SNRNP200* (mRNA processing); *TULP1, RPGRIP, RPGR, RP2, FSCN2, RP1, AIPL1, CEP290* and *LCA5* (transport); *KLHL7* and *TOPORS* (ubiquitin/proteasome pathway); *IMPDH1, CA4* and *IDH3B* (several types of enzymatic activities); and, finally, *RD3, SPATA7* and *PRCD* (unknown function). The sequences upstream (up to 10 kbp) of these genes were extracted and searched for over-represented motifs by running MEME[29]. An initial analysis was performed over the whole set

of sequences; then, in a second round, MEME was run separately for the sequences on each functional class. In order to characterize long and short motifs, two different sets of parameters were used. To search for long motifs, the "*anr"* model was used, with a minimum width of 8 and a maximum of 20, and a total of 200 iterations [-mod anr -nmotifs 20 -minw 8 -maxw 20 -maxiter 200]. To identify the short ones, the same model was applied, but the maximum width was reduced to 10 [-mod anr -nmotifs 10 -minw 8 -maxw 10 -maxiter 200]. Both sets of parameters were applied to the whole dataset analysis and to the split group consisting of 10 different functional classes. For each characterized motif, a loglikelyhood matrix was derived using two background models, the random model (equiprobability for all four nucleotides) and the model considering the GC content bias (40%GC for the whole *CERKL* genomic sequence including the neighbouring loci).

Apart from all those matrices generated by MEME, a set of matrices corresponding to a selection of known transcription initiation factors (including TATA, CAAT, USF, INI, SRF, SP1, and TFIIA) were downloaded from TransFac[30]. Retina-related transcription factor matrices (for PAX6, AP1, ZF5, AP2REP, AP2ALPHA, AP2GAMMA, TBP, MAZR, CRX, GATA4, SP3, ETF, KROX, WT1, NR2E3, V-MAF and WT1) were also gathered from TransFac, Promo[31] and Jaspar[32]. All the matrices were mapped into the analyzed genomic region of *CERKL* with the aid of custom Perl scripts, with the specific purpose of defining potential novel alternative Transcription Starting Sites (TSS) for *CERKL* isoforms. The scores for all the matrices hits on the genomic sequence were normalized between 0 and 1; then, a threshold was defined as the score above 95% of the distribution for all those scores. Only matrices hits having a normalized score equal or greater than that threshold were considered (a summary of those found on the 1 kbp upstream for every reported human and mouse *CERKL* exons which included a TSS is provided on Supplementary Table S4A and S4B, for human and mouse respectively).

Putative translation start sites were evaluated using the Kozak's matrix[33] under the same terms. Moreover, the ENCODE H3K4Me3 track[34] on the UCSC genome browser was also considered as an additional transcriptional evidence, given that histone modification correlates with transcriptionally active sites[35].

RESULTS

Comprehensive identification of alternatively spliced *CERKL* isoforms

Evidences of different alternatively spliced isoforms of *CERKL* have been reported, but a comprehensive prioritized list of the physiologically relevant transcripts is still missing[19, 36]. Furthermore, its wide tissular expression[17, 18] appears to be inconsistent with the tissue-restricted phenotype of *CERKL* mutations, as only the retina is affected. In this case, as it also happens with other retina-associated disease genes, tissue-specific isoforms may reconcile this apparent paradox[14].

Thus, we first aimed to exhaustively characterize the *CERKL* alternatively spliced isoforms generated in human and murine retinas, and perform an interspecific comparative analysis. Two different methods for the synthesis of the cDNAs (detailed description in the Material and Methods section) were used to replicate the experiments, validate the sequences and avoid technical biases. For a comprehensive isoform characterization, we performed 5' and 3' RACE reactions to identify initial and terminal UTRs, and subsequently used a battery of internal PCR primers (listed in Supplementary Table S1 and located in Fig. 1 A2 - human and Fig. 1 B2 - mouse) to unveil the combinatorial network of alternative promoters and exons displayed in *CERKL* transcripts. From these data we designed specific primers to identify fully processed transcripts encompassing from the first to the last exon, and thus depict the complete repertoire of *CERKL* and aligned with the genomic primary structure as a means to validate each transcript variant.

Overall, the retinal *CERKL* isoforms generated by alternative splicing events showed an unexpected complexity, as more than 20 transcripts were identified in human and mouse retinas. The genomic organization of *CERKL* with the splicing events (depicted as angled lines) and 5'UTRs (grey boxes) identified is shown in Fig. 1 A (human) and 1 B (mouse). The most abundant transcripts are indicated by the **#** symbol. For each human and mouse transcript the 3'UTR was unique, although murine transcripts contained a longer 3'UTR than previously reported, pointing to two polyadenylation signals. Notably, in the two species the 5' UTRs showed an unexpected multiplicity of TSSs, which

contributed to the combinatorial complexity of the mature transcripts. This heterogeneity called for a rational and comprehensive nomenclature of all *CERKL* variants in human and mouse. Therefore, sequences from published reports, databases and this work were gathered and systematized. Our proposal is presented in Tables S2A and S2B.

In detail, the analysis of the human 20 fully validated transcripts provided solid evidence of four different *CERKL* TSSs (Fig. 1A). Eleven transcripts were expressed from the previously reported 5' UTR; two, from the starting site of the adjacent upstream *NEUROD1* gene (known to be highly expressed in the CNS and transcribed in the same direction than *CERKL*); six, from an internal previously unknown initiation site within exon 1 (referred as exon 1b along the text and Supplementary material) and, finally the last transcript started from an internal sequence of exon 3 (referred as exon 3a). Of note, the TSS of exon 1b was also supported *in silico* by: i) the First-Exon-Finder, which among other structural features mapped a CpG island within this genomic region, and ii) the clustering of peaks of the H3K4Me3 track, indicative of transcriptionally active chromatin sites (Fig. 3). Yet, we cannot rule out that *CERKL* is transcribed from unknown TSSs in other tissues. In this context, the UCSC genome browser has recently incorporated an ENCODE track that corresponds to manually annotated genes, mostly based on sequenced full-length cDNAs from dbTSS plus reports from independent sources. Twelve out of the 15 ENCODE *CERKL* variants fully overlapped with some retinal transcripts described in this work. Of the remaining three, one (OTTHUMT00000334820) started at a TSS extremely close to the reported *CERKL* 5' UTR and possibly was structurally equivalent; the other two (OTTHUMT00000334817 and OTTHUMT00000334818) started at a completely different internal sites, suggesting two additional TSSs. If the latter two isoforms were validated, the number of *CERKL* TSSs in human would amount to 6.

In contrast, in murine retina only 3 *Cerkl* start sites were experimentally identified (shown in dark grey in Fig. 1B): eleven (out of 23) fully validated transcripts started from the previously reported *Cerkl* site, eleven from the upstream *NeuroD1* gene (as in human), and the last one from the novel exon 3a, located in intron 2. The latter is also supported by the dbTSS database.

Moreover, RT-PCR assays performed in a panel of several tissues provided evidence for an additional TSS within intron 2, which generated exon 3b (not found in retina).

To identify the more abundant transcripts and approach their relative physiological relevance (Fig. 2A-human and 2B-mouse), we used a battery of primers, located either at the different TSSs or the alternative exons at the 5' of *CERKL,* paired with a unique reverse primer in exon 10 (human) or 12 (mouse). The location of the primers is indicated in Fig. 2C. For isoform assignment, each amplified product was isolated and sequenced. The RT-PCRs were replicated several times. The interspecific comparison of the more abundant transcripts in retina revealed a higher number of *CERKL* variants in human (8 transcripts out of 20, with a comparable level of expression) versus mouse (3 transcripts out of 23, with one major variant).

Concerning the CERKL/Cerkl protein isoforms, our data reveals that the combination of TSS multiplicity with the high number of alternative splicing events affecting the first exons (1 to 6), generate a complex pattern of mature transcripts that differ at the 5' end, but share the 3' moiety (exons 6 to 13), as shown in Fig. 2A (human) and 2B (mouse). The alternative 5' exons encode: i) the nuclear localization signals[18, 37], ii) the putative pleckstrin homology (PH) domain and the iii) diacylglycerol kinase (DAGK) signatures[17, 18, 37, 38]. Besides, the human gene includes an in-frame species-specific alternative exon (4b) embedded in the predicted DAGK domain, which interrupts the DAGK consensus signature. The comparison between human and mouse *CERKL* mature mRNAs showed that although the number of isoforms is similar, intron retention is more frequent in mouse than in human (Fig. 1 A2 and 1 B2, isoforms m9, m10, m11). These transcripts bear premature stop-codons and are candidates to be degraded by the nonsense mediated decay mechanisms (NMD), but if translated, would encode a C-terminal truncated protein.

Interestingly, one of the consequences of the use of alternative TSSs is that the previously reported initiation Met codon is not always included in the mature transcript. Then, additional Translation Initiation Sites (TISs) should be considered. *In silico* sequence analyses using motif searches with Kozak's

matrices predicted several TISs along the *CERKL* transcripts (Supplementary Table S5). Of these, only two encoded long peptide sequences, while the remaining TISs yielded a lower score value or would generate very short peptides. Initiation codons with significant TIS scores are indicated in Fig. 1 A2 and 1 B2. For each isoform, only the longest Open Reading Frame (ORF) starting with a high-score Met is depicted (filled boxes).

Exploring the promoter landscape of *CERKL* TSSs.

To shed light on the architecture of the *CERKL* promoters and define *in silico* potential novel alternative TSSs, we aimed to map conserved Transcription Factor Binding Sites (TFBS) on the 1 kb upstream region of every human *CERKL* exon. To this end, we used position weight matrices from: i) reported general transcription initiation motifs, ii) retina related transcription factors, as well as iii) matrices obtained by MEME after the analysis of 49 promoters from genes related to retinal degeneration, in order to underscore conserved retina-specific regulatory motifs (subfunctionalized MEMEs). For a detailed description of these analyses, see the Material and Methods section. The outcome of this search along the upstream sequences of every exon depicted 3 different scenarios, which corresponded to the patterns yielded by: 1) exons with a TSS functional in retina (NeuroD1, 5'CERKL UTR, 1b and 3a); 2) exons with TSSs not found in the retina (corresponding to the starting exons in the ENCODE transcripts OTTHUMT00000334817 and OTTHUMT00000334818), and 3) the remaining internal exons, which are not used as TSSs (Table 1).

Notably, a more focused analysis of the target sites of retina-specific transcription factors revealed several hits that are worth mentioning: a high-scoring hit for PAX6, right upstream exon 3, and some significant hits for CRX upstream *NEUROD1*. However, no hits within the 1 kb upstream region of each exon were found for NR2E3 nor V-MAF (used to detect NRL binding sites), although some were scattered along the *CERKL* genomic region. Overall, the gathered evidences point to distinct promoter architectures concerning TSSs, probably reflecting tissue-specific expression. The detailed list of TFBS, MEME,

and subfunctionalized MEME hits upstream each exon is available on Supplementary Tables S4A and S4B.

Genomic conservation of the *CERKL* region among vertebrates

VISTA tracks on Fig. 3 clearly outline evolutionary conservation of the CERKL syntenic regions among vertebrates (human —*Homo sapiens*—, rhesus chimp —*Macaca mulatta*—, mouse —*Mus musculus*—, chicken —*Gallus gallus*—, and fugu —*Takifugu rubripes*—). The degree of sequence conservation is high, close to 100% between human and rhesus. Among tetrapods, the average degree of conservation is above 70% for all exons, but drops significantly in introns and intergenic regions. However, exon 4b could be an innovation in the ape lineage leading to humans, as it is unique to the human genome. The comparison with fugu reveals an expected lower degree of conservation, as only *NEUROD1* exons rank above 70% while most *CERKL* exons (2, 3, 5, 7, 8, 9, 10, 11, and 12) and *ITGA4* exons (3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 17, 19, 20 and 21) range between 50% and 70% similarity. Surprisingly, *CERKL* exon 1 was among the least conserved. These results agreed with those obtained from bl2seq comparisons between human and mouse syntenic regions both, at the nucleotide (BLASTN) and translated (TBLASTX) levels.

So far, no evidence supporting additional exons for *CERKL* apart from those described in this work could be obtained.

CERKL expression in a collection of human and mouse tissues

Semi-quantitative RT-PCR analysis of the *CERKL* expression was performed in a collection of tissues and cell lines from human and mouse with a pair of primers located in the exons shared by all isoforms (forward in exon 9 and reverse in exon 13, in human, and in exon 12, in mouse, see Fig. 1 A2 and 1 B2 for location, details in the Material and Methods section). The results are shown in Figure 4A and 4C (human) and 4B and 4D (mouse). At least three

independent replicates were performed and quantified for each tissue. *GAPDH* expression was used for normalization.

In humans, the retina was by far the tissue where *CERKL* expression is the highest. In fact, among the other tissues only the brain showed some detectable expression (at levels below 10% of those in retina). Sequence analysis of the brain transcript revealed that gene expression was driven by the *NEUROD1* TSS (data not shown). Of interest for future functional studies, some human cell lines showed detectable levels of *CERKL* transcription, as it is the case of HEK293T and A549 (Fig. 4C).

In mouse, *Cerkl* was also highly expressed in the retina, although the liver showed even slightly higher levels of expression (Fig. 4D). Sequence analysis of the murine liver isoform (marked with an asterisk) showed that it corresponded to m30in variant. This isoform would generate a prematurely truncated protein, as it retained a non-coding fragment of intron 11. Other mouse tissues, such as testis and spleen, also showed high to moderate levels of *Cerkl* expression.

As aforementioned, besides the reported mouse *Cerkl* promoter (heretofore, UTR), retinal transcripts were produced from the *NeuroD1* promoter and an internal TSS in intron 2 (from now on, named 3a). Direct sequencing of RT-PCRs from other tissues led us to identify another TSS, 3b, also within intron 2. We performed RT-PCR assays to assess the relative contribution of these TSSs in retina: UTR, *NeuroD1,* 3a and 3b to *Cerkl* expression (Fig. 5). Tissular comparison showed a wide range of expression from each TSS: the *Cerkl* UTR contribution was indeed major in retina, moderate in kidney, faint in brain, and undetectable in blood and spleen. Also, in agreement with previous reports, *NeuroD1*-driven expression was tissue-restricted, and only observed in retina in our panel. In contrast, the 3a TSS-driven transcript was expressed more widely, but showed very low levels in retina. Although the 3b TSS was silent in mouse retina, it was the most active in liver (Fig. 5). Isoforms m24 to m28 in Fig. 1B, which started at either 3a or 3b TSSs, were isolated and sequenced in spleen and liver, but were undetectable in retina. Of note, in some tissues the RT-PCRs specific for these 4 promoters did not explain the total *Cerkl*

transcriptional levels (as revealed by the amplification of the 9-12 exon region common to all isoforms), again pointing to additional TSSs.

Cerkl localization in mouse retina by immunohistochemistry

Previous results based on *in situ* mRNA hybridization showed that *Cerkl* was mainly expressed in the ganglion cell layer, although a fainter level of expression was detected in other retinal layers, including photoreceptors[17]. To accurately assess the localization of the Cerkl protein in the retina, fluorescent immunohistochemistry using different cell-specific antibodies and markers was performed on serial sagital cryosections of adult mouse retinas (2 months old). An in-house rabbit polyclonal anti-Cerkl antibody raised against an exon 2 peptide sequence was affinity-purified and pre-absorbed before use. Double co-immunodetection with this polyclonal anti-Cerkl antibody and either anti-Rhodopsin (specific for rods) or anti-PKCα (which mainly labels bipolar cells and rods), plus counter-staining with DAPI (nuclei) and AlexaFluor 647 conjugated PNA (which labels cones) were performed in parallel to allow a more detailed localization (Fig. 6).

The Cerkl expression was found at the ganglion cell layer (GCL), in the photoreceptors (PhR) and some cell bodies at the outer and inner nuclear (ONL and INL) layers (Fig. 6). Magnification of the photoreceptor cell layer showed a strong immunodetection of Cerkl in cones and faintly, in rods. Of interest, Cerkl mostly localized in the outer segments of both types of photoreceptors, as shown by co-localization of Cerkl with rhodopsin (rods) and cone staining (Fig. 6H and 6I). Besides, Cerkl also showed a perinuclear staining in some cell bodies at the ONL – extremely close to the photoreceptor layers–, probably corresponding to cones (Fig. 6I and 6J, white arrows). Concerning other neuronal retinal types, Cerkl was detected in a population of bipolar cells (white arrowheads in Fig. 6N) as well as in other cell types at the INL, as yet undetermined.

DISCUSSION

One of the major breakthroughs from interspecific sequence comparisons of whole genomes is that the complexity of a particular organism depends not only on the number of genes, but also on the diversity of the proteins produced and the regulation of transcription. An increasing number of evidences in the human genome supports that alternative splicing is more the rule than the exception, as more than 95% of the multiexon genes undergo alternative splicing events, often related to developmental or tissue differentiation processes and differential physiological functions. Many bioinformatic efforts are now being devoted to decipher "the splicing code", which intends to characterize the regulatory splicing strategies on a genome-wide scale in order to predict the specific transcripts from every gene[39, 40]. However, these *in silico* predictions have to be substantiated *in vivo* to identify the physiological relevant isoforms, their regulation, and eventually their contribution to disease. Within this framework, we have combined both, *in vivo* and *in silico* approaches, to analyze the expression of *CERKL,* a retinitis pigmentosa gene of an as yet unknown function. Our data shows an unexpected high transcriptional complexity in human and mouse tissues arising from the combination of tissue-specific promoters and alternative splicing events. A large multiplicity of isoforms, which reflects a sophisticated regulation in the retina, has also been reported for other genes, such as *RPGR*, *RPGRIP1* or *CPEB3*, but only after an accurate transcriptional characterization[41-43].

This high repertoire of CERKL transcript and protein isoforms suggests distinct roles for the alternatively displayed domains. The first two exons of *CERKL* encode a PH domain and two nuclear localization signals (NLSs), while exons 3 to 6 encompass the DAGK domain[17-19, 37, 38] (Fig. 7). Notably, the use of the different promoters and 5'UTRs affects the inclusion/exclusion of the first exons in the final transcript and generate variability at the N-terminal peptide moiety, with a potential impact in the protein function, which supports a finely tuned regulation of the 5' splicing events. In contrast, the exons encoding the C-terminal domains are maintained in all isoforms, even in the transcripts from non-retinal tissues, arguing in favor of a basic function. The comparison between human and mouse retina major *CERKL* isoforms reveals higher

complexity for the human transcripts. In fact, the most abundant isoforms are species-specific (except for h2 and m1, which are structurally equivalent), e.g. the *NeuroD1* promoter contributes to the highly expressed isoforms in mouse whereas its relevance in human adult retina appears to be minor. This holds true for the least abundant isoforms (e.g. h1, h12-h17 and m5, m7, m8-m11, m13, etc.) (Fig. 1A2 and Fig. 1B2). Interspecific differences in the levels of expression and identification of species-specific isoforms have also been reported for other visual disorder genes, such as *IMPDH1*, *OPA1* or *PRPF31*, suggesting distinct functional requirements for each species[44-46]. Remarkably, one third of the murine isoforms (12 out of 32) versus 1 out of 21 in human are produced by mis-splicing (with partial retention of intron sequences). Most of these mis-spliced transcripts would encode a truncated protein, unless degraded by NMD. If extended to other genes, these results would argue in favor of either a more precise splicing machinery, or a less permissive mRNA integrity control, in human.

One of the relevant findings of our work is the use of tissue-specific TSSs and TISs in mice. Among the tissues analyzed, the *NeuroD1* promoter was only active in retina, where the reported *Cerkl* UTR promoter also showed the highest transcriptional activity. Instead, the additional alternative internal promoters were highly expressed in non-neuronal tissues (see the lanes of liver, testis and kidney in Fig. 5). The combination of different promoters and shared splicing events in both species hinders isoform quantification by real time RT-PCR (which relies on small probes) to evaluate their contribution to the *CERKL* transcript population. Thus, a relative quantification by specific amplification of each isoform was performed (Fig. 2, 4 and 5).

The *in silico* analysis of binding sites for transcription initiation factors across the *CERKL* genomic neighbourhood (around 230 kbp) revealed a high number of hits (more than 15,000). However, they were not randomly distributed but clustered just upstream *ITGA4*, *NEUROD1* and *CERKL* canonical TSSs. If we focus on the retina-specific TFBS, no significant scores for OTX2 or NR2E3 could be found upstream the promoters of the aforementioned genes. In contrast, binding sites for CRX, PAX6 and NRL upstream *NEUROD1* TSS, and for NRL in *CERKL* exon 1b, and PAX6 and CRX upstream exon 3 TSS were

identified. These results provide evidences for retina-specific regulatory enhancers close to *CERKL*. Overall, i) the differential patterns observed for the *in silico* predicted enhancers, ii) the TSSs experimentally confirmed in the retina, and iii) the identification of non-retinal transcriptional products, clearly support a highly tuned distinct tissue-specific regulation of *CERKL* expression.

Notably, Cerkl immunohistochemistry showed high expression in cones and moderate in rods, ganglion cells and in other retinal INL cell types. A specific perinuclear staining was observed at the INL and ONL. Hitherto, *CERKL* mutations have been associated both to conventional RP and cone-rod dystrophy (CRD). Regarding this clinical heterogeneity, our findings of expression in cones and rods are consistent with the two clinical entities, but also highlight the need to establish a more accurate scenario. Therefore, full characterization of the transcriptional map of isoforms, the type and location of the mutations, the accurate subcellular localization of proteins, and the action of modifier genes is required to comprehend the contribution of *CERKL*/CERKL variants to retinal degeneration disorders.

Meanwhile, as more mutations are being identified, a genotype-phenotype correlation pattern is emerging (Fig. 7 and Table 2). The first pathogenic variant described p.R257X, a nonsense homozygous mutation in exon 5, generates a truncated protein that abrogates the putative DAGK domain. Interestingly, only one of the eight major isoforms remains unaffected after alternative splicing. The phenotype associated to this variant ranges from canonical RP to more severe CRD features[47]. Another RP-associated mutation, p.R106S, is localized in one of the two putative NLS, probably compromising its import and function in the nucleus[48]. However, all other protein domains remain unaltered, in accordance to a moderate RP phenotype. Other alleles are associated to more severe retinal disorders, with clear cone-rod dystrophy features and early macular degeneration; one of them, c.238+1G>A[36], affects the splicing of the first intron, abrogating the generation of the protein isoforms produced from exon 1 and 1b, thus only the putative isoforms starting in exon 3 or the spliced variants of exon 1a would be produced. The other mutation, p.C125W[49] (also affecting the conformation of the protein isoforms encoded from the methionine in exon 1) changes an evolutionarily conserved cysteine residue of the

pleckstrin domain. Three other clearly pathogenic alleles, two frameshifts (by indels) and a nonsense mutation, have been also reported but their association to particular features is hindered by their compound heterozygous status. Indeed, this is an ongoing task.

Our comprehensive approach, by characterizing a high number of isoforms expressed in a single tissue, provides an exhaustive transcriptional picture on a hitherto fragmentary collection of data and builds a reference framework to assess the severity of new mutations. Considering the high number of *CERKL* isoforms, undertaking accurate analysis for localization and/or functional specificity at the subcellular level remains a key challenge to understand the contribution of this gene to retinal degeneration.

ACKNOWLEDGMENTS

We would like to acknowledge the generous support from Andrés Mayor (Fundaluce, Hospital Central de Asturias). We appreciate the generous support and technical advice on the use of eye cryosections of Dr. Ana Méndez-Zunzúnegui (IDIBELL, Universitat de Barcelona). We also thank ENCODE project for making publicly available, via UCSC Genome browser, the H3K4Me3 and the GENCODE manual gene annotations (including VEGA) tracks. A. G., M. R. and M.C-M were in receipt of the fellowships FPI BES-2007-15414, FPU AP2007-00805 and FPI BES-2010-030745, respectively. E. P. was under contract by CIBERER. This study was supported by grants SAF2009-08079 (Ministerio de Ciencia e Innovación) and SGR2009-1427 (Generalitat de Catalunya), CIBERER (U718), Fundaluce and ONCE to R.G.-D and BFU2010-15656 to G.M.

REFERENCES

1. Sultan M, Schulz MH, Richard H, et al. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* 2008;321:956-960.
2. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* 2008;40:1413-1415.
3. Nilsen TW, Graveley BR. Expansion of the eukaryotic proteome by alternative splicing. *Nature* 463:457-463.
4. Licatalosi DD, Darnell RB. RNA processing and its regulation: global insights into biological networks. *Nat Rev Genet* 11:75-87.
5. McGlincy NJ, Smith CW. Alternative splicing resulting in nonsense-mediated mRNA decay: what is the meaning of nonsense? *Trends Biochem Sci* 2008;33:385-393.
6. Tazi J, Bakkour N, Stamm S. Alternative splicing and disease. *Biochim Biophys Acta* 2009;1792:14-26.
7. Raponi M, Baralle D. Alternative splicing: good and bad effects of translationally silent substitutions. *Febs J* 277:836-840.
8. Ward AJ, Cooper TA. The pathobiology of splicing. *J Pathol* 220:152-163.
9. Xu X, Liu Y, Weiss S, Arnold E, Sarafianos SG, Ding J. Molecular model of SARS coronavirus polymerase: implications for biochemical functions and drug design. *Nucleic Acids Res* 2003;31:7117-7130.
10. McCullough RM, Cantor CR, Ding C. High-throughput alternative splicing quantification by primer extension and matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Nucleic Acids Res* 2005;33:e99.
11. Licatalosi DD, Darnell RB. Splicing regulation in neurologic disease. *Neuron* 2006;52:93-101.
12. Hartong DT, Berson EL, Dryja TP. Retinitis pigmentosa. *Lancet* 2006;368:1795-1809.
13. Beit-Ya'acov A, Mizrahi-Meissonnier L, Obolensky A, et al. Homozygosity for a novel ABCA4 founder splicing mutation is associated with progressive and severe Stargardt-like disease. *Invest Ophthalmol Vis Sci* 2007;48:4308-4314.
14. Schmid F, Glaus E, Cremers FP, Kloeckener-Gruissem B, Berger W, Neidhardt J. Mutation- and tissue-specific alterations of RPGR transcripts. *Invest Ophthalmol Vis Sci* 51:1628-1635.
15. Riazuddin SA, Iqbal M, Wang Y, et al. A splice-site mutation in a retina-specific exon of BBS8 causes nonsyndromic retinitis pigmentosa. *Am J Hum Genet* 86:805-812.
16. Wang Y, Juranek S, Li H, Sheng G, Tuschl T, Patel DJ. Structure of an argonaute silencing complex with a seed-containing guide DNA and target RNA duplex. *Nature* 2008;456:921-926.
17. Tuson M, Marfany G, Gonzalez-Duarte R. Mutation of CERKL, a novel human ceramide kinase gene, causes autosomal recessive retinitis pigmentosa (RP26). *Am J Hum Genet* 2004;74:128-138.
18. Bornancin F, Mechtcheriakova D, Stora S, et al. Characterization of a ceramide kinase-like protein. *Biochim Biophys Acta* 2005;1687:31-43.

19. Tuson M, Garanto A, Gonzalez-Duarte R, Marfany G. Overexpression of CERKL, a gene responsible for retinitis pigmentosa in humans, protects cells from apoptosis induced by oxidative stress. *Mol Vis* 2009;15:168-180.
20. Rhead B, Karolchik D, Kuhn RM, et al. The UCSC Genome Browser database: update 2010. *Nucleic Acids Res* 38:D613-619.
21. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. VISTA: computational tools for comparative genomics. *Nucleic Acids Res* 2004;32:W273-279.
22. Tatusova TA, Madden TL. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol Lett* 1999;174:247-250.
23. Pruitt KD, Tatusova T, Klimke W, Maglott DR. NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res* 2009;37:D32-36.
24. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. *Nucleic Acids Res* 2009;37:D26-31.
25. Wakaguri H, Yamashita R, Suzuki Y, Sugano S, Nakai K. DBTSS: database of transcription start sites, progress report 2008. *Nucleic Acids Res* 2008;36:D97-101.
26. Harrow J, Denoeud F, Frankish A, et al. GENCODE: producing a reference annotation for ENCODE. *Genome Biol* 2006;7 Suppl 1:S4 1-9.
27. Slater GS, Birney E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 2005;6:31.
28. Davuluri RV, Grosse I, Zhang MQ. Computational identification of promoters and first exons in the human genome. *Nat Genet* 2001;29:412-417.
29. Bailey TL, Boden M, Buske FA, et al. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* 2009;37:W202-208.
30. Matys V, Fricke E, Geffers R, et al. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res* 2003;31:374-378.
31. Messeguer X, Escudero R, Farre D, Nunez O, Martinez J, Alba MM. PROMO: detection of known transcription regulatory elements using species-tailored searches. *Bioinformatics* 2002;18:333-334.
32. Portales-Casamar E, Thongjuea S, Kwon AT, et al. JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res* 38:D105-110.
33. Kozak M. An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res* 1987;15:8125-8148.
34. Celniker SE, Dillon LA, Gerstein MB, et al. Unlocking the secrets of the genome. *Nature* 2009;459:927-930.
35. Ruthenburg AJ, Allis CD, Wysocka J. Methylation of lysine 4 on histone H3: intricacy of writing and reading a single epigenetic mark. *Mol Cell* 2007;25:15-30.
36. Auslender N, Sharon D, Abbasi AH, Garzozi HJ, Banin E, Ben-Yosef T. A common founder mutation of CERKL underlies autosomal recessive retinal degeneration with early macular involvement among Yemenite Jews. *Invest Ophthalmol Vis Sci* 2007;48:5431-5438.
37. Inagaki Y, Mitsutake S, Igarashi Y. Identification of a nuclear localization signal in the retinitis pigmentosa-mutated RP26 protein, ceramide kinase-like protein. *Biochem Biophys Res Commun* 2006;343:982-987.
38. Rovina P, Schanzer A, Graf C, Mechtcheriakova D, Jaritz M, Bornancin F. Subcellular localization of ceramide kinase and ceramide kinase-like protein requires interplay of their Pleckstrin Homology domain-containing N-terminal

regions together with C-terminal domains. *Biochim Biophys Acta* 2009;1791:1023-1030.
39. Barash Y, Calarco JA, Gao W, et al. Deciphering the splicing code. *Nature* 465:53-59.
40. Tejedor JR, Valcarcel J. Gene regulation: Breaking the second genetic code. *Nature* 465:45-46.
41. Lu X, Ferreira PA. Identification of novel murine- and human-specific RPGRIP1 splice variants with distinct expression profiles and subcellular localization. *Invest Ophthalmol Vis Sci* 2005;46:1882-1890.
42. Neidhardt J, Glaus E, Barthelmes D, Zeitz C, Fleischhauer J, Berger W. Identification and characterization of a novel RPGR isoform in human retina. *Hum Mutat* 2007;28:797-807.
43. Wang XP, Cooper NG. Characterization of the transcripts and protein isoforms for cytoplasmic polyadenylation element binding protein-3 (CPEB3) in the mouse retina. *BMC Mol Biol* 2009;10:109.
44. Spellicy CJ, Daiger SP, Sullivan LS, et al. Characterization of retinal inosine monophosphate dehydrogenase 1 in several mammalian species. *Mol Vis* 2007;13:1866-1872.
45. Akepati VR, Muller EC, Otto A, Strauss HM, Portwich M, Alexander C. Characterization of OPA1 isoforms isolated from mouse tissues. *J Neurochem* 2008;106:372-383.
46. Tanackovic G, Rivolta C. PRPF31 alternative splicing and expression in human retina. *Ophthalmic Genet* 2009;30:76-83.
47. Aleman TS, Soumittra N, Cideciyan AV, et al. CERKL mutations cause an autosomal recessive cone-rod dystrophy with inner retinopathy. *Invest Ophthalmol Vis Sci* 2009;50:5944-5954.
48. Ali M, Ramprasad VL, Soumittra N, et al. A missense mutation in the nuclear localization signal sequence of CERKL (p.R106S) causes autosomal recessive retinal degeneration. *Mol Vis* 2008;14:1960-1964.
49. Littink KW, Koenekoop RK, van den Born LI, et al. Homozygosity mapping in patients with cone-rod dystrophy: novel mutations and clinical characterizations. *Invest Ophthalmol Vis Sci*.

Table 1 - Distribution of motifs among 1 kbp upstream of every *CERKL* exon showed a differential pattern depending on the kind of exon.

Scenario	Transfac motifs	MEME motifs	Subfunctionalized MEME motifs	TSS type
1	>40	0	<5	Retinal TSSs (NEUROD1, 1/1a, 1b and 3a)
2	<35	≈10	>75	Non-retinal TSSs (OTTHUMT00000334817 first exon and OTTHUMT00000334818 first exon)
3	≈25	0	<5	No TSSs exons

Table 2 - Genotype-phenotype correlations of reported *CERKL* mutations.

Mutation * / exon	Protein domain / molecular effect	Allelic status	Major affected isoforms (Fig. 1A)	Phenotype †	Allele frequency among CERKL reported mutations	References
p.R257X / exon 5	Lipid kinase / protein truncated	homozygous and compound heterozygous	7 out of 8 (h2, h3, h7, h8, h13, h14, h18)	Retinitis Pigmentosa (RP), with some patients showing phenotypes closer to Cone-Rod Dystrophy (CRD). Peripheral pigment deposits plus macular dystrophy	30/ 40 alleles (20 families or patients). (approx. 75%)	Tuson et al.2004; Pomares et al., 2007; Avila-Fernández et al., 2008; Aleman et al., 2009; Littink et al., 2010 ‡
c.238+1G>A / intron 1	pleckstrin homology / abrogates splicing	homozygous	5 out of 8 (h2, h3, h5, h13, h14)	mixed features of RP and CRD, with early macular degeneration	2/40 alleles (approx. 5%)	Auslender et al., 2007
p.R106S / exon 2	nuclear localization signal / compromises nuclear import	homozygous	3 out of 8 (h2, h3, h5)	RP features (bone-spicules) with cone-rod degeneration leading to peripheral and central vision deficit	2/40 (approx. 5%)	Ali et al., 2008
c.156_157ins / exon 1	pleckstrin homology / frameshift and protein truncation	compound heterozygous with c.758delT	3 out of 8 (h2, h3, h5)	NC	1/40 (approx. 2.5%)	Tang et al., 2009
c.758delT / exon 5	lipid kinase / frameshift and protein truncation	compound heterozygous with c.156_157ins	7 out of 8 (h2, h3, h7, h8, h13, h14, h18)	NC	1/40 (approx. 2.5%)	Tang et al., 2009
p.C362X / exon 8	unknown function / protein truncated	compound heterozygous with p.R257X	All isoforms (h2, h3, h5, h7, h8, h13, h14, h18)	NC	1/40 (approx. 2.5%)	Aleman et al., 2009;
p.C125W / exon 2	pleckstrin homology/ evolutionarily conserved residue	homozygous	3 out of 8 (h2, h3, h5)	CRD (with central scotoma and macular atrophy, retinal thinning)	2/40 (approx. 5%)	Littink et al., 2010

***** The isoform cDNA used for reference is NM_201548.4, corresponding to isoform h2 (Fig. 1A). NC-not considered due to heterozygosity. **†** Only the phenotype for homozygous allelic combination is considered of value in genotype-phenotype correlations. NC-not considered. **‡** The reported mutation p.R283X considered as novel in this study corresponds to the already reported p.R257X variant (this difference is due to the isoform cDNA taken as reference, NM_001030311.2 and NM_201548.4, respectively).

FIGURE LEGENDS

FIGURE 1. Alternatively spliced *CERKL* isoforms in human and mouse retina. Extremely high complexity of the splicing events in human (**A1**) and mouse (**B1**) *CERKL* transcripts. Empty boxes represent exons, whereas black boxes show retained introns or cryptic non-coding exons. Angled lines above and below the gene structure indicate validated splicing events. Scheme depicting all the human (**A2**) and mouse (**B2**) spliced variants observed in retina. Exons are indicated as boxes and the coding sequence (CDS) for each isoform, considering the higher likelihood of first methionine (see text), is shown in black. Dark grey boxes represent the TSS found in retina, whereas light grey is used for non-retinal TSSs. **#** illustrates the main isoforms in each species. Arrows named with a letter indicate the position and direction of the primers used for PCR reactions (complete list and sequence in Supplementary Table S1). **^** depicts the non-retinal isoforms found in mouse liver and spleen. The scores of the Kozak's motif hits containing putative TIS methionines for human are: ★ 12.003; ▲ 5.248; ■ 8.389; ✦ 5.281; ● 8.852. As for mouse: ▼ 13.384; ○ 9.620; * 10.662; ◊ 8.389; ¤ 8.863 (complete list of all Kozak's scores in Supplementary Table S5).

FIGURE 2. Evaluation of *CERKL* main transcripts. RT-PCR from human (**A**) and mouse (**B**) retina total RNA, in order to identify the main isoforms. (**C**) Scheme depicting the structure of *CERKL* in human and mouse, with the location of the primers used to generate the PCR reactions. For the sake of clarity, exons not relevant to this assay are not shown. For all the amplicons, the same reverse oligonucleotide (Human: O; Mouse: b) was used, paired with the corresponding forward primers as follows, for human: Lane 1- C, 2- D, 3- E, 4- F, 5- G, 6- I, 7- J and 8- K; for mouse: lane 1- f, 2- g, 3- d, 4- h, 5- i, 6- j, 7- e and 8- k. The primer sequences are provided in the Supplementary Table S1.

FIGURE 3. Summary of annotated and custom features tracks on UCSC genome browser. (**A**) An overall view of the whole genomic neighborhood of human *CERKL*, including upstream *NEUROD1* (*ITGA4* downstream gene is shown in Supplementary Fig. S3). Homology to various species including

mouse is depicted on topmost tracks. Exonic structure of all the experimentally validated *CERKL* isofoms described in this paper. FirstExonFinder predicted TSS, the ENCODE histone track H3K4Me3, a custom track of hits to different position weight matrices for known and predicted transcription factor binding sites, as well as some further evidences of transcriptional activity on neural tissues are shown. (**B**) and (**C**) correspond to a magnification of the region around exon 1 and 3, respectively, containing a more detailed view of the TFBS sites. Same tracks distribution is depicted on all three panels. Matrices hits overlapping homopolymer stretches larger than 5 bp were discarded.

FIGURE 4. *CERKL* semi-quantitative expression analysis in human and mouse tissues. *CERKL* expression identified by RT-PCR in several tissues and cell lines of human (**A**) and mouse (**B**) origin. Semi-quantitative analysis of all *CERKL* transcripts in human (**C**) and mouse (**D**). At least three replicates were performed. *GAPDH* expression was used for normalization. Maximum *CERKL* levels were arbitrarily set as 100% (retina in human, liver in mouse). *CERKL* was amplified using primers A and B in human, and a and b in mouse, as located in Fig. 1 A2 and B2. The amplicon size is indicated in each case. The asterisk in the murine liver sample corresponds to the alternative isoform m24in. Primer sequences are provided in Supplementary Table S1. Notably, the primers used for the amplification of *CERKL* transcript were located in the common region at 3' of the gene; therefore, the bands observed are the result of the transcripts produced from all TSSs in each tissue.

FIGURE 5. Tissue-specific *Cerkl* promoter in adult mice. RT-PCRs were performed on several murine samples to determine the active promoters in each tissue. Fourty-five cycle amplifications were carried out using the same reverse oligonucleotide in exon 12 and different forward primers located in each TSSs identified (*NeuroD1* UTR, *Cerkl* UTR, 3a and 3b) as well as exon 9 to amplify the common region. *Gapdh* was used to normalize between samples. Primer location is depicted en Fig. 1B and sequences are listed in Supplementary Table S1.

FIGURE 6. Immunohistochemistry on mouse retina cryosections. A-J) Localization of Cerkl in photoreceptor cells. Nuclei are stained with DAPI (in

blue, A); Cerkl (B) and Rhodopsin (C) proteins are detected in green and magenta, respectively; cones appear in red (D) using PNA staining. Two merge images (E and F) and the magnification of some sections show clear localization of Cerkl in cones (in yellow, G) and also, more faintly in rods, colocalizing with rhodopsin (H). Although Cerkl localizes mainly in the outer segments, some perinuclear staining could be also observed in the nuclei of the cones at the ONL, indicated by white arrows in (I and J); J) DAPI counter-staining of the nuclei. **K-N) Expression of Cerkl in other retinal layers.** Nuclei are stained with DAPI (in blue, K), Cerkl protein is detected in green (L), bipolar cells and rods expressing PKCα are labelled in red (M). Cerkl is expressed in the ganglion cells (GCL), some cells at the inner and outer nuclear layers (INL, ONL) as well as in the photoreceptors. The merge image (N) shows expression of Cerkl in some bipolar cells (white arrowheads), while confirming localization in rods. The magnification scale is indicated.

FIGURE 7. Scheme of the reported causative mutations on the CERKL protein. The location of the mutations identified so far is shown on a diagram of the CERKL protein. The CERKL domains described by either sequence homology (Pleckstrin homology- PH; Diacylglycerol kinase domain- DAGK) or functional analysis (Nuclear localization signals- NLS; Nuclear Export signals-NES) are also depicted.

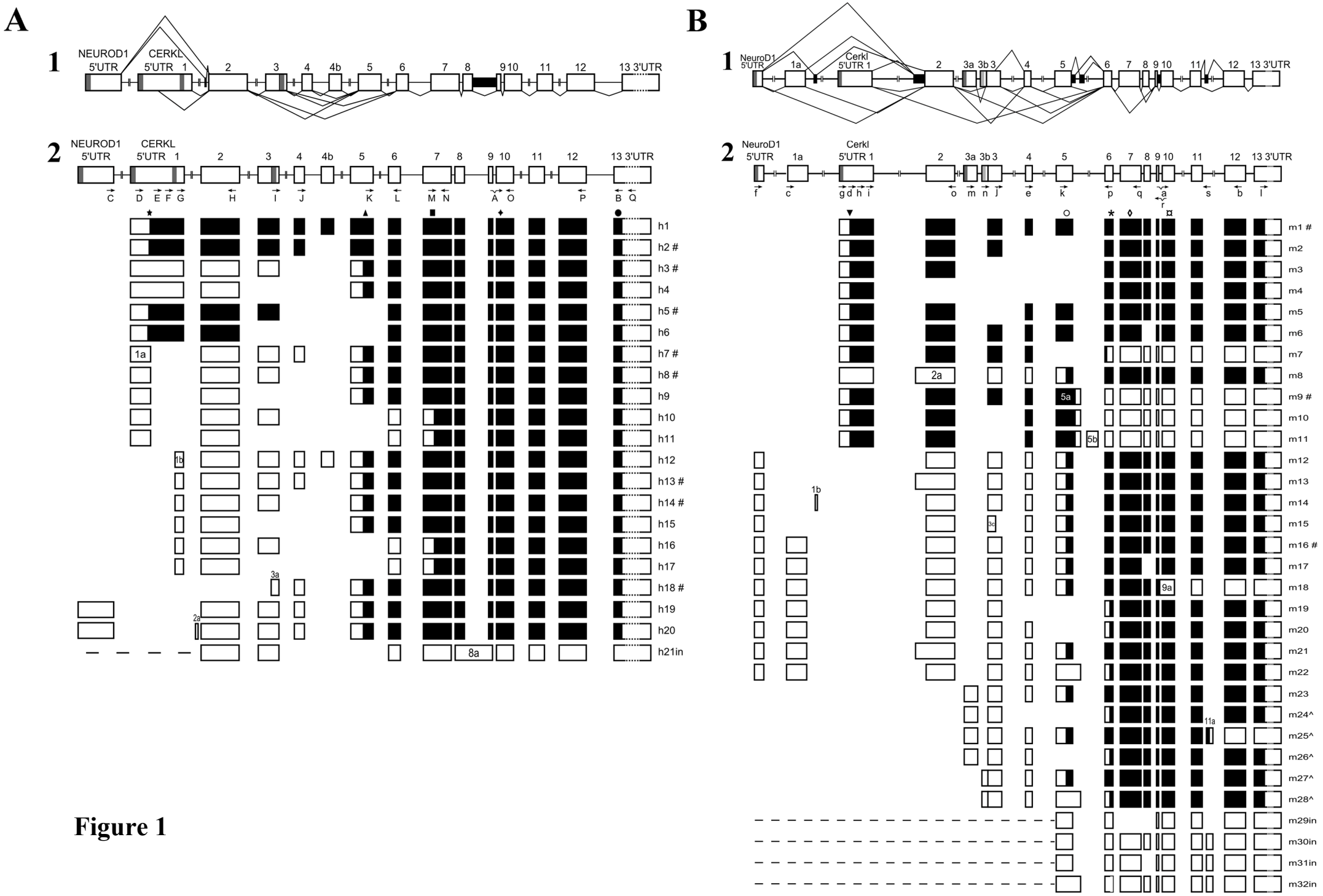
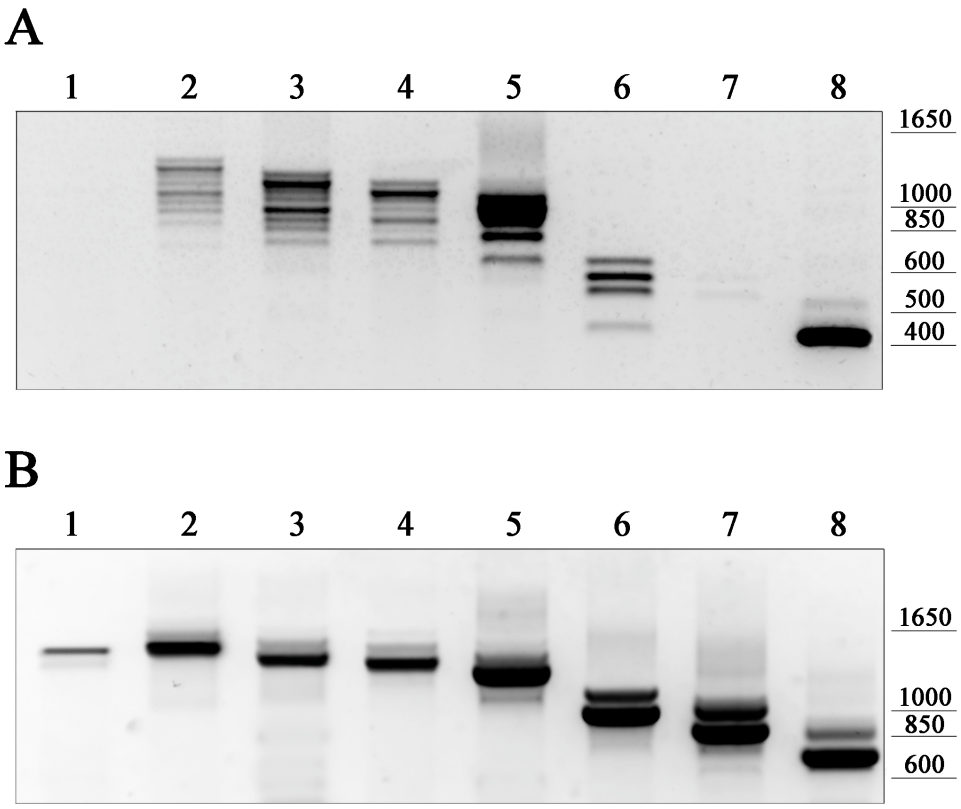

A
1
NEUROD1
5'UTR
CERKL
5'UTR 1
2
3
4
4b
5
6
7
8
9 10
11
12
13 3'UTR
2
NEUROD1
5'UTR
CERKL
5'UTR 1
C
D
E
F
G
H
I
J
K
L
M
N
A
O
P
B
Q
h1
h2 #
h3 #
h4
h5 #
h6
h7 #
h8 #
h9
h10
h11
h12
h13 #
h14 #
h15
h16
h17
h18 #
h19
h20
h21in
1a
1b
2a
3a
8a
B
1
NeuroD1
5'UTR
1a
Cerkl
5'UTR 1
2
3a
3b
3
4
5
6
7
8
9 10
11
12
13 3'UTR
2
NeuroD1
5'UTR
1a
Cerkl
5'UTR 1
f
c
g
d
h
i
o
m
n
j
e
k
p
q
a
r
s
b
l
m1 #
m2
m3
m4
m5
m6
m7
m8
m9 #
m10
m11
m12
m13
m14
m15
m16 #
m17
m18
m19
m20
m21
m22
m23
m24^
m25^
m26^
m27^
m28^
m29in
m30in
m31in
m32in
2a
5a
5b
1b
3c
9a
11a

Figure 1

Figure 2


A
1 2 3 4 5 6 7 8
1650
1000
850
600
500
400
B
1 2 3 4 5 6 7 8
1650
1000
850
600
C
Human
NEUROD1
5'UTR
CERKL
5'UTR 1
2 3 4 4b 5 6 7 8 9 10
C D E F G I J K O
Mouse
NeuroD1
5'UTR
1a
Cerkl
5'UTR 1
2 3 4 5 6 11 12
f g d h i j e k b

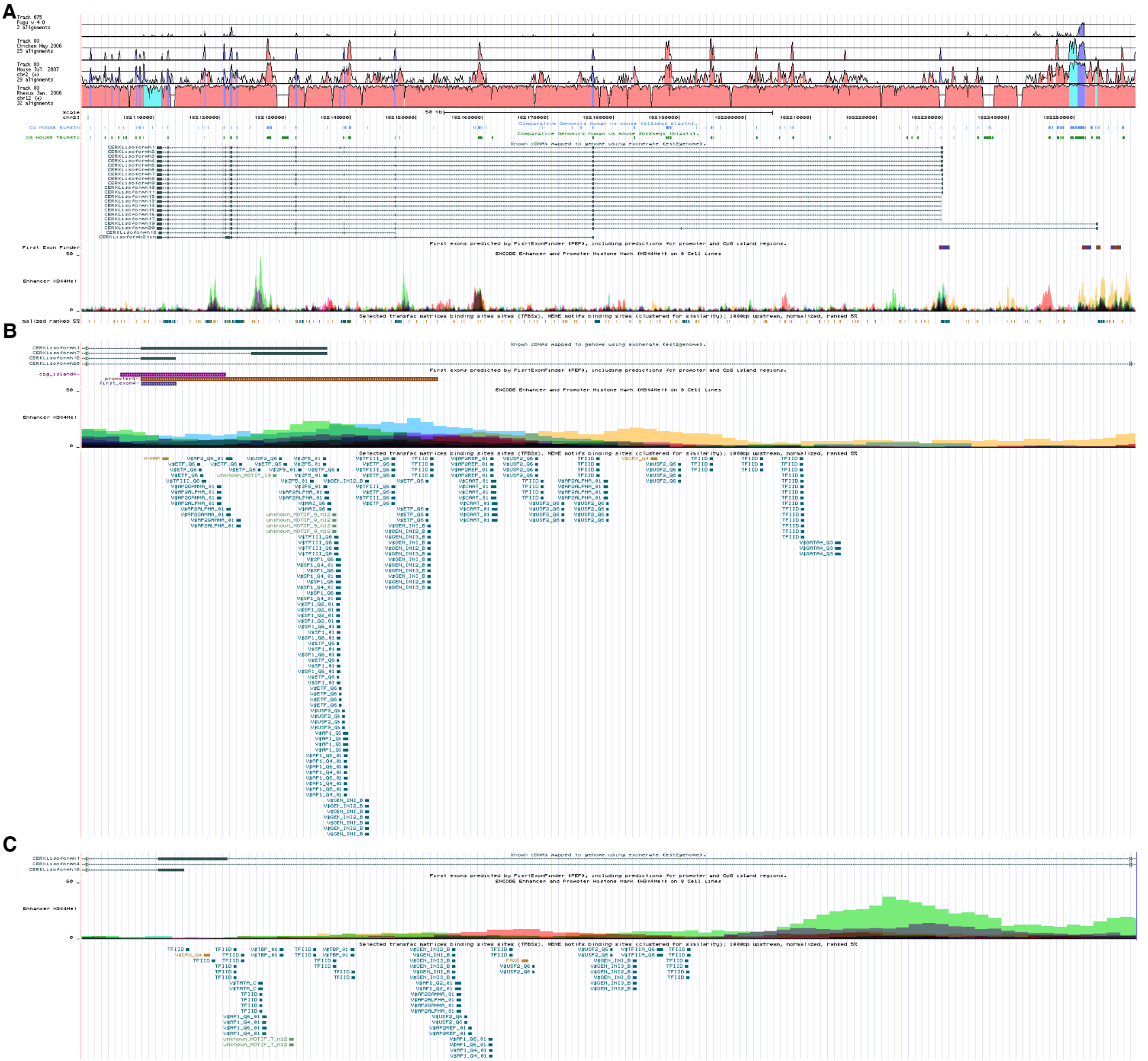

A
B
C
Known cDNAs mapped to genome using exonerate (est2genome).
First exons predicted by FirstExonFinder (FEF), including predictions for promoter and CpG island regions.
ENCODE Enhancer and Promoter Histone Mark (H3K4Me1) on 8 Cell Lines
Enhancer H3K4Me1
Selected transfac matrices binding sites sites (TFBSs), MEME motifs binding sites (clustered for similarity); 1000bp upstream, normalized, ranked 5%

Figure 3

Figure 4

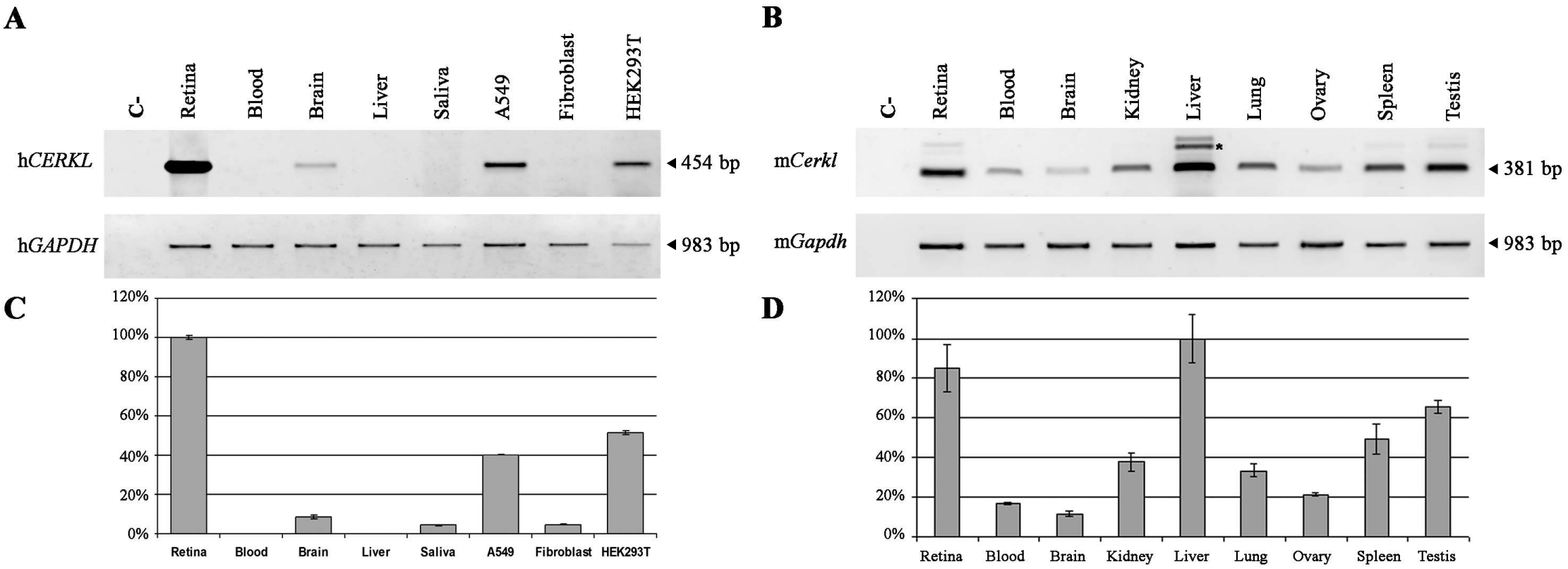

A
C-
Retina
Blood
Brain
Liver
Saliva
A549
Fibroblast
HEK293T
hCERKL
454 bp
hGAPDH
983 bp
B
C-
Retina
Blood
Brain
Kidney
Liver
Lung
Ovary
Spleen
Testis
mCerkl
*
381 bp
mGapdh
983 bp
C
120%
100%
80%
60%
40%
20%
0%
Retina
Blood
Brain
Liver
Saliva
A549
Fibroblast
HEK293T
D
120%
100%
80%
60%
40%
20%
0%
Retina
Blood
Brain
Kidney
Liver
Lung
Ovary
Spleen
Testis

Figure 5

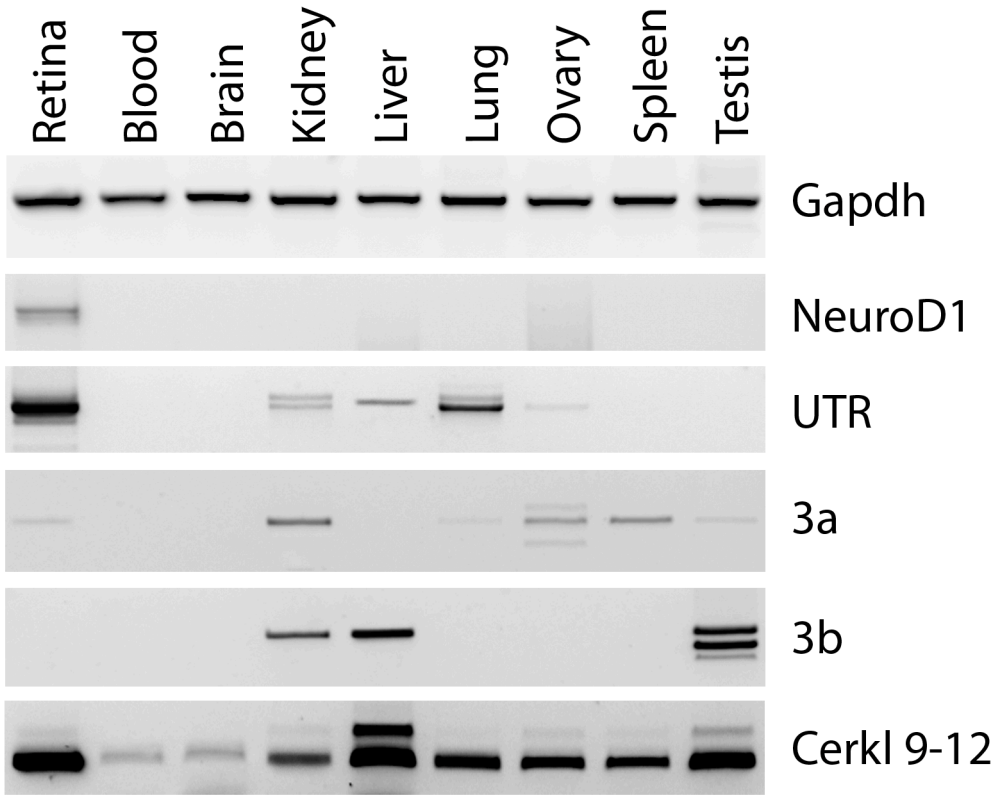

Retina
Blood
Brain
Kidney
Liver
Lung
Ovary
Spleen
Testis
Gapdh
NeuroD1
UTR
3a
3b
Cerkl 9-12

Figure 6

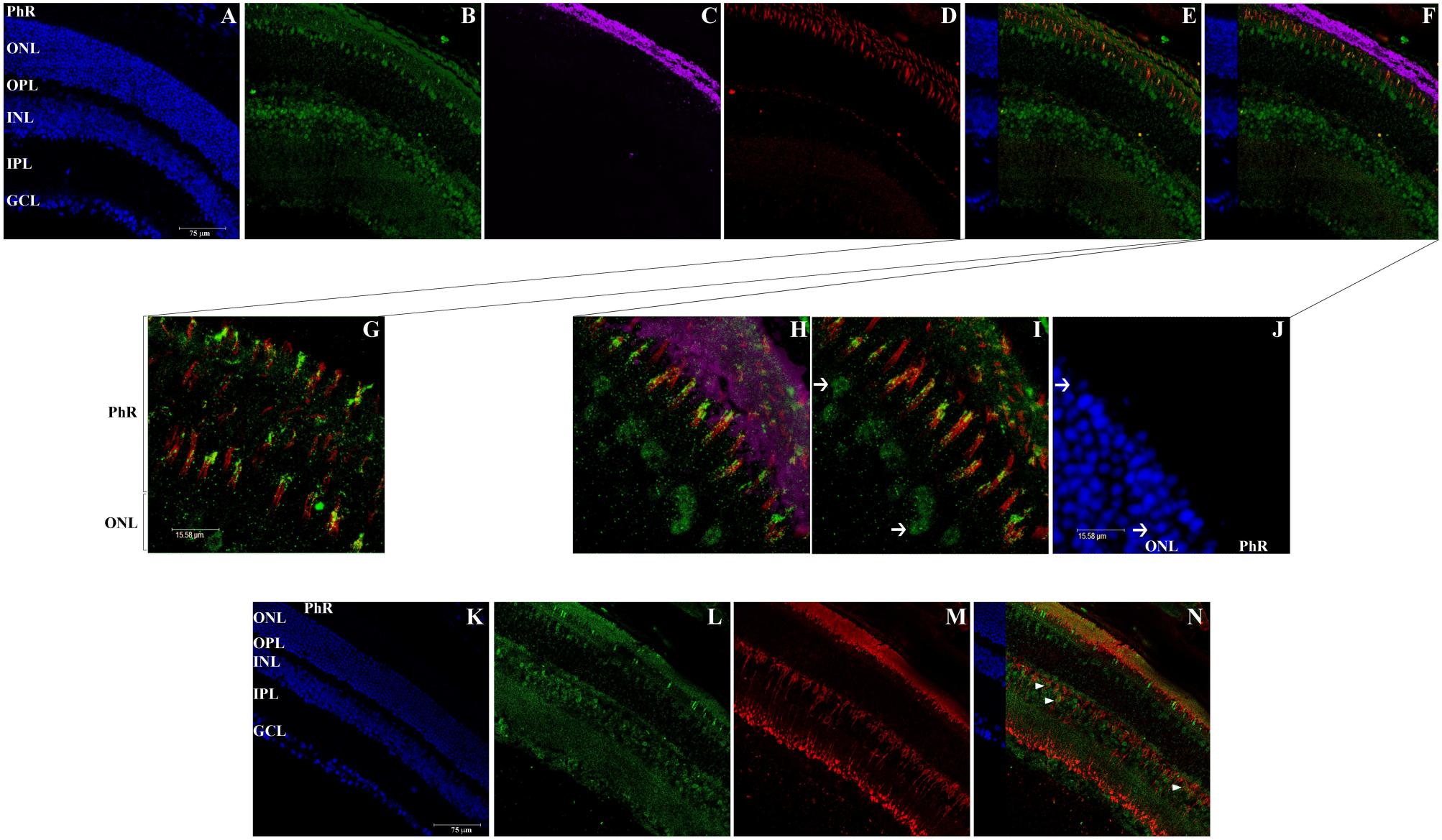
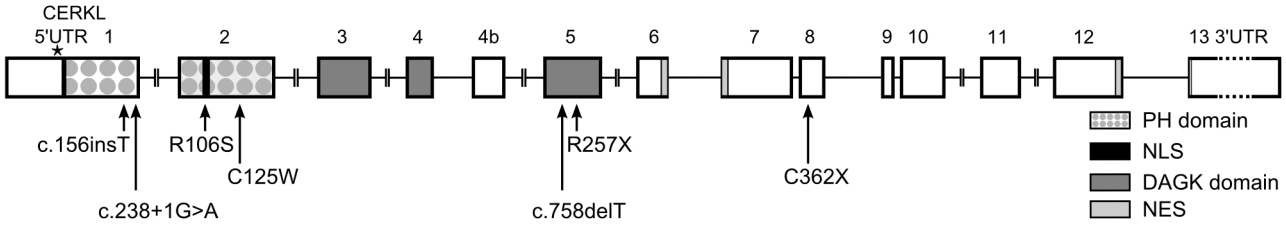

PhR
ONL
OPL
INL
IPL
GCL
A
B
C
D
E
F
75 µm
G
H
I
J
PhR
ONL
15.58 µm
15.58 µm
ONL
PhR
K
L
M
N
ONL
PhR
OPL
INL
IPL
GCL
75 µm

Figure 7


CERKL
5'UTR
1
2
3
4
4b
5
6
7
8
9
10
11
12
13
3'UTR
c.156insT
R106S
C125W
c.238+1G>A
R257X
c.758delT
C362X
PH domain
NLS
DAGK domain
NES