

High Transcriptional Complexity of the Retinitis Pigmentosa *CERKL* Gene in Human and Mouse

Alejandro Garanto,^{1,2,3,4} Marina Riera,^{1,2,3,4} Esther Pomares,^{1,2,3} Jon Permanyer,^{1,2} Marta de Castro-Miró,¹ Florentina Sava,¹ Josep F. Abril,^{1,2} Gemma Marfany,^{1,2,3} and Roser González-Duarte^{1,2,3}

PURPOSE. To shed light on the pathogenicity of the mutations in the retinitis pigmentosa gene *CERKL*, the authors 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 detected primarily in ganglion cells and cones but can also be observed in rods. *Cerkl* is mainly cytosolic. It localizes in the outer segments of photoreceptors and in the perinuclear regions of some cells.

RESULTS. An unexpected multiplicity of *CERKL* transcriptional start sites (four in each species) plus a high variety of alternative splicing events primarily affecting the 5' half of the gene generate >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 highly tuned gene regulation in different tissues, and establishes a framework for the study of *CERKL* genotype-phenotype correlations. (*Invest Ophthalmol Vis Sci.* 2011;52:5202-5214) DOI:10.1167/iovs.10-7101

Spatiotemporal differential splicing, often related to developmental events or tissue differentiation processes, affects >95% of the human genes, as recently unveiled after massive

sequencing of the human transcriptome.^{1,2} Alternative splicing and 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 that are frequently associated with 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 pathologic severity.⁶⁻⁸ This is even more relevant when considering genetic disorders of the mammalian central nervous system (CNS) and derived neurologic tissues, such as the retina, in which 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 nontraumatic adult blindness.¹² Although >45 genes have been identified as causative of RP (Retnet, <http://www.sph.uth.tmc.edu/Retnet/>), approximately 40% of the genetic cases remain unassigned, highlighting the relevance of identifying new candidates because each gene will presumably explain very few cases. The molecular diagnosis 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 abnormalities, decrease the number of unknown RP genes, illuminate new scenarios for tissue-specific gene function, and emphasize the need for accurate characterization of candidate splicing products, particularly because 70% of the exons in the human genome are tissue specific.^{1,16}

Our group first identified *CERKL* as an RP gene¹⁷ 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, some of the *CERKL* isoforms were a priori functional in the patients.¹⁹ These results prompted us to undertake a more accurate characterization of the *CERKL* transcripts in human and mouse. Our work unveils an unexpectedly 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 additional splice sites.

From the ¹Departament de Genètica, Facultat de Biologia, and the ²Institut de Biomedicina, Universitat de Barcelona, Barcelona, Spain; and ³CIBERER, Instituto de Salud Carlos III, Barcelona, Spain.

⁴These authors contributed equally to the work presented here and should therefore be regarded as equivalent authors.

Supported by Grants SAF2009-08079 (Ministerio de Ciencia e Innovación) and SGR2009-1427 (Generalitat de Catalunya), CIBERER (U718), Fundaluce and ONCE (RG-D) and BFU2010-15656 (GM). AG, MR, and MC-M were in receipt of the fellowships FPI BES-2007-15414, FPU AP2007-00805, and FPI BES-2010-030745, respectively. EP was under contract by CIBERER.

Submitted for publication December 20, 2010; revised March 31, 2011; accepted April 11, 2011.

Disclosure: A. Garanto, None; M. Riera, None; E. Pomares, None; J. Permanyer, None; M. de Castro-Miró, None; F. Sava, None; J.F. Abril, None; G. Marfany, None; R. González-Duarte, None

Corresponding author: Roser González-Duarte, Departament de Genètica, Facultat de Biologia, Universitat de Barcelona, Avenida Diagonal 645, 08028 Barcelona, Spain; rgonzalez@ub.edu.

Overall, these results, together with the bioinformatics analysis, strongly support the generation of many protein isoforms and the different roles of CERKL in retinal cells and other tissues, and they provide a molecular framework for genotype-phenotype correlations because the location of the mutation in the *CERKL* gene would affect the number and type of transcripts and, hence, be related to the progression and severity of the disease.

MATERIALS AND METHODS

Animal Handling, Tissue Dissection, and Preparation of Samples

All animal handling and procedures were performed according to the ARVO Statement for the Use of Animals in Ophthalmic and Vision Research and the regulations of the animal care facilities at the University of Barcelona. In brief, C57BL/6J mice (Charles River Laboratories, Davis, CA) were euthanized with CO₂ followed by cervical dislocation, and specific tissues and organs were dissected and immediately frozen in liquid nitrogen. Human blood and saliva samples were collected from nonaffected subjects with RP, after they provided informed consent, in accordance with the tenets of the Declaration of Helsinki. Retina and brain total RNA samples were supplied by Clontech Laboratories, Inc. (Mountain View, CA), and liver cDNA was provided by BD Biosciences (San Jose, CA).

Cell Culture and Constructs

Human embryonic kidney cells 293T (HEK293T, Bethyl Laboratories, Montgomery, TX) and wild-type fibroblasts (kindly provided by Daniel Grinberg and Luüisa Vilageliu) were grown in DMEM with 4 mM L-glutamine. The human lung adenocarcinoma epithelial cell line (A549; Abcam, Cambridge, MA) 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 penicillin, and 100 µg/mL streptomycin (Invitrogen Life Technologies, Carlsbad, CA).

RNA Extraction and RT-PCR

For total RNA extraction, a tissue kit (High Pure RNA Tissue Kit; Roche Diagnostics, Indianapolis, IN) was used in accordance with the manufacturer's instructions. Human and mouse blood RNA was mixed (RNAlater; Ambion/Applied Biosystems, Foster City, CA) before extraction (RiboPure-Blood Kit; Ambion/Applied Biosystems). Saliva samples were treated as indicated (Oragene/RNA protocol; DNA Genotek Inc., Ontario, Canada), and RNA was extracted from human cultured cells (RNeasy kit; Qiagen, Germantown, MD). RT-PCR assays were performed for human and mouse samples (Mint Kit [Evrogen, Moscow, Russia] or Transcriptor High Fidelity cDNA Synthesis Kit [Roche Diagnostics, Indianapolis, IN]). For tissue expression analysis, all reaction mixtures (50 µL) contained 10 µM each primer pair, 2 µM dNTPs, 1.5 mM MgCl₂, and 1 U polymerase (GoTaq; Promega, Madison, WI). Primer localizations are depicted in Figures 1A2 (human) and 1B2 (mouse), and the sequences are given in Supplementary Table S1, <http://www.iovs.org/lookup/suppl/doi:10.1167/iovs.10-7101/-/DCSupplemental>. *CERKL* was amplified using primers A and B for human and a and b for mouse (120 seconds at 94°C followed by 35 cycles of 94°C for 20 seconds, 60°C for 30 seconds, and 72°C for 30/20 seconds). *GAPDH* was used for normalization (120 seconds at 94°C and 30 cycles of 94°C for 20 seconds and 63°C for 120 seconds).

Analysis of the 5' and 3' UTRs of human and murine *CERKL* retina isoforms was performed, using either the Plug adaptor or oligo-d(T) primers (provided in the Mint Kit; Evrogen) paired with suitable *CERKL*-specific internal primers under the indicated PCR conditions. The characterization of alternatively spliced variants and promoters was performed using a combination of the internal primers located in different exons. The primers were designed to share the same amplification conditions: 120 seconds at 94°C followed by 40 cycles of 94°C

for 20 seconds, 58°C for 30 seconds, and 72°C for 90 seconds. All sequences have been submitted to GenBank (accession numbers are shown in Supplementary Tables S2A and S2B, <http://www.iovs.org/lookup/suppl/doi:10.1167/iovs.10-7101/-/DCSupplemental>).

Transfections and Recombinant Protein Expression and Immunodetection

For protein expression, HEK293T cells (2×10^5 cells) were seeded and transfected using reagent (Lipofectamine 2000; Invitrogen Life Technologies, Carlsbad, CA), according to the manufacturer's protocol. The recombinant constructs were obtained by cloning representative human cDNA isoforms (h2, h13, h18 in Figs. 1A and 4A) with and in-frame HA epitope fused at the C terminus into pcDNA 3.1 (Clontech Laboratories, Inc., Mountain View, CA). After 48 hours, cells were lysed with protein loading buffer $\times 1$ and boiled for 5 minutes. Protein lysates were loaded onto 12% SDS-PAGE gels that were transferred and analyzed by Western blot. Immunodetection was performed with a primary monoclonal anti-HA (1:1000) and HRP-conjugated anti-mouse secondary antibody (1:3000). Tubulin immunodetection was used as a loading control.

Immunohistochemistry on Mouse Retina Cryosections

Eyes from 8-week-old C57BL/6J mice were fixed in 4% paraformaldehyde (PFA) and 0.5% glutaraldehyde (2 hours at room temperature), cryoprotected in acrylamide, and embedded in OCT (Tissue-Tek, Sakura Finetek, Torrance, CA). Sixteen-micrometer sections on polylysine-covered slides were used for immunostaining, as described²⁰ with some modifications. Incubation with peanut agglutinin (PNA) conjugated to Alexa Fluor 647 (40 mg/mL; Invitrogen Life Technologies) and the primary antibodies mouse anti-rhodopsin 1:500 (Abcam), mouse anti-PKC α 1:500 (Santa Cruz Biotechnology, Inc., Santa Cruz, CA), and preabsorbed rabbit anti-CERKL 1:50, was performed overnight. Subsequently, slides were incubated with the corresponding secondary antibodies (1:300) conjugated to either Alexa Fluor 488, 546, or 568 (Invitrogen Life Technologies). Sections were mounted with reagent (Fluoprep; Biomérieux, Marcy l'Etoile, France) and photographed with a confocal microscope (SP2; Leica Microsystems, Wetzlar, Germany).

Bioinformatic Analysis of the Genomic Human CERKL Locus

Most of the computational analyses were performed using the genomic sequence of the human *CERKL* locus at chromosome 2 (March 2006 assembly version [NCBI36/hg18]) within the interval 182,029,864 bp to 182,259,440 bp (including the *ITG4* and *NEUROD1* loci), which was retrieved from the UCSC human genome browser.²¹ However, for the purpose of comparative genomics and to determine the conservation among human and other vertebrates (such as *Macaca mulatta*, *Mus musculus*, *Gallus gallus*, and *Takifugu rubripes*), precomputed whole genome alignments were analyzed through the VISTA UCSC browser mirror, which provides the VISTA track feature.²² The syntenic region of the mouse genome was also retrieved. BLASTN and TBLASTX alignments were performed on the syntenic sequences using the NCBI bl2seq algorithm²³ for a more in-depth comparison between human and mouse.

Previously described *CERKL* isoforms were retrieved from several databases: RefSeq,²⁴ GenBank,²⁵ dbTSS,²⁶ and VEGA.²⁷ 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,²⁸ following the est2genome model algorithm for easier comparison of all the exonic structures from both the database and experimental evidence (complete visualization is shown in Supplementary Fig. S1, <http://www.iovs.org/lookup/suppl/doi:10.1167/iovs.10-7101/-/DCSupplemental>).

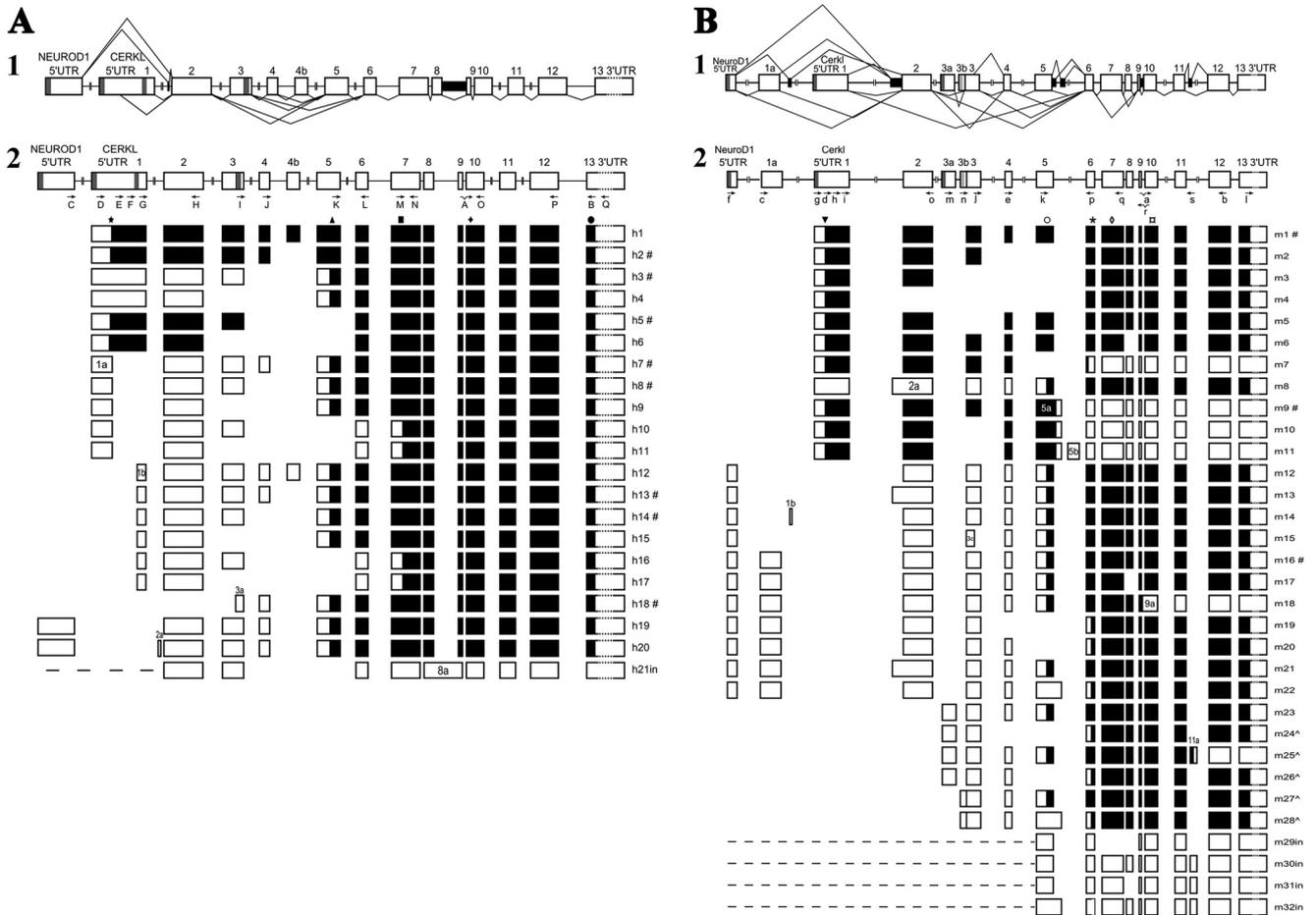


FIGURE 1. Alternately spliced *CERKL* isoforms in human and mouse retina. Extremely high complexity of the splicing events in human (**A1**) and mouse (**B1**) *CERKL* transcripts. *Open boxes*: exons. *Filled boxes*: retained introns or cryptic noncoding exons. *Angled lines* above and below the gene structure indicate validated splicing events. Scheme depicting all the human (**A2**) and mouse (**B2**) spliced variants observed in the retina. Exons are indicated as *boxes* and the coding sequence (CDS) for each isoform, considering the higher likelihood of first methionine, is shown in *black*. *Dark gray*: TSS found in retina. *Light gray*: nonretinal TSS. *Main isoforms in each species. *Arrows*: letters indicate the position and direction of the primers used for PCR reactions (complete list and sequence in Supplementary Table S1, <http://www.iovs.org/lookup/suppl/doi:10.1167/iovs.10-7101/-DCSupplemental>). ^Nonretinal 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. For mouse they are: ▼ 13.384; ○ 9.620; *10.662; ◇ 8.389; ✕ 8.863 (the complete list of all Kozak's scores are contained in Supplementary Table S5, <http://www.iovs.org/lookup/suppl/doi:10.1167/iovs.10-7101/-DCSupplemental>).

Although a track for the First-Exon-Finder program²⁹ on the UCSC genome browser was already available, an additional attempt was performed to predict more CpG islands, promoters, and first exons on the *CERKL* genomic region (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).

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). Up to 10-kbp upstream sequences of these genes were searched for overrepresented motifs by running MEME.³⁰ The first analysis was performed over the whole set of sequences; then, in the following round, MEME was run separately for the sequences of each functional class. Two sets of

parameters were used to characterize long and short motifs. To search for long motifs, the "anr" model was used, with a minimum width of 8 and a maximum width of 20, and a total of 200 iterations (-mod anr -n motifs 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 -n motifs 10 -minw 8 -maxw 10 -maxiter 200). Both sets of parameters were applied to the whole data set analysis and to the split group consisting of 10 different functional classes. For each characterized motif, a log likelihood 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 neighboring loci). We extended the analysis to the promoters (1 kb upstream) of the cone-rod dystrophy (CRD) genes using the TRANSFAC matrices with particular emphasis on the retina-related transcription factor. The genes were grouped according to the disease to which they contributed most: RP (*RHO*, *PDE6A*, *PDE6B*, *CNGA1*, *CNGB1*, *SAG*, *GUCA1B*, *LRAT*, *RPE65*, *RLBP1*, *RGR*, *RDH12*, *RBP3*, *ROM1*, *EYS*, *NR2E3*, *NRL*, *OTX2*, *MERTK*, *CRB1*, *USH2A*, *PRPF3*, *PRPF8*, *PRPF31*, *RP9*, *SNRNP200*, *TULP1*, *RPGR*, *RP2*, *FSCN2*, *RP1*, *CEP290*, *LCA5*, *KLHL7*, *TOPORS*, *IMPDH1*, *CA4*, *IDH3B*, *RD3*, *SPATA7*, and *PRCD*), CRD (*GUCA1A*, *PITPNM3*, *RIMS1*, *UNC119*, *ADAM9*, *CACNA2D4*, *RAX2*,

CDHRI, and *CACNA1F*), and the two retinal disorders (*ABCA4*, *GUCY2D*, *PROM1*, *PRPH2*, *CRX*, *SEMA4A*, *RPGRIP1*, *AIPL1*, and *CERKL*). Unfortunately, no clear pattern of single/clustered transcription factor sites emerged considering any of the three gene groups, either on general or on retina-specific transcription factor matrices (Supplementary Fig. S2, <http://www.iovs.org/lookup/suppl/doi:10.1167/iovs.10-7101/-DCSupplemental>).

In addition to those generated by MEME, a new set of matrices corresponding to a selection of known transcription initiation factors (including TATA, CAAT, USF, INI, SRF, SP1, and TFIIA) was downloaded from TransFac.³¹ 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,³² and Jaspar.³³ All the matrices were mapped into the analyzed genomic region of *CERKL* using custom Perl scripts with the specific purpose of defining potential novel alternative transcription starting sites (TSS) for *CERKL* isoforms. The score 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 hits of matrices showing a normalized score equal to or greater than the threshold were considered (a summary of those found on the 1 kbp upstream for every reported human and mouse *CERKL* exons that included a TSS is provided on Supplementary Tables S4A (human) and S4B (mouse), <http://www.iovs.org/lookup/suppl/doi:10.1167/iovs.10-7101/-DCSupplemental>).

Putative translation start sites were evaluated using the Kozak matrix³⁴ under the same terms. Moreover, the ENCODE H3K4Me3 track³⁵ on the UCSC genome browser was also considered as additional transcriptional evidence, given that histone modification correlates with transcriptionally active sites.³⁶ The distribution of SNPs across the exons of the *CERKL* gene was analyzed using dbSNP31, over the hg19 database.

RESULTS

Comprehensive Identification of Alternatively Spliced *CERKL* Isoforms

Evidence of different alternatively spliced isoforms of *CERKL* have been reported, but a comprehensive prioritized list of the physiologically relevant transcript is still missing.^{19,37} Furthermore, its wide tissular expression^{17,18} appears to be inconsistent with the tissue-restricted phenotype of *CERKL* mutations because only the retina was affected. In this case, as happens with other retina-associated disease genes, tissue-specific isoforms might have reconciled this apparent paradox.¹⁴

Thus, we first aimed to exhaustively characterize the *CERKL* alternatively spliced isoforms generated in human and murine retinas and to perform an interspecific comparative analysis. Two different methods for the synthesis of the cDNAs (detailed description in Materials and Methods) 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 on endogenously expressed retinal transcripts and subsequently used a battery of internal PCR primers (listed in Supplementary Table S1 [<http://www.iovs.org/lookup/suppl/doi:10.1167/iovs.10-7101/-DCSupplemental>] and located in Figs. 1A2 [human] and 1B2 [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 the first to the last exon and thus depict the complete repertoire of *CERKL* 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 because >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 (gray boxes) identified are shown in Figures 1A (human) and 1B (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 TSS that 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 Supplementary Tables S2A and S2B (<http://www.iovs.org/lookup/suppl/doi:10.1167/iovs.10-7101/-DCSupplemental>).

In detail, the analysis of the 20 fully validated human transcripts provided solid evidence of four different *CERKL* TSS (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 to as exon 1b in the text and Supplementary material, <http://www.iovs.org/lookup/suppl/doi:10.1167/iovs.10-7101/-DCSupplemental>); and one started from an internal sequence of exon 3 (referred to as exon 3a). Of note, the TSS of exon 1b was also supported in silico by the First-Exon-Finder, which, among other structural features, mapped a CpG island within this genomic region, and by the clustering of peaks of the H3K4Me3 track, indicative of transcriptionally active chromatin sites (Fig. 2). Yet we cannot rule out that *CERKL* is transcribed from unknown TSS in other tissues. In this context, the UCSC genome browser has recently incorporated an ENCODE track that corresponds to manually annotated genes, based primarily on sequenced full-length cDNAs from dbTSS plus reports from independent sources. Twelve 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 completely different internal sites, suggesting two additional TSS. If the latter two isoforms were validated, the number of *CERKL* TSS in human would amount to six.

In contrast, in the murine retina, only three *Cerkl* start sites were experimentally identified (Fig. 1B, dark gray): 11 (of 23) fully validated transcripts started from the previously reported *Cerkl* site, 11 from the upstream *NeuroD1* gene (as in human), and the last 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 the retina). A complete list specifying the contribution (presence or absence) of every exon in each *CERKL/Cerkl* isoform is presented in Supplementary Table S3, <http://www.iovs.org/lookup/suppl/doi:10.1167/iovs.10-7101/-DCSupplemental>.

To identify the more abundant transcripts and approach their relative physiological relevance (Figs. 3A [human] and 3B [mouse]), we used a battery of primers, located either at the different TSS or the alternative exons at the 5' of *CERKL*, paired with a unique reverse primer in exon 10 (human) or exon 12 (mouse). The location of the primers is indicated in Figure 3C. For isoform assignment, each amplified product was isolated and sequenced. The RT-PCRs were replicated several

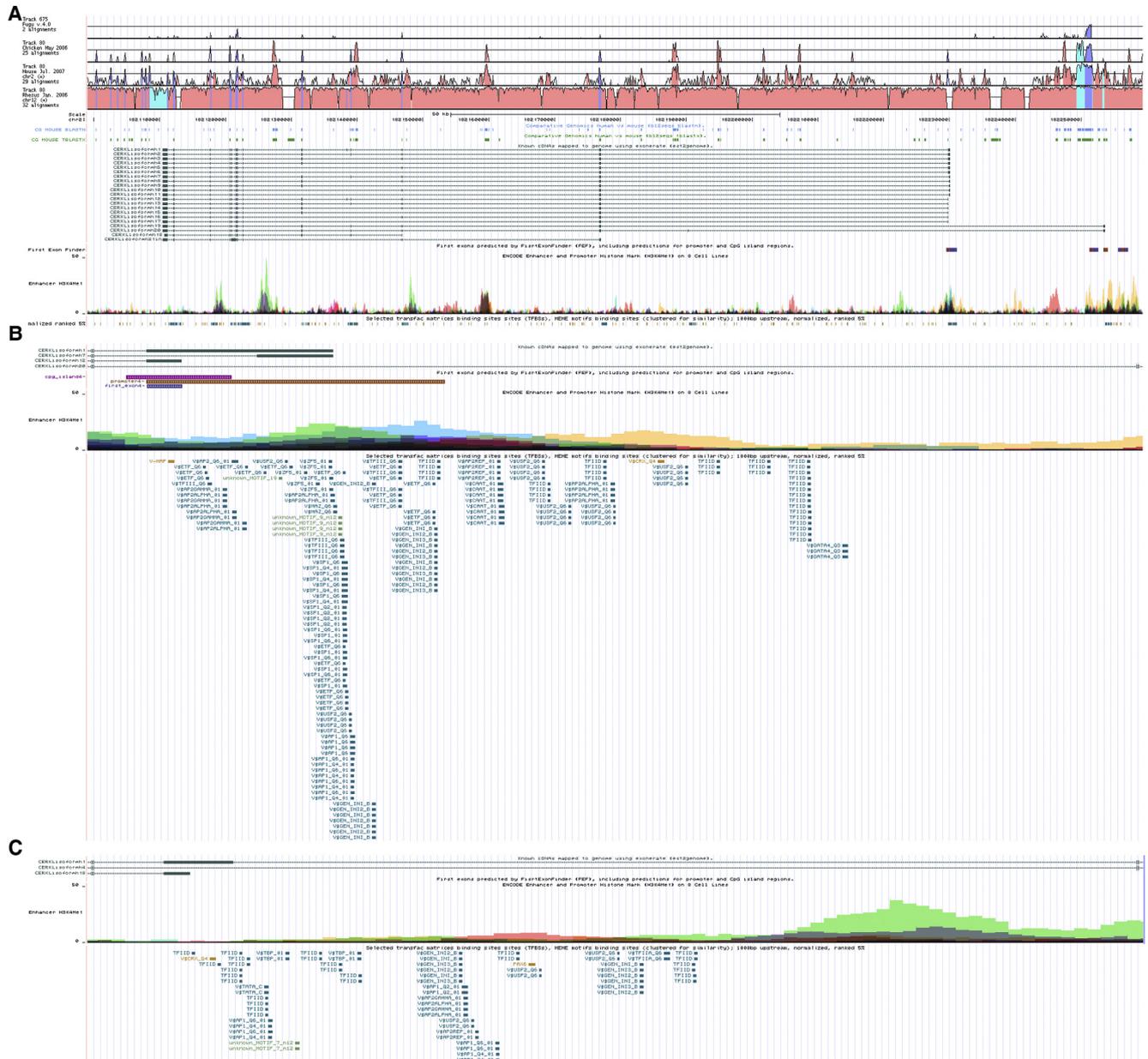


FIGURE 2. Summary of annotated and custom feature tracks on the UCSC genome browser. **(A)** An overall view of the whole genomic neighborhood of human *CERKL*, including upstream *NEUROD1* (*ITGA4*) downstream gene is shown in Supplementary Fig. S1, <http://www.iovs.org/lookup/suppl/doi:10.1167/iovs.10-7101/-/DCSupplemental>). Homology to various species, including mouse, is depicted on the topmost tracks. Exonic structure of all the experimentally validated *CERKL* isoforms described in this article. 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, and some further evidence of transcriptional activity on neural tissues are shown. **(B, C)** Magnifications of the regions around exons 1 and 3, respectively, containing a more detailed view of the TFBS sites. The same track distribution is depicted on all three panels. Matrix hits overlapping homopolymer stretches larger than 5 bp were discarded.

times. The interspecific comparison of the more abundant transcripts in the retina revealed a higher number of *CERKL* variants in human (8 of 20 transcripts, with a comparable level of expression) than mouse (3 of 23 transcripts, with one major variant).

Concerning the *CERKL*/*Cerkl* protein isoforms, our data reveal that the combination of TSS multiplicity with the high number of alternative splicing events affecting the first exons (exons 1–6) generates a complex pattern of mature transcripts that differ at the 5' end but share the 3' moiety (exons 6–13), as shown in Figures 3A (human) and 3B (mouse). The alternative 5' exons encode the nuclear localization signals,^{18,38} the

putative pleckstrin homology (PH) domain, and the diacylglycerol kinase (DAGK) signatures.^{17,18,38,39} In addition, the human gene includes an in-frame species-specific alternative exon (4b) embedded in the predicted DAGK domain that interrupts the DAGK consensus signature. The comparison of 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 (Figs. 1A2, 1B2, isoforms m9, m10, m11). These transcripts bear premature stop codons and may be candidates to be degraded by the nonsense-mediated decay mechanisms (NMD) but, if translated, would encode a C-terminal-truncated protein.

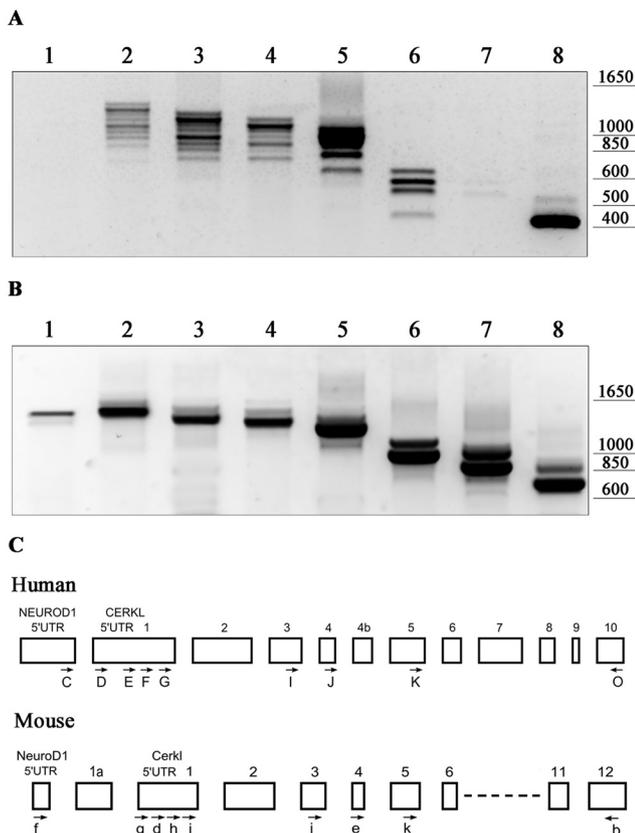


FIGURE 3. Evaluation of *CERKL* main transcripts. RT-PCR from human (A) and mouse (B) retina total RNA, to identify the main isoforms. (C) Scheme depicting the structure of *CERKL* in human and mouse, with the location of the primers used to generate the PCR reactions. For the sake of clarity, exons not relevant to this assay are not shown. For all amplicons, the same reverse oligonucleotide (human: O; mouse: b) was used, paired with the corresponding forward primers. For human: lane 1, C; lane 2, D; lane 3, E; lane 4, F; lane 5, G; lane 6, I; lane 7, J; lane 8, K. For mouse: lane 1, f; lane 2, g; lane 3, d; lane 4, h; lane 5, i; lane 6, j; lane 7, e; lane 8, k. Primer sequences are provided in the Supplementary Table S1, <http://www.iovs.org/lookup/suppl/doi:10.1167/iovs.10-7101/-DCSupplemental>.

Evidence for *CERKL* Alternative Translational Initiation Sites

Interestingly, one of the consequences of the use of alternative TSS is that the previously reported initiation Met codon is not always included in the mature transcript. Then, additional translation initiation sites (TIS) should be considered. In silico sequence analyses using motif searches with a Kozak matrix predicted several TIS along the *CERKL* transcripts (Supplementary Table S5, <http://www.iovs.org/lookup/suppl/doi:10.1167/iovs.10-7101/-DCSupplemental>). Of these, only two encoded long peptide sequences, whereas the remaining putative TIS yielded a lower score value or would generate very short peptides. Initiation codons with significant TIS scores are indicated in Figures 1A2 and 1B2. For each isoform, only the longest open-reading frame starting with a high-score Met is depicted (filled boxes).

As proof of principle, we tried to express three human highly expressed isoforms (h2, h13, and h18) harboring different in-frame methionines with a high Kozak score. The h2 encompassed the complete *CERKL* sequence, starting at the previously described 5'UTR, whereas the h13 and h18 cDNAs started at different TSS. The two latter did not con-

tain the first in-frame methionine in exon 1, but they both shared an in-frame Met residue at exon 5 having a high Kozak score. Of note, other out-of-frame methionines located upstream in exon 5 showed comparable Kozak values (Fig. 4A). For each construct the *CERKL* coding sequence was fused at the 3' end to an HA epitope to facilitate protein immunodetection. HEK293T cells were transfected with each construct, and RT-PCR was performed to assess the level of the recombinant *CERKL* transcription. Notably, we observed a high yield of the *CERKL* protein from 2 of 3 constructs (h2 and h18), each starting from the corresponding highlighted high Met score (Figs. 4B, 4C). Indeed, the size of the expressed *CERKL*-HA proteins was in agreement with their expected molecular mass (60 and 32 kDa).

Exploring the Promoter Landscape of *CERKL* TSS

To shed light on the architecture of the *CERKL* promoters and to define in silico potential novel alternative TSS, 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 reported general transcription initiation motifs, retina-related transcription factors, and matrices obtained by MEME after analysis of the 49 promoters of RP genes to underscore conserved retina-specific regulatory motifs (subfunctionalized MEMEs) (for a detailed description of these analyses, see Materials and Methods). The outcome of this search along the upstream sequences of every exon depicted three different scenarios that corresponded to the patterns yielded by exons with a TSS function in retina (NeuroD1, 5'*CERKL* UTR, 1b, and 3a), exons with TSS not found in the retina (corresponding to the starting exons in the ENCODE transcripts OTTHUMT00000334817 and OTTHUMT00000334818), and the remaining internal exons, which are not used as TSS (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 or V-MAF (used to detect NRL-binding sites), although some were scattered along the *CERKL* genomic region. Overall, the evidence points to distinct promoter architecture concerning TSS, probably reflecting tissue-specific expression. Supplementary Tables S4A and S4B (<http://www.iovs.org/lookup/suppl/doi:10.1167/iovs.10-7101/-DCSupplemental>) show the detailed list of TFBS, MEME, and subfunctionalized MEME hits upstream of each exon.

Given that *CERKL* mutations also contribute to CRD, we extended the analysis to the promoters (1 kb upstream) of the CRD genes using the TRANSFAC matrices, with particular emphasis on the retina-related transcription factor. The genes were grouped according to the disease to which they contributed most: RP (already listed), CRD, and a group of genes involved in both retinal disorders. Unfortunately, no clear pattern of single/clustered—general or retina-specific—transcription factor sites emerged in any of the three gene groups (Supplementary Fig. S2, <http://www.iovs.org/lookup/suppl/doi:10.1167/iovs.10-7101/-DCSupplemental>).

Genomic Conservation of the *CERKL* Region among Vertebrates

VISTA tracks on Figure 2 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

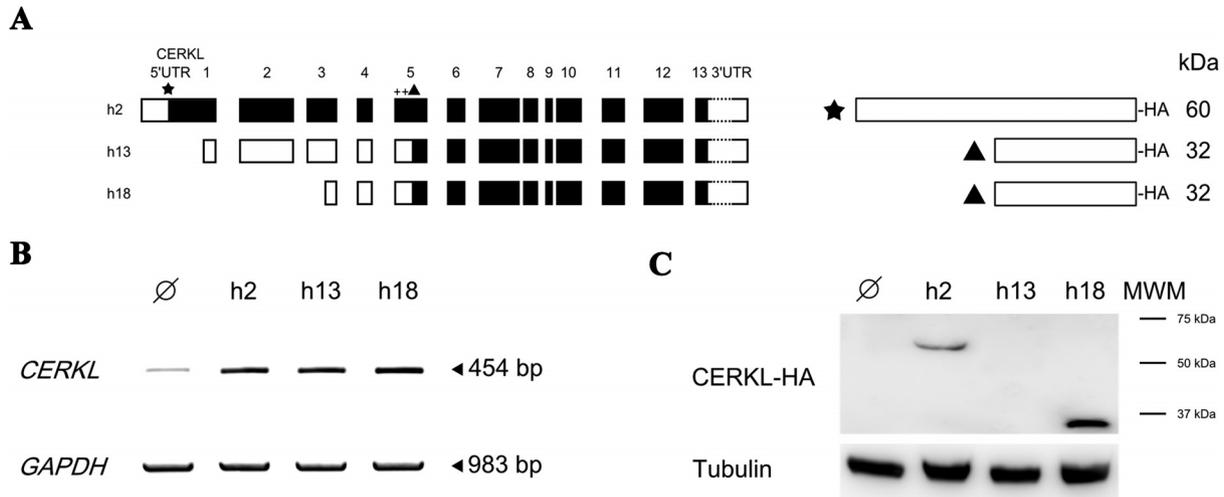


FIGURE 4. Evidence for additional initiating methionines in alternatively spliced human *CERKL* isoforms. (A) Diagram of the three different HA-tagged constructs from isoforms h2, h8, and h13, as well as the structure and molecular mass of the predicted encoded proteins. Methionines showing high Kozak scores are indicated by an *asterisk* (methionine in exon 1) and a *filled triangle* (internal methionine in exon 5), whereas other out-of-frame significantly scored Met are marked with a *cross*. *Filled boxes*: putative CDS. (B) RT-PCR showing expression of the *CERKL* constructs in transfected HEK293T cells. Lower endogenous *CERKL* levels were also detected in nontransfected cells (\emptyset). *GAPDH* gene was used for normalization. (C) *CERKL*-HA-fused proteins were immunodetected with an anti-HA monoclonal antibody. α -Tubulin was used as a loading control.

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 because it is unique to the human genome. The comparison with *fugu* reveals an expected lower degree of conservation because only *NEUROD1* exons rank above 70% whereas 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, the *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 the translated (TBLASTX) levels. Thus 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

Semiquantitative RT-PCR analysis of *CERKL* expression was performed in a collection of tissues and cell lines of 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; exon 12 in mouse; see Figs. 1A2 and 1B2 for locations; details are given in Materials and Methods). The results are shown in Figures 5A and 5C (human) and 5B and 5D (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 in which *CERKL* expression was 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 is the case with HEK293T and A549 (Fig. 5C).

In mouse, *Cerkl* was also highly expressed in the retina, although the liver showed even slightly higher levels of expression (Fig. 5D). 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 because it retained a noncoding fragment of intron 11. Other mouse tissues, such as testis and spleen, also showed high to moderate levels of *Cerkl* expression.

As mentioned, in addition to the reported mouse *Cerkl* promoter (heretofore, UTR), retinal transcripts were produced from the *NeuroD1* promoter and an internal TSS in intron 2 (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 TSS in the retina: UTR, *NeuroD1*, 3a and 3b to *Cerkl* expression (Fig. 6). Tissue comparison showed a wide range of expression from each TSS: the *Cerkl* UTR contribution was indeed major in the retina, moderate in the kidney, faint in the brain, and undetectable in the blood and spleen. In addition, in

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 TSS (<i>NEUROD1</i> , 1/1a, 1b, and 3a)
2	<35	≈10	>75	Nonretinal TSS (OTTHUMT00000334817 first exon and OTTHUMT00000334818 first exon)
3	≈25	0	<5	No TSS exons

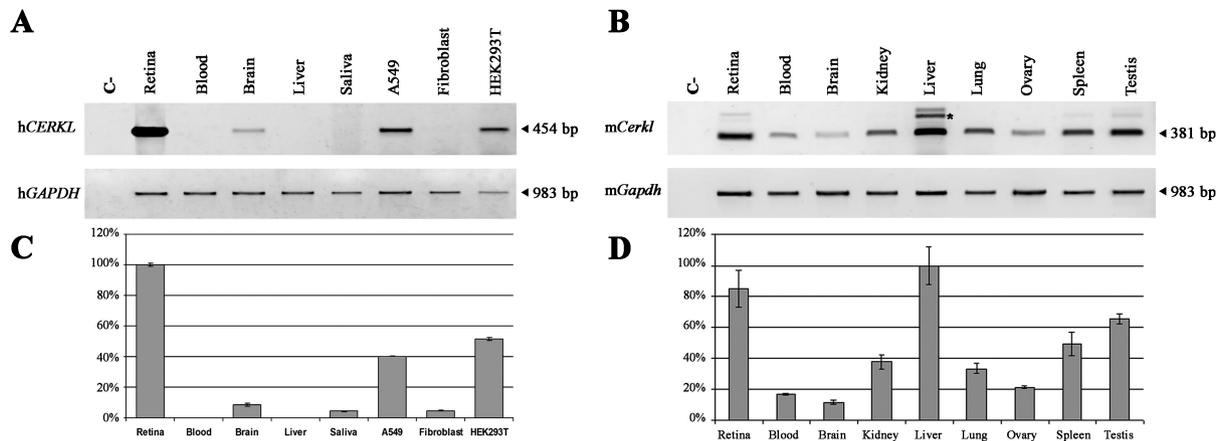


FIGURE 5. *CERKL* semiquantitative expression analysis in human and mouse tissues. *CERKL* expression identified by RT-PCR in several tissues and cell lines of human (A) and mouse (B) origin. Semiquantitative analysis of all *CERKL* transcripts in human (C) and mouse (D). At least three replicates were 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 primers a and b in mouse, as located in Figures 1A2 and 1B2. The amplicon size is indicated in each case. The asterisk in the murine liver sample corresponds to the alternative isoform m30in. Primer sequences are provided in Supplementary Table S1, <http://www.iovs.org/lookup/suppl/doi:10.1167/iovs.10-7101/-DCSupplemental>. 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 TSS in each tissue.

agreement with previous reports, *NeuroD1*-driven expression was tissue restricted and was observed only in the retina in our panel. In contrast, the 3a TSS-driven transcript was expressed more widely but showed very low levels in the retina. Although the 3b TSS was silent in the mouse retina, it was the most active in the liver (Fig. 6). Isoforms m24 to m28 in Figure 1B, which started at either 3a or 3b TSS, were isolated and sequenced in the spleen and liver but were undetectable in the retina. Of note, in some tissues, the RT-PCRs specific for these four promoters did not explain the total *Cerkl* transcriptional levels (as revealed by the amplification of the 9 to 12 exon region common to all isoforms), again pointing to additional TSS.

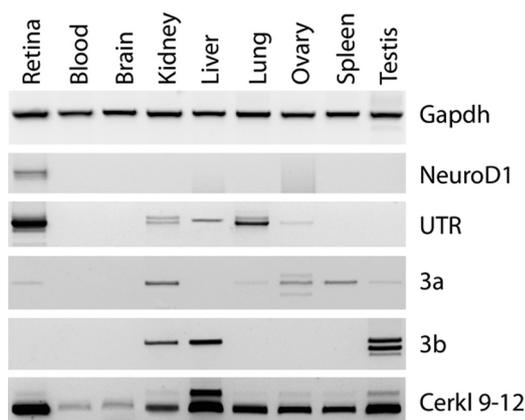


FIGURE 6. Tissue-specific *Cerkl* promoter in adult mice. RT-PCRs were performed on several murine samples to determine the active promoters in each tissue. Forty-five cycle amplifications were carried out using the same reverse oligonucleotide in exon 12 and different forward primers located in each TSS 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 in Figure 1B, and sequences are listed in Supplementary Table S1, <http://www.iovs.org/lookup/suppl/doi:10.1167/iovs.10-7101/-DCSupplemental>.

Cerkl Localization in Mouse Retina by Immunohistochemistry

Previous results based on in situ mRNA hybridization showed that *Cerkl* was expressed mainly in the ganglion cell layer, though 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 sagittal 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 preabsorbed before use. Double coimmunodetection with this polyclonal anti-*Cerkl* antibody and either anti-rhodopsin (specific for rods) or anti-PKC α (which primarily labels bipolar cells and rods), plus counterstaining with DAPI (nuclei) and Alexa Fluor 647-conjugated PNA (which labels cones) were performed in parallel to allow a more detailed localization (Fig. 7).

Cerkl expression was found at the ganglion cell layer (GCL), in the photoreceptors (PhR), and in some cell bodies at the outer nuclear layer (ONL) and inner nuclear layer (INL) (Fig. 7). Magnification of the photoreceptor cell layer showed a strong immunodetection of *Cerkl* in cones and, faintly, in rods. Of interest, *Cerkl* localized primarily in the outer segments of both types of photoreceptors, as shown by its colocalization with rhodopsin (rods) and cone (Figs. 7H, 7I) staining. In addition, *Cerkl* showed perinuclear staining in some cell bodies at the ONL, extremely close to the photoreceptor layers, probably corresponding to cones (Figs. 7I, 7J, white arrows). Concerning other neuronal retinal types, *Cerkl* was detected in a population of bipolar cells (white arrowheads in Fig. 7N) 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 amount of evidence

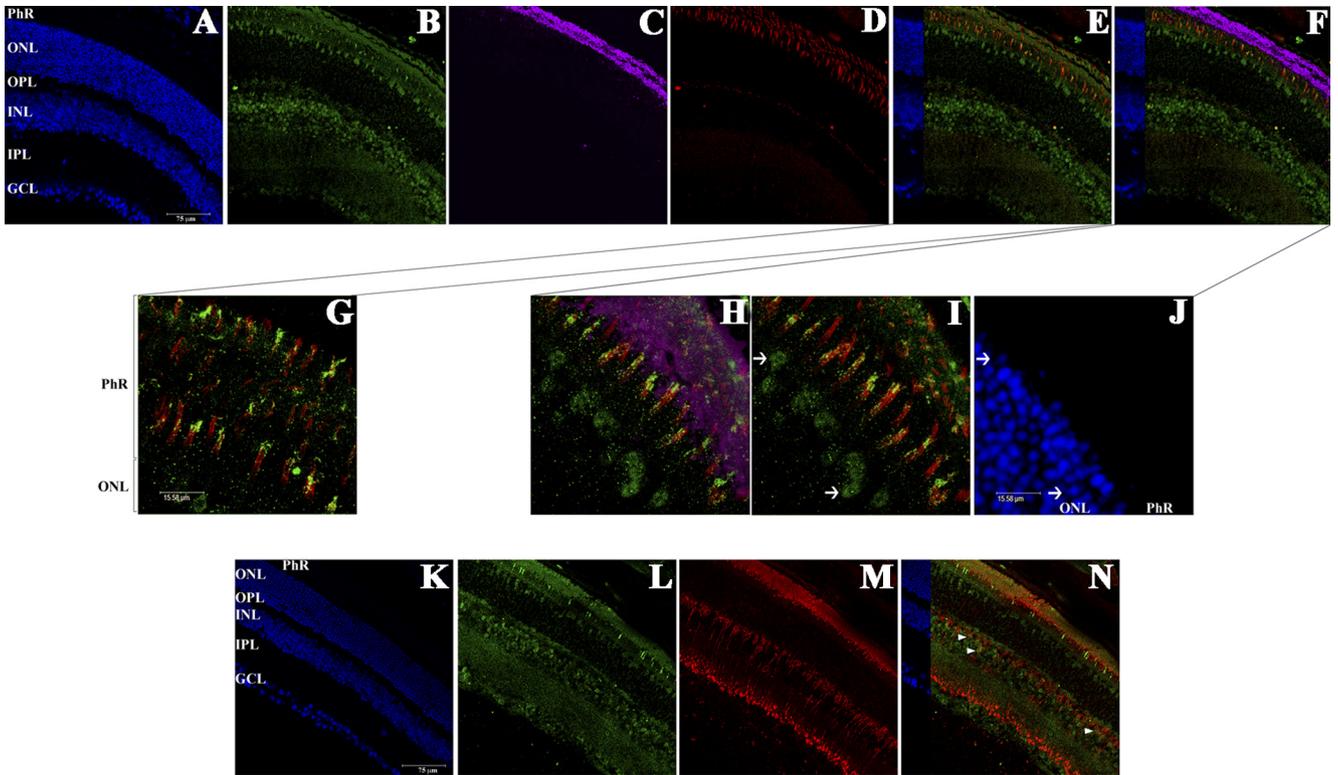


FIGURE 7. Immunohistochemistry on mouse retina cryosections. (A–J) Localization of Cerkl in photoreceptor cells. Nuclei are stained with DAPI (blue, A); Cerkl (B) and Rhodopsin (C) proteins are detected in green and magenta, respectively; cones appear in red (D) using PNA staining. Two merged images (E, F) and the magnification of some sections show clear localization of Cerkl in cones (yellow, G) and, 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 (I, J). (J) DAPI counterstaining of the nuclei. (K–N) Expression of Cerkl in other retinal layers. Nuclei are stained with DAPI (blue, K), Cerkl protein is detected in green (L), bipolar cells and rods expressing PKCa are labeled in red (M). Cerkl is expressed in the ganglion cells (GCL), some cells in the INL and ONL, and in the photoreceptors. The merged image (N) shows expression of Cerkl in some bipolar cells (white arrowheads) while confirming localization in rods. Scale bars show magnifications.

in the human genome supports that alternative splicing is more the rule than the exception because >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 is intended to characterize the regulatory splicing strategies on a genome-wide scale to predict the specific transcripts from every gene.^{40,41} However, these in silico predictions must be substantiated in vivo to identify the physiologically 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 show unexpectedly high transcriptional complexity in human and mouse tissues arising from the combination of tissue-specific promoters and alternative splicing

events, particularly in the retina. A large multiplicity of retina transcripts has also been reported for other genes, such as *RPGR*, *RPGRIP1*, and *CPEB3*^{42–44}. In agreement with these results, a recent accurate transcriptional characterization focused on the PRPF gene family (proteins associated with spliceosome formation and responsible for retinal dystrophies) showed that the processed pre-mRNA levels were highest in the retina than in other tissues and organs. Their results pointed to a particularly increased splicing activity at the base of the high multiplicity of retinal transcripts and called for sophisticated quality control mechanisms.⁴⁵

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, whereas exons 3 to 7 encompass the DAGK domain^{17–19,38,39} (Fig. 8). Notably, the use of the different promoters and 5'UTRs affects the inclu-

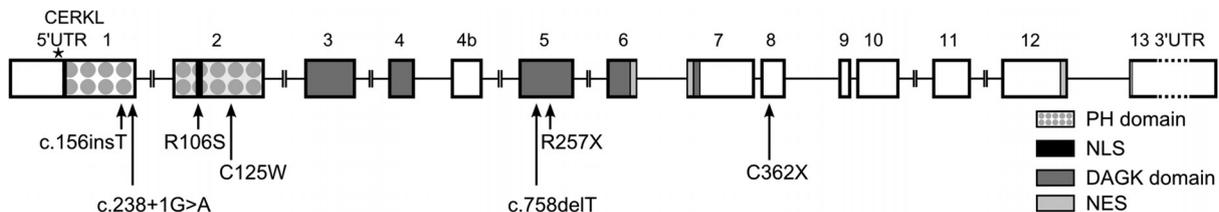


FIGURE 8. Scheme of the reported causative mutations on the *CERKL* gene. The location of the mutations identified thus far is shown on a diagram of the *CERKL* protein. The *CERKL* domains described by either sequence homology (PH, pleckstrin; DAGK, diacylglycerol kinase domain) or functional analysis (NLS, nuclear localization signals; NES, nuclear export signals) are also depicted.

sion/exclusion of the first exons in the final transcript and generates 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 nonretinal 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). For example, the *NeuroD1* promoter contributes to the highly expressed isoforms in mouse, whereas its relevance in the human adult retina appears to be minor. This holds true for the least abundant isoforms (e.g., h1, h12, h15-h17 and m5, m7, m8, m10, m11, m13) (Fig. 1A2, 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*, *OPAI1*, and *PRPF31*, suggesting distinct functional requirements for each species.⁴⁶⁻⁴⁸ Remarkably, one-third of the murine isoforms (12 of 32) compared with 1 of 21 human isoforms are produced by missplicing (with partial retention of intron sequences). Most of these misspliced transcripts would encode a truncated protein, unless degraded by NMD. Other reports analyzing human versus murine transcripts identified other retinal dystrophy genes with preferential or unique intron retention in the mouse, among them *RPGR* (intron 14),⁴⁹ *RPGRIP1* (intron 13),⁴² and *PRPF31* (intron 7).⁴⁸ 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, at least in retina; even though it has been shown that relevant splicing events associated with NMD remain conserved through mammalian genomes, reflecting a common clearance mechanism of transcripts that might compromise cell viability.⁵⁰

One of the relevant findings of our work is the use of tissue-specific TSS in mouse. Among the tissues analyzed, the *NeuroD1* promoter was only active in the retina, where the reported *Cerkl* UTR promoter also showed the highest transcriptional activity. Instead, the additional alternative internal promoters were highly expressed in nonneuronal tissues (Fig. 6, liver, testis, kidney). The combination of different promoters and shared splicing events in both species hindered 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 (Figs. 3, 5, 6). Of note, the retina is the tissue in which higher expression and greater display of *CERKL* transcript variability is observed. In addition to the multiplicity of promoters and alternative splicing events, another layer of complexity is provided by other in-frame methionines, which direct the synthesis of shortened *CERKL* protein isoforms, with a downstream start in exon 5. In vitro experimental evidence strongly supports this starting Met in the h18 isoform, though no expression could be detected for the h13 variant. Whether this apparent discrepancy could be explained by other upstream out-of-frame methionines in h13 (not present on the h18 alternative 5'UTR exons) that affect the translational initiation complex formation remains to be elucidated (Fig. 4). Although the *CERKL* function is as yet undetermined, such a high repertoire of transcripts and proteins, while making functional assignment a real challenge, hints at a very crucial role in retinal cell survival. Indeed, in a more general view, these results open new scenarios for the human proteome complexity associated with a multiplicity of isoforms.

The in silico analysis of binding sites for transcription initiation factors across the *CERKL* genomic neighborhood (approximately 230 kbp) revealed a high number of hits (>15,000). However, they were not randomly distributed but were clustered just upstream of *ITGA4*, *NEUROD1*, and *CERKL* canonical TSS. If we focus on the retina-specific TFBS, no significant scores for OTX2 or NR2E3 could be found upstream of the promoters of these genes. In contrast, binding sites for CRX, PAX6, and NRL upstream of *NEUROD1* TSS, for NRL in *CERKL* exon 1b, and for PAX6 and CRX upstream of exon 3 TSS were identified. These results provide evidence for retina-specific regulatory enhancers close to *CERKL*. Overall, the differential patterns observed for the in silico predicted enhancers, the TSS experimentally confirmed in the retina, and the identification of nonretinal transcriptional products clearly support a highly tuned, tissue-specific regulation of *CERKL* expression.

Notably, *Cerkl* immunohistochemistry showed high expression in cones and moderate expression in rods, ganglion cells, and other retinal INL cell types. A specific perinuclear staining was observed at the INL and ONL. Hitherto, *CERKL* mutations have been associated with both conventional RP and 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.

To establish a more precise relationship between mutations and the relative pathogenicity of each isoform, the distribution of SNPs along the coding gene structure was analyzed in silico. A priori, a homogeneous distribution of both mutations and SNPs should be expected when all the exons and encoded domains contribute equally to function. The results of this analysis showed that not all the domains harbored the same frequency of SNPs because some showed higher SNP frequencies than the observed average, whereas others were devoid of polymorphic variants, thus suggesting differential selection pressures (Supplementary Fig. S3, <http://www.iovs.org/lookup/suppl/doi:10.1167/iovs.10-7101/-/DCSupplemental>). For instance, the alternately spliced exons that encompassed the DAGK domain contained fewer SNPs, whereas the exons that encoded the pleckstrin homology domain showed more SNPs than average. The biological meaning of this differential distribution remains to be assessed.

Meanwhile, as more mutations are being identified, a genotype-phenotype correlation pattern is emerging (Fig. 8, 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 1 of the 8 major isoforms remains unaffected after alternative splicing. The phenotype associated with this variant ranges from canonical RP to more severe CRD features.⁵³ Another RP-associated mutation, p.R106S, is localized in 1 of the 2 putative nuclear localization signals, probably compromising its import and function in the nucleus.⁵⁵ However, all other protein domains remain unaltered, in accordance with a moderate RP phenotype. Other alleles are associated with more severe retinal disorders, with clear cone-rod dystrophy features and early macular degeneration. One of them, c.238+1G>A,³⁷ affects the splicing of the first intron, abrogating the generation of the putative protein isoforms produced from exon 1 and 1b. Thus, only the isoforms starting in exon 3 or the spliced variants of exon 1a would be produced. The other mutation, p.C125W⁵⁴ (also affecting the conformation of the protein isoforms encoded from the methi-

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 n (%)	References
p. R257X/exon 5	Lipid kinase/protein truncated	Homozygous and compound heterozygous	7 of 8 (h2, h3, h7, h8, h13, h14, h18)	RP, with some patients showing phenotypes closer to CRD; peripheral pigment deposits plus macular dystrophy	30/40 alleles (20 families or patients) (~75%)	Tuson et al., ¹⁷ Pomares et al., ⁵¹ Avila-Fernandez et al., ⁵² Aleman et al., ⁵³ Littink et al. ⁵⁴ ‡
c.238+1G>A/intron 1	Pleckstrin homology/abrogates splicing	Homozygous	5 of 8 (h2, h3, h5, h13, h14)	Mixed features of RP and CRD, with early macular degeneration	2/40 alleles (~5%)	Auslender et al. ³⁷
p.R106S/exon 2	Nuclear localization signal/compromises nuclear import	Homozygous	3 of 8 (h2, h3, h5)	RP features (bone-spicules) with CRD leading to peripheral and central vision deficit	2/40 (~5%)	Ali et al. ⁵⁵
c.156_157ins/exon 1	Pleckstrin homology/frameshift and protein truncation	Compound heterozygous with C.758delt	3 of 8 (h2, h3, h5)	NC	1/40 (~2.5%)	Tang et al., ⁵⁶
c.758delt/exon 5	Lipid kinase/frameshift and protein truncation	Compound heterozygous with C.156_157/ins	7 of 8 (h2, h3, h7, h8, h13, h14, h18)	NC	1/40 (~2.5%)	Tang et al., ⁵⁶
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 (~2.5%)	Aleman et al. ⁵³
p.C125W/exon 2	Pleckstrin homology/evolutionarily conserved residue	Homozygous	3 of 8 (h2, h3, h5)	CRD (with central scotoma and macular atrophy, retinal thinning)	2/40 (~5%)	Littink et al., ⁵⁴

* 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).

onine 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 also been reported, but their association with 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 or functional specificity, or both, at the subcellular level remains a key challenge to understand the contribution of this gene to retinal degeneration.

Acknowledgments

The authors thank Andrés Mayor (Fundaluce, Hospital Central de Asturias) for the generous support; Ana Méndez-Zunzunégui (IDIBELL, Universitat de Barcelona) for generous support and technical advice on the use of eye cryosections; and the ENCODE project for making publicly available, through the UCSC Genome browser, the H3K4Me3 and the GENCODE manual gene annotations (including VEGA) tracks.

References

- 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.
- 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.
- Nilsen TW, Graveley BR. Expansion of the eukaryotic proteome by alternative splicing. *Nature*. 463:457–463.
- Licalosi DD, Darnell RB. RNA processing and its regulation: global insights into biological networks. *Nat Rev Genet*. 2010;11:75–87.
- McGlinchy NJ, Smith CW. Alternative splicing resulting in nonsense-mediated mRNA decay: what is the meaning of nonsense? *Trends Biochem Sci*. 2008;33:385–393.
- Tazi J, Bakkour N, Stamm S. Alternative splicing and disease. *Biochim Biophys Acta*. 2009;1792:14–26.
- Raponi M, Baralle D. Alternative splicing: good and bad effects of translationally silent substitutions. *FEBS J*. 2010;277:836–840.
- Ward AJ, Cooper TA. The pathobiology of splicing. *J Pathol*. 2010;220:152–163.
- 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.
- 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.
- Licalosi DD, Darnell RB. Splicing regulation in neurologic disease. *Neuron*. 2006;52:93–101.
- Hartong DT, Berson EL, Dryja TP. Retinitis pigmentosa. *Lancet*. 2006;368:1795–1809.
- 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.
- 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*. 2010;51:1628–1635.
- 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*. 2010;86:805–812.
- Wang Y, Juraneck 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.
- 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.
- Bornancin F, Mechtcheriakova D, Stora S, et al. Characterization of a ceramide kinase-like protein. *Biochim Biophys Acta*. 2005;1687:31–43.
- 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.
- Mendez A, Lem J, Simon M, Chen J. Light-dependent translocation of arrestin in the absence of rhodopsin phosphorylation and transducin signaling. *J Neurosci*. 2003;23:3124–3129.
- Rhead B, Karolchik D, Kuhn RM, et al. The UCSC Genome Browser database: update. *Nucleic Acids Res*. 2010;38:D613–D619.
- Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. VISTA: computational tools for comparative genomics. *Nucleic Acids Res*. 2004;32:W273–W279.
- Tatusova TA, Madden TL. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol Lett*. 1999;174:247–250.
- Pruitt KD, Tatusova T, Klimke W, Maglott DR. NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res*. 2009;37:D32–D36.
- Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. *GenBank Nucleic Acids Res*. 2009;37:D26–D31.
- 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–D101.
- Harrow J, Denoeud F, Frankish A, et al. GENCODE: producing a reference annotation for ENCODE. *Genome Biol*. 2006;7(suppl 1):S41–S49.
- Slater GS, Birney E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*. 2005;6:31.
- Davuluri RV, Grosse I, Zhang MQ. Computational identification of promoters and first exons in the human genome. *Nat Genet*. 2001;29:412–417.
- Bailey TL, Boden M, Buske FA, et al. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res*. 2009;37:W202–W208.
- Matys V, Fricke E, Geffers R, et al. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res*. 2003;31:374–378.
- 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.
- Portales-Casamar E, Thongjuea S, Kwon AT, et al. JASPAR: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res*. 2010;38:D105–D110.
- Kozak M. An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res*. 1987;15:8125–8148.
- Celniker SE, Dillon LA, Gerstein MB, et al. Unlocking the secrets of the genome. *Nature*. 2009;459:927–930.
- 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.
- Auslender N, Sharon D, Abbasi AH, Garzoni 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.
- 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.
- 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.
- Barash Y, Calarco JA, Gao W, et al. Deciphering the splicing code. *Nature*. 2010;465:53–59.
- Tejedor JR, Valcarcel J. Gene regulation: Breaking the second genetic code. *Nature*. 2010;465:45–46.

42. 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.
43. 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.
44. 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.
45. Tanackovic G, Ransijn A, Thibault P, et al. PRPF mutations are associated with generalized defects in spliceosome formation and pre-mRNA splicing in patients with retinitis pigmentosa. *Hum Mol Genet.* 2011;20:2116-2130.
46. Spellacy CJ, Daiger SP, Sullivan LS, et al. Characterization of retinal inosine monophosphate dehydrogenase 1 in several mammalian species. *Mol Vis.* 2007;13:1866-1872.
47. 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.
48. Tanackovic G, Rivolta C. PRPF31 alternative splicing and expression in human retina. *Ophthalmic Genet.* 2009;30:76-83.
49. Kirschner R, Rosenberg T, Schultz-Heienbrok R, et al. RPGR transcription studies in mouse and human tissues reveal a retina-specific isoform that is disrupted in a patient with X-linked retinitis pigmentosa. *Hum Mol Genet.* 1999;8:1571-1578.
50. de Lima Morais DA, Harrison PM. Large-scale evidence for conservation of NMD candidature across mammals. *PLoS One.* 5:e11695.
51. Pomares E, Marfany G, Brión MJ, et al. Novel high-throughput SNP genotyping cosegregation analysis for genetic diagnosis of autosomal recessive retinitis pigmentosa and Leber congenital amaurosis. *Hum Mutat.* 2007;28:511-516.
52. Avila-Fernandez A, Riveiro-Alvarez R, Vallespin E, et al. CERKL mutations and associated phenotypes in seven Spanish families with autosomal recessive retinitis pigmentosa. *Invest Ophthalmol Vis Sci.* 2008;49:2709-2713.
53. Aleman TS, Soumitra 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.
54. 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.* 2010;51:5943-5951.
55. Ali M, Ramprasad VL, Soumitra 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.
56. Tang Z, Wang Z, Wang Z, et al. Novel compound heterozygous mutations in CERKL cause autosomal recessive retinitis pigmentosa in a nonconsanguineous Chinese family. *Arch Ophthalmol.* 2009;127:1077-1078.