


Defective splicing of the *RB1* transcript is the dominant cause of retinoblastomas

Kamil J. Cygan^{1,2}  · Rachel Soemedi^{1,2} · Christy L. Rhine² · Abraham Profeta³ · Eileen L. Murphy² · Michael F. Murray⁴ · William G. Fairbrother^{1,2,5}

Received: 24 April 2017 / Accepted: 20 July 2017 / Published online: 5 August 2017
© Springer-Verlag GmbH Germany 2017

Abstract Defective splicing is a common cause of genetic diseases. On average, 13.4% of all hereditary disease alleles are classified as splicing mutations with most mapping to the critical GT or AG nucleotides within the 5' and 3' splice sites. However, splicing mutations are underreported and the fraction of splicing mutations that compose all disease alleles varies greatly across disease gene. For example, there is a great excess (46%; ~threefold) of hereditary disease alleles that map to splice sites in *RB1* that cause retinoblastoma. Furthermore, mutations in the exons and deeper intronic position may also affect splicing. We recently developed a high-throughput method that assays reported disease mutations for their ability to disrupt pre-mRNA splicing. Surprisingly, 27% of *RB1*-coding mutations tested also disrupt splicing. High-throughput in vitro spliceosomal assembly assay reveals heterogeneity in which stage of spliceosomal assembly is affected by splicing mutations.

58% of exonic splicing mutations were primarily blocked at the A complex in transition to the B complex and 33% were blocked at the B complex. Several mutants appear to reduce more than one step in the assembly. As *RB1* splicing mutants are enriched in retinoblastoma disease alleles, additional priority should be allocated to this class of allele while interpreting clinical sequencing experiments. Analysis of the spectrum of *RB1* variants observed in 60,706 exomes identifies 197 variants that have enough potential to disrupt splicing to warrant further consideration.

Kamil J. Cygan and Rachel Soemedi contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s00439-017-1833-4) contains supplementary material, which is available to authorized users.

✉ William G. Fairbrother
william_fairbrother@brown.edu

¹ Center for Computational Molecular Biology, Brown University, Providence, RI, USA

² Department of Molecular Biology, Cell Biology and Biochemistry, Brown University, Providence, RI, USA

³ College Hill Research, Barrington, RI, USA

⁴ Geisinger Health System, Danville, PA 17822, USA

⁵ Hassenfeld Child Health Innovation Institute, Brown University, Providence, RI, USA

Introduction

Retinoblastoma, the most common form of cancer in children, has a complex etiology. While most instances of retinoblastoma occur in a sporadic form, a third of all cases start with an inherited mutation in one copy of the *RB1* gene. Although retinoblastoma is a recessive disorder, the disease appears to follow a dominant mode of transmission as a second loss-of-function mutation is acquired somatically. Retinoblastomas are treated effectively with surgery though survivors of retinoblastoma face an increased risk of other cancers later in life.

There are 293 single nucleotide mutations in *RB1* reported to cause retinoblastoma currently listed in the public version of the Human Gene Mutation Database (HGMD) (Stenson et al. 2003). Many of these mutations were reported to disrupt splicing. Some *RB1* splicing mutations have been associated with lower penetrance (Lefevre et al. 2002; Schaffer et al. 2000; Schubert et al. 1997) as well as incomplete penetrance that varies between paternal and maternal inheritances (Klutzn et al. 2002).

Across all of the disease mutations in 2314 intron-containing genes reported in the HGMD, on average 13.4%

disrupt splicing. Splicing mutations most often result in exon skipping, but other consequences of splice-site misrecognition include the usage of cryptic 3' or 5' splice site (ss) or intron retention (IR). A common feature of these aberrant isoforms is an insertion or deletion (in/del) in the mRNA, which often results in a disruption of reading frame and message decay by the nonsense-mediated decay (NMD) pathway.

The cost of sequencing has fallen several thousandfold in the last 10 years and has driven application of whole-genome sequencing (WGS) and whole-exome sequencing (WES) into personal genomics and clinical medicine. A typical exome will reveal thousands of variants of unknown significance. The effects of coding variants on protein function are particularly difficult to interpret, as individual functional assays do not exist for most proteins. In contrast, splicing function can be determined by polymerase chain reaction (PCR) analysis of complementary DNA (cDNA) from RNA extracted from mutant and wild-type cells (i.e., RT-PCR).

Recently, our group developed a *massively parallel* splicing assay (MaPSy) to screen the effects of coding variants on splicing (Soemedi et al. 2017). This dual *in vivo/in vitro* assay identifies previously annotated coding mutations as splicing mutations and pinpoints the stage in splicing that is disrupted. Here, we present the results of this assay on 30 coding variants in *RBI* that cause retinoblastoma. A total of eight mutations significantly disrupt splicing *in vivo* and *in vitro*. Numerous variants carried in the population at low levels have the potential to disrupt splicing and potentially contribute to elevated cancer risk in their carriers. We have created an online tool that displays these results.

Results

Global analysis suggests that retinoblastoma belongs to a distinct class of diseases driven by splicing mutations

To better understand the spectrum of mutations that cause retinoblastoma, the role of splicing mutations in disease was explored for all reported disease-causing mutations. For each of the hereditary disease genes reported in HGMD, the complete set of point mutations was downloaded and separated according to location (i.e., splice site and exonic). The fraction of point mutations that alter splicing was displayed graphically [Fig. 1a, adapted from Soemedi et al. (2017)]. Overall, the fraction of all disease-causing mutations that altered splicing was estimated to be 13.4% on average. A permutation approach was used to determine the 99.9% confidence interval for each disease gene. The retinoblastoma causing *RBI* was significantly enriched for mutations that fall within the splice sites. A total of 130 mutations,

comprising 46% of all reported point mutations, were localized to canonical splice sites.

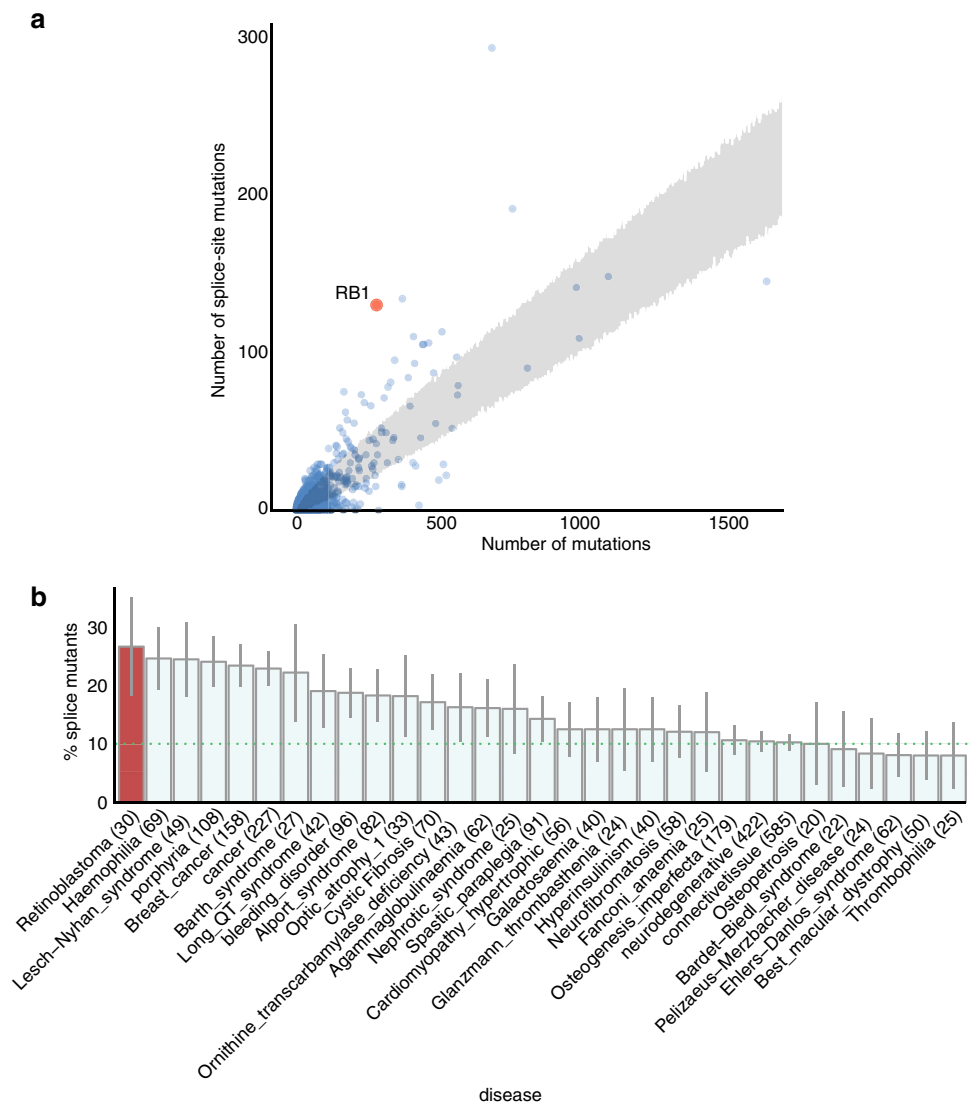
The remaining coding mutations (Table 1) were analyzed with MaPSy. Briefly, this assay resynthesizes the wild type and the mutant sequences and combines thousand of these allelic reporters in a single pool, which was tested *in vivo* (via transfection of HEK293 cells) and *in vitro* (via incubation in HeLa nuclear extract) for splicing efficiency. A contingency table was created for each mutant/wild-type pair and included the counts obtained from deep sequencing of the input pool as well as the output-spliced fractions. To determine pairs with significant allelic skew, we required at least 1.5-fold change and a two-sided Fisher's exact test adjusted with 5% false discovery rate (FDR). Of the 5000 exonic mutations tested in this panel, approximately 10% disrupted splicing *in vivo* and *in vitro* (dotted green line, Fig. 1b) (Soemedi et al. 2017). Of all the diseases in HGMD, the highest fraction (27%) of splicing phenotypes seen in exonic mutations was found in the retinoblastoma gene, *RBI*. This represents a 2.5-fold enrichment in splicing disrupting mutations in coding mutations and a threefold enrichment in mutations that coincide with splice sites. Taken together, these data suggest that *RBI* is especially prone to being disrupted by splicing mutants.

RBI has a high fraction of exonic mutations that alter splicing *in vitro* and *in vivo*

To place these coding splicing mutations in their genomic context, *RBI* and exon loci are depicted in Fig. 2a. *RBI* is a relatively large gene (200 Kb, 26 introns) that contains more introns than an average transcript (mean number of introns in a transcript is ~11). A total of eight exons were selected for mutational analysis (Table 1). Of the eight exons examined, five of which contained exonic mutations that affected splicing (Fig. 2; Table 1). The criteria utilized for identifying exonic splicing mutants was a fold difference of at least 1.5 between weaker and stronger alleles, two-sided Fisher's exact test adjusted with 5% FDR. To allow for comparison of splicing performance between any two species in the pool, an individual splicing index was calculated using the following formula: $\log_2 \left(\frac{\text{spl}_i / \sum_{i=1}^n \text{spl}_i}{\text{inp}_i / \sum_{i=1}^n \text{inp}_i} \right)$, where spl_i is the count of

reads in the spliced-output fraction for species i , inp_i is the count of reads in the input for species i , and n is the total number of species in the pool. In the majority of cases (7/8), the mutant allele was the weaker substrate, splicing on average at 25% of the efficiency of the wild-type allele (Fig. 2b; Soemedi et al. 2017). It is typical that splicing mutations result in exon skipping. Less frequently, usage of cryptic 3'ss or 5'ss is observed and occasional cases of IR has been reported. The consequence of these aberrant processing events is to create a transcript in/del in the message. This

Fig. 1 *RB1* mutations frequently disrupt splicing. **a** HGMD non-synonymous, nonsense, and splice-site mutations were grouped by genes ($n = 2314$) and analyzed. Mutations that fell in the canonical splice sites (y axis) were plotted against total mutations (x-axis). The gray area represents region of 99.9% confidence interval (see Electronic Supplementary Material Methods). **b** Proportions of coding mutations that disrupt splicing in different hereditary disorders were ranked according to their degrees of enrichment of splicing mutations. Red bar indicates 27% of *RB1* mutations disrupt splicing in vivo and in vitro. Error bars represent 95% confidence interval. The number inside each set of round brackets following each disease name represents a total number of mutations tested by MaPSy for that disease



modification disrupts the reading frame unless the size of the in/del is a multiple of three. In the case of the 30 *RB1* mutations screened, none of the mutations were associated with a cryptic splicing event and 5 out of 8 of significant MaPSy mutations in *RB1* would be predicted to create a frame-shifting event if the exon was skipped. While there is good correlation between the relative strength of wild type and mutant splicing in vitro and in vivo, the in vivo assay is a more sensitive assay and tends to result in more extreme enrichment or depletion of an allele in the successfully spliced fraction (Soemedi et al. 2017).

The transition between A and B complexes is the major point of disruption for *RB1*-coding mutants

In addition to offering a highly sensitive test of a pre-mRNA's ability to serve as a substrate for splicing, the in vitro splicing assay can be utilized to follow spliceosome

assembly (Padgett et al. 1986). The in vitro splicing/spliceosome assembly was done by performing splicing reaction by incubating in vitro transcribed RNA in cell nuclear extract. The spliceosome assembles in a stepwise fashion on in vitro transcribed RNA to form the A, the B, and the C complexes (Das and Reed 1999; Konarska and Sharp 1986). The spliceosome assembly and the formation of the A complex begin with U1 and U2 small nuclear ribonucleoproteins (snRNPs) binding to the 5' splice site and the branch point, respectively. In the next step, U4, U5, and U6 tri-snRNP is recruited to form pre-catalytic B complex. Following the formation of the B complex, a series of conformational changes results in the release of U1 and U4 subunits and establishes the C complex (Fig. 3a). While the B and C complexes on the reporter substrates are biochemically inseparable, the A and B/C complexes can be biochemically separated into distinct fractions by glycerol gradient centrifugation (Soemedi et al. 2017). The resulting fractions

Table 1 Variants in *RBI* analyzed with MaPSy

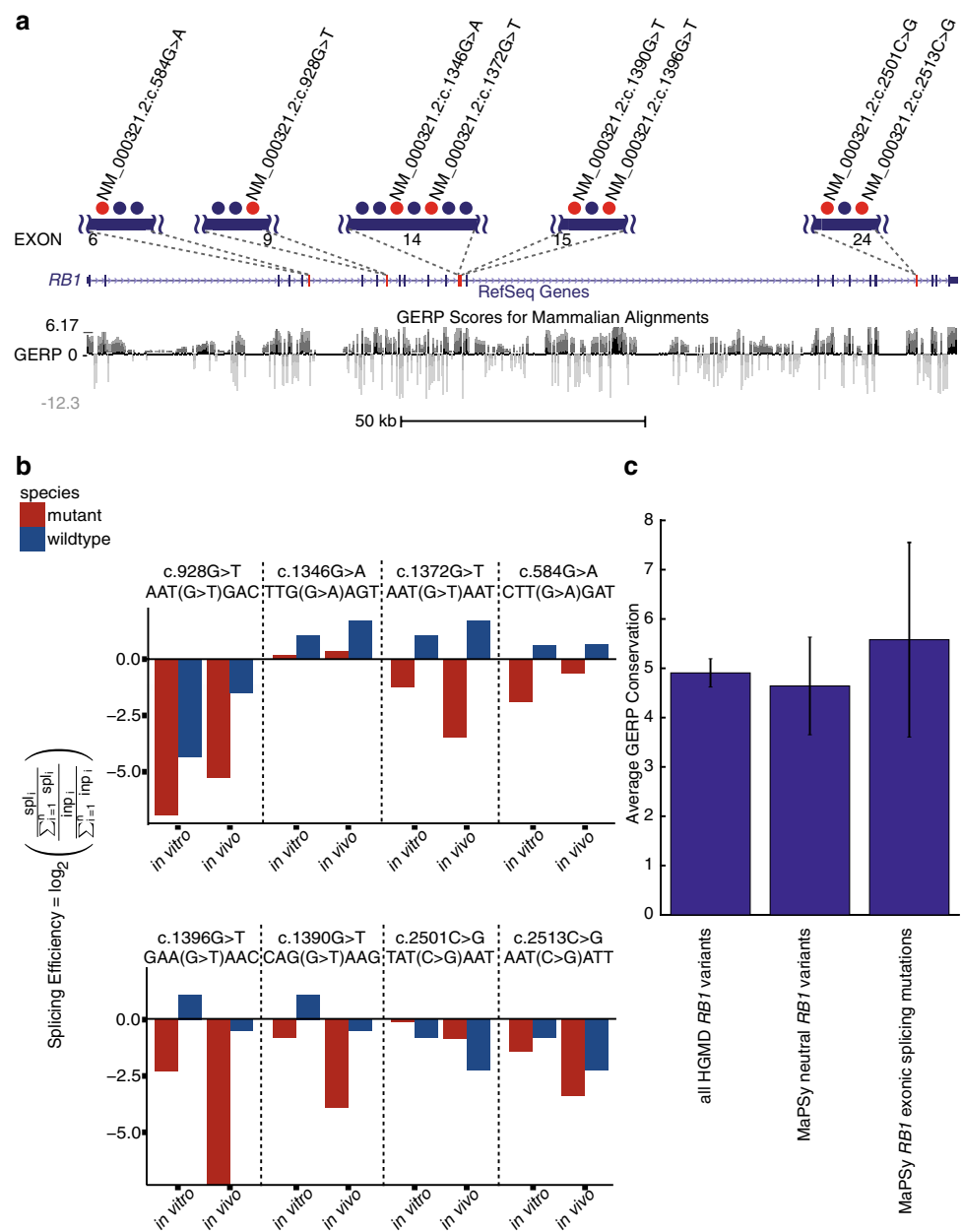
HGMD ID	Significant in MaPSy	Exon number	Frame-shifting if exon skipped	Variant description
CM016042	No	14	No	NM_000321.2:c.1339A>T
CM016043	No	16	Yes	NM_000321.2:c.1449T>G
CM022059	No	6	Yes	NM_000321.2:c.604A>T
CM025388	No	9	No	NM_000321.2:c.908T>A
CM025389	No	16	Yes	NM_000321.2:c.1494T>A
CM025390	No	16	Yes	NM_000321.2:c.1467C>A
CM030499	No	9	No	NM_000321.2:c.920C>T
CM030500	Yes	9	No	NM_000321.2:c.928G>T
CM030504	Yes	14	No	NM_000321.2:c.1346G>A
CM030505	Yes	14	No	NM_000321.2:c.1372G>T
CM032660	No	16	Yes	NM_000321.2:c.1494T>G
CM034897	Yes	6	Yes	NM_000321.2:c.584G>A
CM040261	Yes	15	Yes	NM_000321.2:c.1396G>T
CM044254	No	12	Yes	NM_000321.2:c.1129A>T
CM063089	Yes	15	Yes	NM_000321.2:c.1390G>T
CM071074	No	12	Yes	NM_000321.2:c.1166T>A
CM117842	No	16	Yes	NM_000321.2:c.1447C>T
CM900192	No	14	No	NM_000321.2:c.1333C>T
CM942037	No	12	Yes	NM_000321.2:c.1150C>T
CM951103	No	6	Yes	NM_000321.2:c.554T>C
CM951105	No	11	No	NM_000321.2:c.1072C>T
CM951106	No	15	Yes	NM_000321.2:c.1399C>T
CM952105	Yes	24	Yes	NM_000321.2:c.2501C>G
CM952423	Yes	24	Yes	NM_000321.2:c.2513C>G
CM961225	No	11	No	NM_000321.2:c.1060C>T
CM961226	No	11	No	NM_000321.2:c.1072C>G
CM961227	No	12	Yes	NM_000321.2:c.1190C>A
CM961228	No	14	No	NM_000321.2:c.1363C>T
CM973041	No	14	No	NM_000321.2:c.1339A>C
CM981700	No	14	No	NM_000321.2:c.1388C>G

can be sequenced and the representation of the wild type and mutant alleles can be ascertained from the read counts. The following formula was used to calculate allelic skew at each stage of spliceosome assembly: $\log_2 \left(\frac{\text{mut}_s/\text{mut}_i}{\text{wt}_s/\text{wt}_i} \right)$, where mut_s

is the count of reads in the stage for the mutant, mut_i is the count of reads in the input for the mutant, wt_s is the count of reads in the stage for the wild type, and wt_i is the count of reads in the input for the wild type. The eight *RBI* mutations were analyzed with the experiment described above. The results reveal diverse patterns of enrichment across the different fractions (Fig. 3b; Soemedi et al. 2017). The map of *RBI* is shown in Fig. 3b, top. Exons with exonic splicing mutations (ESMs) are colored red, blue represents exons that were tested with MaPSy, but have no ESM, and exons that were not tested are colored black (Fig. 3b, top). Each mutant to wild-type ratios in A, BC, and spliced fractions is illustrated as a heat map (Fig. 3b). Degrees of enrichment are

indicated in red color spectrum and degrees of depletion in blue color spectrum (ratios in log₂ scale indicated in color spectrum is shown in the bar legend below each heat map). The relative positions of ESM (red dots) and control (black dots) in each exon are illustrated to the left of the heat maps. The first group represented cases, where the mutation principally blocked A complex formation. In practice, this step was not the only step of splicing affected, but mutant NM_000321.2:c.1396G>T (HGMD ID CM040261) clearly illustrates an underrepresentation of mutant allele in the A complex fraction, suggesting a failure in the events leading up to U1 snRNP and U2 snRNP recognizing the 5' ss and 3' ss. Mutant NM_000321.2:c.1372G>T (HGMD ID CM030505) has a minor defect in A complex assembly, and while the mechanism of splicing defect in NM_000321.2:c.1372G>T was similar to NM_000321.2:c.1396G>T, NM_000321.2:c.1372G>T has more dramatic distortion in allelic ratio during the transition

Fig. 2 *RB1* mutations that disrupt splicing. Out of 30 *RB1* mutations analyzed with MaPSy, 8 disrupted splicing both in vivo and in vitro. **a** *RB1* exons that contain splicing mutations are shown in red and other exons in blue. Each exon with splicing mutations is shown enlarged with mutations depicted as dots (red mutations that disrupt splicing and blue mutations that do not disrupt splicing). **b** Relative representation of the mutant (red bar) and wild type (blue bar) in the spliced product. Negative values indicate a species that lost representation in the fully spliced (i.e., output) pool relative to the starting library (i.e., input) pool; sp_i is the count of reads in the spliced-output fraction for species i , inp_i is the count of reads in the input for species i , and n is the total number of species in the pool. **c** Average Genomic Evolutionary Rate Profiling (GERP) conservation for all HGMD variants in *RB1*, MaPSy neutral variants in *RB1*, and significant MaPSy variants in *RB1*



from the A complex to the B/C complex. This transition was the most common point of misregulation of exonic splicing mutants in *RB1*. The remaining mutations appear to also block one of the events leading up to the entry of U4/U6.U5 tri-snRNP into the spliceosome and the formation of the catalytically active complex (Fig. 3a).

Deep sampling of genetic variation in the human population suggests the presence of rare deleterious alleles that may alter splicing

Disruption of *RB1* splicing often leads to retinoblastoma. As demonstrated earlier, a high fraction of non-synonymous variants cause aberrant splicing of *RB1*. Currently,

the whole-exome sequencing technology has been increasingly used to detect causal variants and diagnose diseases. Intronic variants can also be captured by exome sequencing as the library fragments often extend into the intron. Because non-coding variation can contribute to disease risk, we examined a typical exome capture [NA12891WEX data set 1000 Genomes Project (Genomes Project et al. 2015)]. The reanalysis reveals how the power to detect intronic variants declines as a function of distance from splice site (Fig. 4a). While there is no strict consensus, typical exome sequencing run will be designed at 20- to 100-fold mean transcriptome coverage (MTC) (Meynert et al. 2013). An intronic position 75 nucleotides away from

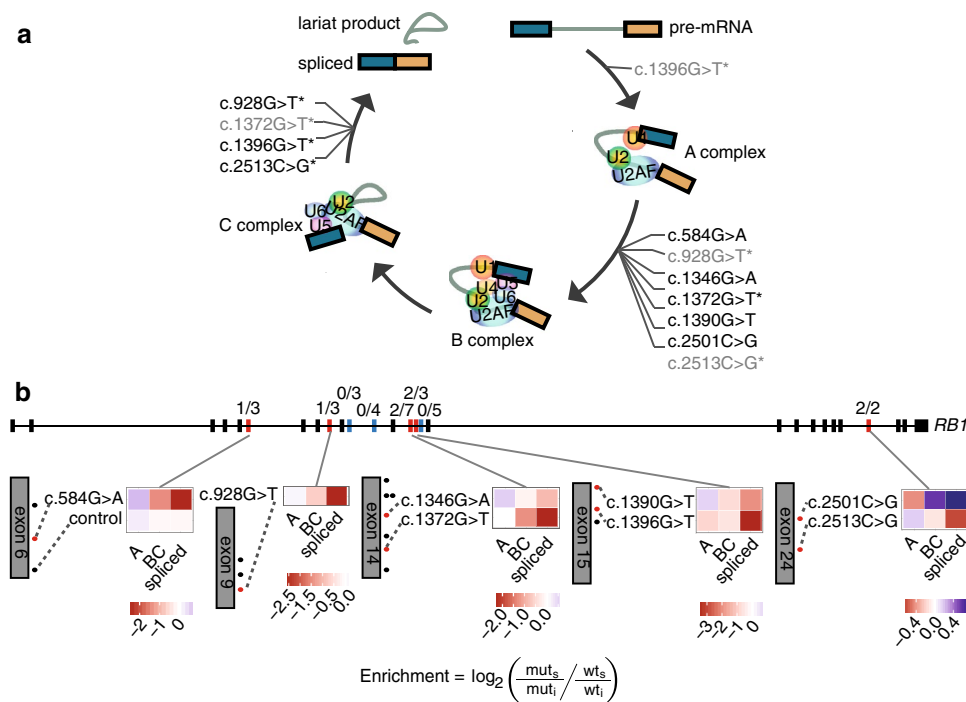


Fig. 3 *RB1* mutations disrupt splicing in various stages of the spliceosomal assembly. **a** *RB1* mutations alter splicing at different stages of spliceosomal assembly. The mechanism of spliceosome assembly from A, BC, and spliced is illustrated. ESM that impact each transition of the assembly are indicated (*black font* indicates major disruption and *gray font* indicates minor disruption). ESM that act at multiple stages of spliceosome assembly are marked with *asterisks*. **b** Relative positions of all exons in *RB1* gene are shown with exons containing ESM shown in *red*, exons with no ESM that were tested with MaPSy in *blue* and exons that were not tested in *black* (*top*). Numbers on top of each exon represent the number of ESM out of total number of mutations tested in the respective exon. The heat-map

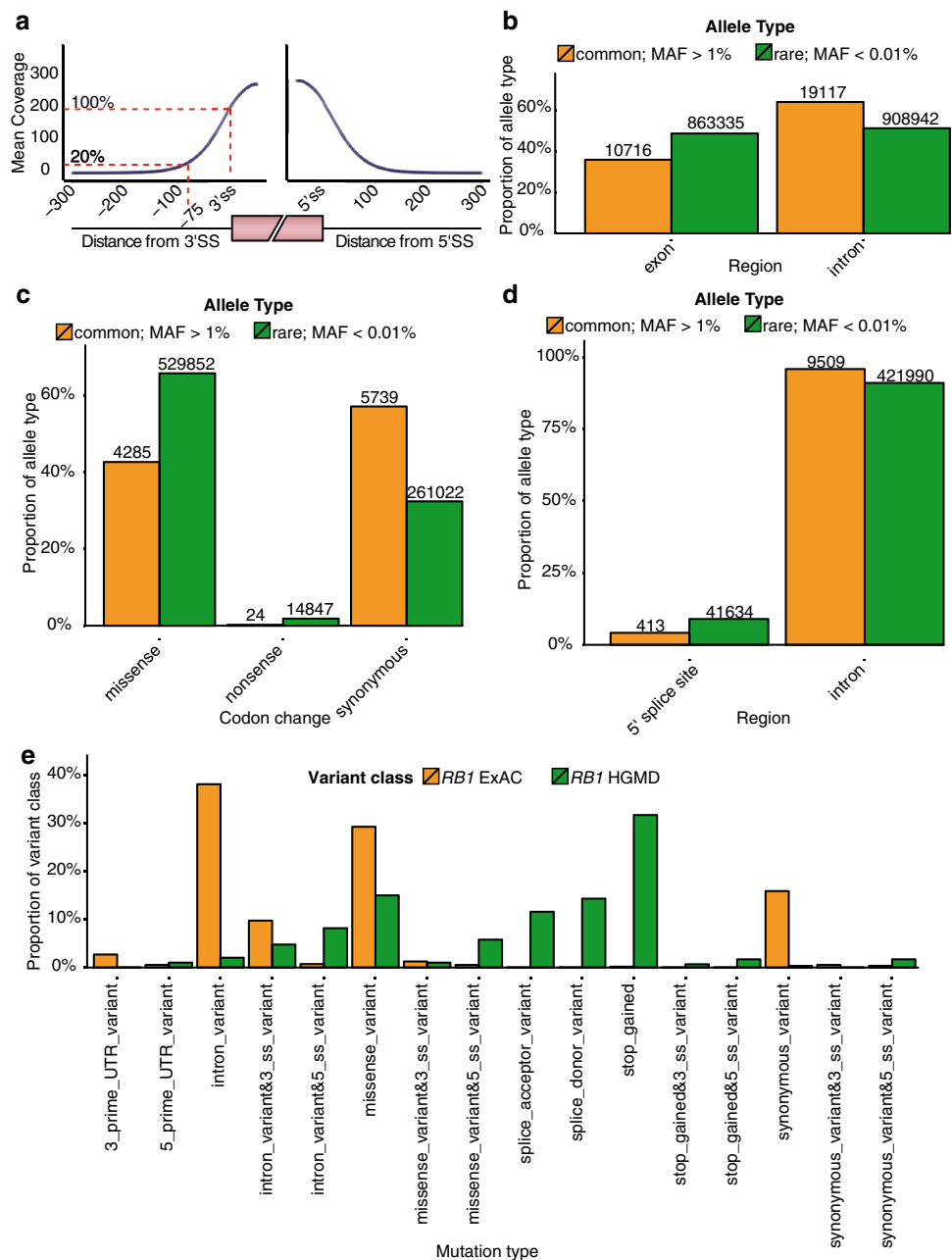
representation of the degree of enrichment in the spliceosomal fractions (A, BC, and spliced) for each mutation in the exon is shown, together with the relative position of the respective mutation in each exon (*gray box* to the left of each heat map). ESM are indicated as *red dots* and non-ESM as *black dots*, with the *bottom end of the gray box* representing the 5' end of the exon and the *top* representing the 3' end of the exon. Log₂ scale of the mutant to wild-type ratios in each spliceosomal fraction is indicated in the *color bar* legend below each heat map; mut_s is the count of reads in the fraction for the mutant, wt_s is the count of reads in the input for the mutant, mut_t is the count of reads in the fraction for the wild type, and wt_t is the count of reads in the input for the wild type

the splice site has about 20% of full coverage (e.g., 4X coverage for an experiment with 20X MTC) (Fig. 4a).

Several studies have suggested variable penetrance of *RB1* splicing mutants, so it is entirely possible for asymptomatic individuals to be carriers of retinoblastoma alleles (Harbour 2001; Lefevre et al. 2002; Scheffer et al. 2000; Schubert et al. 1997) or potentially be more susceptible to other types of cancers (Dommering et al. 2012). To better understand the potential contribution of natural variants of *RB1* to retinoblastoma disease risk, as well as, to identify and estimate a proportion of variants that could be potential future targets in understanding the disease mechanisms, the aggregated data of all publicly available exome sequencing experiments from the Exome Aggregation Consortium (ExAC) (Lek et al. 2016) were downloaded for analysis. According to prevailing models of population genetics, variations that accumulate to an appreciable frequency in the population will mostly be neutral as variation that have a deleterious or advantageous affect on fitness will either

be rapidly eliminated or fixed in the population by natural selection (Kimura 1983). The resulting prediction is that common variants will be more likely to be neutral than rare variants. However, it is possible that other factors like linkage disequilibrium can complicate inferences of selection drawn from data derived from single alleles. To explore the agreement of the observed human polymorphism data with theoretical models of selection, all ExAC variants were separated into classes based on the severity of their predicted effect on gene function (i.e., nonsense > missense > synonymous). Variants were also separated by minor allele frequency (MAF) into rare (MAF < 0.01%) and common (MAF > 1%). Common variants were depleted in all the functional categories tested (e.g., depleted in exons relative to introns, Fig. 4b). Rare variants, on the other hand, were more evenly distributed. Selection against protein coding changes was observed in the enrichment of common variants in synonymous changes and depletion in the missense and nonsense categories (Fig. 4c).

Fig. 4 Low-frequency variants predicted to disrupt splicing in *RB1*. **a** Power to discover intronic variants in exome sequencing experiment declines as a function of distance from exon boundaries. **b** Common alleles' underrepresentation in exons shows the mark of selection (*salmon color*). Rare alleles are more uniformly distributed across functional and non-functional categories (*teal color*). **c** Common exonic alleles (*salmon color*) are underrepresented in deleterious categories (missense and nonsense) of mutations relative to rare variants (*teal color*). **d** Common intronic variants (*salmon color*) are underrepresented in 5'ss relative to rare alleles (*teal color*). **e** Comparison of the distribution of polymorphisms (ExAC variants, *salmon color*) in *RB1* to retinoblastoma causing *RB1* disease alleles (HGMD, *teal color*)



To determine if selection was evident against splicing signals, variants that disrupted splice sites were evaluated relative to variants in other parts of the intron. Again, rare variants occurred at almost twice the rate of common variants in 5'ss (Fig. 4d). As all common variants were initially rare variants, this result suggests that half of all rare single nucleotide polymorphisms (SNPs) that fall within splice-site regions are eliminated by natural selection.

At least 553 *RB1* variants exist in the human population

Given the previous analysis, there exist a substantial proportion of rare variants found in asymptomatic individuals that

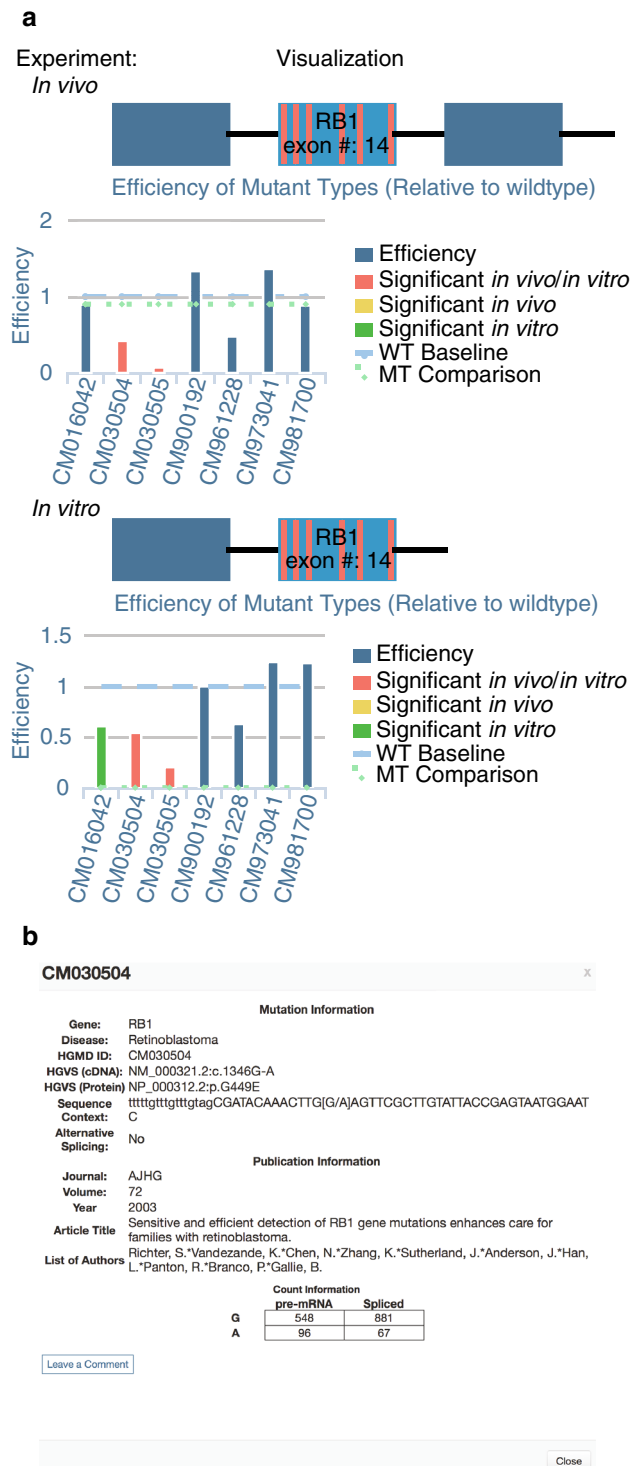
might in fact carry an increased disease risk for retinoblastoma. We decided to evaluate all reported variants and create a curated list of the ones that can potentially cause splicing aberration in *RB1*. Among *RB1* variants discovered in the ExAC data set of 60,706 exomes, all but 4 are rare variants (MAF < 0.01%). 284 variants occur in the coding region, and 269 occur in the intronic region. Further analysis shows that 13.4% of the rare variants fall within splice sites (65 variants fall within -20 to +3 window at the 3' splice site and 9 variants fall within -3 to +6 window at the 5' splice site). This distribution of low-frequency alleles is confirmed in another data set of asymptomatic individuals (the Geisinger cohort contains 423 allele/50,000 people). The distribution

Fig. 5 Online browser enables navigation through *RB1* mutations that disrupt splicing. <http://fairbrother.biomed.brown.edu/RB1.htm> contains an interactive online tool to browse the results of in vitro and in vivo splicing assays on 30 *RB1* mutations. **a** Main browser depicts the in vivo substrate (above) and the in vitro substrate (below) for each exon in *RB1* that was tested for splicing efficiency. Histogram bars indicate the normalized efficiency of mutation as a fraction of the wild type (dashed blue line). Red bars indicate significant results. Blue bars indicate results that failed to meet significance threshold (a fold difference of at least 1.5 between weaker and stronger allele, two-sided Fisher's exact test adjusted with 5% FDR). Options convert HGMD labels to amino acid changes. **b** Mouse over buttons displays raw data for each mutation tested

of these variants across different categories (e.g., splice-site variants, synonymous, non-synonymous, etc.) differs significantly from the 293 HGMD mutations reported to cause retinoblastoma reflecting the fact that most variations are not as deleterious as disease alleles (Fig. 4e). However, the exonic regions of *RB1* are highly conserved across vertebrates implying that the coding sequence is evolving under purifying selection (Fig. 2a, c). It is very likely that some of these rare variants affect *RB1* function or splicing and potentially contribute to disease risk. While the HGMD disease alleles arose through spontaneous mutation, many of the rare alleles appeared to be inherited. As rare alleles are more likely to be deleterious, we have compiled a table of all variants in *RB1* annotated by splicing location (intron, 5'ss, branch-point region, 3'ss, and exon) (Supplementary Table 1). To construct the final list of mutations that have enough potential to disrupt splicing, *RB1* variants were analyzed by a variety of predictive tools (Ke et al. 2011; Lim and Fairbrother 2012; Taggart et al. 2012; Xiong et al. 2015; Yeo and Burge 2004). 197 variants were set aside for further consideration. These variants either: created or disrupted splicing elements, modified the branch-point region, substantially changed inclusion/exclusion ratio of the closest exon, or fell within a splice site (Supplementary Table 1).

Online visualization tool enables splicing phenotype to be added to annotations of disease alleles

Finally, a visualization tool is available to access the splicing data for disease-causing variants in *RB1*. An online mutation browser was developed that diagrams the reporter constructs used in the assay (Fig. 5). The location of each analyzed *RB1* variant is depicted and can be used to check for clustering of mutations or the distance from known splicing signals (splice sites) (Fig. 5a). In addition, the visualization tool shows splicing efficiency of each variant relative to the wild-type counterpart and allows for quick interpretation of severity of the splicing phenotype (Fig. 5a). Finally, the online tool contains the in vivo and in vitro results for all 30 of mutations: read counts in vivo and in vitro for the starting libraries, A,



B/C, and spliced complexes for a total of 30×6 experiments (Fig. 5b). Mutations can be searched by mutant ID or author and references link mutation to original reports. The online tool allows researchers to analyze novel variants and submit alleles for analysis with MaPSy. The tool is freely available at <http://fairbrother.biomed.brown.edu/RB1.htm> and supports all major browsers.

Discussion

With the development of tools to screen thousand of variants for splicing defects, it has become clear that retinoblastoma is highly biased towards splicing mutations. Indeed, for all the exonic disease mutations that were included in a recent high-throughput screen of disease variants, retinoblastoma had the highest fraction of ESMs. This suggests that diseases like retinoblastoma function mainly by a loss-of-function mechanism. As many splicing defects result in mRNA indels that disrupt reading frame, splicing mutations can be highly deleterious and, therefore, warrant additional attention during the variant classification process.

In addition to discovering alleles that confer a splicing defect, the mechanism of the defect was uncovered. The spliceosome assembles in a progressive, step-wise fashion. While early work on splicing indicated that many spliceosomes were committed to a splicing relatively early (i.e., during the formation of the A complex), defective alleles in *RB1* frequently block splicing at later points in the assembly. Frequently, it is the formation of the B or C complex that is prevented by a mutation, perhaps by distorting enhancer signals used by the SR family and related proteins which have been shown to drive early stage spliceosome formations into active ones (Rosciigno and Garcia-Blanco 1995). A recent systematic depletion of splicing factors also found substrate specific effects when targeting core components of the spliceosome. Components that were thought to be required for B complex or even mature catalytic spliceosomes affected splice-site selection (Papasaikas et al. 2015).

As sequencing becomes less expensive, exome sequencing will be used more and more in precision medicine applications. Individual genomes will be sequenced more frequently by exome sequencing. We demonstrated that the capture technology can be used to discover variants well into the intron. Unlike disease alleles, variants discovered in this fashion are, for the most part, biologically inert. Despite this we can demonstrate the mark of selection on splicing signals such as the 5' splice site (Fig. 4d) and estimate the proportion of variants that are in fact deleterious. By comparing common to rare alleles, the data suggests that many *RB1* rare alleles are deleterious and subject to negative selection. As the vast majority of variants in a genome are rare variants, a major focus going forward will be to screen these variants for alleles that contribute to disease risk. This goal is especially important for genes like *RB1* that cause diseases, which seem to be especially biased towards splicing mutations. Finally, we have created an online search tool for splicing defective *RB1* variants and will systematically screen the population *RB1* variants using MaPSy. In conclusion, by leveraging the power of biochemical assays, computational tools, and whole-exome capturing technologies, it is entirely possible to deeply characterize sequence variations

and assign risk alleles for diseases like retinoblastoma. This task is more urgent than ever to keep pace with variant discovery that is taking place at the clinics.

Acknowledgements R.S. was supported by Postdoctoral Fellowship from Center for Computational Molecular Biology (CCMB), Brown University. C.L.R. was supported by Graduate Research Fellowship from National Science Foundation (NSF). This work was supported by the National Institutes of Health (NIH) Grants R01GM095612 (to W.G.F.), R01GM105681 (to W.G.F.) and R21HG007905 (to W.G.F.) and by SFARI award 342705 (to W.G.F.).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

- Das R, Reed R (1999) Resolution of the mammalian E complex and the ATP-dependent spliceosomal complexes on native agarose mini-gels. *RNA* 5:1504–1508
- Dommering CJ et al (2012) RB1 mutations and second primary malignancies after hereditary retinoblastoma. *Fam Cancer* 11:225–233. doi:[10.1007/s10689-011-9505-3](https://doi.org/10.1007/s10689-011-9505-3)
- Genomes Project C et al (2015) A global reference for human genetic variation. *Nature* 526:68–74. doi:[10.1038/nature15393](https://doi.org/10.1038/nature15393)
- Harbour JW (2001) Molecular basis of low-penetrance retinoblastoma. *Arch Ophthalmol* 119:1699–1704
- Ke S et al (2011) Quantitative evaluation of all hexamers as exonic splicing elements. *Genome Res* 21:1360–1374. doi:[10.1101/gr.119628.110](https://doi.org/10.1101/gr.119628.110)
- Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge
- Klutzb M, Brockmann D, Lohmann DR (2002) A parent-of-origin effect in two families with retinoblastoma is associated with a distinct splice mutation in the RB1 gene. *Am J Hum Genet* 71:174–179. doi:[10.1086/341284](https://doi.org/10.1086/341284)
- Konarska MM, Sharp PA (1986) Electrophoretic separation of complexes involved in the splicing of precursors to mRNAs. *Cell* 46:845–855
- Lefevre SH et al (2002) A T to C mutation in the polypyrimidine tract of the exon 9 splicing site of the RB1 gene responsible for low penetrance hereditary retinoblastoma. *J Med Genet* 39:E21
- Lek M et al (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536:285–291. doi:[10.1038/nature19057](https://doi.org/10.1038/nature19057)
- Lim KH, Fairbrother WG (2012) Spliceman—a computational web server that predicts sequence variations in pre-mRNA splicing. *Bioinformatics* 28:1031–1032. doi:[10.1093/bioinformatics/bts074](https://doi.org/10.1093/bioinformatics/bts074)
- Meynert AM, Bicknell LS, Hurler ME, Jackson AP, Taylor MS (2013) Quantifying single nucleotide variant detection sensitivity in exome sequencing. *BMC Bioinform* 14:195. doi:[10.1186/1471-2105-14-195](https://doi.org/10.1186/1471-2105-14-195)
- Padgett RA, Grabowski PJ, Konarska MM, Seiler S, Sharp PA (1986) Splicing of messenger RNA precursors. *Annu Rev Biochem* 55:1119–1150. doi:[10.1146/annurev.bi.55.070186.005351](https://doi.org/10.1146/annurev.bi.55.070186.005351)
- Papasaikas P, Tejedor JR, Vigevani L, Valcarcel J (2015) Functional splicing network reveals extensive regulatory potential of the core spliceosomal machinery. *Mol Cell* 57:7–22. doi:[10.1016/j.molcel.2014.10.030](https://doi.org/10.1016/j.molcel.2014.10.030)
- Rosciigno RF, Garcia-Blanco MA (1995) SR proteins escort the U4/U6.U5 tri-snRNP to the spliceosome. *RNA* 1:692–706

- Scheffer H, Van Der Vlies P, Burton M, Verlind E, Moll AC, Imhof SM, Buys CH (2000) Two novel germline mutations of the retinoblastoma gene (RB1) that show incomplete penetrance, one splice site and one missense. *J Med Genet* 37:E6
- Schubert EL, Strong LC, Hansen MF (1997) A splicing mutation in RB1 in low penetrance retinoblastoma. *Hum Genet* 100:557–563
- Soemedi R et al (2017) Pathogenic variants that alter protein code often disrupt splicing. *Nat Genet*. doi:[10.1038/ng.3837](https://doi.org/10.1038/ng.3837)
- Stenson PD et al (2003) Human gene mutation database (HGMD): 2003 update. *Hum Mutat* 21:577–581. doi:[10.1002/humu.10212](https://doi.org/10.1002/humu.10212)
- Taggart AJ, DeSimone AM, Shih JS, Filloux ME, Fairbrother WG (2012) Large-scale mapping of branchpoints in human pre-mRNA transcripts in vivo. *Nat Struct Mol Biol* 19:719–721. doi:[10.1038/nsmb.2327](https://doi.org/10.1038/nsmb.2327)
- Xiong HY et al (2015) RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. *Science* 347:1254806. doi:[10.1126/science.1254806](https://doi.org/10.1126/science.1254806)
- Yeo G, Burge CB (2004) Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biology* 11:377–394. doi:[10.1089/1066527041410418](https://doi.org/10.1089/1066527041410418)