Winter 1-13-2014

Validation of predicted mRNA splicing mutations using high-throughput transcriptome data

Coby Viner
University of Western Ontario, cviner2@uwo.ca

Stephanie Dorman
University of Western Ontario, sdorman@uwo.ca

Ben Shirley
Cytognomix Inc, ben.shirley@cytognomix.com

Peter Rogan
The University of Western Ontario, progan@uwo.ca

Citation of this paper:

Viner C, Dorman SN, Shirley BC and Rogan PK (2014) Validation of predicted mRNA splicing mutations using high-throughput transcriptome data [v1; ref status: indexed, http://f1000r.es/2no] F1000Research 2014, 3:8 (doi: 10.12688/f1000research.3-8.v1)

WEB TOOL

Validation of predicted mRNA splicing mutations using high-throughput transcriptome data [v1; ref status: indexed, http://f1000r.es/2no]

Coby Viner[1], Stephanie N. Dorman[2], Ben C. Shirley[3], Peter K. Rogan[1-3]

[1]Department of Computer Science, University of Western Ontario, London, Ontario, N6A 5B7, Canada
[2]Department of Biochemistry, University of Western Ontario, London, Ontario, N6A 5C1, Canada
[3]Cytognomix, Inc., London, Ontario, N6G 4X8, Canada

v1 **First published:** 13 Jan 2014, **3**:8 (doi: 10.12688/f1000research.3-8.v1)
Latest published: 13 Jan 2014, **3**:8 (doi: 10.12688/f1000research.3-8.v1)

Abstract

Interpretation of variants present in complete genomes or exomes reveals numerous sequence changes, only a fraction of which are likely to be pathogenic. Mutations have been traditionally inferred from allele frequencies and inheritance patterns in such data. Variants predicted to alter mRNA splicing can be validated by manual inspection of transcriptome sequencing data, however this approach is intractable for large datasets. These abnormal mRNA splicing patterns are characterized by reads demonstrating either exon skipping, cryptic splice site use, and high levels of intron inclusion, or combinations of these properties. We present, Veridical, an *in silico* method for the automatic validation of DNA sequencing variants that alter mRNA splicing. Veridical performs statistically valid comparisons of the normalized read counts of abnormal RNA species in mutant versus non-mutant tissues. This leverages large numbers of control samples to corroborate the consequences of predicted splicing variants in complete genomes and exomes.

Article Status Summary

Referee Responses

Referees	1	2	3	4
v1 published 13 Jan 2014	? report	✓ report	? report	✓ report

1 **Stefania Bortoluzzi**, University of Padova Italy

2 **Francesc Xavier Roca**, Nanyang Technological University Singapore

3 **Liliana Florea**, Johns Hopkins University USA

4 **Peter Robinson**, Universitätsklinikum Charité Germany

Latest Comments

No Comments Yet

Corresponding author: Peter K. Rogan (progan@uwo.ca)

How to cite this article: Viner C, Dorman SN, Shirley BC and Rogan PK (2014) **Validation of predicted mRNA splicing mutations using high-throughput transcriptome data [v1; ref status: indexed, http://f1000r.es/2no]** *F1000Research* 2014, **3**:8 (doi: 10.12688/f1000research.3-8.v1)

Grant information: PKR is supported by the Canadian Breast Cancer Foundation, Canadian Foundation for Innovation, Canada Research Chairs Secretariat and the Natural Sciences and Engineering Research Council of Canada (NSERC Discovery Grant 371758-2009). SND received fellowships from the Ontario Graduate Scholarship Program, the Pamela Greenaway-Kohlmeier Translational Breast Cancer Research Unit, and the CIHR Strategic Training Program in Cancer Research and Technology Transfer, and the University of Western Ontario.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: PKR is the inventor of US Patent 5,867,402 and other patents pending, which underlie the prediction and validation of mutations. He is one of the founders of Cytognomix, Inc. which is developing software based on this technology for complete genome or exome splicing mutation analysis. BCS is an employee of Cytognomix, Inc. CV and SND declare that they have no competing interests. Both trial and licensed versions of Veridical are fully functional, however the length of the trial period is limited.

First published: 13 Jan 2014, **3**:8 (doi: 10.12688/f1000research.3-8.v1)
First indexed: 07 Mar 2014, **3**:8 (doi: 10.12688/f1000research.3-8.v1)

Introduction

DNA variant analysis of complete genome or exome data has typically relied on filtering of alleles according to population frequency and alterations in coding of amino acids. Numerous variants of unknown significance (VUS) in both coding and non-coding gene regions cannot be categorized with these approaches. To address these limitations, *in silico* methods that predict biological impact of individual sequence variants on protein coding and gene expression have been developed, which exhibit varying degrees of sensitivity and specificity[1]. These approaches have generally not been capable of objective, efficient variant analysis on a genome-scale.

Splicing variants, in particular, are known to be a significant cause of human disease[2–5] and indeed have even been hypothesized to be the most frequent cause of hereditary disease[6]. Computational identification of mRNA splicing mutations within DNA sequencing (DNA-Seq) data has been implemented to varying degrees of sensitivity, with most software only evaluating conservation solely at the intronic dinucleotides adjacent to the junction (i.e.7). Other approaches are capable of detecting significant mutations at other positions with constitutive, and in certain instances, cryptic, splice sites[5,8,9] which can result in aberrations in mRNA splicing. Presently, only information theory-based mRNA splicing mutation analysis has been implemented on a genome scale[10]. Splicing mutations can abrogate recognition of natural, constitutive splice sites (inactivating mutation), weaken their binding affinity (leaky mutation), or alter splicing regulatory protein binding sites that participate in exon definition. The abnormal molecular phenotypes of these mutations comprise: (a) complete exon skipping, (b) reduced efficiency of splicing, (c) failure to remove introns (also termed intron retention or intron inclusion), or (d) cryptic splice site activation, which may define abnormal exon boundaries in transcripts using non-constitutive, proximate sequences, extending or truncating the exon. Some mutations may result in combinations of these molecular phenotypes. Nevertheless, novel or strengthened cryptic sites can be activated independently of any direct effect on the corresponding natural splice site. The prevalence of these splicing events has been determined by ourselves and others[5,11–13]. The diversity of possible molecular phenotypes makes such aberrant splicing challenging to corroborate at the scale required for complete genome (or exome) analyses. This has motivated the development of statistically robust algorithms and software to comprehensively validate the predicted outcomes of splicing mutation analysis.

Putative splicing variants require empirical confirmation based on expression studies from appropriate tissues carrying the mutation, compared with control samples lacking the mutation. In mutations identified from complete genome or exome sequences, corresponding transcriptome analysis based on RNA sequencing (RNA-Seq) is performed to corroborate variants predicted to alter splicing. Manually inspecting a large set of splicing variants of interest with reference to the experimental samples' RNA-Seq data in a program like the Integrative Genomics Viewer (IGV)[14], or simply performing database searches to find existing evidence would be time-consuming for large-scale analyses. Checking control samples would be required to ensure that the variant is not a result of alternative splicing, but is actually causally linked to the variant of interest. Manual inspection of the number of control samples required for statistical power to verify that each displays normal splicing would be laborious and does not easily lend itself to statistical analyses. This may lead to either missing contradictory evidence or to discarding a variant due to the perceived observation of statistically insignificant altered splicing within control samples. In addition, a list of putative splicing variants returned by variant prediction software can often be extremely large. The validation of such a significant quantity of variants may not be feasible, for example, in certain types of cancer, in instances where the genomic mutational load is high and only manual annotation is performed. We have therefore developed Veridical, a software program that automatically searches all given experimental and control RNA-Seq data to validate DNA-derived splicing variants. When adequate expression data are available at the locus carrying the mutation, this approach reveals a comprehensive set of genes exhibiting mRNA splicing defects in complete genomes and exomes. Veridical and its associated software programs are available at: www.veridical.org.

Methods

The program Veridical was developed to allow high-throughput validation of predicted splicing mutations using RNA sequencing data. Veridical requires at least three files to operate: a DNA variant file containing putative mRNA splicing mutations, a file listing of corresponding transcriptome (RNA-Seq) BAM files, and a file annotating exome structure. A separate file listing RNA-Seq BAM files for control samples (i.e. normal tissue) can also be provided. Here, we predict mutations in a set of breast tumours which are validated with RNA-Seq data from the same individuals, with RNA-Seq data from normal breast tissues as controls. However, in principle, potential splicing mutations for any disease state with available RNASeq data can be investigated. In each tumour, every variant is analyzed by checking the informative sequencing reads from the corresponding RNA-Seq experiment for non-constitutive splice isoforms, and comparing these results with the same type of data from all other tumour and normal samples that do not carry the variant in their exomes.

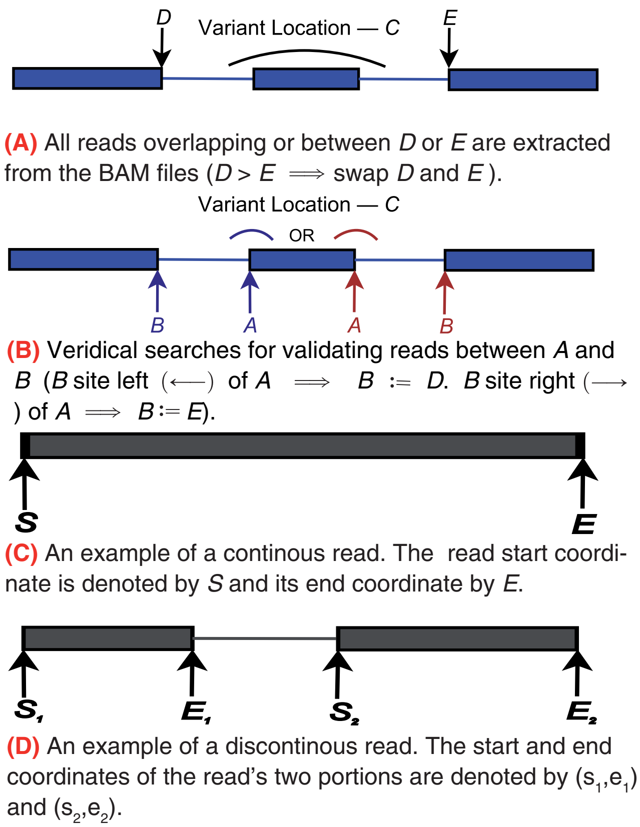
Veridical concomitantly evaluates control samples, providing for an unbiased assessment of splicing variants of potentially diverse phenotypic consequences. Note that control samples include all non-variant containing files, as well any normal samples provided. Maximizing the set of control samples, while computationally more expensive, increases the statistical robustness of the results obtained.

For each variant, Veridical directly analyzes sequence reads aligned to the exons and introns that are predicted to be affected by the genomic variant. We elected to avoid indirect measures of exon skipping, such as loss of heterozygosity in the transcript, because of the possibility of confusion with other molecular etiologies (i.e. deletion or gene conversion), unrelated to the splicing mutations. The nearest natural site is found using the exome annotation file provided, based upon the directionality of the variant, as defined within Table 1. The genomic coordinates of the neighboring exon boundaries are then found and the program proceeds, iterating over all known transcript variants for the given gene. A diagram of this procedure is provided in Figure 1. The variant location, *C*, is specifically referring to the, variant-induced, location of the predicted

Table 1. Definitions used within Veridical to determine the direction in which reads are checked. ***A*** **and** ***B*** **represent natural site positions, defined in Figure 1(b).**

Pertinent Splice Site			
A	***B***	**Strand**	**Direction**
Exonic	Donor$^{\alpha}$	+	→
Exonic	Donor$^{\alpha}$	-	←
Intronic	Acceptor$^{\beta}$	+	←
Intronic	Acceptor$^{\beta}$	-	→

$^{\alpha}$ – 5′ splice site $^{\beta}$ – 3′ splice site



(A) All reads overlapping or between D or E are extracted from the BAM files ($D > E \implies$ swap D and E).

(B) Veridical searches for validating reads between A and B (B site left ($\longleftarrow$) of $A \implies B := D$. B site right ($\longrightarrow$) of $A \implies B := E$).

(C) An example of a continous read. The read start coordinate is denoted by S and its end coordinate by E.

(D) An example of a discontinous read. The start and end coordinates of the read's two portions are denoted by (s_1,e_1) and (s_2,e_2).

Figure 1. A diagram portraying the definitions used within Veridical to specify genic variant position and read coordinates. We employ the same conventions as IGV[14]. Blue lines denote genes, wherein thick lines represent exons and thin lines represent introns. Grey lines denote reads, wherein thick lines denote a read mapping to some particular location in the genome and thin lines represent connecting segments of reads that are split across spliced-in regions (i.e. exons or included introns).

mRNA splice site, which is often proximate to, but distinct from the coordinate of the actual genomic mutation itself.

The program uses the BamTools API[15] to iterate over all of the reads within a given genomic region across experimental and control samples. Individual reads are then assessed for their corroborating value towards the analysis of the variant being processed, as outlined in the flowchart in Figure 2. Validating reads are based on whether they alter either the location of the splice junction (i.e. junction-spanning) or the abundance of the transcript, particularly in intronic regions (i.e. read-abundance). Junction-spanning reads contain DNA sequences from two adjacent exons or are reads that extend into the intron. These reads directly show whether the intronic sequence is removed or retained by the spliceosome, respectively. Read-abundance validated reads are based upon sequences predicted to be found in the mutated transcript in comparison with sequences that are expected to be excised from the mature transcript in the absence of a mutation. Both types of reads can be used to validate cryptic splicing, exon skipping, or intron inclusion. A read is only said to corroborate cryptic splicing if and only if the variant under consideration is expected to activate cryptic splicing. Junction-spanning, cryptic splicing reads are those in which a read is exactly split from the cryptic splice site to the adjacent exon junction. For read-abundance cryptic splicing, we define the concept of a read fraction, which is the ratio of the number of reads corroborating the cryptically spliced isoform and the number of reads that do not support the use of the cryptic splice site (i.e. non-cryptic corroborating) in the same genomic region of a sample. Cryptic corroborating reads are those which occur within the expected region where cryptic splicing occurs (i.e. spliced-in regions). This region is bounded by the variant splice site location and the adjacent (direction dependent) splice junction. Non-cryptic corroborating reads, which we also term "anti-cryptic" reads, are those that do not lie within this region, but would still be retained within the portion that would be excised, had cryptic splicing occurred. To identify instances of exon skipping, Veridical only employs junction-spanning reads. A read is considered to corroborate exon skipping if the connecting read segments are split such that it connects two exon boundaries, skipping an exon in between. A read is considered to corroborate intron inclusion when the read is continuous and either overlaps with the intron-exon boundary (and is then said to be junction-spanning) or if the read is within an intron (and is then said to be based upon read-abundance).

We proceed to formalize the above descriptions as follows. A given read is denoted by r, with start and end coordinates (r_s, r_e), if the read is continuous, as diagrammed within Figure 1(c), or otherwise, with start and end coordinate pairs, (r_{s_1}, r_{e_1}) and (r_{s_2}, r_{e_2}), as diagrammed within Figure 1(d). Let l be the length of the read. The set ζ denotes the totality of validating reads. The criterion for $r \in \zeta$ is detailed below. It is important to note that validating reads are necessary but not sufficient to validate a variant. Sufficiency is achieved only if the number of validating reads is statistically significant relative to those present in control samples. ζ itself is partitioned into three sets: ζ_c, ζ_e, and ζ_i for evidence of cryptic splicing, exon skipping, and intron inclusion, respectively. We allow partitions to be empty. Let S denote the relevant splice junctions as defined by Figure 1 and Table 1. Without loss of generality, we consider only the red (i.e. direction is right) set of labels within Figure 1(b). Then the (splice consequence) partitions of ζ are given by:

$$r \in \zeta_c \iff \text{variant is cryptic} \wedge r_{s_2} - r_{e_1} = C - S \vee (r_s > A \wedge r_e < S)$$

$$r \notin \zeta_c \wedge \text{variant is cryptic} \wedge \neg\,(\,r_{s_2} - r_{e_1} = C - S) \Rightarrow r \in \mathit{anticryptic}$$

$$r \in \zeta_e \iff (r_{e_1} = A \wedge r_{s_2} = B)$$

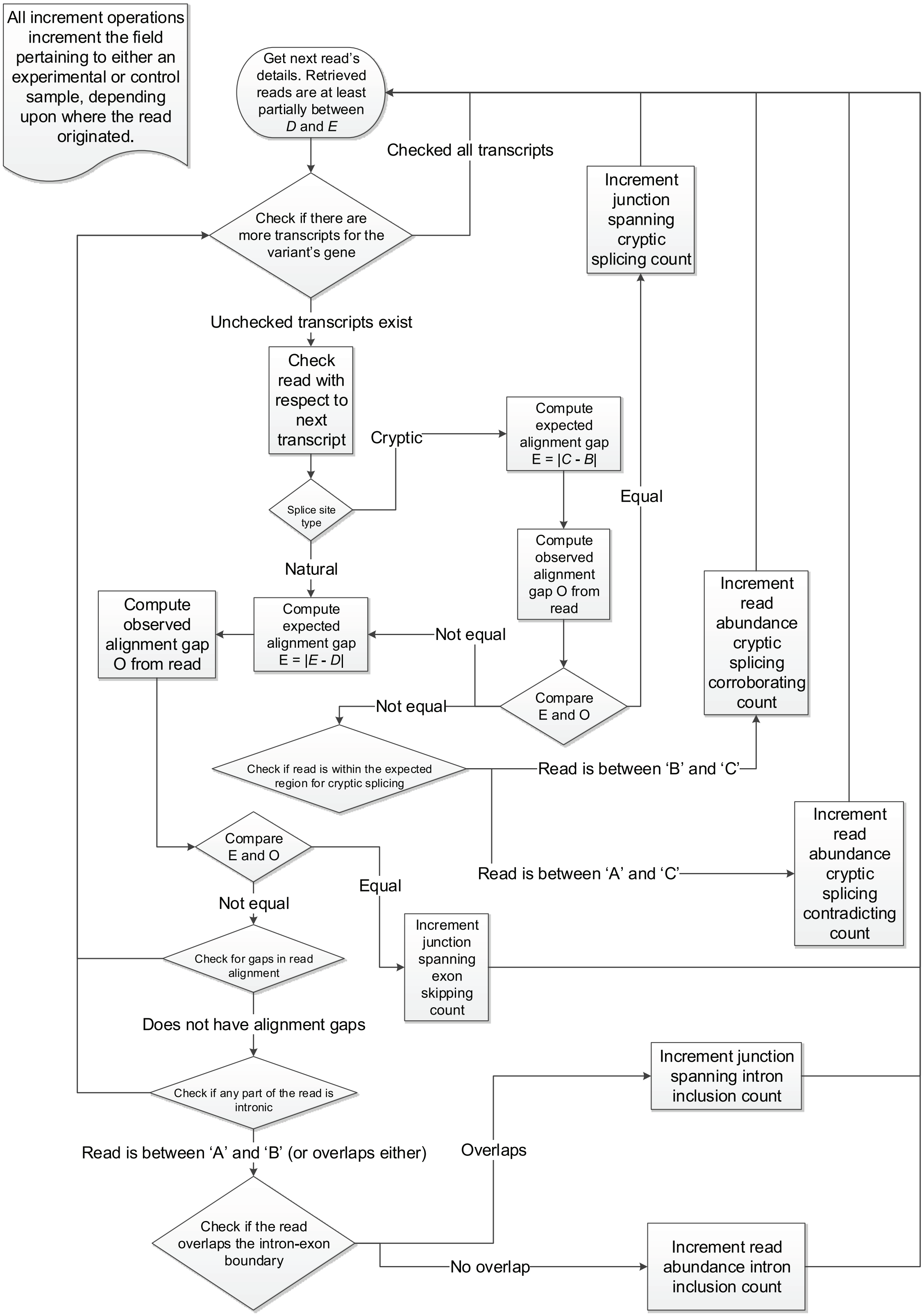

All increment operations increment the field pertaining to either an experimental or control sample, depending upon where the read originated.
Get next read's details. Retrieved reads are at least partially between *D* and *E*
Checked all transcripts
Check if there are more transcripts for the variant's gene
Unchecked transcripts exist
Check read with respect to next transcript
Cryptic
Compute expected alignment gap E = |*C* - *B*|
Splice site type
Natural
Compute observed alignment gap O from read
Compute expected alignment gap E = |*E* - *D*|
Compute observed alignment gap O from read
Not equal
Compare E and O
Equal
Increment junction spanning cryptic splicing count
Not equal
Check if read is within the expected region for cryptic splicing
Read is between 'B' and 'C'
Increment read abundance cryptic splicing corroborating count
Read is between 'A' and 'C'
Increment read abundance cryptic splicing contradicting count
Compare E and O
Equal
Not equal
Increment junction spanning exon skipping count
Check for gaps in read alignment
Does not have alignment gaps
Check if any part of the read is intronic
Read is between 'A' and 'B' (or overlaps either)
Check if the read overlaps the intron-exon boundary
Overlaps
Increment junction spanning intron inclusion count
No overlap
Increment read abundance intron inclusion count

Figure 2. The algorithm employed by Veridical to validate variants. Refer to Table 1 for definitions concerning direction and Figure 1 for variable definitions.

$r \in \zeta_i \Longleftrightarrow (A \in [r_s, r_e] \vee B \in [r_s, r_e]) \vee [(A \notin [r_s, r_e] \wedge B \notin [r_s, r_e]) \wedge (r_s < A - l \wedge r_e > B \wedge r \notin (A, B))]$

We separately partition ζ by its evidence type, the set of junction-spanning reads, δ and read-abundance reads, α, as follows:

$r \in \delta \Longleftrightarrow (A \in [r_s, r_e] \vee B \in [r_s, r_e]) \vee [r \in \zeta_c \wedge r_{s_2} - r_{e_1} = C - S]$.
$r \in \alpha \Longleftrightarrow r \notin \delta$

Once all validating reads are tallied for both the experimental and control samples, a p-value is computed. This is determined by computing a z-score upon Yeo-Johnson (YJ)[16] transformed data. This transformation, shown in Equation 1, ensures that the data is sufficiently normally distributed to be amenable to parametric testing.

$$\psi(x,\lambda) = \begin{cases} \dfrac{(x+1)^{\lambda} - 1}{\lambda} & \text{if } x \geq 0 \ \wedge \lambda \neq 0 \\ \log(x+1) & \text{if } x \geq 0 \ \wedge \lambda = 0 \\ -\dfrac{(-x+1)^{2-\lambda} - 1}{2-\lambda} & \text{if } x < 0 \ \wedge \lambda \neq 2 \\ -\log(-x+1) & \text{if } x < 0 \ \wedge \lambda = 2 \end{cases} \qquad (1)$$

The transform is similar to the Box-Cox power transformation, but obviates the requirement of inputting strictly positive values and has more desirable statistical properties. Furthermore, this transformation allowed us to avoid the use of non-parametric testing, which has its own pitfalls regarding assumptions of the underlying data distribution[17]. We selected $\lambda = \frac{1}{2}$, because Veridical's untransformed output is skewed left, due to their being, in general, less validating reads in control samples and the fact that there are, by design, vastly more control samples than experimental samples. We found that this value for λ generally made the distribution much more normal.

A comparison of the distributions of untransformed and transformed data is provided in Figure S1. We were not concerned about small departures from normality as a z-test with a large number of samples is robust to such deviations[18]. It is important to realize, therefore, that the p-values given by Veridical are much more robust when the program is provided with a large number of samples.

Thus, we can compute the p-value of the pairwise unions of the two sets of partitions of ζ, except the irrelevant $\zeta_e \cup \alpha = \varnothing$. We only provide p-values for these pairwise unions and do not attempt to provide p-values for the partitions for the different consequences of the mutations on splicing. While such values would be useful, we do not currently have a robust means to compute them. Our previous work provides guidance on interpretation of splicing mutation outcomes[3–5,10]. Thus for $\zeta_x \in \{\zeta_c, \zeta_e, \zeta_i\}$, let $\Phi_Z(z)$ represent the cumulative distribution function of the one-sided (right-tailed — i.e. $P[X > x]$) standard normal distribution. Let N represent the total number of samples and let V represent the set of all ζ_x validations, across all samples. Then:

$$\mu = \frac{\sum_{j=1}^{N} V_j}{N} \qquad \sigma = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(V_j - \overline{V})^2}$$

$$z = \frac{|\zeta_x| - \mu}{\sigma} \qquad p = \Phi(\psi(z, \frac{1}{2}))$$

The program outputs two tables, along with summaries thereof. The first table lists all validated read counts across all categories for experimental samples, while the second table does the same for the control samples. P-values are shown in parentheses within the experimental table, which refer to the column-dependent (i.e. the category is given in the column header) p-value for that category with respect to that same category in control samples. The program produces three files: a log file containing all details regarding validated variants, an output file with the programs progress reports and summaries, and a filtered validated variant file. The filtered file contains all validated variants of statistical significance (set as $p < 0.05$, by default), defined as variants with one or more validating read categories which achieve statistical significance in a relevant category (i.e. a cryptic variant for which $p = 0.04$ in the junction-spanning cryptic column would meet this criteria).

We elected to use RefSeq[19] genes for the exome annotation, as opposed to, the more permissive exome annotation sets, UCSC Known Genes[20] or Ensembl[21]. The large number of transcript variants within Ensembl, in particular, caused many spurious intron inclusion validation events. This occurred because reads were found to be intronic in many cases, when in actuality they were exonic with respect to the more common transcript variant. In addition, the inclusion of the large number of rare transcripts in Ensembl significantly increased program runtime and made validation events much more challenging to interpret unequivocally. The use of RefSeq, which is a conservative annotation of the human exome, resolves these issues.

We also provide an R program[22] which produces publication quality histograms displaying embedded Q-Q plots and p-values, to evaluate for normality of the read distribution and statistical significance, respectively. The R program performs the YJ transformation as implemented in the `car` package[23]. The histograms generated by the program use the Freedman-Draconis[24] rule for break determination, and the Q-Q plots use algorithm Type 8 for their quantile function, as recommended by Hyndman and Fan[25]. Lastly, a Perl program was implemented to automatically retrieve and correctly format an exome annotation file from the UCSC database[20] for use in Veridical. All data uses hg19/GRCh37, however when new versions of the genome become available, this program can be used to update the annotation file.

Results

Veridical validates predicted mRNA splicing mutations using high-throughput RNA sequencing data. The performance of the software is affected by the number of predicted splicing mutations, the number of abnormal samples containing mutations and control samples and the corresponding RNA-Seq data for each type of sample. Veridical has the ability to analyze approximately 3000 variants in approximately 4 hours assuming an input of 100 BAM files of RNA-Seq data. The relationship between time and numbers of BAM files and variants are plotted in Figure 3 for a 2.27 GHz processor. Veridical uses memory in linear proportion to the number and size of the input BAM files. In our tests, using RNA-Seq BAM files with an average size of approximately 6 GB, Veridical used approximately 0.7 GB for ten files to 1 GB for 100 files. Currently, splicing consequences that are reported include intron inclusion, exon skipping, and cryptic splicing, which are validated through junction-spanning reads, or based on read-abundance in the region circumscribing the variant (see Methods for details). For example, a cryptic splicing junction-spanning read will show that the mRNA contains a truncated or extended exon at the predicted location, which is directly attached to the sequence of the corresponding adjacent exon. For mutations that alter read-abundance, each read within the genomic location assessed (i.e. intron for intron inclusion) is counted for the variant-containing samples and then compared with the number of reads in the control files. For each input variant, Veridical outputs the number of validating reads (i.e. RNA-Seq reads which corroborate the predicted splicing consequence) for a given splice consequence within the variant-containing tumour samples and within control samples (i.e. non-variant containing tumour samples and normal samples). The program provides read counts for the different categories for all experimental and control samples as tab-delimited tables, along with the relevant p-values, indicating the statistical probability that the predicted mutation exhibits a normal expression pattern.

We demonstrate how Veridical and its associated R program are used to validate predicted splicing mutations in somatic breast cancer. Each example depicts a particular variant-induced splicing consequence, analyzed by Veridical, with its corresponding significance level. The relevant primary RNA-Seq data are displayed in IGV, along with histograms and Q-Q plots showing the read distributions for each example. The source data are obtained from controlled-access breast carcinoma data from The Cancer Genome Atlas (TCGA)[28]. Tumour-normal matched DNA sequencing data from the TCGA consortium was used to predict a set of splicing mutations, and a subset of corresponding RNA sequencing data was analyzed to confirm these predictions with Veridical. The following examples demonstrate the utility of Veridical to identify potentially pathogenic mutations from a much larger subset of predicted variants.

Input, output, and explanatory files for Veridical

5 Data Files

http://dx.doi.org/10.6084/m9.figshare.894971

Leaky mutations are those variants that reduce, but not abolish, the spliceosome's ability to recognize the intron/exon boundary[3]. This can lead to the mis-splicing (intron inclusion and/or exon skipping) of many but not all transcripts. An example of a predicted leaky mutation (chr5:162905690 G>T) in the *HMMR* gene in which both junction-spanning exon skipping ($p < 0.01$) and read-abundance-based intron inclusion ($p = 0.04$) are observed is provided within Figure 4. We predict this mutation to be leaky because its final R_i exceeds 1.6 bits — the minimal individual information required to recognize a splice site and produce correctly spliced mRNA[4]. Indeed, the natural site, while weakened by 2.16 bits, remains strong — 10.67 bits. This prediction is validated by the variant-containing sample's RNA-Seq data (Figure 4), in which both exon skipping (5 reads) and intron inclusion (14 reads) are observed, along with 70 reads portraying wild-type splicing. Only a single normally spliced read contains the G→T mutation. These results are consistent with an imbalance of expression of the two alleles,

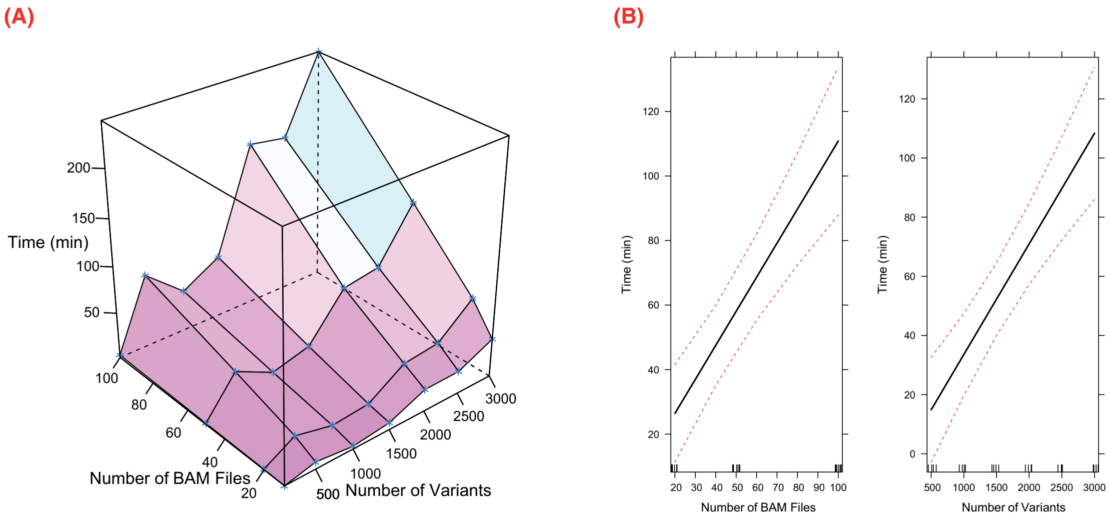


Figure 3. Profiling data for Veridical runtime. Tests were conducted upon an Intel Xeon @2.27 GHz. Visualizations were generated with R[22] using Lattice[26] and Effects[27]. A surface plot of time vs. numbers of BAM files and variants is provided in (**A**). Effect plots are given in (**B**) and demonstrate the effects of the numbers of BAM files and variants upon runtime. The effect plots were generated using a linear regression model (R^2 = 0.7525).

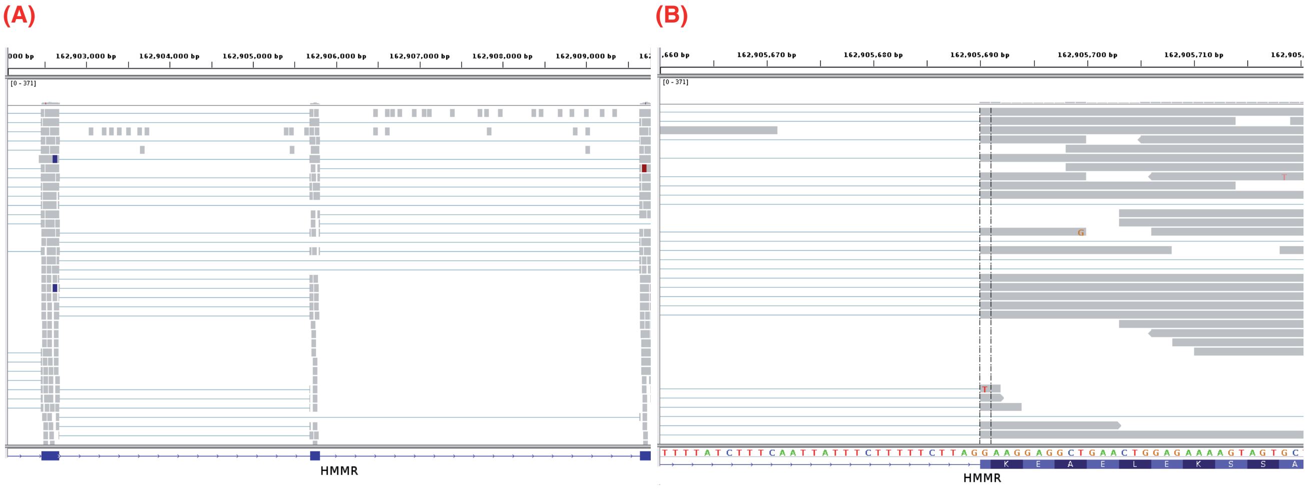


Figure 4. IGV images depicting a predicted leaky mutation (chr5:162905690 G>T) within the natural acceptor site of exon 12 (162905689–162905806) of *HMMR*. This gene has four transcript variants and the given exon number pertains to isoforms *a* and *b* (reference sequences NM_001142556 and NM_012484). RNA-Seq reads are shown in the centre panel. The bottom blue track depicts RefSeq genes, wherein each blue rectangle denotes an exon and blue connecting lines denote introns. In the middle panel, each rectangle (grey by default) denotes an aligned read, while thin lines are segments of reads split across exons. Red and blue coloured rectangles in the middle panel denote aligned reads of inserts that are larger or smaller than expected, respectively. (**A**) depicts a genomic region of chromosome 5: 162902054–162909787. The variant occurs in the middle exon. Intron inclusion can be seen in this image, represented by the reads between the first and middle exon (since the direction is right, as described within Table 1). These 14 reads are read-abundance-based, since they do not span the intron-exon junction. (**B**) depicts a closer view of the region shown in (**A**) — 162905660–162905719. The dotted vertical black lines are centred upon the first base of the variant-containing exon. The thin lines in the middle panel that span the entire exon fragment are evidence of exon skipping. These 5 reads are split across the exon before and after the variant-containing exon, as seen in (**A**).

as expected for a leaky variant. Figure 5 shows that for the distribution of read-abundance-based intron inclusion is statistically significant ($p = 0.04$).

Variants that inactivate splice sites have negative final R_i values[3] with only rare exceptions[4], indicating that splice site recognition is essentially abolished in these cases. We present the analysis of two inactivating mutations within the *PTEN* and *TMTC2* genes from different tumour exomes, namely: chr10:89711873 A>G and chr12:83359523 G>A, respectively. The *PTEN* variant displays junction-spanning exon skipping events ($p < 0.01$), while the *TMTC2* gene portrays both junction-spanning and read-abundance-based intron inclusion (both splicing consequences with $p < 0.01$). In addition, all intron inclusion reads in the experimental sample contain the mutation itself, while only one such read exists across all control samples analyzed ($p < 0.01$). The *PTEN* variant contains numerous exon skipping reads (32 versus an average of 2.466 such reads per control sample). The *TMTC2* variant contains many junction-spanning intron inclusion reads with the G→A mutation (all of its junction-spanning intron inclusion reads: 22 versus an average of 0.002 such reads per control sample). IGV screenshots for these variants are provided within Figure 6. This figure also shows an example of junction-spanning cryptic splice site activated by the mutation (chr1:985377 C>T) within the *AGRN* gene. The concordance between the splicing outcomes generated by these mutations and the Veridical results indicates that the proposed method detects both mutations that inactivate splice sites and cryptic splice site activation.

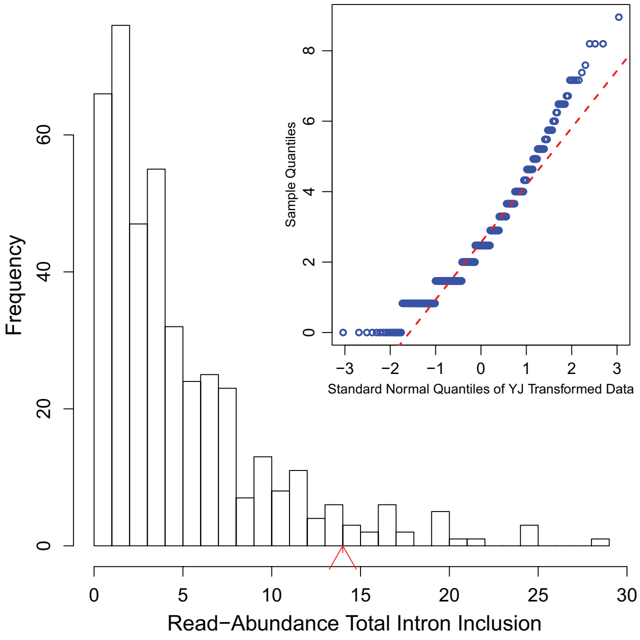


Figure 5. Histogram of read-abundance-based intron inclusion with embedded Q-Q plots of the predicted leaky mutation (chr5:162905690 G>T) within *HMMR***,** as shown in Figure 4. The arrowhead denotes the number of reads (14 in this case) in the variant-containing file which is more than observed in the control samples ($p = 0.04$).

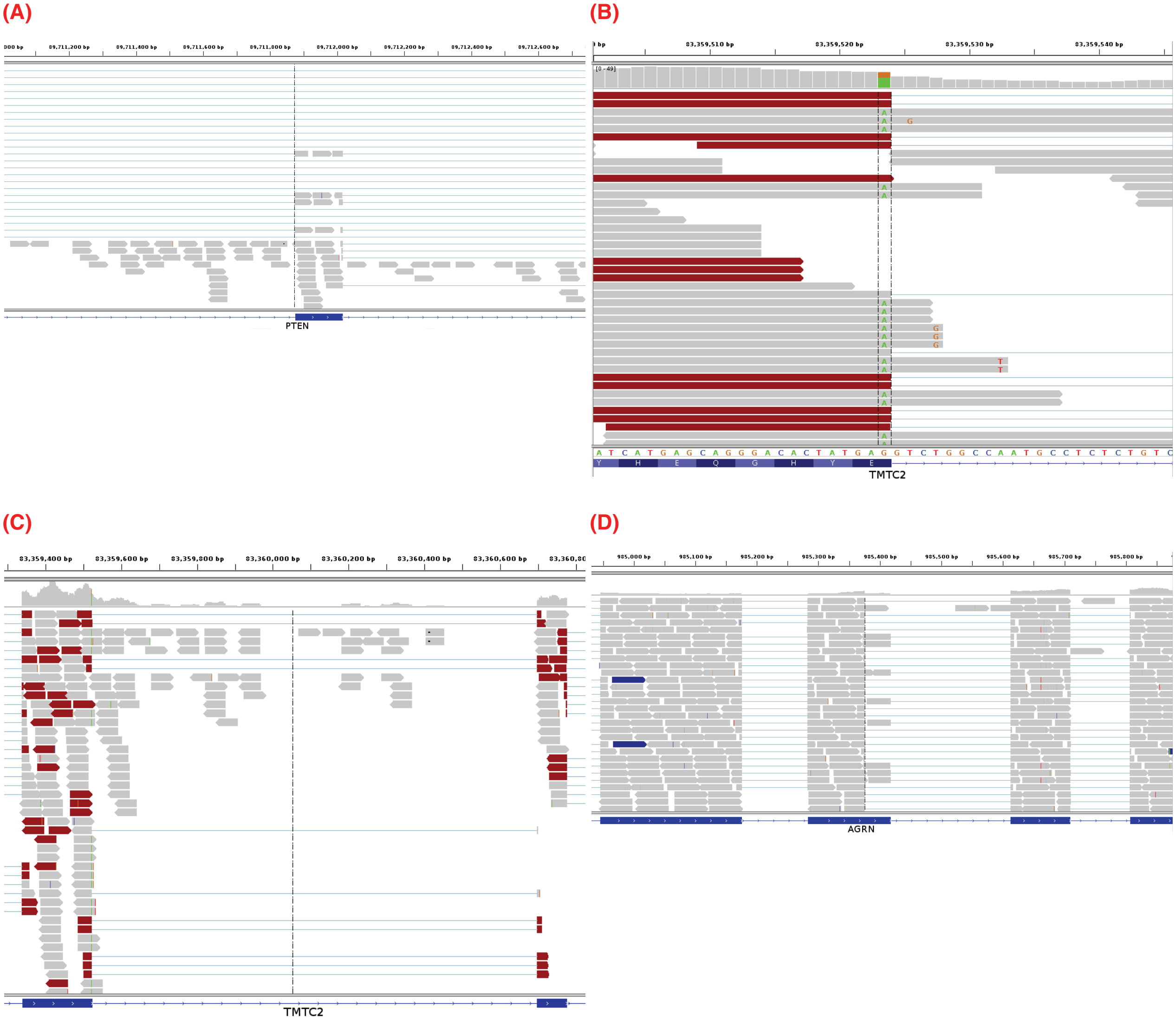


Figure 6. (**A**) depicts an inactivating mutation (chr10:89711873 A>G) within the natural acceptor site of exon 6 (89711874–89712016) of *PTEN*. The dotted vertical black line denotes the location of the relevant splice site. The region displayed is 89711004–89712744 on chromosome 10. Many of the 32 exon skipping reads are evident, typified by the thin lines in the middle panel that span the entire exon. There is also a significant amount of read-abundance-based intron inclusion, shown by the reads to the left of the dotted vertical line. Exon skipping was statistically significant ($p < 0.01$), while read-abundance-based intron inclusion was not ($p = 0.53$). Panels (**B**) and (**C**) depict an inactivating mutation (chr12:83359523 G>A) within the natural donor site of exon 6 (83359338–83359523) of *TMTC2*. (**B**) depicts a closer view (83359501–83359544) of the region shown in (**C**) and only shows exon 6. Some of the 22 junction-spanning intron inclusion reads can be seen. In this case, all of these reads contain the mutation, shown by the green adenine base in each read, between the two vertical dotted lines. (**C**) depicts a genomic region of chromosome 12: 83359221–83360885, *TMTC2* exons 6–7. The variant occurs in the left exon. 65 read-abundance-based intron inclusion can be seen in this image, represented by the reads between the two exons. Panel (**D**) depicts a mutation (chr1:985377 C>T) causing a cryptic donor to be activated within exon 27 (the second from left, 985282–985417) of *AGRN*. The region displayed is 984876–985876 on chromosome 1 (exons 26–29 are visible). Some of the 34 cryptic (junction-spanning) reads are portrayed. The dotted black vertical line denotes the cryptic splice site, at which cryptic reads end. Refer to the caption of Figure 4 for IGV graphical element descriptions.

Recurrent genetic mutations in some oncogenes have been reported among tumours within the same, or different, tissues of origin. Common recurrent mutations present in multiple abnormal samples are recognized by Veridical. This avoids including a variant-containing sample among the control group, and outputs the results of all of the variant-containing samples. A relevant example is shown in Figure 7. The mutation (chr1:46726876 G>T) causes activation of a cryptic splice site within *RAD54L* in multiple tumours. Upon computation of the p-values for each of the variant-containing tumours, relative to all non-variant containing tumours and normal controls, not all variant-containing tumours displayed splicing abnormalities at statistically significant levels. Of the six variant-containing

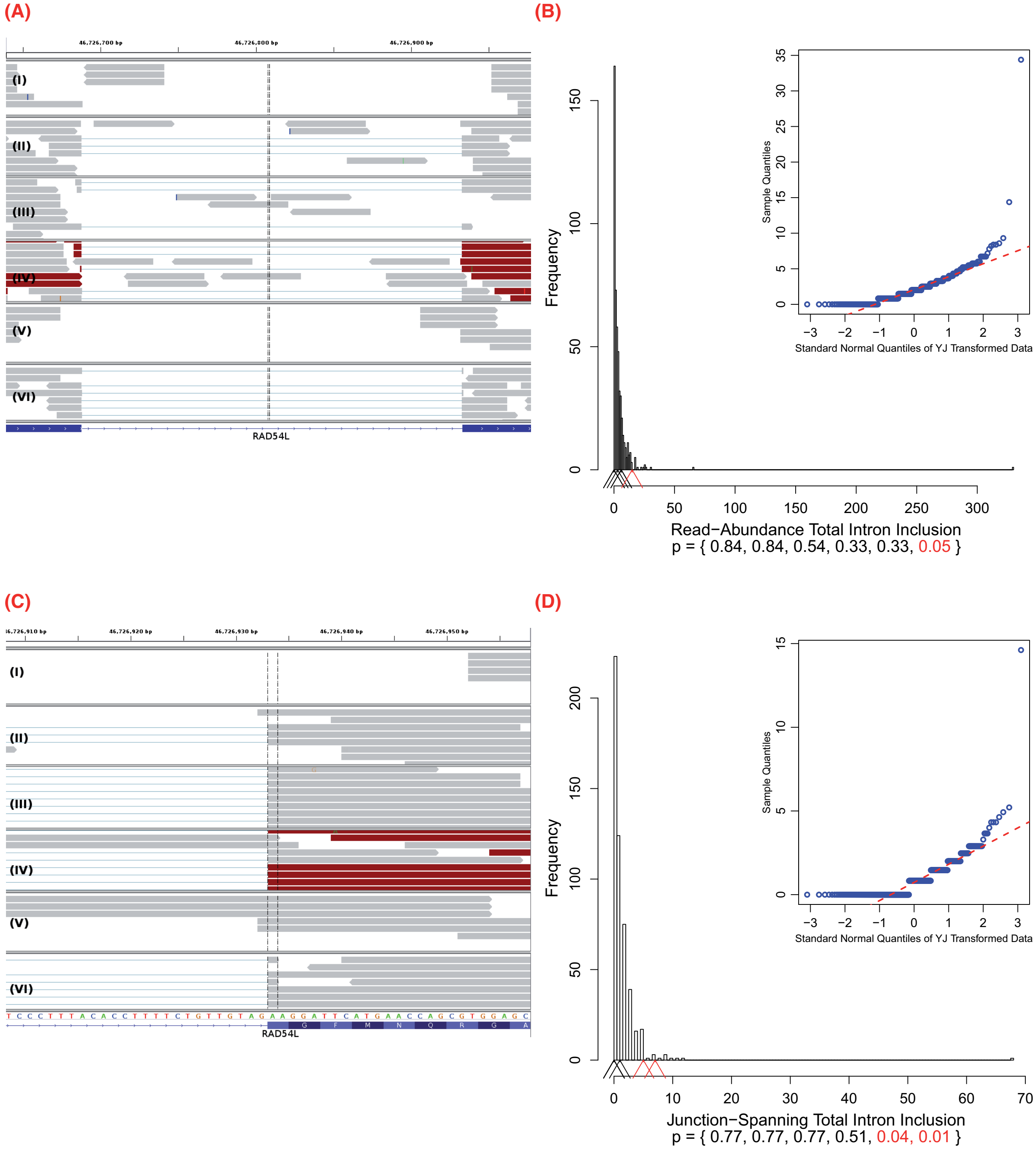


Figure 7. IGV images and their corresponding histograms with embedded Q-Q plots depicting all six variant containing files with a mutation (chr1:46726876 G>T) which, in some cases, causes a cryptic donor to be activated within the intron between exons 7 and 8 of *RAD54L*. This results in the extension of the downstream natural donor (the 5´ end of exon 8). This gene has two transcript variants and the given exon numbers pertain to isoform *a* (reference sequence NM_003579). Only samples IV and V have statistically significant intron inclusion relative to controls. read-abundance-based intron inclusion can be seen in (**A**), between the two exons. The region displayed is on chromosome 1: 46726639–46726976. (**B**) depicts the corresponding histogram for the 15 read-abundance-based intron inclusion reads (p = 0.05) that are present in sample IV. The intron-exon boundary on the right is the downstream natural donor. (**C**) typifies some of the 13 junction-spanning intron inclusion reads that are a direct result of the intronic cryptic site's activation. In these instances, reads extending past the intron-exon boundary are being spliced at the cryptic site, instead of the natural donor. In particular, samples IV and V both have a statistically significant numbers of such reads, 7 (p = 0.01) and 5 (p = 0.04), respectively. This is further typified by the corresponding histogram in (**D**). (**C**) focuses upon exon 8 from (**A**) and displays the genomic positions 46726908–46726957. Refer to the caption of Figure 4 for IGV graphical element descriptions. In the histograms, arrowheads denote numbers of reads in the variant-containing files. The bottom of the plots provide p-values for each respective arrowhead. Statistically significant p-values and their corresponding arrowheads are denoted in red.

tumours, two had significant levels of junction-spanning intron inclusion, and one showed statistically significant read-abundance-based intron inclusion. Details for all of the aforementioned variants, including a summary of read counts pertaining to each relevant splicing consequence, for experimental versus control samples, are provided in Table 2.

Discussion

We have implemented Veridical, a software program that automates confirmation of mRNA splicing mutations by comparing sequence read-mapped expression data from samples containing variants that are predicted to cause defective splicing with control samples lacking these mutations. The program objectively evaluates each mutation with statistical tests that determine the likelihood of and exclude normal splicing. To our knowledge, no other software currently validates splicing mutations with RNA-Seq data on a genome-wide scale, although many applications can accurately detect conventional alternative splice isoforms (i.e.29). Veridical is intended for use with large data sets derived from many samples, each containing several hundred variants that have been previously prioritized as likely splicing mutations, regardless of how the candidate mutations are selected. It is not practical to analyze all variants present in an exome or genome, rather only a filtered subset, due to the extensive computations required for statistical validation. As such, Veridical is a key component of an end-to-end, hypothesis-based, splicing mutation analysis framework that also includes the Shannon splicing mutation pipeline[10] and the Automated Splice Site Analysis and Exon Definition server[5].

There is a trade-off between lengthy run-times and statistical robustness of Veridical, especially when there are either a large number of variants or a large number of RNA-Seq files. As with most statistical methods, those employed here are not amenable to small sample sets, but become quite powerful when a large number of controls are employed. In order to ensure that mutations can be validated, we recommend an excess of control transcriptome data relative to those from samples containing mutations (> 5 : 1), regardless of the computational expense. We *do not* recommend the use of a single control to corroborate a sample containing a putative mutation. Not surprisingly, we have found that junction-spanning reads have the greatest value for corroborating cryptic splicing and exon skipping. Even a single such read is almost always sufficient to merit the validation a variant, provided that sufficient control samples are used. For intron inclusion, both junction-spanning and read-abundance-based reads are useful and a variant can readily be validated with either, provided that the variant-containing experimental sample(s) show a statistically significant increase in the presence of either form of intron inclusion corroborating reads.

Veridical is able to automatically process variants from multiple different experimental samples, and can group the variant information if any given mutation is present in more than one sample. The use of a large sample size allows for robust statistical analyses to be performed, which aid significantly in the interpretation of results. The main utility of Veridical is to filter through large data sets of predicted splicing mutations to prioritize the variants. This helps to predict which variants will have a deleterious effect upon the protein product. Veridical is able to avoid reporting splicing changes that are naturally occurring through checking all variant-containing and non-containing control samples for the predicted splicing consequence. In addition, running multiple tumour samples at once allows for manual inspection to discover samples that contained the alternative splicing pattern, and consequently, permits the identification of DNA mutations in the same location which went undetected during genome sequencing.

The statistical power of Veridical is dependent upon the quality of the RNA-Seq data used to validate putative variants. In particular, a lack of sufficient coverage at a particular locus will cause Veridical to be unable to report any significant results. A coverage of at least 20 reads should be sufficient. This estimate is based upon alternative splicing analyses in which this threshold was found to imply concordance with microarray and RT-PCR measurements[30–33]. There are many potential legitimate reasons why a mutation may not be validated: (a) nonsense-mediated decay may result in a loss of expression of the entire transcript, (b) the gene itself may have multiple paralogs and reads may not be unambiguously mapped, (c) other non-splicing mutations could account for a loss of expression, and (d) confounding natural alternative splicing isoforms may result in a loss of statistical significance during read mapping of the control samples. The prevalence of loci with insufficient data is dependent upon the coverage of the sequencing technology used. As sequencing technologies improve, the proportion of validated mutations is expected to increase. Such an increase would mirror that observed for the prevalence of alternative splicing events[34]. It is important to note that acceptance of the null hypothesis, due to an absence of evidence required to disprove it, does not imply that the underlying prediction of a mutation at a particular locus is incorrect, but merely that the current empirical methods employed were insufficient to corroborate it.

While there is considerable prior evidence for splicing mutations that alter natural and cryptic splice site recognition, we were somewhat surprised at the apparent high frequency of statistically significant intron inclusion revealed by Veridical. In fact, evidence indicates that a significant portion of the genome is transcribed[34], and it is estimated that 95% of known genes are alternatively spliced[30]. Defective mRNA splicing can lead to multiple alternative transcripts including those with retained introns, cassette exons, alternate promoters/terminators, extended or truncated exons, and reduced exons[35]. In breast cancer, exon skipping and intron retention were observed to be the most common form of alternative splicing in triple negative, non-triple negative, and HER2 positive breast cancer[36]. In normal tissue, intron retention and exon skipping has been predicted to affect 2572 exons in 2127 genes and 50 633 exons in 12 797 genes, respectively[37]. In addition, previous studies suggest that the order of intron removal can influence the final mRNA transcript composition of exons and introns[38]. Intron inclusion observed in normal tissue may result from those introns that are removed from the transcript at the end of mRNA splicing. Given that these splicing events are relatively common in normal tissues, it becomes all the more important to distinguish expression patterns that are clearly due to the effects of splicing mutations — one of the guiding principles of the Veridical method.

Table 2. **Examples of variants validated by Veridical. Header abbreviations Chr, C_v, C_s, #, SC, and ET, denote chromosome, variant coordinate, splice site coordinate, sample number (where applicable), splicing consequence, and evidence type, respectively.** Headers containing R with some subscript denote numbers of validated reads for the specified variant's splicing consequence(s) and evidence type(s). R_E denotes reads within variant-containing tumour samples. R_T and R_N denote control samples, for tumours and normal cells, respectively. R_μ is the per sample mean of R_T and R_N. Splicing consequences: CS denotes cryptic splicing, ES denotes exon skipping, and II denotes intron inclusion. Evidence types: JS denotes junction-spanning and RA denotes read-abundance.

Gene	Chr	C_v	C_s	Variant	Type	Initial R_i	Final R_i	ΔR_i	#	SC	ET	p-value	R_E	R_T	R_N	R_μ	Figure
HMMR	chr5	162905690	162905689	G/T	Leaky	12.83	10.67	-2.16		ES	JS	< 0.01	5	11	0	0.020	[4],[5]
										II	RA	0.04	14	2133	103	4.051	
PTEN	chr10	89711873	89711874	A/G	Inactivating	12.09	-2.62	-14.71		ES	JS	< 0.01	32	975	386	2.466	[6(A)]
TMTC2	chr12	83359523	83359524	G/A	Inactivating	1.74	-1.27	-3.01		II	JS	< 0.01	22	2241	383	4.754	[6(B)]
										II	JSwM	< 0.01	22	0	1	0.002	
										II	RA	< 0.01	65	7293	1395	15.739	[6(C)]
AGRN	chr1	985377	985376	C/T	Cryptic	-2.24	4.79	7.03		CS	JS	< 0.01	34	97	23	0.217	[6(D)]
RAD54L	chr1	46726876	46726895	G/T	Cryptic	13.4	14.84	1.44	I	II	JS	N/A	0	645	58	1.274	[7]
										II	RA	0.54	3	2171	290	4.458	
									II	II	JS	0.51	1	645	58	1.274	
										II	RA	0.33	6	2171	290	4.458	
									III	II	JS	N/A	0	645	58	1.274	
										II	RA	0.33	6	2171	290	4.458	
									IV	II	JS	0.01	7	645	58	1.274	
										II	RA	0.05	15	2171	290	4.458	
									V	II	JS	0.04	5	645	58	1.274	
										II	RA	N/A	0	2171	290	4.458	
									VI	II	JS	N/A	0	645	58	1.274	
										II	RA	N/A	0	2171	290	4.458	

Veridical is an important analytical resource for unsupervised, thorough validation of splicing mutations through the use of companion RNA-Seq data from the same samples. The approach will be broadly applicable for many types of genetic abnormalities, and should reveal numerous, previously unrecognized, mRNA splicing mutations in exome and complete genome sequences.

Data availability

figshare: Input, output, and explanatory files for Veridical, http://dx.doi.org/10.6084/m9.figshare.894971[39].

Author contributions

PKR elaborated the problem and conceived of the analytic solution. CV implemented the algorithm, wrote, and tested the Veridical software and its accompanying Perl and R scripts. CV and SND designed the methods used, generated, and interpreted mutation results. BCS wrote and tested the Shannon pipeline, which provided the predictions evaluated by Veridical. BCS and CV implemented procedures and code for software distribution. CV, SND, and PKR wrote the manuscript, which has been approved by all of the authors.

Competing interests

PKR is the inventor of US Patent 5,867,402 and other patents pending, which underlie the prediction and validation of mutations. He is one of the founders of Cytognomix, Inc. which is developing software based on this technology for complete genome or exome splicing mutation analysis. BCS is an employee of Cytognomix, Inc. CV and SND declare that they have no competing interests. Both trial and licensed versions of Veridical are fully functional, however the length of the trial period is limited.

Grant information

PKR is supported by the Canadian Breast Cancer Foundation, Canadian Foundation for Innovation, Canada Research Chairs Secretariat and the Natural Sciences and Engineering Research Council of Canada (NSERC Discovery Grant 371758-2009). SND received fellowships from the Ontario Graduate Scholarship Program, the Pamela Greenaway-Kohlmeier Translational Breast Cancer Research Unit, and the CIHR Strategic Training Program in Cancer Research and Technology Transfer, and the University of Western Ontario.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgments

We acknowledge the TCGA consortium for providing the data used in this article. The data used in this article is part of TCGA project #988: "Predicting common genetic variants that alter the splicing of human gene transcripts". This work was made possible by the facilities of the Shared Hierarchical Academic Research Computing Network (SHARCNET) and Compute/Calcul Canada.

Supplementary materials

Veridical variant input format

This input format most easily accepts formatted output from the Shannon Pipeline. In particular, all variants of interest should be concatenated into a single file. Once a, tab-delimited, concatenated file has been generated, it can easily be formatted correctly by using **`FilterShannonPipelineResults.pl.`** One can also manually ensure the following: the header line has no quotation marks or special characters, empty columns have been replaced by a period (**.**) and each variant line contains only a single gene (comma-delimited gene lists must be split such that there is only one gene per line). If one wishes Veridical to consider variants pertaining to more than one experimental sample, a comma-delimited list of experimental samples, in the form of BAM file names, must be provided as the **`key`** column. The **`key`** column must always contain at least one file name that is present as the base name of one of the files listed in the BAM file list that must be passed to Veridical.

Alternatively, one can prepare the input format as follows. The header must contain at least the following, case-insensitive, values to which the file's columns must adhere to: chromosome, splice&coordinate, strand, type, gene, location, location_type, heterozygosity, variant, input, key. The column headers need only contain the given text (i.e. a column labeled **`gene_name`** would be sufficient to satisfy the above requirement for a "gene" column). Column headers with ampersands (&) denote that all words joined by this symbol must be present for that column (i.e. **`Splice_site_coordinate`** satisfies the "splice&coordinate" requirement). The order of the columns is immaterial. The input column can contain any identifier for the variant and need not be unique. The **`location`** column specifies if the site is natural or cryptic. For Veridical, all that matters is that cryptic variants contain the word "cryptic" as part of their value in this column and that non-cryptic variants do not. The **`location_type`** column is only used for cryptic variants and specifies if the variant is intronic or exonic. It is not currently used by the program. This column must be present but can always be set to null (i.e.).

A few rows from a sample variant file is provided below (text wrapped for readability):

```
Chromosome Splice_site_coordinate Strand
Ri-initial Ri-final ΔRi Type Gene_Name
Location Location_Type Loc._Rel._to_exon
Dist._from_nearest_nat._site
Loc._of_nearest_nat._site
Ri_of_nearest_nat Cryptic_Ri_rel._nat.
rsID Average_heterozygosity
Variant_coordinate Input_variant Input_ID
RNASeqDirectory_ID RNA_Seq_BAM_ID_KEY
chr10 89711874 + 12.09 -2.62 -14.71
ACCEPTOR PTEN NATURALSITE . . . . . . .
. 89711873 A/G ID1 dir file
chr10 89712017 + 5.18 -1.85 -7.03 DONOR
PTEN NATURALSITE . . . . . . . . 89712018
T/C ID1 dir file
chrX 9621719 + -4.78 2.25 7.03 DONOR
TBL1X CRYPTICSITE EXONIC . 11 9621730 2.24
GREATER . . 9621720 C/T ID1 dir file
```

Veridical exome annotation input format

This input format can be generated via **`ConvertToExomeAnotation.pl.`** The file must be tab-delimited, excepting its header, which must be comma-delimited. It must have the following, case-insensitive, header columns, to which its data must adhere: transcript, chromosome, exon chr start, exon chr end, exon rank, gene. The column headers need only contain the given text (i.e. a column labeled **`gene_name`** would be sufficient to satisfy the above requirement for a "gene" column). The order of the columns is immaterial.

A few rows from a sample exome annotation file is provided below (text wrapped for readability):

```
Transcript ID,ID,ID,Chromosome Name,Strand,
Exon Chr Start,Exon Chr End,
Exon Rank in Transcript,Transcript Start,
Transcript End, Associated Gene Name
NM_213590 NM_213590 NM_213590 chr13 +
50571142 50571899 1 50571142 50592603
TRIM13
NM_213590 NM_213590 NM_213590 chr13 +
50586070 50592603 2 50571142 50592603
TRIM13
NM_198318 NM_198318 NM_198318 chr19 +
50180408 50180573 1 50180408 50191707
PRMT1
```

Veridical output

If a variant contains any validating reads, Veridical outputs the variant in question, along with some summary information and a table specifying the numbers of each validating read type detected for both the experimental and control samples. Within the output of Veridical, the phrase: "Validated (x) variant n times" means that the variant was validated mainly for splicing consequence x and has n validating reads. The variant will only appear within the **`*.filtered`** output file if the p-value for either junction-spanning or read-abundance-based reads for splicing consequence x was statistically significant (defined, by default, as: $p < 0.05$). After the variant being validated is provided, along with its primary predicted splicing consequence, the output is divided into two sections with identical contents: one for the experimental sample(s) and another for control samples. The summary enumerates the number of reads of each splicing consequence, partitioned by evidence type (junction-spanning or read-abundance-based), and by sample type (tumour or normal for control samples, and only tumour for experimental samples). A table describing the number of each read type for every file follows this summary. An example of this output, for the variant within RAD54L, as shown by Figure 7 and the last portion of Table 2, is provided. While Veridical outputs this as plain text, with the table in a tab-delimited format, we provide this output as an Excel document with descriptions of the meaning of each table heading, to clarify the presentation of the data. All input and output files for the five variants presented are provided. **`VeridicalOutExample.xls`** contains the output for the variant within *RAD54L*, along with descriptions of the terms used and the output format. all.vin contains the input variant file. **`allTumoursBAMFileList.txt`** and **`allNormalsBAMFileList.txt`** are the BAM file lists for tumour and normal samples, respectively. all.vout contains the Veridical output. The exome file can be retrieved using **`ConvertToExomeAnnotation.pl`**, available with the other programs at: www.veridical.org. The BAM file lists contain the TCGA file UUID, followed by a slash, followed by the file name. The RNA-Seq data itself can be downloaded from TCGA at: https://tcga-data.nci.nih.gov/tcga/.

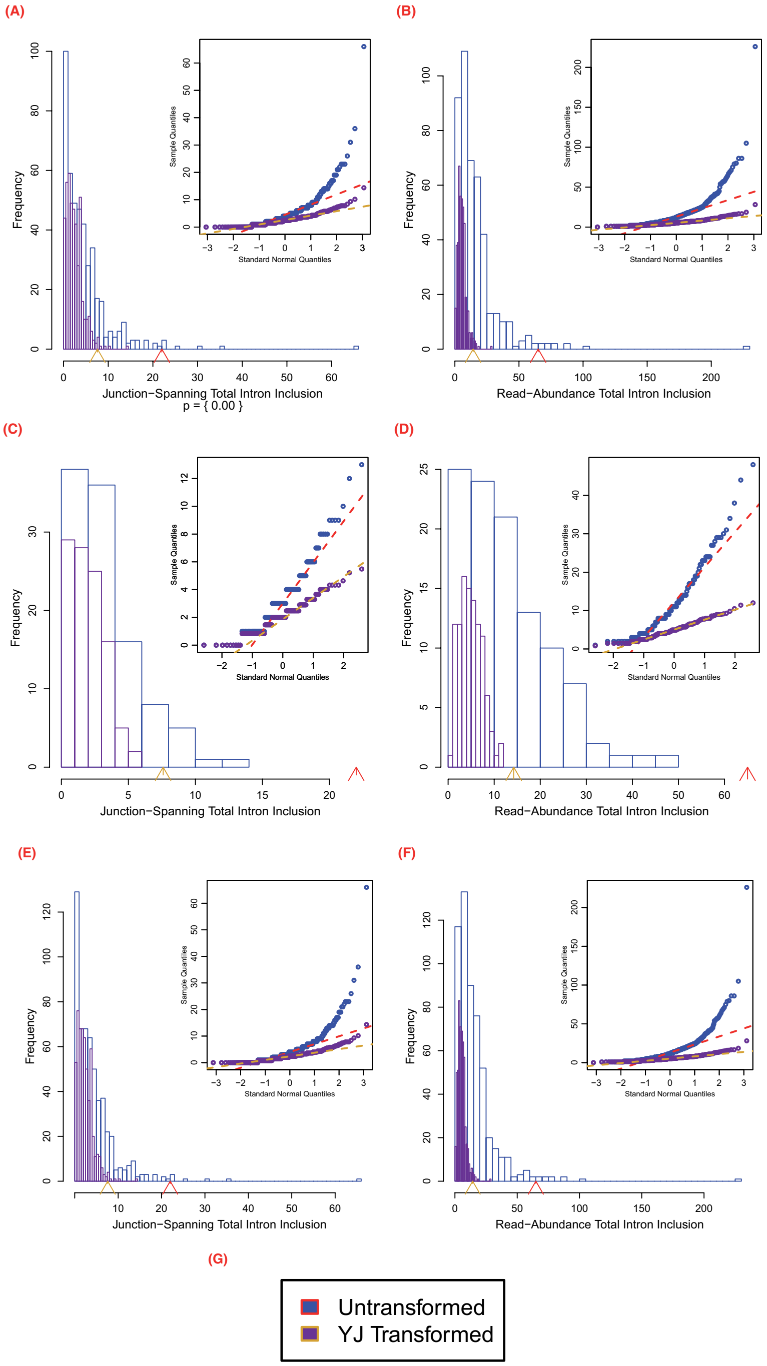


Figure S1. Histogram and embedded Q-Q plots portraying the difference between untransformed and Yeo-Johnson (YJ) transformed data. The plots depict intron inclusion for the inactivating mutation (chr12:83359523 G>A) within *TMTC2*, as shown in Figure 6(B) and 6(C). The arrowheads denote the number of reads in the variant-containing file, which is, in all cases, more than observed in the control samples ($p < 0.01$). The figure legend for all panels is provided in (**G**), which shows that blue and red plot elements correspond to untransformed data, while yellow and purple correspond to YJ transformed elements. Dotted lines in the Q-Q plots are lines passing through the first and third quantiles for a normal reference distribution. (**A**), (**C**), and (**E**) show junction-spanning based reads, while (**B**), (**D**), and (**F**) show read-abundance-based reads. (**A**)-(**B**) depict tumour sample distributions, (**B**)-(**C**) depict normal sample distributions, and (**E**)-(**F**) depict combined tumour and normal sample distributions. This figure is demonstrative of the general trend we have observed. Only data from normal samples resemble a Gaussian distribution and the YJ transformation greatly improves the Gaussian nature of all distributions.

References

1. Rogan PK, Zou GY: **Best practices for evaluating mutation prediction methods.** *Hum Mutat.* 2013; **34**(11): 1581–1582.
PubMed Abstract | Publisher Full Text

2. Krawczak M, Reiss J, Cooper DN: **The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: Causes and consequences.** *Hum Genet.* 1992; **90**(1–2): 41–54.
PubMed Abstract | Publisher Full Text

3. Rogan PK, Faux BM, Schneider TD: **Information analysis of human splice site mutations.** *Hum Mutat.* 1998; **12**(3): 153–171.
PubMed Abstract | Publisher Full Text

4. Rogan PK, Svojanovsky S, Leeder JS: **Information theory-based analysis of CYP2C19, CYP2D6 and CYP3A5 splicing mutations.** *Pharmacogenetics.* 2003; **13**(4): 207–218.
PubMed Abstract | Publisher Full Text

5. Mucaki EJ, Shirley BC, Rogan PK: **Prediction of mutant mRNA splice isoforms by information theory-based exon definition.** *Hum Mutat.* 2013; **34**(4): 557–565.
PubMed Abstract | Publisher Full Text

6. López-Bigas N, Audit B, Ouzounis C, *et al.*: **Are splicing mutations the most frequent cause of hereditary disease?** *FEBS Lett.* 2005; **579**(9): 1900–1903.
PubMed Abstract | Publisher Full Text

7. Wang K, Li M, Hakonarson H: **ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data.** *Nucleic Acids Res.* 2010; **38**(16): e164.
PubMed Abstract | Publisher Full Text | Free Full Text

8. Churbanov A, Vorechovský I, Hicks C: **A method of predicting changes in human gene splicing induced by genetic variants in context of cis-acting elements.** *BMC Bioinformatics.* 2010; **11**(1): 22.
PubMed Abstract | Publisher Full Text | Free Full Text

9. Pertea M, Lin X, Salzberg SL: **GeneSplicer: A new computational method for splice site prediction.** *Nucleic Acids Res.* 2001; **29**(5): 1185–1190.
PubMed Abstract | Publisher Full Text | Free Full Text

10. Shirley BC, Mucaki EJ, Whitehead T, *et al.*: **Interpretation, stratification and evidence for sequence variants affecting mRNA splicing in complete human genome sequences.** *Genomics Proteomics Bioinformatics.* 2013; **11**(2): 77–85.
PubMed Abstract | Publisher Full Text

11. Eswaran J, Cyanam D, Mudvari P, *et al.*: **Transcriptomic landscape of breast cancers through mRNA sequencing.** *Sci Rep.* 2012; **2**: 264.
PubMed Abstract | Publisher Full Text | Free Full Text

12. Eswaran J, Horvath A, Godbole S, *et al.*: **RNA sequencing of cancer reveals novel splicing alterations.** *Sci Rep.* 2013; **3**: 1689.
PubMed Abstract | Publisher Full Text | Free Full Text

13. Kwan T, Benovoy D, Dias C, *et al.*: **Genome-wide analysis of transcript isoform variation in humans.** *Nat Genet.* 2008; **40**(2): 225–231.
PubMed Abstract | Publisher Full Text

14. Thorvaldsdóttir H, Robinson JT, Mesirov JP: **Integrative genomics viewer (IGV): High-performance genomics data visualization and exploration.** *Brief Bioinform.* 2013; **14**(2): 178–192.
PubMed Abstract | Publisher Full Text | Free Full Text

15. Barnett DW, Garrison EK, Quinlan AR, *et al.*: **BamTools: A C++ API and toolkit for analyzing and managing BAM files.** *Bioinformatics.* 2011; **27**(12): 1691–1692.
PubMed Abstract | Publisher Full Text | Free Full Text

16. Yeo IK, Johnson RA: **A new family of power transformations to improve normality or symmetry.** *Biometrika.* 2000; **87**(4): 954–959.
Publisher Full Text

17. Johnson DH: **Statistical sirens: the allure of nonparametrics.** *Ecology.* 1995; **76**: 1998–2000.
Reference Source

18. Hubbard R: **The probable consequences of violating the normality assumption in parametric statistical analysis.** *Area.* 1978; **10**(5): 393–398.
Reference Source

19. Pruitt KD, Tatusova T, Maglott DR: **NCBI reference sequence (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins.** *Nucleic Acids Res.* 2005; **33**(Database Issue): D501–D504.
PubMed Abstract | Publisher Full Text | Free Full Text

20. Hsu F, Kent JW, Clawson H, *et al.*: **The UCSC known genes.** *Bioinformatics.* 2006; **22**(9): 1036–1046.
PubMed Abstract | Publisher Full Text

21. Hubbard T, Barker D, Birney E, *et al.*: **The Ensembl genome database project.** *Nucleic Acids Res.* 2002; **30**(1): 38–41.
PubMed Abstract | Publisher Full Text | Free Full Text

22. **RDC Team *R: A Language and Environment for Statistical Computing*.** Vienna, Austria: R Foundation for Statistical Computing, 2008.
Reference Source

23. Fox J, Weisberg S: ***An R Companion to Applied Regression*, 2nd ed**. Thousand Oaks CA: Sage, 2011.
Reference Source

24. Freedman D, Diaconis P: **On the histogram as a density estimator: L_2 theory.** *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete.* 1981; **57**(4): 453–476.
Publisher Full Text

25. Hyndman RJ, Fan Y: **Sample quantiles in statistical packages.** *American Statistician.* 1996; **50**(4): 361–365.
Publisher Full Text

26. Sarkar D: **Lattice: Multivariate Data Visualization with R**. New York: Springer, 2008.
Reference Source

27. Fox J: **Effect displays in R for generalised linear models.** *J Stat Softw.* 2003; **8**(15): 1–27.
Reference Source

28. Koboldt DC, Fulton RS, McLellan MD, *et al.*: **Comprehensive molecular portraits of human breast tumours.** *Nature.* 2012; **490**(7418): 61–70.
PubMed Abstract | Publisher Full Text | Free Full Text

29. Shen S, Park JW, Huang J, *et al.*: **MATS: A bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data.** *Nucleic Acids Res.* 2012; **40**(8): e61.
PubMed Abstract | Publisher Full Text | Free Full Text

30. Pan Q, Shai O, Lee LJ, *et al.*: **Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing.** *Nat Genet.* 2008; **40**(12): 1413–1415.
PubMed Abstract | Publisher Full Text

31. Griffith M, Griffith OL, Mwenifumbo J, *et al.*: **Alternative expression analysis by RNA sequencing.** *Nat Methods.* 2010; **7**(10): 843–847.
PubMed Abstract | Publisher Full Text

32. Katz Y, Wang ET, Airoldi EM, *et al.*: **Analysis and design of RNA sequencing experiments for identifying isoform regulation.** *Nat Methods.* 2010; **7**(12): 1009–1015.
PubMed Abstract | Publisher Full Text | Free Full Text

33. Shen S, Lin L, Cai JJ, *et al.*: **Widespread establishment and regulatory impact of Alu exons in human genes.** *Proc Natl Acad Sci U S A.* 2011; **108**(7): 2837–2842.
PubMed Abstract | Publisher Full Text | Free Full Text

34. Kapranov P, Willingham AT, Gingeras TR: **Genome-wide transcription and the implications for genomic organization.** *Nat Rev Genet.* 2007; **8**(6): 413–423.
PubMed Abstract | Publisher Full Text

35. Feng H, Qin Z, Zhang X: **Opportunities and methods for studying alternative splicing in cancer with RNA-Seq.** *Cancer Lett.* 2013; **340**(2): 179–191.
PubMed Abstract | Publisher Full Text

36. Eswaran J, Horvath A, Godbole S, *et al.*: **RNA sequencing of cancer reveals novel splicing alterations.** *Sci Rep.* 2013; **3**: 1689.
PubMed Abstract | Publisher Full Text | Free Full Text

37. Pal S, Gupta R, Davuluri RV: **Alternative transcription and alternative splicing in cancer.** *Pharmacol Ther.* 2012; **136**(3): 283–294.
PubMed Abstract | Publisher Full Text

38. Takahara K, Schwarze U, Imamura Y, *et al.*: **Order of intron removal influences multiple splice outcomes, including a two-exon skip, in a COL5A1 acceptor-site mutation that results in abnormal pro-a1 (V) N-propeptides and Ehlers-Danlos syndrome type I.** *Am J Hum Genet.* 2002; **71**(3): 451–465.
PubMed Abstract | Publisher Full Text | Free Full Text

39. Viner C, Dorman SN, Shirley BC, *et al.*: **Input, output, and explanatory files for Veridical**. figshare, 2013.
Data Source

Current Referee Status: ? ☑ ? ☑

Referee Responses for Version 1

☑ **Peter Robinson**
Institute for Medical Genetics, Universitätsklinikum Charité, Berlin, Germany

Approved: 18 March 2014

Referee Report: 18 March 2014
This paper from the Rogan group presents a methodology for validation of DNA sequencing variants that alter mRNA splicing. While variants of the most conserved splice site nucleotides at the intron-exon boundary can be predicted to cause splice defects with high reliability, it remains difficult to predict whether variants deeper in the intron or those that potentially affect exonic splicing enhancers actually cause splice defects. RNA-seq data, when coupled with variant data, potentially provide a means of correlating variation data with observations of (mis-)splicing patterns.

The program fulfils an important need in the community, the results appear promising and will be of special interest to groups performing RNA-seq analysis in medical settings. I have only some minor suggestions that the authors may like to consider.

Suggestions:

1. The explanation of the methodology is relatively difficult to follow, and I wonder if it might not be better to simplify Figures 1 and 2 for didactic sake. For instance, in Figure 1A, it is unclear where the location of variant C is. Does the curved line mean that it could be anywhere in the middle exon? Also, I assume that exons are being shown in blue and reads shown in gray?

   Also, the legend text is overly complicated: $D > E$ swap D and E. While aficionados of first order logic will follow without problems, I would suggest that it would be better for didactic purposes to delete this and to implicitly assume that $D<E$ for the sake of this figure. Figure 1B is confusing at this point in the manuscript because the motivation for switching the variable A,B, D, and E is not yet clear. On the other hand, panel C and panel D are trivial and do not add anything. I would suggest using Figure 1 to provide one concrete example one a simple level, and stating in the text that the variables are to be switched if the candidate mutation is located on the other side of the exon.

   Also, the explanations of the method that are couched in first order logic-like notation are difficult to follow, because it is not stated whether the variant C can precede the start of the read (in which case C-S would be negative). The subscripts for r in turn have the subscript s1 but the variable S in the formula does not.

   Although in the end, I think I follow the overall method, the reader is forced to make arbitrary assumptions in order to interpret the formulae being used to explain the method. A similar comment pertains to the flow chart in Figure 2.

Therefore, I would suggest the authors take some pains to improve the clarity of the explanation of the method. I would suggest that they show one of two concrete examples and provide English language specifications of the FOL-like formulae that describe the partitioning of reads.

2. I am a little unclear on the use of control samples vs experimental samples. Assuming the experimental samples come from different individuals, what is the reason to assume that they will have the same distribution of splice mutations? And given that one finds dozens of splice variants in normal individuals, what exactly is meant by a control sample? Will control samples not also have lots of splice mutations? How does the method deal with this? And if we are dealing with cancer samples, why not user a paired control to detect cancer-specific mutations? In light of this, the statement "*Maximizing the set of control samples, while computationally more expensive, increases the statistical robustness of the results obtained.",* does not appear to be supported by evidence presented in the manuscript.

3. It would be interesting to see a comparison of the distribution of *Ri* values and results of Veridical analysis?

4. How does Veridical decide which sequence variant is causative if there are multiple variations located in the vicinity of a given mis-spliced exon?

5. The mutation nomenclature chr1:985377 C>T should not have a space between the position and the nucleotides.

6. It is unclear to me why a linear regression model was used to show the performance of the method. The authors could provide timings from real runs.

7. It would be interesting to see a plot on the relationship of the p-values called by Veridical and the sequencing depth covered. The authors state "*In particular, a lack of sufficient coverage at a particular locus will cause Veridical to be unable to report any significant results. A coverage of at least 20 reads should be sufficient.*", but they do not provide evidence for this assertion. This is an important question given that low-expressed genes are thus likely to be systematically under-represented in the results of Veridcal, and this should be commented on somewhere in the paper.

8. It would be good if the authors provided Sanger validation of at least some of the mis-splicing events reported in the paper.

9. The input format for Veridical is described as *"This input format most easily accepts formatted output from the Shannon Pipeline."* Why not allow VCF files and filter them for potential splice variants informatically prior to Veridcal analysis? It was unclear to me how the variants are to be selected and whether Veridical can be easily used outside of the Shannon pipeline?

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Liliana Florea
McKusick-Nathans Institute of Human Genetics, Johns Hopkins University, Baltimore, MD, USA

Approved with reservations: 07 March 2014

Referee Report: 07 March 2014
The authors describe a method and the associated software, Veridical, for assessing the effects on pre-mRNA splicing of predicted splicing-affecting mutations. To do so the program compares splicing effects, measured by the supporting read counts, in variant-containing (disease) samples against a distribution derived from very large numbers of 'normals', either normal tissue from the same individual or samples from healthy individuals.

The idea is ingenious and novel as applied to mutations affecting splicing, although not in general (see VAAST [Yandell et al. 2011], which exploits the availability of large numbers of samples to identify likely deleterious variants; it is also the premise for the HapMap and 1000 Genomes projects). The software is fast and practical, being able to test thousands of variants in hundreds of samples within hours. This is the first software of its kind, and if accurate it will be a very valuable resource for clinical genomics.

That being said, while the article provides proof-of-concept and clearly demonstrates the potential of the tool with specific examples, there are several missing pieces that are needed to provide the readers with a view of its overall performance and limitations and to help them use it effectively.

Major comments:

1. The article shows numerous positive examples, however there is no indication of the tool's performance in general. The authors should include the results from running the tool on a full data set, to give potential users an idea of the expected outcome.

   Also, several other tools (e.g., MATS, Miso, SpliceTrap) have been developed for the related problem of discovering alternative splicing events and comparing them among samples. MATS in particular, allows differential splicing analyses with multiple replicates. Ideally the paper would include a comparison with MATS on the data set analyzed; this comparison is informative even if MATS is used with only a subset of the samples.

2. The method uses the YJ-transformed distribution of supporting read counts across the 'normals' to determine a p-value for the variant, and thus judge its significance and impact on splicing. This is an interesting concept that assumes that with large numbers of 'normals' sample and batch effects will even out; hence, large numbers of samples are required to ensure accuracy. Since these are absolute (non-normalized) counts, however, the method may not work if the variant sample is obtained with a different method, e.g. by rRNA depletion of total RNA whereas most normal samples would come from polyA+ libraries. The authors should clearly discuss this and other possible limitations of their approach.

3. Related to the above, the authors mention on several occasions the difficulty in identifying intron inclusion (II) events, in particular the large number of false positives. Indeed, IIs are generally difficult to predict due to the presence of intronic reads ('noise') from unspliced RNA. The levels can vary from sample to sample and across the genome, depending on the sample preparation, gene expression level, splicing efficiency, etc. By comparing read counts exclusively among samples and without taking into account the gene- or genome-level background, Veridical is likely to produce many false positives.

   In particular, the 14 supporting reads in the left intron on Figure 4 seem hardly sufficient to indicate an II event, all the more as there is a larger number of reads in the neighboring intron (not predicted to be II). The authors should provide other type of evidence for this event.

4. The mathematical formulas for the various classes of supporting reads and their locations (page 4, continued on page 6) are hard to understand. It would greatly help the readers to include a figure showing schematically the event and read location with respect to the introns and exons.

Minor comments:

1. As another reviewer pointed out, the software requires a registration to obtain a temporary license for 30 days, after which the availability and terms of use are unclear. This mode of distribution is not a problem, but the terms should be clearly stated in the manuscript. Also, this is a stand-alone software and not a web tool as implied by the article.

2. The authors use the term 'cryptic' splice sites throughout the manuscript (I assume meaning 'aberrantly activated'), but some of the events discovered could be alternative exon ends. It would be helpful to clarify in the context.

This is a potentially very powerful and useful tool. I gave the article an 'Approved with reservation' because it is critical to include results in the aggregate to complement the showcased examples, as well as to discuss its limitations. I will gladly change once these few issues are addressed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Competing Interests: No competing interests were disclosed.

Francesc Xavier Roca
Division of Molecular Genetics and Cell Biology, Nanyang Technological University, Singapore, Singapore

Approved: 28 January 2014

Referee Report: 28 January 2014
This manuscript describes a new computational tool named Veridical, which detects mutant-allele specific splicing changes from large RNAseq datasets. This outstanding tool appears very useful to screen the wealth of transcriptomic data for effects in splicing due to mutations in disease samples, and I think that it will potentially be of interest for many if not all such RNAseq-based studies. In addition, this could spur further efforts to derive similar tools with improved efficiencies. Use of this method should help establish the importance of aberrant splicing in disease as well as the effects of genomic mutations at the RNA level. I only have two comments, that do not diminish my overall rating of this work as of high value:

1. I personally disagree with the widespread use of the word “validation” in the title, abstract and text. Authors describe Veridical as a tool to “validate” DNA sequence variants that alter splicing. Indeed, I think that this tool provides an “association” between the variants and splicing, but not a formal proof of their connection. As the genomic and RNA samples usually come from different individuals with many confounding variables, the possibility that the splicing changes arise from factors other than the individual DNA mutations cannot be ruled out. In other words, changes in the levels of

trans-acting splicing factors could account in part or totally for the splicing changes across samples. The statistical tests properly conducted in Veridical are designed to minimize such possibility but do not rule it out. In addition, the inherent noisy nature of RNAseq datasets also prompts for caution in the conclusions. To me, the direct proof that a DNA mutation changes splicing of its pre-mRNA can only be provided using minigenes and cell transfection (or in vitro splicing), in which the substrate sequences and cellular context are under almost absolute control. Indeed, the Veridical method is reminiscent of GWAS (Genome-Wide Association Studies), in which the genotype in the DNA, wild-type or mutant, is associated to its phenotype, such as normal versus disease (or other traits) in GWAS, or normal versus aberrant splicing in this study. Thus, for me Veridical provides strong associations – but not validations – between DNA mutations and their effects on splicing.

2. As mentioned briefly at the beginning of Discussion, Veridical has built-in prediction tools to prioritize the mutations that are more likely to affect splicing, such as those mapping to splice sites. Even if other sources and tools are cited, a more extensive explanation of these components of Veridical would help the reader/user.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Stefania Bortoluzzi
Department of Biology, University of Padova, Padova, Italy

Approved with reservations: 27 January 2014

Referee Report: 27 January 2014
The paper "*Validation of predicted mRNA splicing mutations using high-throughput transcriptome data*" by Viner *et al*. presents Veridical, a new software for the interpretation and validation of genetic variants identified by DNA sequencing that alter mRNA splicing, leveraging RNA-seq data. The method is based on statistical comparisons of the normalized read counts of abnormally-spliced RNA species in mutant versus non-mutant tissues.

Actually, the interpretation of genetic variants is a difficult and key issue in current research.
The integration of genomic and transcriptomic data, namely the use of RNA-seq-based transcriptome characterization as a "molecular phenotype" of cells is useful and meaningful.

The software is standalone (not a web-tool) and it is completed by perl scripts, facilitating data management.
The manuscript declare that "Veridical and its associated software programs are available at: www.veridical.org".

Actually, Veridical is commercially available to the scientific community. A trial version lasting 30 days can be downloaded by the website, but in order to obtain binaries, the website requests a registration with an institutional email address - they reserve the right to deny access to users who register with third-party mail servers (Gmail, Yahoo, Hotmail, etc.).

No pricing information is included in the manuscript and, more importantly, in the webpages accessible to download the software, either before or after registration.

After downloading the software, I was not able to find R scripts that can be useful to generate some plots, as indicated in the manuscript.

Saying that, the paper is written in a clear language and it is quite complete.

I propose a few revisions that in my opinion can improve the manuscript readability and clarity.

Introduction

- Line 13 (minor): indicate which hereditary disease (colon cancer?).

Methods

- 2nd Par, Line 5 (minor): “Maximising” is used, but probably the meaning is “increasing” (the number of).

- Figure 1 (major): I feel that the info provided by points C and D is trivial, whilst point A's images, sentences and legends can be improved. Figure 1 C and D shows simply examples of reads that are mapped continuously and discontinuously to the reference genome. I think that every potential user of this type of software known well this concept. On the other hand, regarding A and B (upper part of the figure) there is not clear correspondence between the text in the legend and the image, and between the image and the text below (the arch overlap the point A in the figure B, whereas the text says "reads between A and B").

- In general, in many cases in the manuscript, the correspondence between legend and figure can be improved, by indicating more clearly the points specific sentences in the legend refer to. Regarding this issue, for instance in Figure 4 I can see indicated neither the “exon 12” nor the “14 reads” mentioned in the legend. Please indicate (using colors, boxes, arrows or overlapping text) key elements in the figure, and revise all figures using the same criterion.

- Page 5 (minor): Consider revising the sentence “Furthermore, this transformation allowed us to avoid the use of non-parametric testing, which has its own pitfalls regarding assumptions of the underlying data distribution”, since normally it is assumed that parametric tests ground on assumptions on data distributions, but non-parametric tests by definition can be used without information about data distribution.

- End of the next paragraph (major): “It is important to realize, therefore, that the p-values given by Veridical are much more robust when the program is provided with a large number of samples.” This is a pretty clear concept. Please, indicate a general rule to the user/reader: How many samples are required? Setting a reasonable minimum can be more useful for experimental design than saying the larger the sample size the most robust the result.

Results

I have two important criticisms about the Results section:

1. The section is not organized in paragraphs, and mixes performance info (run time using different number of samples and variants) with example results.

2. Not clearly saying how these results were obtained. This is important to guarantee repeatability.

- (Major) I propose to reorganize the results (considering skipping less important examples; retain surely Fig. 4 and 6) and insert a first paragraph providing information about the dataset used for variants validation (how many samples, how many controls) and about the variant calling (BAM files can be obtained with different settings and criteria and the same apply to calling and filtering of variants). Moreover, please explain how RNA-seq data are treated, and particularly how they are normalized to guarantee cross-samples comparability.

- (Major) Also, a brief discussion about the impact of disease samples not carrying the given mutation can be useful, as well as regarding the possibility that a tumour sample not carrying the considered variant can present altered transcriptome since other variants (or factors) impact on the "molecular phenotype".

- Figure 4 (minor): Please comment about the possible existence of intronic transcripts (totally unknown or also annotated in Ensemble, but not displayed in the more conservative RefSeq annotations).

- Figure 5 (minor): Please define better the measure "Read–Abundance Total Intron Inclusion".

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Competing Interests: No competing interests were disclosed.