ORIGINAL RESEARCH

Interpretation, Stratification and Evidence for Sequence Variants Affecting mRNA Splicing in Complete Human Genome Sequences

Ben C. Shirley 1, Eliseos J. Mucaki 2, Tyson Whitehead 3, Paul I. Costea 4, Pelin Akan 4, Peter K. Rogan 1,2,5,*

1 Department of Computer Science, Middlesex College, The University of Western Ontario, London, ON N6A 5B7, Canada
2 Department of Biochemistry, Schulich School of Medicine and Dentistry, The University of Western Ontario, London, ON N6A 5C1, Canada
3 SHARCNET, London, ON N6A 5B7, Canada
4 Royal Institute of Technology, Science for Life Laboratory, Solna 17165, Sweden
5 Cytoagnomix Inc., London, ON N6G 4X8, Canada

Received 6 December 2012; revised 16 January 2013; accepted 21 January 2013
Available online 14 March 2013

Abstract Information theory-based methods have been shown to be sensitive and specific for predicting and quantifying the effects of non-coding mutations in Mendelian diseases. We present the Shannon pipeline software for genome-scale mutation analysis and provide evidence that the software predicts variants affecting mRNA splicing. Individual information contents (in bits) of reference and variant splice sites are compared and significant differences are annotated and prioritized. The software has been implemented for CLC-Bio Genomics platform. Annotation indicates the context of novel mutations as well as common and rare SNPs with splicing effects. Potential natural and cryptic mRNA splicing variants are identified, and null mutations are distinguished from leaky mutations. Mutations and rare SNPs were predicted in genomes of three cancer cell lines (U2OS, U251 and A431), which were supported by expression analyses. After filtering, tractable numbers of potentially deleterious variants are predicted by the software, suitable for further laboratory investigation. In these cell lines, novel functional variants comprised 6–17 inactivating mutations, 1–5 leaky mutations and 6–13 cryptic splicing mutations. Predicted effects were validated by RNA-seq analysis of the three aforementioned cancer cell lines, and expression microarray analysis of SNPs in HapMap cell lines.

Introduction

The volume of human next-generation sequencing (NGS) data requiring bioinformatic analysis has necessitated development of high-performance software for genome scale assembly and analysis [1]. Genomic variations found in these analyses,
particularly single nucleotide polymorphisms (SNPs), have traditionally been interpreted in terms of amino acid modifications in coding regions. Clinically-significant non-coding variants are a relatively unexplored source of pathogenic mutations and lack a general, high-throughput method to interpret their effects. We present genome-scale software to quantify the effect of mutations in the common classes of splice donor (U1) or acceptor (U2)-type sites in a high-throughput manner. Mutations predicted with this method will be useful for pinpointing potentially deleterious variants suitable for further laboratory investigation.

Clinical studies have deemed the vast majority of known variants in patients with Mendelian disorders to be of uncertain pathogenic significance (VUS) [2,3]. cis mutations can affect protein translation, mRNA processing and initiation of transcription. In silico methods have been developed for the first two of these cases (e.g., [4,5]), but have only been routinely applied for protein coding changes in genome-scale applications (e.g., [6]). Many NGS studies classify mutations at only the highly conserved dinucleotides within each splice junction (e.g., [7]). Although more sensitive methods have been developed which assess other conserved sequence elements [8–12], none have been scaled for the large numbers of variants generated by NGS and nor have they been validated for these data. Exonic variants in close proximity to splice junctions but outside of this window may be classified as synonymous, missense or nonsense substitutions, yet still have profound effects on splicing, which may be the predominant contributor to the phenotype. Unless multiple affected patients are reported with the same mutation, the mutations are transmitted through pedigrees, and functional assays verify their effects, these variants in patients are generally be classified as VUS. mRNA splicing mutations are common in Mendelian diseases [13,14], and it is likely that they contribute to many complex disorders. Clearly, genome-scale predictive methods that filter alteration mutations not found in dbSNP135 (a list of ~54 million known nucleotide polymorphisms) and those with less than 1% average heterozygosity are reported (Tables S1–3).

The Automated Splice Site Analysis (ASSA) [5] server evaluates single mutations that change splice site strength with information theory-based models [15]. The average information, $R_{\text{sequence}}$, of a set of binding sites recognized by the same protein (such as U1 or U2) describes the conservation of these sequences. Sequences are ranked according to their individual information content ($R_i$ in bits) [15–17]. Individual information content is a portable, universal measure which allows direct comparison of binding sites across the genome or transcriptome, regardless of the sequence or protein recognized. Functional binding sites have $R_i > 0$, corresponding to $\Delta G < 0$ kcal/mol [18]. Strong binding sites have $R_i \gg R_{\text{sequence}}$ while weak sites have $R_i \ll R_{\text{sequence}}$. Any sequence variation may change its protein binding affinity, which is reflected by a change in the computed $R_i$ of that binding site. A 1-bit change in information content ($\Delta R_i$) corresponds to a $\geq 2$-fold change in binding affinity (100$^{\Delta R_i}$). The ASSA server has been widely used and its sensitivity and specificity have previously been extensively validated in hundreds of studies of individual mutations (http://tinyurl.com/splice-server-citations). However, it requires approximately 30 s to examine a single variant and is therefore not suitable for comprehensive analysis of whole-genome sequencing data. The Shannon pipeline was developed using the same mathematical approach and information weight matrices as ASSA to carry out batch information theory-based analysis of thousands of mutations from the BRCA1 and BRCA2 genes in Breast Cancer Information Core Database [19]. In the present study, the software has been adapted to perform a single matrix algebraic calculation across a genome with an efficient state machine that significantly increased computational speed over ASSA. Here we describe this software tool and analyze predicted mutations with RNA-seq data from genomes of 3 cancer cell lines.

**Results**

**Performance of the Shannon pipeline software**

We implemented an efficient algorithm for high-throughput detection and interpretation of mRNA splicing mutations based on information theory-based position weight matrices of a genome-wide set of curated splice donor and acceptor sites [20]. The present study focuses on software performance, interpretation of contextual changes identified from genomic annotations and supported by genome-scale RNA-seq data. The strategy underlying the Shannon splicing pipeline is to evaluate many sequence changes by information analysis quickly; this is followed by implementation of a set of heuristics based on these results combined with genome annotations to distinguish normal splice sites from those with diminished binding and cryptic sites with competitive binding affinities.

To assess performance, all point mutations detected in the complete genomes of the three cancer cell lines were analyzed using the pipeline. Variants in the cell lines U2OS (osteosarcoma-derived), A431 (epidermoid squamous carcinoma-derived) and U251 (glioblastoma-derived) were examined and filtered to create tractable sets of variants. Predicted splice-altering mutations not found in dbSNP135 (a list of ~54 million known nucleotide polymorphisms) and those with less than 1% average heterozygosity are reported (Tables S1–3).

The software processes single nucleotide variants (SNVs) to identify and annotate putative splicing mutations with sufficient speed to analyze single or multiple genomes within a few hours. Analysis of all single nucleotide substitutions detected in the genome of the U2OS cell line – 211,049 variants – is completed in 1 h 12 min on an IBM-based CPU in either Linux or Mac OSX. The speed of a genome analysis is dependent on the number of chromosomes represented in the input data. The state machine facilitates the analysis of all variants on a single chromosome with the highest efficiency because genomic data for each chromosome must be read and parsed. A complete analysis of 300 variants on a single small chromosome (e.g., chromosome 22) can be completed in 5 min. Variants distributed throughout all chromosomes require at least 1 h to process. The Shannon pipeline should be executed on a machine with sufficient RAM to store the entire human genome ($\geq 4$ Gb). When all chromosomes are represented, increasing the number of mutations results in an approximately linear increase in actual computation time, after accounting for the overhead required for memory management of genome sequences and annotations. For example, 2 h 35 min is required...
The diagram illustrates potential outcomes of mRNA splicing mutations predicted by the Shannon pipeline. Variation within splice donor and/or acceptor sites may lead to altered splicing events such as exon skipping (○), exonic (□) or intronic (△) cryptic site use, and/or reduction in the abundance of normally spliced mRNA forms, termed leaky mutations (○).

The software sorts the results into categories of inactivated, leaky and cryptic splicing mutations (Figures 1 and 2). However, further in silico filtering by the user is recommended to select the variants most likely to be deleterious to normal mRNA splicing. Results are sorted according to the degree of change information (i.e., mutation severity) or by heterozygosity, and to de-emphasize or eliminate common SNPs that affect mRNA splicing in cases where the phenotype is known to be rare. Filtering of cryptic splice sites exceeding the strength of and close to adjacent natural sites of the same phase eliminates many predicted unused cryptic sites with changes in R
values. Predicted mutations in known genes with open reading frames exclude poorly characterized genes lacking HUGO-approved names or encoding non-coding RNAs, and pseudogenes. Post-hoc filtering (especially of cryptic splicing mutations) significantly enriches for likely mutations in the genomes of these cancer cell lines by the order of 10,000 fold (Table 1).

**Comparison of predicted splicing mutations with expression data**

Several variants detected in genomes of U2OS, U251 and A431, which were predicted to affect splicing, were compared to the distribution of RNA-seq reads in their respective regions of the transcriptome. When interpreting these data, we assume that predicted mutations are present in a genetic background, in which the other parentally derived allele lacks the same variant (i.e., heterozygous). Abnormal reads or exon skipping of the mutant allele is viewed in the context of a single allele and expected normal splicing of the corresponding exon. For mutations that are predicted to inactivate a splice site, we assume a binomial distribution in the number of expected reads, based on the wild type allele. Natural splice site mutations are expected to significantly reduce the number of splice junction-spanning reads in relation to those in the adjacent exons, consistent with exon skipping. In some cases, intron inclusion adjacent to a splice site variant with lower R
value may also be evidence of a splicing mutation. In U2OS, 10 of 13 novel inactivating variants found in mutated natural splice sites met these criteria, along with 2 additional probable mutations (Table S1). The same criteria were met by 2 of 4 (with 1 additional probable) novel inactivating variants in U251 (Table S2), and 4 of 7 (with 1 additional probable) variants in A431 (Table S3).

Expression data support our predictions for 1 of 7 activated cryptic site variants in U2OS, 1 of 14 variants in A431 and 0 of 10 in U251. Many of the predicted splice sites reside in intronic regions or alternative exons that map far upstream or downstream of constitutively expressed exons. They are unlikely to displace constitutive isoforms, since donor site recognition is processive [21] and the increased lengths of such cryptic exons would probably be suboptimal [22]. Often, these sites are associated with rare, alternatively spliced ESTs expressed in other tissues than these cell lines. Because these variants are often extra-exonic, changes in expression must be inferred indirectly from decreased read count, intron inclusion or increased exon skipping. Changes in reading frame from inclusion of out-of-phase intronic sequences may induce nonsense-mediated decay (NMD). Reads mapping to adjacent introns are expected to be reduced in number as a result of NMD. Sequencing reads that are concentrated in the intronic region adjacent to exon of interest are considered support for predicted mutations. NMD may also affect transcript read counts associated with severe leaky or inactivated natural donor sites, which produce exon skipping with frame-shifting. Several predicted splicing mutations confirmed by RNA-seq are well-known driver mutations that contribute to tumor phenotypes.

We highlight a unique natural donor site mutation within **RBBP8** (NM_203291.1:c.248G > A or chr18:20529676G > A; 6.2 →3.2 bits, indicating the change in the R
value of the donor site, before and after it is mutated) in A431, a tumor suppressor gene mutated in numerous neoplasias with a role in endonucleolytic processing of a covalent topoisomerase-DNA complexes. The mutation weakens but does not abolish the natural donor site from 6.2 to 3.2 bits. A cryptic mRNA splice form using a pre-existing donor site 24 bp downstream to the weakened natural site is confirmed by RNA-seq (Figure 3A). The ASSA server predicts the activation of this intronic cryptic donor site, as well as a second site of equal strength further downstream to the mutated donor site (Figure 3B). There are a total of 56 reads that both encroach into the intron and overlap this variant. Forty-one of these cover the cryptic exon splice junction of interest (the aligned reads stop at the 3.2 bit cryptic site, which is 24 nt downstream of the natural site, and continue into the next natural exon). Thirty-one junction spanning reads also contain the A-allele. There are an additional 23 reads that cross into the intron, but do not extend as far as the cryptic site of interest. In 19 cases, these reads contain the A-allele. The remaining 4 intron-crossing reads which contain the G-allele appear to be misaligned, as they contain short matches (of ≤3 nt) to the downstream exon. There are an additional 2 reads that span the junction between the downstream cryptic exon junction and the adjacent exon (31 nt downstream; also 3.2 bits). Finally, 12 reads are correctly spliced and contain the mutant A-allele, suggesting that the natural site is not completely inactivated by this nucleotide substitution, which is consistent with leaky splicing.

Changes in expression are also noted in other genes. **DDX11** is inactivated in U2OS (chr12:31242087T > G;
6.89 \text{ bits}). DDX11 is a component of the cohesin complex which has a crucial role in chromosome segregation, and is essential for survival of advanced melanoma [23]. In U2OS, \textit{WWOX}, a tumor suppressor gene in osteosarcoma [24], contains a leaky mutation (chr16:78312497C > A; 10.24 \text{ bits}). Both alleles of \textit{APIP}, an apoptosis associated gene, are inactivated in U251 (chr11:34905054G > C; 9.32 \text{ bits}). Gene expression of \textit{APIP} is down regulated in non-small cell lung carcinoma [25]. Amplification of \textit{METTL2B}, which harbors a leaky mutation in U251 (chr7:128117227G > A; 5.48 \text{ bits}), has been demonstrated in several cancers, including glioblastoma [26]. In A431, leaky mutations are also confirmed in the glioblastoma-initiating gene \textit{TRRAP} (chr7:128117227G > A; 5.48 \text{ bits}) and \textit{USF1} (chr1:161013165G > T; 4.89 \text{ bits}), which encodes a transcription regulator.

### Table 1: Performance of Shannon pipeline for human mRNA splicing mutation prediction

| Source of variants | Number of variants analyzed | Running time a |
|--------------------|-----------------------------|----------------|
| U2OS cell line     | 211,049                     | 1 h 12 min     |
| A431 cell line     | 290,589                     | 1 h 17 min     |
| U251 cell line     | 314,637                     | 1 h 20 min     |
| ESP 6500 Exomes    | 1,872,893                   | 2 h 35 min     |

Note: aIntel I7 CPU with 16 Gb RAM.

Figure 2 Sample output of the Shannon pipeline software

The Shannon pipeline software generates the following types of output. A. Tabular results showing the first 12 of 134 changes in \( R_i \) values at different genomic coordinates predicted to be significant, after filtering for cryptic splicing mutations from all variants \((n = 22,197)\) in a complete genome sequence. The first filter eliminates exonic cryptic sites, the second selects cryptic sites with increased \( R_i \) values, the third ensures that the cryptic site is stronger than the corresponding natural site of the same phase and the final filter ensures that all remaining sites exceed the minimum \( R_i \) value of a functional splice site. B. Manhattan-like plot indicating the locations and changes in \( R_i \) of all variants which alter splice site information in a region within intron 1 of \textit{BRCA1} (chr17:41277500-41288500) from different individuals with increased breast cancer risk. C. Custom track illustrating a cryptic splicing mutation detected in an ovarian serous carcinoma that inactivates the acceptor site of exon 4 in \textit{STXBP4}, resulting in the activation of a pre-existing, in frame, alternative splice site 6 nucleotides downstream.
important for TGFβ2 expression in glioblastoma [28]. SYNE2, which is mutated in a significant percentage of head and neck squamous cell carcinomas [29], contains an inactivating splice site variant in A431 (chr1:64669514T > A; 1.89 → 0.83 bits). RRM2B, an inducible DNA repair gene that has been implicated in squamous cell carcinoma [30], contains an inactivating mutation in A431 (chr8:103250667A > C; 3.6 → 15.02 bits). SMARCD1, encoding a chromatin modulator that interacts with nuclear receptor transcription factors, is also inactivated in A431 (chr12:50480538G > C; 8.46 → 3.21 bits), and has been shown to be mutated in hepatocellular and other carcinomas [31].

Several mutations were found in potential tumor-associated genes, with either suggestive or little supporting expression data. However, defects in many of these genes have been implicated in various neoplasias including glioblastoma, osteosarcoma, and epidermoid squamous carcinoma. In general, these were predicted leaky mutations, where effects (diminished read counts and exon skipping) were inferred against the confounding background of a presumably intact allele. Natural site mutations in FANCD2 (NM_033084.3:c.3106-9T > A; 6.0 → 3.5 bits; delayed activation of the DNA damage response in gliomas [32]) and MDC1 (NM_014641.2:c.2129-8G > C; 6.4 → 4.7 bits; mediator of the DNA damage checkpoint and underexpressed in many cancers [33]) were found in the U251 cells.

Characterization of defective biochemical and functional pathways

Potential driver mutations affecting protein coding of genes from the A431, U2OS, and U251 cell lines have recently been
reported [34]. Functionally significant driver mutations affecting splicing are expected to comprise many of the same pathways implicated by protein coding mutations that are predicted to be damaging. The gene set with combined driver point and copy number alteration was examined using Reactome [35]. Shannon pipeline results, supported by RNA-seq data, were added to gene sets proposed by [34] and the expanded gene set was examined with the overrepresentation analysis tool in Reactome. Of the genes containing transcript-validated splicing mutations, both datasets were consistent in 2 of 5 pathways in A431 (interferon signaling and cytokine signaling in immune system), 8 of 8 pathways in U2OS (cell cycle mitotic, cell cycle, DNA replication, mitotic M–M/G1 phases, M phase, kinetochore capture of astral microtubules, mitotic prometaphase and apoptosis) and 0 of 2 pathways in U251. The gene set including all inactivating and leaky variants (regardless of verification status) was found in 5 of 7 of the same pathways in A431 (additionally, a variant was found in the semaphorin interaction pathway), 8 of 12 of the same pathways in U2OS and 0 of 11 pathways in U251. In A431 and U2OS, these splicing mutation predictions enhance and strengthen the pathway analysis based on protein coding mutations alone.

### Discussion

Complete genome and exome sequencing detects numerous rare, non-recurrent mutations in different individuals with the same disease diagnosis. Making sense of genetically heterogeneous results requires detection and interpretation of mutations in many genomes. The identification of significant mutations in different driver genes, followed by a gene set or pathway analysis can reveal common, essential pathways in otherwise genetically heterogeneous diseases, such as cancer. Incomplete detection or reclassification of coding mutations will most likely impact the sensitivity of these analyses. Most existing methods to predict the effects of splice site variation lack scalability, transparency or portability, with respect to their scoring systems. Information content can be applied to any region of any adequately annotated genome. Change in information ($\Delta R$) is a portable measure and its thermodynamic basis meaningfully estimates the effects of splicing variation. By contrast, other systems (e.g., [36]) are not suited for genome scale analysis and produce results that are not directly related to splice site strength.

A recent study reported the genomic, transcriptomic and protein sequences in the cell lines that were the source of the data that we analyzed [34]. It described the same single splicing mutation in the APIP gene identified in the present study, but none of the others that we predicted. Further, there was no overlap between the genes containing predicted protein coding mutations in [34] and those indicated from the current study. This was somewhat surprising, as we anticipated that some loss of function mutations in tumor suppressor genes would arise from compound heterozygosity. Instead, mutant genes from both studies tended to occur in the same pathways (for U2OS and A431).

Many of our predictions were supported by the same RNA-seq data that identified only APIP [34]. Conventional splice junction mutation analysis of NGS data, which tends to emphasize only the significance of changes in conserved splice junction, intronic dinucleotides does not appear to be as sensitive or comprehensive as the information theory-based approach we describe [7]. Assuming the cell line genotypes faithfully reflect the tumor genetics, likely driver mutations in the tumors were missed. These genes contribute to the tumor signatures and in most instances, belong to major pathways that are dysfunctional in the tumor. The caveat is that many of these cancer-associated genes have been uncovered in other tumor types, rather than the tumors that gave rise to the cell lines studied here.

Many of the predicted mutations that are supported by expression data make sense in light of independent studies, which have suggested the same driver genes and pathways that are defective in these tumor types [29,37–39]. We note that the recommended filtering procedures eliminate and/or minimize inclusion of mutations in gene classes with no known connection to neoplastic disease etiology. The sensitivity and specificity of these predictions support use of the Shannon pipeline in other somatic genomic analyses, and possibly for a wider spectrum of heritable genetic disorders.

The interpretation of potential splicing mutations in complete genome data is also challenging because the source of annotations, Ensembl, contains many accurate but apparently irrelevant genomic features. These comprise of exons called on the basis of a single or a few ESTs with deep intronic locations (relative to constitutive exons) [40,41], and predicted mutant ESTs that are in fact present in non- or low expression genes (due to tissue specificity of the gene). Where the RNA-seq data are either insufficient or irrelevant, pseudogenes (or genes which are members of families containing pseudogenes) may contain mismapped reads for the non-functional copies that can produce false positive mutation calls. Automatic filtering of genes from the RNA-seq data prior to validating information theory-based predictions would significantly simplify post hoc processing of the Shannon pipeline. Until such a workflow is available, individual predicted mutations have to be assessed manually, because cryptic sites that alter the strength of a “decoy” exon, while a technically legitimate result, is probably irrelevant as a potential disease-causing mutation.

Accurate genome-scale mutation analysis of bulk sequencing data in a timeframe suitable for integration with prediction tools for other types of mutations will be needed to discover disease-related genes and pathways in large-scale genomic studies of many patients. The need to distinguish the probable pathogenic from benign sequence changes has become acute [42]. Computing efficiency is essential for concurrent analysis of large sets of genome sequences [43]. The processing speed we have attained has distinct advantages for identifying functional non-coding variants detected in large multi-genomic analyses. However, increased speed comes at the expense of diminished ability to analyze complex mutations on the fly, such as insertions and deletions or multinucleotide substitutions. Such variation is significantly less common than SNPs in wild type genome and exome sequences [44], but nevertheless can have consequences on gene function and phenotype. The ASSA server is capable of analyzing these categories of mutations; however it is considerably slower than the Shannon pipeline (30 s per variant). In the future, the Shannon pipeline will be integrated with the ASSA server to examine complex variants seamlessly.
Materials and methods

Shannon pipeline software architecture

We have implemented the Shannon pipeline software plugin using the CLC-Bio genomics developer toolkit to simplify access to this technology and interpretation by novice users. The same plugin can be executed on a single client computer, a remote server or a grid system, and benefits from automated software updates. The server version uses an architecture in which a Workbench client transmits variant data to the server, which performs the computations, and returns results that can be filtered and formatted on the client. A standalone version of the fully functional Genome Workbench plugin is also available. By contrast, the splicing mutation feature that is native in CLC-Bio Genomics products is limited to detecting changes in dinucleotides at the exon boundaries, which represent fewer than 5% of all splicing mutations detected by the Shannon pipeline.

The Shannon Human Splicing Pipeline uses an efficient algorithm coded in C++ to quickly analyze genome-scale data sources for information changes (Figure S1). Methods for computing \( R_i \) and \( \Delta R_i \) values determine the dot product of an information weight matrix and the unitary sequence vector for each genomic window and compare the resultant scalar of an information weight matrix and the unitary sequence vectors. Changes in \( R_i \) for each position (\(|R_i| > \delta \)) values determine the dot product of a cryptic intron within these exons if a second pre-existing acceptor (or donor) could potentially activate could potentially (but rarely) lead to the formation of a cryptic donor in the first exon or acceptor in the last exon could lead to a truncated exon. The Shannon pipeline considers the exonic cryptic sites of the opposite polarity (acceptors in first exons and donors for last exons), as their activation could potentially (but rarely) lead to the formation of a cryptic intron within these exons if a second pre-existing cryptic site of opposite polarity is present in the proper orientation.

Although the Shannon pipeline output contains a vastly reduced number of potentially significant variants, further filtering is usually necessary to obtain the final set of functionally relevant variants. Pipeline output is generated for variants that result in \( |\Delta R_i| > \pm 1 \) bit. Each variant is annotated with the distance from nearest natural site, \( R_i \) between the natural and nearby cryptic sites. Thus, we remove those natural sites with positive \( |\Delta R_i| \) values as well as cryptic sites with reductions in \( R_i \) value. Pipeline generated annotations that were found in the tabular output help simplify the data filtering process. To predict those variants with significant changes, tables were filtered to distinguish natural and cryptic splicing mutations (Tables S1–3). The filters used for cryptic splicing mutations were based on criteria given in [47] (a) \( |\Delta R_i| > 0 \), (b) either occurring within an exon or within an intron less than 300 bp from nearest natural site, (c) cryptic

Stratification of variants

Similar to ASSA, the pipeline analysis produces summary tables (e.g., Figure 2A) for different types of mutations (assuming each type is represented): (1) complete sets of all splicing variants, (2) mutations predicted to inactivate splice sites, (3) leaky splicing mutations that reduce but do not abolish splicing and (4) cryptic splice sites that are either activated, inactivated or reduced in strength. We define inactivating variants as those that reduce the \( R_i \) of the affected binding site below 1.6 bits [20]. Binding sites containing a leaky variant are defined as those, in which initial \( R_i \) is decreased upon mutation to \( R_i > 1.6 \). Finally, candidate cryptic sites encompass all sites with higher affinity for binding than a corresponding natural site based on comparison of their respective \( R_i \) values (see below). Tabular data can be sorted by clicking the column header of each column. Data can be exported and viewed without modification in a spreadsheet program.

The \( 5' \) end of the first exon and the \( 3' \) end of the last exon of a gene are not splice sites. Therefore, the Shannon pipeline does not report mutations that affect their \( R_i \) at these positions; the exception is genes that encode alternate splice forms using further upstream/downstream exons present in Ensembl 66. Variants which alter the strength of cryptic splice sites within the first and last exons are also considered. Use of a strengthened cryptic donor in the first exon or acceptor in the last exon could lead to a truncated exon. The Shannon pipeline considers the exonic cryptic sites of the opposite polarity (acceptors in first exons and donors for last exons), as their activation could potentially (but rarely) lead to the formation of a cryptic intron within these exons if a second pre-existing cryptic site of opposite polarity is present in the proper orientation.
splice site $R_i$ value exceeding the strength of the nearest natural site $R_i$ of the same type and (d) intronic cryptic splice sites are selected 5' to the exon if acceptors and 3' to the exon, if donors. All reported variants were further categorized according to whether they had been previously reported or were novel. In Table 2, only novel and known variants of <1% average heterozygosity in dbSNP are reported, which are more likely to be functionally significant due to selection. Nevertheless, the threshold for filtering based on heterozygosity can be specified by the user.

Pipeline results are also displayed visually. Δ$R_i$ and final $R_i$ values are plotted by chromosome location, similar to Manhattan-style representations, for either individual chromosomes or entire genomes (Figure 2B). Hovering the cursor over data points generates tooltips containing information needed to find the complete entry within the corresponding tabular data. To locate interesting data points, a zoom function allows closer inspection of the plot. This visualization allows patterns to be observed and data points which stand out to be easily located and inspected more closely in a tabular format or on the ASSA server.

Chromosome-specific, custom browser tracks indicating Δ$R_i$ values in the BED format are created during analysis. This enables visualization of predicted mutations in the context of other genome annotations, for example, mapped reads from RNA-seq, spliced ESTs and known mRNAs. Figure 2C displays a result on the CLC Genomics Workbench browser along with the Ensembl genes v.66 track.

Experimental validation

RNA-seq analysis using published data from these cell lines [34] was used to compare our methods with expression data. TopHat [48] was executed with the following command-line options: -g 5 –solexa1.3-quals -p 8, and examined with the Integrative Genomics viewer (IGV) [49] to interrogate predictions made with the Shannon pipeline.

Authors’ contributions

PKR conceived of and directed the project. BCS implemented these methods, which involved modifying previous software developed by TW and EJM, creation of new modules, and integration with the CLC-Bio Genomics Workbench. EJM created and updated genome annotation databases required by the software. BCS and PKR compared variants predicted to be potentially deleterious with RNA-seq data provided by PIC and PA. BCS, EJM, 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 founder of Cytognomix, which is developing the software presented here.

Acknowledgements

The authors would like to acknowledge support from Natural Sciences and Engineering Research Council (Discovery Grant 371758-2009) (PKR), Canadian Breast Cancer Foundation (PKR), Compute Canada (PKR), Canadian Foundation for Innovation, Canada Research Chairs, MITACs Accelerate (BCS), the Ontario Graduate Scholarship Programs (BCS) and Cytognomix Inc. We thank Mikael Flensborg, Paul Fleischer and Alex Andersen for assistance with CLC-Bio products integration.

Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.gpb.2013.01.008.

References

[1] Gullapalli RR, Desai KV, Santana-Santos L, Kant JA, Becich MJ. Next generation sequencing in clinical medicine: challenges and lessons for pathology and biomedical informatics. J Pathol Inform 2012;3:40.
[2] Kavanagh D, Anderson HE. Interpretation of genetic variants of uncertain significance in atypical hemolytic uremic syndrome. Kidney Int 2012;81:11–3.
[3] Spurdle AB, Healey S, Devereau A, Hogervorst FB, Monteiro AN, Nathanson KL, et al. ENIGMA—evidence-based network for the interpretation of germline mutant alleles: an international initiative to evaluate risk and clinical significance associated with sequence variation in BRCA1 and BRCA2 genes. Hum Mutat 2012;33:2–7.
[4] Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods 2010;7:248–9.
[5] Nalla VK, Rogan PK. Automated splicing mutation analysis by information theory. Hum Mutat 2005;25:334–42.
[6] Kumar A, White TA, MacKenzie AP, Clegg N, Lee C, Dumpit RF, et al. Exome sequencing identifies a spectrum of mutation
### Supplemental Table 1.

#### A. U2OS Splicing Mutations

| Leaky | \( R_{i,\text{initial}} \) | \( R_{i,\text{final}} \) | \( \Delta R_i \) | Site | Gene       | rsID     | Av. Het. | Validated* |
|-------|-----------------|-----------------|----------------|-------|------------|----------|----------|------------|
| **Novel Variants** |                  |                 |                |       |            |          |          |            |
| chr5:148630982G>A | 10.04           | 7.03            | -3.01          | DONOR | ABLIM3     |          |          | Y          |
| chr9:140507906C>G | 9.45            | 6.83            | -2.62          | ACCEPTOR | ARRDC1    |          |          | LF         |
| chr12:46355654A>C | 7.66            | 5.61            | -2.06          | ACCEPTOR | SCAF11    |          |          | P          |
| chr10:115470783A>C | 8.72            | 6.80            | -1.93          | ACCEPTOR | RP11-211N11.5.1 |          |          | NE         |
| chr16:81060105T>G | 3.99            | 2.87            | -1.12          | ACCEPTOR | CENPN     |          |          | Y          |
| **Known Variants < 1% Av. Het.** |                  |                 |                |       |            |          |          |            |
| chr16:78312497C>A | 10.24           | 6.67            | -3.57          | ACCEPTOR | WWOX      | rs8050128 | 0        | Y          |
| chr2:98177124T>G | 7.96            | 4.70            | -3.26          | DONOR | ANKRD36B  | rs11681640 | 0        | PS         |
| chr5:78076488A>T | 9.60            | 7.14            | -2.45          | ACCEPTOR | AR5B      | rs183651028 | 0        |            |
| chr22:36657628T>G | 8.94            | 6.88            | -2.06          | ACCEPTOR | APOL1     | rs41368549 | 0        | N          |
| chr2:32961745T>A | 8.99            | 7.14            | -1.85          | ACCEPTOR | TTC27     | rs10200333 | 0        | P          |
| chr15:68445912T>A | 16.57           | 14.72           | -1.85          | ACCEPTOR | PIAS1     | rs11633620 | 0.005    | Y\(^{\text{H}}\) |
| chr1:33318561T>C | 5.86            | 4.55            | -1.32          | DONOR | S100PBP   | rs702836  | 0        | N          |
| chr17:73698557C>T | 9.81            | 8.70            | -1.12          | ACCEPTOR | SAP30BP   | rs820232  | 0.009    | Y          |

| Inactivating | \( R_{i,\text{initial}} \) | \( R_{i,\text{final}} \) | \( \Delta R_i \) | Site | Gene       | rsID     | Av. Het. | Validated |
|--------------|-----------------|-----------------|----------------|-------|------------|----------|----------|-----------|
| **Novel Variants** |                  |                 |                |       |            |          |          |           |
| chrX:91518144T>G | 9.64            | -8.99           | -18.62         | DONOR | PCDH11X    |          |          | NE        |
| chr17:47286205A>C | 9.57            | -9.06           | -18.62         | DONOR | GNGT2      |          |          | Y         |
| chr3:41277214G>A | 8.87            | -2.02           | -10.88         | ACCEPTOR | CTNNB1    |          |          | Y         |
| chr12:120613944A>C | 8.78            | -9.85           | -18.62         | DONOR | GCN1L1     |          |          | Y         |
| chr3:52181002A>C | 7.88            | -10.75          | -18.62         | DONOR | POC1A      |          |          | Y         |
| chr9:1311133696T>G | 7.88            | -10.75          | -18.62         | DONOR | URM1       |          |          | P         |
| chrX:49689926T>G | 7.72            | -10.90          | -18.62         | DONOR | CLCN5      |          |          | Y         |
| Gene     | Location | rsID       | Av. Het. | Validated |
|----------|----------|------------|----------|-----------|
| CCDC93   |          |            | Y        |           |
| DDX11    |          |            | Y        |           |
| COBLL1   |          |            | Y        |           |
| WDR46    |          |            | P        |           |
| GPR133   |          |            | NE       |           |
| NAV2-AS1 |          |            | N        |           |
| RPL10    |          |            | N        |           |
| TIMMDC1  |          |            | Y        |           |
| C17orf80 |          | rs113825288| 0.007    | Y         |
| C8orf59  |          | rs67573812 | 0        | NE        |
| CFL1     |          | rs667555   | 0        | N         |
| PRMT3    |          | rs6483700  | 0.007    | NE        |
| RARS2    |          | rs77773960 | 0        | Y         |
| RP11-403P17.5.1 | rs2241619 | 0        | Y         |
| SPTLC1   | EXON     |            |          |           |
| RPSA     | INTRON   |            |          |           |
| AC068580.6.1 | EXON    |          |          |           |
| MPI      | EXON     |            |          |           |
| S100A1   | EXON     |            |          |           |
| B4GALT2  | INTRON   |            |          |           |
| IKBKE    | EXON     | rs1539242  | 0.006    | Y         |
| UNC5B    | INTRON   | rs41278006 | 0.006    | N         |
| GOLGA2   | EXON     | rs74686374 | 0.006    | AE        |
| PPARG    | EXON     | rs1797895  | 0.005    | N         |
## B. U251 Splicing Mutations

| Leaky | \( R_{i,\text{initial}} \) | \( R_{i,\text{final}} \) | \( \Delta R_i \) | Site | Gene | rsID | Av. Het. | Validated |
|-------|-----------------|-----------------|----------------|------|------|------|----------|-----------|
| **Novel Variants** |
| chr3:10123021T>A | 5.95 | 3.54 | -2.41 | ACCEPTOR | FANCD2 | | | N |
| chr16:78312497C>A | 10.24 | 6.67 | -3.57 | ACCEPTOR | WWOX | rs8050128 | 0 | NE |
| chr3:123554710C>T | 8.47 | 5.29 | -3.18 | DONOR | MYLK | rs144796555 | 0 | NE |
| chr7:128117227G>A | 5.48 | 2.47 | -3.01 | DONOR | METTL2B | rs76349929 | 0 | Y |
| chr6:121767966T>A | 14.37 | 11.77 | -2.60 | ACCEPTOR | GJA1 | rs56199702 | 0 | N |
| chr16:88051014A>G | 6.87 | 4.34 | -2.52 | DONOR | BANP | rs6540151 | 0 | NE |
| chr7:98533187T>G | 9.09 | 7.16 | -1.93 | ACCEPTOR | TRRAP | rs62472016 | 0 | Y |
| chr15:68445912T>A | 16.57 | 14.72 | -1.85 | ACCEPTOR | PIA51 | rs11633620 | 0.005 | AM |
| chr2:32961745T>A | 8.99 | 7.14 | -1.85 | ACCEPTOR | TTC27 | rs10200333 | 0 | P |
| chr6:30679289G>C | 6.44 | 4.72 | -1.72 | ACCEPTOR | MDC1 | rs147822906 | 0 | P |
| chr1:65830299T>G | 16.58 | 15.06 | -1.52 | ACCEPTOR | DNAJC6 | rs2296479 | 0 | P |
| chr6:30679289G>C | 3.41 | 1.96 | -1.45 | ACCEPTOR | MDC1 | rs147822906 | 0 | NE |
| chrX:119402264A>C | 11.63 | 10.30 | -1.33 | ACCEPTOR | FAM70A | rs41300936 | 0 | Y |
| chr1:33318561T>C | 5.86 | 4.55 | -1.32 | DONOR | S100PB | rs702836 | 0 | NE |
| chr2:37265193A>C | 10.76 | 9.64 | -1.12 | ACCEPTOR | HEATR5B | rs139662639 | 0 | NE |
| chr17:73698557C>T | 9.81 | 8.70 | -1.12 | ACCEPTOR | SAP30BP | rs820232 | 0.009 | NE |
| chr7:74152376T>C | 7.73 | 6.64 | -1.09 | ACCEPTOR | GTF2I | rs810377 | 0 | N |
| **Known Variants < 1% Av. Het.** |
| chr14:32142782T>G | 10.35 | -8.27 | -18.62 | DONOR | NUBPL | | | NE |
| chr19:33904479A>C | 10.07 | -8.55 | -18.62 | DONOR | PEPD | | | N |
| chr11:34905054G>C | 9.32 | 0.54 | -8.78 | ACCEPTOR | APIP | | | Y |
| chr14:43308444A>C | 9.07 | -9.55 | -18.62 | DONOR | RP11-342M1.4.1 | | | NE |
| chr12:4700466T>G | 6.83 | -11.79 | -18.62 | DONOR | Dyrk4 | | | P |
| chr3:40570938T>G | 5.73 | -12.90 | -18.62 | DONOR | ZNF621 | | | N |
| Chr     | Gene     | Site   | RsID  | Av. Het | Validated |
|---------|----------|--------|-------|---------|-----------|
| chr14:75128861G>C | KIAA0317 | ACCEPTOR | -2.43 | 9.24 | 11.67 | N |
| chr21:45539300A>G | PWP2 | DONOR | 4.77 | 7.77 | 3.01 | N |
| chr19:620369T>G | POLRMT | ACCEPTOR | 5.92 | 7.76 | 1.84 | N |
| chr2:73228685T>G | SFXN5 | ACCEPTOR | 4.16 | 6.00 | 1.84 | N |
| chr13:73369248A>T | POLRMT | ACCEPTOR | 5.92 | 7.76 | 1.84 | N |
| chr17:30303021A>G | SUZ12 | EXON | 1.45 | 4.63 | 3.18 | N |
| chr22:50954300T>G | NCAPH2 | DONOR | -14.18 | 4.45 | 18.63 | N |
| chr8:144876251G>C | CYP20A1 | ACCEPTOR | 0.15 | 3.88 | 3.73 | NE |
| chr10:135104007G>A | TUBGCP2 | DONOR | 1.62 | 2.71 | 1.09 | NE |
| chr10:70661701T>G | DDX50 | INTRON | -16.58 | 2.06 | 18.63 | N |
| chr2:204150427G>C | CYP20A1 | ACCEPTOR | 4.03 | 5.50 | 1.47 | N |
| chr3:12447814C>G | PPARG | DONOR | 0.26 | 4.13 | 3.87 | NE |
| chr13:20611221G>A | ZMYM2 | INTRON | -0.56 | 2.04 | 2.61 | N |

**C. A431 Mutations**

| Chr     | Gene     | Site   | RsID  | Av. Het | Validated |
|---------|----------|--------|-------|---------|-----------|
| chr18:20529676G>A | RBBP8 | DONOR | 6.17 | 3.16 | -3.01 | Y |
| chr1:179310449A>G | SOAT1 | DONOR | 10.01 | 7.40 | -2.61 | P |
| chr1:12318166T>C | VPS13D | DONOR | 11.69 | 9.97 | -1.72 | Y |
| chr1:161013165G>T | USF1 | ACCEPTOR | 4.89 | 3.59 | -1.30 | Y |
| chr1:36360707C>A | EIF2C1 | ACCEPTOR | 8.67 | 7.56 | -1.11 | N |
| chr2:136148401A>T | ZRANB3 | ACCEPTOR | 21.77 | 18.95 | -2.82 | P |
| chr1:169272451A>T | NME7 | ACCEPTOR | 9.96 | 8.49 | -1.47 | P |
| Inactivating | \( R_{i,\text{initial}} \) | \( R_{i,\text{final}} \) | \( \Delta R_i \) | Site | Gene | rsID | Av. Het. | Validated |
|-------------|-----------------|-----------------|-----------|------|------|------|--------|-----------|
| **Novel Variants** | | | | | | | | |
| chr1:12064177T>G | 9.57 | -9.06 | -18.62 | DONOR | MFN2 | Y | | |
| chr12:50480538G>C | 8.46 | -3.21 | -11.67 | ACCEPTOR | SMARC1D | Y | | |
| chr11:62341301A>C | 7.84 | -10.78 | -18.62 | DONOR | EEF1G | P | | |
| chr2:213886617A>C | 7.74 | -10.88 | -18.62 | DONOR | AC093865.1.1 | PS | | |
| chr2:165600195A>C | 6.36 | -12.26 | -18.62 | DONOR | COBLL1 | N | | |
| chr11:20142118A>C | 4.52 | -14.11 | -18.62 | DONOR | NAV2-AS1 | PS | | |
| chr7:98591160G>C | 4.33 | -7.34 | -11.67 | ACCEPTOR | TRRAP | Y | | |
| chr8:103250667A>C | 3.60 | -15.02 | -18.62 | DONOR | RRM2B | Y | | |
| chr20:55045807G>A | 3.41 | 0.40 | -3.01 | DONOR | C20orf43 | PS | | |
| chr22:36722714G>C | 2.90 | 1.18 | -1.72 | ACCEPTOR | MYH9 | N | | |
| chr12:107374398C>T | 2.11 | -16.52 | -18.63 | DONOR | MTERFD3 | N | | |
| **Known Variants < 1% Av. Het.** | | | | | | | | |
| chr12:122740009C>T | 4.77 | -6.11 | -10.88 | ACCEPTOR | RP11-512M8.6.1 | rs10744155 | 0 | PS | |
| chr11:20529886G>A | 4.71 | -13.92 | -18.63 | DONOR | PRMT3 | rs6483700 | 0.007 | PS | |
| chr14:64669514T>A | 1.89 | -0.83 | -2.72 | ACCEPTOR | SYNE2 | rs189611387 | 0 | Y | |

| Cryptic Splicing | \( R_{i,\text{initial}} \) | \( R_{i,\text{final}} \) | \( \Delta R_i \) | Site | Gene | Location | rsID | Av. Het. | Validated |
|-----------------|-----------------|-----------------|-----------|------|------|--------|------|--------|-----------|
| **Novel Variants** | | | | | | | | | |
| chr5:177637135C>G | 5.83 | 9.57 | 3.73 | DONOR | HNRPAB | EXON | N | | |
| chr16:57180998G>C | -0.56 | 8.22 | 8.78 | ACCEPTOR | CPNE2 | EXON | N | | |
| Chromosome:Position | Variant Description | 
|---------------------|--------------------|
| chr16:27790837G>C | -1.56 7.22 8.78 ACCEPTOR KIAA0556 EXON | N |
| chr19:14000568T>G | 5.85 7.15 1.30 ACCEPTOR C19orf57 EXON | N |
| chr8:133822931G>C | 5.02 6.47 1.45 ACCEPTOR PHF20L1 EXON | N |
| chr1:101437592A>C | 4.63 6.36 1.73 ACCEPTOR SLC30A7 INTRON | N |
| chr10:3191408C>A | 5.02 6.35 1.33 DONOR PITRM1 EXON | N |
| chr14:10387139A>G | -4.68 6.20 10.88 ACCEPTOR MARK3 INTRON | N |
| chr21:47676584A>G | 4.29 5.89 1.60 DONOR MCM3AP EXON | N |
| chr16:27790837G>C | 3.15 5.77 2.62 ACCEPTOR KIAA0556 EXON | N |
| chr22:20113022T>G | -12.96 5.67 18.63 DONOR RANBP1 EXON | N |
| chr6:7227103T>G | -14.30 4.34 18.63 DONOR RREB1 INTRON | N |
| chr16:90051055G>A | 1.05 3.58 2.52 DONOR AFG3L1P EXON | PS |
| chr1:206647742A>G | 6.67 9.68 3.01 DONOR IKBKE EXON rs1539242 0.006 N |
| chr3:12447814C>G | 0.26 4.13 3.87 DONOR PPARG EXON rs1797895 0.005 LF |
| chr2:232087474A>G | -15.64 3.00 18.63 DONOR ARMC9 EXON rs1626450 0.0004 N |

**Bolded** Gene names are previously established tumor driver genes. Legend to symbols in Validation column: Y: **yes validated**; Y*: **yes validated**, based on HapMap Phase II Affymetrix Exon array; P: probable: consistent with mutation altering splicing, but not conclusive; N: not validated; NE: not expressed in cell line. Certain RNAseq results were not interpretable: LF: low fidelity splicing – significant intron inclusion; PS: multiple pseudogenes or duplicated exons; AE: rare alternate exon – insufficient data; AM: ambiguous due to significant exon skipping/alternative splicing throughout gene.