**Title:** An algorithm to detect abnormal mRNA splicing and assess its clinical impact in individuals undergoing genetic testing for hereditary cancer syndromes

**Authors:**

Nick Kamps-Hughes,[1] Victoria Carlton,[1] Laure Fresard,[1] Steve Osazuwa,[1] Elizabeth Starks,[1] John J. Vincent,[1] Sarah Albritton,[1] Robert L. Nussbaum,[1] Keith Nykamp[1]*

[1]Invitae Corporation, San Francisco, CA, 94103, USA

*Correspondence: Keith Nykamp, keith.nykamp@invitae.com

**ABSTRACT**

Nearly 14% of germline variants that cause disease do so by disrupting normal mRNA splicing. Most (67%) of the DNA variants predicted to disrupt splicing may be in some genetic test's reportable range (coding exons +/- 20 bp of flanking intron) but not in the canonical dinucleotide donor or acceptor splice sites, and are often considered variants of uncertain significance (VUS). We developed and validated an RNA sequencing protocol—Splice Effect Event Resolver (SEER)—to detect statistically significant alterations in the splicing patterns of specific genes relative to normal controls. SEER was used to analyze leukocyte RNA of 63 hereditary cancer syndrome genes in 20,317 individuals undergoing clinical genetic testing; although 3,563 (17.5%) had at least one DNA variant predicted to affect splicing, only 971 demonstrated altered splicing—a 73% false positive rate. Integrating RNA results into variant interpretation allowed reclassification of VUS to P/LP in 0.4% and B/LB in 5.9% of 20,317 patients. An additional 0.2% of patients were found to have a DNA variant outside of the test's reportable range but associated with a significant splicing change. RNA analysis had a significantly higher impact on definitive classifications (i.e., P/LP and B/LB) in non-White individuals than in non-Hispanic White individuals, illustrating that RNA splicing analysis may reduce ancestral and ethnic disparities in genetic testing utility.

**KEYWORDS:** RNA sequencing; splicing; variant interpretation; diagnostic testing; splice site prediction

## INTRODUCTION

Disruption of normal mRNA splicing is a common cause of genetic disease. Approximately 14% of all germline variants currently classified as pathogenic or likely pathogenic (P/LP) likely cause disease through misplicing,[1] and the vast majority (>90%) of these P/LP splicing variants disrupt the essential dinucleotides (±1-2) of the canonical donor and acceptor splice sites (referred to as CSS). In contrast, many DNA variants outside of the CSS dinucleotides are predicted by commonly used algorithms to affect splicing, but most (69%) of these potential splicing variants (PSpVs) are classified as variants of uncertain significance (VUS),[1] demonstrating the challenge of classifying non-CSS PSpVs in the absence of confirmatory RNA testing. Knowing which of these PSpVs have a clinically significant impact on splicing could result in fewer VUS and increased actionability for genetic testing.

Here, we report findings from a two-phase study. In the first phase, we developed and validated the Splice Effect Event Resolver (SEER) algorithm, which used RNA sequencing data as an input to aid in the classification of PSpVs. In the second phase, we used SEER to analyze RNA from more than 20,000 individuals undergoing paired DNA and RNA sequencing of hereditary cancer syndrome genes.

SEER builds upon several overlapping RNA analysis methods[2–6] to quantify exon-exon junctions in short-read sequencing of cDNA and measure statistically significant deviations in the splicing patterns of specific genes relative to a panel of control individuals. Significant alterations in the transcripts of a gene are then linked together into a localized splicing event (referred to as an ‘abnormal splicing event group’), which can be further evaluated for its potential impact on translation and protein function with high fidelity. SEER distinguishes aberrant splicing from the naturally occurring alternative splicing that generates protein diversity and functional specificity

across tissue types.[7–9] Our results also show that SEER allows a definitive interpretation of a significant fraction of rare PSpVs and supports the discovery of previously undetected variants outside the reportable range of next-generation sequencing (NGS) DNA tests.

## METHODS

### Cohorts

In phase 1 of this two-phase study (Figure S1), we developed and validated the SEER algorithm and NGS methods using a retrospective cohort of 532 research participants who had prior genetic testing for hereditary cancer syndromes performed at Invitae and had either 1) a germline DNA variant that was known or predicted to alter splicing or 2) a personal or family history strongly suggestive of a hereditary cancer syndrome but no detected P/LP variants. Informed written consent was obtained under a protocol approved by the WCG Institutional Review Board (#20190811). As a reference panel, we used 273 anonymous samples from presumed healthy male and female blood donors (BioIVT, NewYork) enriched for self-reported ancestries underrepresented in genetic studies, including African American (51.2%) and Hispanic (30.4%).

In Phase 2, a prospective cohort of 20,317 individuals referred to Invitae for germline hereditary cancer gene testing underwent paired DNA and RNA sequencing between July 2021 and May 2022. SEER was used to analyze the leukocyte RNA of 63 hereditary cancer genes in these individuals. Use of de-identified samples and data was approved by an independent institutional review board (WCG IRB #20161796).

### Selection of genes for RNA analysis

A total of 85 hereditary cancer genes were initially considered for RNA analysis (Table S1). Twenty-two of these were ultimately excluded from the final assay because either 1) they were

expressed at too low levels in leukocyte transcripts to allow statistically significant demonstration of a splicing change, or 2) they were genes for which loss-of-function, most commonly caused by abnormal splicing, was not a known mechanism for an increased risk for cancer.

**Detection of statistically significant splicing changes**

Targeted RNA-Seq data generated from blood samples (Supplemental Methods) was used to extract junction counts across 63 genes of interest for input into SEER (Table S1). To assess the statistical significance of splice junction usage in a patient sample, SEER (Figure 1 A,B) measured usage of known and novel splice junctions using "percent spliced in" (PSI).[11,12] The PSI of each junction was defined as the ratio of the number of reads supporting the junction divided by the total number of reads overlapping the junction (i.e., supporting reads plus non-supporting reads). SEER assessed the statistical significance of each splicing change in a patient sample by comparing it to the splice junction usage from the 273-sample reference panel using a beta-binomial test (Supplemental Methods). We called PSI changes with $p \leq 10^{-3}$ as highly statistically significant and flagged these junctions for follow-up (described below).

To evaluate the magnitude and functional impact of abnormal splicing events on protein function, we derived the PSI fold change (PSI-X) metric, which is the ratio of the PSI for the most *relevant canonical junction* in a patient sample divided by the PSI for the same junction in control individuals. We chose PSI-X as our primary metric for this evaluation because loss of canonical junctions, rather than gain of abnormal junctions, correlated more reliably with loss of a normally spliced transcript, which in turn was expected to correlate with loss of functional protein. Moreover, the genes included in SEER cause disease through a loss-of-function mechanism, so a metric that provides a relative reduction (i.e., fold change) of the normally spliced transcript for these genes is expected to correlate with the likelihood that an abnormal splicing event causes disease. For intronic variants, the most relevant canonical junction was

defined as the junction that spliced out the intron containing the variant; for exonic variants, we identified all canonical junctions that shared a breakpoint with the newly created junction with the highest PSI value. If more than one canonical junction was selected, the PSI values were averaged.

**Assessing the clinical significance of splicing changes**

To assist with interpreting the clinical impact of flagged junctions, SEER assembled an "abnormal splicing event group" for each statistically significant junction. Specifically, SEER used flagged junctions as seeds to recursively find other significantly altered ($p \leq 0.05$) junctions that shared breakpoints. Trained variant scientists reviewed all event groups that had at least one significant ($p \leq 10^{-3}$) breakpoint within 50 bp of a PSpV (Supplemental Methods). PSpVs were predicted using MaxEntScan,[13] SpliceSiteFinder-Like,[14] and the Alamut Splicing Module (Interactive Biosoftware, Rouen, France, version 1.4.4_2016.02.03).

Once we identified event groups with a significant reduction in PSI-X, we assessed the predicted impact of the abnormal junctions on protein function. This was an essential step for interpretation of the DNA variants associated with significant event groups since abnormal splicing may result in translation of a protein that retains some level of normal function, thereby compensating for the loss of the normally spliced mRNA. As such, each junction in an event group was labeled as *canonical* (i.e., joining two adjacent exons in the canonical transcript); *exon skipping* (i.e., joining two non-adjacent exons in the canonical transcript with at least one exon being skipped); *partial exon exclusion,* in which a portion of a canonical exon was not included in the transcript; or *partial intron inclusion,* in which a portion of an intron was included in a transcript. At the event group level, the label for the abnormal junction with the highest PSI value was used, with two exceptions: the event group was labeled (1) *cryptic exon* if two abnormal junctions flanked an intronic sequence and (2) *complex* if the PSI of the abnormal

junction with the highest PSI value was less than double the PSI of the abnormal junction with the next highest PSI value in the event group. All event groups, other than complex, were also labeled according to whether they were expected to cause nonsense-mediated decay (NMD). Event groups causing a frameshift and/or premature termination codon upstream of the last coding exon were labeled as *NMD+.* Those expected to result in in-frame insertions anywhere in the transcript, or in deletions or premature termination codons downstream of the last exon junction, were labeled as *NMD-*.

**Detection and classification of deep intronic variants**

Deep intronic variants are rare variants outside a test's reportable range that are missed by standard panel-based NGS tests. What constitutes a deep intronic variant depends on how much intron flanking an exon is included in a test's reportable range. When SEER identified an abnormal splicing event group at the variant discovery significance threshold (i.e., $p \leq 10e^{-5}$) *and* one or more abnormal junctions exhibited a fold-change in PSI ≥10,000, we manually examined the primary NGS data for a DNA variant that may explain the abnormal splicing event. If none were found in the primary data, we performed long-read DNA sequencing (PacBio, CA) throughout the region surrounding the abnormal splicing event. If a PSpV was identified, it was added to the clinical report and interpreted by the variant scientist team. If a PSpV could not be identified, the abnormal splicing change was presumed to be a technical false positive or the result of natural variation in splicing and not included in the clinical report.

**SEER validation**

SEER was validated using 50 positive control samples from the Phase 1 research cohort containing 46 unique positive control variants: 10 samples with 7 unique PSpVs associated with a splicing change in the literature; 13 samples with 12 unique PSpVs in the essential GT or AG dinucleotides but without published functional evidence; and 27 samples with 27 unique PSpVs

found to be associated with a splicing alteration during Phase 1. True positives were defined as a significant abnormal splicing event ($p \leq 0.001$) within 50 nucleotides of the PSpV. False positives were any significant abnormal splicing event ($p \leq 0.001$) detected throughout the 63-gene panel for a canonical junction not within 50 bp of a known PSpV. All junctions not near a PSpV from the 50 positive control samples plus all junctions in an additional 16 negative control samples that had variants shown in Phase 1 not to impact splicing constituted a total of 66 negative control samples containing 66 negative control variants. All 66 control samples were used to calculate the sensitivity, specificity, accuracy, reproducibility, and repeatability of the SEER algorithm (Figure 1 C,D and Table S2). Reproducibility and repeatability were defined as the ratio of identical positive or negative splice calls adjacent to a DNA variant across dates (reproducibility) and replicates (repeatability).

**Integration of RNA splicing changes as evidence in DNA variant classification**

To assist with classification of PSpVs, we created a new category of evidence, referred to as *Observed RNA Effects* (Table S3), to incorporate into our semi-quantitative variant classification framework, Sherloc.[10] Each criterion within this new evidence category was assigned pathogenic (P) or benign (B) point scores. Point scores were assigned and calibrated during Phase 1 by applying Observed RNA Effects criteria to more than 300 DNA variants known or expected to impact splicing. Final classifications were assigned based on the sum of all available evidence, including the Observed RNA Effects criterion when applicable. Differences in reclassification between individuals undergoing RNA analysis from self-reported ethnicities were tested using a one-sided Fisher Exact Test comparing the proportion of individuals with reclassified variants in individuals from the most common self-reported ethnicities.

## RESULTS

**Validation of targeted RNA sequencing and detection of altered splicing with SEER**

In the research cohort (Phase 1), SEER detected a significant splicing change in 45 of 46 positive control DNA variants in the 50 positive control samples (97.8% sensitivity; Figure 1C) while maintaining a specificity of 99.7%. We used two of the remaining four positive control samples, along with eight of the negative control samples, to perform reproducibility and repeatability testing in triplicate across 10 samples. All 30 (100%) of the reproducibility samples and 29 (96.7%) of 30 repeatability samples yielded the expected results (Figure 1C). True positive and true negative calls demonstrated strong separation at the p-value threshold used for detecting splicing effects (Figure 1D), and variants found in multiple samples demonstrated consistent PSI values (data not shown). SEER also enabled the detection of deep intronic variants outside of the reportable range in four individuals from the phase 1 research cohort, all of whom had presented with a persuasive personal and/or family history of hereditary cancer syndrome (Figure S1).

**Prospective analysis of paired DNA and RNA testing**

In Phase 2 of the study, 20,317 individual blood samples were processed for RNA extraction and sequencing. Across the entire population, 3,563 (17.5%) of the patients had at least one DNA variant identified by NGS that was a PSpV (Table 1). Abnormal splicing, defined by a statistically significant event group, coincided with 605 unique PSpVs (Table 2, Row 1) in a total of 971 patients (4.8% of all prospective samples) (Table 1). These 605 PSpVs were detected in 54 of the 63 genes analyzed by SEER (Table S4). We observed normal splicing associated with a PSpV in 2,487 patients (73% of patients with a PSpV), illustrating the high false positive rate for the algorithms and heuristics used to identify PSpVs.

**Interpretation of DNA variants associated with abnormal splicing event groups**

SEER detected abnormal splicing events associated with PSpVs with very high accuracy (99.7%). The next step for interpreting PSpVs required determining the degree to which normal

splicing was disrupted, for which we derived the canonical PSI-X metric. Importantly, plotting the distribution of PSI-X for all definitively classified P/LP and B/LB PSpVs (Figure 2A) and computing a receiver operating characteristic (ROC) curve (Figure 2B) for these data resulted in an area under the ROC curve (AUROC) of 0.91 for PSI-X, demonstrating good discriminatory power of PSI-X alone for separating between known disease-causing PSpVs and benign splicing variants.

We then assessed PSI-X for heterozygous DNA deletions encompassing one or more exons observed in the prospective cohort, with the assumption that these exonic deletions represented complete loss of junction usage for the flanking splice sites. With these deletion variants we found a distribution in PSI-X from 0.3 to 0.75 (Figure 2C), suggesting that a PSI-X <0.8 confidently correlated with complete loss of a canonical splice junction from one allele. If we were to classify all PSpVs with a PSI-X <0.8 as P/LP, and all PSpVs with a PSI-X ≳0.8 as B/LB, our overall accuracy for variant interpretation by PSI-X alone would be 86%. However, as a subsequent step in interpreting PSpVs, we included in the analysis the likelihood that an abnormal splicing event group would lead to NMD and, if not, the predicted impact on protein function. In this way, we could eliminate six of nine B/LB variants with a PSXI <0.8 as false positives, since the abnormal transcripts created were not expected to result in NMD; five event groups resulted in small in-frame (≤ 6 amino acids) deletions or insertions, and one event group resulted in alternative usage of the first coding exon. By combining a requirement that the PSI-X be <0.8 *with* the requirement that the abnormal transcripts lead to NMD, we could achieve a positive predictive value (PPV) of 98.9% for P/LP variants. For P/LP variants ≳0.8, we could eliminate 9 of 42 false negatives because they were expected to disrupt the protein directly through coding sequence changes (i.e., nonsense, missense, in-frame deletion or insertion) and not through abnormal splicing. Therefore, by combining a PSI-X ≳0.8 with the requirement that the PSpV not be a missense variant or directly change the coding sequence, we could achieve

a negative predictive value (NPV) of 61.2%. However, since many P/LP variants with PSI-X ≳0.8 have been reported in the published literature to cause disease by a splicing mechanism, more work needs to be done to differentiate between P/LP and B/LB variants associated with a PSI-X ≳0.8.

**Characterization of abnormal splicing events in the prospective cohort**

Each PSpV associated with an abnormal splicing event group identified by SEER was classified using Observed RNA Effects evidence within the Sherloc interpretation framework. The most abundant splicing change in the prospective cohort was exon skipping, with partial exon exclusion and partial intron inclusion events being the next most abundant events. Partial exon exclusion, partial intron inclusion, and complex events displayed the largest discrepancies between PSpV definitive classifications (i.e., P, LP, LB, B) and VUS, indicating that these were the most challenging event groups to interpret (Figure 3A).

By evaluating the location of PSpVs in relation to the exonic sequence, we found that the largest number of P/LP variants relative to VUS were within the CSS (-1, -2, +1, +2; Figure 3B). In addition, there was an enrichment at positions +3, +4, and +5 from the canonical donor splice site, and a cluster at -11, -10 and -9, upstream of the canonical acceptor site that most often creates a novel acceptor splice site. Overall, we observed abnormal splicing events associated with 605 PSpVs (Table 2), with 93% of these PSpVs located within our test's reportable range of +/- 20 bp flanking the exons.

For all observed PSpVs with RNA analysis results, 20.2% of the DNA variants were ultimately classified as P/LP, 43.3% classified as a VUS, and 36.5% classified as benign/likely benign (B/LB) (Table 2, row 3). As expected, abnormal splicing correlated with a much higher rate of P/LP variants (42.9%) and much lower rate of B/LB variants (11.6%) (Table 2, row 1).

Conversely, normal splicing correlated with a much lower rate of P/LP variants (9.4%) and much higher rate of B/LB variants (48.3%) (Table 2, row 2). Importantly, we found that 6.3% of all patients from the phase 2 cohort had a P/LP or B/LB variant that would have been a VUS by our interpretation framework if it were not for the added RNA splicing data from our study: 75 patients (0.4% of all samples) would have received a VUS instead of a P/LP classification, and 1,204 patients (5.9% of all samples) would have received a VUS instead of a B/LB classification (Table 1). When comparing these data across self-reported ancestries, we found that individuals with self-reported Hispanic (9.4%, p=8.2e-12), Black/African American (10.4%, p=1.8e-16), or Asian (9.1%, p=1.4e-5) ancestries had statistically significantly higher rates of definitive classifications (P/LP or B/LB) than Non-Hispanic White populations (4.8%) due to the addition of supplemental RNA splicing data (Table 3).

We also found that 42 (0.2%) of all 20,317 patients carried 35 unique variants associated with abnormal splicing outside of the reportable range (Table 1). Of these 35 variants, 6 (17.1%), seen in nine patients, could be classified as P/LP, representing a 0.04% yield of P/LP deep intronic variants in the entire cohort (9/20,317 patients). An additional two variants, seen in two patients (4.8%), were classified as benign, and 27 of the 35 unique variants, present in 31 patients (73.8%), were classified as VUS (Figure 4).

## DISCUSSION

We describe SEER, an accurate and scalable approach for detecting and quantifying splicing alterations in peripheral blood leukocyte transcripts associated with rare DNA variants in patients undergoing genetic testing for clinical hereditary cancer syndromes. We demonstrate that adding RNA analysis to sequence analysis for hereditary cancer syndrome genes facilitates the interpretation of DNA variants expected or predicted to alter splicing. In addition, SEER can

detect the 7% of deep intronic DNA variants affecting splicing as defined by a reportable range of +/- 20 bp flanking the exons of our NGS panel tests.

We found that supplemental RNA analysis led to a more definitive classification of DNA variants in 6.3% of individuals in a prospective cohort undergoing testing for hereditary cancer. Combined DNA and RNA testing led to a positive result in an additional 0.4% of patients, while 5.9% of patients tested had a B/LB variant that would have been classified as a VUS without RNA analysis, thereby effectively reducing VUS rates. Furthermore, the rate of definitive classifications (i.e., P, LP, LB, B) dependent on RNA analysis was significantly higher in Black/African American, Asian, and Hispanic individuals combined than in non-Hispanic White individuals. The preferential impact of RNA analysis on VUS reclassification in these less-studied populations probably resulted from RNA analysis providing direct, objective evidence that was not limited by the insufficient representation of individuals of non-European origin in variant databases or in the published literature. This indicates that including RNA analysis in the interpretation of DNA variants predicted to impact splicing may preferentially reduce VUS rates for less-studied ethnic populations, thereby reducing health disparities.

These results also highlight the nuances and difficulties of incorporating sample-based RNA splicing analysis into variant interpretation frameworks used in routine genetic testing. Alternative splicing is very common, and it can be challenging to pinpoint specific splicing alterations associated with a hereditary disease. For example, greater than 96% of all patient samples in our prospective cohort had at least one statistically significant splicing change at the $p \leq 0.001$ threshold compared to the reference panel of normal samples. This statistical threshold was not multiple hypothesis-corrected because it was meant to accurately determine whether a specific PSpV altered normal splicing. By focusing scientific review on splicing alterations with a PSpV within 50 bp of the nearest altered junction, we reduced the frequency

of abnormal splicing event groups to 4.8% of all patient samples, while maintaining 99.7% accuracy for detecting abnormal splicing events previously known to be associated with DNA variants in our validation of SEER.

Even with supplemental RNA splicing data, the biological impact and clinical relevance of DNA variants associated with a significant splicing alteration are difficult to assess, often still resulting in VUS classifications. This is particularly true for variants found outside of the reportable range of the test since they are often observed in a single individual and lack the additional corroborating evidence needed to make a definitive classification, such as population frequency data in publicly available databases.

Another complicating factor is that not all abnormal splicing events are complete. A heterozygous DNA variant may result in a reduction of normal mRNA splicing that ranges from 0% (i.e., no reduction) to 50% (i.e., complete loss of normal splicing from the variant allele), with everything in between considered partial loss, also referred to as 'leakiness.'[15,16] Importantly, because a PSpV that leads to a complete loss of normal splicing is more likely to cause disease than a PSpV that leads to a partial loss, accurate classification of PSpVs using abnormal splicing data requires a confident measure of the amount of reduction. We developed such a measurement to quantify the reduction in normal mRNA splicing—PSI-X—and determined that splicing alterations in blood samples associated with a PSI-X <0.80 correlate with sufficient loss of normal splicing from one allele to cause disease and qualify as pathogenic evidence for autosomal dominant hereditary cancer syndromes within our variant classification framework. Although a PSI-X threshold of 0.80 demonstrates high discriminatory performance (AUROC = 0.91) for PSpVs with abnormal splicing events, we still find that 15% of B/LB variants fall below the threshold and 14% of P/LP variants are above the threshold. To achieve a level of confidence in the final classification of PSpVs associated with abnormal splicing suitable for

clinical testing, it is essential that trained variant scientists review the expected effect of abnormal splicing on translation and protein function, and then integrate these findings into a DNA variant interpretation framework. In this way, we demonstrated that by pairing DNA and RNA testing, we can increase the number of patients who have an actionable finding while achieving a PPV of 98.9%. On the other hand, even with a careful review of the abnormal event groups associated with PSI-X ≳0.8, it was still difficult to discriminate between P/LP and B/LB variants, as demonstrated by a much lower NPV (61.2%). As a result, to limit the likelihood of a false negative result for our clinical testing patients, we do not award P or B points in the Sherloc framework for statistically significant abnormal splicing event groups with PSI-X ≳0.8. PSpVs associated with these higher PSI-X values will generally be VUS unless complementary evidence (i.e., family segregation, population frequency, additional splicing studies) support a more definitive classification.

Although this study clearly demonstrates that RNA splicing data can aid in variant classification and identification, the results should be interpreted in the context of the following technical and biological limitations. First, the 63 hereditary cancer genes included in this study are tumor suppressor genes that are not highly expressed in whole blood leukocyte RNA. We expect each gene to have tissue-specific expression and alternatively spliced products that differ from what can be detected in a blood sample. As a result, abnormal splicing observed from whole blood leukocytes may not reflect a biological change in the tissue of interest, while the absence of a splicing alteration in whole blood may not always correlate with an absence of abnormal splicing in the disease-relevant tissue. Second, the SEER algorithm is designed to evaluate changes in junction usage. A small fraction of splicing alterations result in complete intron retention, which escapes detection by the SEER algorithm in its current form. This likely explains why, in our study, several known P/LP DNA variants had PSI-X ≳0.80 with transcripts that underwent NMD. Further development is underway to include a method for evaluating the relative frequency of

complete intron inclusion in the SEER algorithm, which will better resolve the uncertainty of variants with high PSI-X. Third, although PSI values and aberrant junctions were highly consistent for the same variant across multiple samples, a few DNA variants showed high levels of variability. Most of these correlated with the presence of *cis*-acting elements, such as secondary DNA variants, suspected to modify the splicing impact of the primary DNA variant. Currently it is not possible to predict *a priori* which splicing changes are impacted by modifiers, but with more samples sequenced for the same DNA variants we may be able to identify patterns in the DNA sequence that modulate the PSI and alternative splicing events within these tumor suppressor genes.

In summary, RNA analysis in whole blood is a valuable tool for finding and assessing the impact of DNA variants suspected of affecting RNA splicing, thereby altering gene expression. It helps to reduce VUS rates in a fraction of individuals who have DNA variants predicted to affect splicing, and supports the identification of deep intronic variants outside the test's reportable range. However, interpretation of the splicing information is challenging and there are limitations in the techniques and in our biological understanding of splicing. By combining comprehensive DNA and RNA sequencing with the validated SEER algorithm, we will continue to collect data from many thousands of individuals referred for hereditary cancer syndrome testing, develop deeper insights into the clinical relevance of splicing alterations, and improve the interpretation of inherited DNA variants— leading to a further reduction in uncertainty for hereditary cancer gene testing.

**Declaration of Interests**

All authors are employed by and shareholders of Invitae, Inc.

## REFERENCES


1. Truty, R., Ouyang, K., Rojahn, S., Garcia, S., Colavin, A., Hamlington, B., Freivogel, M., Nussbaum, R.L., Nykamp, K., and Aradhya, S. (2021). Spectrum of splicing variants in disease genes and the ability of RNA analysis to reduce uncertainty in clinical interpretation. Am. J. Hum. Genet. *108*, 696–708.

2. Li, Y.I., Knowles, D.A., Humphrey, J., Barbeira, A.N., Dickinson, S.P., Im, H.K., and Pritchard, J.K. (2018). Annotation-free quantification of RNA splicing using LeafCutter. Nat. Genet. *50*, 151–158.

3. Frésard, L., Smail, C., Ferraro, N.M., Teran, N.A., Li, X., Smith, K.S., Bonner, D., Kernohan, K.D., Marwaha, S., Zappala, Z., et al. (2019). Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts. Nat. Med. *25*, 911–919.

4. Mertes, C., Scheller, I.F., Yépez, V.A., Çelik, M.H., Liang, Y., Kremer, L.S., Gusic, M., Prokisch, H., and Gagneur, J. (2021). Detection of aberrant splicing events in RNA-seq data using FRASER. Nat. Commun. *12*, 529.

5. Jenkinson, G., Li, Y.I., Basu, S., Cousin, M.A., Oliver, G.R., and Klee, E.W. (2020). LeafCutterMD: an algorithm for outlier splicing detection in rare diseases. Bioinformatics *36*, 4609–4615.

6. Vaquero-Garcia, J., Barrera, A., Gazzara, M.R., González-Vallinas, J., Lahens, N.F., Hogenesch, J.B., Lynch, K.W., and Barash, Y. (2016). A new view of transcriptome complexity and regulation through the lens of local splicing variations. Elife *5*, e11752.

7. Sibley, C.R., Blazquez, L., and Ule, J. (2016). Lessons from non-canonical splicing. Nat. Rev. Genet. *17*, 407–421.

8. Baralle, F.E., and Giudice, J. (2017). Alternative splicing as a regulator of development and tissue identity. Nat. Rev. Mol. Cell Biol. *18*, 437–451.

9. Ule, J., and Blencowe, B.J. (2019). Alternative Splicing Regulatory Networks: Functions, Mechanisms, and Evolution. Mol. Cell *76*, 329–345.

10. Nykamp, K., Anderson, M., Powers, M., Garcia, J., Herrera, B., Ho, Y.-Y., Kobayashi, Y., Patil, N., Thusberg, J., Westbrook, M., et al. (2017). Sherloc: a comprehensive refinement of the ACMG-AMP variant classification criteria. Genet. Med. *19*, 1105–1117.

11. Katz, Y., Wang, E.T., Airoldi, E.M., and Burge, C.B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat. Methods *7*, 1009–1015.

12. Schafer, S., Miao, K., Benson, C.C., Heinig, M., Cook, S.A., and Hubner, N. (2015). Alternative splicing signatures in RNA-seq data: Percent spliced in (PSI). Curr. Protoc. Hum. Genet. *87*, 11.16.1–11.16.14.

13. Yeo, G., and Burge, C.B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J. Comput. Biol. *11*, 377–394.

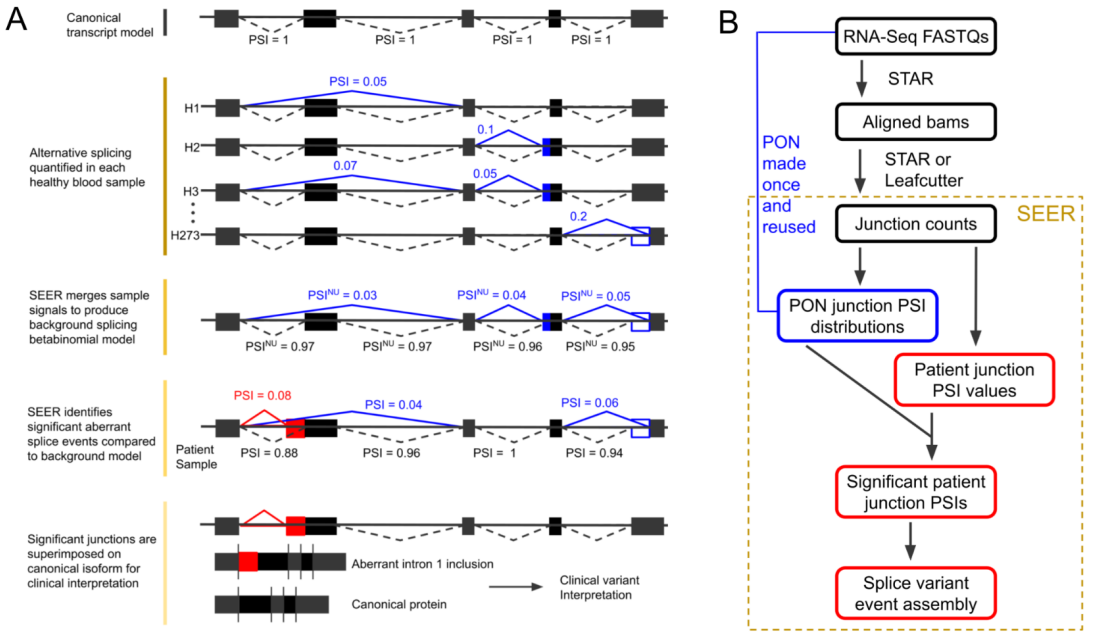
14. Shapiro, M.B., and Senapathy, P. (1987). RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. Nucleic Acids Res. *15*, 7155–7174.

15. Schröder, S., Wieland, B., Ohlenbusch, A., Yigit, G., Altmüller, J., Boltshauser, E., Dörk, T., and Brockmann, K. (2021). Evidence of Pathogenicity for the Leaky Splice Variant c.1066-6T>G in ATM in a Patient with Variant Ataxia Telangiectasia. Abstracts of the 46th Annual Meeting of the Society for Neuropediatrics.

16. Bergsma, A.J., Kroos, M., Hoogeveen-Westerveld, M., Halley, D., van der Ploeg, A.T., and

Pijnappel, W.W. (2015). Identification and characterization of aberrant GAA pre-mRNA splicing in pompe disease using a generic approach. Hum. Mutat. *36*, 57–68.

## FIGURES




C

| Validation Metrics | |
|---|---|
| True positive samples | 50 |
| True negative samples | 66 |
| Anonymous blood samples for PON | 273 |
| Accuracy | 99.7% |
| Sensitivity | 97.8% |
| Specificity | 99.7% |
| Repeatability/ Reproducibility | 98.3% |

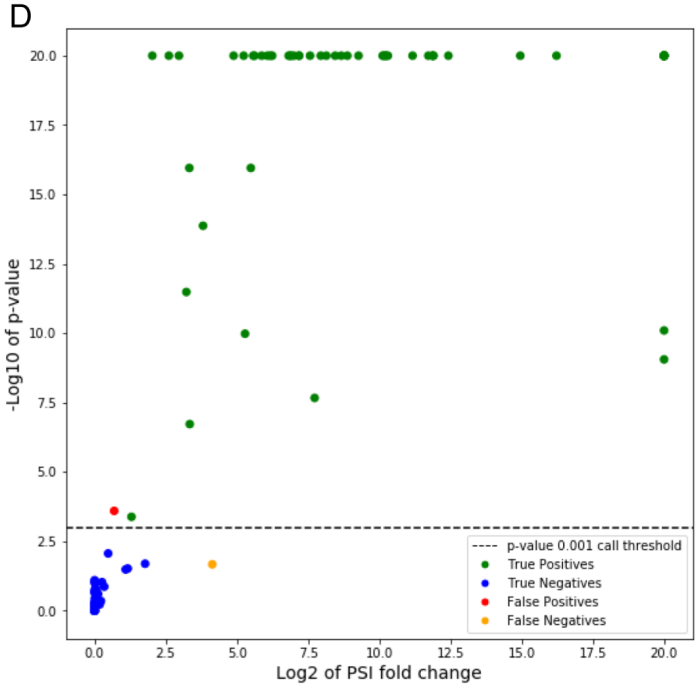



**Figure 1. SEER method and validation**. A) SEER concept. SEER quantifies the annotated and unannotated splicing variation in the healthy population by modeling the beta-binomial distribution of junction usage for all junctions observed in the healthy normal samples. Patient junction usage is then tested against this distribution to identify junctions that are aberrantly

used, filtering out common but unannotated variation. Significant splice aberrations are then mapped to the canonical transcript to infer protein effects. Canonical junctions are shown in dashed black lines, common alternatively-spliced junctions modeled by the normal reference panel are shown in solid blue lines, and the aberrant splice junction detected in the patient sample is shown with a solid red line. B) Bioinformatic workflow. FASTQ files from RNA-Seq experiments are aligned and split reads are extracted and converted to raw junction counts. SEER uses the raw junction counts from the normal samples to compute the percent spliced in (PSI) distributions for each junction in the healthy normal samples using a beta binomial count model. Beta binomial parameters for each junction are saved in a panel of normal splicing (PONS) file, which is calculated once and reused for every process-matched patient sample. Patient junction counts are converted to PSI values, which are tested against the panel of normal samples to identify significantly aberrant junction usage. Significant patient junctions are positionally linked to assemble splice variant events. C) Validation metrics. SEER was validated using a set of positive and negative control blood samples from patients with known and expected splicing events. Anonymous blood samples were used to create the PONS file. SEER showed high accuracy and reproducibility in detecting abnormal splicing events in patient samples. D) Quantitative features of abnormal splicing events used for validation. SEER p-value and PSI fold change versus PONS are plotted for known positive and negative control splice junctions. Positive and negative calls are strongly separated and inaccurate calls fall near the p-value threshold. Infinite values are set to 20 on both axes for display purposes and the top right point contains 11 samples with infinite fold change and a p-value of 0.

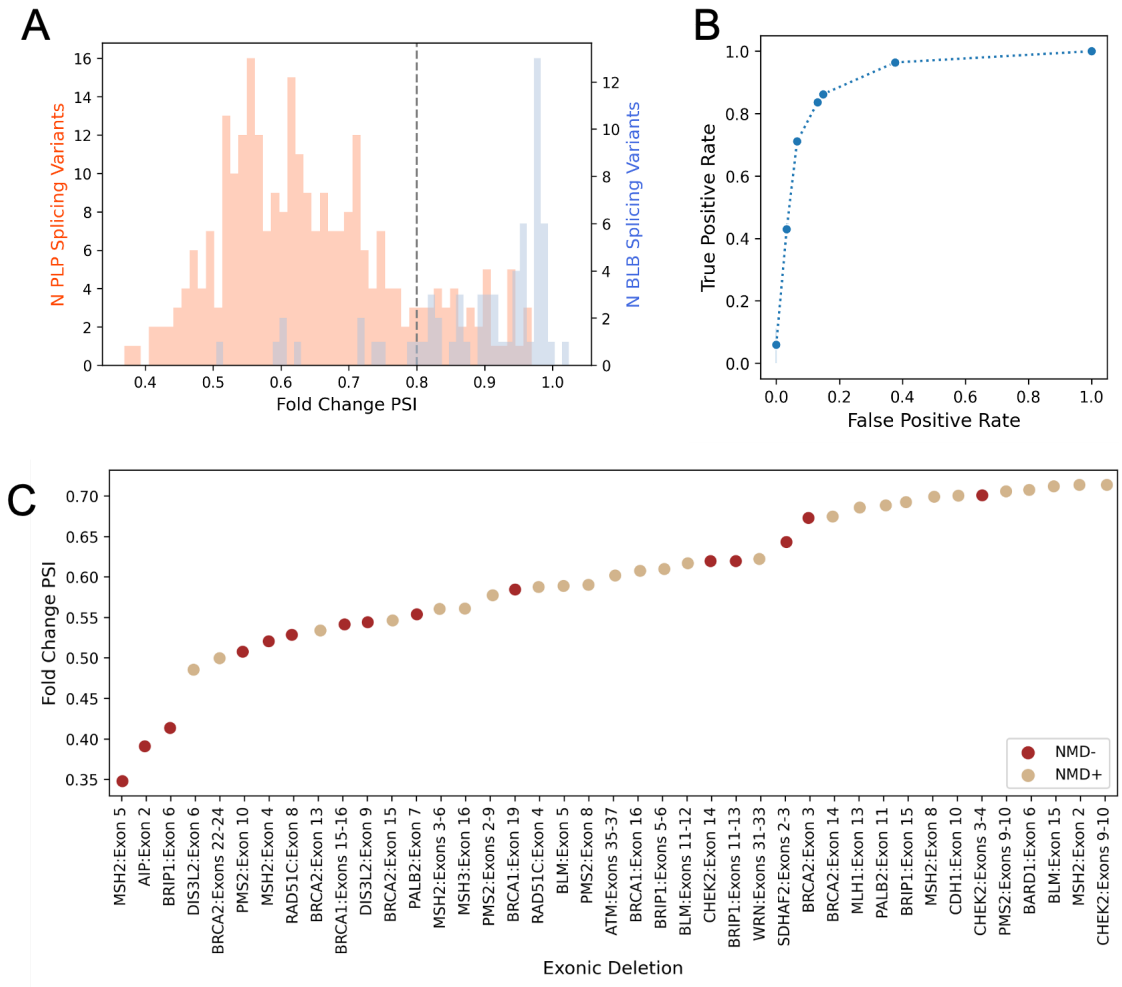



**Figure 2. Establishing a PSI-X threshold for interpreting abnormal splice events as pathogenic**. A) PSI-X (percent spliced in fold change) distribution for PSpVs (potential splicing variants) classified as pathogenic or likely pathogenic (PLP, beige) or benign or likely benign (BLB, blue). Dotted line represents PSI-X of 0.80. B) Receiver operating characteristic (ROC) curve for the data in panel A. The area under the ROC curve is 0.91. C) PSI-X values for heterozygous exon deletions detected by DNA sequencing. Although the PSI-X for deletions expected to cause NMD (nonsense mediated decay, beige circles, NMD+) is statistically different (one-sided t-test, p=0.0026) than the PSI-X for deletions expected to escape NMD (red circles, NMD-), all events are below 0.8, demonstrating the ability to correctly identify them independently of NMD status.

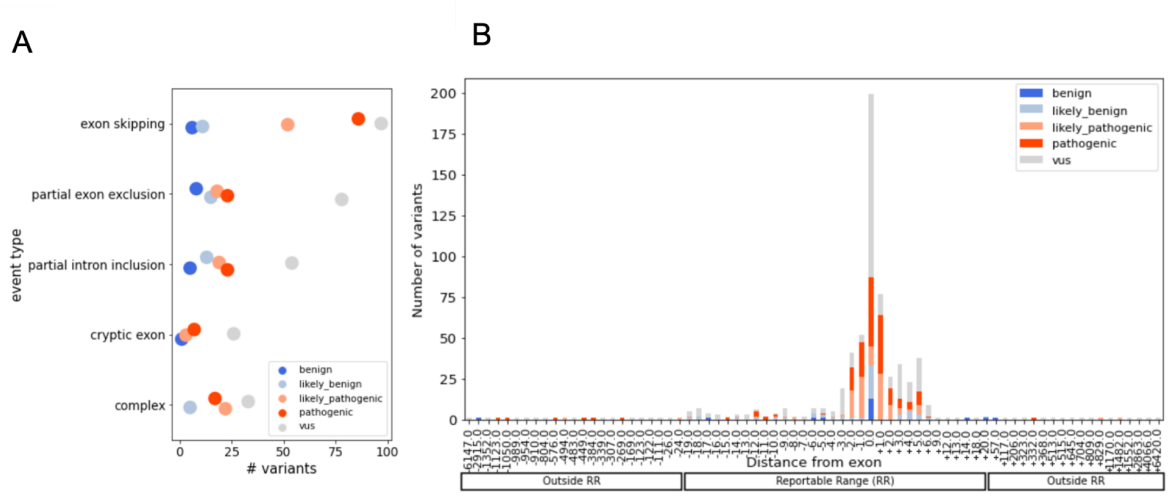



**Figure 3. Characterization of splicing events observed in the prospective cohort**. A) Splicing event types plotted by the number of potential splicing variants (PSpVs) observed. Colors represent classification of PSpVs following interpretation of abnormal splicing events. Most PSpVs classified as pathogenic or likely pathogenic are associated with changes leading to exon-skipping, while PSpVs associated with partial exon exclusion, partial intron inclusion, and complex events show the largest discrepancy between definitive classifications and variants of uncertain significance. B) Position of variants associated with splicing events in regards to the transcript exon (0 is exonic). Although there is a clear enrichment of likely pathogenic and pathogenic variants at +1, +2 and +3, we do observe both benign and pathogenic variants at other positions. Overall, 93% of PSpVs associated with abnormal splicing events are within 20 bps of the exon, corresponding with the reportable range of the NGS DNA test.

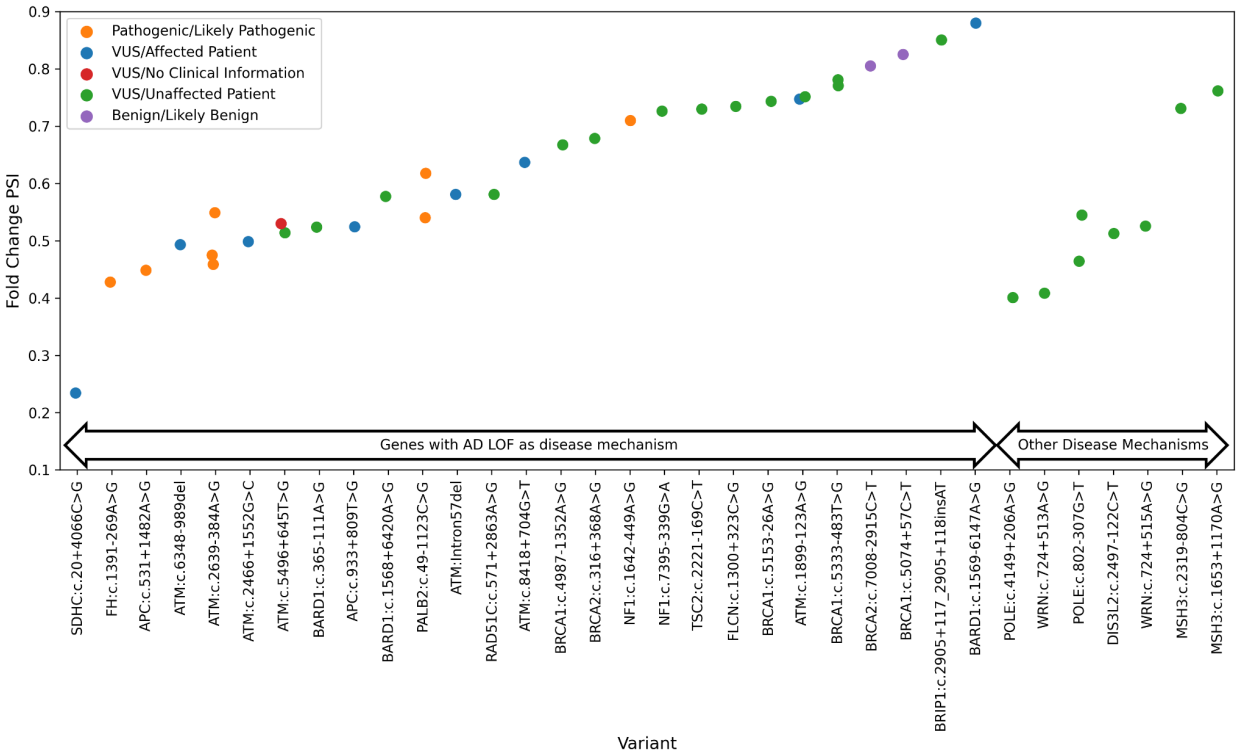



**Figure 4. Discovery of deep intronic variants with SEER** Thirty-five deep intronic DNA variants identified in this study. Five DNA variants were classified as pathogenic or likely pathogenic (orange), two DNA variants were classified as benign or likely benign (purple), and the remaining were classified as variants of uncertain significance (VUS). The affected status of individuals with VUS is represented by color: green = unaffected, blue = affected, and red = unknown due to lack of clinical information provided. The canonical PSI fold change (PSI-X) showed a large distribution for these abnormal splicing events, ranging from 0.23 to 0.9.

**TABLES**

**Table 1: Results from Phase 2**

| | Unique Individuals | % of All Samples |
|---|---|---|
| DNA + RNA sequencing | 20,317 | 100 |
| With a Potential Splicing Variant (PSpV) | 3,563 | 17.5 |
| PSpV & Normal Splicing | 2,592 | 12.8 |
| PSpV & Abnormal Splicing | 971 | 4.8 |
| PSpV upgrade VUS to PLP | 75 | 0.4 |
| PSpV downgrade VUS to BLB | 1,204 | 5.9 |
| Variants outside the reportable range | 42 | 0.2 |

**Table 2: Classification of PSpVs in the Phase 2 study**

| | PLP | VUS | BLB | Total PSpVs |
|---|---|---|---|---|
| PSpV & Abnormal splicing | 260 (42.9%) | 275 (45.5%) | 70 (11.6%) | 605 (100%) |
| PSpV & Normal splicing | 120 (9.4%) | 541 (42.3%) | 617 (48.3%) | 1278 (100%) |
| Total PSpVs | 380 (20.2%) | 816 (43.3%) | 687 (36.5%) | 1883 (100%) |

**Table 3: Reclassification rates among self-reported ethnic populations**

| | Unique individuals | % of ethnic population |
|---|---|---|
| Non-Hispanic White | 608 | 4.8 |
| Black/African American | 158 | 10.4* |
| Asian | 55 | 9.12* |
| Hispanic | 137 | 9.4* |
| Other** | 312 | 9.7 |

*p-value<0.05 after one-sided Fisher Exact test. **Includes Ashkenazi Jewish, Native American, Multiple and Unknown.