*Journal of Biomedical Research and Practice*
*Volume 2, Issue 2*
*Article ID: 100011*

*Research Article*

# Splice Site Variant Analyzer: Determining the Pathogenicity of Splice Site Variants

**Corinne E. Sexton[1†], Mark E. Wadsworth[1†], Justin B. Miller[1†], Michael J. Cormier[1], and Perry G. Ridge[1*]**

[1]Department of Biology, Brigham Young University, USA

***Corresponding author**: Perry G. Ridge, Department of Biology, Brigham Young University, Provo, Utah 84602, USA, E-mail: perry.ridge@byu.edu

[†]Contributed Equally to this work

**Received**: June 10, 2018; **Accepted**: July 18, 2018; **Published**: July 25, 2018

**Citation:** Sexton CE, Wadsworth ME, Miller JB, Cormier MJ, Ridge PG (2018) Splice Site Variant Analyzer: Determining the Pathogenicity of Splice Site Variants. J Biomed Res Prac 2(2): 100012.

## Abstract

We present Splice Site Variant Analyzer (SSVA) to simplify the characterization of deleterious and benign variants in or around splice sites. SSVA uses a Variant Call Format (VCF) file to query variants in humans against the Annovar database, MaxEntScan software, and the Conserved Domain Database. From Annovar, SSVA calculates the GERP score, the Exac score for each population, the allele frequency from the 1000 Genomes Project, and the likelihood score that the variant affects splicing. From MaxEntScan, SSVA calculates a splice site efficiency score based on the sequence. Finally, SSVA uses the Conserved Domain Database through rpsblast to determine if conserved domains are affected by the variant. SSVA presents each of these scores in a single output file that allows researchers to easily classify each splice site variant as pathogenic or benign. SSVA fills a void in splice site variant analysis by merging the output from several databases to provide researchers with a free and comprehensive analysis of the pathogenicity of splice site variants in a single step at runtime.

**Keywords:** splice sites, annotation, MaxEntScan, variant analysis, human genetics

## Introduction

In eukaryotes, splicing allows short coding regions (exons) that are separated by noncoding regions (introns) to join and form a larger gene sequence. The spliceosome recognizes conserved splice site features at the 5' and 3' ends of the coding sequences, and facilitates splicing [1]. While these regions are conserved, different base positions, different mutations, and different alternative splice sites affect the likelihood that a mutation is deleterious to a gene [2]. Therefore, correctly interpreting the significance of splice site variants is difficult because various factors provide a partial representation of variant effect. Furthermore, synonymous mutations sometimes affect splice site regulation. For instance, research in ataxia-telan-giectasia mutated (ATM) serine/threonine kinase in patients with ataxia-telangiectasia (immunodeficiency disease) [3] and cystic fibrosis transmembrane conductance regulator (CFTR) [4] gene systems has shown that seemingly benign mis-sense mutations affect splicing at splice sites and are therefore pathogenic. Unfortunately, most current tools cannot accurately identify pathogenic variants that affect splicing regulatory elements [5].

We present the Splice Site Variant Analyzer (SSVA), which takes a Variant Call Format (VCF) file and returns a tab-separated file with scores evaluating each variant transcript. SSVA is implemented in Java 1.8 and queries several known databases to output a comprehensive annotation for splice site variants included in a VCF file. All source code, executables, and annotated examples are freely available at https://github.com/ridgelab/ssva.

## Materials and Methods

SSVA provides researchers with several scores to evaluate splice site mutations. The three main analysis tools are: Annovar databases, the MaxEntScan algorithm, and the Conserved Domain Database.

### Annovar databases

First, SSVA uses databases included in Annovar [6]. SSVA provides researchers with the option of using either GRCh37/hg19 or GRCh38/hg38 as human genome references. Since GERP scores [7] are not yet compatible with hg38 coordinates, SSVA queries four databases for hg19 coordinates, or three databases for hg38 coordinates. All databases are described below:

**gerp++gt2 (hg19 only):** A score above 2 indicates that a variant is evolutionarily conserved and potentially functional [8].

**exac03:** This score represents the allele frequency in exomes across the following populations: ALL, AFR (African), AMR (Admixed American), EAS (East Asian), FIN (Finnish), NFE (Non-Finnish European), OTH (other populations), and SAS (South Asian) populations [9].

**1000g2015aug:** The allele frequency of whole-genome variants across the following populations: ALL, AFR (African), AMR (Admixed American), EAS (East Asian), EUR (European), and SAS (South Asian) [10].

**dbscsnv11:** A likelihood score that a variant affects splicing. This score is calculated using AdaBoost and Random Forest algorithms. A score of greater than 0.6 is considered splice altering [11].

### MaxEntScan algorithm

The MaxEntScan (MES) algorithm was developed in 2004 and remains one of the best performing tests of splice site variants [12]. In fact, the dbscsnv database, which was developed in 2014, uses MES as a base model for its ensemble learning methods and Random Forest scoring [11]. MES was also used to classify seemingly benign variants in the ATM gene as pathogenic [13]. SSVA provides the wild type MES score as well as the variant MES score for each variant. Higher MES scores are typically associated with efficient splicing. The MES percentage is then calculated as the percent change from the wild type sequence score to the variant sequence score. Although the scoring significance cut off is not explicitly named in the original publication of MES, we suggest that a percentage difference of less than negative 60% implies splice site deletion.

### Querying the conserved domain database

Finally, SSVA queries the conserved domain database (CDD) through rpsblast and provides the potential domains that would be lost if a variant were deleterious to a splice site. The CDD is curated by the National Center for Biotechnology Information (NCBI) [14] and contains 56 066 protein domain models from several different databases [15]. SSVA provides information for each transcript that comes from inferring a lost splice site at the location of the variant. Included for each query that passes the e-value standard (default 0.005) is the CDD domain identifier, the percent of the domain that would be lost if that splice site were skipped, the corresponding e-value of the rpsblast, and the annotation associated with that domain. If the user queries more than a thousand variants, this step may considerably slow SSVA runtime. To forgo rpsblast, users may set the '-p' flag to false.

## Results & Discussion

SSVA provides researchers with information about the role of splice site variants through a single-command at runtime. We tested the program by querying splice site variants in the ClinVar database [16] that are characterized as pathogenic, pathogenic or likely pathogenic, benign, and benign or likely benign. We present those results in Table 1.

**Table 1:** ClinVar variants correctly identified by SSVA.

| Classification | ClinVar Variants | MES | dbscsnv | MES and dbscsnv | MES or dbscsnv | Percent of Variants Identified (%) |
|---|---|---|---|---|---|---|
| **Pathogenic or Likely Pathogenic** | 580 | 502 | 512 | 499 | 515 | 88.8 |
| **Pathogenic** | 501 | 435 | 444 | 432 | 447 | 89.2 |
| **Pathogenic in the Splice Site** | 421 | 388 | 385 | 385 | 388 | 92.2 |
| **Pathogenic or Likely Pathogenic in the Splice Site** | 487 | 445 | 442 | 442 | 445 | 91.4 |
| **Benign in the Splice Site** | 2 | 2 | 1 | 1 | 2 | 100 |
| **Benign or Likely Benign in the Splice Site** | 2 | 2 | 1 | 1 | 2 | 100 |
| **Benign or Likely Benign** | 1685 | 1583 | 980 | 973 | 1590 | 94.4 |
| **Benign** | 1465 | 1373 | 836 | 829 | 1380 | 94.2 |

*Single nucleotide polymorphisms (SNPs) from ClinVar were filtered with the following parameters: expert panel, multiple submitters, clinical testing, and single nucleotide. The first column is the classification of the variant as either benign or pathogenic. We also separated variants based on their location within the splice site or surrounding the splice site. The second column is the number of variants from ClinVar in each classification (benign or pathogenic). Columns three through six are the number of variants that were correctly identified as either pathogenic or benign for each group classification based on MES, dbscsnv, MES and dbscsnv, and MES or dbscsnv, respectively. The seventh column is the percent accuracy of SSVA, calculated by using variants identified by MES or dbscsnv (column six) divided by the number of variants in ClinVar (column two) for each group classification.*

Our results suggest that by considering both the MES score and the dbscsnv score, SNPs are characterized more accurately as benign or pathogenic. Therefore, we propose that providing both scores in one concise program allows researchers to better predict effects of splice site variants than by using either the dbscsnv database or the MES software alone. We show that SSVA correctly identifies 88.8-100% of pathogenic or benign variants in or around splice sites. SSVA has a slight bias toward labelling variants around splice sites as benign (94.2% accuracy) versus likely pathogenic (88.8% accuracy). Within splice sites, SSVA is more accurate, with 100% accuracy for benign variants and 91.4% accurate for likely pathogenic.

Since SSVA uses various databases, we propose that as these databases become more comprehensive, SSVA will also become more comprehensive and accurate. SSVA allows researchers to quickly query many databases and evaluate scores derived from each source. Furthermore, we show that SSVA accurately evaluates splice site variants as either benign or pathogenic. SSVA provides the most comprehensive and user-friendly open source pipeline to analyze the deleterious effects of splice site variants.

## Availability

SSVA is freely available on GitHub at https://github.com/ridgelab/ssva

## Acknowledgements

The authors acknowledge Brigham Young University and the Fulton Supercomputing Lab staff for their continued support of this research. Specifically, we appreciate the contributions of Levi Morrison for his help optimizing and debugging parts of SSVA.

## Funding

The authors thank Brigham Young University for providing funding for this research, as well as providing a facility in which to work.

## Conflict of Interest

None of the authors have any conflicts of interest to be declared

## References

1. Maniatis T (1991) Mechanisms of alternative pre-mRNA splicing. Science 251: 33-34.
2. Sorek R, Ast G (2003) Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. Genome Res 13: 1631-1637.
3. Teraoka SN, Telatar M, Becker-Catania S, Liang T, Önengüt S, et al. (1999) Splicing Defects in the Ataxia-Telangiectasia Gene, ATM: Underlying Mutations and Consequences. Am J Hum Genet 64: 1617-1631.
4. Pagani F, Stuani C, Tzetis M, Kanavakis E, Efthymiadou A, et al. (2003) New type of disease causing mutations: the example of the composite exonic regulatory elements of splicing in CFTR exon 12. Hum Mol Genet 12: 1111-1120.
5. Grodecká L, Buratti E, Freiberger T (2017) Mutations of Pre-mRNA Splicing Regulatory Elements: Are Predictions Moving Forward to Clinical Diagnostics? Int J Mol Sci 18: 1668.
6. Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38: e164.
7. Cooper GM, Goode DL, Ng SB, Sidow A, Bamshad MJ, et al. (2010) Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. Nat Methods 7: 250-251.
8. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, et al. (2010) Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. PLOS Computational Biology 6: e1001025.
9. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E et al. (2016) Analysis of protein-coding genetic variation in 60,706 humans. Nature 536: 285-291.
10. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, et al. (2015) A global reference for human genetic variation. Nature 526: 68-74.
11. Jian X, Boerwinkle E, Liu X (2014) *In silico* prediction of splice-altering single nucleotide variants in the human genome. Nucleic Acids Res 42: 13534-13544.
12. Yeo G, Burge CB (2004) Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals. J Comput Biol 11: 377-394.
13. Eng L, Coutinho G, Nahas S, Yeo G, Tanouye R, et al. (2004) Nonclassical splicing mutations in the coding and noncoding regions of the ATM Gene: Maximum entropy estimates of splice junction strengths. Hum Mutat 23: 67-76.
14. Coordinators NR. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 46: D8-D13.
15. Marchler-Bauer A, Bo Y, Han L, He J, Lanczycki CJ, et al. (2017) CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. Nucleic Acids Res 45: D200-D203.
16. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, et al. (2018) ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res 46: D1062-D1067.