*Bioinformatics*, 35(21), 2019, 4405–4407
doi: 10.1093/bioinformatics/btz263
Advance Access Publication Date: 16 April 2019
Applications Note



Genome analysis

# Spliceogen: an integrative, scalable tool for the discovery of splice-altering variants

**Steven Monger[1], Michael Troup[1], Eddie Ip[1,2], Sally L. Dunwoodie[1,2,3] and Eleni Giannoulatou[1,2,*]**

[1]Victor Chang Cardiac Research Institute, Sydney, Australia, [2]St Vincent's Clinical School and [3]School of Biotechnology and Biomolecular Sciences, UNSW Sydney, Australia

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on December 24, 2018; revised on March 31, 2019; editorial decision on April 5, 2019; accepted on April 10, 2019

## Abstract

**Motivation:** *In silico* prediction tools are essential for identifying variants which create or disrupt *cis*-splicing motifs. However, there are limited options for genome-scale discovery of splice-altering variants.

**Results:** We have developed Spliceogen, a highly scalable pipeline integrating predictions from some of the individually best performing models for splice motif prediction: MaxEntScan, GeneSplicer, ESRseq and Branchpointer.

**Availability and implementation:** Spliceogen is available as a command line tool which accepts VCF/BED inputs and handles both single nucleotide variants (SNVs) and indels (https://github.com/VCCRI/Spliceogen). SNV databases with prediction scores are also available, covering all possible SNVs at all genomic positions within all Gencode-annotated multi-exon transcripts.

**Contact:** e.giannoulatou@victorchang.edu.au

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

Splicing defects occur in approximately one-third of disease-associated genetic variants (Lim *et al.*, 2011). Variants may alter splicing by directly impacting *trans*-acting splicing factors or more commonly, by creating or disrupting instances of the *cis*-acting motifs which guide splice site definition: donors, acceptors, branch-points, enhancers and silencers. These motifs, which are bound by components of the spliceosome and other splicing factors, exhibit substantial heterogeneity. Many prediction algorithms are available which provide scores that reflect the strength of a motif, or confidence that a motif will be bound *in vivo*. The American College of Medical Genetics and Genomics guidelines for the interpretation of splicing variants recommends employing multiple prediction algorithms to account for their individual strengths and biases (Richards *et al.*, 2015). Several web and graphical interfaces provide multi-algorithm consensus predictions for influencing any of the *cis* motifs of splicing, and dbscSNV (Jian *et al.*, 2014) provides a database of ensemble predictions for single nucleotide variant (SNVs) within splice sites. However, the options for comprehensive, genome-scale assessment of variant spliceogenicity are limited. We have developed Spliceogen, a highly scalable tool for the discovery of splice-altering variants which integrates predictions from MaxEntScan (Yeo and Burge, 2004), GeneSplicer (Pertea *et al.*, 2001), ESRseq (Ke *et al.*, 2011) and Branchpointer (Signal *et al.*, 2018).

## 2 Methods and results

### 2.1 Algorithm integration and adaptation

Spliceogen is a command line tool that accepts VCF/BED inputs and provides motif scores for both SNVs and indels calculated by multiple prediction algorithms (detailed workflow is shown in Supplementary Fig. S1). The algorithms we selected for Spliceogen cover all the major *cis* motifs which guide splicing (Fig. 1A). In order to integrate these algorithms, it was necessary to develop several modifications and extensions to the command line implementations of MaxEntScan and GeneSplicer to allow their use in batch variant analysis. First, neither MaxEntScan nor GeneSplicer outputs variant information alongside their predictions or allows for direct comparison between reference and alternative allele scores. We implemented

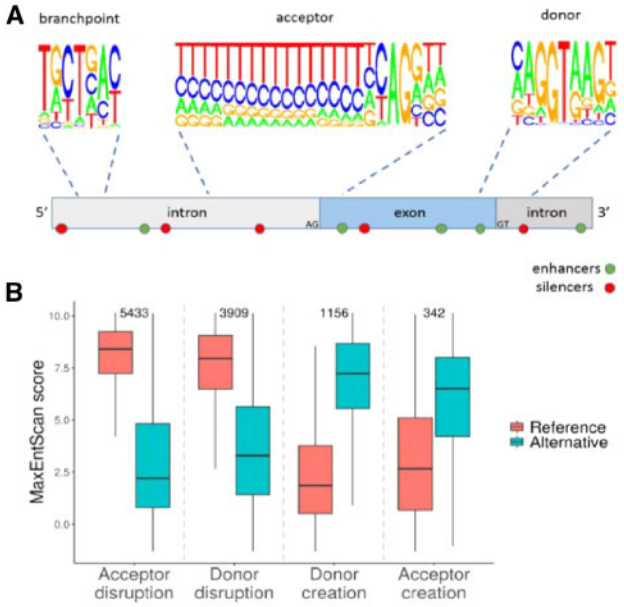


**Fig. 1.** (**A**) Splicing is guided by donor, acceptor, branchpoint, enhancer and silencer motifs. The motif logo plots were created by deriving the nucleotide frequencies for donor and acceptor motifs from 391 464 internal exon junctions from Gencode-annotated transcripts. Branchpoint motif nucleotide frequencies were derived from 8759 human branchpoints annotated as 'canonical' (Taggart *et al.*, 2017). Enhancer and silencer motifs are dispersed throughout. (**B**) MaxEntScan reference and alternative scores for donor/acceptor creating and disrupting variants. Sample sizes are indicated above. Variants which create new motifs within existing splice sites were excluded

solutions to these issues, e.g. by modifying GeneSplicer to read variant information from a FASTA header and output it alongside predictions. Second, since a variant can occur within any position of a motif, it is necessary to scan the sequence flanking a variant to identify potential motifs. MaxEntScan lacks this functionality, requiring a 9 or 23 bp input string aligned with the respective motifs a priori. We developed scanning functionality for MaxEntScan and ESRseq (detailed in Supplementary Fig. S2), which is similar to the sliding window algorithm recently implemented for a MaxEntScan Ensembl variant effect predictor plugin (Shamsani *et al.*, 2018). Third, GeneSplicer scans the input sequence as well as the reverse complement, regardless of transcript orientation. We modified GeneSplicer to read strand information and restrict scanning only to the given orientation. Comparing the predictions of the original and adapted versions of GeneSplicer in a variant analysis revealed that 26% (12/45) of the original top candidate variants were false positives arising from donor/acceptor-like sequences present on the non-coding strand. Additionally, GeneSplicer reads only one FASTA line per file, substantially limiting its scalability. We adapted GeneSplicer to handle large input files, enabling a 50-fold speed improvement (Supplementary Table S1).

### 2.2 Identification of splice-altering variants

We applied Spliceogen to a set of 14 438 reported cancer-associated splice-altering variants (Shiraishi *et al.*, 2018). We investigated the reference and alternative MaxEntScan scores separately for donor/acceptor creating and disrupting variants (Fig. 1B). In order to identify potential donor/acceptor disrupting variants, we developed an annotation-based approach for identifying all variants that overlap the extended (9 and 23 bp) donor and acceptor motifs of splice sites, based on the user-provided GTF.

To aid in variant interpretation, we provide ranked candidate variants for different modes of splice disruption. In order to refine our classification of donor/acceptor creating variants, we developed a logistic regression model (Supplementary Material) based on the MaxEntScan and GeneSplicer scores of Shiraishi *et al.* variants, using random selections of variants outside of splice sites from 1000 Genomes Project (Auton *et al.*, 2015) as a negative dataset. We achieved area under the curve values of 0.952 (donor) and 0.914 (acceptor). Variants are assigned a probability value reflecting their potential to create donor or acceptor splice sites. Supplementary Figure S3 further details our approach for identifying variants which create or disrupt acceptors, donors, branchpoints, enhancers and silencers.

### 2.3 Scalability and database

Benchmarking was performed on a single compute node with 1 CPU allocated using multiple VCF inputs containing up to 25 million variants. Predictions were generated at a rate of 2.3 million variants per compute hour, with peak memory usage <500 MB (Supplementary Table S2). Benchmarking was performed without including Branchpointer predictions, as it required no adaptation for batch variant analysis.

We used Spliceogen to assess 4.9 billion SNVs, covering all exonic and intronic genomic positions. By selecting all SNVs that either overlap an annotated splice site or receive a high logistic regression score, we provide a comprehensive database of predictions for SNVs with the potential to alter splicing via the creation or disruption of donor/acceptor motifs. In contrast, the coverage of dbscSNV is restricted to 5 million positions adjacent to splice sites.

## 3 Conclusion

Spliceogen is an integrative, all-in-one pipeline for comprehensive discovery of variants with the potential to alter splicing by creating or disrupting splicing *cis* motifs. It is available as a highly scalable command line tool as well as a genome-wide SNV database suitable for ANNOVAR annotation (Wang *et al.*, 2010).

## Funding

This work was supported by the Chain Reaction (The Ultimate Corporate Bike Challenge to S.L.D.), the Office of Health and Medical Research, NSW State Government to S.L.D., the National Health and Medical Research Council Principal Research Fellowship [1135886 to S.L.D.], the NSW Health Early-Mid Career Fellowship to E.G. and the National Heart Foundation of Australia Future Leader Fellowship [101204 to E.G.].

*Conflict of Interest*: none declared.

## References

Auton,A. *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.

Jian,X. *et al.* (2014) In silico prediction of splice-altering single nucleotide variants in the human genome. *Nucleic Acids Res.*, **42**, 13534–13544.

Ke,S. *et al.* (2011) Quantitative evaluation of all hexamers as exonic splicing elements. *Genome Res.*, **21**, 1360–1374.

Lim,K.H. *et al.* (2011) Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. *Proc. Natl. Acad. Sci. USA*, **108**, 11093–11098.

Pertea,M. *et al.* (2001) GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Res.*, **29**, 1185–1190.

Richards,S. *et al.* (2015) Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College

of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.*, **17**, 405–424.

Shamsani,J. *et al.* (2018) A plugin for the Ensembl Variant Effect Predictor that uses MaxEntScan to predict variant spliceogenicity. *Bioinformatics*, Epub ahead of print.

Shiraishi,Y. *et al.* (2018) A comprehensive characterization of *cis*-acting splicing-associated variants in human cancer. *Genome Res.*, **28**, 1111–1125.

Signal,B. *et al.* (2018) Machine learning annotation of human branchpoints. *Bioinformatics*, **34**, 920–927.

Taggart,A.J. *et al.* (2017) Large-scale analysis of branchpoint usage across species and cell lines. *Genome Res.*, **27**, 639–649.

Wang,K. *et al.* (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, **38**, e164.

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.