DEEP LEARNING-BASED ABERRANT SPLICING DETECTION

APPENDIX

[0001] The Appendix includes a bibliography of potentially relevant references listed in a paper authored by the inventors. The subject matter of the paper is covered in the US Provisionals to which this Application claims priority to/benefit of. These references can be made available by the Counsel upon request or may be accessible via Global Dossier.

PRIORITY APPLICATIONS

[0002] This application claims priority to or the benefit of US Provisional Patent Application No. 62/573,125, titled, "Deep Learning-Based Splice Site Classification," by Kishore Jaganathan, Kai-How Farh, Sofia Kyriazopoulou Panagiotopoulou and Jeremy Francis McRae, filed October 16, 2017 (Attorney Docket No. ILLM 1001-1/IP-1610-PRV); US Provisional Patent Application No. 62/573,131, titled, "Deep Learning-Based Aberrant Splicing Detection," by Kishore Jaganathan, Kai-How Farh, Sofia Kyriazopoulou Panagiotopoulou and Jeremy Francis McRae, filed October 16, 2017 (Attorney Docket No. ILLM 1001-2/IP-1614-PRV); US Provisional Patent Application No. 62/573,135, titled, "Aberrant Splicing Detection Using Convolutional Neural Networks (CNNs)," by Kishore Jaganathan, Kai-How Farh, Sofia Kyriazopoulou Panagiotopoulou and Jeremy Francis McRae, filed October 16, 2017 (Attorney Docket No. ILLM 1001-3/IP-1615-PRV); and US Provisional Patent Application No. 62/726,158, titled, "Predicting Splicing from Primary Sequence with Deep Learning," by Kishore Jaganathan, Kai-How Farh, Sofia Kyriazopoulou Panagiotopoulou and Jeremy Francis McRae, filed August 31, 2018 (Attorney Docket No. ILLM 1001-10/IP-1749-PRV). The provisional applications are hereby incorporated by reference for all purposes.

INCORPORATIONS

[0003] The following are incorporated by reference for all purposes as if fully set forth herein:

[0004] PCT Patent Application No. PCT/US18/55915, titled "Deep Learning-Based Splice Site Classification," by Kishore Jaganathan, Kai-How Farh, Sofia Kyriazopoulou Panagiotopoulou and Jeremy Francis McRae, filed on October 15, 2018 (Attorney Docket No. ILLM 1001-7/ IP-1610-PCT), subsequently published as PCT Publication No. WO 2019/079198.

[0005] PCT Patent Application No. PCT/US18/ 55923, titled "Aberrant Splicing Detection Using Convolutional Neural Networks (CNNs)," by Kishore Jaganathan, Kai-How Farh, Sofia Kyriazopoulou Panagiotopoulou and Jeremy Francis McRae, filed contemporaneously October 15, 2018 (Attorney Docket No. ILLM 1001-9/IP-1615-PCT) , subsequently published as PCT Publication No. WO 2019/079202.

[0006] US Nonprovisional Patent Application titled "Deep Learning-Based Splice Site Classification," by Kishore Jaganathan, Kai-How Farh, Sofia Kyriazopoulou Panagiotopoulou and Jeremy Francis McRae, (Attorney Docket No. ILLM 1001-4/IP-1610-US) filed contemporaneously.

[0007] US Nonprovisional Patent Application titled "Deep Learning-Based Aberrant Splicing Detection," by Kishore Jaganathan, Kai-How Farh, Sofia Kyriazopoulou Panagiotopoulou and Jeremy Francis McRae, (Attorney Docket No. ILLM 1001-5/IP-1614-US) filed contemporaneously.

[0008] US Nonprovisional Patent Application titled "Aberrant Splicing Detection Using Convolutional Neural Networks (CNNs)," by Kishore Jaganathan, Kai-How Farh, Sofia Kyriazopoulou Panagiotopoulou and Jeremy Francis McRae, (Attorney Docket No. ILLM 1001-6/IP-1615-US) filed contemporaneously.

CLAIMS

1. A system for aberrant splicing determination, including numerous processors operating in parallel, coupled to memory, the memory loaded with computer instructions to determine aberrant splicing, comprising:

a trained atrous convolutional neural network, abbreviated ACNN, running on the numerous processors, that processes pre-mRNA sequences, including:

an input module that

receives a variant sequence of at least 101 target nucleotides and flanking nucleotides, comprising at least 901 nucleotides with the target nucleotide flanked by at least 400 nucleotides on each side; and

accesses a reference sequence comprising at least 901 nucleotides;

layers that perform atrous convolutions on the target nucleotides in the variant sequence and the reference sequence, and assign splice site scores for likelihood that each of the target nucleotides is a donor splice site, an acceptor splice site, or a non-splicing site; and

an output module that determines, from position-wise differences in the splice site scores of the target nucleotides in the variant sequence and the reference sequence, whether a variant in the variant sequence causes aberrant splicing at any of the target nucleotides in the variant sequence and is therefore pathogenic.

2. The system of claim 1, wherein, for at least one target nucleotide position, at least one position-wise difference in the splice site scores is above a predetermined threshold.

3. The system of claim 2, wherein a maximum difference in the splice site scores at a position among the target nucleotides is below a predetermined threshold, further configured to classify the variant sequence as not causing aberrant splicing and therefore benign.

4. The system of any of claims 2-3, wherein the predetermined threshold is determined by:

for a plurality of candidate thresholds,

processing a first set of reference and variant sequence pairs generated by benign common variants to produce a first set of aberrant splicing detections; and

processing a second set of reference and variant sequence pairs generated by pathogenic rare variants to produce a second set of aberrant splicing detections; and

selecting at least one threshold- that maximizes a count of aberrant splicing detections in the second set and minimizes a count of aberrant splicing detections in the first set.

5. The system of any of claims 2-4, wherein the reference sequence and the variant sequence each has at least 101 target nucleotides and each target nucleotide is flanked by at least 5000 nucleotides on each side.

6. The system of any of claims 2-5, wherein the splice site scores of the target nucleotides in the reference sequence are encoded in a first output of the ACNN and the splice site scores of the target nucleotides in the variant sequence are encoded in a second output of the ACNN.

7. The system of claim 6, wherein the first output is encoded as a first 101 x 3 matrix and the second output is encoded as a second 101 x 3 matrix.

8. The system of claim 7, wherein each row in the first 101 x 3 matrix uniquely represents splice site scores for a likelihood that a target nucleotide in the reference sequence is a donor splice site, an acceptor splice site, or a non-splicing site.

9. The system of claim 7 or 8, wherein each row in the second 101 x 3 matrix uniquely represents splice site scores for a likelihood that a target nucleotide in the variant sequence is a donor splice site, an acceptor splice site, or a non-splicing site.

10. The system of claim 9, wherein splice site scores in each row of the first 101 x 3 matrix and the second 101 x 3 matrix are exponentially normalized and sum to unity.

11. The system of any of claims 8-10, wherein the output module performs a row-to-row comparison of the first 101 x 3 matrix and the second 101 x 3 matrix and determines, on a row-wise basis, changes in distribution of splice site scores.

12. The system of claim 11, wherein, for at least one instance of the row-to-row comparison, the change in distribution is above a predetermined threshold, further configured to classify the variant as causing aberrant splicing and therefore pathogenic.

13. The system of any of claims 1, 2-8 and 10-12, wherein the reference sequence and the variant sequence are one-hot encoded.

14. A computer-implemented method of determining aberrant splicing in a pre-mRNA sequence, comprising:

processing a pre-mRNA sequence using a trained atrous convolutional neural network, abbreviated ACNN, including:

receiving a variant sequence of at least 101 target nucleotides and flanking nucleotides, comprising at least 901 nucleotides with the target nucleotides flanked by at least 400 nucleotides on each side and accessing a reference sequence comprising at least 901 nucleotides;

performing atrous convolutions on the target nucleotides in the variant sequence and the reference sequence and assigning splice site scores for likelihood that each of the target nucleotides is a donor splice site, an acceptor splice site, or a non-splicing site; and

determining position-wise differences in the splice site scores of the target nucleotides in the variant sequence and the reference sequence and classifying, from the position-wise differences, whether the variant sequence causes aberrant splicing at any of the target nucleotides in the variant sequence and is therefore pathogenic.

15. The method of claim 14, wherein the determining identifies positional distribution of occurrence of splicing events in a target sub-sequence.

16. The method of claim 15, wherein the splicing events include at least one of cryptic splicing, exon skipping, mutually exclusive exons, alternative donor site, alternative acceptor site, and intron retention.

17. The method of any of claims 14-16 wherein the reference sequence and the variant sequence are aligned with respect to nucleotide positions and differ by at least one variant nucleotide.

18. The method of any of claims 14-17, wherein the target nucleotides in the variant sequence and the reference sequence are each flanked by at least 1000 nucleotides on each side.

19. The method of any of claims 14-18, wherein the target nucleotides in the variant sequence and the reference sequence are each flanked by at least 5000 nucleotides on each side.

20. The method of any of claims 14-19, wherein the variant sequence includes two variants.

21. The method of any of claims 14-20, applied to per-gene enrichment analysis for a particular gene sampled from a diseased cohort of individuals with a genetic disorder, further including:

applying the trained ACNN system to identify candidate variants from the diseased cohort in the sampled particular gene that cause aberrant splicing;

identifying de novo candidate variants among the candidate variants;

determining a baseline number of mutations at the de novo candidate variants for the particular gene based on summing observed trinucleotide mutation rates of the de novo candidate variants in a population of normal individuals, and multiplying the sum of the trinucleotide mutation rates with a transmission count and a size of the diseased cohort;

and

comparing the baseline number of mutations with a count of observed variants at the de novo candidate variants in the diseased cohort, and based on the comparison, determining that the particular gene is associated with the genetic disorder.

22. The method of claim 21, further including comparing using a statistical test that produces a p-value as the output.

23. The method of any of claims 21-22, wherein the genetic disorder is autism spectrum disorder, ASD.

24. The method of any of claims 21-22, wherein the diseased cohort is a developmental delay disorder, DDD, cohort.

25. The method of any of claims 21-24, wherein at least some of the candidate variants are protein-truncating variants.

26. The method of any of claims 21-25, wherein at least some of the candidate variants are missense variants.

27. A non-transitory computer readable storage medium impressed with computer program instructions to detect aberrant splicing, the instructions, when executed on a processor, implement a method comprising:

processing a pre-mRNA sequence using a trained atrous convolutional neural network, ACNN, including:

receiving a variant sequence of at least 101 target nucleotides and flanking nucleotides, comprising at least 901 nucleotides with the target nucleotides flanked by at least 400 nucleotides on each side and accessing a reference sequence comprising at least 901 nucleotides;

performing atrous convolutions on the target nucleotides in the variant sequence and the reference sequence and assigning splice site scores for likelihood that each of the target nucleotides is a donor splice site, an acceptor splice site, or a non-splicing site;

determining position-wise differences in the splice site scores of the target nucleotides in the variant sequence and the reference sequence; and

classifying, from the nucleotide-by-nucleotide differences, whether the variant sequence causes aberrant splicing at any of the target nucleotides in the variant sequence and is therefore pathogenic.