CLAIMS

[Claim 1].A computer-implemented method to predict likelihood of splice sites in pre-mRNA genomic sequences, the method including:

training a convolutional neural network, abbreviated CNN, on training examples of pre-mRNA nucleotide sequences, including at least 50000 training examples of donor splice sites, at least 50000 training examples of acceptor splice sites, and at least 100000 training examples of non-splicing sites, wherein the training includes:

inputting one-hot encoded training examples of the nucleotide sequences, wherein each nucleotide sequence comprises at least 201 nucleotides, with at least one target nucleotide flanked by at least 100 upstream context nucleotides and at least 100 downstream context nucleotides; and

adjusting, by backward propagation, parameters of filters in the CNN to predict scores for likelihood that each target nucleotide in the nucleotide sequence is a donor splice site, an acceptor splice site, or a non-splicing site;

whereby the trained CNN is configured to accept as input a pre-mRNA nucleotide sequence of at least 201 nucleotides that is one-hot encoded and that includes at least one target nucleotide and to score the likelihood that the target nucleotide is a donor splice site, an acceptor splice site, or a non-splicing site.

[Claim 2].The computer-implemented method of claim 1, wherein the target nucleotide is flanked by 200 upstream context nucleotides and 200 downstream context nucleotides.

[Claim 3].The computer-implemented method of claim 1 or 2, further including training the CNN on at least 150000 training examples of donor splice sites, 150000 training examples of acceptor splice sites, and 1000000 training examples of non-splicing sites.

[Claim 4].The computer-implemented method of any of claims 1-3, wherein the CNN is parameterized by a number of convolution layers, a number of convolution filters, and a number of subsampling layers.

[Claim 5].The computer-implemented method of any of claims 1-4, wherein the CNN includes one or more fully-connected layers and a terminal classification layer.

[Claim 6].The computer-implemented method of any of claims 1-5, wherein the CNN comprises dimensionality altering layers that reshape spatial and feature dimensions of a preceding input.

[Claim 7].The computer-implemented method of any of claims 1-6, wherein the CNN is parameterized by a number of residual blocks, a number of skip connections, and a number of residual connections.

[Claim 8].The computer-implemented method of claim 7, wherein each residual block comprises at least one batch normalization layer, at least one rectified linear unit, ReLU, layer, at least one dimensionality altering layer, and at least one residual connection.

[Claim 9].The computer-implemented method of claim 7, wherein each residual block comprises two batch normalization layers, two rectified linear unit, ReLU, non-linearity layers, two dimensionality altering layers, and one residual connection.

[Claim 10]. The computer-implemented method of any of claims 1-9, wherein the CNN batch-wise evaluates the training examples during an epoch.

[Claim 11]. The computer-implemented method of any of claims 1-10, wherein the training examples are randomly sampled into batches, wherein each batch has a predetermined batch size.

[Claim 12]. The computer-implemented method of claim 10, wherein the CNN iterates evaluation of the training examples over at least 10 epochs.

[Claim 13]. The computer-implemented method of any of claims 1- 12, wherein the input comprises a target nucleotide sequence having two adjacent target nucleotides.

[Claim 14]. The computer-implemented method of claim 13, wherein the two adjacent target nucleotides are adenine, abbreviated A, and guanine abbreviated G.

[Claim 15]. The computer-implemented method of claim 13, wherein the two adjacent target nucleotides are guanine abbreviated G, and uracil abbreviated D.

[Claim 16]. A system for predicting likelihood of splice sites in pre-mRNA genomic sequences, including one or more processors coupled to memory, the memory loaded with computer instructions that, when executed on the processors, implement actions comprising:

training a convolutional neural network, abbreviated CNN, on training examples of pre-mRNA nucleotide sequences, including at least 50000 training examples of donor splice sites, at least

50000 training examples of acceptor splice sites, and at least 100000 training examples of non-splicing sites, wherein the training includes:

inputting one-hot encoded examples of the nucleotide sequences, wherein each nucleotide sequence comprises at least 401 nucleotides, with at least one target nucleotide flanked by at least 200 flanking nucleotides on each side, upstream and downstream; and

adjusting, by backward propagation, parameters of filters in the CNN to predict, as output, triplet scores for likelihood that each target nucleotide in the nucleotide sequence is a donor splice site, an acceptor splice site, or a non-splicing site;

whereby the trained CNN is configured to accept as input a pre-mRNA nucleotide sequence of at least 401 nucleotides that is one-hot encoded and that includes at least one target nucleotide, and to score the likelihood that the target nucleotide is a donor splice site, an acceptor splice site, or a non-splicing site.

[Claim 17]. The system of claim 16, wherein the CNN is trained on 150000 training examples of donor splice sites, 150000 training examples of acceptor splice sites, and 1000000 training examples of non-splicing sites.

[Claim 18]. 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 convolutional neural network, abbreviated CNN, 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 convolutions on and assign to the target nucleotides in the variant sequence and the reference sequence, 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.

[Claim 19]. The system of claim 18, wherein the output module is configured to determine that the variant in the variant sequence causes aberrant splicing when, for at least one target nucleotide position, a difference in the splice site scores is above a predetermined threshold.

[Claim 20]. The system of any of claims 18-19, wherein a maximum difference in the splice site scores at any 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.

[Claim 21]. The system of any of claims 18-20, 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, for use by the output module, that maximizes a count of aberrant splicing detections in the second set and minimizes a count of aberrant splicing detections in the first set.

[Claim 22]. The system of any of claims 18-21, further configured to identify variants that cause autism spectrum disorder, abbreviated ASD.

[Claim 23]. The system of any of claims 18-21, further configured to identify variants that cause developmental delay disorder, abbreviated DDD.

[Claim 24]. The system of any of claims 18-21, wherein each sequence of at least 101 successive target nucleotides is flanked by an expanded context at least 1000 nucleotides on each side.

[Claim 25]. The system of any of claims 18-24, wherein the splice site scores of the target nucleotides in the reference sequence are encoded in a first output of the CNN and the splice site

scores of the target nucleotides in the variant sequence are encoded in a second output of the CNN.

[Claim 26]. The system of claim 25, 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.

[Claim 27]. The system of claim 26, 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.

[Claim 28]. The system of claim 27, 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.

[Claim 29]. The system of any of claims 26-28, 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.

[Claim 30]. The system of any of claims 26-29, wherein the classifier 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.

[Claim 31]. The system of claim 30, 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.

[Claim 32]. The system of any of claims 18-31, wherein the reference sequence and the variant sequence are one-hot encoded.

[Claim 33]. A method of detecting aberrant splicing, the method including:

using a trained convolutional neural network, CNN, accepting as input pre-mRNA nucleotide sequences comprising at least 501 nucleotides, including at least 101 target nucleotides flanked by at least 200 nucleotides on each side;

processing a reference sequence and a variant sequence through the CNN to produce splice site scores for a likelihood that each target nucleotide in the reference sequence and in the variant sequence is a donor splice site, an acceptor splice site, or a non-splicing site, wherein the

reference sequence and the variant sequence each has at least 101 target nucleotides flanked by at least 200 nucleotides on each side, upstream and downstream; and

determining, from differences in the splice site scores of the target nucleotides in the reference sequence and in the variant sequence, whether a variant in the variant sequence causes aberrant splicing and is therefore pathogenic.

[Claim 34]. 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:

using a trained convolutional neural network, CNN, accepting as input pre-mRNA nucleotides in an input sequences comprising at least 501 nucleotides, including at least 101 target nucleotides flanked by at least 200 nucleotides on each side;

processing a reference sequence and a variant sequence through the CNN to produce splice site scores for a likelihood that each target nucleotide in the reference sequence and in the variant sequence is a donor splice site, an acceptor splice site, or a non-splicing site, wherein the reference sequence and the variant sequence each has at least 101 target nucleotides and each target nucleotide is flanked by at least 200 nucleotides on each side; and

determining, from differences in the splice site scores of the target nucleotides in the reference sequence and in the variant sequence, whether a variant in the variant sequence causes aberrant splicing and is therefore pathogenic.