CLAIMS

What is claimed is:

1. A system for identifying repeat patterns that cause sequence-specific errors in nucleotide sequencing data, comprising:

one or more processors and one or more storage devices storing instructions that, when executed on the one or more processors cause the one or more processors to implement:

an input preparation subsystem configured to:

computationally overlay repeat patterns under test on numerous nucleotide sequences and produce overlaid samples,

wherein each repeat pattern represents a particular nucleotide composition that has a particular length and appears in an overlaid sample at a particular offset position,

wherein each overlaid sample has a target position considered to be a variant nucleotide, and

wherein for each combination of the particular nucleotide composition, the particular length, and the particular offset position, a set of the overlaid samples is computationally generated;

a pre-trained variant filter subsystem configured to:

process the overlaid samples through a convolutional neural network and, based on detection of nucleotide patterns in the overlaid samples by convolution filters of the convolutional neural network, generate classification scores for likelihood that the variant nucleotide in each of the overlaid samples is a true variant or a false variant;

a repeat pattern output subsystem configured to:

output distributions of the classification scores generated by the pre-trained variant filter subsystem for each repeat factor of the respective repeat patterns, wherein each repeat pattern is a homopolymer of a single base or a copolymer of at least two bases selected from A, C, T, G and wherein the repeat factor specifies the number of repetitions of the one or more bases in the repeat pattern; and

a sequence-specific error correlation subsystem configured to:

specify, based on a threshold, a subset of the classification scores in the distributions as indicative of the false variant classifications, and

classify those repeat patterns which are associated with the subset of the classification scores that are indicative of the false variant classifications as causing the sequence-specific errors.

2. The system of claim 1, wherein the sequence-specific error correlation subsystem is further configured to:

classify lengths and offset positions of the repeat patterns which are associated with the subset of the classification scores as causing the sequence-specific errors.

3. The system of claim 1 or 2, wherein the variant nucleotide is at the target position flanked by at least 20 nucleotides on each side.

4. The system of any of claims 1-3, wherein the pre-trained variant filter subsystem is configured to process each combination of the repeat patterns overlaid on at least 100 nucleotide sequences in at least 100 overlaid samples.

5. The system of any of claims 1-4, wherein the repeat patterns comprise the at least one base from four bases (A, C, G, and T) with at least 6 repeat factors.

6. The system of claim 5, wherein the repeat patterns are homopolymers of a single base (A, C, G, or T) with the at least 6 repeat factors.

7. The system of any of claims 1-6, wherein the repeat patterns are copolymers of at least two bases from four bases (A, C, G, and T) with the at least 6 repeat factors.

8. The system of any of claims 1-7, wherein the offset positions vary in terms of a position at which the repeat patterns are overlaid on the nucleotide sequences, measurable as an offset

between an origin position of the repeat patterns and an origin position of the nucleotide sequences, and at least ten offsets are used to produce the overlaid samples.

9. The system of any of claims 1-8, wherein the repeat patterns are right of a center nucleotide in the overlaid samples and not overlapping the center nucleotide.

10. The system of any of claims 1-8, wherein the repeat patterns are left of a center nucleotide in the overlaid samples and not overlapping the center nucleotide.

11. The system of any of claims 1-8, wherein the repeat patterns are overlaid on a center nucleotide in the overlaid samples.

12. The system of any of claims 1-11, wherein the repeat factors are integers in a range of five to one-quarter of a count of nucleotides in the overlaid samples.

13. The system of claim 6, further configured to apply to repeat patterns that are the homopolymers of the single base for each of four bases (A, C, G, and T).

14. The system of claim 13, wherein the input preparation subsystem is further configured to produce the repeat patterns and the overlaid samples for the homopolymers for each of the four bases.

15. The system of claim 14, wherein the repeat patterns are positioned right of a center nucleotide in the overlaid samples and the classification score distributions for each of the homopolymers are displayed in juxtaposition.

16. The system of claim 14, wherein the repeat patterns are positioned left of a center nucleotide in the overlaid samples and the classification score distributions for each of the homopolymers are displayed in juxtaposition.

17. The system of any of claims 1-16, wherein the nucleotide sequences on which the repeat patterns are overlaid are randomly generated.

18. The system of any of claims 1-17, wherein the nucleotide sequences on which the repeat patterns are overlaid are randomly selected from naturally occurring DNA nucleotide sequences.

19. The system of any of claims 1-18, wherein an analysis subsystem is configured to cause display of the distributions of the classification scores for each of the repeat factors.

20. The system of any of claims 1-19, wherein the pre-trained variant filter subsystem is trained on at least 500000 training examples of true variants and at least 50000 training examples of false variants; and

wherein each training example is a nucleotide sequence with a variant nucleotide at a target position flanked by at least 20 nucleotides on each side.

21. The system of any of claims 1-20, wherein the pre-trained variant filter subsystem has convolutional layers, a fully-connected layer, and a classification layer.

22. A computer-implemented method of identifying repeat patterns that cause sequence-specific errors in nucleotide sequencing data, including:

computationally overlaying repeat patterns under test on numerous nucleotide sequences and producing overlaid samples, wherein each repeat pattern represents a particular nucleotide composition that has a particular length and appears in an overlaid sample at a particular offset position, wherein each overlaid sample has a target position considered to be a variant

nucleotide, and wherein for each combination of the particular nucleotide composition, the particular length, and the particular offset position, a set of the overlaid samples is computationally generated;

processing the overlaid samples through a convolutional neural network and, based on detection of nucleotide patterns in the overlaid samples by convolution filters of the convolutional neural network, generating classification scores for likelihood that the variant nucleotide in each of the overlaid samples is a true variant or a false variant;

outputting distributions of the classification scores generated by the pre-trained variant filter subsystem for each repeat factor of the respective repeat patterns, wherein each repeat pattern is a homopolymer of a single base or a copolymer of at least two bases selected from A, C, T, G and wherein the repeat factor specifies the number of repetitions of one or more bases in the repeat pattern; and

specifying, based on a threshold, a subset of the classification scores in the distributions as indicative of the false variant classifications and classifying those repeat patterns which are associated with the subset of the classification scores that are indicative of the false variant classifications as causing the sequence-specific errors.

23. The computer-implemented method of claim 22, when applied to the system according to any one of claims 3, 5-13 or 15-18.

24. A non-transitory computer readable storage medium impressed with computer program instructions to identify repeat patterns that cause sequence-specific errors in nucleotide sequencing data, the instructions, when executed on a processor, implement a computer-implemented method comprising:

computationally overlaying repeat patterns under test on numerous nucleotide sequences and producing overlaid samples, wherein each repeat pattern represents a particular nucleotide composition that has a particular length and appears in an overlaid sample at a particular offset position, wherein each overlaid sample has a target position considered to be a variant nucleotide, and wherein for each combination of the particular nucleotide composition, the particular length, and the particular offset position, a set of the overlaid samples is computationally generated;

processing the overlaid samples through a convolutional neural network and, based on detection of nucleotide patterns in the overlaid samples by convolution filters of the convolutional neural network, generating classification scores for likelihood that the variant nucleotide in each of the overlaid samples is a true variant or a false variant;

outputting distributions of the classification scores generated by the pre-trained variant filter subsystem for each repeat factor of the respective t repeat patterns, wherein each repeat pattern is a homopolymer of a single base or a copolymer of at least two bases selected from A, C, T, G and wherein the repeat factor specifies the number of repetitions of one or more bases in the repeat pattern; and

specifying, based on a threshold, a subset of the classification scores in the distributions as indicative of the false variant classifications and classifying those repeat patterns which are associated with the subset of the classification scores that are indicative of the false variant classifications as causing the sequence-specific errors.

25. The non-transitory computer readable storage medium of claim 24, when applied to the system according to any one of claims 3, 5-13 or 15-18.