

CLAIMS

What is claimed is:

1. A neural network-implemented system, comprising:
a variant classifier, running on one or more processors operating in parallel and coupled to memory, that has
a convolutional neural network having at least two convolution layers and each of the convolution layers having at least five convolution filters trained over one thousand to millions of gradient update iterations to
process an input sequence with a variant at a target position flanked by at least ten bases on each side, and
produce an intermediate convolved feature;
a metadata correlator that correlates the variant with a set of metadata features which represent
mutation characteristics of the variant,
read mapping statistics of the variant, and
occurrence frequency of the variant; and
a fully-connected neural network having at least two fully-connected layers trained over the one thousand to millions of gradient update iterations to
process a feature sequence derived from a combination of the intermediate convolved feature and the metadata features, and
output classification scores for likelihood that the variant is a somatic variant, a germline variant, or noise.
2. The neural network-implemented system of claim 1, wherein the metadata correlator is further configured to correlate the variant with an amino acid impact feature that specifies whether the variant is a nonsynonymous variant that changes a codon so as to produce a new codon which codes for a different amino acid.
3. The neural network-implemented system of any of claims 1-2, wherein the metadata correlator is further configured to correlate the variant with a variant type feature that specifies type whether the variant is a single-nucleotide polymorphism, an insertion, or a deletion.
4. The neural network-implemented system of any of claims 1-3, wherein the metadata correlator is further configured to correlate the variant with a read mapping statistic feature that specifies quality parameters of read mapping that identified the variant.
5. The neural network-implemented system of any of claims 1-4, wherein the metadata correlator is further configured to correlate the variant with a population frequency feature that specifies allele frequencies of the variant in sequenced populations.
6. The neural network-implemented system of any of claims 1-5, wherein the metadata correlator is further configured to correlate the variant with a sub-population frequency feature that specifies allele frequencies of the variant in ethnic sub-populations stratified from sequenced populations.

7. The neural network-implemented system of any of claims 1-6, wherein the metadata correlator is further configured to correlate the variant with an evolutionary conservation feature that specifies conservativeness of the target position across multiple species.
8. The neural network-implemented system of any of claims 1-7, wherein the metadata correlator is further configured to correlate the variant with a clinical significance feature that specifies the variant's clinical effect, drug sensitivity, and histocompatibility as determined from clinical tests.
9. The neural network-implemented system of any of claims 1-8, wherein the metadata correlator is further configured to correlate the variant with a functional impact feature that specifies the variant's impact on functionality of a protein resulting from an amino acid substitution caused by the variant.
10. The neural network-implemented system of any of claims 1-9, wherein the metadata correlator is further configured to correlate the variant with an ethnicity prediction feature that specifies likelihoods identifying ethnic makeup of an individual that provided a tumor sample associated with the variant.
11. The neural network-implemented system of any of claims 1-10, wherein the metadata correlator is further configured to correlate the variant with a tumor frequency feature that specifies frequency of the variant in sequenced cancerous tumors.
12. The neural network-implemented system of any of claims 1-11, wherein the metadata correlator is further configured to correlate the variant with an alternative allele feature that specifies at least one base mutated by the variant at the target position in a reference sequence.
13. The neural network-implemented system of any of claims 1-12, wherein the convolutional neural network and the fully-connected neural network of the variant classifier are trained together end-to-end on five hundred thousand training examples from a first dataset of cancer-causing mutations, followed by training on fifty thousand training examples from a second dataset of cancer-causing mutations.
14. The neural network-implemented system of any of claims 1-13, wherein the convolutional neural network and the fully-connected neural network of the variant classifier are tested together end-to-end on validation data held-out only from the second dataset.
15. The neural network-implemented system of any of claims 1-14, wherein each of the convolution layers and the fully-connected layers is followed by at least one rectified linear unit layer.
16. The neural network-implemented system of any of claims 1-15, wherein each of the convolution layers and the fully-connected layers are followed by at least one batch normalization layer.
17. The neural network-implemented system of any of claims 1-16, wherein the variant is flanked by at least 19 bases on each side.
18. The neural network-implemented system of any of claims 1-17, further configured to comprise a concatenator that derives the feature sequence by concatenating the intermediate feature with the metadata features.

19. The neural network-implemented system of any of claims 1-18, wherein the metadata features are encoded in a one-dimensional array.
20. The neural network-implemented system of any of claims 1-19, wherein the input sequence is encoded in an n-dimensional array, where $n \geq 2$.
21. The neural network-implemented system of any of claims 1-20, wherein each of the convolution layers has at least six convolution filters.
22. A neural network-implemented method of variant classification, including:
processing an input sequence through a convolutional neural network to produce an intermediate convolved feature, wherein
the convolutional neural network has at least two convolution layers and each of the convolution layers has at least five convolution filters trained over one thousand to millions of gradient update iterations, and
the input sequence has a variant at a target position flanked by at least ten bases on each side;
correlating the variant with a set of metadata features which represent
mutation characteristics of the variant,
read mapping statistics of the variant, and
occurrence frequency of the variant; and
processing a feature sequence through a fully-connected neural network to output classification scores for likelihood that the variant is a somatic variant, a germline variant, or noise, wherein
the fully-connected neural network has at least two fully-connected layers trained over the one thousand to millions of gradient update iterations, and
the feature sequence is derived from a combination of the intermediate convolved feature and the metadata features.
23. The neural network-implemented method of claim 22, implementing each of the claims which ultimately depend from claim 1.
24. A non-transitory computer readable storage medium impressed with computer program instructions to classify variants, the instructions, when executed on a processor, implement a method comprising:
processing an input sequence through a convolutional neural network to produce an intermediate convolved feature, wherein
the convolutional neural network has at least two convolution layers and each of the convolution layers has at least five convolution filters trained over one thousand to millions of gradient update iterations, and
the input sequence has a variant at a target position flanked by at least ten bases on each side;
correlating the variant with a set of metadata features which represent
mutation characteristics of the variant,
read mapping statistics of the variant, and
occurrence frequency of the variant; and

processing a feature sequence through a fully-connected neural network to output classification scores for likelihood that the variant is a somatic variant, a germline variant, or noise, wherein

the fully-connected neural network has at least two fully-connected layers trained over the one thousand to millions of gradient update iterations, and

the feature sequence is derived from a combination of the intermediate convolved feature and the metadata features.

25. The non-transitory computer readable storage medium of claim 24, implementing each of the claims which ultimately depend from claim 1.

26. A neural network-implemented system, comprising:

a variant classifier, running on one or more processors operating in parallel and coupled to memory, that has

a convolutional neural network trained to process an input sequence and produce an intermediate convolved feature, wherein

the convolutional neural network has at least two convolution layers and each of the convolution layers has at least five convolution filters trained over one thousand to millions of gradient update iterations,

the input sequence has a variant at a target position flanked by at least ten bases on each side and has a set of metadata features correlated with the variant, and

the metadata features represent (A) mutation characteristics of the variant including at least three of (i) a chromosome feature specifying a chromosome on which a DNA segment spanning the variant occurs, (ii) an alternative allele feature specifying at least one base mutated by the variant at the target position, (iii) a mutation type feature specifying whether the variant is a single-nucleotide variant, an insertion or a deletion, (iv) a functional impact feature specifying an impact of the variant on functionality of a protein resulting from an amino acid substitution caused by the variant and (vi) a clinical significant feature specifying a clinical effect, a drug sensitivity and a histocompatibility of the variant, (B) read mapping statistics of the variant including at least two of (i) an allele frequency of the variant, (ii) a read depth of the variant, (iii) a base call quality score of the variant, (iv) a read number of the variant, and (C) occurrence frequency of the variant including at least one of (i) a tumor frequency specifying a frequency of the variant in sequenced cancerous tumors and (ii) a sub-population frequency specifying allele frequencies of the variant in ethnic sub-populations stratified from sequenced populations; and

a fully-connected neural network trained to process the intermediate convolved feature and output classification scores for likelihood that the variant is a somatic variant, a germline variant, or noise, wherein

the fully-connected neural network has at least two fully-connected layers trained over the one thousand to millions of gradient update iterations.

27. The neural network-implemented system of claim 26, further configured to comprise a metadata correlator that correlates the variant with the metadata features.

28. The neural network-implemented system of any of claims 26-27, implementing each of the claims 1-17.

29. A neural network-implemented method of variant classification, including:
processing an input sequence through a convolutional neural network to produce an intermediate convolved feature, wherein

the convolutional neural network has at least two convolution layers and each of the convolution layers has at least five convolution filters trained over one thousand to millions of gradient update iterations,

the input sequence has a variant at a target position flanked by at least ten bases on each side and has a set of metadata features correlated with the variant, and

- the metadata features represent (A) mutation characteristics of the variant including at least three of (i) a chromosome feature specifying a chromosome on which a DNA segment spanning the variant occurs, (ii) an alternative allele feature specifying at least one base mutated by the variant at the target position, (iii) a mutation type feature specifying whether the variant is a single-nucleotide variant, an insertion or a deletion, (iv) a functional impact feature specifying an impact of the variant on functionality of a protein resulting from an amino acid substitution caused by the variant and (vi) a clinical significant feature specifying a clinical effect, a drug sensitivity and a histocompatibility of the variant, (B) read mapping statistics of the variant including at least two of (i) an allele frequency of the variant, (ii) a read depth of the variant, (iii) a base call quality score of the variant, (iv) a read number of the variant, and (C) occurrence frequency of the variant including at least one of (i) a tumor frequency specifying a frequency of the variant in sequenced cancerous tumors and (ii) a sub-population frequency specifying allele frequencies of the variant in ethnic sub-populations stratified from sequenced populations; and processing the intermediate convolved feature through a fully-connected neural network to output classification scores for likelihood that the variant is a somatic variant, a germline variant, or noise, wherein
- the fully-connected neural network has at least two fully-connected layers trained over the one thousand to millions of gradient update iterations.

30. The neural network-implemented method of claim 29, implementing each of the claims 22-23.

31. A non-transitory computer readable storage medium impressed with computer program instructions to classify variants, the instructions, when executed on a processor, implement a method comprising:
processing an input sequence through a convolutional neural network to produce an intermediate convolved feature, wherein

the convolutional neural network has at least two convolution layers and each of the convolution layers has at least five convolution filters trained over one thousand to millions of gradient update iterations,

the input sequence has a variant at a target position flanked by at least ten bases on each side and has a set of metadata features correlated with the variant, and

the metadata features represent (A) mutation characteristics of the variant including at least three of (i) a chromosome feature specifying a chromosome on which a DNA segment spanning the variant occurs, (ii) an alternative allele feature specifying at least one base mutated by the variant at the target position, (iii) a mutation type feature specifying whether the variant is a single-nucleotide variant, an insertion or a deletion, (iv) a functional impact feature specifying an impact of the variant on functionality of a protein resulting from an amino acid substitution caused by the variant and (vi) a clinical significant

feature specifying a clinical effect, a drug sensitivity and a histocompatibility of the variant, (B) read mapping statistics of the variant including at least two of (i) an allele frequency of the variant, (ii) a read depth of the variant, (iii) a base call quality score of the variant, (iv) a read number of the variant, and (C) occurrence frequency of the variant including at least one of (i) a tumor frequency specifying a frequency of the variant in sequenced cancerous tumors and (ii) a sub-population frequency specifying allele frequencies of the variant in ethnic sub-populations stratified from sequenced populations; and processing the intermediate convolved feature through a fully-connected neural network to output classification scores for likelihood that the variant is a somatic variant, a germline variant, or noise, wherein the fully-connected neural network has at least two fully-connected layers trained over the one thousand to millions of gradient update iterations.

32. The non-transitory computer readable storage medium of claim 31, implementing the method according to one or more of the claims 22, 23, 29-30.