

CLAIMS

What is claimed is:

1. A sequence variation analysis method for human Mendelian genetic disease-related variation, comprising:

acquiring sequence variation data to be analyzed;

obtaining a first variation feature set and a second variation feature set by performing feature extraction on the sequence variation data to be analyzed;

inputting the first variation feature set into a trained first phenotype relationship prediction model, to obtain a first phenotype relationship prediction result, and inputting the second variation feature set into a trained second phenotype relationship prediction model, to obtain a second phenotype relationship prediction result; and

obtaining a third phenotype relationship prediction result by taking a union of the first phenotype relationship prediction result and the second phenotype relationship prediction result, wherein:

the first phenotype relationship prediction model is constructed by:

A2: constructing a first data set, the first data comprising results of pathogenicity determination recorded in a public variation database and results of pathogenicity determination made by an interpretation expert;

B2: randomly dividing data in the first data set into a first training set and a first test set according to a first preset ratio;

C2: training first machine learning models using data in the first training set and evaluating the trained first machine learning models using the first test set to obtain evaluation results; and

D2: determining an optimal first machine learning model from the trained first machine learning models based on the evaluation results as the first phenotype relationship prediction model, and

the second phenotype relationship prediction model is constructed by:

A3: constructing a second data set, the second data set comprising allele frequency data of a variation site, functional prediction data, and reliable phenotype relationship determination results;

B3: randomly dividing data in the second data set into a second training set and a second test

set according to a second preset ratio;

C3: training second machine learning models using data in the second training set and evaluating the trained second machine learning models using the second test set to obtain evaluation results; and

D3: determining an optimal second machine learning model from the trained second machine learning models based on the evaluation result as the second phenotype relationship prediction model.

2. The sequence variation analysis method of claim 1, further comprising, subsequent to said obtaining the third phenotype relationship prediction result:

comparing the third phenotype relationship prediction result with a corresponding variation interpretation result;

determining, in response to that the third phenotype relationship prediction result is consistent with the variation interpretation result, the variation interpretation result to be reliable; and

determining, in response to that the third phenotype relationship prediction result is inconsistent with the variation interpretation result, the variation interpretation result to be unreliable.

3. The sequence variation analysis method of claim 1, further comprising:

filtering a first variation site in the sequence variation data to be analyzed, wherein:

the first variation site is located in a simple repeat region of a reference sequence, the first variation site has a weak correlation with a phenotypic alteration, and the first variation site is a non-single base variation.

4. The sequence variation analysis method of claim 3, wherein the weak correlation with the phenotypic alteration means that the variation is located in a non-coding region or an intron region and an allele frequency of the variation is greater than 0.05.

5. The sequence variation analysis method of claim 3, wherein the first variation site further satisfies any one of the following conditions:

condition 1: the total number of detections being greater than a first preset value, and the number of low-quality detections being greater than a second preset value; and

condition 2: located within a third preset value of upstream and downstream of a physical position of a reference sequence of a variation site satisfying condition 1.

6. The sequence variation analysis method of claim 1, further comprising:
acquiring an amino acid level annotation result of the sequence variation data to be analyzed;
and

judging, based on the amino acid level annotation result, whether a complex variation exists in the sequence variation data to be analyzed.

7. The sequence variation analysis method of claim 6, wherein the variation sites where the complex variation exists satisfy the following conditions:

at least two of variation site reference sequence coordinates resulting in an amino acid change being overlapped; or

at least two of the variation site reference sequence coordinates resulting in the amino acid change being adjacent and affecting the same encoded amino acid.

8. The sequence variation analysis method of claim 1, further comprising:

obtaining a third variation feature set by performing feature extraction on the sequence variation data to be analyzed;

inputting the third variation feature set into a trained variation type prediction model, to obtain a prediction probability of each variation type to which each variation site in the sequence variation data to be analyzed belongs; and

determining a variation type prediction result of the corresponding variation site based on the prediction probability.

9. The sequence variation analysis method of claim 8, wherein said determining the variation type prediction result of the corresponding variation site based on the prediction probability comprises:

selecting a variation type corresponding to the maximum prediction probability greater than a preset threshold as the variation type prediction result of the variation site.

10. The sequence variation analysis method of claim 9, wherein said determining the variation type prediction result of the corresponding variation site further comprises:

determining, when the variation type prediction result only comprises a reference sequence genotype, the variation site to be unreliable; and

determining, when the variation type prediction result comprises at least one non-reference sequence genotype, the variation site to be reliable.

11. The sequence variation analysis method of claim 8, wherein:

a value from the third variation feature set comprises variation support data and allele frequency; and

the variation type comprises homozygous variation, heterozygous variation, and no variation.

12. The sequence variation analysis method of claim 1, wherein:

each first variation feature value in the first variation feature set is a published phenotype relationship determination result of a variation site; and

each second variation feature value in the second variation feature set is allele frequency data of the variation site and functional prediction data.

13. The sequence variation analysis method of claim 12, wherein the functional prediction data in the second variation feature value comprises at least one of protein conservation prediction data, nucleic acid conservation prediction data, cleavage hazard prediction data, and variation site harmful degree prediction data.

14. A computer-readable storage medium, having a computer program stored thereon, wherein the computer program, when being executed by a processor, implements the sequence variation analysis method of any one of claims 1 to 13.

15. A sequence variation analysis system for human Mendelian genetic disease-related variation, comprising:

an acquisition module configured to acquire sequence variation data to be analyzed; and

a first analysis module configured to:

obtain a first variation feature set and a second variation feature set by performing feature extraction on the sequence variation data to be analyzed;

input the first variation feature set into a trained first phenotype relationship prediction model, to obtain a first phenotype relationship prediction result, and input the second variation feature set into a trained second phenotype relationship prediction model, to obtain a second phenotype relationship prediction result; and

obtain a third phenotype relationship prediction result by taking a union of the first phenotype relationship prediction result and the second phenotype relationship prediction result, wherein:

the first phenotype relationship prediction model is constructed by:

A2: constructing a first data set, the first data comprising results of pathogenicity determination recorded in a public variation database and results of pathogenicity determination made by an interpretation expert;

B2: randomly dividing data in the first data set into a first training set and a first test set according to a first preset ratio;

C2: training first machine learning models using data in the first training set and evaluating the trained first machine learning models using the first test set to obtain evaluation results; and

D2: determining an optimal first machine learning model from the trained first machine learning models based on the evaluation results as the first phenotype relationship prediction model, and

the second phenotype relationship prediction model is constructed by:

A3: constructing a second data set, the second data set comprising allele frequency data of a variation site, functional prediction data, and reliable phenotype relationship determination results;

B3: randomly dividing data in the second data set into a second training set and a second test set according to a second preset ratio;

C3: training second machine learning models using data in the second training set and evaluating the trained second machine learning models using the second test set to obtain evaluation results; and

D3: determining an optimal second machine learning model from the trained second machine learning models based on the evaluation result as the second phenotype relationship prediction model.

16. The sequence variation analysis system of claim 15, further comprising a second analysis module configured to filter a first variation site in the sequence variation data to be analyzed, wherein:

the first variation site is located in a simple repeat region of a reference sequence, the first variation site has a weak correlation with a phenotypic alteration, and the first variation site is a non-single base variation.

17. The sequence variation analysis system of claim 15, further comprising a third analysis module configured to:

acquire an amino acid level annotation result of the sequence variation data to be analyzed;

and

judge, based on the amino acid level annotation result, whether a complex variation exists in the sequence variation data to be analyzed.

18. The sequence variation analysis system of claim 15, further comprising a fourth analysis module configured to:

obtain a third variation feature set by performing feature extraction on the sequence variation data to be analyzed;

input the third variation feature set into a trained variation type prediction model, to obtain a prediction probability of each variation type to which each variation site in the sequence variation data to be analyzed belongs; and

determine a variation type prediction result of the corresponding variation site based on the prediction probability.