WHAT IS CLAIMED IS:

1. A computer-implemented method for determining an ancestral origin for different parts of a genome of a subject, comprising:

storing a trained machine learning model, the machine learning model comprising a predictor sub-model and a smoothing sub-model, the machine learning model being trained based on segments of training genomic sequences that have known ancestral origins;

receiving data representing an input genomic sequence of the subject, the input genomic sequence covering a plurality of segments including a plurality of single nucleotide polymorphisms (SNP) sites of the genome of the subject, wherein each segment comprises a sequence of SNP values at the SNP sites, each SNP value specifying a variant at the SNP site;

determining, using the predictor sub-model and based on the data, an initial ancestral origin estimate of each segment of SNP values; and

for each segment of the plurality of segments:

identifying a subset of neighboring segments that neighbor the segment in the genome;

inputting, to the smoothing sub-model, the initial ancestral origin estimates for the subset of neighboring segments; and

performing, by the smoothing sub-model, a smoothing operation over the segment and the subset of neighboring segments using the initial ancestral origin estimates to obtain a final prediction result for the ancestral origin of the segment.

2. The method of claim 1, wherein the predictor sub-model comprises one or more predictor units; and

wherein determining the initial ancestral origin estimate of each segment comprises inputting a sequence of SNP values of a different segment of the plurality of segments to the one or more predictor units to generate an initial ancestral origin estimate, the initial ancestral origin estimate comprising one of a classification output or a coordinates output;

wherein the classification output indicates an ancestral origin category, out of a plurality of candidate ancestral origin categories, of the segment input to the predictor unit; and

wherein the coordinates output includes coordinates indicative of an ancestral origin or a breed of the segment.

3. The method of claim 2, wherein the coordinates comprise geographical coordinates of a locale of the ancestral origin in a physical space.

4. The method of claims 2, wherein the coordinates comprise breed coordinates; and

wherein subjects of different breeds have different breed coordinates generated from genomic sequences of the subjects of the different breeds.

5. The method of claim 4, wherein the breed coordinates are defined in a multi-dimensional space, the multi-dimensional space being defined by dimensions obtained from a dimension reduction operation on an encoding of SNP sites; and

wherein the predictor sub-model is trained using vectors representing genomic sequences of reference subjects and reference breed coordinates obtained from performing the dimension reduction operation on the vectors.

6. The method of any one of claims 2-5, wherein determining the initial ancestral origin estimate of each segment comprises:

determining, for each of the plurality of candidate ancestral origins, a probability of the segment being classified into the candidate ancestral origin; and

selecting the candidate ancestral origin having the highest probabilities as the ancestral origin of the segment.

7. The method of any one of claims 2-6, wherein the predictor sub-model comprises a single predictor unit; and

wherein determining the initial ancestral origin estimate of each segment comprises inputting a sequence of SNP values of a different segment of the plurality of segments, as well as a segment index associated with the sequence of SNP values, to the predictor unit to generate an initial ancestral origin estimate.

8. The method of any of claims 2-7, wherein:

each predictor unit comprises a neural network model;

the neural network model comprises:

an input layer including a plurality of input nodes, each input node receiving an SNP value of the segment input to the predictor unit, and each input node scaling the SNP value with weights of the input layer, and

an output layer including a plurality of output nodes, each of the output nodes receiving inputs based on sums of the scaled SNP values and generating an output based on the inputs; and

the initial ancestral origin estimate from each predictor unit is generated based on the outputs of the output nodes of the predictor unit.

9. The method of claim 8, wherein, in each predictor unit:

each output node of the output layer corresponds to one of the plurality of candidate ancestral origins; and

the output of each output node indicates the segment is classified into the corresponding candidate ancestral origin.

10. The method of any of claims 8 or 9, wherein, in each predictor unit:

each output node of the output layer outputs a component of geographical coordinates of an ancestral origin locale; and

wherein the scaling of the SNP values by the input nodes is based on a regression model.

11. The method of any of claims 8-10, wherein each output node further comprises an activation function; and

the output is generated based on applying the activation function to the inputs.

12. The method of any of claims 8-11, wherein, in each predictor unit, each of the output nodes receives a sum of the scaled SNP values from each of the plurality of input nodes as the inputs.

13. The method of any of claims 8-12, wherein, in each predictor unit:

the neural network model further comprises a hidden layer comprising a plurality of intermediate nodes;

each of the intermediate nodes receives a sum of the scaled SNP values from each of the plurality of input nodes and generates an intermediate output based on scaling the sums with weights of the hidden layer; and

each of output nodes receives the intermediate outputs from each of the intermediate nodes as inputs.

14. The method of claim 13, wherein each of candidate ancestral origins is associated with a plurality of plurality of locales;

wherein each of the intermediate nodes corresponds to a locale of the plurality of locales of the candidate ancestral origins; and

wherein the intermediate output of an intermediate node indicates whether the segment is classified into the corresponding locale.

15. The method of any of claims 1-14, wherein the smoothing sub-model comprises a convolutional neural network model; and

wherein the smoothing operation is based on performing a convolution operation between a kernel and subsets of the initial ancestral origin estimates included in a sliding window by the convolution neural network model.

16. The method of claim 15, wherein each initial ancestral origin estimate included in the sliding window is scaled by a weight indicative of a pre-determined usefulness metric of the segment from which the initial ancestral origin is generated; and

wherein the convolution operation is performed between the kernel and the scaled initial ancestral origin estimates.

17. The method of claim 16, wherein the usefulness metric of a segment is determined based on a prediction error rate associated with the sites of the genome at which the segment is located.

18. The method of any of claims 2-17, wherein:

each predictor unit of the one or more predictor units comprises a plurality of decision trees;

each decision tree receives a different subset of SNPs of the segment input to the predictor unit and generates a decision tree output based on the subset of SNPs; and

the initial ancestral origin estimate from each predictor unit is generated based on the decision tree outputs of the plurality of decision trees.

19. The method of claim 17, wherein, in each predictor unit, the decision tree outputs are combined to generate, as the initial ancestral origin estimate, geographical coordinates of an ancestral origin locale based on a regression model.

20. The method of any of claims 18 or 19, wherein, in each predictor unit, the decision tree outputs are combined to generate, for each of the plurality of candidate ancestral origins, a probability of the segment input to the predictor unit having the candidate ancestral origin; and

wherein the candidate ancestral origin being associated with the highest probabilities among the plurality of candidate ancestral origins is selected as the initial ancestral origin estimate of the predictor unit.

21. The method of any of claims 17-20, wherein the smoothing sub-model comprises a plurality of decision trees.

22. The method of any of claims 1-20, wherein different segments are non-overlapping.

23. The method of any of claims 1-22, wherein:

the machine learning model is trained in a training operation based on a combined cross-entropy loss function, the combined cross-entropy loss function being a linear combination of a first cross-entropy loss function and a second cross-entropy loss function;

the first cross-entropy loss function compares ground truth ancestral origins of segments of a training sequence of SNPs with initial ancestral origin estimates output by the predictor sub-model for the segments of the training sequence of SNPs to generate first comparison results;

weights of the predictor sub-model are updated based on the first comparison results;

the second cross-entropy loss function compares the ground truth ancestral origins of the segments of the training sequence of SNPs with final prediction results output

by the smoothing sub-model for the segments of the training sequence of SNPs to generate second comparison results; and

weights of the smoothing sub-model are updated based on the second comparison results.

24. The method of any of claims 1-21, wherein the machine learning model comprises a plurality of decision trees; and

wherein the machine learning model is trained based on a gradient tree boosting operation that fits the plurality of decision trees to a regression model that relates patterns of SNP sequences to ancestral origins.

25. The method of any of claims 1-24, further comprising: forwarding the final prediction results to an application to enable the application to determine, based on the final prediction results and the sequence of SNPs, a risk of a disease of the subject, and outputs the risk.

26. A computer product comprising a computer readable medium storing a plurality of instructions for controlling a computer system to perform an operation of any of the methods above.

27. A system comprising:

the computer product of claim 26; and

one or more processors for executing instructions stored on the computer readable medium.