Claims

Claim 1. An artificial intelligence-based method of base calling, the artificial intelligence-based method including:

training a first neural network base caller by using a first set of cluster images as training data,

wherein cluster images depict intensity emissions generated as a result of nucleotide incorporation in clusters of associated analytes on a substrate during a sequence cycle of a sequencing run and

wherein the first set of cluster images are annotated with first ground truth data that uses discrete valued labels to identify a correct base call;

evaluating a second set of cluster images as inference data by applying the trained first neural network base caller to the second set of cluster images and generating base call predictions,

wherein the base call predictions are represented by continuous valued weights that identify a predicted base call;

generating hybrid ground truth data for annotating the second set of cluster images based on a combination of a second set of discrete valued labels corresponding to the second set of cluster images and the continuous valued weights generated for the second set of cluster images by the trained first neural network base caller;

training a second neural network base caller using the second set of cluster images and the hybrid ground truth data generated for the second set of cluster images as training data;

wherein the second neural network base caller has fewer processing modules and parameters than the first neural network base caller; and

evaluating a third set of cluster images as inference data by applying the trained second neural network base caller on the third set of cluster images and generating base call predictions.

Claim 2. The artificial intelligence-based method of claim 1, wherein the discrete valued labels are one-hot encoded with a one-value for a correct base and zero-values for incorrect bases.

Claim 3. The artificial intelligence-based method of claim 2, wherein the discrete valued labels have a near-one-value for the correct base and near-zero-values for the incorrect bases.

Claim 4. The artificial intelligence-based method of claim 1, wherein the continuous valued weights are part of a probability distribution for a correct base being Adenine (A), Cytosine (C), Thymine (T), and Guanine (G).

Claim 5. The artificial intelligence-based method of claim 1, wherein one of the processing modules is neural network layers.

Claim 6. The artificial intelligence-based method of claim 5, wherein one of the parameters is interconnections between the neural network layers.

Claim 7. The artificial intelligence-based method of claim 1, wherein one of the processing modules is neural network filters.

Claim 8. The artificial intelligence-based method of claim 1, wherein one of the processing modules is neural network kernels.

Claim 9. The artificial intelligence-based method of claim 1, wherein one of the parameters is multiplication and addition operations.

Claim 10. The artificial intelligence-based method of claim 1, wherein a cluster image depicts intensity emissions of clusters, and

wherein the intensity emissions are captured during a sequencing cycle of a sequencing run.

Claim 11. The artificial intelligence-based method of claim 10, wherein the cluster image further depicts intensity emissions of background surrounding the clusters.

Claim 12. The artificial intelligence-based method of claim 1, wherein the first, second, and third sets of cluster images share one or more common cluster images.

Claim 13. The artificial intelligence-based method of claim 1, further including:

training an ensemble of the first neural network base caller by using the first set of cluster images as training data, wherein the first set of cluster images are annotated with the first ground truth data that uses the discrete valued labels to identify the correct base call, and

wherein the ensemble comprises two or more instances of the first neural network base caller;

evaluating the second set of cluster images as inference data by applying the trained first neural network base caller on the second set of cluster images and generating the base call predictions, wherein the base call predictions are represented by the continuous valued weights that identify the predicted base call;

training the second neural network base caller using the second set of cluster images as training data, wherein the second set of cluster images are annotated with the second ground truth data that identifies the correct base call based on

(i) the discrete valued labels, and

(ii) the continuous valued weights;

wherein the second neural network base caller has fewer processing modules and parameters than the ensemble of the first neural network base caller; and

evaluating the third set of cluster images as inference data by applying the trained second neural network base caller on the third set of cluster images and generating the base call predictions.

Claim 14. The artificial intelligence-based method of claim 1, further including:

implementing the trained second neural network base caller on one or more parallel processors of a sequencing instrument for real-time base calling.

Claim 15. A system for artificial intelligence-based base calling, comprising:

a neural network base caller trained on cluster images that are annotated with hybrid ground truth data that identifies correct base calls for one or more nucleic acid clusters depicted by the cluster images based on a combination of:

(i) discrete valued labels of ground truth data used to train an additional neural network base caller, and

(ii) continuous valued weights of base call predictions generated by the additional neural network base caller for the cluster images during inference;

wherein the neural network base caller has fewer processing modules and parameters than the additional neural network base caller; and

wherein the neural network base caller is configured to evaluate additional cluster images and generate, for the additional cluster images, base call predictions.

Claim 16. The system of claim 15, wherein the discrete valued labels are one-hot encoded with a one-value for a correct base and zero-values for incorrect bases.

Claim 17. The system of claim 15, wherein the continuous valued weights are part of a probability distribution for a correct base being Adenine (A), Cytosine (C), Thymine (T), and Guanine (G).

Claim 18. The system of claim 15, wherein:

one of the processing modules is neural network layers,

one of the parameters is interconnections between the neural network layers,

one of the processing modules is neural network filters,

one of the processing modules is neural network kernels, and

one of the parameters is multiplication and addition operations.