Claims

1. A system for prediction of somatic mutations based on data of a potentially tumorous sample, the system comprising:

   at least one processor(s);

   memory comprising program code executable by the processor(s) to configure the processor(s) to:

   receive nucleic acid sequencing data of the potentially tumorous sample;

   transform the nucleic acid sequencing data into an image-like representation;

   process the image-like representation by a trained first neural network to predict somatic mutations in the nucleic acid sequencing data;

   wherein the first neural network is trained using a training dataset comprising:

      nucleic acid sequencing data of a plurality of training samples including either tumor only samples or both tumor samples and non-tumor samples; and

      pseudo-labels for the plurality of training samples generated by processing the sequencing data of the plurality of training samples using an ensemble model for somatic variant prediction.

2. The system of claim 1, wherein the ensemble model comprises a plurality of models for prediction of somatic variant, and the pseudo-labels are generated based on a combination of outputs of the plurality of models.

3. The system of claim 1, wherein the trained first neural network is trained to

   predict: single nucleotide variants; or

   insertions and/or deletions in the received nucleic acid sequencing data.

4. The system of claim 1, wherein the received nucleic acid data comprises data of a plurality of candidate sites; and

the image-like representation is generated for at least a subset of the candidate sites using information comprising one or more of: raw alignment data, base quality, mapping quality, strand bias and reference base data.

5. The system of claim 1, wherein the trained first neural network processes an image-like representation of nucleic acid sequencing data of a non-tumorous sample associated with the potentially tumorous sample in combination with the image-like representation of the received nucleic acid sequencing data of the potentially tumorous sample to identify the somatic mutations.
6. The system of claim 1, wherein the processor(s) is further configured to generate a heat map in relation to the image-like representation;

   wherein the heat map indicates relative importance parts of the image-like representation in prediction of somatic mutation.
7. The system of claim 6, wherein the heat map is generated by performing Guided Backpropagation on the trained first neural network.
8. The system of claim 6 or claim 7, wherein the heat map indicates one or more variant alleles or one or more candidate sites in the nucleic acid data predicted to comprise somatic mutations.
9. The system of claim 1, wherein the processor(s) is further configured to:

   train a second neural network for prediction of somatic mutations in Formalin-Fixed Paraffin-Embedded (FFPE) sample;

   wherein the second neural network is trained in part based on application of transfer learning using the trained first neural network.

   (i) The system of claim 9, wherein application of transfer learning comprises:performing prediction of somatic mutations in nucleic acid sequencing data obtained from frozen, FFPE samples obtained from

tumor samples and corresponding non-tumor samples using the trained first neural network;

(ii) treating somatic mutation predictions obtained based on the frozen tumor samples and corresponding non-tumor samples as positive ground truth labels;

(iii) treating somatic mutation predictions for the FFPE tumor samples not overlapping with the somatic mutation predictions obtained based on the frozen tumor samples and corresponding non-tumor samples as negative ground truth labels;

(iv) training the second neural network using the positive and negative ground truth labels.

10. A method for prediction of somatic mutations based on data of a potentially tumorous sample, the method comprising:

receiving nucleic acid sequencing data of the potentially tumorous sample;

transforming the nucleic acid sequencing data into an image-like representation;

processing the image-like representation by a trained neural network to predict somatic mutations in the nucleic acid sequencing data;

wherein the first neural network is trained using a training dataset comprising:

nucleic acid sequencing data of a plurality of training samples including either tumor only samples or both tumor samples and non-tumor samples; and

pseudo-labels for the plurality of training samples generated by processing the sequencing data of the plurality of training samples using an ensemble model for somatic variant prediction.

11. The method of claim 11, wherein the ensemble model comprises a plurality of models for prediction of somatic variant, and the pseudo-labels are generated based on a combination of outputs of the plurality of models.

12. The method of claim 11, wherein the trained first neural network is trained to predict:

single nucleotide variants; or

predict insertions and/or deletions in the received nucleic acid sequencing data.

13. The method of claim 11, wherein the received nucleic acid data comprises data of a plurality of candidate sites; and

the image-like representation is generated for at least a subset of the candidate sites using information comprising one or more of: raw alignment data, base quality, mapping quality, strand bias and reference base data.

14. The method of claim 11, wherein the trained first neural network processes an image-like representation of nucleic acid sequencing data of a non-tumorous sample corresponding to the potentially tumorous sample in combination with the image-like representation of the received nucleic acid sequencing data of the potentially tumorous sample to identify the somatic mutations.

15. The method of claim 11, wherein the method further comprises generating a heat map in relation to the image-like representation;

wherein the heat map indicates relative importance parts of the image-like representation in prediction of somatic mutation.

16. The method of claim 16, wherein the heat map is generated by performing Guided Backpropagation on the trained first neural network.

17. The method of claim 16 or claim 17, wherein the heat map indicates one or more variant alleles or one or more candidate sites in the nucleic acid data predicted to comprise somatic mutations.

18. The method of claim 11, wherein the method further comprises:

training a second neural network for prediction of somatic mutations in Formalin-Fixed Paraffin-Embedded (FFPE) sample;

wherein the second neural network is trained in part based on application of transfer learning using the trained first neural network.

The method of claim 19, wherein application of transfer learning comprises:

performing prediction of somatic mutations in nucleic acid sequencing data obtained from frozen, FFPE samples obtained from tumor samples and corresponding non-tumor samples using the trained first neural network;

(i) treating somatic mutation predictions obtained based on the frozen tumor samples and corresponding non-tumor samples as positive ground truth labels;

(ii) treating somatic mutation predictions for the FFPE tumor samples not overlapping with the somatic mutation predictions obtained based on the frozen tumor samples and corresponding non-tumor samples as negative ground truth labels;

(iii) training the second neural network using the positive and negative ground truth labels.