CLAIMS

1. A computer-implemented method of progressively training a base caller, including:

initially training a base caller, and generating labelled training data using the initially trained base caller; and

iteratively further training the base caller with analyte comprising organism base sequences, and generating labelled training data using the further trained base caller, by repeating N iterations comprising:

further training the base caller for N1 iterations of the N iterations with one or more analytes by (a) culling a first organism base sequence of the one or more analytes to a first plurality of base subsequences and (b) mapping the first plurality of base subsequences to unique known subsequences of the first organism base sequence, and

further training the base caller for N2 iterations of the N iterations with one or more analytes by (a) culling a second organism base sequence of the one or more analytes to a second plurality of base subsequences comprising a higher number of bases than the first plurality of base subsequences for the first organism base sequence and (b) mapping the second plurality of base subsequences to unique known subsequences of the second organism base sequence with a lower likelihood of matching than the mapping of the first plurality of base subsequences to the unique known subsequences of the first organism base sequence,

wherein a complexity of neural network configurations loaded in the base caller monotonically increases with the N iterations by increasing one or more of a number of layers, a number of weights, or a number of parameters to account for the higher number of bases of the second plurality of base subsequences and the lower likelihood of matching the second plurality

of base subsequences to the unique known subsequences of the second organism base sequence, and

wherein the labelled training data generated during an iteration of the N iterations is used to train the base caller during an immediate subsequent iteration of the N iterations.

2. The computer-implemented method of claim 1, wherein initially training the base caller comprises:

initially training the base caller with analyte comprising one or more oligo base sequences, and generating labelled training data using the initially trained base caller.

3. The computer-implemented method of claim 1 or 2, wherein:

mapping the first plurality of base subsequences to the unique known subsequences of the first organism base sequence comprises mapping at least ten consecutive bases from the first plurality of base subsequences to at least ten consecutive bases from the unique known subsequences of the first organism base sequence; and

mapping the second plurality of base subsequences to the unique known subsequences of the second organism base sequence comprises mapping at least ten consecutive bases from the second plurality of base subsequences to at least ten consecutive bases from the unique known subsequences of the second organism base sequence with a lower likelihood of matching than the mapping of the at least ten consecutive bases from the first plurality of base subsequences to the at least ten consecutive bases from the unique known subsequences of the first organism base sequence.

4. The computer-implemented method of any of claims 1-3, wherein further training the base caller for the N1 iterations comprises, during one iteration of the N1 iterations:

populating (i) a first cluster of a plurality of clusters of a flow cell with a first base subsequence of the first plurality of base subsequences of the first organism base sequence, (ii) a second cluster of the plurality of clusters of the flow cell with a second base subsequence of the first plurality of base subsequences of the first organism base sequence, and (iii) a third cluster of the plurality of clusters of the flow cell with a third base subsequence of the first plurality of base subsequences of the first organism base sequence;

receiving (i) a first sequence signal from the first cluster indicative of the first base subsequence populated in the first cluster, (ii) a second sequence signal from the second cluster indicative of the second base subsequence populated in the second cluster, and (iii) a third sequence signal from the third cluster indicative of the third base subsequence populated in the third cluster;

generating (i) a first predicted base subsequence, based on the first sequence signal, (ii) a second predicted base subsequence, based on the second sequence signal, and (iii) a third predicted base subsequence, based on the third sequence signal;

mapping (i) the first predicted base subsequence with a first section of the first organism base sequence and (ii) the second predicted base subsequence with a second section of the first organism base sequence, while failing to map the third predicted base subsequence with any section of the first organism base sequence; and

generating labelled training data comprising (i) the first predicted base subsequence mapped to the first section of the first organism base sequence, where the first section of the first organism base sequence is ground truth for the first predicted base subsequence, and (ii) the second predicted base subsequence mapped to the second section of the first organism base sequence,

where the second section of the first organism base sequence is ground truth for the second predicted base subsequence.

5. The computer-implemented method of claim 4, wherein further training the base caller for the N1 iterations comprises, during the one iteration of the N1 iterations:

prior to generating the first, second, and third predicted base subsequences, training the base caller using labelled training data generated during initially training the base caller.

6. The computer-implemented method of claim 4 or 5, wherein:

the first predicted base subsequence has L1 number of bases; and

one or more bases of the L1 number of bases of the first predicted base subsequence does not match with corresponding bases of the first section of the first organism base sequence, due to errors in base calling predictions by the base caller.

7. The computer-implemented method of any of claims 4-6, the first predicted base subsequence has L1 number of bases, wherein the L1 number of bases of the first predicted base subsequence comprises initial L2 bases, followed by subsequent L3 bases, and wherein mapping the first predicted base subsequence with the first section of the first organism base sequence comprises:

substantially and uniquely matching the initial L2 bases of the first predicted base subsequence with consecutive L2 bases of the first organism base sequence;

identifying the first section of the first organism base sequence, such that the first section (i) includes the consecutive L2 bases as initial bases and (ii) includes L1 number of bases; and

mapping the first predicted base subsequence with the identified first section of the first organism base sequence.

8. The computer-implemented method of claim 7, further comprising:

while substantially and uniquely matching the initial L2 bases of the first predicted base subsequence, refraining from aiming to match the subsequent L3 bases of the first predicted base subsequence with any base of the first organism base sequence.

9. The computer-implemented method of claim 7 or 8, wherein the initial L2 bases of the first predicted base subsequence is substantially matched with the consecutive L2 bases of the first organism base sequence, such that at least a threshold number of bases of the initial L2 bases of the first predicted base subsequence is matched with the consecutive L2 bases of the first organism base sequence.

10. The computer-implemented method of any of claims 7-9, wherein the initial L2 bases of the first predicted base subsequence is uniquely matched with consecutive L2 bases of the first organism base sequence, such that the initial L2 bases of the first predicted base subsequence is substantially matched with only the consecutive L2 bases of the first organism base sequence, and with no other consecutive L2 bases of the first organism base sequence.

11. The computer-implemented method of any of claims 4-10, the third predicted base subsequence has L1 number of bases, and wherein failing to map the third predicted base subsequence with any section of the first organism base sequence comprises:

failing to substantially and uniquely match (i) an initial L2 bases of the L1 number of bases of the third predicted base subsequence with consecutive L2 bases of the first organism base sequence.

12. A system comprising:

at least one processor; and

a non-transitory computer readable storage medium impressed with computer program instructions that, when executed on the at least one processor, implement actions comprising:

initially training a base caller, and generating labelled training data using the initially trained base caller; and

iteratively further training the base caller with analyte comprising organism base sequences, and generating labelled training data using the further trained base caller, by repeating N iterations comprising:

further training the base caller for N1 iterations of the N iterations with one or more analytes by (a) culling a first organism base sequence of the one or more analytes to a first plurality of base subsequences and (b) mapping the first plurality of base subsequences to unique known subsequences of the first organism base sequence, and

further training the base caller for N2 iterations of the N iterations with one or more analytes by (a) culling a second organism base sequence of the one or more analytes to a second plurality of base subsequences comprising a higher number of bases than the first

plurality of base subsequences for the first organism base sequence and (b) mapping the second plurality of base subsequences to unique known subsequences of the second organism base sequence with a lower likelihood of matching than the mapping of the first plurality of base subsequences to the unique known subsequences of the first organism base sequence,

wherein a complexity of neural network configurations loaded in the base caller monotonically increases with the N iterations by increasing one or more of a number of layers, a number of weights, or a number of parameters to account for the higher number of bases of the second plurality of base subsequences and the lower likelihood of matching the second plurality of base subsequences to the unique known subsequences of the second organism base sequence, and

wherein the labelled training data generated during an iteration of the N iterations is used to train the base caller during an immediate subsequent iteration of the N iterations.

13. The system of claim 12, wherein further training the base caller for the N1 iterations comprises, during one iteration of the N1 iterations:

populating (i) a first cluster of a plurality of clusters of a flow cell with a first base subsequence of the first plurality of base subsequences of the first organism base sequence, (ii) a second cluster of the plurality of clusters of the flow cell with a second base subsequence of the first plurality of base subsequences of the first organism base sequence, and (iii) a third cluster of the plurality of clusters of the flow cell with a third base subsequence of the first plurality of base subsequences of the first organism base sequence;

receiving (i) a first sequence signal from the first cluster indicative of the first base subsequence populated in the first cluster, (ii) a second sequence signal from the second cluster indicative of the second base subsequence populated in the second cluster, and (iii) a third sequence signal from the third cluster indicative of the third base subsequence populated in the third cluster;

generating (i) a first predicted base subsequence, based on the first sequence signal, (ii) a second predicted base subsequence, based on the second sequence signal, and (iii) a third predicted base subsequence, based on the third sequence signal;

mapping (i) the first predicted base subsequence with a first section of the first organism base sequence and (ii) the second predicted base subsequence with a second section of the first organism base sequence, while failing to map the third predicted base subsequence with any section of the first organism base sequence; and

generating labelled training data comprising (i) the first predicted base subsequence mapped to the first section of the first organism base sequence, where the first section of the first organism base sequence is ground truth for the first predicted base subsequence, and (ii) the second predicted base subsequence mapped to the second section of the first organism base sequence, where the second section of the first organism base sequence is ground truth for the second predicted base subsequence.

14. The system of claim 13, wherein the one iteration of the N1 iterations is a first iteration of the N1 iterations, and wherein further training the base caller for a second iteration of the N1 iterations comprises:

training the base caller using the labelled training data generated during the first iteration of the N1 iterations;

using the base caller trained with the labelled training data generated during the first iteration of the N1 iterations, generating (i) a further first predicted base subsequence, based on the first sequence signal, (ii) a further second predicted base subsequence, based on the second sequence signal, and (iii) a further third predicted base subsequence, based on the third sequence signal;

mapping (i) the further first predicted base subsequence with the first section of the first organism base sequence, (ii) the further second predicted base subsequence with the second section of the first organism base sequence, and (iii) the further third predicted base subsequence with a third section of the first organism base sequence; and

generating further labelled training data comprising (i) the further first predicted base subsequence mapped to the first section of the first organism base sequence, where the first section of the first organism base sequence is ground truth for the further first predicted base subsequence, (ii) the further second predicted base subsequence mapped to the second section of the first organism base sequence, where the further second section of the first organism base sequence is ground truth for the further second predicted base subsequence, and (iii) the further third predicted base subsequence mapped to the third section of the first organism base sequence, where the further third section of the first organism base sequence is ground truth for the further third predicted base subsequence.

15. The system of claim 14, further comprising:

generating a first error between (i) the first predicted base subsequence generated during the first iteration of the N1 iterations and (ii) the first section of the first organism base sequence; and

generating a second error between (i) the further first predicted base subsequence generated during the second iteration of the N1 iterations and (ii) the first section of the first organism base sequence,

wherein the second error is less than the first error, as the base caller is better trained during the second iteration relative to the first iteration.

16. The system of claim 14 or 15, wherein:

a neural network configuration of the base caller is unchanged during the first iteration of the N1 iterations and the second iteration of the N1 iterations; and

the neural network configuration of the base caller is reused for multiple iterations, until a convergence condition is satisfied; or

a neural network configuration of the base caller during the first iteration of the N1 iterations is different from, and more complex than, a neural network configuration of the base caller during the second iteration of the N1 iterations.

17. The system of any of claims 12-16, wherein further training the base caller for the N1 iterations of the N iterations with the analyte comprising the first organism base sequence comprises:

for a first subset of the N1 iterations, further training the base caller with a first neural network configuration loaded in the base caller; and

for a second subset of the N1 iterations, further training the base caller with a second neural network configuration loaded in the base caller, the second neural network configuration different from the first neural network configuration.

18. The system of claim 17, wherein the second neural network configuration has a greater number of layers, weights, or parameters than the first neural network configuration.

19. A non-transitory computer readable storage medium impressed with computer program instructions to progressively train a base caller, the computer program instructions, when executed on a processor, implement a method comprising:

initially training a base caller, and generating labelled training data using the initially trained base caller; and

iteratively further training the base caller with analyte comprising organism base sequences, and generating labelled training data using the further trained base caller, by repeating N iterations comprising:

further training the base caller for N1 iterations of the N iterations with one or more analytes by (a) culling a first organism base sequence of the one or more analytes to a first plurality of base subsequences and (b) mapping the first plurality of base subsequences to unique known subsequences of the first organism base sequence, and

further training the base caller for N2 iterations of the N iterations with one or more analytes by (a) culling a second organism base sequence of the one or more analytes to a second plurality of base subsequences comprising a higher number of bases than the first plurality of base subsequences for the first organism base sequence and (b) mapping the second plurality of base subsequences to unique known subsequences of the second organism base sequence with a lower likelihood of matching than the mapping of the first plurality of base subsequences to the unique known subsequences of the first organism base sequence,

wherein a complexity of neural network configurations loaded in the base caller monotonically increases with the N iterations by increasing one or more of a number of layers, a number of weights, or a number of parameters to account for the higher number of bases of the second plurality of base subsequences and the lower likelihood of matching the second plurality of base subsequences to the unique known subsequences of the second organism base sequence, and

wherein the labelled training data generated during an iteration of the N iterations is used to train the base caller during an immediate subsequent iteration of the N iterations.

20. The non-transitory computer readable storage medium of claim 19, further impressed with computer program instructions that, when executed on the processor, implement actions comprising iteratively further training the base caller by:

for one or more iterations of the N1 iterations with analyte comprising the first organism base sequence, loading a first neural network configuration in the base caller; and

for one or more iterations of the N2 iterations with analyte comprising the second organism base sequence, loading a second neural network configuration in the base caller, the second neural network configuration different from the first neural network configuration.