

ARTIFICIAL INTELLIGENCE-BASED MANY-TO-MANY BASE CALLING

FIELD OF THE TECHNOLOGY DISCLOSED

[0001] The technology disclosed relates to artificial intelligence type computers and digital data processing systems and corresponding data processing methods and products for emulation of intelligence (*i.e.*, knowledge based systems, reasoning systems, and knowledge acquisition systems); and including systems for reasoning with uncertainty (*e.g.*, fuzzy logic systems), adaptive systems, machine learning systems, and artificial neural networks. In particular, the technology disclosed relates to using deep neural networks such as deep convolutional neural networks for analyzing data.

PRIORITY APPLICATION

[0002] This application claims priority to and benefit of U.S. Provisional Patent Application No. 62/979,414, titled “ARTIFICIAL INTELLIGENCE-BASED MANY-TO-MANY BASE CALLING,” filed 20 February 2020 (Attorney Docket No. ILLM 1016-1/IP-1858-PRV) and U.S. Patent Application No. 17/180,542, titled “ARTIFICIAL INTELLIGENCE-BASED MANY-TO-MANY BASE CALLING,” filed 19 February 2021 (Attorney Docket No. ILLM 1016-2/IP-1858-US). The priority applications are hereby incorporated by reference for all purposes as if fully set forth herein.

INCORPORATIONS

[0003] The following are incorporated by reference as if fully set forth herein:

[0004] U.S. Provisional Patent Application No. 62/979,384, titled “ARTIFICIAL INTELLIGENCE-BASED BASE CALLING OF INDEX SEQUENCES,” filed 20 February 2020 (Attorney Docket No. ILLM 1015-1/IP-1857-PRV);

[0005] U.S. Provisional Patent Application No. 62/979,385, titled “KNOWLEDGE DISTILLATION-BASED COMPRESSION OF ARTIFICIAL INTELLIGENCE-BASED BASE CALLER,” filed 20 February 2020 (Attorney Docket No. ILLM 1017-1/IP-1859-PRV);

[0006] U.S. Provisional Patent Application No. 63/072,032, titled “DETECTING AND FILTERING CLUSTERS BASED ON ARTIFICIAL INTELLIGENCE-PREDICTED BASE CALLS,” filed 28 August 2020 (Attorney Docket No. ILLM 1018-1/IP-1860-PRV);

[0007] U.S. Provisional Patent Application No. 62/979,412, titled “MULTI-CYCLE CLUSTER BASED REAL TIME ANALYSIS SYSTEM,” filed 20 February 2020 (Attorney Docket No. ILLM 1020-1/IP-1866-PRV);

- [0008] U.S. Provisional Patent Application No. 62/979,411, titled “DATA COMPRESSION FOR ARTIFICIAL INTELLIGENCE-BASED BASE CALLING,” filed 20 February 2020 (Attorney Docket No. ILLM 1029-1/IP-1964-PRV);
- [0009] U.S. Provisional Patent Application No. 62/979,399, titled “SQUEEZING LAYER FOR ARTIFICIAL INTELLIGENCE-BASED BASE CALLING,” filed 20 February 2020 (Attorney Docket No. ILLM 1030-1/IP-1982-PRV);
- [0010] U.S. Nonprovisional Patent Application No. 16/825,987, titled “TRAINING DATA GENERATION FOR ARTIFICIAL INTELLIGENCE-BASED SEQUENCING,” filed 20 March 2020 (Attorney Docket No. ILLM 1008-16/IP-1693-US);
- [0011] U.S. Nonprovisional Patent Application No. 16/825,991 titled “ARTIFICIAL INTELLIGENCE-BASED GENERATION OF SEQUENCING METADATA,” filed 20 March 2020 (Attorney Docket No. ILLM 1008-17/IP-1741-US);
- [0012] U.S. Nonprovisional Patent Application No. 16/826,126, titled “ARTIFICIAL INTELLIGENCE-BASED BASE CALLING,” filed 20 March 2020 (Attorney Docket No. ILLM 1008-18/IP-1744-US);
- [0013] U.S. Nonprovisional Patent Application No. 16/826,134, titled “ARTIFICIAL INTELLIGENCE-BASED QUALITY SCORING,” filed 20 March 2020 (Attorney Docket No. ILLM 1008-19/IP-1747-US); and
- [0014] U.S. Nonprovisional Patent Application No. 16/826,168, titled “ARTIFICIAL INTELLIGENCE-BASED SEQUENCING,” filed 21 March 2020 (Attorney Docket No. ILLM 1008-20/IP-1752-PRV-US).

BACKGROUND

[0015] The subject matter discussed in this section should not be assumed to be prior art merely as a result of its mention in this section. Similarly, a problem mentioned in this section or associated with the subject matter provided as background should not be assumed to have been previously recognized in the prior art. The subject matter in this section merely represents different approaches, which in and of themselves can also correspond to implementations of the claimed technology.

[0016] Deep neural networks are a type of artificial neural networks that use multiple nonlinear and complex transforming layers to successively model high-level features. Deep neural networks provide feedback via backpropagation which carries the difference between observed and predicted output to adjust parameters. Deep neural networks have evolved with the availability of large training datasets, the power of parallel and distributed computing, and sophisticated training algorithms. Deep neural networks have facilitated major advances in

numerous domains such as computer vision, speech recognition, and natural language processing.

[0017] Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are components of deep neural networks. Convolutional neural networks have succeeded particularly in image recognition with an architecture that comprises convolution layers, nonlinear layers, and pooling layers. Recurrent neural networks are designed to utilize sequential information of input data with cyclic connections among building blocks like perceptrons, long short-term memory units, and gated recurrent units. In addition, many other emergent deep neural networks have been proposed for limited contexts, such as deep spatio-temporal neural networks, multi-dimensional recurrent neural networks, and convolutional auto-encoders.

[0018] The goal of training deep neural networks is optimization of the weight parameters in each layer, which gradually combines simpler features into complex features so that the most suitable hierarchical representations can be learned from data. A single cycle of the optimization process is organized as follows. First, given a training dataset, the forward pass sequentially computes the output in each layer and propagates the function signals forward through the network. In the final output layer, an objective loss function measures error between the inferred outputs and the given labels. To minimize the training error, the backward pass uses the chain rule to backpropagate error signals and compute gradients with respect to all weights throughout the neural network. Finally, the weight parameters are updated using optimization algorithms based on stochastic gradient descent. Whereas batch gradient descent performs parameter updates for each complete dataset, stochastic gradient descent provides stochastic approximations by performing the updates for each small set of data examples. Several optimization algorithms stem from stochastic gradient descent. For example, the Adagrad and Adam training algorithms perform stochastic gradient descent while adaptively modifying learning rates based on update frequency and moments of the gradients for each parameter, respectively.

[0019] Another core element in the training of deep neural networks is regularization, which refers to strategies intended to avoid overfitting and thus achieve good generalization performance. For example, weight decay adds a penalty term to the objective loss function so that weight parameters converge to smaller absolute values. Dropout randomly removes hidden units from neural networks during training and can be considered an ensemble of possible subnetworks. To enhance the capabilities of dropout, a new activation function, maxout, and a variant of dropout for recurrent neural networks called rnnDrop have been proposed. Furthermore, batch normalization provides a new regularization method through normalization

of scalar features for each activation within a mini-batch and learning each mean and variance as parameters.

[0020] Given that sequenced data are multi- and high-dimensional, deep neural networks have great promise for bioinformatics research because of their broad applicability and enhanced prediction power. Convolutional neural networks have been adapted to solve sequence-based problems in genomics such as motif discovery, pathogenic variant identification, and gene expression inference. Convolutional neural networks use a weight-sharing strategy that is especially useful for studying deoxyribonucleic acid (DNA) because it can capture sequence motifs, which are short, recurring local patterns in DNA that are presumed to have significant biological functions. A hallmark of convolutional neural networks is the use of convolution filters.

[0021] Unlike traditional classification approaches that are based on elaborately-designed and manually-crafted features, convolution filters perform adaptive learning of features, analogous to a process of mapping raw input data to the informative representation of knowledge. In this sense, the convolution filters serve as a series of motif scanners, since a set of such filters is capable of recognizing relevant patterns in the input and updating themselves during the training procedure. Recurrent neural networks can capture long-range dependencies in sequential data of varying lengths, such as protein or DNA sequences.

[0022] Therefore, an opportunity arises to use a principled deep learning-based framework for template generation and base calling.

[0023] In the era of high-throughput technology, amassing the highest yield of interpretable data at the lowest cost per effort remains a significant challenge. Cluster-based methods of nucleic acid sequencing, such as those that utilize bridge amplification for cluster formation, have made a valuable contribution toward the goal of increasing the throughput of nucleic acid sequencing. These cluster-based methods rely on sequencing a dense population of nucleic acids immobilized on a solid support, and typically involve the use of image analysis software to deconvolve optical signals generated in the course of simultaneously sequencing multiple clusters situated at distinct locations on a solid support.

[0024] However, such solid-phase nucleic acid cluster-based sequencing technologies still face considerable obstacles that limit the amount of throughput that can be achieved. For example, in cluster-based sequencing methods, determining the nucleic acid sequences of two or more clusters that are physically too close to one another to be resolved spatially, or that in fact physically overlap on the solid support, can pose an obstacle. For example, current image analysis software can require valuable time and computational resources for determining from which of two overlapping clusters an optical signal has emanated. As a consequence,

compromises are inevitable for a variety of detection platforms with respect to the quantity and/or quality of nucleic acid sequence information that can be obtained.

[0025] High density nucleic acid cluster-based genomics methods extend to other areas of genome analysis as well. For example, nucleic acid cluster-based genomics can be used in sequencing applications, diagnostics and screening, gene expression analysis, epigenetic analysis, genetic analysis of polymorphisms, and the like. Each of these nucleic acid cluster-based genomics technologies, too, is limited when there is an inability to resolve data generated from closely proximate or spatially overlapping nucleic acid clusters.

[0026] Clearly there remains a need for increasing the quality and quantity of nucleic acid sequencing data that can be obtained rapidly and cost-effectively for a wide variety of uses, including for genomics (*e.g.*, for genome characterization of any and all animal, plant, microbial or other biological species or populations), pharmacogenomics, transcriptomics, diagnostics, prognostics, biomedical risk assessment, clinical and research genetics, personalized medicine, drug efficacy and drug interactions assessments, veterinary medicine, agriculture, evolutionary and biodiversity studies, aquaculture, forestry, oceanography, ecological and environmental management, and other purposes.

[0027] The technology disclosed provides neural network-based methods and systems that address these and similar needs, including increasing the level of throughput in high-throughput nucleic acid sequencing technologies, and offers other related advantages.

BRIEF DESCRIPTION OF THE DRAWINGS

[0028] In the drawings, like reference characters generally refer to like parts throughout the different views. Also, the drawings are not necessarily to scale, with an emphasis instead generally being placed upon illustrating the principles of the technology disclosed. In the following description, various implementations of the technology disclosed are described with reference to the following drawings, in which:

[0029] Figures 1A, 1B, and 1C show the disclosed many-to-many base calling.

[0030] Figures 1D and 1E illustrate different examples of the disclosed many-to-many base calling.

[0031] Figures 2, 3, and 4 show different implementations of a base call generator.

[0032] Figure 5 shows one implementation of the disclosed multi-cycle gradient backpropagation.

[0033] Figure 6 is a flowchart of one implementation of the technology disclosed.

[0034] Figure 7 illustrates technical effect and advantages of the technology disclosed.

[0035] Figures 8A and 8B depict one implementation of a sequencing system. The sequencing system comprises a configurable processor.

[0036] Figure 9 is a simplified block diagram of a system for analysis of sensor data from the sequencing system, such as base call sensor outputs.

[0037] Figure 10 is a simplified diagram showing aspects of the base calling operation, including functions of a runtime program executed by a host processor.

[0038] Figure 11 is a simplified diagram of a configuration of a configurable processor such as that of Figure 9.

[0039] Figure 12 is a computer system that can be used by the disclosed sequencing system to implement the base calling techniques disclosed herein.

DETAILED DESCRIPTION

[0040] The following discussion is presented to enable any person skilled in the art to make and use the technology disclosed and is provided in the context of a particular application and its requirements. Various modifications to the disclosed implementations will be readily apparent to those skilled in the art, and the general principles defined herein may be applied to other implementations and applications without departing from the spirit and scope of the technology disclosed. Thus, the technology disclosed is not intended to be limited to the implementations shown but is to be accorded the widest scope consistent with the principles and features disclosed herein.

Sequencing Images

[0041] Base calling is the process of determining the nucleotide composition of a sequence. Base calling involves analyzing image data, *i.e.*, sequencing images, produced during a sequencing run (or sequencing reaction) carried out by a sequencing instrument such as Illumina's iSeq, HiSeqX, HiSeq 3000, HiSeq 4000, HiSeq 2500, NovaSeq 6000, NextSeq 550, NextSeq 1000, NextSeq 2000, NextSeqDx, MiSeq, and MiSeqDx.

[0042] The following discussion outlines how the sequencing images are generated and what they depict, in accordance with one implementation.

[0043] Base calling decodes the intensity data encoded in the sequencing images into nucleotide sequences. In one implementation, the Illumina sequencing platforms employ cyclic reversible termination (CRT) chemistry for base calling. The process relies on growing nascent strands complementary to template strands with fluorescently-labeled nucleotides, while tracking the emitted signal of each newly added nucleotide. The fluorescently-labeled nucleotides have a 3' removable block that anchors a fluorophore signal of the nucleotide type.

[0044] Sequencing occurs in repetitive cycles, each comprising three steps: (a) extension of a nascent strand by adding the fluorescently-labeled nucleotide; (b) excitation of the fluorophore using one or more lasers of an optical system of the sequencing instrument and imaging through different filters of the optical system, yielding the sequencing images; and (c) cleavage of the fluorophore and removal of the 3' block in preparation for the next sequencing cycle. Incorporation and imaging cycles are repeated up to a designated number of sequencing cycles, defining the read length. Using this approach, each cycle interrogates a new position along the template strands.

[0045] The tremendous power of the Illumina sequencers stems from their ability to simultaneously execute and sense millions or even billions of clusters (also called “analytes”) undergoing CRT reactions. A cluster comprises approximately one thousand identical copies of a template strand, though clusters vary in size and shape. The clusters are grown from the template strand, prior to the sequencing run, by bridge amplification or exclusion amplification of the input library. The purpose of the amplification and cluster growth is to increase the intensity of the emitted signal since the imaging device cannot reliably sense fluorophore signal of a single strand. However, the physical distance of the strands within a cluster is small, so the imaging device perceives the cluster of strands as a single spot.

[0046] Sequencing occurs in a flow cell (or biosensor) – a small glass slide that holds the input strands. The flow cell is connected to the optical system, which comprises microscopic imaging, excitation lasers, and fluorescence filters. The flow cell comprises multiple chambers called lanes. The lanes are physically separated from each other and may contain different tagged sequencing libraries, distinguishable without sample cross contamination. In some implementations, the flow cell comprises a patterned surface. A “patterned surface” refers to an arrangement of different regions in or on an exposed layer of a solid support.

[0047] The imaging device of the sequencing instrument (*e.g.*, a solid-state imager such as a charge-coupled device (CCD) or a complementary metal–oxide–semiconductor (CMOS) sensor) takes snapshots at multiple locations along the lanes in a series of non-overlapping regions called tiles. For example, there can be sixty four or ninety six tiles per lane. A tile holds hundreds of thousands to millions of clusters.

[0048] The output of the sequencing run is the sequencing images. Sequencing images depict intensity emissions of the clusters and their surrounding background using a grid (or array) of pixelated units (*e.g.*, pixels, superpixels, subpixels). The intensity emissions are stored as intensity values of the pixelated units. The sequencing images have dimensions $w \times h$ of the grid of pixelated units, where w (width) and h (height) are any numbers ranging from 1 and 100,000 (*e.g.*, 115×115 , 200×200 , 1800×2000 , 2200×25000 , 2800×3600 , 4000×400). In some

implementations, w and h are the same. In other implementations, w and h are different. The sequencing images depict intensity emissions generated as a result of nucleotide incorporation in the nucleotide sequences during the sequencing run. The intensity emissions are from associated clusters and their surrounding background.

Neural Network-Based Base Calling

[0049] The following discussion focuses on a neural network-based base caller 102 described herein. First, the input to the neural network-based base caller 102 is described, in accordance with one implementation. Then, examples of the structure and form of the neural network-based base caller 102 are provided. Finally, the output of the neural network-based base caller 102 is described, in accordance with one implementation.

[0050] A data flow logic provides the sequencing images to the neural network-based base caller 102 for base calling. The neural network-based base caller 102 accesses the sequencing images on a patch-by-patch basis (or a tile-by-tile basis). Each of the patches is a sub-grid (or sub-array) of pixelated units in the grid of pixelated units that forms the sequencing images. The patches have dimensions $q \times r$ of the sub-grid of pixelated units, where q (width) and r (height) are any numbers ranging from 1 and 10000 (*e.g.*, 3×3 , 5×5 , 7×7 , 10×10 , 15×15 , 25×25 , 64×64 , 78×78 , 115×115). In some implementations, q and r are the same. In other implementations, q and r are different. In some implementations, the patches extracted from a sequencing image are of the same size. In other implementations, the patches are of different sizes. In some implementations, the patches can have overlapping pixelated units (*e.g.*, on the edges).

[0051] Sequencing produces m sequencing images per sequencing cycle for corresponding m image channels. That is, each of the sequencing images has one or more image (or intensity) channels (analogous to the red, green, blue (RGB) channels of a color image). In one implementation, each image channel corresponds to one of a plurality of filter wavelength bands. In another implementation, each image channel corresponds to one of a plurality of imaging events at a sequencing cycle. In yet another implementation, each image channel corresponds to a combination of illumination with a specific laser and imaging through a specific optical filter. The image patches are tiled (or accessed) from each of the m image channels for a particular sequencing cycle. In different implementations such as 4-, 2-, and 1-channel chemistries, m is 4 or 2. In other implementations, m is 1, 3, or greater than 4.

[0052] Consider, for example, that a sequencing run is implemented using two different image channels: a blue channel and a green channel. Then, at each sequencing cycle, the sequencing run produces a blue image and a green image. This way, for a series of k sequencing

cycles of the sequencing run, a sequence of k pairs of blue and green images is produced as output and stored as the sequencing images. Accordingly, a sequence of k pairs of blue and green image patches is generated for the patch-level processing by the neural network-based base caller 102.

[0053] The input image data to the neural network-based base caller 102 for a single iteration of base calling (or a single instance of forward pass or a single forward traversal) comprises data for a sliding window of multiple sequencing cycles. The sliding window can include, for example, a current sequencing cycle, one or more preceding sequencing cycles, and one or more successive sequencing cycles.

[0054] In one implementation, the input image data comprises data for three sequencing cycles, such that data for a current (time t) sequencing cycle to be base called is accompanied with (i) data for a left flanking/context/previous/preceding/prior (time $t-1$) sequencing cycle and (ii) data for a right flanking/context/next/successive/subsequent (time $t+1$) sequencing cycle.

[0055] In another implementation, the input image data comprises data for five sequencing cycles, such that data for a current (time t) sequencing cycle to be base called is accompanied with (i) data for a first left flanking/context/previous/preceding/prior (time $t-1$) sequencing cycle, (ii) data for a second left flanking/context/previous/preceding/prior (time $t-2$) sequencing cycle, (iii) data for a first right flanking/context/next/successive/subsequent (time $t+1$), and (iv) data for a second right flanking/context/next/successive/subsequent (time $t+2$) sequencing cycle.

[0056] In yet another implementation, the input image data comprises data for seven sequencing cycles, such that data for a current (time t) sequencing cycle to be base called is accompanied with (i) data for a first left flanking/context/previous/preceding/prior (time $t-1$) sequencing cycle, (ii) data for a second left flanking/context/previous/preceding/prior (time $t-2$) sequencing cycle, (iii) data for a third left flanking/context/previous/preceding/prior (time $t-3$) sequencing cycle, (iv) data for a first right flanking/context/next/successive/subsequent (time $t+1$), (v) data for a second right flanking/context/next/successive/subsequent (time $t+2$) sequencing cycle, and (vi) data for a third right flanking/context/next/successive/subsequent (time $t+3$) sequencing cycle. In other implementations, the input image data comprises data for a single sequencing cycle. In yet other implementations, the input image data comprises data for 10, 15, 20, 30, 58, 75, 92, 130, 168, 175, 209, 225, 230, 275, 318, 325, 330, 525, or 625 sequencing cycles.

[0057] The neural network-based base caller 102 processes the image patches through its convolution layers and produces an alternative representation, according to one implementation. The alternative representation is then used by an output layer (*e.g.*, a softmax layer) for generating a base call for either just the current (time t) sequencing cycle or each of the

sequencing cycles, *i.e.*, the current (time t) sequencing cycle, the first and second preceding (time $t-1$, time $t-2$) sequencing cycles, and the first and second succeeding (time $t+1$, time $t+2$) sequencing cycles. The resulting base calls form the sequencing reads.

[0058] In one implementation, the neural network-based base caller 102 outputs a base call for a single target cluster for a particular sequencing cycle. In another implementation, the neural network-based base caller 102 outputs a base call for each target cluster in a plurality of target clusters for the particular sequencing cycle. In yet another implementation, the neural network-based base caller 102 outputs a base call for each target cluster in a plurality of target clusters for each sequencing cycle in a plurality of sequencing cycles, thereby producing a base call sequence for each target cluster.

[0059] In one implementation, the neural network-based base caller 102 is a multilayer perceptron (MLP). In another implementation, the neural network-based base caller 102 is a feedforward neural network. In yet another implementation, the neural network-based base caller 102 is a fully-connected neural network. In a further implementation, the neural network-based base caller 102 is a fully convolution neural network. In yet further implementation, the neural network-based base caller 102 is a semantic segmentation neural network. In yet another further implementation, the neural network-based base caller 102 is a generative adversarial network (GAN).

[0060] In one implementation, the neural network-based base caller 102 is a convolution neural network (CNN) with a plurality of convolution layers. In another implementation, the neural network-based base caller 102 is a recurrent neural network (RNN) such as a long short-term memory network (LSTM), bi-directional LSTM (Bi-LSTM), or a gated recurrent unit (GRU). In yet another implementation, the neural network-based base caller 102 includes both a CNN and an RNN.

[0061] In yet other implementations, the neural network-based base caller 102 can use 1D convolutions, 2D convolutions, 3D convolutions, 4D convolutions, 5D convolutions, dilated or atrous convolutions, transpose convolutions, depthwise separable convolutions, pointwise convolutions, 1×1 convolutions, group convolutions, flattened convolutions, spatial and cross-channel convolutions, shuffled grouped convolutions, spatial separable convolutions, and deconvolutions. The neural network-based base caller 102 can use one or more loss functions such as logistic regression/log loss, multi-class cross-entropy/softmax loss, binary cross-entropy loss, mean-squared error loss, L1 loss, L2 loss, smooth L1 loss, and Huber loss. The neural network-based base caller 102 can use any parallelism, efficiency, and compression schemes such TFRecords, compressed encoding (*e.g.*, PNG), sharding, parallel calls for map transformation, batching, prefetching, model parallelism, data parallelism, and

synchronous/asynchronous stochastic gradient descent (SGD). The neural network-based base caller 102 can include upsampling layers, downsampling layers, recurrent connections, gates and gated memory units (like an LSTM or GRU), residual blocks, residual connections, highway connections, skip connections, peephole connections, activation functions (*e.g.*, non-linear transformation functions like rectifying linear unit (ReLU), leaky ReLU, exponential liner unit (ELU), sigmoid and hyperbolic tangent (tanh)), batch normalization layers, regularization layers, dropout, pooling layers (*e.g.*, max or average pooling), global average pooling layers, and attention mechanisms.

[0062] The neural network-based base caller 102 is trained using backpropagation-based gradient update techniques. Example gradient descent techniques that can be used for training the neural network-based base caller 102 include stochastic gradient descent, batch gradient descent, and mini-batch gradient descent. Some examples of gradient descent optimization algorithms that can be used to train the neural network-based base caller 102 are Momentum, Nesterov accelerated gradient, Adagrad, Adadelta, RMSprop, Adam, AdaMax, Nadam, and AMSGrad.

[0063] In one implementation, the neural network-based base caller 102 uses a specialized architecture to segregate processing of data for different sequencing cycles. The motivation for using the specialized architecture is described first. As discussed above, the neural network-based base caller 102 processes image patches for a current sequencing cycle, one or more preceding sequencing cycles, and one or more successive sequencing cycles. Data for additional sequencing cycles provides sequence-specific context. The neural network-based base caller 102 learns the sequence-specific context during training and base calls them. Furthermore, data for pre and post sequencing cycles provides second order contribution of pre-phasing and phasing signals to the current sequencing cycle.

[0064] However, images captured at different sequencing cycles and in different image channels are misaligned and have residual registration error with respect to each other. To account for this misalignment, the specialized architecture comprises spatial convolution layers that do not mix information between sequencing cycles and only mix information within a sequencing cycle.

[0065] Spatial convolution layers (or spatial logic) use so-called “segregated convolutions” that operationalize the segregation by independently processing data for each of a plurality of sequencing cycles through a “dedicated, non-shared” sequence of convolutions. The segregated convolutions convolve over data and resulting feature maps of only a given sequencing cycle, *i.e.*, intra-cycle, without convolving over data and resulting feature maps of any other sequencing cycle.

[0066] Consider, for example, that the input image data comprises (i) current image patch for a current (time t) sequencing cycle to be base called, (ii) previous image patch for a previous (time $t-1$) sequencing cycle, and (iii) next image patch for a next (time $t+1$) sequencing cycle. The specialized architecture then initiates three separate convolution pipelines, namely, a current convolution pipeline, a previous convolution pipeline, and a next convolution pipeline. The current data processing pipeline receives as input the current image patch for the current (time t) sequencing cycle and independently processes it through a plurality of spatial convolution layers to produce a so-called “current spatially convolved representation” as the output of a final spatial convolution layer. The previous convolution pipeline receives as input the previous image patch for the previous (time $t-1$) sequencing cycle and independently processes it through the plurality of spatial convolution layers to produce a so-called “previous spatially convolved representation” as the output of the final spatial convolution layer. The next convolution pipeline receives as input the next image patch for the next (time $t+1$) sequencing cycle and independently processes it through the plurality of spatial convolution layers to produce a so-called “next spatially convolved representation” as the output of the final spatial convolution layer.

[0067] In some implementations, the current, previous, and next convolution pipelines are executed in parallel. In some implementations, the spatial convolution layers are part of a spatial convolution network (or subnetwork) within the specialized architecture.

[0068] The neural network-based base caller 102 further comprises temporal convolution layers (or temporal logic) that mix information between sequencing cycles, *i.e.*, inter-cycles. The temporal convolution layers receive their inputs from the spatial convolution network and operate on the spatially convolved representations produced by the final spatial convolution layer for the respective data processing pipelines.

[0069] The inter-cycle operability freedom of the temporal convolution layers emanates from the fact that the misalignment property, which exists in the image data fed as input to the spatial convolution network, is purged out from the spatially convolved representations by the stack, or cascade, of segregated convolutions performed by the sequence of spatial convolution layers.

[0070] Temporal convolution layers use so-called “combinatory convolutions” that groupwise convolve over input channels in successive inputs on a sliding window basis. In one implementation, the successive inputs are successive outputs produced by a previous spatial convolution layer or a previous temporal convolution layer.

[0071] In some implementations, the temporal convolution layers are part of a temporal convolution network (or subnetwork) within the specialized architecture. The temporal convolution network receives its inputs from the spatial convolution network. In one implementation, a first temporal convolution layer of the temporal convolution network

groupwise combines the spatially convolved representations between the sequencing cycles. In another implementation, subsequent temporal convolution layers of the temporal convolution network combine successive outputs of previous temporal convolution layers. The output of the final temporal convolution layer is fed to an output layer that produces an output. The output is used to base call one or more clusters at one or more sequencing cycles.

[0072] The data flow logic provides per-cycle cluster data to the neural network-based base caller 102. The per-cycle cluster data is for a plurality of clusters and for a first subset of sequencing cycles of a sequencing run. Consider, for example, that the sequencing run has 150 sequencing cycles. The first subset of sequencing cycles can then include any subset of the 150 sequencing cycles, for example, the first 5, 10, 15, 25, 35, 40, 50, or 100 sequencing cycles of the 150-cycle sequencing run. Also, each sequencing cycle produces sequencing images that depict intensity emissions of clusters in the plurality of clusters. This way, the per-cycle cluster data for the plurality of clusters and for the first subset of sequencing cycles of the sequencing run includes sequencing images only for the first 5, 10, 15, 25, 35, 40, 50, or 100 sequencing cycles of the 150-cycle sequencing run and does not include sequencing images for the remainder sequencing cycles of the 150-cycle sequencing run.

[0073] The neural network-based base caller 102 base calls each cluster in the plurality of clusters at each sequencing cycle in the first subset of sequencing cycles. To do so, the neural network-based base caller 102 processes the per-cycle cluster data and generates intermediate representations of the per-cycle cluster data. Then, the neural network-based base caller 102 processes the intermediate representations through an output layer and produces a per-cluster, per-cycle probability quadruple for each cluster and for each sequencing cycle. Examples of the output layer include a softmax function, a log-softmax function, an ensemble output average function, a multi-layer perceptron uncertainty function, a Bayes Gaussian distribution function, and a cluster intensity function. The per-cluster, per-cycle probability quadruples are stored as the probability quadruples and referred to herein as “base-wise likelihoods” because there are four nucleotide bases A, C, T, and G.

[0074] Softmax function is a preferred function for multi-class classification. The softmax function calculates the probabilities of each target class over all possible target classes. The output range of the softmax function is between zero and one and the sum of all the probabilities is equal to one. The softmax function computes the exponential of the given input value and the sum of exponential values of all the input values. The ratio of the exponential of the input value and the sum of exponential values is the output of the softmax function, referred to herein as “exponential normalization.”

[0075] Formally, training a so-called softmax classifier is regression to a class probability, rather than a true classifier as it does not return the class but rather a confidence prediction of each class's probability. The softmax function takes a class of values and converts them to probabilities that sum to one. The softmax function squashes a n -dimensional vector of arbitrary real values to n -dimensional vector of real values within the range zero to one. Thus, using the softmax function ensures that the output is a valid, exponentially normalized probability mass function (nonnegative and summing to one).

[0076] Intuitively, the softmax function is a "soft" version of the maximum function. The term "soft" derives from the fact that the softmax function is continuous and differentiable. Instead of selecting one maximal element, it breaks the vector into parts of a whole with the maximal input element getting a proportionally larger value, and the other getting a less proportion of the value. The property of outputting a probability distribution makes the softmax function suitable for probabilistic interpretation in classification tasks.

[0077] Let us consider z as a vector of inputs to the softmax layer. The softmax layer units are the number of nodes in the softmax layer and therefore, the length of the z vector is the number of units in the softmax layer (if we have ten output units, then there are ten z elements).

[0078] For an n -dimensional vector $Z = [z_1, z_2, \dots, z_n]$, the softmax function uses exponential normalization (exp) to produce another n -dimensional vector $p(Z)$ with normalized values in the range $[0, 1]$ and that add to unity:

$$Z = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix} \text{ and, } p(Z) \rightarrow \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{bmatrix}$$

$$p_j = \frac{\exp^{z_j}}{\sum_{k=1}^n \exp^{z_k}} \forall j \in 1, 2, \dots, n$$

[0079] For example, a softmax function is applied to three classes as

$z \mapsto \text{softmax}([z; \frac{z}{10}; -2z])$. Note that the three outputs always sum to one. They thus define a discrete probability mass function.

[0080] A particular per-cluster, per-cycle probability quadruple identifies probabilities of a base incorporated in a particular cluster at a particular sequencing cycle being A, C, T, and G. When the output layer of the neural network-based base caller 102 uses a softmax function, the

probabilities in the per-cluster, per-cycle probability quadruple are exponentially normalized classification scores that sum to unity.

[0081] In one implementation, the method includes processing the convolved representation through the output layer to produce likelihoods of a base incorporated in the target analyte at the current sequencing cycle being A, C, T, and G, and classifying the base as A, C, T, or G based on the likelihoods. In one implementation, the likelihoods are exponentially normalized scores produced by a softmax layer.

[0082] In one implementation, the method includes deriving, from the output, an output pair for the target analyte that identifies a class label of a base incorporated in the target analyte at the current sequencing cycle being A, C, T, or G, and base calling the target analyte based on the class label. In one implementation, a class label of 1, 0 identifies an A base, a class label of 0, 1 identifies a C base, a class label of 1, 1 identifies a T base, and a class label of 0, 0 identifies a G base. In another implementation, a class label of 1, 1 identifies an A base, a class label of 0, 1 identifies a C base, a class label of 0.5, 0.5 identifies a T base, and a class label of 0, 0 identifies a G base. In yet another implementation, a class label of 1, 0 identifies an A base, a class label of 0, 1 identifies a C base, a class label of 0.5, 0.5 identifies a T base, and a class label of 0, 0 identifies a G base. In yet further implementation, a class label of 1, 2 identifies an A base, a class label of 0, 1 identifies a C base, a class label of 1, 1 identifies a T base, and a class label of 0, 0 identifies a G base. In one implementation, the method includes deriving, from the output, a class label for the target analyte that identifies a base incorporated in the target analyte at the current sequencing cycle being A, C, T, or G, and base calling the target analyte based on the class label. In one implementation, a class label of 0.33 identifies an A base, a class label of 0.66 identifies a C base, a class label of 1 identifies a T base, and a class label of 0 identifies a G base. In another implementation, a class label of 0.50 identifies an A base, a class label of 0.75 identifies a C base, a class label of 1 identifies a T base, and a class label of 0.25 identifies a G base. In one implementation, the method includes deriving, from the output, a single output value, comparing the single output value against class value ranges corresponding to bases A, C, T, and G, based on the comparing, assigning the single output value to a particular class value range, and base calling the target analyte based on the assigning. In one implementation, the single output value is derived using a sigmoid function, and the single output value ranges from 0 to 1. In another implementation, a class value range of 0-0.25 represents an A base, a class value range of 0.25-0.50 represents a C base, a class value range of 0.50-0.75 represents a T base, and a class value range of 0.75-1 represents a G base.

[0083] Additional details about the neural network-based base caller 102 can be found in US Provisional Patent Application No. 62/821,766, titled “ARTIFICIAL INTELLIGENCE-BASED

SEQUENCING,” (Attorney Docket No. ILLM 1008-9/IP-1752-PRV), filed on March 21, 2019, which is incorporated herein by reference.

Many-to-Many Base Calling

[0084] The technology disclosed causes the neural network-based base caller 102 to generate, for a given window of input, base calls not only for the center sequencing cycle but also for the flanking sequencing cycles, in accordance with one implementation. That is, in one implementation, the technology disclosed simultaneously generates base calls for cycle N, cycle N+1, cycle N-1, cycle N+2, cycle N-2, and so on for a given input window. That is, a single forward propagation/traversal/base calling iteration of the neural network-based base caller 102 generates base calls for multiple sequencing cycles in the input window of sequencing cycles, which is referred to herein as “many-to-many base calling.”

[0085] The technology disclosed then uses the disclosed many-to-many base calling to generate multiple base calls for a same target sequencing cycle that appeared across multiple sliding windows of inputs. For example, the target sequencing cycle can appear at different positions in the multiple sliding windows of inputs (*e.g.*, starting at position N+2 in the first sliding window, progressing to position N+1 in the second sliding window, and finishing at position N in the third sliding window).

[0086] Base calling the target sequencing cycle multiple times produces multiple candidates of the correct base call for the target sequencing cycle. The technology disclosed then evaluates the multiple candidates of the correct base call as an aggregate and determines a final base call for the target sequencing cycle. The final base call for the target sequencing cycle can be selected using aggregate-analysis techniques like averaging, consensus, and weighted consensus.

[0087] Figures 1A, 1B, and 1C show the disclosed many-to-many base calling 100. The neural network-based base caller 102 (*i.e.*, the base caller 102) processes at least right flanking, center, and left flanking inputs, and produces at least right flanking, center, and left flanking outputs, in accordance with one implementation of the technology disclosed.

[0088] The many-to-many base calling 100 is configured to provide data for n number of sequencing cycles as input to the base caller 102 and generate base calls for any number of cycles in the n number of cycles in one iteration of base calling (*i.e.*, one forward pass instance). A target sequencing cycle 108 can be base called n number of times and can appear/occur/fall at various positions in the n number of times of the base calling iterations.

[0089] The target sequencing cycle 108 can be the center sequencing cycle (Figure 1B) in some base calling iterations. In other iterations, the target sequencing cycle 108 can be the right flanking/context sequencing cycle (Figure 1A) adjacent to the center sequencing cycle or can be

the left flanking/context sequencing cycle adjacent (Figure 1C) to the center sequencing cycle. The right or left offset from the center sequencing cycle can also vary. That is, the target sequencing cycle 108 in the n number of times of the base calling iterations can fall either at the center position, immediately to the right of the center position, immediately to the left of the center position, at any offset to the right of the center position, at any offset to the left of the center position, or at any other position in the n number of times of the base calling iterations. The base calling iterations for the target sequencing cycle can have inputs of varying lengths of sequencing cycles and also multiple base calling outputs for various lengths of sequencing cycles in a given input window of sequencing cycles.

[0090] In one implementation, the technology disclosed includes accessing a progression of per-cycle analyte channel sets generated for sequencing cycles of a sequencing run, processing, through the neural network-based base caller 102, windows of per-cycle analyte channel sets in the progression for the windows of sequencing cycles of the sequencing run such that the neural network-based base caller processes 102 a subject window of per-cycle analyte channel sets in the progression for the subject window of sequencing cycles of the sequencing run and generates provisional base call predictions for three or more sequencing cycles in the subject window of sequencing cycles, from multiple windows in which a particular sequencing cycle appeared at different positions, using the neural network-based base caller 102 to generate provisional base call predictions for the particular sequencing cycle, and determining a base call for the particular sequencing cycle based on the provisional base call predictions.

[0091] In one implementation, the technology disclosed includes accessing a series of per-cycle analyte channel sets generated for sequencing cycles of a sequencing run, processing, through the neural network-based base caller 102, windows of per-cycle analyte channel sets in the series for the windows of sequencing cycles of the sequencing run such that the neural network-based base caller 102 processes a subject window of per-cycle analyte channel sets in the series for the subject window of sequencing cycles of the sequencing run and generates base call predictions for two or more sequencing cycles in the subject window of sequencing cycles, and processing, through the neural network-based base caller 102, a plurality of the windows of per-cycle analyte channel sets in the series for the plurality of the windows of sequencing cycles of the sequencing run and generating an output for each window in the plurality of the windows.

[0092] Each window in the plurality of the windows can include a particular per-cycle analyte channel set for a particular sequencing cycle of the sequencing run. The output for each window in the plurality of the windows includes (i) a base call prediction for the particular sequencing cycle and (ii) one or more additional base call predictions for one or more additional sequencing cycles of the sequencing run, thereby generating a plurality of base call predictions

for the particular sequencing cycle across the plurality of the windows (for example, generated in parallel or simultaneously by an output layer). Finally, the technology disclosed includes determining a base call for the particular sequencing cycle based on the plurality of base call predictions.

[0093] The right flanking input 132 comprises current image data 108 for a current sequencing cycle (*e.g.*, cycle 4) of a sequencing run, supplemented with previous image data 104 and 106 for one or more previous sequencing cycles (*e.g.*, cycles 2 and 3) preceding the current sequencing cycle. The right flanking output 142 comprises a right flanking base call prediction 114 for the current sequencing cycle and base call predictions 110 and 112 for the previous sequencing cycles.

[0094] The center input 134 comprises the current image data 108, supplemented with the previous image data 106 (*e.g.*, cycle 3) and successive image data 116 for one or more successive sequencing cycles (*e.g.*, cycle 5) succeeding the current sequencing cycle. The center output 144 comprises a center base call prediction 120 for the current sequencing cycle and base call predictions 118 and 122 for the previous sequencing cycles and the successive sequencing cycles.

[0095] The left flanking input 136 comprises the current image data 108, supplemented with the successive image data 116 and 124. The left flanking output 146 comprises a left flanking base call prediction 126 for the current sequencing cycle and base call predictions 128 and 130 for the successive sequencing cycles (*e.g.*, cycles 5 and 6).

[0096] Figures 1D and 1E illustrate different examples of the disclosed many-to-many base calling. In Figures 1D and 1E, the blue box represents the particular or target sequencing cycle (or data therefor). The particular sequencing cycle is also considered a current sequencing cycle in various implementations of the technology disclosed. The orange boxes represent sequencing cycles (or data therefor) different than the particular sequencing cycle. The green circles represent one or more base calls generated for the particular sequencing cycle. The base calls can be generated by any base caller, such as Illumina's Real-Time Analysis (RTA) software or the disclosed neural network-based base caller 102. The data for the sequencing cycles can be images or some other type of input data, such as current readings, voltage changes, pH scale data, and so on.

[0097] Turning to Figure 1D, the first many-to-many base calling example 180 shows three base calling iterations 180a, 180b, and 180c and corresponding three input windows/groups of sequencing cycles w1, w2, and w3 (or data therefore). In one implementation, the base calling iterations produce base calls for each sequencing cycle in the corresponding input window of sequencing cycles. In another implementation, the base calling iterations produce base calls for

only some of the sequencing cycles in the corresponding input window of sequencing cycles (*e.g.*, only the particular sequencing cycle). Also, the particular sequencing cycle can appear at different positions in the input windows/groups of sequencing cycles w1, w2, and w3. In other implementations (not shown), two or more input windows/groups of sequencing cycles can have the particular sequencing cycle at a same position. Furthermore, the input windows/groups of sequencing cycles w1, w2, and w3 have the particular sequencing cycle as at least one overlapping cycle, and also have one or more non-overlapping cycles. That is, orange boxes at different positions in different input windows/group of sequencing cycles represent different non-overlapping cycles. Finally, the three base calling iterations 180a, 180b, and 180c generate three base calls (*i.e.*, the three green circles) for the particular sequencing cycle, which can be considered provisional base calls and subsequently analyzed as an aggregate to make a final base call for the particular sequencing cycle. Different examples of analysis are described later in Figures 2, 3, and 4.

[0098] The second and third examples of many-to-many base calling 181 and 182 illustrate that the particular sequencing cycle can be at any position in the input windows/group of sequencing cycles and have any number of right and left flanking cycles or no flanking cycles at all (*e.g.*, the third window (w3) in the third many-to-many base calling example 182. The three base calling iterations 181a, 181b, and 181c generate three base calls (*i.e.*, the three green circles) for the particular sequencing cycle, which can be considered provisional base calls and subsequently analyzed as an aggregate to make a final base call for the particular sequencing cycle. Different examples of analysis are described later in Figures 2, 3, and 4. The three base calling iterations 182a, 182b, and 182c generate three base calls (*i.e.*, the three green circles) for the particular sequencing cycle, which can be considered provisional base calls and subsequently analyzed as an aggregate to make a final base call for the particular sequencing cycle. Different examples of analysis are described later in Figures 2, 3, and 4.

[0099] Figure 1E illustrates a many-to-many base calling example 183 with five base calling iterations 183a-e, each of which generates a base call prediction for the particular sequencing cycles by processing five respective windows/sets/groups of input data in which data for the particular sequencing cycle occurs at different positions. The five base calling iterations 183a-e generate five base calls (*i.e.*, the five green circles) for the particular sequencing cycle, which can be considered provisional base calls and subsequently analyzed as an aggregate to make a final base call for the particular sequencing cycle. Different examples of analysis are described later in Figures 2, 3, and 4.

[0100] Figures 2, 3, and 4 show different implementations of a base call generator. The base call generator 202 (*e.g.*, running on a host processor) is coupled (*e.g.*, via a PCIe bus or Ethernet

or InfiniBand (IB)) to the neural network-based base caller 102 (*e.g.*, running on a chip) and is configured to generate a base call for the current sequencing cycle (*e.g.*, cycle 4) based on the right flanking, center, and left flanking base call predictions for the current sequencing cycle.

[0101] The current image data for the current sequencing cycle depicts intensity emissions of analytes and their surrounding background captured at the current sequencing cycle. The right flanking 114, center 120, and left 126 flanking base call predictions for the current sequencing cycle (*e.g.*, cycle 4) identify likelihoods of a base incorporated in one or more of the analytes at the current sequencing cycle being A, C, T, and G. In one implementation, the likelihoods are exponentially normalized scores produced by a softmax layer used as an output layer by the base caller 102.

[0102] In one implementation, the right flanking base call prediction 114 for the current sequencing cycle accounts for prephasing effect between the current sequencing cycle (*e.g.*, cycle 4) and the previous sequencing cycles. In one implementation, the center base call prediction 120 for the current sequencing cycle (*e.g.*, cycle 4) accounts for the prephasing effect between the current sequencing cycle and the previous sequencing cycles and phasing effect between the current sequencing cycle and the successive sequencing cycles. In one implementation, the left flanking base call prediction 126 for the current sequencing cycle (*e.g.*, cycle 4) accounts for the phasing effect between the current sequencing cycle and the successive sequencing cycles.

[0103] As shown in Figure 2, the base call generator is further configured to comprise an averager 204 that base-wise sums the likelihoods across the right flanking 114, center 120, and left 126 flanking base call predictions for the current sequencing cycle (*e.g.*, cycle 4), determines base-wise averages 212 based on the base-wise sums, and generates the base call 214 for the current sequencing cycle (*e.g.*, cycle 4) based on a highest one of the base-wise averages (*e.g.*, 0.38).

[0104] As shown in Figure 3, the base call generator is further configured to comprise a consensor 304 that determines a preliminary base call for each of the right flanking 114, center 120, and left flanking 126 base call predictions for the current sequencing cycle (*e.g.*, cycle 4) based on a highest one of the likelihoods, thereby producing a sequence 306 of preliminary base calls, and generates the base call for the current sequencing cycle based on a most common base call 308 in the sequence of preliminary base calls.

[0105] As shown in Figure 4, the base call generator is further configured to comprise a weighted consensor 404 that determines a preliminary base call for each of the right flanking, center, and left flanking base call predictions for the current sequencing cycle based on a highest one of the likelihoods, thereby producing a sequence 406 of preliminary base calls, applies base-

wise weights 408 to respective ones of the preliminary base calls in the sequence of preliminary base calls and produces a sequence 410 of weighted preliminary base calls, and generates the base call for the current sequencing cycle (*e.g.*, cycle 4) based on a most weighted base call 412 in the sequence of weighted preliminary base calls. In some implementations, the base-wise weights 408 are preset, for example, on a cycle-by-cycle basis. In other implementations, the base-wise weights 408 are learned, for example, using a least-square approach.

[0106] In one implementation illustrated in Figure 6, the technology disclosed includes accessing current image data for a current sequencing cycle of a sequencing run (action 602), previous image data for one or more previous sequencing cycles preceding the current sequencing cycle, and successive image data for one or more successive sequencing cycles succeeding the current sequencing cycle, processing different groupings of the current image data, the previous image data, and the successive image data through a neural network-based base caller and producing first, second, and third base call predictions for the current sequencing cycle (action 612), and generating a base call for the current sequencing cycle based on the first, second, and third base call predictions (action 622).

[0107] In one implementation, the different groupings include a first grouping comprising the current image data and the previous image data, a second grouping comprising the current image data, the previous image data, and the successive image data, and a third grouping comprising the current image data and the successive image data.

[0108] In one implementation, the technology disclosed includes processing the first grouping through the neural network-based base caller to produce the first base call prediction, processing the second grouping through the neural network-based base caller to produce the second base call prediction, and processing the third grouping through the neural network-based base caller to produce the third base call prediction.

[0109] In one implementation, the first, second, and third base call predictions for the current sequencing cycle identify likelihoods of a base incorporated in one or more of the analytes at the current sequencing cycle being A, C, T, and G.

[0110] In one implementation, the technology disclosed includes generating the base call for the current sequencing cycle by base-wise summing the likelihoods across the first, second, and third base call predictions for the current sequencing cycle, determining base-wise averages based on the base-wise summing, and generating the base call for the current sequencing cycle based on a highest one of the base-wise averages.

[0111] In one implementation, the technology disclosed includes generating the base call for the current sequencing cycle by determining a preliminary base call for each of the first, second, and third base call predictions for the current sequencing cycle based on a highest one of the

likelihoods, thereby producing a sequence of preliminary base calls, and generating the base call for the current sequencing cycle based on a most common base call in the sequence of preliminary base calls.

[0112] In one implementation, the technology disclosed includes generating the base call for the current sequencing cycle by determining a preliminary base call for each of the first, second, and third base call predictions for the current sequencing cycle based on a highest one of the likelihoods, thereby producing a sequence of preliminary base calls, applying base-wise weights to respective ones of the preliminary base calls in the sequence of preliminary base calls and producing a sequence of weighted preliminary base calls, and generating the base call for the current sequencing cycle based on a most weighted base call in the sequence of weighted preliminary base calls.

[0113] In one implementation called “multi-cycle training, single-cycle inference,” the base caller 102 is trained to produce two or more base call predictions for two or more sequencing cycles during training using the base caller generator, but during inference produces base call prediction only for a single sequencing cycle.

[0114] In one implementation called “multi-cycle training, multi-cycle inference,” the base caller 102 is trained to produce two or more base call predictions for two or more sequencing cycles during training and do the same during the inference using the base caller generator 202.

Multi-Cycle Gradient Backpropagation

[0115] Figure 5 shows one implementation of the disclosed “multi-cycle gradient backpropagation 500.” As shown in Figure 5, the many-to-many base calling 100 is further configured to comprise a trainer that, computes errors 512, 532, and 552 between base calls generated by the base call generator 202 for the current sequencing cycle (*e.g.*, cycle 3), the previous sequencing cycles (*e.g.*, cycle 2), and the successive sequencing cycles (*e.g.*, cycle 4) based on the right flanking 506, center 504, and left flanking 502 outputs of the neural network-based base caller 102 and respective ground truth base calls 554, 534, and 514, determines respective gradients 542, 522, and 562 for the current sequencing cycle, the previous sequencing cycles, and the successive sequencing cycles based on the errors, and updates parameters of the neural network-based base caller by backpropagating the gradients.

Technical Effect/Advantage

[0116] Figure 7 illustrates technical effect and advantages of the technology disclosed.

[0117] The “multi-cycle training, single-cycle inference” implementation is called “DL 3C Middle” in Figure 7 and improves the base calling error rate by 8% over the traditional non-neural network-based real time analysis base calling software.

[0118] The “multi-cycle training, multi-cycle inference” implementation is called “DL 3C Average” in Figure 7 and further improves the base calling error rate by another 8% over the DL 3C Middle implementation.

[0119] Base calling sequencing cycles multiple times improves the base calling accuracy and detects and resolves base calling discrepancies and ambiguous base calls.

[0120] The multi-cycle gradient backpropagation also improves the gradients of the base caller 102 and its base calling accuracy over the training task of base calling.

Sequencing System

[0121] Figures 8A and 8B depict one implementation of a sequencing system 800A. The sequencing system 800A comprises a configurable processor 846. The configurable processor 846 implements the base calling techniques disclosed herein. The sequencing system is also referred to as a “sequencer.”

[0122] The sequencing system 800A can operate to obtain any information or data that relates to at least one of a biological or chemical substance. In some implementations, the sequencing system 800A is a workstation that may be similar to a bench-top device or desktop computer. For example, a majority (or all) of the systems and components for conducting the desired reactions can be within a common housing 802.

[0123] In particular implementations, the sequencing system 800A is a nucleic acid sequencing system configured for various applications, including but not limited to de novo sequencing, resequencing of whole genomes or target genomic regions, and metagenomics. The sequencer may also be used for DNA or RNA analysis. In some implementations, the sequencing system 800A may also be configured to generate reaction sites in a biosensor. For example, the sequencing system 800A may be configured to receive a sample and generate surface attached clusters of clonally amplified nucleic acids derived from the sample. Each cluster may constitute or be part of a reaction site in the biosensor.

[0124] The exemplary sequencing system 800A may include a system receptacle or interface 810 that is configured to interact with a biosensor 812 to perform desired reactions within the biosensor 812. In the following description with respect to Figure 8A, the biosensor 812 is loaded into the system receptacle 810. However, it is understood that a cartridge that includes the biosensor 812 may be inserted into the system receptacle 810 and in some states the cartridge can be removed temporarily or permanently. As described above, the cartridge may include, among other things, fluidic control and fluidic storage components.

[0125] In particular implementations, the sequencing system 800A is configured to perform a large number of parallel reactions within the biosensor 812. The biosensor 812 includes one or

more reaction sites where desired reactions can occur. The reaction sites may be, for example, immobilized to a solid surface of the biosensor or immobilized to beads (or other movable substrates) that are located within corresponding reaction chambers of the biosensor. The reaction sites can include, for example, clusters of clonally amplified nucleic acids. The biosensor 812 may include a solid-state imaging device (*e.g.*, CCD or CMOS imager) and a flow cell mounted thereto. The flow cell may include one or more flow channels that receive a solution from the sequencing system 800A and direct the solution toward the reaction sites. Optionally, the biosensor 812 can be configured to engage a thermal element for transferring thermal energy into or out of the flow channel.

[0126] The sequencing system 800A may include various components, assemblies, and systems (or sub-systems) that interact with each other to perform a predetermined method or assay protocol for biological or chemical analysis. For example, the sequencing system 800A includes a system controller 806 that may communicate with the various components, assemblies, and sub-systems of the sequencing system 800A and also the biosensor 812. For example, in addition to the system receptacle 810, the sequencing system 800A may also include a fluidic control system 808 to control the flow of fluid throughout a fluid network of the sequencing system 800A and the biosensor 812; a fluid storage system 814 that is configured to hold all fluids (*e.g.*, gas or liquids) that may be used by the bioassay system; a temperature control system 804 that may regulate the temperature of the fluid in the fluid network, the fluid storage system 814, and/or the biosensor 812; and an illumination system 816 that is configured to illuminate the biosensor 812. As described above, if a cartridge having the biosensor 812 is loaded into the system receptacle 810, the cartridge may also include fluidic control and fluidic storage components.

[0127] Also shown, the sequencing system 800A may include a user interface 818 that interacts with the user. For example, the user interface 818 may include a display 820 to display or request information from a user and a user input device 822 to receive user inputs. In some implementations, the display 820 and the user input device 822 are the same device. For example, the user interface 818 may include a touch-sensitive display configured to detect the presence of an individual's touch and also identify a location of the touch on the display. However, other user input devices 822 may be used, such as a mouse, touchpad, keyboard, keypad, handheld scanner, voice-recognition system, motion-recognition system, and the like. As will be discussed in greater detail below, the sequencing system 800A may communicate with various components, including the biosensor 812 (*e.g.*, in the form of a cartridge), to perform the desired reactions. The sequencing system 800A may also be configured to analyze data obtained from the biosensor to provide a user with desired information.

[0128] The system controller 806 may include any processor-based or microprocessor-based system, including systems using microcontrollers, reduced instruction set computers (RISC), application specific integrated circuits (ASICs), field programmable gate array (FPGAs), coarse-grained reconfigurable architectures (CGRAs), logic circuits, and any other circuit or processor capable of executing functions described herein. The above examples are exemplary only, and are thus not intended to limit in any way the definition and/or meaning of the term system controller. In the exemplary implementation, the system controller 806 executes a set of instructions that are stored in one or more storage elements, memories, or modules in order to at least one of obtain and analyze detection data. Detection data can include a plurality of sequences of pixel signals, such that a sequence of pixel signals from each of the millions of sensors (or pixels) can be detected over many base calling cycles. Storage elements may be in the form of information sources or physical memory elements within the sequencing system 800A.

[0129] The set of instructions may include various commands that instruct the sequencing system 800A or biosensor 812 to perform specific operations such as the methods and processes of the various implementations described herein. The set of instructions may be in the form of a software program, which may form part of a tangible, non-transitory computer readable medium or media. As used herein, the terms “software” and “firmware” are interchangeable, and include any computer program stored in memory for execution by a computer, including RAM memory, ROM memory, EPROM memory, EEPROM memory, and non-volatile RAM (NVRAM) memory. The above memory types are exemplary only, and are thus not limiting as to the types of memory usable for storage of a computer program.

[0130] The software may be in various forms such as system software or application software. Further, the software may be in the form of a collection of separate programs, or a program module within a larger program or a portion of a program module. The software also may include modular programming in the form of object-oriented programming. After obtaining the detection data, the detection data may be automatically processed by the sequencing system 800A, processed in response to user inputs, or processed in response to a request made by another processing machine (*e.g.*, a remote request through a communication link). In the illustrated implementation, the system controller 806 includes an analysis module 844. In other implementations, system controller 806 does not include the analysis module 844 and instead has access to the analysis module 844 (*e.g.*, the analysis module 844 may be separately hosted on cloud).

[0131] The system controller 806 may be connected to the biosensor 812 and the other components of the sequencing system 800A via communication links. The system controller 806

may also be communicatively connected to off-site systems or servers. The communication links may be hardwired, corded, or wireless. The system controller 806 may receive user inputs or commands, from the user interface 818 and the user input device 822.

[0132] The fluidic control system 808 includes a fluid network and is configured to direct and regulate the flow of one or more fluids through the fluid network. The fluid network may be in fluid communication with the biosensor 812 and the fluid storage system 814. For example, select fluids may be drawn from the fluid storage system 814 and directed to the biosensor 812 in a controlled manner, or the fluids may be drawn from the biosensor 812 and directed toward, for example, a waste reservoir in the fluid storage system 814. Although not shown, the fluidic control system 808 may include flow sensors that detect a flow rate or pressure of the fluids within the fluid network. The sensors may communicate with the system controller 806.

[0133] The temperature control system 804 is configured to regulate the temperature of fluids at different regions of the fluid network, the fluid storage system 814, and/or the biosensor 812. For example, the temperature control system 804 may include a thermocycler that interfaces with the biosensor 812 and controls the temperature of the fluid that flows along the reaction sites in the biosensor 812. The temperature control system 804 may also regulate the temperature of solid elements or components of the sequencing system 800A or the biosensor 812. Although not shown, the temperature control system 804 may include sensors to detect the temperature of the fluid or other components. The sensors may communicate with the system controller 806.

[0134] The fluid storage system 814 is in fluid communication with the biosensor 812 and may store various reaction components or reactants that are used to conduct the desired reactions therein. The fluid storage system 814 may also store fluids for washing or cleaning the fluid network and biosensor 812 and for diluting the reactants. For example, the fluid storage system 814 may include various reservoirs to store samples, reagents, enzymes, other biomolecules, buffer solutions, aqueous, and non-polar solutions, and the like. Furthermore, the fluid storage system 814 may also include waste reservoirs for receiving waste products from the biosensor 812. In implementations that include a cartridge, the cartridge may include one or more of a fluid storage system, fluidic control system or temperature control system. Accordingly, one or more of the components set forth herein as relating to those systems can be contained within a cartridge housing. For example, a cartridge can have various reservoirs to store samples, reagents, enzymes, other biomolecules, buffer solutions, aqueous, and non-polar solutions, waste, and the like. As such, one or more of a fluid storage system, fluidic control system or temperature control system can be removably engaged with a bioassay system via a cartridge or other biosensor.

[0135] The illumination system 816 may include a light source (*e.g.*, one or more LEDs) and a plurality of optical components to illuminate the biosensor. Examples of light sources may include lasers, arc lamps, LEDs, or laser diodes. The optical components may be, for example, reflectors, dichroics, beam splitters, collimators, lenses, filters, wedges, prisms, mirrors, detectors, and the like. In implementations that use an illumination system, the illumination system 816 may be configured to direct an excitation light to reaction sites. As one example, fluorophores may be excited by green wavelengths of light, as such the wavelength of the excitation light may be approximately 532 nm. In one implementation, the illumination system 816 is configured to produce illumination that is parallel to a surface normal of a surface of the biosensor 812. In another implementation, the illumination system 816 is configured to produce illumination that is off-angle relative to the surface normal of the surface of the biosensor 812. In yet another implementation, the illumination system 816 is configured to produce illumination that has plural angles, including some parallel illumination and some off-angle illumination.

[0136] The system receptacle or interface 810 is configured to engage the biosensor 812 in at least one of a mechanical, electrical, and fluidic manner. The system receptacle 810 may hold the biosensor 812 in a desired orientation to facilitate the flow of fluid through the biosensor 812. The system receptacle 810 may also include electrical contacts that are configured to engage the biosensor 812 so that the sequencing system 800A may communicate with the biosensor 812 and/or provide power to the biosensor 812. Furthermore, the system receptacle 810 may include fluidic ports (*e.g.*, nozzles) that are configured to engage the biosensor 812. In some implementations, the biosensor 812 is removably coupled to the system receptacle 810 in a mechanical manner, in an electrical manner, and also in a fluidic manner.

[0137] In addition, the sequencing system 800A may communicate remotely with other systems or networks or with other bioassay systems 800A. Detection data obtained by the bioassay system(s) 800A may be stored in a remote database.

[0138] Figure 8B is a block diagram of a system controller 806 that can be used in the system of Figure 8A. In one implementation, the system controller 806 includes one or more processors or modules that can communicate with one another. Each of the processors or modules may include an algorithm (*e.g.*, instructions stored on a tangible and/or non-transitory computer readable storage medium) or sub-algorithms to perform particular processes. The system controller 806 is illustrated conceptually as a collection of modules, but may be implemented utilizing any combination of dedicated hardware boards, DSPs, processors, etc. Alternatively, the system controller 806 may be implemented utilizing an off-the-shelf PC with a single processor or multiple processors, with the functional operations distributed between the processors. As a further option, the modules described below may be implemented utilizing a

hybrid configuration in which certain modular functions are performed utilizing dedicated hardware, while the remaining modular functions are performed utilizing an off-the-shelf PC and the like. The modules also may be implemented as software modules within a processing unit.

[0139] During operation, a communication port 850 may transmit information (*e.g.*, commands) to or receive information (*e.g.*, data) from the biosensor 812 (Figure 8A) and/or the sub-systems 808, 814, 804 (Figure 8A). In implementations, the communication port 850 may output a plurality of sequences of pixel signals. A communication link 834 may receive user input from the user interface 818 (Figure 8A) and transmit data or information to the user interface 818. Data from the biosensor 812 or sub-systems 808, 814, 804 may be processed by the system controller 806 in real-time during a bioassay session. Additionally or alternatively, data may be stored temporarily in a system memory during a bioassay session and processed in slower than real-time or off-line operation.

[0140] As shown in Figure 8B, the system controller 806 may include a plurality of modules 824-848 that communicate with a main control module 824, along with a central processing unit (CPU) 852. The main control module 824 may communicate with the user interface 818 (Figure 8A). Although the modules 824-848 are shown as communicating directly with the main control module 824, the modules 824-848 may also communicate directly with each other, the user interface 818, and the biosensor 812. Also, the modules 824-848 may communicate with the main control module 824 through the other modules.

[0141] The plurality of modules 824-848 include system modules 828-832, 826 that communicate with the sub-systems 808, 814, 804, and 816, respectively. The fluidic control module 828 may communicate with the fluidic control system 808 to control the valves and flow sensors of the fluid network for controlling the flow of one or more fluids through the fluid network. The fluid storage module 830 may notify the user when fluids are low or when the waste reservoir is at or near capacity. The fluid storage module 830 may also communicate with the temperature control module 832 so that the fluids may be stored at a desired temperature. The illumination module 826 may communicate with the illumination system 816 to illuminate the reaction sites at designated times during a protocol, such as after the desired reactions (*e.g.*, binding events) have occurred. In some implementations, the illumination module 826 may communicate with the illumination system 816 to illuminate the reaction sites at designated angles.

[0142] The plurality of modules 824-848 may also include a device module 836 that communicates with the biosensor 812 and an identification module 838 that determines identification information relating to the biosensor 812. The device module 836 may, for example, communicate with the system receptacle 810 to confirm that the biosensor has

established an electrical and fluidic connection with the sequencing system 800A. The identification module 838 may receive signals that identify the biosensor 812. The identification module 838 may use the identity of the biosensor 812 to provide other information to the user. For example, the identification module 838 may determine and then display a lot number, a date of manufacture, or a protocol that is recommended to be run with the biosensor 812.

[0143] The plurality of modules 824-848 also includes an analysis module 844 (also called signal processing module or signal processor) that receives and analyzes the signal data (*e.g.*, image data) from the biosensor 812. Analysis module 844 includes memory (*e.g.*, RAM or Flash) to store detection/image data. Detection data can include a plurality of sequences of pixel signals, such that a sequence of pixel signals from each of the millions of sensors (or pixels) can be detected over many base calling cycles. The signal data may be stored for subsequent analysis or may be transmitted to the user interface 818 to display desired information to the user. In some implementations, the signal data may be processed by the solid-state imager (*e.g.*, CMOS image sensor) before the analysis module 844 receives the signal data.

[0144] The analysis module 844 is configured to obtain image data from the light detectors at each of a plurality of sequencing cycles. The image data is derived from the emission signals detected by the light detectors and process the image data for each of the plurality of sequencing cycles through the base caller 102 and produce a base call for at least some of the analytes at each of the plurality of sequencing cycle. The light detectors can be part of one or more over-head cameras (*e.g.*, Illumina's GAIIX's CCD camera taking images of the clusters on the biosensor 812 from the top), or can be part of the biosensor 812 itself (*e.g.*, Illumina's iSeq's CMOS image sensors underlying the clusters on the biosensor 812 and taking images of the clusters from the bottom).

[0145] The output of the light detectors is the sequencing images, each depicting intensity emissions of the clusters and their surrounding background. The sequencing images depict intensity emissions generated as a result of nucleotide incorporation in the sequences during the sequencing. The intensity emissions are from associated analytes and their surrounding background. The sequencing images are stored in memory 848.

[0146] Protocol modules 840 and 842 communicate with the main control module 824 to control the operation of the sub-systems 808, 814, and 804 when conducting predetermined assay protocols. The protocol modules 840 and 842 may include sets of instructions for instructing the sequencing system 800A to perform specific operations pursuant to predetermined protocols. As shown, the protocol module may be a sequencing-by-synthesis (SBS) module 840 that is configured to issue various commands for performing sequencing-by-synthesis processes. In SBS, extension of a nucleic acid primer along a nucleic acid template is

monitored to determine the sequence of nucleotides in the template. The underlying chemical process can be polymerization (*e.g.*, as catalyzed by a polymerase enzyme) or ligation (*e.g.*, catalyzed by a ligase enzyme). In a particular polymerase-based SBS implementation, fluorescently labeled nucleotides are added to a primer (thereby extending the primer) in a template dependent fashion such that detection of the order and type of nucleotides added to the primer can be used to determine the sequence of the template. For example, to initiate a first SBS cycle, commands can be given to deliver one or more labeled nucleotides, DNA polymerase, etc., into/through a flow cell that houses an array of nucleic acid templates. The nucleic acid templates may be located at corresponding reaction sites. Those reaction sites where primer extension causes a labeled nucleotide to be incorporated can be detected through an imaging event. During an imaging event, the illumination system 816 may provide an excitation light to the reaction sites. Optionally, the nucleotides can further include a reversible termination property that terminates further primer extension once a nucleotide has been added to a primer. For example, a nucleotide analog having a reversible terminator moiety can be added to a primer such that subsequent extension cannot occur until a deblocking agent is delivered to remove the moiety. Thus, for implementations that use reversible termination a command can be given to deliver a deblocking reagent to the flow cell (before or after detection occurs). One or more commands can be given to effect wash(es) between the various delivery steps. The cycle can then be repeated *n* times to extend the primer by *n* nucleotides, thereby detecting a sequence of length *n*. Exemplary sequencing techniques are described, for example, in Bentley et al., *Nature* 456:53-59 (2008); WO 04/018497; US 7,057,026; WO 91/06678; WO 07/123744; US 7,329,492; US 7,211,414; US 7,315,019; US 7,405,281, and US 2008/014708082, each of which is incorporated herein by reference.

[0147] For the nucleotide delivery step of an SBS cycle, either a single type of nucleotide can be delivered at a time, or multiple different nucleotide types (*e.g.*, A, C, T and G together) can be delivered. For a nucleotide delivery configuration where only a single type of nucleotide is present at a time, the different nucleotides need not have distinct labels since they can be distinguished based on temporal separation inherent in the individualized delivery. Accordingly, a sequencing method or apparatus can use single color detection. For example, an excitation source need only provide excitation at a single wavelength or in a single range of wavelengths. For a nucleotide delivery configuration where delivery results in multiple different nucleotides being present in the flow cell at one time, sites that incorporate different nucleotide types can be distinguished based on different fluorescent labels that are attached to respective nucleotide types in the mixture. For example, four different nucleotides can be used, each having one of four different fluorophores. In one implementation, the four different fluorophores can be

distinguished using excitation in four different regions of the spectrum. For example, four different excitation radiation sources can be used. Alternatively, fewer than four different excitation sources can be used, but optical filtration of the excitation radiation from a single source can be used to produce different ranges of excitation radiation at the flow cell.

[0148] In some implementations, fewer than four different colors can be detected in a mixture having four different nucleotides. For example, pairs of nucleotides can be detected at the same wavelength, but distinguished based on a difference in intensity for one member of the pair compared to the other, or based on a change to one member of the pair (*e.g.*, via chemical modification, photochemical modification or physical modification) that causes apparent signal to appear or disappear compared to the signal detected for the other member of the pair. Exemplary apparatus and methods for distinguishing four different nucleotides using detection of fewer than four colors are described for example in US Pat. App. Ser. Nos. 61/538,294 and 61/619,878, which are incorporated herein by reference in their entireties. U.S. Application No. 13/624,200, which was filed on September 21, 2012, is also incorporated by reference in its entirety.

[0149] The plurality of protocol modules may also include a sample-preparation (or generation) module 842 that is configured to issue commands to the fluidic control system 808 and the temperature control system 804 for amplifying a product within the biosensor 812. For example, the biosensor 812 may be engaged to the sequencing system 800A. The amplification module 842 may issue instructions to the fluidic control system 808 to deliver necessary amplification components to reaction chambers within the biosensor 812. In other implementations, the reaction sites may already contain some components for amplification, such as the template DNA and/or primers. After delivering the amplification components to the reaction chambers, the amplification module 842 may instruct the temperature control system 804 to cycle through different temperature stages according to known amplification protocols. In some implementations, the amplification and/or nucleotide incorporation is performed isothermally.

[0150] The SBS module 840 may issue commands to perform bridge PCR where clusters of clonal amplicons are formed on localized areas within a channel of a flow cell. After generating the amplicons through bridge PCR, the amplicons may be “linearized” to make single stranded template DNA, or sstDNA, and a sequencing primer may be hybridized to a universal sequence that flanks a region of interest. For example, a reversible terminator-based sequencing by synthesis method can be used as set forth above or as follows.

[0151] Each base calling or sequencing cycle can extend an sstDNA by a single base which can be accomplished for example by using a modified DNA polymerase and a mixture of four

types of nucleotides. The different types of nucleotides can have unique fluorescent labels, and each nucleotide can further have a reversible terminator that allows only a single-base incorporation to occur in each cycle. After a single base is added to the sstDNA, excitation light may be incident upon the reaction sites and fluorescent emissions may be detected. After detection, the fluorescent label and the terminator may be chemically cleaved from the sstDNA. Another similar base calling or sequencing cycle may follow. In such a sequencing protocol, the SBS module 840 may instruct the fluidic control system 808 to direct a flow of reagent and enzyme solutions through the biosensor 812. Exemplary reversible terminator-based SBS methods which can be utilized with the apparatus and methods set forth herein are described in US Patent Application Publication No. 2007/0166705 A1, US Patent Application Publication No. 2006/0188901 A1, US Patent No. 7,057,026, US Patent Application Publication No. 2006/0240439 A1, US Patent Application Publication No. 2006/0281471 A1, PCT Publication No. WO 05/065814, US Patent Application Publication No. 2005/0147009 A1, PCT Publication No. WO 06/08B199 and PCT Publication No. WO 07/01470251, each of which is incorporated herein by reference in its entirety. Exemplary reagents for reversible terminator-based SBS are described in US 7,541,444; US 7,057,026; US 7,414,147; US 7,427,673; US 7,566,537; US 7,592,435 and WO 07/148353, each of which is incorporated herein by reference in its entirety.

[0152] In some implementations, the amplification and SBS modules may operate in a single assay protocol where, for example, template nucleic acid is amplified and subsequently sequenced within the same cartridge.

[0153] The sequencing system 800A may also allow the user to reconfigure an assay protocol. For example, the sequencing system 800A may offer options to the user through the user interface 818 for modifying the determined protocol. For example, if it is determined that the biosensor 812 is to be used for amplification, the sequencing system 800A may request a temperature for the annealing cycle. Furthermore, the sequencing system 800A may issue warnings to a user if a user has provided user inputs that are generally not acceptable for the selected assay protocol.

[0154] In implementations, the biosensor 812 includes millions of sensors (or pixels), each of which generates a plurality of sequences of pixel signals over successive base calling cycles. The analysis module 844 detects the plurality of sequences of pixel signals and attributes them to corresponding sensors (or pixels) in accordance to the row-wise and/or column-wise location of the sensors on an array of sensors.

[0155] Figure 9 is a simplified block diagram of a system for analysis of sensor data from the sequencing system 800A, such as base call sensor outputs. In the example of Figure 9, the

system includes the configurable processor 846. The configurable processor 846 can execute a base caller (*e.g.*, the neural network-based base caller 102) in coordination with a runtime program executed by the central processing unit (CPU) 852 (*i.e.*, a host processor). The sequencing system 800A comprises the biosensor 812 and flow cells. The flow cells can comprise one or more tiles in which clusters of genetic material are exposed to a sequence of analyte flows used to cause reactions in the clusters to identify the bases in the genetic material. The sensors sense the reactions for each cycle of the sequence in each tile of the flow cell to provide tile data. Genetic sequencing is a data intensive operation, which translates base call sensor data into sequences of base calls for each cluster of genetic material sensed in during a base call operation.

[0156] The system in this example includes the CPU 852, which executes a runtime program to coordinate the base call operations, memory 848B to store sequences of arrays of tile data, base call reads produced by the base calling operation, and other information used in the base call operations. Also, in this illustration the system includes memory 848A to store a configuration file (or files), such as FPGA bit files, and model parameters for the neural networks used to configure and reconfigure the configurable processor 846, and execute the neural networks. The sequencing system 800A can include a program for configuring a configurable processor and in some embodiments a reconfigurable processor to execute the neural networks.

[0157] The sequencing system 800A is coupled by a bus 902 to the configurable processor 846. The bus 902 can be implemented using a high throughput technology, such as in one example bus technology compatible with the PCIe standards (Peripheral Component Interconnect Express) currently maintained and developed by the PCI-SIG (PCI Special Interest Group). Also in this example, a memory 848A is coupled to the configurable processor 846 by bus 906. The memory 848A can be on-board memory, disposed on a circuit board with the configurable processor 846. The memory 848A is used for high speed access by the configurable processor 846 of working data used in the base call operation. The bus 906 can also be implemented using a high throughput technology, such as bus technology compatible with the PCIe standards.

[0158] Configurable processors, including field programmable gate arrays FPGAs, coarse grained reconfigurable arrays CGRAs, and other configurable and reconfigurable devices, can be configured to implement a variety of functions more efficiently or faster than might be achieved using a general purpose processor executing a computer program. Configuration of configurable processors involves compiling a functional description to produce a configuration file, referred to sometimes as a bitstream or bit file, and distributing the configuration file to the configurable

elements on the processor. The configuration file defines the logic functions to be executed by the configurable processor, by configuring the circuit to set data flow patterns, use of distributed memory and other on-chip memory resources, lookup table contents, operations of configurable logic blocks and configurable execution units like multiply-and-accumulate units, configurable interconnects and other elements of the configurable array. A configurable processor is reconfigurable if the configuration file may be changed in the field, by changing the loaded configuration file. For example, the configuration file may be stored in volatile SRAM elements, in non-volatile read-write memory elements, and in combinations of the same, distributed among the array of configurable elements on the configurable or reconfigurable processor. A variety of commercially available configurable processors are suitable for use in a base calling operation as described herein. Examples include Google's Tensor Processing Unit (TPU)TM, rackmount solutions like GX4 Rackmount SeriesTM, GX9 Rackmount SeriesTM, NVIDIA DGX-1TM, Microsoft's Stratix V FPGATM, Graphcore's Intelligent Processor Unit (IPU)TM, Qualcomm's Zeroth PlatformTM with Snapdragon processorsTM, NVIDIA's VoltaTM, NVIDIA's DRIVE PXTM, NVIDIA's JETSON TX1/TX2 MODULETM, Intel's NirvanaTM, Movidius VPUTM, Fujitsu DPITM, ARM's DynamicIQTM, IBM TrueNorthTM, Lambda GPU Server with Testa V100sTM, Xilinx AlveoTM U200, Xilinx AlveoTM U250, Xilinx AlveoTM U280, Intel/Altera StratixTM GX2800, Intel/Altera StratixTM GX2800, and Intel StratixTM GX10M. In some examples, a host CPU can be implemented on the same integrated circuit as the configurable processor.

[0159] Embodiments described herein implement the neural network-based base caller 102 using the configurable processor 846. The configuration file for the configurable processor 846 can be implemented by specifying the logic functions to be executed using a high level description language HDL or a register transfer level RTL language specification. The specification can be compiled using the resources designed for the selected configurable processor to generate the configuration file. The same or similar specification can be compiled for the purposes of generating a design for an application-specific integrated circuit which may not be a configurable processor.

[0160] Alternatives for the configurable processor configurable processor 846, in all embodiments described herein, therefore include a configured processor comprising an application specific ASIC or special purpose integrated circuit or set of integrated circuits, or a system-on-a-chip SOC device, or a graphics processing unit (GPU) processor or a coarse-grained reconfigurable architecture (CGRA) processor, configured to execute a neural network based base call operation as described herein.

[0161] In general, configurable processors and configured processors described herein, as configured to execute runs of a neural network, are referred to herein as neural network processors.

[0162] The configurable processor 846 is configured in this example by a configuration file loaded using a program executed by the CPU 852, or by other sources, which configures the array of configurable elements 916 (*e.g.*, configuration logic blocks (CLB) such as look up tables (LUTs), flip-flops, compute processing units (PMUs), and compute memory units (CMUs), configurable I/O blocks, programmable interconnects), on the configurable processor to execute the base call function. In this example, the configuration includes data flow logic 908 which is coupled to the buses 902 and 906 and executes functions for distributing data and control parameters among the elements used in the base call operation.

[0163] Also, the configurable processor 846 is configured with base call execution data flow logic 908 to execute the neural network-based base caller 102. The data flow logic 908 comprises multi-cycle execution clusters (*e.g.*, 914) which, in this example, includes execution cluster 1 through execution cluster X. The number of multi-cycle execution clusters can be selected according to a trade-off involving the desired throughput of the operation, and the available resources on the configurable processor 846.

[0164] The multi-cycle execution clusters are coupled to the data flow logic 908 by data flow paths 910 implemented using configurable interconnect and memory resources on the configurable processor 846. Also, the multi-cycle execution clusters are coupled to the data flow logic 908 by control paths 912 implemented using configurable interconnect and memory resources for example on the configurable processor 846, which provide control signals indicating available execution clusters, readiness to provide input units for execution of a run of the neural network-based base caller 102 to the available execution clusters, readiness to provide trained parameters for the neural network-based base caller 102, readiness to provide output patches of base call classification data, and other control data used for execution of the neural network-based base caller 102.

[0165] The configurable processor 846 is configured to execute runs of the neural network-based base caller 102 using trained parameters to produce classification data for the sensing cycles of the base calling operation. A run of the neural network-based base caller 102 is executed to produce classification data for a subject sensing cycle of the base calling operation. A run of the neural network-based base caller 102 operates on a sequence including a number N of arrays of tile data from respective sensing cycles of N sensing cycles, where the N sensing cycles provide sensor data for different base call operations for one base position per operation in time sequence in the examples described herein. Optionally, some of the N sensing cycles can be

out of sequence if needed according to a particular neural network model being executed. The number N can be any number greater than one. In some examples described herein, sensing cycles of the N sensing cycles represent a set of sensing cycles for at least one sensing cycle preceding the subject sensing cycle and at least one sensing cycle following the subject cycle in time sequence. Examples are described herein in which the number N is an integer equal to or greater than five.

[0166] The data flow logic 908 is configured to move tile data and at least some trained parameters of the model parameters from the memory 848A to the configurable processor 846 for runs of the neural network-based base caller 102, using input units for a given run including tile data for spatially aligned patches of the N arrays. The input units can be moved by direct memory access operations in one DMA operation, or in smaller units moved during available time slots in coordination with the execution of the neural network deployed.

[0167] Tile data for a sensing cycle as described herein can comprise an array of sensor data having one or more features. For example, the sensor data can comprise two images which are analyzed to identify one of four bases at a base position in a genetic sequence of DNA, RNA, or other genetic material. The tile data can also include metadata about the images and the sensors. For example, in embodiments of the base calling operation, the tile data can comprise information about alignment of the images with the clusters such as distance from center information indicating the distance of each pixel in the array of sensor data from the center of a cluster of genetic material on the tile.

[0168] During execution of the neural network-based base caller 102 as described below, tile data can also include data produced during execution of the neural network-based base caller 102, referred to as intermediate data, which can be reused rather than recomputed during a run of the neural network-based base caller 102. For example, during execution of the neural network-based base caller 102, the data flow logic 908 can write intermediate data to the memory 848A in place of the sensor data for a given patch of an array of tile data. Embodiments like this are described in more detail below.

[0169] As illustrated, a system is described for analysis of base call sensor output, comprising memory (*e.g.*, 848A) accessible by the runtime program storing tile data including sensor data for a tile from sensing cycles of a base calling operation. Also, the system includes a neural network processor, such as configurable processor 846 having access to the memory. The neural network processor is configured to execute runs of a neural network using trained parameters to produce classification data for sensing cycles. As described herein, a run of the neural network is operating on a sequence of N arrays of tile data from respective sensing cycles of N sensing cycles, including a subject cycle, to produce the classification data for the subject

cycle. The data flow logic 908 is provided to move tile data and the trained parameters from the memory to the neural network processor for runs of the neural network using input units including data for spatially aligned patches of the N arrays from respective sensing cycles of N sensing cycles.

[0170] Also, a system is described in which the neural network processor has access to the memory, and includes a plurality of execution clusters, the execution clusters in the plurality of execution clusters configured to execute a neural network. The data flow logic 908 has access to the memory and to execution clusters in the plurality of execution clusters, to provide input units of tile data to available execution clusters in the plurality of execution clusters, the input units including a number N of spatially aligned patches of arrays of tile data from respective sensing cycles, including a subject sensing cycle, and to cause the execution clusters to apply the N spatially aligned patches to the neural network to produce output patches of classification data for the spatially aligned patch of the subject sensing cycle, where N is greater than 1.

[0171] As shown in Figures 9 and 10, in one implementation, the technology disclosed comprises an artificial intelligence-based system for base calling. The system comprises a host processor, memory accessible by the host processor storing image data for sequencing cycles of a sequencing run, wherein current image data for a current sequencing cycle of the sequencing run depicts intensity emissions of analytes and their surrounding background captured at the current sequencing cycle, and a configurable processor having access to the memory, the configurable processor including a plurality of execution clusters, the execution clusters in the plurality of execution clusters configured to execute a neural network, and data flow logic having access to the memory and to the execution clusters in the plurality of execution clusters, configured to provide the current image data, previous image data for one or more previous sequencing cycles preceding the current sequencing cycle, and successive image data for one or more successive sequencing cycles succeeding the current sequencing cycle to available execution clusters in the plurality of execution clusters, cause the execution clusters to apply different groupings of the current image data, the previous image data, and the successive image data to the neural network to produce first, second, and third base call predictions for the current sequencing cycle, and to feedback the first, second, and third base call predictions for the current sequencing cycle to the memory for use in generation of a base call for the current sequencing cycle based on the first, second, and third base call predictions.

[0172] In one implementation, the different groupings include a first grouping comprising the current image data and the previous image data, a second grouping comprising the current image data, the previous image data, and the successive image data, and a third grouping comprising the current image data and the successive image data.

[0173] In one implementation, the execution clusters apply the first grouping to the neural network to produce the first base call prediction, the second grouping to the neural network to produce the second base call prediction, and the third grouping to the neural network to produce the third base call prediction.

[0174] In one implementation, the first, second, and third base call predictions for the current sequencing cycle identify likelihoods of a base incorporated in one or more of the analytes at the current sequencing cycle being A, C, T, and G.

[0175] In one implementation, the data flow logic is further configured to generate the base call for the current sequencing cycle by base-wise summing the likelihoods across the first, second, and third base call predictions for the current sequencing cycle, determining base-wise averages based on the base-wise summing, and generating the base call for the current sequencing cycle based on a highest one of the base-wise averages.

[0176] In one implementation, the data flow logic is further configured to generate the base call for the current sequencing cycle by determining a preliminary base call for each of the first, second, and third base call predictions for the current sequencing cycle based on a highest one of the likelihoods, thereby producing a sequence of preliminary base calls, and generating the base call for the current sequencing cycle based on a most common base call in the sequence of preliminary base calls.

[0177] In one implementation, the data flow logic is further configured to generate the base call for the current sequencing cycle by determining a preliminary base call for each of the first, second, and third base call predictions for the current sequencing cycle based on a highest one of the likelihoods, thereby producing a sequence of preliminary base calls, applying base-wise weights to respective ones of the preliminary base calls in the sequence of preliminary base calls and producing a sequence of weighted preliminary base calls, and generating the base call for the current sequencing cycle based on a most weighted base call in the sequence of weighted preliminary base calls.

[0178] Figure 10 is a simplified diagram showing aspects of the base calling operation, including functions of a runtime program executed by a host processor. In this diagram, the output of image sensors from a flow cell are provided on lines 1000 to image processing threads 1001, which can perform processes on images such as alignment and arrangement in an array of sensor data for the individual tiles and resampling of images, and can be used by processes which calculate a tile cluster mask for each tile in the flow cell, which identifies pixels in the array of sensor data that correspond to clusters of genetic material on the corresponding tile of the flow cell. The outputs of the image processing threads 1001 are provided on lines 1002 to a dispatch logic 1010 in the CPU which routes the arrays of tile data to a data cache 1004 (e.g.,

SSD storage) on a high-speed bus 1003, or on high-speed bus 1005 to the neural network processor hardware 1020, such as the configurable processor 846 of Figure 9, according to the state of the base calling operation. The processed and transformed images can be stored on the data cache 1004 for sensing cycles that were previously used. The hardware 1020 returns classification data output by the neural network to the dispatch logic 1010, which passes the information to the data cache 1004, or on lines 1011 to threads 1002 that perform base call and quality score computations using the classification data, and can arrange the data in standard formats for base call reads. The outputs of the threads 1002 that perform base calling and quality score computations are provided on lines 1012 to threads 1003 that aggregate the base call reads, perform other operations such as data compression, and write the resulting base call outputs to specified destinations for utilization by the customers.

[0179] In some embodiments, the host can include threads (not shown) that perform final processing of the output of the hardware 1020 in support of the neural network. For example, the hardware 1020 can provide outputs of classification data from a final layer of the multi-cluster neural network. The host processor can execute an output activation function, such as a softmax function, over the classification data to configure the data for use by the base call and quality score threads 1002. Also, the host processor can execute input operations (not shown), such as batch normalization of the tile data prior to input to the hardware 1020.

[0180] Figure 11 is a simplified diagram of a configuration of a configurable processor 846 such as that of Figure 9. In Figure 11, the configurable processor 846 comprises an FPGA with a plurality of high speed PCIe interfaces. The FPGA is configured with a wrapper 1100 which comprises the data flow logic 908 described with reference to Figure 9. The wrapper 1100 manages the interface and coordination with a runtime program in the CPU across the CPU communication link 1109 and manages communication with the on-board DRAM 1102 (*e.g.*, memory 848A) via DRAM communication link 1110. The data flow logic 908 in the wrapper 1100 provides patch data retrieved by traversing the arrays of tile data on the on-board DRAM 1102 for the number N cycles to a cluster 1101, and retrieves process data 1115 from the cluster 1101 for delivery back to the on-board DRAM 1102. The wrapper 1100 also manages transfer of data between the on-board DRAM 1102 and host memory, for both the input arrays of tile data, and for the output patches of classification data. The wrapper transfers patch data on line 1113 to the allocated cluster 1101. The wrapper provides trained parameters, such as weights and biases on line 1112 to the cluster 1101 retrieved from the on-board DRAM 1102. The wrapper provides configuration and control data on line 1111 to the cluster 1101 provided from, or generated in response to, the runtime program on the host via the CPU communication link 1109. The cluster can also provide status signals on line 1116 to the wrapper 1100, which are used in cooperation

with control signals from the host to manage traversal of the arrays of tile data to provide spatially aligned patch data, and to execute the multi-cycle neural network over the patch data using the resources of the cluster 1101.

[0181] As mentioned above, there can be multiple clusters on a single configurable processor managed by the wrapper 1100 configured for executing on corresponding ones of multiple patches of the tile data. Each cluster can be configured to provide classification data for base calls in a subject sensing cycle using the tile data of multiple sensing cycles described herein.

[0182] In examples of the system, model data, including kernel data like filter weights and biases can be sent from the host CPU to the configurable processor, so that the model can be updated as a function of cycle number. A base calling operation can comprise, for a representative example, on the order of hundreds of sensing cycles. Base calling operation can include paired end reads in some embodiments. For example, the model trained parameters may be updated once every 20 cycles (or other number of cycles), or according to update patterns implemented for particular systems and neural network models. In some embodiments including paired end reads in which a sequence for a given string in a genetic cluster on a tile includes a first part extending from a first end down (or up) the string, and a second part extending from a second end up (or down) the string, the trained parameters can be updated on the transition from the first part to the second part.

[0183] In some examples, image data for multiple cycles of sensing data for a tile can be sent from the CPU to the wrapper 1100. The wrapper 1100 can optionally do some pre-processing and transformation of the sensing data and write the information to the on-board DRAM 1102. The input tile data for each sensing cycle can include arrays of sensor data including on the order of 4000 x 3000 pixels per sensing cycle per tile or more, with two features representing colors of two images of the tile, and one or two bytes per feature per pixel. For an embodiment in which the number N is three sensing cycles to be used in each run of the multi-cycle neural network, the array of tile data for each run of the multi-cycle neural network can consume on the order of hundreds of megabytes per tile. In some embodiments of the system, the tile data also includes an array of DFC data, stored once per tile, or other type of metadata about the sensor data and the tiles.

[0184] In operation, when a multi-cycle cluster is available, the wrapper allocates a patch to the cluster. The wrapper fetches a next patch of tile data in the traversal of the tile and sends it to the allocated cluster along with appropriate control and configuration information. The cluster can be configured with enough memory on the configurable processor to hold a patch of data including patches from multiple cycles in some systems, that is being worked on in place, and a

patch of data that is to be worked on when the current patch of processing is finished using a ping-pong buffer technique or raster scanning technique in various embodiments.

[0185] When an allocated cluster completes its run of the neural network for the current patch and produces an output patch, it will signal the wrapper. The wrapper will read the output patch from the allocated cluster, or alternatively the allocated cluster will push the data out to the wrapper. Then the wrapper will assemble output patches for the processed tile in the DRAM 1102. When the processing of the entire tile has been completed, and the output patches of data transferred to the DRAM, the wrapper sends the processed output array for the tile back to the host/CPU in a specified format. In some embodiments, the on-board DRAM 1102 is managed by memory management logic in the wrapper 1100. The runtime program can control the sequencing operations to complete analysis of all the arrays of tile data for all the cycles in the run in a continuous flow to provide real time analysis.

Computer System

[0186] Figure 12 is a computer system 1200 that can be used by the sequencing system 800A to implement the base calling techniques disclosed herein. Computer system 1200 includes at least one central processing unit (CPU) 1272 that communicates with a number of peripheral devices via bus subsystem 1255. These peripheral devices can include a storage subsystem 1210 including, for example, memory devices and a file storage subsystem 1236, user interface input devices 1238, user interface output devices 1276, and a network interface subsystem 1274. The input and output devices allow user interaction with computer system 1200. Network interface subsystem 1274 provides an interface to outside networks, including an interface to corresponding interface devices in other computer systems.

[0187] In one implementation, the system controller 806 is communicably linked to the storage subsystem 1210 and the user interface input devices 1238.

[0188] User interface input devices 1238 can include a keyboard; pointing devices such as a mouse, trackball, touchpad, or graphics tablet; a scanner; a touch screen incorporated into the display; audio input devices such as voice recognition systems and microphones; and other types of input devices. In general, use of the term “input device” is intended to include all possible types of devices and ways to input information into computer system 1200.

[0189] User interface output devices 1276 can include a display subsystem, a printer, a fax machine, or non-visual displays such as audio output devices. The display subsystem can include an LED display, a cathode ray tube (CRT), a flat-panel device such as a liquid crystal display (LCD), a projection device, or some other mechanism for creating a visible image. The display subsystem can also provide a non-visual display such as audio output devices. In general, use of

the term “output device” is intended to include all possible types of devices and ways to output information from computer system 1200 to the user or to another machine or computer system.

[0190] Storage subsystem 1210 stores programming and data constructs that provide the functionality of some or all of the modules and methods described herein. These software modules are generally executed by deep learning processors 1278.

[0191] Deep learning processors 1278 can be graphics processing units (GPUs), field-programmable gate arrays (FPGAs), application-specific integrated circuits (ASICs), and/or coarse-grained reconfigurable architectures (CGRAs). Deep learning processors 1278 can be hosted by a deep learning cloud platform such as Google Cloud Platform™, Xilinx™, and Cirrascale™. Examples of deep learning processors 1278 include Google’s Tensor Processing Unit (TPU)™, rackmount solutions like GX4 Rackmount Series™, GX12 Rackmount Series™, NVIDIA DGX-1™, Microsoft’s Stratix V FPGA™, Graphcore’s Intelligent Processor Unit (IPU)™, Qualcomm’s Zeroth Platform™ with Snapdragon processors™, NVIDIA’s Volta™, NVIDIA’s DRIVE PX™, NVIDIA’s JETSON TX1/TX2 MODULE™, Intel’s Nirvana™, Movidius VPU™, Fujitsu DPI™, ARM’s DynamicIQ™, IBM TrueNorth™, Lambda GPU Server with Testa V100s™, and others.

[0192] Memory subsystem 1222 used in the storage subsystem 1210 can include a number of memories including a main random access memory (RAM) 1232 for storage of instructions and data during program execution and a read only memory (ROM) 1234 in which fixed instructions are stored. A file storage subsystem 1236 can provide persistent storage for program and data files, and can include a hard disk drive, a floppy disk drive along with associated removable media, a CD-ROM drive, an optical drive, or removable media cartridges. The modules implementing the functionality of certain implementations can be stored by file storage subsystem 1236 in the storage subsystem 1210, or in other machines accessible by the processor.

[0193] Bus subsystem 1255 provides a mechanism for letting the various components and subsystems of computer system 1200 communicate with each other as intended. Although bus subsystem 1255 is shown schematically as a single bus, alternative implementations of the bus subsystem can use multiple busses.

[0194] Computer system 1200 itself can be of varying types including a personal computer, a portable computer, a workstation, a computer terminal, a network computer, a television, a mainframe, a server farm, a widely-distributed set of loosely networked computers, or any other data processing system or user device. Due to the ever changing nature of computers and networks, the description of computer system 1200 depicted in Figure 12 is intended only as a specific example for purposes of illustrating the preferred implementations of the present

invention. Many other configurations of computer system 1200 are possible having more or less components than the computer system depicted in Figure 12.

Clauses

[0195] We disclose the following clauses:

1. An artificial intelligence-based system for base calling, the system comprising:
 - a neural network-based base caller that processes at least right flanking, center, and left flanking inputs, and produces at least right flanking, center, and left flanking outputs;
 - wherein the right flanking input comprises current image data for a current sequencing cycle of a sequencing run, supplemented with previous image data for one or more previous sequencing cycles preceding the current sequencing cycle, and wherein the right flanking output comprises a right flanking base call prediction for the current sequencing cycle and base call predictions for the previous sequencing cycles;
 - wherein the center input comprises the current image data, supplemented with the previous image data and successive image data for one or more successive sequencing cycles succeeding the current sequencing cycle, and wherein the center output comprises a center base call prediction for the current sequencing cycle and base call predictions for the previous sequencing cycles and the successive sequencing cycles;
 - wherein the left flanking input comprises the current image data, supplemented with the successive image data, and wherein the left flanking output comprises a left flanking base call prediction for the current sequencing cycle and base call predictions for the successive sequencing cycles; and
 - a base call generator coupled to the neural network-based base caller and configured to generate a base call for the current sequencing cycle based on the right flanking, center, and left flanking base call predictions for the current sequencing cycle.
2. The artificial intelligence-based system of clause 1, wherein the current image data for the current sequencing cycle depicts intensity emissions of analytes and their surrounding background captured at the current sequencing cycle.
3. The artificial intelligence-based system of clause 2, wherein the right flanking, center, and left flanking base call predictions for the current sequencing cycle identify likelihoods of a base incorporated in one or more of the analytes at the current sequencing cycle being A, C, T, and G.
4. The artificial intelligence-based system of clause 3, wherein the base call generator is further configured to comprise an averager that

base-wise sums the likelihoods across the right flanking, center, and left flanking base call predictions for the current sequencing cycle;

determines base-wise averages based on the base-wise sums; and

generates the base call for the current sequencing cycle based on a highest one of the base-wise averages.

5. The artificial intelligence-based system of clause 3, wherein the base call generator is further configured to comprise a consensor that

determines a preliminary base call for each of the right flanking, center, and left flanking base call predictions for the current sequencing cycle based on a highest one of the likelihoods, thereby producing a sequence of preliminary base calls; and

generates the base call for the current sequencing cycle based on a most common base call in the sequence of preliminary base calls.

6. The artificial intelligence-based system of clause 3, wherein the base call generator is further configured to comprise a weighted consensor that

determines a preliminary base call for each of the right flanking, center, and left flanking base call predictions for the current sequencing cycle based on a highest one of the likelihoods, thereby producing a sequence of preliminary base calls;

applies base-wise weights to respective ones of the preliminary base calls in the sequence of preliminary base calls and produces a sequence of weighted preliminary base calls; and

generates the base call for the current sequencing cycle based on a most weighted base call in the sequence of weighted preliminary base calls.

7. The artificial intelligence-based system of clause 3, wherein the likelihoods are exponentially normalized scores produced by a softmax layer.

8. The artificial intelligence-based system of clause 1, further configured to comprise a trainer that, during training,

computes errors between base calls generated by the base call generator for the current sequencing cycle, the previous sequencing cycles, and the successive sequencing cycles based on the right flanking, center, and left flanking outputs of the neural network-based base caller and ground truth base calls;

determines gradients for the current sequencing cycle, the previous sequencing cycles, and the successive sequencing cycles based on the errors; and

updates parameters of the neural network-based base caller by backpropagating the gradients.

9. The artificial intelligence-based system of clause 1, wherein the right flanking base call prediction for the current sequencing cycle accounts for prephasing effect between the current sequencing cycle and the previous sequencing cycles.
10. The artificial intelligence-based system of clause 9, wherein the center base call prediction for the current sequencing cycle accounts for the prephasing effect between the current sequencing cycle and the previous sequencing cycles and phasing effect between the current sequencing cycle and the successive sequencing cycles.
11. The artificial intelligence-based system of clause 10, wherein the left flanking base call prediction for the current sequencing cycle accounts for the phasing effect between the current sequencing cycle and the successive sequencing cycles.
12. An artificial intelligence-based system for base calling, the system comprising:
a host processor;
memory accessible by the host processor storing image data for sequencing cycles of a sequencing run, wherein current image data for a current sequencing cycle of the sequencing run depicts intensity emissions of analytes and their surrounding background captured at the current sequencing cycle; and
a configurable processor having access to the memory, the configurable processor including:
a plurality of execution clusters, the execution clusters in the plurality of execution clusters configured to execute a neural network; and
data flow logic having access to the memory and to the execution clusters in the plurality of execution clusters, configured to provide the current image data, previous image data for one or more previous sequencing cycles preceding the current sequencing cycle, and successive image data for one or more successive sequencing cycles succeeding the current sequencing cycle to available execution clusters in the plurality of execution clusters, cause the execution clusters to apply different groupings of the current image data, the previous image data, and the successive image data to the neural network to produce first, second, and third base call predictions for the current sequencing cycle, and to feedback the first, second, and third base call predictions for the current sequencing cycle to the memory for use in generation of a base call for the current sequencing cycle based on the first, second, and third base call predictions.
13. The artificial intelligence-based system of clause 12, wherein the different groupings include a first grouping comprising the current image data and the previous image data, a second grouping comprising the current image data, the previous image data, and the successive image data, and a third grouping comprising the current image data and the successive image data.

14. The artificial intelligence-based system of clause 13, wherein the execution clusters apply the first grouping to the neural network to produce the first base call prediction, the second grouping to the neural network to produce the second base call prediction, and the third grouping to the neural network to produce the third base call prediction.

15. The artificial intelligence-based system of clause 12, wherein the first, second, and third base call predictions for the current sequencing cycle identify likelihoods of a base incorporated in one or more of the analytes at the current sequencing cycle being A, C, T, and G.

16. The artificial intelligence-based system of clause 15, wherein the data flow logic is further configured to generate the base call for the current sequencing cycle by

- base-wise summing the likelihoods across the first, second, and third base call predictions for the current sequencing cycle;
- determining base-wise averages based on the base-wise summing; and
- generating the base call for the current sequencing cycle based on a highest one of the base-wise averages.

17. The artificial intelligence-based system of clause 15, wherein the data flow logic is further configured to generate the base call for the current sequencing cycle by

- determining a preliminary base call for each of the first, second, and third base call predictions for the current sequencing cycle based on a highest one of the likelihoods, thereby producing a sequence of preliminary base calls; and
- generating the base call for the current sequencing cycle based on a most common base call in the sequence of preliminary base calls.

18. The artificial intelligence-based system of clause 15, wherein the data flow logic is further configured to generate the base call for the current sequencing cycle by

- determining a preliminary base call for each of the first, second, and third base call predictions for the current sequencing cycle based on a highest one of the likelihoods, thereby producing a sequence of preliminary base calls;
- applying base-wise weights to respective ones of the preliminary base calls in the sequence of preliminary base calls and producing a sequence of weighted preliminary base calls; and
- generating the base call for the current sequencing cycle based on a most weighted base call in the sequence of weighted preliminary base calls.

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

- accessing current image data for a current sequencing cycle of a sequencing run, previous

image data for one or more previous sequencing cycles preceding the current sequencing cycle, and successive image data for one or more successive sequencing cycles succeeding the current sequencing cycle;

processing different groupings of the current image data, the previous image data, and the successive image data through a neural network-based base caller and producing first, second, and third base call predictions for the current sequencing cycle; and

generating a base call for the current sequencing cycle based on the first, second, and third base call predictions.

20. The artificial intelligence-based method of clause 19, wherein the different groupings include
a first grouping comprising the current image data and the previous image data,
a second grouping comprising the current image data, the previous image data, and the successive image data, and
a third grouping comprising the current image data and the successive image data.

21. The artificial intelligence-based method of clause 20, further including:
processing the first grouping through the neural network-based base caller to produce the first base call prediction,
processing the second grouping through the neural network-based base caller to produce the second base call prediction, and
processing the third grouping through the neural network-based base caller to produce the third base call prediction.

22. The artificial intelligence-based method of clause 19, wherein the first, second, and third base call predictions for the current sequencing cycle identify likelihoods of a base incorporated in one or more of the analytes at the current sequencing cycle being A, C, T, and G.

23. The artificial intelligence-based method of clause 22, further including generating the base call for the current sequencing cycle by
base-wise summing the likelihoods across the first, second, and third base call predictions for the current sequencing cycle;
determining base-wise averages based on the base-wise summing; and
generating the base call for the current sequencing cycle based on a highest one of the base-wise averages.

24. The artificial intelligence-based method of clause 22, further including generating the base call for the current sequencing cycle by

determining a preliminary base call for each of the first, second, and third base call predictions for the current sequencing cycle based on a highest one of the likelihoods, thereby producing a sequence of preliminary base calls; and

generating the base call for the current sequencing cycle based on a most common base call in the sequence of preliminary base calls.

25. The artificial intelligence-based method of clause 22, further including generating the base call for the current sequencing cycle by

determining a preliminary base call for each of the first, second, and third base call predictions for the current sequencing cycle based on a highest one of the likelihoods, thereby producing a sequence of preliminary base calls;

applying base-wise weights to respective ones of the preliminary base calls in the sequence of preliminary base calls and producing a sequence of weighted preliminary base calls; and

generating the base call for the current sequencing cycle based on a most weighted base call in the sequence of weighted preliminary base calls.

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

processing at least right flanking, center, and left flanking inputs through a neural network-based base caller, and producing at least right flanking, center, and left flanking outputs;

wherein the right flanking input comprises current image data for a current sequencing cycle of a sequencing run, supplemented with previous image data for one or more previous sequencing cycles preceding the current sequencing cycle, and wherein the right flanking output comprises a right flanking base call prediction for the current sequencing cycle and base call predictions for the previous sequencing cycles;

wherein the center input comprises the current image data, supplemented with the previous image data and successive image data for one or more successive sequencing cycles succeeding the current sequencing cycle, and wherein the center output comprises a center base call prediction for the current sequencing cycle and base call predictions for the previous sequencing cycles and the successive sequencing cycles;

wherein the left flanking input comprises the current image data, supplemented with the successive image data, and wherein the left flanking output comprises a left flanking base call prediction for the current sequencing cycle and base call predictions for the successive sequencing cycles; and

generating a base call for the current sequencing cycle based on the right flanking, center, and left flanking base call predictions for the current sequencing cycle.

27. The artificial intelligence-based method of clause 26, wherein the current image data for the current sequencing cycle depicts intensity emissions of analytes and their surrounding background captured at the current sequencing cycle.

28. The artificial intelligence-based method of clause 26, wherein the right flanking, center, and left flanking base call predictions for the current sequencing cycle identify likelihoods of a base incorporated in one or more of the analytes at the current sequencing cycle being A, C, T, and G.

29. The artificial intelligence-based method of clause 28, further including generating the base call for the current sequencing cycle by

base-wise summing the likelihoods across the right flanking, center, and left flanking base call predictions for the current sequencing cycle;

determining base-wise averages based on the base-wise summing; and

generating the base call for the current sequencing cycle based on a highest one of the base-wise averages.

30. The artificial intelligence-based method of clause 28, further including generating the base call for the current sequencing cycle by

determining a preliminary base call for each of the right flanking, center, and left flanking base call predictions for the current sequencing cycle based on a highest one of the likelihoods, thereby producing a sequence of preliminary base calls; and

generating the base call for the current sequencing cycle based on a most common base call in the sequence of preliminary base calls.

31. The artificial intelligence-based method of clause 28, further including generating the base call for the current sequencing cycle by

determining a preliminary base call for each of the right flanking, center, and left flanking base call predictions for the current sequencing cycle based on a highest one of the likelihoods, thereby producing a sequence of preliminary base calls;

applying base-wise weights to respective ones of the preliminary base calls in the sequence of preliminary base calls and producing a sequence of weighted preliminary base calls; and

generating the base call for the current sequencing cycle based on a most weighted base call in the sequence of weighted preliminary base calls.

32. The artificial intelligence-based method of clause 28, wherein the likelihoods are exponentially normalized scores produced by a softmax layer.

33. The artificial intelligence-based method of clause 26, further including, during training,
- computing errors between base calls generated by the base call generator for the current sequencing cycle, the previous sequencing cycles, and the successive sequencing cycles based on the right flanking, center, and left flanking outputs of the neural network-based base caller and ground truth base calls;
 - determining gradients for the current sequencing cycle, the previous sequencing cycles, and the successive sequencing cycles based on the errors; and
 - updating parameters of the neural network-based base caller by backpropagating the gradients.
34. The artificial intelligence-based method of clause 26, wherein the right flanking base call prediction for the current sequencing cycle accounts for prephasing effect between the current sequencing cycle and the previous sequencing cycles.
35. The artificial intelligence-based method of clause 34, wherein the center base call prediction for the current sequencing cycle accounts for the prephasing effect between the current sequencing cycle and the previous sequencing cycles and phasing effect between the current sequencing cycle and the successive sequencing cycles.
36. The artificial intelligence-based method of clause 35, wherein the left flanking base call prediction for the current sequencing cycle accounts for the phasing effect between the current sequencing cycle and the successive sequencing cycles.
37. An artificial intelligence-based method of base calling, the method including:
- processing at least first, second, and third inputs through a neural network-based base caller, and producing at least first, second, and third outputs;
 - wherein the first input comprises particular image data for a particular sequencing cycle of a sequencing run, supplemented with previous image data for one or more previous sequencing cycles preceding the particular sequencing cycle, and wherein the first output comprises a first base call prediction for the particular sequencing cycle and base call predictions for the previous sequencing cycles;
 - wherein the second input comprises the particular image data, supplemented with the previous image data and successive image data for one or more successive sequencing cycles succeeding the particular sequencing cycle, and wherein the second output comprises a second base call prediction for the particular sequencing cycle and base call predictions for the previous sequencing cycles and the successive sequencing cycles;
 - wherein the third input comprises the particular image data, supplemented with the

successive image data, and wherein the third output comprises a third base call prediction for the particular sequencing cycle and base call predictions for the successive sequencing cycles; and

generating a base call for the particular sequencing cycle based on the first, second, and third base call predictions for the particular sequencing cycle.

38. The artificial intelligence-based method of clause 37, implementing each of the clauses which ultimately depend from clause 1.

39. A non-transitory computer readable storage medium impressed with computer program instructions to perform artificial intelligence-based base calling, the instructions, when executed on a processor, implement a method comprising:

accessing current image data for a current sequencing cycle of a sequencing run, previous image data for one or more previous sequencing cycles preceding the current sequencing cycle, and successive image data for one or more successive sequencing cycles succeeding the current sequencing cycle;

processing different groupings of the current image data, the previous image data, and the successive image data through a neural network-based base caller and producing first, second, and third base call predictions for the current sequencing cycle; and

generating a base call for the current sequencing cycle based on the first, second, and third base call predictions.

40. The non-transitory computer readable storage medium of clause 39, implementing each of the clauses which ultimately depend from clause 1.

41. A non-transitory computer readable storage medium impressed with computer program instructions to perform artificial intelligence-based base calling, the instructions, when executed on a processor, implement a method comprising:

processing at least first, second, and left inputs through a neural network-based base caller, and producing at least first, second, and left outputs;

wherein the first input comprises particular image data for a particular sequencing cycle of a sequencing run, supplemented with previous image data for one or more previous sequencing cycles preceding the particular sequencing cycle, and wherein the first output comprises a first base call prediction for the particular sequencing cycle and base call predictions for the previous sequencing cycles;

wherein the second input comprises the particular image data, supplemented with the previous image data and successive image data for one or more successive sequencing cycles succeeding the particular sequencing cycle, and wherein the second output comprises a second

base call prediction for the particular sequencing cycle and base call predictions for the previous sequencing cycles and the successive sequencing cycles;

wherein the left input comprises the particular image data, supplemented with the successive image data, and wherein the left output comprises a left base call prediction for the particular sequencing cycle and base call predictions for the successive sequencing cycles; and

generating a base call for the particular sequencing cycle based on the first, second, and left base call predictions for the particular sequencing cycle.

44. The non-transitory computer readable storage medium of clause 43, implementing each of the clauses which ultimately depend from clause 1.

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

accessing a progression of per-cycle analyte channel sets generated for sequencing cycles of a sequencing run;

processing, through a neural network-based base caller, windows of per-cycle analyte channel sets in the progression for the windows of sequencing cycles of the sequencing run such that

the neural network-based base caller

processes a subject window of per-cycle analyte channel sets in the progression for the subject window of sequencing cycles of the sequencing run and

generates provisional base call predictions for three or more sequencing cycles in the subject window of sequencing cycles;

from multiple windows in which a particular sequencing cycle appeared at different positions, using the neural network-based base caller to generate provisional base call predictions for the particular sequencing cycle; and

determining a base call for the particular sequencing cycle based on the plurality of base call predictions.

46. The artificial intelligence-based method of clause 45, implementing each of the clauses which ultimately depend from clause 1.

47. A system including one or more processors coupled to memory, the memory loaded with computer instructions to perform artificial intelligence-based base calling, the instructions, when executed on the processors, implement actions comprising:

accessing a progression of per-cycle analyte channel sets generated for sequencing cycles of a sequencing run;

processing, through a neural network-based base caller, windows of per-cycle analyte

channel sets in the progression for the windows of sequencing cycles of the sequencing run such that

the neural network-based base caller

processes a subject window of per-cycle analyte channel sets in the progression for the subject window of sequencing cycles of the sequencing run and

generates provisional base call predictions for three or more sequencing cycles in the subject window of sequencing cycles;

from multiple windows in which a particular sequencing cycle appeared at different positions, using the neural network-based base caller to generate provisional base call predictions for the particular sequencing cycle; and

determining a base call for the particular sequencing cycle based on the plurality of base call predictions.

48. The system of clause 47, implementing each of the clauses which ultimately depend from clause 1.

49. A non-transitory computer readable storage medium impressed with computer program instructions to perform artificial intelligence-based base calling, the instructions, when executed on a processor, implement a method comprising:

accessing a progression of per-cycle analyte channel sets generated for sequencing cycles of a sequencing run;

processing, through a neural network-based base caller, windows of per-cycle analyte channel sets in the progression for the windows of sequencing cycles of the sequencing run such that

the neural network-based base caller

processes a subject window of per-cycle analyte channel sets in the progression for the subject window of sequencing cycles of the sequencing run and

generates provisional base call predictions for three or more sequencing cycles in the subject window of sequencing cycles;

from multiple windows in which a particular sequencing cycle appeared at different positions, using the neural network-based base caller to generate provisional base call predictions for the particular sequencing cycle; and

determining a base call for the particular sequencing cycle based on the plurality of base call predictions.

50. The non-transitory computer readable storage medium of clause 49, implementing each of the clauses which ultimately depend from clause 1.

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

accessing a series of per-cycle analyte channel sets generated for sequencing cycles of a sequencing run;

processing, through a neural network-based base caller, windows of per-cycle analyte channel sets in the series for the windows of sequencing cycles of the sequencing run such that the neural network-based base caller

processes a subject window of per-cycle analyte channel sets in the series for the subject window of sequencing cycles of the sequencing run and

generates base call predictions for two or more sequencing cycles in the subject window of sequencing cycles;

processing, through the neural network-based base caller,

a plurality of the windows of per-cycle analyte channel sets in the series for the plurality of the windows of sequencing cycles of the sequencing run and

generating an output for each window in the plurality of the windows,

wherein each window in the plurality of the windows includes a particular per-cycle analyte channel set for a particular sequencing cycle of the sequencing run, and

wherein the output for each window in the plurality of the windows includes

(i) a base call prediction for the particular sequencing cycle and

(ii) one or more additional base call predictions for one or more additional sequencing cycles of the sequencing run, thereby generating a plurality of base call predictions for the particular sequencing cycle across the plurality of the windows; and

determining a base call for the particular sequencing cycle based on the plurality of base call predictions.

52. A system including one or more processors coupled to memory, the memory loaded with computer instructions to perform artificial intelligence-based base calling, the instructions, when executed on the processors, implement actions comprising:

accessing a series of per-cycle analyte channel sets generated for sequencing cycles of a sequencing run;

processing, through a neural network-based base caller, windows of per-cycle analyte channel sets in the series for the windows of sequencing cycles of the sequencing run such that the neural network-based base caller

processes a subject window of per-cycle analyte channel sets in the series for the subject window of sequencing cycles of the sequencing run and
generates base call predictions for two or more sequencing cycles in the subject window of sequencing cycles;
processing, through the neural network-based base caller,
a plurality of the windows of per-cycle analyte channel sets in the series for the plurality of the windows of sequencing cycles of the sequencing run and
generating an output for each window in the plurality of the windows,
wherein each window in the plurality of the windows includes a particular per-cycle analyte channel set for a particular sequencing cycle of the sequencing run, and
wherein the output for each window in the plurality of the windows includes
(i) a base call prediction for the particular sequencing cycle and
(ii) one or more additional base call predictions for one or more additional sequencing cycles of the sequencing run, thereby generating a plurality of base call predictions for the particular sequencing cycle across the plurality of the windows; and
determining a base call for the particular sequencing cycle based on the plurality of base call predictions.

53. The system of clause 52, implementing each of the clauses which ultimately depend from clause 1.

54. A non-transitory computer readable storage medium impressed with computer program instructions to perform artificial intelligence-based base calling, the instructions, when executed on a processor, implement a method comprising:

accessing a series of per-cycle analyte channel sets generated for sequencing cycles of a sequencing run;
processing, through a neural network-based base caller, windows of per-cycle analyte channel sets in the series for the windows of sequencing cycles of the sequencing run such that
the neural network-based base caller
processes a subject window of per-cycle analyte channel sets in the series for the subject window of sequencing cycles of the sequencing run and
generates base call predictions for two or more sequencing cycles in the subject window of sequencing cycles;
processing, through the neural network-based base caller,
a plurality of the windows of per-cycle analyte channel sets in the series for the plurality of the windows of sequencing cycles of the sequencing run and

generating an output for each window in the plurality of the windows,
wherein each window in the plurality of the windows includes a particular per-cycle analyte channel set for a particular sequencing cycle of the sequencing run, and
wherein the output for each window in the plurality of the windows includes
(i) a base call prediction for the particular sequencing cycle and
(ii) one or more additional base call predictions for one or more additional sequencing cycles of the sequencing run, thereby generating a plurality of base call predictions for the particular sequencing cycle across the plurality of the windows; and
determining a base call for the particular sequencing cycle based on the plurality of base call predictions.

55. The non-transitory computer readable storage medium of clause 54, implementing each of the clauses which ultimately depend from clause 1.

[0196] Other implementations of the method described above can include a non-transitory computer readable storage medium storing instructions executable by a processor to perform any of the methods described above. Yet another implementation of the method described in this section can include a system including memory and one or more processors operable to execute instructions, stored in the memory, to perform any of the methods described above.

[0197] What is claimed is:

CLAIMS

Claim 1. An artificial intelligence-based system for base calling, the system comprising:

a neural network-based base caller including parallel convolutional neural network pipelines respectively trained to process at least right flanking, center, and left flanking inputs of images depicting clusters and their surrounding backgrounds, and produce at least right flanking, center, and left flanking outputs for the clusters in a current sequencing cycle; wherein the right flanking input comprises current image data for the current sequencing cycle of a sequencing run, supplemented with previous image data for one or more previous sequencing cycles preceding the current sequencing cycle, and wherein the right flanking output comprises a right flanking base call prediction for the current sequencing cycle and base call predictions for the previous sequencing cycles;

wherein the center input comprises the current image data, supplemented with the previous image data and successive image data for one or more successive sequencing cycles succeeding the current sequencing cycle, and wherein the center output comprises a center base call prediction for the current sequencing cycle and base call predictions for the previous sequencing cycles and the successive sequencing cycles;

wherein the left flanking input comprises the current image data, supplemented with the successive image data, and wherein the left flanking output comprises a left flanking base call prediction for the current sequencing cycle and base call predictions for the successive sequencing cycles;

wherein training of the trained parallel convolutional neural network pipelines comprises providing a training dataset of the right flanking, center, and left flanking inputs of images depicting clusters and their surrounding backgrounds together with ground truth base calls of clusters in the current sequencing cycle and applying backpropagation-based gradient updates; and

a base call generator coupled to the neural network-based base caller and configured to generate a base call for the clusters in the current sequencing cycle based on the right flanking, center, and left flanking base call predictions for the current sequencing cycle.

Claim 2. The artificial intelligence-based system of claim 1, wherein the current image data for the current sequencing cycle depicts intensity emissions of analytes and their surrounding background captured at the current sequencing cycle.

Claim 3. The artificial intelligence-based system of claims 1 or 2, wherein the right flanking, center, and left flanking base call predictions for the current sequencing cycle identify likelihoods of a base incorporated in one or more analytes at the current sequencing cycle being A, C, T, and G.

Claim 4. The artificial intelligence-based system of claim 3, wherein the base call generator is further configured to comprise an averager that

base-wise sums the likelihoods across the right flanking, center, and left flanking base call predictions applicable to clusters in the current sequencing cycle;
determines base-wise averages based on the base-wise sums; and
generates the base call for the current sequencing cycle based on a highest one of the base-wise averages.

Claim 5. The artificial intelligence-based system of claim 3 or claim 4, wherein the base call generator is further configured to comprise a consensor that

determines a preliminary base call for each of the right flanking, center, and left flanking base call predictions applicable to clusters in the current sequencing cycle based on a highest one of the likelihoods, thereby producing a sequence of preliminary base calls; and
generates the base call applicable to the clusters for the current sequencing cycle based on a most common base call in the sequence of preliminary base calls.

Claim 6. The artificial intelligence-based system of any of claims 3-5, wherein the base call generator is further configured to comprise a weighted consensor that

determines a preliminary base call for each of the right flanking, center, and left flanking base call predictions for the current sequencing cycle based on a highest one of the likelihoods, thereby producing a sequence of preliminary base calls;
applies base-wise weights to respective ones of the preliminary base calls in the sequence of preliminary base calls and produces a sequence of weighted preliminary base calls; and

generates the base call for the current sequencing cycle based on a most weighted base call in the sequence of weighted preliminary base calls.

Claim 7. The artificial intelligence-based system of any of claims 3-6, wherein the likelihoods are exponentially normalized scores produced by a softmax layer.

Claim 8. The artificial intelligence-based system of any of claims 1-7, further configured to comprise a trainer that, during training,

computes errors between base calls generated by the base call generator for the current sequencing cycle, the previous sequencing cycles, and the successive sequencing cycles based on the right flanking, center, and left flanking outputs of the neural network-based base caller and ground truth base calls;

determines gradients for the current sequencing cycle, the previous sequencing cycles, and the successive sequencing cycles based on the errors; and

updates parameters of the neural network-based base caller by backpropagating the gradients.

Claim 9. The artificial intelligence-based system of any of claims 1-8, wherein the right flanking base call prediction for the current sequencing cycle accounts for a prephasing effect between the current sequencing cycle and the previous sequencing cycles.

Claim 10. The artificial intelligence-based system of claim 9, wherein the center base call prediction for the current sequencing cycle accounts for the prephasing effect between the current sequencing cycle and the previous sequencing cycles and a phasing effect between the current sequencing cycle and the successive sequencing cycles.

Claim 11. The artificial intelligence-based system of claim 10, wherein the left flanking base call prediction for the current sequencing cycle accounts for the phasing effect between the current sequencing cycle and the successive sequencing cycles.

Claim 12. An artificial intelligence-based system for base calling, the system comprising:

a host processor;

memory accessible by the host processor storing image data depicting clusters and their

surrounding backgrounds for sequencing cycles of a sequencing run, wherein current image

data for a current sequencing cycle of the sequencing run depicts intensity emissions of analytes and their surrounding background captured at the current sequencing cycle; and a configurable processor having access to the memory, the configurable processor including: a plurality of execution clusters, the execution clusters in the plurality of execution clusters configured to execute a neural network; and data flow logic having access to the memory and to the execution clusters in the plurality of execution clusters, configured to provide the current image data, previous image data for one or more previous sequencing cycles preceding the current sequencing cycle, and successive image data for one or more successive sequencing cycles succeeding the current sequencing cycle to available execution clusters in the plurality of execution clusters, and cause the execution clusters to apply different groupings of the current image data, the previous image data, and the successive image data, the groupings comprising windows of cycles with the current sequencing cycle appearing in different positions in the windows, to parallel convolutional neural network pipelines in the neural network that have been trained using a training dataset of the windows of cycles with the current sequencing cycle appearing in different positions in the windows together with ground truth base calls of clusters in the current sequencing cycle and applying backpropagation-based gradient updates, to produce from the groupings first, second, and third base call predictions applicable to the clusters for the current sequencing cycle, and to feed the first, second, and third base call predictions for the current sequencing cycle to the memory for use in generation of a base call applicable to the cluster for the current sequencing cycle based on the first, second, and third base call predictions.

Claim 13. The artificial intelligence-based system of claim 12, wherein the different groupings include a first grouping comprising the current image data and the previous image data, a second grouping comprising the current image data, the previous image data, and the successive image data, and a third grouping comprising the current image data and the successive image data.

Claim 14. The artificial intelligence-based system of claim 13, wherein the execution clusters apply the first grouping to the neural network to produce the first base call prediction, the second grouping to the neural network to produce the second base call prediction, and the third grouping to the neural network to produce the third base call prediction.

Claim 15. The artificial intelligence-based system of any of claims 12-14, wherein the first, second, and third base call predictions for the current sequencing cycle identify likelihoods of a base incorporated in one or more of the analytes at the current sequencing cycle being A, C, T, and G.

Claim 16. The artificial intelligence-based system of claim 15, wherein the data flow logic is further configured to generate the base call for the current sequencing cycle by
base-wise summing the likelihoods across the first, second, and third base call predictions for the current sequencing cycle;
determining base-wise averages based on the base-wise summing; and
generating the base call for the current sequencing cycle based on a highest one of the base-wise averages.

Claim 17. An artificial intelligence-based method of base calling, the method including:
accessing current image data depicting clusters and their surrounding backgrounds for a current sequencing cycle of a sequencing run, previous image data for one or more previous sequencing cycles preceding the current sequencing cycle, and successive image data for one or more successive sequencing cycles succeeding the current sequencing cycle;
processing different groupings of the current image data, the previous image data, and the successive image data through a neural network-based base caller including trained parallel convolutional neural network pipelines respectively, the groupings comprising of windows of cycles with the current sequencing cycle appearing in different positions in the windows, and
producing a plurality of base call predictions applicable to the clusters for the current sequencing cycle;
wherein training of the trained parallel convolutional neural network pipelines comprises
providing a training dataset of right flanking, center, and left flanking inputs of images depicting clusters and their surrounding backgrounds together with ground truth base calls of

clusters in the current sequencing cycle and applying backpropagation-based gradient updates, wherein the right flanking, center, and left flanking inputs correspond to successive, current, and previous inputs; and
generating base calls applicable to clusters for the current sequencing cycle based on the plurality of base call predictions for the clusters.

Claim 18. An artificial intelligence-based method of base calling, the method including:
generating respective provisional base calls for images of clusters and their surrounding backgrounds in a particular sequencing cycle in response to executing respective iterations of a base caller;
wherein the respective iterations process respective input sets for respective windows of sequencing cycles with the particular sequencing cycle appearing in different positions in the respective windows; and
wherein the respective windows of sequencing cycles have the particular sequencing cycle as at least one overlapping cycle, and one or more non-overlapping cycles; and
generating a final base call for the particular sequencing cycle based on the respective provisional base calls.