CLAIMS

What is claimed is:

[Claim 1]. A method for improving the accuracy of a variant call by accounting for indications of correlated error events, the method comprising:

accessing, by one or more computers and from one or more memory devices, a pileup of a plurality of sequence reads aligned to a first region of a reference genome;

obtaining, by the one or more computers, information describing characteristics of each of the plurality of reads of the pileup corresponding to a first position of the reference genome, wherein the information describing the characteristics of the respective reads include information describing:

(i) a mapping quality score for each sequence read in the pileup at the first position,

(ii) a read-allele score for each sequence read in the pileup at the first position for each candidate allele at the first position,

(iii) a read orientation for each sequence read of the pileup at the first position,

(iv) a position of each base at the first position within each sequence read of the pileup at the first position with reference to the 5′ end of the sequence read,

(v) a read-allele score for each sequence read of the plurality of sequence reads for each candidate allele at the reference position, and

(vi) a base quality score for each read for the base at the first position;

providing, by the one or more computers and based on the obtained information, one or more inputs to a probability model describing the characteristics of the plurality of reads of the pileup, wherein the probability model is configured to determine a score, for each hypothesis of one or more hypotheses selected based on the one or more inputs, that indicates whether the hypothesis is true;

obtaining, by the one or more computers, output information for each of the one or more hypotheses, wherein the output information for each of the one or more hypotheses is (i) generated by the probability model based on the probability model's processing of the one or more inputs to the probability model describing the one or more characteristics of the respective reads of the pileup and (ii) indicative of a score that indicates whether the hypothesis is true, wherein the output information includes:

first output information for a first hypothesis of the one or more hypotheses that includes a likelihood that the sequence reads at the first position indicate the occurrence of a homozygous reference with a foreign allele that matches an alternative allele (alt);

second output information for a second hypothesis of the one or more hypotheses that includes a likelihood that the sequence reads at the first position indicate the occurrence of a homozygous alt with a foreign allele that matches a reference allele,

third output information for a third hypothesis of the one or more hypotheses that includes a likelihood that the sequence reads at the first position indicate the occurrence of a homozygous reference with a sequencing error that matches the alt allele, and

fourth output information for a fourth hypothesis of the one or more hypotheses that includes a likelihood that the sequence reads at the first position indicate the occurrence of a homozygous alt with a sequencing error that matches the reference allele; and

determining, by the one or more computers and based on the obtained output information generated by the probability model for each of the plurality of hypotheses, a likelihood that a true variant exists at the first position.

[Claim 2]. The method of claim 1, wherein determining, by the one or more computers and based on the obtained output information generated by the probability model for each of the plurality of hypotheses, a likelihood that a true variant exists at the first position comprises:

determining, by the one or more computers, a collective score that is based on the output information generated by the probability model for each of the plurality of hypothesis, wherein the collective score indicates a likelihood that the true variant exists;

determining, by the one or more computers, whether the score generated by the collective score satisfies a predetermined threshold; and

based on determining, by the one or more computers, that the collective score satisfies the predetermined threshold, adding information to a VCF file indicating that a true variant exists at the first position, optionally wherein the information indicating that a true variant exists at the first position includes information identifying (i) the first position, (ii) a candidate alt allele at the first position, and (iii) the collective score.

[Claim 3]. The method of claim 1, wherein the read-allele score for each sequence read of the pileup at the first position is based on an output produced by a pair Hidden Markov Model (P-HMM) model, for each of the sequence reads at the first position, which indicates a probability of observing a sequence read r_i given a particular candidate allele $G_{m,\phi}$.

[Claim 4]. The method of claim 1, wherein the one or more memory devices received the pileup of aligned sequence reads from a Field Programmable Gate Array (FPGA) device, wherein the FPGA includes one or more configurable digital logic gates that have been configured as a mapping and alignment unit to perform read mapping and alignment.

[Claim 5]. The method of claim 4,

wherein the computer is configured to access the one or more memory devices using one or more wired or wireless networks,

wherein a Field Programmable Gate Array (FPGA) device and the one or more memory devices are housed in an expansion card that has been coupled to a circuit board of a sequencer,

wherein the sequencer is configured to generate sequence reads based on an input sample and store the generated sequence reads in the one or more memory devices, and

wherein the mapping and alignment unit of the FPGA is configured to access the one or more memory devices to obtain the generated sequence reads; or

wherein the computer and the sequencer are each configured to access the one or more memory devices using one or more wired or wireless networks,

wherein the Field Programmable Gate Array (FPGA) device and the one or more memory devices are housed in an expansion card that has been coupled to a circuit board of server that is located remotely from the computer and the sequencer,

wherein the sequencer is configured to generate sequence reads based on an input sample, provide the generated sequence reads to the server using the one or more wired or wireless networks for storage, of the generated sequence reads, in the one or more memory devices, and

wherein the mapping and alignment unit of the FPGA is configured to access the one or more memory devices to obtain the generated sequence reads.

[Claim 6]. A system comprising:

one or more computers and one or more storage devices storing instructions that are operable, when executed by the one or more computers, to cause the one or more computers to perform operations comprising:

accessing, by one or more computers and from one or more memory devices, a pileup of a plurality of sequence reads aligned to a first region of a reference genome;

obtaining, by the one or more computers, information describing one or more characteristics of each of the plurality of reads of the pileup corresponding to a first position of the reference genome, wherein the information describing the characteristics of the respective reads include information describing:

(i) a mapping quality score for each sequence read in the pileup at the first position and

(ii) a read-allele score for each sequence read in the pileup at the first position for each candidate allele at the first position,

(i) a read orientation for each sequence read of the pileup at the first position,

(ii) a position of each base at the first position within each sequence read of the pileup at the first position with reference to the 5′ end of the sequence read,

(iii) a read-allele score for each sequence read of the plurality of sequence reads for each candidate allele at the reference position, and

(iv) a base quality score for each read for the base at the first position;

providing, by the one or more computers and based on the obtained information, one or more inputs to a probability model describing the characteristics of the plurality of reads of the pileup, wherein the probability model is configured to determine a score, for each hypothesis of one or more hypotheses selected based on the one or more inputs, that indicates whether the hypothesis is true;

obtaining, by the one or more computers, output information for each of the one or more hypotheses, wherein the output information for each of the one or more hypotheses is (i) generated by the probability model based on the probability model's processing of the one or more inputs to the probability model describing the one or more characteristics of the respective reads of the pileup and (ii) indicative of a score that indicates whether the hypothesis is true, wherein the output information includes:

first output information for a first hypothesis of the one or more hypotheses that includes a likelihood that the sequence reads at the first position indicate the occurrence of a homozygous reference with a foreign allele that matches an alternative allele (alt);

second output information for a second hypothesis of the one or more hypotheses that includes a likelihood that the sequence reads at the first position indicate the occurrence of a homozygous alt with a foreign allele that matches a reference allele,

third output information for a third hypothesis of the one or more hypotheses that includes a likelihood that the sequence reads at the first position indicate the occurrence of a homozygous reference with a sequencing error that matches the alt allele, and

fourth output information for a fourth hypothesis of the one or more hypotheses that includes a likelihood that the sequence reads at the first position indicate the occurrence of a homozygous alt with a sequencing error that matches the reference allele; and

determining, by the one or more computers and based on the obtained output information generated by the probability model for each of the plurality of hypotheses, a likelihood that a true variant exists at the first position.

[Claim 7]. The system of claim 6, wherein determining, by the one or more computers and based on the obtained output information generated by the probability model for each of the plurality of hypotheses, a likelihood that a true variant exists at the first position comprises:

determining, by the one or more computers, a collective score that is based on the output information generated by the probability model for each of the plurality of hypothesis, wherein the collective score indicates a likelihood that the true variant exists;

determining, by the one or more computers, whether the score generated by the collective score satisfies a predetermined threshold; and

based on determining, by the one or more computers, that the collective score satisfies the predetermined threshold, adding information to a Variant Call Format (VCF) file indicating that a true variant exists at the first position, optionally wherein the information indicating that a true

variant exists at the first position includes information identifying (i) the first position, (ii) a candidate alt allele at the first position, (iii) the collective score.

[Claim 8]. The system of claim 6, wherein the read-allele score for each sequence read of the pileup at the first position is based on an output produced by a pair Hidden Markov Model (P-HMM) model, for each of the sequence reads at the first position, which indicates a probability of observing a sequence read r_i given a particular candidate allele $G_{m,\phi}$.

[Claim 9]. The system of claim 6, wherein the one or more memory devices received the pileup of aligned sequence reads from a Field Programmable Gate Array (FPGA) device, wherein the FPGA includes one or more configurable digital logic gates that have been configured as a mapping and alignment unit to perform read mapping and alignment.

[Claim 10]. The system of claim 9,

wherein the computer is configured to access the one or more memory devices using one or more wired or wireless networks,

wherein a Field Programmable Gate Array (FPGA) device and the one or more memory devices are housed in an expansion card that has been coupled to a circuit board of a sequencer,

wherein the sequencer is configured to generate sequence reads based on an input sample and store the generated sequence reads in the one or more memory devices, and

wherein the mapping and alignment unit of the FPGA is configured to access the one or more memory devices to obtain the generated sequence reads; or

wherein the computer and the sequencer are each configured to access the one or more memory devices using one or more wired or wireless networks,

wherein the Field Programmable Gate Array (FPGA) device and the one or more memory devices are housed in an expansion card that has been coupled to a circuit board of server that is located remotely from the computer and the sequencer,

wherein the sequencer is configured to generate sequence reads based on an input sample, provide the generated sequence reads to the server using the one or more wired or wireless networks for storage, of the generated sequence reads, in the one or more memory devices, and

wherein the mapping and alignment unit of the FPGA is configured to access the one or more memory devices to obtain the generated sequence reads.

[Claim 11]. A non-transitory computer-readable medium storing software comprising instructions executable by one or more computers which, upon such execution, cause the one or more computers to perform operations comprising:

accessing, by the one or more computers and from one or more memory devices, a pileup of a plurality of sequence reads aligned to a first region of a reference genome;

obtaining, by the one or more computers, information describing one or more characteristics of each of the plurality of reads of the pileup corresponding to a first position of the reference genome, wherein the information describing the characteristics of the respective reads include information describing:

(i) a mapping quality score for each sequence read in the pileup at the first position,

(ii) a read-allele score for each sequence read in the pileup at the first position for each candidate allele at the first position,

(iii) a read orientation for each sequence read of the pileup at the first position,

(iv) a position of each base at the first position within each sequence read of the pileup at the first position with reference to the 5´ end of the sequence read,

(v) a read-allele score for each sequence read of the plurality of sequence reads for each candidate allele at the reference position, and

(vi) a base quality score for each read for the base at the first position;

providing, by the one or more computers and based on the obtained information, one or more inputs to a probability model describing the characteristics of the plurality of reads of the pileup, wherein the probability model is configured to determine a score, for each hypothesis of one or more hypotheses selected based on the one or more inputs, that indicates whether the hypothesis is true;

obtaining, by the one or more computers, output information for each of the one or more hypotheses, wherein the output information for each of the one or more hypotheses is (i) generated by the probability model based on the probability model's processing of the one or more inputs to the probability model describing the one or more characteristics of the respective reads of the pileup and (ii) indicative of a score that indicates whether the hypothesis is truewherein the output information includes:

first output information for a first hypothesis of the one or more hypotheses that includes a likelihood that the sequence reads at the first position indicate the occurrence of a homozygous reference with a foreign allele that matches an alternative allele (alt);

second output information for a second hypothesis of the one or more hypotheses that includes a likelihood that the sequence reads at the first position indicate the occurrence of a homozygous alt with a foreign allele that matches a reference allele,

third output information for a third hypothesis of the one or more hypotheses that includes a likelihood that the sequence reads at the first position indicate the occurrence of a homozygous reference with a sequencing error that matches the alt allele, and

fourth output information for a fourth hypothesis of the one or more hypotheses that includes a likelihood that the sequence reads at the first position indicate the occurrence of a homozygous alt with a sequencing error that matches the reference allele; and

determining, by the one or more computers and based on the obtained output information generated by the probability model for each of the plurality of hypotheses, a likelihood that a true variant exists at the first position.

[Claim 12]. The computer-readable medium of claim 11, wherein determining, by the one or more computers and based on the obtained output information generated by the probability model for each of the plurality of hypotheses, a likelihood that a true variant exists at the first position comprises:

determining, by the one or more computers, a collective score that is based on the output information generated by the probability model for each of the plurality of hypothesis, wherein the collective score indicates a likelihood that the true variant exists;

determining, by the one or more computers, whether the score generated by the collective score satisfies a predetermined threshold; and

based on determining, by the one or more computers, that the collective score satisfies the predetermined threshold, adding information to a VCF file indicating that a true variant exists at the first position, optionally wherein the information indicating that a true variant exists at the first position includes information identifying (i) the first position, (ii) a candidate alt allele at the first position, (iii) the collective score.

[Claim 13]. The computer-readable medium of claim 11, wherein the read-allele score for each sequence read of the pileup at the first position is based on an output produced by a pair Hidden Markov Model (P-HMM) model, for each of the sequence reads at the first position, which indicates a probability of observing a sequence read r_i given a particular candidate allele $G_{m,\phi}$.

[Claim 14]. The computer-readable medium of claim 11, wherein the one or more memory devices received the pileup of aligned sequence reads from a Field Programmable Gate Array (FPGA) device, wherein the FPGA includes one or more configurable digital logic gates that have been configured as a mapping and alignment unit to perform read mapping and alignment.

[Claim 15]. The computer-readable medium of claim 14,

wherein the computer is configured to access the one or more memory devices using one or more wired or wireless networks,

wherein a Field Programmable Gate Array (FPGA) device and the one or more memory devices are housed in an expansion card that has been coupled to a circuit board of a sequencer,

wherein the sequencer is configured to generate sequence reads based on an input sample and store the generated sequence reads in the one or more memory devices, and

wherein the mapping and alignment unit of the FPGA is configured to access the one or more memory devices to obtain the generated sequence reads; or

wherein the computer and the sequencer are each configured to access the one or more memory devices using one or more wired or wireless networks,

wherein the Field Programmable Gate Array (FPGA) device and the one or more memory devices are housed in an expansion card that has been coupled to a circuit board of server that is located remotely from the computer and the sequencer,

wherein the sequencer is configured to generate sequence reads based on an input sample, provide the generated sequence reads to the server using the one or more wired or wireless networks for storage, of the generated sequence reads, in the one or more memory devices, and

wherein the mapping and alignment unit of the FPGA is configured to access the one or more memory devices to obtain the generated sequence reads.