SYSTEMS AND METHODS FOR CORRELATED ERROR EVENT MITIGATION FOR VARIANT CALLING

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims the benefit of U.S. Provisional Application No. 62/710,348, filed on February 16, 2018, and titled "Methods, Devices, and Systems for Performing Foreign Read Detection and Burst Error Detection," the disclosure of which is incorporated herein by reference in its entirety.

SEQUENCE LISTING

This application contains a Sequence Listing that has been submitted electronically as an ASCII text file named Sequence_Listing.txt. The ASCII text file, created on March 10, 2022, is 877 bytes in size. The material in the ASCII text file is hereby incorporated by reference in its entirety.

BACKGROUND

[0002] A nucleic acid sequencer is an instrument that is configured to automate the process of sequencing nucleic acids such as deoxyribonucleic acid (DNA) or ribonucleic acid (RNA). Nucleic acid sequencing is a process of determining an order of nucleotides in a genetic sequence.

[0003] The nucleic acid sequencer is configured to receive a nucleic acid sample and generate output data, referred to as one or more “reads,” that represents an order of nucleotides in the nucleic acid sample. The nucleotides in a DNA sample can include one or more of guanine (G), cytosine (C), adenine (A), and thymine (T) in any combination. The nucleotides in an RNA sample can include one or more of G, C, A, and uracil (U) in any combination in any combination.

[0004] The DNA reads generated by the DNA sequencer can be mapped and aligned to a known DNA sequence of a reference genome. Once mapped and aligned to a reference genome, the mapped and aligned sequence of reads can be analyzed in view of the reference genome to identify potential variations that exist between the mapped and aligned sequence of reads and the reference genome.

SUMMARY

[0005] Aspects of the present disclosure are directed towards methods, systems, and apparatus, including computer programs, for accounting for correlated error events that are problematic for variant callers used to determine whether an alternative (hereinafter “alt”) allele identified in a pileup of mapped and aligned sequence reads is a true variant.

[0032] FIG. 9A is an example of summary of experimental results from execution of a process for correlated error mitigation for variant calling that has been performed on a pileup of sequencing reads whose result indicates an example of a true positive results (SEQ ID NO: 1).

[0033] FIG. 9B is an example of a modified set of probability results for the pileup of FIG. 9A.

[0034] FIG. 10A is an example of summary of experimental results from execution of a process for correlated error mitigation for variant calling that has been performed on a pileup of sequencing reads whose result indicates a low likelihood of an occurrence of a mapping error (SEQ ID NO: 2).

[0035] FIG. 10B is an example of a modified set of probability results for the pileup of FIG. 10A.

[0036] FIG. 11A is an example of summary of experimental results from execution of a process for correlated error mitigation for variant calling that has been performed on a pileup of sequencing reads whose result indicates a high likelihood that the candidate alt allele is unlikely to be a true variant due to a sequencing error (SEQ ID NO: 3).

[0037] FIG. 11B is an example of a modified set of probability results for the pileup of FIG. 11A.

[0038] FIG. 12A is an example of summary of experimental results from execution of a process for correlated error mitigation for variant calling that has been performed on a pileup of sequencing reads whose result indicates a high likelihood that the candidate alt allele is unlikely to be a true variant due to a sequencing error on both read orientations (SEQ ID NO: 4).

[0039] FIG. 12B is an example of a modified set of probability results for the pileup of FIG. 12.

DETAILED DESCRIPTION

account for this potential error, indicators in the probability model for determining a likelihood that an occurrence of one or more mapping errors exits, let θ_i indicate the strand direction for read *i*, with values 0 and 1 indicating the forward and reverse strand directions, respectively. With a strand-aware modified probability mapping model for determining a likelihood that an occurrence of one or more mapping errors exist, three hypotheses can be evaluated. These three hypothesis include that the foreign reads occur exclusively on the forward strand, exclusively on the reverse strand, or on both strands. A solution includes a hypothesis that maximizes Ψ_m .

Starting with equation ~~**Error! Reference source not found**~~ (5), this variation is as follows:

$$\Psi_m = P(\mathbf{G}_m) \max_{n,\beta,p_F} p_F \left(\prod_{i \notin \Lambda_n} P(r_i \mid G_m) \right) \prod_{i \in \Lambda_n} \left(\beta P(r_i \mid F_m) U\left(-10\log_{10}\left(\frac{p_F}{P_0(F_m)} \right) - \mu_i \right) + (1-\beta) P(r_i \mid G_m) \right), \quad \text{Equation (8)}$$

where $\Lambda_0 = \{i : \theta_i = 0\}$, $\Lambda_1 = \{i : \theta_i = 1\}$, $\Lambda_2 = \{i : \theta_i = 0 \text{ or } 1\}$.

[0133] FIG. 5 is a flowchart of an example of a process for sequencing error mitigation for variant calling. The process 500 is described below as being performed by a computer such as the variant calling unit of FIG. 1. In some implementations, the variant calling unit may be implemented using hardware such as one or more FPGAs, ASICs, CPUs, GPUs, or any combination thereof, using software, or any combination thereof.

[0134] A computer may begin performance of the process 500 by accessing a pileup of aligned sequence reads stored in one or more memory devices (step 510). The aligned sequence reads may have been generated using a mapping and aligning unit implemented using configurable logic gates of an FGPA device. In some implementations, the accessed reads are stored in the one or more memory devices and can include a first set of sequence reads that are forward aligned and a second set of sequence reads that are reverse aligned. Each respective set of sequence read corresponds to a particular read orientation or direction.

[0135] The computer may continue performance of the process 500 by obtaining information describing (i) a read orientation for each of the reads of the pileup of aligned sequence reads stored in the one or more memories, (ii) a position of each base at the reference location with

[0155] Let us define the extended candidate genotype $\mathbf{G}'_m = \begin{bmatrix} G_{m,1} & G_{m,2} & E_{m,0} & E_{m,1} \end{bmatrix}$, where $E_{m,\theta}$ is the error allele for strand direction θ .

[0156] Let $L_{E,\theta}$ be the length of the homopolymer matching base, $E_{m,\theta}$ immediately preceding the error in strand direction θ.

[0157] For each extended candidate genotype **G'$_m$** , we calculate:

$$\Psi_m = P(\mathbf{G}_m) \prod_{\theta=0,1} \max_{n_\theta,\alpha_\theta} f\left(\overline{q}_{\theta,n_\theta}, L_{E,\theta}\right) \left(\prod_{i>n_\theta} \frac{P\left(r_{\theta,i} \mid G_{m,1}\right) + P\left(r_{\theta,i} \mid G_{m,2}\right)}{2} \right) \prod_{i \le n_\theta} \left(\alpha_\theta P\left(r_{\theta,i} \mid E_{m,\theta}\right) + (1-\alpha_\theta) \frac{P\left(r_{\theta,i} \mid G_{m,1}\right) + P\left(r_{\theta,i} \mid G_{m,2}\right)}{2} \right), \quad \text{Equation (9)}$$

where $f(\overline{q}, L_E)$ indicates the prior probability of an error event affecting a subset of reads as a function of the mean base quality over the subset and the length of the homopolymer matching the error.

[0158] The quantity Ψ_m represents an estimate of the joint probability $P(G'_m, R)$ under the following assumptions: (1) the error events affect a contiguous subset of reads when ordered by decreasing distance from the 5' end, starting with the first, and do not affect the reads outside of that subset, (2) the error events happen independently for each strand direction, and (3) the prior probability of such error events is a function of the mean base quality over the subset of reads affected by the error event and the length of the homopolymer matching base, $E_{m,\theta}$ immediately preceding the error in strand direction θ.

[0159] In some implementations, the number of candidates can be reduced to only test evaluating cases where $G_{m,1} = G_{m,2} = Gm$. In this case the expression simplifies to

$$\Psi_m = P(\mathbf{G}_m) \prod_{\theta=0,1} \max_{n_\theta,\alpha_\theta} f\left(\overline{q}_{\theta,n_\theta}, L_{E,\theta}\right) \left(\prod_{i>n_\theta} P\left(r_{\theta,i} \mid G_m\right) \right) \prod_{i \le n_\theta} \left(\alpha_\theta P\left(r_{\theta,i} \mid E_{m,\theta}\right) + (1-\alpha_\theta) P\left(r_{\theta,i} \mid G_m\right) \right). \quad \text{Equation (10)}$$

[0160] In some implementations, the value of α_θ can be fixed as $\alpha_\theta = 0.5$. Thus, we can rewrite ~~**Error! Reference source not found**~~ (7) as:

$$\Psi_m = P(\mathbf{G}_m)\prod_{\theta=0,1}\max_{n_\theta} f\left(\overline{q}_{\theta,n_\theta}, L_{E,\theta}\right)\left(\prod_{i>n_\theta} P\left(r_{\theta,i} \mid G_m\right)\right)\prod_{i\le n_\theta}\left(\frac{P\left(r_{\theta,i} \mid E_{m,\theta}\right)+P\left(r_{\theta,i} \mid G_m\right)}{2}\right). \quad \text{Equation (11)}$$

[0161] In some implementations, the prior probability function $f(\overline{q}_{\theta,n_\theta,L_{E\theta}})$ in **~~Error! Reference source not found~~** (6) is expressed as a general function that could have a wide range of shapes. In theory, this function could be trained on real data. However, given limitations on the use of training this function on real data, some implementations of the present probability model for determining sequencing errors may use $f\left(\overline{q}, L_E\right) = 10^{-\max\left(0, 2.5\overline{q} - \min\left(20, 5L_E\right)\right)/10}$.

[0162] In some implementations, the equations above may accommodate the possibility of different error alleles for each strand direction. However, in some implementations, the probability model for determining a likelihood of one or more sequencing errors may only consider hypotheses where $E_{m,0} = E_{m,1}$. In such implementations, the θ subscript can be dropped and then the error allele can be denoted as E_m.

[0163] FIG. 6 is another flowchart of an example of a process 600 for correlated error mitigation for variant calling. The process 600 is described below as being performed by a computer such as the variant calling unit of FIG. 1. In some implementations, the variant calling unit may be implemented using hardware such as one or more FPGAs, ASICs, CPUs, GPUs, or any combination thereof, using software, or any combination thereof.

[0164] A computer may begin performance of the process 600 by accessing a pileup of aligned sequence reads stored in one or more memory devices (step 610). The aligned sequence reads may have been generated using a mapping and aligning unit implemented using configurable logic gates of an FGPA device. In some implementations, the accessed reads are stored in the one or more memory devices and can include a first set of sequence reads that are forward aligned and a second set of sequence reads that are reverse aligned. Each respective set of sequence read corresponds to a particular read orientation or direction.