República Federativa do Brasil
Ministério da Economia
Instituto Nacional da Propriedade Industrial

(21) BR 112019027637-8 A2

(22) Data do Depósito: 09/07/2019

(43) Data da Publicação Nacional: 07/07/2020

(54) Título: SISTEMA PARA IDENTIFICAR PADRÕES DE REPETIÇÃO QUE CAUSAM ERROS ESPECÍFICOS DE SEQUÊNCIA, MÉTODO IMPLEMENTADO POR COMPUTADOR E MEIO DE ARMAZENAMENTO LEGÍVEL POR COMPUTADOR NÃO TRANSITÓRIO

(51) Int. Cl.: G16B 40/20; G16B 20/20; G06N 3/04.

(30) Prioridade Unionista: 08/07/2019 US 16/505,100; 11/07/2018 US 62/696,699; 16/08/2018 NL 2021473.

(71) Depositante(es): ILLUMINA, INC..

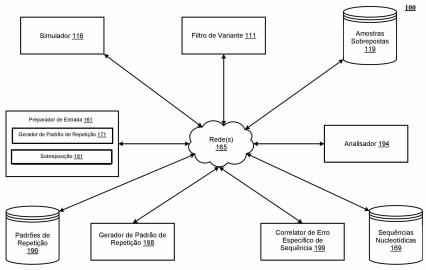
(72) Inventor(es): DORNA KASHEFHAGHIGHI; AMIRALI KIA; KAI-HOW FARH.

(86) Pedido PCT: PCT US2019041078 de 09/07/2019

(87) Publicação PCT: WO 2020/014280 de 16/01/2020

(85) Data da Fase Nacional: 20/12/2019

(57) Resumo: A tecnologia divulgada apresenta uma estrutura baseada em aprendizado profundo, que identifica padrões de sequência que causam erros específicos de sequência (SSEs). Os sistemas e métodos treinam um filtro de variantes em dados de variante em larga escala para aprender dependências causais entre padrões de sequência e chamadas de variante falsa. O filtro variante possui uma estrutura hierárquica construída em redes neurais profundas, como redes neurais convolucionais e redes neurais totalmente conectadas. Os sistemas e métodos implementam uma simulação que usa o filtro de variante para testar padrões de sequência conhecidos quanto a seus efeitos na filtragem de variante. A premissa da simulação é a seguinte: quando um par de um padrão de repetição em teste e uma chamada variante são alimentados para o filtro de variante como parte de uma sequência de entrada simulada e o filtro de variante classifica a chamada variante como uma chamada de variante falsa, então o padrão de repetição é considerado como causador da chamada de variante falsa e identificado como causador de SSE.



SISTEMA PARA IDENTIFICAR PADRÕES DE REPETIÇÃO QUE CAUSAM ERROS ESPECÍFICOS DE SEQUÊNCIA, MÉTODO IMPLEMENTADO POR COMPUTADOR E MEIO DE ARMAZENAMENTO LEGÍVEL POR COMPUTADOR NÃO TRANSITÓRIO

PEDIDOS PRIORITÁRIOS

[0001] Este pedido reivindica prioridade a ou o benefício dos seguintes pedidos:

[0002] O Pedido de Patente Provisória US Nº 62/696,699, intitulado"DEEP LEARNING-BASED FRAMEWORK FOR IDENTIFYING SEQUENCE PATTERNS THAT CAUSE SEQUENCE-SPECIFIC ERRORS (SSEs)", depositado em 11 de julho de 2018 (Nº de Registro do Procurador ILLM 1006-1/IP-1650-PRV);

[0003] O Pedido da Holanda Nº 2021473, intitulado “DEEP LEARNING-BASED FRAMEWORK FOR IDENTIFYING SEQUENCE PATTERNS THAT CAUSE SEQUENCE-SPECIFIC ERRORS (SSEs)”, depositado em 16 de agosto de 2018 (Nº de Registro do Procurador ILLM 1006-4/IP-1650-NL); e

[0004] O Pedido de Patente Provisória US Nº 16/505,100, intitulado"DEEP LEARNING-BASED FRAMEWORK FOR IDENTIFYING SEQUENCE PATTERNS THAT CAUSE SEQUENCE-SPECIFIC ERRORS (SSEs)", depositado em 08 de julho de 2019 (Nº de Registro do Procurador ILLM 1006-2/IP-1650-US).

[0005] Os pedidos prioritários são incorporados por referência neste documento para todos os fins.

INCORPORAÇÕES

[0006] Os seguintes são incorporados por referência para todos os propósitos, como se totalmente estabelecido neste documento:

[0007] Aplicativo Strelka™ da Illumina Inc. hospedado em https://github.com/Illumina/strelka e descrito no artigo T Saunders,

Christopher & Wong, Wendy & Swamy, Sajani & Becq, Jennifer & J Murray, Lisa & Cheetham, Keira. (2012). Strelka: Accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics (Oxford, England). 28. 1811-7;

[0008] Aplicativo Strelka2 ™ da Illumina Inc. hospedado em https://github.com/Illumina/strelka e descrito no artigo Kim, S., Scheffler, K., Halpern, AL, Bekritsky, MA, Noh, E., Källberg, M., Chen, X., Beyter, D., Krusche, P. e Saunders, CT (2017);

[0009] A. van der Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior e K. Kavukcuoglu, "WAVENET: A GENERATIVE MODEL FOR RAW AUDIO", arXiv: 1609.03499, 2016;

[0010] S. Ö. Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, A. Ng, J. Raiman, S. Sengupta and M. Shoeybi, "DEEP VOICE: REAL-TIME NEURAL TEXT-TO-SPEECH," arXiv:1702.07825, 2017;

[0011] F. Yu e V. Koltun, "MULTI-SCALE CONTEXT AGGREGATION BY DILATED CONVOLUTIONS", arXiv: 1511.07122, 2016;

[0012] K. He, X. Zhang, S. Ren e J. Sun, "DEEP RESIDUAL LEARNING FOR IMAGE RECOGNITION", arXiv: 1512.03385, 2015;

[0013] R.K. Srivastava, K. Greff e J. Schmidhuber, "HIGHWAY NETWORKS", arXiv: 1505.00387, 2015;

[0014] G. Huang, Z. Liu, L. van der Maaten e K.Q. Weinberger, "DENSELY CONNECTED CONVOLUTIONAL NETWORKS", arXiv: 1608.06993, 2017;

[0015] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke e A. Rabinovich, "GOING DEEPER WITH CONVOLUTIONS", arXiv: 1409.4842, 2014;

[0016] S. Ioffe e C. Szegedy, "BATCH NORMALIZATION: ACCELERATING DEEP NETWORK TRAINING BY REDUCING INTERNAL

COVARIATE SHIFT”, arXiv: 1502.03167, 2015;

[0017] Srivastava, Nitish, Hinton, Geoffrey, Krizhevsky, Alex, Sutskever, Ilya e Salakhutdinov, Ruslan, "DROPOUT: A SIMPLE WAY TO PREVENT NEURAL NETWORKS FROM OVERFITTING", The Journal of Machine Learning Research, 15 (1): 1929-1958, 2014;

[0018] J.M. Wolterink, T. Leiner, M.A. Viergever e I. Išgum, “DILATED CONVOLUTIONAL NEURAL NETWORKS FOR CARDIOVASCULAR MR SEGMENTATION IN CONGENITAL HEART DISEASE”, arXiv: 1704.03669, 2017;

[0019] L.C. Piqueras, “AUTOREGRESSIVE MODEL BASED ON A DEEP CONVOLUTIONAL NEURAL NETWORK FOR AUDIO GENERATION,” Tampere University of Technology, 2016;

[0020] J. Wu, “Introduction to Convolutional Neural Networks”, Nanjing University, 2017;

[0021] Documento 12 - I.J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville e Y. Bengio, “CONVOLUTIONAL NETWORKS”, Deep Learning, MIT Press, 2016;

[0022] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang e G. Wang, “RECENT ADVANCES IN CONVOLUTIONAL NEURAL NETWORKS”, arXiv: 1512.07108, 2017;

[0023] M. Lin, Q. Chen e S. Yan, "Network in Network", em Proc. do ICLR, 2014;

[0024] L. Sifre, “Rigid-motion Scattering for Image Classification, tese de Ph.D., 2014;

[0025] L. Sifre e S. Mallat, "Rotation, Scaling and Deformation Invariant Scattering for Texture Discrimination", em Proc. da CVPR, 2013;

[0026] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions", em Proc. da CVPR, 2017;

[0027] X. Zhang, X. Zhou, M. Lin e J. Sun, “ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices”, em

arXiv: 1707.01083, 2017;

[0028] K. Ele, X. Zhang, S. Ren e J. Sun, "Deep Residual Learning for Image Recognition", em Proc. da CVPR, 2016;

[0029] S. Xie, R. Girshick, P. Dollár, Z. Tu e K. He, "Aggregated Residual Transformations for Deep Neural Networks", em Proc. da CVPR, 2017;

[0030] AG Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto e H. Adam, “Mobilenets: Efficient Convolutional Neural Networks for Mobile Vision Applications”, em arXiv: 1704.04861, 2017;

[0031] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov e L. Chen, “MobileNetV2: Inverted Residuals and Linear Bottlenecks”, em arXiv: 1801.04381v3, 2018;

[0032] Z. Qin, Z. Zhang, X. Chen e Y. Peng, “FD-MobileNet: Improved MobileNet with a Fast Downsampling Strategy”, em arXiv: 1802.03750, 2018;

[0033] Pedido de Patente Internacional PCT Nº PCT/US17/61554, intitulado “Validation Methods and Systems for Sequence Variant Calls”, depositado em 14 de novembro de 2017;

[0034] Pedido de Patente Provisório US Nº 62/447,076, intitulado "Validation Methods and Systems for Sequence Variant Calls", depositado em 17 de janeiro de 2017;

[0035] Pedido de Patente Provisório US Nº 62/422,841, intitulado "Methods and Systems to Improve Accuracy in Variant Calling", depositado em 16 de novembro de 2016; e

[0036] N. ten DIJKE, “Convolutional Neural Networks for Regulatory Genomics”, Tese de Mestrado, Universiteit Leiden Opleiding Informatica, 17 de junho de 2017.

CAMPO DA TECNOLOGIA DIVULGADA

[0037] A tecnologia divulgada refere-se a computadores do tipo

inteligência artificial e sistemas de processamento de dados digitais e métodos de processamento de dados correspondentes e produtos para emulação de inteligência (isto é, sistemas baseados em conhecimento, sistemas de raciocínio e sistemas de aquisição de conhecimento); e incluindo sistemas de raciocínio com incerteza (por exemplo, sistemas de lógica difusa), sistemas adaptativos, sistemas de aprendizagem de máquina e redes neurais artificiais. Em particular, a tecnologia divulgada refere-se ao uso de redes neurais profundas, como redes neurais convolucionais (CNNs) e redes neurais totalmente conectadas (FCNNs) para análise de dados.

FUNDAMENTOS

[0038] O assunto discutido nesta seção não deve ser considerado como estado da técnica apenas como resultado de sua menção nesta seção. Da mesma forma, não se deve presumir que um problema mencionado nesta seção ou associado ao assunto fornecido como fundamento tenha sido reconhecido anteriormente no estado da técnica. O assunto desta seção representa apenas abordagens diferentes, que por si só também podem corresponder a implementações da tecnologia reivindicada.

[0039] O sequenciamento de última geração disponibilizou grandes quantidades de dados sequenciados para filtragem de variantes. Os dados sequenciados são altamente correlacionados e têm interdependências complexas, o que dificultou a aplicação de classificadores tradicionais, como a máquina de vetores de suporte, à tarefa de filtragem de variantes. Classificadores avançados capazes de extrair recursos de alto nível de dados sequenciados são, portanto, desejados.

[0040] As redes neurais profundas são um tipo de redes neurais artificiais que usam várias camadas transformadoras não-lineares e complexas para modelar sucessivamente recursos de alto nível. As redes neurais profundas fornecem feedback via retropropagação, que carrega a diferença entre a saída observada e a prevista para ajustar os parâmetros. As redes neurais profundas evoluíram com a disponibilidade de grandes

conjuntos de dados de treinamento, o poder da computação distribuída e paralela e algoritmos sofisticados de treinamento. As redes neurais profundas facilitaram grandes avanços em vários domínios, como visão computacional, reconhecimento de fala e processamento de linguagem natural.

[0041] Redes neurais convolucionais (CNNs) e redes neurais recorrentes (RNNs) são componentes de redes neurais profundas. As redes neurais convolucionais tiveram sucesso, particularmente, no reconhecimento de imagens com uma arquitetura que compreende camadas de convolução, camadas não lineares e camadas de pool. As redes neurais recorrentes são projetadas para utilizar informações sequenciais de dados de entrada com conexões cíclicas entre blocos de construção como perceptrons, unidades de memória de longo prazo e unidades recorrentes fechadas. Além disso, muitas outras redes neurais profundas emergentes foram propostas para contextos limitados, como redes neurais espaço-temporais profundas, redes neurais recorrentes multidimensionais e auto-codificadores convolucionais.

[0042] O objetivo de treinar redes neurais profundas é a otimização dos parâmetros de peso em cada camada, que combina gradualmente recursos mais simples em recursos complexos, para que as representações hierárquicas mais adequadas possam ser aprendidas com os dados. Um único ciclo do processo de otimização é organizado da seguinte maneira. Primeiro, dado um conjunto de dados de treinamento, o passo para frente calcula sequencialmente a saída em cada camada e propaga os sinais da função para frente através da rede. Na camada de saída final, uma função de perda objetiva mede o erro entre as saídas inferidas e os rótulos fornecidos. Para minimizar o erro de treinamento, o passo para trás usa a regra da cadeia para retropropagar sinais de erro e calcular gradientes em relação a todos os pesos ao longo de toda a rede neural. Finalmente, os parâmetros de peso são atualizados usando algoritmos de otimização baseados na descida do gradiente estocástico. Enquanto a descida do

gradiente em lote executa atualizações de parâmetros para cada conjunto de dados completo, a descida do gradiente estocástico fornece aproximações estocásticas executando as atualizações para cada pequeno conjunto de exemplos de dados. Vários algoritmos de otimização decorrem da descida do gradiente estocástico. Por exemplo, os algoritmos de treinamento Adagrad e Adam executam descida do gradiente estocástico enquanto modificam adaptativamente as taxas de aprendizagem com base na frequência de atualização e nos momentos dos gradientes para cada parâmetro, respectivamente.

[0043] Outro elemento central no treinamento de redes neurais profundas é a regularização, que se refere a estratégias destinadas a evitar o sobreajuste e, assim, alcançar um bom desempenho de generalização. Por exemplo, a redução de peso adiciona um termo de penalidade à função de perda objetiva, para que os parâmetros de peso convirjam para valores absolutos menores. O dropout remove aleatoriamente unidades ocultas das redes neurais durante o treinamento e pode ser considerado um conjunto de possíveis sub-redes. Para aprimorar os recursos de dropout, foram propostas uma nova função de ativação, maxout e uma variante de dropout para redes neurais recorrentes denominadas rnnDrop. Além disso, a normalização de lote fornece um novo método de regularização através da normalização de recursos escalares para cada ativação dentro de um mini lote e aprendendo cada média e variação como parâmetros.

[0044] Dado que os dados sequenciados são multidimensionais e de alta dimensão, as redes neurais profundas têm grandes promessas para a pesquisa em bioinformática devido à sua ampla aplicabilidade e poder de previsão aprimorado. As redes neurais convolucionais foram adaptadas para resolver problemas genômicos baseados em sequências, como descoberta de motivos, identificação de variantes patogênicas e inferência de expressão gênica. Uma característica das redes neurais convolucionais é o uso de filtros de convolução. Diferentemente das abordagens de classificação tradicionais

baseadas em recursos elaborados e criados manualmente, os filtros de convolução realizam uma aprendizagem adaptável dos recursos, análogo a um processo de mapeamento de dados brutos de entrada para a representação informativa do conhecimento. Nesse sentido, os filtros de convolução servem como uma série de scanners de motivos, pois um conjunto desses filtros é capaz de reconhecer padrões relevantes na entrada e se atualizar durante o procedimento de treinamento. Redes neurais recorrentes podem capturar dependências de longo alcance em dados sequenciais de comprimentos variados, como sequências de proteínas ou DNA.

[0045] Portanto, surge uma oportunidade de usar uma estrutura baseada em princípios de aprendizagem profunda que associa padrões de sequência a erros de sequência.

BREVE DESCRIÇÃO DOS DESENHOS

[0046] Nos desenhos, caracteres de referência semelhantes geralmente se referem a partes semelhantes ao longo das diferentes vistas. Além disso, os desenhos não estão necessariamente em escala, com ênfase sendo geralmente colocada na ilustração dos princípios da tecnologia divulgada. Na descrição a seguir, várias implementações da tecnologia divulgada são descritas com referência aos seguintes desenhos, nos quais:

[0047] A **FIGURA 1** é um diagrama de blocos que mostra vários aspectos do DeepPOLY, um framework baseado em aprendizagem profunda para identificar padrões de sequência que causam erros específicos de sequência (SSEs). A **FIGURA1** inclui módulos como um filtro de variante, um simulador e um analisador. A **FIGURA 1** também inclui bancos de dados que armazenam amostras sobrepostas, sequências de nucleotídeos e padrões de repetição.

[0048] A **FIGURA 2** ilustra um exemplo de arquitetura do filtro de variante. O filtro de variante possui uma estrutura hierárquica construída em uma rede neural convolucional (CNN) e uma rede neural totalmente

conectada (FCNN). O DeepPOLY usa o filtro de variantes para testar padrões de sequência conhecidos quanto a seus efeitos na filtragem de variantes.

[0049] A **FIGURA 3** mostra uma implementação do pipeline de processamento do filtro de variante.

[0050] A **FIGURA 4A** mostra plotagens positivas verdadeiras e falsas que ilustram graficamente o desempenho do filtro de variante em dados retidos.

[0051] As **FIGURAS 4B** e **4C** mostram imagens acumuladas de leituras alinhadas que validam a precisão do filtro de variante.

[0052] A **FIGURA 5** mostra uma implementação de codificação one-hot usada para codificar a amostra sobreposta que tem uma variante chamada em uma posição de destino flanqueada por 20-50 bases em cada lado.

[0053] A **FIGURA 6** ilustra exemplos de amostras sobrepostas produzidas pelo preparador de entrada sobrepondo os padrões de repetição nas sequências de nucleotídeos.

[0054] A **FIGURA 7A** usa um diagrama de caixa e bigodes para identificar a causa dos erros específicos da sequência por padrões de repetição à esquerda do nucleotídeo variante na posição alvo nas amostras sobrepostas.

[0055] A **FIGURA 7B** usa um diagrama de caixa e bigodes para identificar a causa dos erros específicos da sequência por padrões de repetição à direita do nucleotídeo variante na posição alvo nas amostras sobrepostas.

[0056] A **FIGURA 7C** usa um diagrama de caixa e bigodes para identificar a causa dos erros específicos da sequência por padrões de repetição incluindo um nucleotídeo variante na posição alvo nas amostras sobrepostas.

[0057] A **FIGURA 8A** usa um diagrama de caixa e bigode para

identificar a causa de erros específicos de sequência por padrões de repetição de homopolímeros de uma única base "C" sobrepostos a desvios variados nas sequências de nucleotídeos.

[0058] A **FIGURA 8B** usa um diagrama de caixa e bigode para identificar a causa de erros específicos de sequência por padrões repetidos de homopolímeros de uma única base "G" sobrepostos a desvios variados nas sequências de nucleotídeos.

[0059] A **FIGURA 8C** usa um diagrama de caixa e bigode para identificar a causa de erros específicos da sequência por padrões de repetição de homopolímeros de uma única base “A” sobrepostos a desvios variados nas sequências de nucleotídeos.

[0060] A **FIGURA 8D** usa um diagrama de caixa e bigode para identificar a causa de erros específicos de sequência por padrões de repetição de homopolímeros de uma única base "C" sobrepostos a desvios variados nas sequências de nucleotídeos.

[0061] A **FIGURA 9** exibe pontuações de classificação como uma distribuição para a probabilidade de que um nucleotídeo variante seja uma variante verdadeira ou falsa quando padrões repetidos de homopolímeros de uma única base são colocados um por um "antes" e "depois" de um nucleotídeo variante de cada uma das quatro bases na posição alvo.

[0062] As **FIGURAS 10A** a **10C** exibem uma representação de padrões de repetição de copolímeros que ocorrem naturalmente em cada uma das sequências nucleotídicas da amostra que contribuem para uma classificação falsa de variantes.

[0063] A **FIGURA 11** é um diagrama de blocos simplificado de um sistema de computador que pode ser usado para implementar o filtro de variante.

[0064] A **FIGURA 12** ilustra uma implementação de como os erros específicos de sequência (SSEs) são correlacionados para repetir padrões com base em classificações de variantes falsas.

DESCRIÇÃO DETALHADA

[0065] A discussão a seguir é apresentada para permitir que qualquer pessoa versada na técnica faça e utilize a tecnologia divulgada, e é fornecida no contexto de um pedido particular e seus requisitos. Várias modificações às implementações divulgadas serão prontamente evidentes para os versados na técnica, e os princípios gerais definidos neste documento podem ser aplicados a outras implementações e pedidos sem se afastar do espírito e âmbito da tecnologia divulgada. Assim, a tecnologia divulgada não se destina a ser limitada às implementações apresentadas, mas deve receber o escopo mais amplo consistente com os princípios e características divulgados neste documento.

Introdução

[0066] Erros específicos de sequência (SSEs) são erros de chamada base causados por padrões de sequência específicos. Por exemplo, verificou-se que os padrões de sequência 'GGC' e 'GGCNG' e suas repetições invertidas causam grandes quantidades de chamadas erradas. SSEs levam a falhas de montagem e artefatos de mapeamento. Além disso, como qualquer erro de chamada pode ser confundido com uma variante, as SSEs resultam em chamadas de variantes falsas e são um grande obstáculo para a chamada de variante precisa.

[0067] Divulgamos um framework baseado em aprendizagem profunda, DeepPOLY, que identifica padrões de sequência que causam SSEs. O DeepPOLY treina um filtro de variante em dados de variante em larga escala para aprender dependências causais entre padrões de sequência e chamadas de variantes falsas. O filtro de variantes possui uma estrutura hierárquica construída em redes neurais profundas que avaliam uma sequência de entrada em várias escalas espaciais e realizam a filtragem de variantes, ou seja, prever se uma variante chamada na sequência de entrada é uma chamada de variante verdadeira ou uma chamada de variante

falsa. Os dados de variantes em larga escala incluem variantes de árvore genealógica, das quais variantes herdadas são usadas como exemplos de treinamento de chamadas de variantes verdadeiras e variantes de novo observadas em apenas um filho são usadas como exemplos de treinamento de chamadas de variantes falsas. Em algumas implementações, pelo menos algumas das variantes de novo observadas em apenas um filho são usadas como exemplos de treinamento de chamadas de variantes verdadeiras.

[0068] Durante o treinamento, os parâmetros das redes neurais profundas são otimizados para maximizar a precisão da filtragem usando uma abordagem de descida gradiente. O filtro de variantes resultante aprende a associar chamadas de variantes falsas a padrões de sequência nas sequências de entrada.

[0069] O DeepPOLY então implementa uma simulação que usa o filtro de variantes para testar padrões de sequência conhecidos quanto a seus efeitos na filtragem de variantes. Os padrões de sequência conhecidos são padrões de repetição (ou copolímeros) que diferem na composição de base, comprimento do padrão e fator de repetição. Os padrões de repetição são testados em diferentes deslocamentos das variantes chamadas.

[0070] A premissa da simulação é a seguinte: quando um par de um padrão de repetição em teste e uma variante chamada é alimentado ao filtro de variante como parte de uma sequência de entrada simulada e o filtro de variante classifica a variante chamada como uma chamada falsa de variante, então o padrão de repetição é considerado como causador da chamada falsa de variante e identificado como causador de SSE. Sob essa premissa, o DeepPOLY testa centenas e milhares de padrões de repetição para identificar quais causam SSE, com sensibilidade ao deslocamento.

[0071] O DeepPOLY também descobre padrões de sequência que ocorrem naturalmente que causam SSEs processando sequências de entrada que ocorrem naturalmente através do filtro de variantes e analisando ativações de parâmetros das redes neurais profundas durante o

processamento. Esses padrões de sequência são identificados como causadores de SSE, para os quais os neurônios de entrada das redes neurais profundas produzem as ativações de parâmetro mais altas e os neurônios de saída produzem uma classificação de chamada de variante falsa.

[0072] O DeepPOLY confirma padrões de sequência que causam SSE conhecidos anteriormente e relata novos padrões mais específicos.

[0073] O DeepPOLY é independente da química de sequenciamento subjacente, plataforma de sequenciamento e polimerases de sequenciamento e pode produzir perfis abrangentes de padrões de sequência que causam SSE para diferentes químicas de sequenciamento, plataformas de sequenciamento e polimerases de sequenciamento. Esses perfis podem ser usados para melhorar as químicas de sequenciamento, criar plataformas de sequenciamento de qualidade superior e criar polimerases de sequenciamento diferentes. Eles também podem ser usados para recalcular as pontuações de qualidade da chamada de base e melhorar a precisão da chamada de variante.

[0074] O filtro de variante possui duas redes neurais profundas: uma rede neural convolucional (CNN) seguida por uma rede neural totalmente conectada (FCNN). Um padrão de repetição em teste é sobreposto em uma sequência de nucleotídeos para produzir uma amostra sobreposta. A amostra sobreposta possui uma variante de chamada na posição de destino, flanqueada por 20 a 50 bases de cada lado. Consideramos a amostra sobreposta como uma imagem com vários canais que codificam numericamente os quatro tipos de bases, A, C, G e T. A amostra sobreposta, abrangendo a chamada variante, é codificada com um hot hot para conservar as informações específicas da posição de cada base individual na amostra sobreposta.

[0075] A rede neural convolucional recebe a amostra sobreposta superaquecida porque é capaz de preservar as relações de localização

espacial dentro da amostra superposta. A rede neural convolucional processa a amostra sobreposta por várias camadas de convolução e produz um ou mais recursos convoluídos intermediários. As camadas de convolução utilizam filtros de convolução para detectar padrões de sequência na amostra sobreposta. Os filtros de convolução atuam como detectores de motivo que examinam a amostra sobreposta em busca de motivos de baixo nível e produzem sinais de diferentes forças, dependendo dos padrões de sequência subjacentes. Os filtros de convolução são aprendidos automaticamente após o treinamento em centenas e milhares de exemplos de treinamento de chamadas de variantes verdadeiras e falsas.

[0076] A rede neural totalmente conectada processa os recursos intermediários convoluídos através de várias camadas totalmente conectadas. Os neurônios densamente conectados das camadas totalmente conectadas detectam padrões de sequência de alto nível codificados nos recursos envolvidos. Finalmente, uma camada de classificação da rede neural totalmente conectada gera probabilidades para a variante chamada, sendo uma chamada de variante verdadeira ou uma chamada de variante falsa.

[0077] Além de usar o dropout, pares de normalização do lote e não linearidade da unidade linear retificada são intercalados entre as camadas convolucionais e as camadas totalmente conectadas para melhorar as taxas de aprendizado e reduzir o sobreajuste.

Terminologia

[0078] Toda a literatura e material semelhante citado neste pedido, incluindo, mas não limitada a, patentes, pedidos de patente, artigos, livros, tratados e páginas da web, independentemente do formato dessa literatura e materiais similares, são expressamente incorporados por referência em seus totalidade. No caso de uma ou mais da literatura incorporada, patentes e materiais semelhantes diferirem ou contradizerem este pedido, incluindo, mas não limitado a termos definidos, uso de termos, técnicas descritas ou

semelhantes, esse pedido prevalece.

[0079] Conforme usado neste documento, os seguintes termos têm os significados indicados.

[0080] Uma base refere-se a uma base nucleotídica ou nucleotídico, A (adenina), C (citosina), T (timina) ou G (guanina).

[0081] O termo "cromossomo" refere-se ao carreador de genes portadores de hereditariedade de uma célula viva, que é derivada de cadeias de cromatina que compreendem componentes de DNA e proteínas (especialmente histonas). O sistema convencional de numeração de cromossomos do genoma humano internacionalmente reconhecido é empregado neste documento.

[0082] O termo "sítio" refere-se a uma posição única (por exemplo, ID do cromossomo, posição e orientação do cromossomo) em um genoma de referência. Em algumas implementações, um sítio pode ser um resíduo, uma tag de sequência ou a posição de um segmento em uma sequência. O termo "locus" pode ser usado para se referir à localização específica de uma sequência de ácido nucleico ou polimorfismo em um cromossomo de referência.

[0083] O termo "amostra" neste documento refere-se a uma amostra, tipicamente derivada de um fluido biológico, célula, tecido, órgão ou organismo contendo um ácido nucleico ou uma mistura de ácidos nucleicos contendo pelo menos uma sequência de ácido nucleico que deve ser sequenciada e/ou faseado. Essas amostras incluem, mas não são limitadas a, expectoração/fluido oral, líquido amniótico, sangue, uma fração sanguínea, amostras de biópsia por agulha fina (por exemplo, biópsia cirúrgica, biópsia por agulha fina, etc.), urina, líquido peritoneal, líquido pleural, tecido de explante, cultura de órgãos e qualquer outra preparação de tecido ou célula, ou fração ou derivado ou isolados a partir deles. Embora a amostra seja frequentemente retirada de um sujeito humano (por exemplo, paciente), amostras podem ser coletadas de qualquer organismo com

cromossomos, incluindo, mas não limitados a, cães, gatos, cavalos, cabras, ovelhas, gado, porcos, etc. A amostra pode ser usada diretamente conforme obtida da fonte biológica ou após um pré-tratamento para modificar o caráter da amostra. Por exemplo, esse pré-tratamento pode incluir a preparação de plasma a partir do sangue, diluição de fluidos viscosos e assim por diante. Os métodos de pré-tratamento também podem envolver, mas não estão limitados a, filtração, precipitação, diluição, destilação, mistura, centrifugação, congelamento, liofilização, concentração, amplificação, fragmentação de ácidos nucleicos, inativação de componentes interferentes, adição de reagentes, lisação, etc.

[0084] O termo "sequência" inclui ou representa uma cadeia de nucleotídeos acoplados um ao outro. Os nucleotídeos podem ser baseados em DNA ou RNA. Deve ser entendido que uma sequência pode incluir múltiplas sub-sequências. Por exemplo, uma única sequência (por exemplo, de um amplificon de PCR) pode ter 350 nucleotídeos. A leitura da amostra pode incluir múltiplas sub-sequências dentro destes 350 nucleotídeos. Por exemplo, a leitura da amostra pode incluir primeira e segunda subsequências de flanqueamento com, por exemplo, 20-50 nucleotídeos. A primeira e a segunda sequências de flanqueamento podem estar localizadas em ambos os lados de um segmento repetitivo que possui uma sub-sequência correspondente (por exemplo, 40-100 nucleotídeos). Cada uma das sub-sequências de flanqueamento pode incluir (ou incluir porções de) uma sub-sequência de primer (por exemplo, 10-30 nucleotídeos). Para facilitar a leitura, o termo "sub-sequência" será referido como "sequência", mas entende-se que duas sequências não são necessariamente separadas uma da outra em uma cadeia comum. Para diferenciar as várias sequências descritas neste documento, as sequências podem receber rótulos diferentes (por exemplo, sequência alvo, sequência primer, sequência flanqueadora, sequência de referência e similares). Outros termos, como "alelo", podem receber rótulos diferentes para diferenciar objetos semelhantes.

[0085] O termo "sequenciamento de extremidade pareada" refere-se a métodos de sequenciamento que sequenciam as duas extremidades de um fragmento alvo. O sequenciamento de extremidade pareada pode facilitar a detecção de rearranjos genômicos e segmentos repetitivos, bem como fusões de genes e novos transcritos. A metodologia para o sequenciamento de extremidade pareada é descrita na publicação PCT WO07010252, pedido PCT Nº de série PCTGB2007/003798 e publicação do pedido de patente US US 2009/0088327, cada uma das quais é incorporada por referência neste documento. Em um exemplo, uma série de operações pode ser executada da seguinte maneira; (a) gerar grupos de ácidos nucleicos; (b) linearizar os ácidos nucleicos; (c) hibridizar um primeiro primer de sequenciamento e realizar ciclos repetidos de extensão, varredura e desbloqueio, conforme estabelecido acima; (d) "inverter" os ácidos nucleicos alvo na superfície da célula de fluxo sintetizando uma cópia complementar; (e) linearizar a cadeia ressintetizada; e (f) hibridizar um segundo primer de sequenciamento e realizar ciclos repetidos de extensão, varredura e desbloqueio, conforme estabelecido acima. A operação de inversão pode ser realizada administrando reagentes conforme estabelecido acima para um único ciclo de amplificação em ponte.

[0086] O termo "genoma de referência" ou "sequência de referência" refere-se a qualquer sequência específica do genoma conhecido, parcial ou completa, de qualquer organismo que possa ser usado para referenciar sequências identificadas de um sujeito. Por exemplo, um genoma de referência usado para seres humanos, assim como muitos outros organismos, é encontrado no National Center for Biotechnology Information, em ncbi.nlm.nih.gov. Um "genoma" refere-se à informação genética completa de um organismo ou vírus, expressa em sequências de ácidos nucleicos. Um genoma inclui os genes e as sequências não codificadoras do DNA. A sequência de referência pode ser maior que as leituras alinhadas a ela. Por exemplo, pode ser pelo menos cerca de 100 vezes maior, ou pelo menos

cerca de 1000 vezes maior, ou pelo menos cerca de 10,000 vezes maior, ou pelo menos cerca de 105 vezes maior, ou pelo menos cerca de 106 vezes maior, ou pelo menos cerca de 107 vezes maior. Em um exemplo, a sequência do genoma de referência é a de um genoma humano completo. Em outro exemplo, a sequência do genoma de referência é limitada a um cromossomo humano específico, como o cromossomo 13. Em algumas implementações, um cromossomo de referência é uma sequência cromossômica da versão hg19 do genoma humano. Tais sequências podem ser referidas como sequências de referência cromossômica, embora o termo genoma de referência se destine a cobrir tais sequências. Outros exemplos de sequências de referência incluem genomas de outras espécies, bem como cromossomos, regiões sub-cromossômicas (como cadeias), etc., de qualquer espécie. Em várias implementações, o genoma de referência é uma sequência de consenso ou outra combinação derivada de vários indivíduos. No entanto, em certas aplicações, a sequência de referência pode ser obtida de um indivíduo em particular.

[0087] O termo "leitura" refere-se a uma coleção de dados de sequência que descreve um fragmento de uma referência ou amostra de nucleotídeo. O termo "leitura" pode se referir a uma leitura de amostra e/ou uma leitura de referência. Normalmente, embora não necessariamente, uma leitura representa uma sequência curta de pares de bases contíguas na amostra ou referência. A leitura pode ser representada simbolicamente pela sequência de pares de bases (em ATCG) da amostra ou fragmento de referência. Ela pode ser armazenada em um dispositivo de memória e processada conforme apropriado para determinar se a leitura corresponde a uma sequência de referência ou se atende a outros critérios. Uma leitura pode ser obtida diretamente de um aparelho de sequenciamento ou indiretamente a partir de informações de sequência armazenadas relativas à amostra. Em alguns casos, uma leitura é uma sequência de DNA de comprimento suficiente (por exemplo, pelo menos cerca de 25 bp) que pode

ser usada para identificar uma sequência ou região maior, por exemplo, que pode ser alinhada e atribuída especificamente a um cromossomo ou região genômica ou gene.

[0088] Os métodos de sequenciamento de última geração incluem, por exemplo, sequenciamento por tecnologia de síntese (Illumina), pirosequenciamento (454), tecnologia de semicondutores de íons (sequenciamento Ion Torrent), sequenciamento em tempo real de molécula única (Pacific Biosciences) e sequenciamento por ligação (sequenciamento SOLiD). Dependendo dos métodos de sequenciamento, o comprimento de cada leitura pode variar de cerca de 30 bp a mais de 10,000 bp. Por exemplo, o método de sequenciamento Illumina usando o sequenciador SOLiD gera leituras de ácido nucleico de cerca de 50 bp. Por outro exemplo, o Sequenciamento Ion Torrent gera leituras de ácido nucleico de até 400 bp e o pirosequenciamento 454 gera leituras de ácido nucleico de cerca de 700 bp. Por outro exemplo, os métodos de sequenciamento em tempo real de molécula única podem gerar leituras de 10,000 a 15,000 bp. Portanto, em certas implementações, as leituras da sequência de ácido nucleico têm um comprimento de 30-100 bp, 50-200 bp ou 50-400 bp.

[0089] Os termos “leitura da amostra”, “sequência da amostra” ou “fragmento da amostra” se referem aos dados da sequência para uma sequência genômica de interesse de uma amostra. Por exemplo, a leitura da amostra compreende dados de sequência de um amplicon de PCR tendo uma sequência de primers foward e reverse. Os dados da sequência podem ser obtidos a partir de qualquer metodologia de sequência selecionada. A leitura da amostra pode ser, por exemplo, de uma reação de sequenciamento por síntese (SBS), uma reação de sequenciamento por ligação ou qualquer outra metodologia de sequenciamento adequada para a qual se deseja determinar o comprimento e/ou a identidade de um elemento repetitivo. A leitura da amostra pode ser uma sequência de consenso (por exemplo, média ou ponderada) derivada de várias leituras da amostra. Em certas

implementações, o fornecimento de uma sequência de referência compreende a identificação de um locus de interesse com base na sequência pimer do amplicon de PCR.

[0090] O termo "fragmento bruto" refere-se a dados de sequência para uma porção de uma sequência genômica de interesse que se sobrepõe pelo menos parcialmente a uma posição designada ou a uma posição secundária de interesse dentro de uma amostra de leitura ou fragmento de amostra. Exemplos não limitativos de fragmentos brutos incluem um fragmento concatenado duplex, um fragmento concatenado simplex, um fragmento não concatenado duplex e um fragmento não concatenado simplex. O termo "bruto" é usado para indicar que o fragmento bruto inclui dados de sequência que têm alguma relação com os dados de sequência em uma leitura de amostra, independentemente de o fragmento bruto exibir uma variante de suporte que corresponda e autentique ou confirme uma variante em potencial em uma leitura de amostra. O termo "fragmento bruto" não indica que o fragmento inclui necessariamente uma variante de suporte que valida uma chamada de variante em uma leitura de amostra. Por exemplo, quando uma leitura de amostra é determinada por um aplicativo de chamada de variante para exibir uma primeira variante, o aplicativo de chamada de variante pode determinar que um ou mais fragmentos brutos não possuem um tipo correspondente de variante "de suporte" que, de outra forma, pode ser esperado que ocorra, dada a variante na leitura de amostra.

[0091] Os termos "mapeamento", "alinhado", "alinhamento" ou "ordenamento" referem-se ao processo de comparar uma leitura ou tag a uma sequência de referência e, assim, determinar se a sequência de referência contém a sequência de leitura. Se a sequência de referência contiver a leitura, a leitura poderá ser mapeada para a sequência de referência ou, em certas implementações, para um local específico na sequência de referência. Em alguns casos, o alinhamento simplesmente informa se uma leitura é ou não um membro de uma sequência de referência

específica (ou seja, se a leitura está presente ou ausente na sequência de referência). Por exemplo, o alinhamento de uma leitura com a sequência de referência para o cromossomo humano 13 indicará se a leitura está presente na sequência de referência para o cromossomo 13. Uma ferramenta que fornece essas informações pode ser chamada de testador de associação estabelecida. Em alguns casos, um alinhamento indica adicionalmente um local na sequência de referência onde a leitura ou o tag é mapeado. Por exemplo, se a sequência de referência é sequência do genoma humano completo, um alinhamento pode indicar que uma leitura está presente no cromossomo 13 e pode ainda indicar que a leitura está em uma cadeia e/ou sítio específico do cromossomo 13.

[0092] O termo "indel" refere-se à inserção e/ou deleção de bases no DNA de um organismo. Um micro-indel representa um indel que resulta em uma alteração líquida de 1 a 50 nucleotídeos. Nas regiões codificadoras do genoma, a menos que o comprimento de um indel seja um múltiplo de 3, ele produzirá uma mutação de deslocamento de quadro. Indels podem ser contrastados com mutações pontuais. Um indel insere e deleta nucleotídeos de uma sequência, enquanto uma mutação pontual é uma forma de substituição que substitui um dos nucleotídeos sem alterar o número geral no DNA. Os indels também podem ser contrastados com uma Mutação de Base Tandem (TBM), que pode ser definida como substituição em nucleotídeos adjacentes (principalmente substituições em dois nucleotídeos adjacentes, mas foram observadas substituições em três nucleotídeos adjacentes).

[0093] O termo "variante" refere-se a uma sequência de ácido nucleico que é diferente de uma referência de ácido nucleico. A variante da sequência de ácidos nucleicos típica inclui, sem limitação, o polimorfismo de nucleotídeo único (SNP), polimorfismos de deleção e inserção curtos (Indel), variação do número de cópias (CNV), marcadores microssatélites ou repetições em tandem curtas e variação estrutural. A chamada de variante

somática é o esforço para identificar variantes presentes em baixa frequência na amostra de DNA. A chamada de variantes somática é de interesse no contexto do tratamento do câncer. O câncer é causado por um acúmulo de mutações no DNA. Uma amostra de DNA de um tumor é geralmente heterogênea, incluindo algumas células normais, algumas células em um estágio inicial da progressão do câncer (com menos mutações) e algumas células em estágio avançado (com mais mutações). Devido a essa heterogeneidade, ao sequenciar um tumor (por exemplo, a partir de uma amostra de FFPE), mutações somáticas geralmente aparecem em baixa frequência. Por exemplo, um SNV pode ser visto em apenas 10% das leituras que abrangem uma determinada base. Uma variante que deve ser classificada como somática ou de linhagem germinativa pelo classificador de variantes também é referida neste documento como a "variante em teste".

[0094] O termo "ruído" refere-se a uma chamada de variante incorreta resultante de um ou mais erros no processo de sequenciamento e/ou no pedido de chamada de variante.

[0095] O termo "frequência de variante" representa a frequência relativa de um alelo (variante de um gene) em um locus específico de uma população, expresso como uma fração ou porcentagem. Por exemplo, a fração ou porcentagem pode ser a fração de todos os cromossomos da população que carrega esse alelo. A título de exemplo, a frequência da variante da amostra representa a frequência relativa de um alelo/variante em um determinado locus/posição ao longo de uma sequência genômica de interesse sobre uma "população" correspondente ao número de leituras e/ou amostras obtidas para a sequência genômica de interesse de um indivíduo. Como outro exemplo, uma frequência de variante de linha de base representa a frequência relativa de um alelo/variante em um locus/posição específica ao longo de uma ou mais sequências genômicas de linha de base em que a "população" corresponde ao número de leituras e/ou amostras obtidas para o um ou mais sequências genômicas de linha de base de uma

população de indivíduos normais.

[0096] O termo "frequência alélica da variante (VAF)" refere-se à porcentagem de leituras sequenciadas observadas correspondentes à variante dividida pela cobertura geral na posição alvo. VAF é uma medida da proporção de leituras sequenciadas que carregam a variante.

[0097] Os termos "posição", "posição designada" e "locus" se referem a um local ou coordenada de um ou mais nucleotídeos dentro de uma sequência de nucleotídeos. Os termos "posição", "posição designada" e "locus" também se referem a um local ou coordenada de um ou mais pares de bases em uma sequência de nucleotídeos.

[0098] O termo "haplótipo" refere-se a uma combinação de alelos em locais adjacentes em um cromossomo que são herdados juntos. Um haplótipo pode ser um locus, vários loci ou um cromossomo inteiro, dependendo do número de eventos de recombinação que ocorreram entre um determinado conjunto de loci, se houver algum.

[0099] O termo "limiar" neste documento refere-se a um valor numérico ou não numérico que é usado como ponto de corte para caracterizar uma amostra, um ácido nucleico ou uma porção da mesma (por exemplo, uma leitura). Um limiar pode variar com base na análise empírica. O limiar pode ser comparado a um valor medido ou calculado para determinar se a fonte que gera esse valor sugere que deve ser classificada de uma maneira específica. Os valores de limiar podem ser identificados empiricamente ou analiticamente. A escolha de um limiar depende do nível de confiança que o usuário deseja ter para fazer a classificação. O limiar pode ser escolhido para uma finalidade específica (por exemplo, equilibrar a sensibilidade e seletividade). Conforme usado neste documento, o termo "limiar" indica um ponto no qual um curso de análise pode ser alterado e/ou um ponto no qual uma ação pode ser acionada. Não é necessário que um limiar seja um número predeterminado. Em vez disso, o limiar pode ser, por exemplo, uma função baseada em uma pluralidade de fatores. O limiar pode

ser adaptável às circunstâncias. Além disso, um limiar pode indicar um limite superior, um limite inferior ou um intervalo entre os limites.

[00100] Em algumas implementações, uma métrica ou pontuação baseada em dados de sequenciamento pode ser comparada ao limiar. Conforme usado neste documento, os termos "métrica" ou "pontuação" podem incluir valores ou resultados que foram determinados a partir dos dados de sequenciamento ou podem incluir funções baseadas nos valores ou resultados que foram determinados a partir dos dados de sequenciamento. Como um limiar, a métrica ou a pontuação pode ser adaptável às circunstâncias. Por exemplo, a métrica ou a pontuação pode ser um valor normalizado. Como exemplo de uma pontuação ou métrica, uma ou mais implementações podem usar pontuações de contagem ao analisar os dados. Uma pontuação de contagem pode ser baseada no número de leituras de amostra. As leituras de amostra podem ter passado por um ou mais estágios de filtragem, de modo que as leituras de amostra tenham pelo menos uma característica ou qualidade comum. Por exemplo, cada uma das leituras de amostra usadas para determinar uma pontuação de contagem pode ter sido alinhada com uma sequência de referência ou pode ser atribuída como um alelo em potencial. O número de leituras de amostra com uma característica comum pode ser contado para determinar uma contagem de leituras. As pontuações de contagem podem ser baseadas na contagem de leitura. Em algumas implementações, a pontuação da contagem pode ser um valor igual à contagem de leitura. Em outras implementações, a pontuação da contagem pode ser baseada na contagem de leitura e em outras informações. Por exemplo, uma pontuação de contagem pode ser baseada na contagem de leitura de um alelo específico de um locus genético e um número total de leituras para o locus genético. Em algumas implementações, a pontuação da contagem pode ser baseada na contagem de leitura e nos dados obtidos anteriormente para o locus genético. Em algumas implementações, as pontuações de contagem podem ser

pontuações normalizadas entre valores predeterminados. A pontuação da contagem também pode ser uma função das contagens de leitura de outros loci de uma amostra ou uma função das contagens de leitura de outras amostras que foram executadas simultaneamente com a amostra de interesse. Por exemplo, a pontuação da contagem pode ser uma função da contagem de leitura de um alelo específico e das contagens de leitura de outros loci na amostra e/ou das contagens de outras amostras. Como um exemplo, as contagens de leitura de outros loci e/ou as contagens de leitura de outras amostras podem ser usadas para normalizar a pontuação de contagem para o alelo específico.

[00101] Os termos "cobertura" ou "cobertura de fragmento" se referem a uma contagem ou outra medida de um número de leituras de amostra para o mesmo fragmento de uma sequência. Uma contagem de leitura pode representar uma contagem do número de leituras que cobrem um fragmento correspondente. Como alternativa, a cobertura pode ser determinada pela multiplicação da contagem de leituras por um fator designado que se baseia no conhecimento histórico, no conhecimento da amostra, conhecimento do lócus etc.

[00102] O termo “profundidade de leitura” (convencionalmente um número seguido de “×”) refere-se ao número de leituras sequenciadas com alinhamento sobreposto na posição alvo. Isso geralmente é expresso como uma média ou porcentagem que excede um ponto de corte em um conjunto de intervalos (como éxons, genes ou painéis). Por exemplo, um relatório clínico pode dizer que a cobertura média do painel é 1,105 × com 98% das bases direcionadas cobertas >100 ×.

[00103] Os termos “pontuação de qualidade da chamada de base” ou “pontuação Q” se referem a uma probabilidade em escala PHRED variando de 0 a 20 inversamente proporcional à probabilidade de que uma única base sequenciada esteja correta. Por exemplo, uma chamada de base T com Q de 20 é considerada provavelmente correta com um valor P de

confiança de 0,01. Qualquer geração de base com Q <20 deve ser considerada de baixa qualidade, e qualquer variante identificada onde uma proporção substancial de leituras sequenciadas que suportam a variante são de baixa qualidade deve ser considerada potencialmente falsa positiva.

[00104] Os termos "leituras de variantes" ou "número de leitura de variantes" se referem ao número de leituras sequenciadas que suportam a presença da variante.

DeepPOLY

[00105] Descrevemos o DeepPOLY, um framework baseado em aprendizagem profunda para identificar padrões de sequência que causam erros específicos de sequência (SSEs). O sistema e os processos são descritos com referência à **FIGURA 1**. Como a **FIGURA 1** é um diagrama arquitetural, certos detalhes são intencionalmente omitidos para melhorar a clareza da descrição. A discussão da **FIGURA 1** está organizado da seguinte forma. Primeiro, os módulos da figura são introduzidos, seguidos por suas interconexões. Em seguida, o uso dos módulos é descrito em mais detalhes.

[00106] A **FIGURA1** inclui o sistema **100**. O sistema **100** inclui um filtro de variante **111** (também referido neste documento como um subsistema de filtro de variante), um preparador de entrada **161**(também referido neste documento como um subsistema de preparação de entrada), um simulador **116** (também referido neste documento como um subsistema de simulação), um analisador **194** (também referido neste documento como um subsistema de análise), um banco de dados de padrões de repetição **196**, um banco de dados de sequências de nucleotídeos **169**, um banco de dados de amostras sobrepostas **119** e um emissor de padrões de repetição **198** (também referido neste documento como um subsistema de saída de padrões de repetição).

[00107] Os mecanismos de processamento e bancos de dados da **FIGURA 1**, designado como módulos, pode ser implementado em hardware ou software e não precisa ser dividido exatamente nos mesmos blocos

conforme mostrados na **FIGURA 1**. Alguns dos módulos também podem ser implementados em diferentes processadores, computadores ou servidores ou distribuídos por vários processadores, computadores ou servidores diferentes. Além disso, será apreciado que alguns dos módulos podem ser combinados, operados em paralelo ou em uma sequência diferente da mostrada na **FIGURA 1** sem afetar as funções alcançadas. Os módulos na **FIGURA 1** também pode ser pensado como etapas do fluxograma em um método. Um módulo também não precisa necessariamente ter todo o seu código disposto contiguamente na memória; algumas partes do código podem ser separadas de outras partes do código com o código de outros módulos ou outras funções dispostas no meio.

[00108] As interconexões dos módulos do ambiente **100** são agora descritas. A(s) rede(s) **114** acopla(m) os mecanismos de processamento e os bancos de dados, todos em comunicação entre si (indicados por linhas sólidas de setas duplas). O caminho de comunicação real pode ser ponto a ponto através de redes públicas e/ou privadas. As comunicações podem ocorrer em uma variedade de redes, por exemplo, redes privadas, VPN, circuito MPLS ou Internet e podem usar interfaces de programação de aplicativos (APIs) apropriadas e formatos de intercâmbio de dados, por exemplo, Transferência de Estado Representacional (REST), Notificação de Objeto JavaScript (JSON), Linguagem de Marcação Extensível (XML), Protocolo Simples de Acesso a Objetos (SOAP), Serviço de Mensagens do Java (JMS) e/ou Sistema Modular de Plataforma. Todas as comunicações podem ser criptografadas. A comunicação geralmente é feita através de uma rede como LAN (rede local), WAN (rede ampla), rede telefônica (Rede Telefônica Pública Comutada (PSTN)), Protocolo de Início de Sessão (SIP), rede sem fio, rede ponto a ponto, rede em estrela, rede de token ring, rede de hub, Internet, inclusive a Internet móvel, por meio de protocolos como EDGE, 3G, 4G LTE, Wi-Fi e WiMAX. Além disso, uma variedade de técnicas de autorização e autenticação, como nome de usuário/senha, autorização

aberta (OAuth), Kerberos, SecureID, certificados digitais e muito mais, podem ser usadas para proteger as comunicações.

Processo de Sequenciamento

[00109] As implementações estabelecidas neste documento podem ser aplicáveis à análise de sequências de ácidos nucleicos para identificar variações de sequência. As implementações podem ser usadas para analisar possíveis variantes/alelos de uma posição/locus genéticos e determinar um genótipo do locus genético ou, em outras palavras, fornecer uma geração de genótipo para o locus. A título de exemplo, as sequências de ácidos nucleicos podem ser analisadas de acordo com os métodos e sistemas descritos na Publicação de Pedido de Patente US Nº 2016/0085910 e na Publicação de Pedido de Patente US Nº 2013/0296175, cujo objeto completo é expressamente incorporado por referência neste documento em sua totalidade.

[00110] Em uma implementação, um processo de sequenciamento inclui o recebimento de uma amostra que inclui ou é suspeita de incluir ácidos nucleicos, como o DNA. A amostra pode ser de uma fonte conhecida ou desconhecida, como um animal (por exemplo, humano), planta, bactéria ou fungo. A amostra pode ser coletada diretamente da fonte. Por exemplo, sangue ou saliva podem ser coletados diretamente de um indivíduo. Alternativamente, a amostra pode não ser obtida diretamente da fonte. Em seguida, um ou mais processadores direcionam o sistema para preparar a amostra para o sequenciamento. A preparação pode incluir remover material estranho e/ou isolar certo material (por exemplo, DNA). A amostra biológica pode ser preparada para incluir características para um ensaio particular. Por exemplo, a amostra biológica pode ser preparada para sequenciamento por síntese (SBS). Em certas implementações, a preparação pode incluir amplificação de certas regiões de um genoma. Por exemplo, a preparação pode incluir amplificar loci genéticos predeterminados que são conhecidos por incluir STRs e/ou SNPs. Os loci genéticos podem ser amplificados

utilizando sequências iniciadoras predeterminadas.

[00111] Em seguida, os um ou mais processadores direcionam o sistema para sequenciar a amostra. O sequenciamento pode ser realizado através de uma variedade de protocolos conhecidos de sequenciamento. Em implementações específicas, o sequenciamento inclui SBS. No SBS, uma pluralidade de nucleotídeos marcados com fluorescência é usada para sequenciar uma pluralidade de aglomerados de DNA amplificado (possivelmente milhões de aglomerados) presentes na superfície de um substrato óptico (por exemplo, uma superfície que pelo menos parcialmente define um canal em uma célula de fluxo). As células de fluxo podem conter amostras de ácido nucleico para sequenciamento, onde as células de fluxo são colocadas dentro dos suportes de células de fluxo apropriados.

[00112] Os ácidos nucleicos podem ser preparados de modo a compreender uma sequência primer conhecida que é adjacente a uma sequência alvo desconhecida. Para iniciar o primeiro ciclo de sequenciamento de SBS, um ou mais nucleotídeos marcados de maneira diferente e DNA polimerase, etc., podem ser escoados para/através da célula de fluxo por um sub-sistema de fluxo fluido. Um único tipo de nucleotídeo pode ser adicionado de cada vez, ou os nucleotídeos usados no procedimento de sequenciamento podem ser especialmente projetados para possuir uma propriedade de terminação reversível, permitindo assim que cada ciclo da reação de sequenciamento ocorra simultaneamente na presença de vários tipos de nucleotídeos marcados (por exemplo, A, C, T, G). Os nucleotídeos podem incluir porções marcadoras detectáveis, como fluoróforos. Onde os quatro nucleotídeos são misturados, a polimerase é capaz de selecionar a base correta a incorporar e cada sequência é estendida por uma única base. Os nucleotídeos não incorporados podem ser removidos por lavagem, fluindo uma solução de lavagem através da célula de fluxo. Um ou mais lasers podem excitar os ácidos nucleicos e induzir fluorescência. A fluorescência emitida a partir dos ácidos nucleicos é

baseada nos fluoróforos da base incorporada e diferentes fluoróforos podem emitir diferentes comprimentos de onda da luz de emissão. Um reagente de desbloqueio pode ser adicionado à célula de fluxo para remover grupos terminadores reversíveis das cadeias de DNA que foram estendidas e detectadas. O reagente de desbloqueio pode então ser lavado fluindo uma solução de lavagem através da célula de fluxo. A célula de fluxo está então pronta para um ciclo adicional de sequenciamento começando com a introdução de um nucleotídeo marcado conforme estabelecido acima. As operações fluídicas e de detecção podem ser repetidas várias vezes para concluir uma execução de sequenciamento. Exemplos de métodos de sequenciamento são descritos, por exemplo, em Bentley et al., Nature 456: 53-59 (2008), Publicação Internacional No. WO 04/018497; Pat. U.S. Nº 7,057,026; Publicação Internacional Nº WO 91/06678; Publicação Internacional Nº WO 07/123744; Pat. U.S. Nº 7,329,492; Patente US Nº 7,211,414; Patente US Nº 7,315,019; Patente US Nº 7,405,281 e Publicação de Pedido de Patente US Nº 2008/0108082, cada uma das quais é incorporada neste documento por referência.

[00113] Em algumas implementações, os ácidos nucleicos podem ser ligados a uma superfície e amplificados antes ou durante o sequenciamento. Por exemplo, a amplificação pode ser realizada usando a amplificação em ponte para formar grupos de ácidos nucleicos em uma superfície. Métodos úteis de amplificação em ponte são descritos, por exemplo, na Patente US Nº 5,641,658; Publicação do Pedido de Patente US Nº 2002/0055100; Patente US Nº 7,115,400; Publicação do Pedido de Patente US Nº 2004/0096853; Publicação do Pedido de Patente US Nº 2004/0002090; Publicação do Pedido de Patente US Nº 2007/0128624; e Publicação do Pedido de Patente US Nº 2008/0009420, cada uma das quais é incorporada neste documento por referência em sua totalidade. Outro método útil para amplificar ácidos nucleicos em uma superfície é a amplificação por círculo rolante (RCA), por exemplo, conforme descrito em

Lizardi et al., Nat. Genet. 19:225-232 (1998) e Publicação do Pedido de Patente US Nº 2007/0099208 A1, cada um dos quais é incorporado neste documento por referência.

[00114] Um exemplo de protocolo SBS explora nucleotídeos modificados com blocos 3' removíveis, por exemplo, conforme descrito na Publicação Internacional Nº WO 04/018497, Publicação do Pedido de Patente US Nº 2007/0166705A1 e Patente US Nº 7,057,026, cada uma das quais é incorporada neste documento por referência. Por exemplo, ciclos repetidos de reagentes SBS podem ser entregues a uma célula de fluxo com ácidos nucleicos alvo ligados a eles, por exemplo, como resultado do protocolo de amplificação em ponte. Os aglomerados de ácidos nucleicos podem ser convertidos na forma de cadeia simples usando uma solução de linearização. A solução de linearização pode conter, por exemplo, uma endonuclease de restrição capaz de clivar uma cadeia de cada agrupamento. Outros métodos de clivagem podem ser usados como uma alternativa às enzimas de restrição ou enzimas de corte, incluindo, entre outros, a clivagem química (por exemplo, clivagem de uma ligação diol com periodato), clivagem de sítios abásicos por clivagem com endonuclease (por exemplo, "USER", conforme fornecido por NEB, Ipswich, Mass., EUA, número de peça M5505S), por exposição ao calor ou álcalis, clivagem de ribonucleotídeos incorporados em produtos de amplificação compreendidos de outro modo por desoxirribonucleotídeos, clivagem fotoquímica ou clivagem de um ligante peptídico. Após a operação de linearização, um primer de sequenciamento pode ser entregue à célula de fluxo sob condições para hibridação do primer de sequenciamento com os ácidos nucleicos alvo que devem ser sequenciados.

[00115] Uma célula de fluxo pode então ser contatada com um reagente de extensão SBS possuindo nucleotídeos modificados com blocos removíveis 3' e marcadores fluorescentes sob condições para estender um primer hibridizado com cada ácido nucleico alvo por uma única adição de

nucleotídeo. Apenas um único nucleotídeo é adicionado a cada primer, porque uma vez que o nucleotídeo modificado foi incorporado à cadeia polinucleotídica crescente complementar à região do molde que está sendo sequenciado, não há grupo 3'-OH livre disponível para direcionar a extensão da sequência adicional e, portanto, a polimerase não pode adicionar mais nucleotídeos. O reagente de extensão SBS pode ser removido e substituído por reagente de varredura contendo componentes que protegem a amostra sob excitação com radiação. Exemplos de componentes para reagentes de varredura são descritos na Publicação do Pedido de Patente US Nº 2008/0280773 A1 e no Pedido de Patente US Nº 13/018,255, cada um dos quais é incorporado neste documento por referência. Os ácidos nucleicos estendidos podem então ser detectados por fluorescência na presença do reagente de varredura. Uma vez detectada a fluorescência, o bloco 3' pode ser removido usando um reagente de desbloqueio adequado ao grupo de blocos utilizado. Exemplos de reagentes de desbloqueio que são úteis para os respectivos grupos de blocos são descritos em WO004018497, US 2007/0166705A1 e Patente US Nº 7,057,026, cada uma das quais é incorporada neste documento por referência. O reagente de desbloqueio pode ser lavado, deixando os ácidos nucleicos alvo hibridizados com primers estendidos com grupos 3'-OH que são agora competentes para a adição de um nucleotídeo adicional. Consequentemente, os ciclos de adição de reagente de extensão, reagente de varredura e reagente de desbloqueio, com lavagens opcionais entre uma ou mais das operações, podem ser repetidos até que uma sequência desejada seja obtida. Os ciclos acima podem ser realizados usando uma operação de entrega de reagente de extensão única por ciclo quando cada um dos nucleotídeos modificados tem um marcador diferente ligado a ele, conhecido por corresponder à base particular. Os diferentes marcadores facilitam a discriminação entre os nucleotídeos adicionados durante cada operação de incorporação. Alternativamente, cada ciclo pode incluir operações separadas de entrega de

reagentes de extensão seguidas por operações separadas de entrega e detecção de reagentes de varredura, caso em que dois ou mais dos nucleotídeos podem ter o mesmo marcador e podem ser distinguidos com base na ordem de entrega conhecida.

[00116] Embora a operação de sequenciamento tenha sido discutida acima com relação a um protocolo SBS específico, será entendido que outros protocolos para sequenciar qualquer uma de uma variedade de outras análises moleculares podem ser realizados conforme desejado.

[00117] Em seguida, um ou mais processadores do sistema recebem os dados de sequenciamento para análise subsequente. Os dados de sequenciamento podem ser formatados de várias maneiras, como em um arquivo .BAM. Os dados de sequenciamento podem incluir, por exemplo, um número de leituras de amostra. Os dados de sequenciamento podem incluir uma pluralidade de leituras de amostra que possuem sequências de amostra correspondentes dos nucleotídeos. Embora apenas uma leitura de amostra seja discutida, deve-se entender que os dados de sequenciamento podem incluir, por exemplo, centenas, milhares, centenas de milhares ou milhões de leituras de amostra. Diferentes leituras de amostras podem ter diferentes números de nucleotídeos. Por exemplo, uma amostra de leitura pode variar entre 10 nucleotídeos e cerca de 500 nucleotídeos ou mais. As leituras de amostra podem abranger todo o genoma da(s) fonte(s). Como um exemplo, as leituras da amostra são direcionadas para loci genéticos predeterminados, como aqueles loci genéticos com suspeita de STRs ou suspeita de SNPs.

[00118] Cada leitura de amostra pode incluir uma sequência de nucleotídeos, que pode ser referida como uma sequência de amostra, fragmento de amostra ou uma sequência alvo. A sequência de amostra pode incluir, por exemplo, sequências primer, sequências de flanqueamento e uma sequência alvo. O número de nucleotídeos dentro da sequência de amostra pode incluir 30, 40, 50, 60, 70, 80, 90, 100 ou mais. Em algumas implementações, uma ou mais das leituras de amostra (ou sequências de

amostras) incluem pelo menos 150 nucleotídeos, 200 nucleotídeos, 300 nucleotídeos, 400 nucleotídeos, 500 nucleotídeos ou mais. Em algumas implementações, as leituras de amostra podem incluir mais de 1000 nucleotídeos, 2000 nucleotídeos ou mais. As leituras da amostra (ou as sequências da amostra) podem incluir sequências primer em uma ou nas duas extremidades.

[00119] Em seguida, os um ou mais processadores analisam os dados de sequenciamento para obter chamadas de variante em potencial e uma frequência de variante de amostra dos chamadas de variante de amostra. A operação também pode ser referida como um aplicativo de chamada de variante ou chamador de variante. Assim, o chamador de variantes identifica ou detecta variantes e o classificador de variantes classifica as variantes detectadas como somáticas ou de linhagens germinativas. Os chamadores de variantes alternativos podem ser utilizados de acordo com as implementações deste documento, em que os diferentes chamadores de variantes podem ser utilizados com base no tipo de operação de sequenciamento que está sendo executada, com base em características da amostra que são de interesse e similares. Um exemplo não limitativo de um aplicativo de chamada de variante, como o aplicativo Pisces™ da Illumina Inc. (San Diego, CA) hospedado em https://github.com/Illumina/Pisces e descrito no artigo Dunn, Tamsen & Berry, Gwenn & Emig-Agius, Dorothea & Jiang, Yu & Iyer, Anita & Udar, Nitin & Strömberg, Michael. (2017). Pisces: An Accurate and Versatile Single Sample Somatic and Germline Variant Caller. 595-595. 10.1145/3107411.3108203, cujo objeto completo é expressamente incorporado no presente documento por referência na sua totalidade.

[00120] Esse aplicativo de chamada de variante pode compreender quatro módulos executados sequencialmente:

[00121] (1) Concatenador de Leitura Pisces: reduz o ruído ao concatenar leituras pareadas em um BAM (leitura uma e leitura dois da

mesma molécula) em leituras de consenso. A saída é um BAM concatenado.

[00122] (2) Chamador de Variante Pisces: gera SNVs pequenos, inserções e deleções. Pisces inclui um algoritmo de colapso de variante para unir variantes divididas por limites de leitura, algoritmos básicos de filtragem e um algoritmo simples de pontuação de confiança baseado em Poisson. A saída é um VCF.

[00123] (3) Recalibrador de Qualidade de Variante Pisces (VQR): no caso dos chamadores de variantes seguirem predominantemente um padrão associado a danos térmicos ou desaminação de FFPE, a etapa VQR rebaixará a pontuação de variante Q dos chamadores de variantes suspeitas. A saída é um VCF ajustado.

[00124] (4) Faseador de Variante Pisces (Scylla): usa um método de agrupamento ganancioso com suporte de leitura para montar pequenas variantes em alelos complexos a partir de subpopulações clonais. Isso permite a determinação mais precisa das consequências funcionais pelas ferramentas a jusante. A saída é um VCF ajustado.

[00125] Adicionalmente ou alternativamente, a operação pode utilizar o aplicativo Strelka™ do aplicativo de geração variante da Illumina Inc. hospedado em https://github.com/Illumina/strelka e descrito no artigo T Saunders, Christopher & Wong, Wendy & Swamy, Sajani & Becq, Jennifer e J. Murray, Lisa e Cheetham, Keira. (2012). Strelka: Accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics (Oxford, England). 28. 1811-7. 10.1093/bioinformatics/bts271, cujo objeto completo é expressamente incorporado no presente documento por referência na sua totalidade. Além disso, adicionalmente ou alternativamente, a operação pode utilizar o aplicativo Strelka2™ do aplicativo de geração de variante da Illumina Inc. hospedado em https://github.com/Illumina/strelka e descrito no artigo Kim, S., Scheffler, K., Halpern, A.L., Bekritsky, M.A., Noh, E., Källberg, M., Chen, X., Beyter, D., Krusche, P., and Saunders, C.T. (2017). Strelka2: Fast and accurate variant

calling for clinical sequencing applications, the complete subject matter of which is expressly incorporated herein by reference in its entirety. Além disso, adicionalmente ou alternativamente, a operação pode utilizar uma ferramenta de anotação/geração de variante, como o aplicativo Nirvana™ da Illumina Inc. hospedado em https://github.com/Illumina/Nirvana/wiki e descrito no artigo Stromberg, Michael & Roy, Rajat & Lajugie, Julien & Jiang, Yu & Li, Haochen & Margulies, Elliott. (2017). Nirvana: Clinical Grade Variant Annotator. 596-596. 10.1145/3107411.3108204, cujo objeto completo é expressamente incorporado no presente documento por referência na sua totalidade.

[00126] Essa ferramenta de anotação/geração de variante pode aplicar diferentes técnicas algorítmicas, como as divulgadas no Nirvana:

[00127] a. Identificação de todas as transcrições sobrepostas com Matriz de Intervalo: Para anotação funcional, podemos identificar todas as transcrições sobrepostas a uma variante e uma árvore de intervalo pode ser usada. No entanto, como um conjunto de intervalos pode ser estático, conseguimos otimizá-lo ainda mais para uma Matriz de Intervalo. Uma árvore de intervalo retorna todas as transcrições sobrepostas em O (min (n, k lg n)) tempo, em que n é o número de intervalos na árvore e k é o número de intervalos sobrepostos. Na prática, como k é realmente pequeno comparado a n para a maioria das variantes, o tempo de execução efetivo na árvore de intervalos seria O (k lg n). Aprimoramos para O (lg n + k) criando uma matriz de intervalos em que todos os intervalos são armazenados em uma matriz classificada, de forma que apenas precisamos encontrar o primeiro intervalo sobreposto e depois enumerar até o restante (k-1).

[00128] b. CNVs/SVs (Yu): podem ser fornecidas anotações para variação do número de cópias e variantes estruturais. Semelhante à anotação de pequenas variantes, transcrições sobrepostas ao SV e também variantes estruturais relatadas anteriormente podem ser anotadas em bancos de dados online. Diferentemente das pequenas variantes, nem todas

as transcrições sobrepostas precisam ser anotadas, pois muitas transcrições serão sobrepostas com SVs grandes. Em vez disso, podem ser anotados todos os transcritos sobrepostos que pertencem a um gene sobreposto parcial. Especificamente, para essas transcrições, os íntrons, exons e as consequências danificadas causadas pelas variantes estruturais podem ser relatados. Está disponível uma opção para permitir a saída de todas as transcrições sobrepostas, mas a informação básica para essas transcrições pode ser relatada, como símbolo do gene, sinalizando se é sobreposição canônica ou parcialmente sobreposta às transcrições. Para cada SV/CNV, também é interessante saber se essas variantes foram estudadas e suas frequências em diferentes populações. Portanto, relatamos SVs sobrepostos em bancos de dados externos, como 1000 genomas, DGV e ClinGen. Para evitar o uso de um ponto de corte arbitrário para determinar qual SV é sobreposto, em vez disso, todas as transcrições sobrepostas podem ser usadas e a sobreposição recíproca pode ser calculada, ou seja, o comprimento da sobreposição dividido pelo comprimento mínimo dessas duas SVs.

[00129] c. Relatar anotações suplementares: as anotações suplementares são de dois tipos: variantes pequenas e estruturais (SVs). As SVs podem ser modeladas como intervalos e usar a matriz de intervalos discutida acima para identificar SVs sobrepostas. Pequenas variantes são modeladas como pontos e correspondidas por posição e (opcionalmente) alelo. Como tal, elas são pesquisadas usando um algoritmo de pesquisa tipo binário. Como o banco de dados de anotação suplementar pode ser bastante grande, um índice muito menor é criado para mapear as posições dos cromossomos para localização de arquivos onde a anotação suplementar reside. O índice é uma matriz classificada de objetos (composta de posição do cromossomo e localização do arquivo) que podem ser pesquisados binariamente usando a posição. Para manter o tamanho do índice pequeno, várias posições (até uma certa contagem máxima) são compactadas em um

objeto que armazena os valores para a primeira posição e apenas deltas para posições subsequentes. Como usamos a pesquisa binária, o tempo de execução é O(lg n), onde n é o número de itens no banco de dados.

[00130] d. Arquivos em cache VEP

[00131] e Banco de dados de transcrição: os arquivos de banco de dados suplementar (SAdb) e cache de transcrição (cache) e são despejos serializados de objetos de dados, como transcrições e anotações suplementares. Usamos o cache Ensembl VEP como nossa fonte de dados para o cache. Para criar o cache, todas as transcrições são inseridas em uma matriz de intervalo e o estado final da matriz é armazenado nos arquivos de cache. Assim, durante a anotação, precisamos apenas carregar uma matriz de intervalos pré-calculada e realizar pesquisas nela. Como o cache é carregado na memória e a pesquisa é muito rápida (descrita acima), encontrar transcrições sobrepostas é extremamente rápido no Nirvana (com perfil de menos de 1% do tempo de execução total?).

[00132] f. Banco de dados suplementar: as fontes de dados SAdb estão listadas sob material suplementar. O SAdb para pequenas variantes é produzido por uma mesclagem k -way de todas as fontes de dados, de modo que cada objeto no banco de dados (identificado pelo nome e posição de referência) mantenha todas as anotações adicionais relevantes. Os problemas encontrados durante a análise dos arquivos da fonte de dados foram documentados em detalhes na home page do Nirvana. Para limitar o uso da memória, apenas o índice SA é carregado na memória. Esse índice permite uma pesquisa rápida do local do arquivo para uma anotação suplementar. No entanto, como os dados precisam ser buscados no disco, a adição de anotação suplementar foi identificada como o maior gargalo do Nirvana (com perfil de ~ 30% do tempo de execução total).

[00133] g. Ontologia de Sequência e Consequência: A anotação funcional do Nirvana (quando fornecida) segue as diretrizes da Ontologia de Sequência (SO) (http://www.sequenceontology.org/). Em algumas ocasiões,

tivemos a oportunidade de identificar problemas na SO atual e colaborar com a equipe da SO para melhorar o estado da anotação.

[00134] Essa ferramenta de anotação de variantes pode incluir pré-processamento. Por exemplo, o Nirvana incluiu um grande número de anotações de fontes de dados externas, como ExAC, EVS, projeto 1000 Genomes, dbSNP, ClinVar, Cosmic, DGV e ClinGen. Para fazer pleno uso desses bancos de dados, precisamos sanear as informações deles. Implementamos diferentes estratégias para lidar com diferentes conflitos que existem em diferentes fontes de dados. Por exemplo, no caso de várias entradas do dbSNP para a mesma posição e alelo alternativo, juntamos todos os IDs em uma lista de IDs separados por vírgula; se houver várias entradas com diferentes valores de CAF para o mesmo alelo, usamos o primeiro valor de CAF. Para entradas ExAC e EVS conflitantes, consideramos o número de contagens de amostras e a entrada com maior contagem de amostras é usada. No projeto 1000 Genome, removemos a frequência alélica do alelo conflitante. Outro problema são informações imprecisas. Extraímos principalmente as informações de frequências alélicas do 1000 Genome Projects, no entanto, observamos que, para GRCh38, a frequência alélica relatada no campo info não excluiu amostras com o genótipo não disponível, levando a frequências deflacionadas para variantes que não estão disponíveis para todas as amostras. Para garantir a precisão de nossa anotação, usamos todo o genótipo a nível individual para calcular as verdadeiras frequências alélicas. Como sabemos, as mesmas variantes podem ter representações diferentes com base em alinhamentos diferentes. Para garantir que possamos relatar com precisão as informações das variantes já identificadas, precisamos pré-processar as variantes de diferentes recursos para que elas tenham uma representação consistente. Para todas as fontes de dados externas, aparamos alelos para remover nucleotídeos duplicados no alelo de referência e no alelo alternativo. Para o ClinVar, analisamos diretamente o arquivo xml, realizamos um alinhamento

de cinco números primos para todas as variantes, que geralmente é usado no arquivo vcf. Bancos de dados diferentes podem conter o mesmo conjunto de informações. Para evitar duplicatas desnecessárias, removemos algumas informações duplicadas. Por exemplo, removemos variantes no DGV que possui fonte de dados como os projetos 1000 Genome, pois já relatamos essas variantes em 1000 genomes com informações mais detalhadas.

[00135] De acordo com pelo menos algumas implementações, o aplicativo de chamada de variante fornece chamadores para variantes de baixa frequência, chamada de linhagem germinativa e similares. Como exemplo não limitativo, o aplicativo de chamada de variante pode ser executado em amostras apenas de tumor e/ou amostras pareadas normais de tumor. O aplicativo de chamada de variante pode procurar variações de nucleotídeo único (SNV), múltiplas variações de nucleotídeo (MNV), indels e similares. O aplicativo de chamada de variante identifica variantes, enquanto filtra as incompatibilidades devido a erros de sequenciamento ou erro de preparação de amostras. Para cada variante, o chamadores de variante identifica a sequência de referência, uma posição da variante e a(s) sequência(s) de variante em potencial (por exemplo, SNV de A a C, ou deleção AG a A). O aplicativo de chamada de variante identifica a sequência de amostra (ou fragmento de amostra), uma sequência/fragmento de referência e uma chamada de variante como uma indicação de que uma variante está presente. O aplicativo de chamada de variante pode identificar fragmentos não processados e gerar uma designação dos fragmentos não processados, uma contagem do número de fragmentos não processados que verificam a chamada de variante potencial, a posição dentro do fragmento não processado em que ocorreu uma variante de suporte e outras informações relevantes. Exemplos não limitativos de fragmentos brutos incluem um fragmento concatenado duplex, um fragmento concatenado simplex, um fragmento não concatenado duplex e um fragmento não concatenado simplex.

[00136] O aplicativo de chamada de variante pode gerar as chamadas em vários formatos, como em um arquivo .VCF ou .GVCF. Apenas a título de exemplo, o aplicativo de chamada de variante pode ser incluído em um pipeline MiSeqReporter (por exemplo, quando implementado no instrumento sequenciador MiSeq®). Opcionalmente, o aplicativo pode ser implementado com vários fluxos de trabalho. A análise pode incluir um único protocolo ou uma combinação de protocolos que analisam as leituras da amostra de uma maneira designada para obter as informações desejadas.

[00137] Em seguida, os um ou mais processadores executam uma operação de validação em conexão com a chamada de variante potencial. A operação de validação pode ser baseada em uma pontuação de qualidade e/ou em uma hierarquia de testes em camadas, conforme explicado a seguir. Quando a operação de validação autentica ou verifica a chamada de variante potencial, a operação de validação passa as informações da chamada de variante (do aplicativo de chamada de variante) para o gerador de relatório de amostra. Como alternativa, quando a operação de validação invalida ou desqualifica a chamada de variante potencial, a operação de validação passa uma indicação correspondente (por exemplo, um indicador negativo, um indicador de não chamada, um indicador de chamada inválido) para o gerador de relatório de amostra. A operação de validação também pode passar uma pontuação de confiança relacionada a um grau de confiança que a chamada de variante está correta ou a designação de chamada inválida está correta.

[00138] Em seguida, os um ou mais processadores geram e armazenam um relatório de amostra. O relatório de amostra pode incluir, por exemplo, informações sobre uma pluralidade de loci genéticos em relação à amostra. Por exemplo, para cada locus genético de um conjunto predeterminado de loci genéticos, o relatório de amostra pode pelo menos um dentre fornecer uma chamada de genótipo; indicar que uma chamada de genótipo não pode ser feita; fornecer uma pontuação de confiança em uma

certeza da chamada do genótipo; ou indicar possíveis problemas com um ensaio em relação a um ou mais loci genéticos. O relatório de amostra também pode indicar o gênero de um indivíduo que forneceu uma amostra e/ou indicar que a amostra inclui várias fontes. Conforme usado neste documento, um "relatório de amostra" pode incluir dados digitais (por exemplo, um arquivo de dados) de um locus genético ou conjunto predeterminado de locus genético e/ou um relatório impresso do locus genético ou dos conjuntos de loci genéticos. Assim, gerar ou fornecer pode incluir a criação de um arquivo de dados e/ou impressão do relatório de amostra ou exibição do relatório de amostra.

[00139] O relatório de amostra pode indicar que uma chamada de variante foi determinada, mas não foi validada. Quando uma chamada de variante é considerada inválida, o relatório de amostra pode indicar informações adicionais sobre a base para a determinação de não validar a chamada de variante. Por exemplo, as informações adicionais no relatório podem incluir uma descrição dos fragmentos brutos e uma extensão (por exemplo, uma contagem) na qual os fragmentos brutos suportam ou contradizem a chamada de variante. Adicional ou alternativamente, as informações adicionais no relatório podem incluir a pontuação de qualidade obtida de acordo com as implementações descritas neste documento.

Aplicativo de Chamada de Variante

[00140] As implementações divulgadas neste documento incluem a análise de dados de sequenciamento para identificar chamadas de variantes potenciais. A chamada de variante pode ser realizada com dados armazenados para uma operação de sequenciamento realizada anteriormente. Adicionalmente ou alternativamente, pode ser realizado em tempo real enquanto uma operação de sequenciamento está sendo executada. Cada uma das leituras da amostra é atribuída aos loci genéticos correspondentes. As leituras da amostra podem ser atribuídas aos loci genéticos correspondentes com base na sequência dos nucleotídeos da

amostra lida ou, em outras palavras, na ordem dos nucleotídeos na leitura da amostra (por exemplo, A, C, G, T). Com base nessa análise, a amostra lida pode ser designada como incluindo uma possível variante/alelo de um locus genético específico. A leitura da amostra pode ser coletada (ou agregada ou agrupada) com outras leituras da amostra que foram designadas como incluindo possíveis variantes/alelos do locus genético. A operação de atribuição também pode ser referida como uma operação de chamada na qual a leitura da amostra é identificada como possivelmente associada a uma posição/locus genético específico. As leituras de amostra podem ser analisadas para localizar uma ou mais sequências de identificação (por exemplo, sequências primer) de nucleotídeos que diferenciam a leitura de amostra de outras leituras de amostra. Mais especificamente, a(s) sequência(s) de identificação pode(m) identificar a leitura de amostra de outras amostras como estando associada a um locus genético específico.

[00141] A operação de atribuição pode incluir a análise da série de n nucleotídeos da sequência de identificação para determinar se a série de n nucleotídeos da sequência de identificação combina efetivamente com uma ou mais das sequências selecionadas. Em implementações particulares, a operação de atribuição pode incluir a análise dos primeiros n nucleotídeos da sequência de amostra para determinar se os primeiros n nucleotídeos da sequência de amostra correspondem efetivamente a uma ou mais das sequências selecionadas. O número n pode ter uma variedade de valores, que podem ser programados no protocolo ou inseridos por um usuário. Por exemplo, o número n pode ser definido como o número de nucleotídeos da menor sequência de seleção dentro do banco de dados. O número n pode ser um número predeterminado. O número predeterminado pode ser, por exemplo, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29 ou 30 nucleotídeos. No entanto, menos ou mais nucleotídeos podem ser usados em outras implementações. O número n também pode ser

selecionado por um indivíduo, como um usuário do sistema. O número n pode ser baseado em uma ou mais condições. Por exemplo, o número n pode ser definido como o número de nucleotídeos da sequência primer mais curta dentro do banco de dados ou de um número designado, o que for menor em número. Em algumas implementações, um valor mínimo para n pode ser usado, como 15, de modo que qualquer sequência primer que seja menor que 15 nucleotídeos possa ser designada como uma exceção.

[00142] Em alguns casos, a série de n nucleotídeos de uma sequência de identificação pode não corresponder exatamente aos nucleotídeos da sequência de seleção. No entanto, a sequência de identificação pode efetivamente corresponder à sequência de seleção se a sequência de identificação for quase idêntica à sequência de seleção. Por exemplo, a leitura da amostra pode ser chamada para o locus genético se a série de n nucleotídeos (por exemplo, os primeiros n nucleotídeos) da sequência de identificação corresponderem a uma sequência selecionada com não mais que um número designado de incompatibilidades (por exemplo, 3) e/ou um número designado de deslocamentos (por exemplo, 2). Regras podem ser estabelecidas de modo que cada incompatibilidade ou deslocamento possa contar como uma diferença entre a leitura da amostra e a sequência primer. Se o número de diferenças for menor que um número designado, a leitura da amostra poderá ser chamada para o locus genético correspondente (ou seja, atribuído ao locus genético correspondente). Em algumas implementações, uma pontuação correspondente pode ser determinada com base no número de diferenças entre a sequência de identificação da leitura da amostra e a sequência selecionada associada a um locus genético. Se a pontuação correspondente ultrapassar um limiar de correspondência designado, o locus genético que corresponde à sequência selecionada pode ser designado como um locus potencial para a leitura da amostra. Em algumas implementações, análises subsequentes podem ser realizadas para determinar se a leitura da amostra é chamada para o locus

genético.

[00143] Se a leitura da amostra corresponder efetivamente a uma das sequências selecionadas no banco de dados (ou seja, corresponde exatamente ou quase corresponde conforme descrito acima), a leitura da amostra é atribuída ou designada ao locus genético que se correlaciona com a sequência selecionada. Isso pode ser chamado de chamada de locus ou chamada de locus provisória, em que a leitura da amostra é chamada para o locus genético que se correlaciona com a sequência selecionada. No entanto, como discutido acima, uma leitura de amostra de pode ser solicitada para mais de um locus genético. Em tais implementações, análises adicionais podem ser realizadas para chamar ou atribuir a leitura de amostra para apenas um dos potenciais loci genéticos. Em algumas implementações, a leitura de amostra comparada ao banco de dados de sequências de referência é a primeira leitura do sequenciamento de extremidade pareada. Ao executar o sequenciamento de extremidade pareada, é obtida uma segunda leitura (representando um fragmento bruto) que se correlaciona com a leitura de amostra. Após a atribuição, a análise subsequente que é realizada com as leituras atribuídas pode ser baseada no tipo de locus genético que foi chamado para a leitura atribuída.

[00144] Em seguida, as leituras de amostra são analisadas para identificar possíveis variantes chamadas. Entre outras coisas, os resultados da análise identificam a variante chamada potencial, uma frequência de variante de amostra, uma sequência de referência e uma posição na sequência genômica de interesse em que a variante ocorreu. Por exemplo, se um locus genético for conhecido por incluir SNPs, as leituras atribuídas que foram chamadas para o locus genético poderão ser analisadas para identificar os SNPs das leituras atribuídas. Se o locus genético for conhecido por incluir elementos de DNA repetitivo polimórfico, as leituras atribuídas poderão ser analisadas para identificar ou caracterizar os elementos de DNA repetitivo polimórfico nas leituras da amostra. Em algumas implementações,

se uma leitura atribuída corresponder efetivamente a um locus STR e um locus SNP, um aviso ou sinalizador poderá ser atribuído à leitura de amostra. A leitura da amostra pode ser designada como um locus STR e um locus SNP. A análise pode incluir o alinhamento das leituras atribuídas de acordo com um protocolo de alinhamento para determinar sequências e/ou comprimentos das leituras atribuídas. O protocolo de alinhamento pode incluir o método descrito no Pedido de Patente Internacional Nº PCT/US2013/030867 (Publicação Nº WO 2014/142831), depositado em 15 de março de 2013, que é incorporado neste documento por referência na sua totalidade.

[00145] Em seguida, um ou mais processadores analisam fragmentos brutos para determinar se existem variantes de suporte nas posições correspondentes nos fragmentos brutos. Vários tipos de fragmentos brutos podem ser identificados. Por exemplo, o chamador de variante pode identificar um tipo de fragmento bruto que exibe uma variante que valida a variante chamada original. Por exemplo, o tipo de fragmento bruto pode representar um fragmento concatenado duplex, um fragmento concatenado simplex, um fragmento não concatenado duplex ou um fragmento não concatenado simplex. Opcionalmente, outros fragmentos brutos podem ser identificados em vez de ou além dos exemplos anteriores. Em conexão com a identificação de cada tipo de fragmento bruto, o chamador da variante também identifica a posição, dentro do fragmento bruto, na qual a variante de suporte ocorreu, bem como uma contagem do número de fragmentos brutos que exibiram a variante de suporte. Por exemplo, o chamador variante pode gerar uma indicação de que 10 leituras de fragmentos brutos foram identificadas para representar fragmentos concatenados duplex com uma variante de suporte em uma posição específica X. O chamador variante também pode gerar indicação de que cinco leituras de fragmentos brutos foram identificadas para representar fragmentos não concatenados simplex com uma variante de suporte em uma

posição específica Y. O chamador da variante também pode gerar um número de fragmentos brutos que corresponderam às seqências de referência e, portanto, não incluiu uma variante de suporte que, de outra forma, forneceria evidências para validar a variante chamada potencial na sequência genômica de interesse.

[00146] Em seguida, é mantida uma contagem dos fragmentos brutos que incluem variantes de suporte, bem como a posição na qual a variante de suporte ocorreu. Adicionalmente ou alternativamente, pode ser mantida uma contagem dos fragmentos brutos que não incluíram variantes de suporte na posição de interesse (em relação à posição da chamada de variante potencial na leitura da amostra ou fragmento da amostra). Adicional ou alternativamente, uma contagem pode ser mantida de fragmentos brutos que correspondem a uma sequência de referência e não autenticam ou confirmam a chamada de variante potencial. As informações determinadas são geradas para o aplicativo de validação de chamada de variante, incluindo uma contagem e o tipo de fragmentos brutos que suportam a chamada de variante potencial, posições da variância de suporte nos fragmentos brutos, uma contagem dos fragmentos brutos que não suportam a chamada de variante potencial e afins.

[00147] Quando uma chamada de variante potencial é identificada, o processo gera uma indicação da chamada de variante potencial, a sequência da variante, a posição da variante e uma sequência de referência associada a ela. A chamada de variante é designada para representar uma variante "potencial", pois erros podem fazer com que o processo de chamada identifique uma variante falsa. De acordo com as implementações deste documento, a chamada de variante em potencial é analisada para reduzir e eliminar variantes falsas ou falsos positivos. Adicional ou alternativamente, o processo analisa um ou mais fragmentos brutos associados a uma leitura de amostra e gera uma chamada de variante correspondente associada aos fragmentos brutos.

Filtro de variante

[00148] O filtro de variante **111** inclui uma rede neural convolucional (CNN) e uma rede neural totalmente conectada (FCNN). A entrada para o filtro variante **111** são amostras sobrepostas de sequências de nucleotídeos do banco de dados de amostras sobrepostas **119**. As sequências de nucleotídeos do banco de dados de sequências de nucleotídeos **169** são sobrepostas com padrões de repetição do banco de dados de padrões de repetição **196** para gerar amostras sobrepostas. Uma sobreposição **181** sobrepõe padrões de repetição em sequências de nucleotídeos do banco de dados **169** para produzir amostras sobrepostas que são armazenadas no banco de dados de amostras sobrepostas **119**. O simulador **116** alimenta combinações de padrões de repetição sobrepostas em pelo menos 100 sequências de nucleotídeos em pelo menos 100 amostras sobrepostas ao filtro de variante para análise. Quando amostras sobrepostas com padrão de repetição em teste são fornecidas como entrada do filtro variante **111**, o filtro variante **111** gera pontuações de classificação para a probabilidade de que o nucleotídeo variante em cada uma das amostras sobrepostas seja uma variante verdadeira ou uma variante falsa. Finalmente, o analisador **194** causa a exibição das pontuações de classificação como uma distribuição para cada um dos fatores de repetição para apoiar a avaliação da causa de erro específica da sequência pelos padrões de repetição.

Padrões de repetição

[00149] Um gerador de padrões de repetição **171** gera padrões de repetição "rp" usando padrões de homopolímeros ou copolímeros de comprimento "n" com fatores de repetição distintos "m". Os padrões de repetição do homopolímero compreendem uma única base (A, C, G ou T) enquanto os padrões de repetição do copolímero compreendem mais de uma base. Um "padrão de repetição" é gerado pela aplicação de um "fator de repetição (m)" a um "padrão". A relação entre um padrão de comprimento

(n), um fator de repetição (m) e um padrão de repetição (rp) é representada pela equação (1) como:

[00150] padrão * m = rp (1)

[00151] A Tabela 1 apresenta exemplos de padrões de repetição de homopolímeros. O comprimento dos padrões de homopolímeros é um, ou seja, "n = 1".

n = comprimento do padrão	Padrão	m = fator de repetição	Padrão de repetição (rp)
1	A	5	AAAAA (5 As)
1	A	9	AAAAAAAAA (9 As)
1	A	13	AAAAAAAAAAAAA (13 As)
1	A	17	AAAAAAAAAAAAAAAAA (17 As)
1	A	21	AAAAAAAAAAAAAAAAAAAAA (21 As)
1	A	25	AAAAAAAAAAAAAAAAAAAAAAAAA (25 As)
1	C	5	CCCCC (5 Cs)
1	C	9	CCCCCCCCC (9 Cs)
1	C	13	CCCCCCCCCCCCC (13 Cs)
1	C	17	CCCCCCCCCCCCCCCCC (17 Cs)
1	C	21	CCCCCCCCCCCCCCCCCCCCC (21 Cs)
1	C	25	CCCCCCCCCCCCCCCCCCCCCCCCC (25 Cs)
1	T	5	TTTTT (5 Ts)
1	T	9	TTTTTTTTT (9 Ts)
1	T	13	TTTTTTTTTTTTT (13 Ts)
1	T	17	TTTTTTTTTTTTTTTTT (17 Ts)
1	T	21	TTTTTTTTTTTTTTTTTTTTT (21 Ts)

1	T	25	TTTTTTTTTTTTTTTTTTTTTTTTT (25 Ts)
1	G	5	GGGGG (5 Gs)
1	G	9	GGGGGGGGG (9 Gs)
1	G	13	GGGGGGGGGGGGG (13 Gs)
1	G	17	GGGGGGGGGGGGGGGGG (17 Gs)
1	G	21	GGGGGGGGGGGGGGGGGGGGG (21 Gs)
1	G	25	GGGGGGGGGGGGGGGGGGGGGGGGG (25 Gs)

[00152] A tabela 2 apresenta exemplos de padrões de repetição de copolímeros. O comprimento dos padrões de copolímero é maior que um, ou seja, "n> 1".

n = comprimento do padrão	Padrão	m = fator de repetição	Padrão de repetição (rp)
2	AC	1	AC (1 AC)
2	AC	3	ACACAC (3 ACs)
2	AC	5	ACACACACAC (5 ACs)
2	AC	7	ACACACACACACAC (7 ACs)
2	AC	9	ACACACACACACACACAC (9 ACs)
2	AC	11	ACACACACACACACACACACAC (11 ACs)
2	TA	1	TA (1 TA)
2	TA	3	TATATA (3 TAs)
2	TA	5	TATATATATA (5 TAs)
2	TA	7	TATATATATATATA (7 TAs)
2	TA	9	TATATATATATATATATA (9 TAs)
2	TA	11	TATATATATATATATATATATA (11 TAs)

3	AAT	1	AAT (1 AAT)
3	AAT	2	AATAAT (2 AATs)
3	AAT	3	AATAATAAT (3 AATs)
3	AAT	4	AATAATAATAAT (4 AATs)
3	AAT	5	AATAATAATAATAAT (5 AATs)
3	AAT	6	AATAATAATAATAATAAT (6 AATs)
4	CTAT	1	CTAT (1 CTAT)
4	CTAT	2	CTATCTAT (2 CTATs)
4	CTAT	3	CTATCTATCTAT (3 CTATs)
4	CTAT	4	CTATCTATCTATCTAT (4 CTATs)
4	CTAT	5	CTATCTATCTATCTATCTAT (5 CTATs)
4	CTAT	6	CTATCTATCTATCTATCTATCTAT (5 CTATs)

Filtro de variante

[00153] A **FIGURA 2** ilustra um exemplo de arquitetura **200** do filtro de variante **111**. O filtro de variante **111** possui uma estrutura hierárquica construída em uma rede neural convolucional (CNN) e uma rede neural totalmente conectada (FCNN). O DeepPOLY usa o filtro de variantes **111** para testar padrões de sequência conhecidos quanto a seus efeitos na filtragem de variantes. A entrada para o filtro DE variante **111** compreende sequências nucleotídicas de comprimento 101 tendo um nucleotídeo variante no centro e flanqueado à esquerda e à direita por 50 nucleotídeos. Entende-se que sequências nucleotídicas de diferentes comprimentos podem ser usadas como entradas para o filtro de variante **111**.

[00154] A rede neural convolucional compreende camadas de convolução que realizam a operação de convolução entre os valores de entrada e os filtros de convolução (matriz de pesos) que são aprendidos ao longo de muitas iterações de atualização de gradiente durante o treinamento.

[00155] Seja (*m*, *n*) o tamanho do filtro e *W* seja a matriz de pesos,

então uma camada de convolução realiza uma convolução do *W* com a entrada *X* calculando o produto escalar *W* • *x* + *b*, onde *x* é uma instância de *X* e *b* é o viés. O tamanho da etapa pela qual os filtros de convolução deslizam pela entrada é chamado de passada e a área do filtro (*m* × *n*) é chamada de campo receptivo. Um mesmo filtro de convolução é aplicado em diferentes posições da entrada, o que reduz o número de pesos aprendidos. Ele também permite a aprendizagem invariável da localização, ou seja, se existe um padrão importante na entrada, os filtros de convolução o aprendem, não importa onde esteja na sequência. Detalhes adicionais sobre a rede neural convolucional podem ser encontrados em I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio, “CONVOLUTIONAL NETWORKS,” Deep Learning, MIT Press, 2016; J. Wu, “INTRODUCTION TO CONVOLUTIONAL NEURAL NETWORKS,” Nanjing University, 2017; e N. ten DIJKE, “Convolutional Neural Networks for Regulatory Genomics,” Tese de Mestrado, Universiteit Leiden Opleiding Informatica, 17 de junho de 2017, cujo assunto completo é expressamente incorporado neste documento por referência em sua totalidade. A arquitetura de rede neural convolucional ilustrada na **FIGURA 2** tem duas camadas de convolução. A primeira camada de convolução processa a entrada usando 64 filtros de tamanho 3 cada. A saída da primeira camada de convolução é passada por uma camada de normalização de lote.

[00156] A distribuição de cada camada da rede neural convolucional muda durante o treinamento e varia de uma camada para outra. Isso reduz a velocidade de convergência do algoritmo de otimização. A normalização de lotes (Ioffe e Szegedy 2015) é uma técnica para superar esse problema. Denotando a entrada de uma camada de normalização em lote com *x* e sua saída usando *z*, a normalização de lote aplica a seguinte transformação em *x*:

$$z = \frac{x - \mu}{\sqrt{\sigma^2 + \varepsilon}} \gamma + \beta$$

[00157] A normalização de lote aplica a normalização de variação

média na entrada *x* usando μ e σ e a escala linearmente e a desloca usando γ e β. Os parâmetros de normalização μ e σ são calculados para a camada atual no conjunto de treinamento usando um método chamado média móvel exponencial. Em outras palavras, eles não são parâmetros treináveis. Por outro lado, γ e β são parâmetros treináveis. Os valores para μ e σ calculados acima durante o treinamento são usados no passo para frente durante a produção. Uma função de não linearidade da unidade linear retificada (ReLU) é aplicada à saída da camada de normalização de lote para produzir uma saída normalizada. Outros exemplos de funções de não linearidade incluem sigmóide, tangente hiperbólica (tanh) e ReLU com vazamento.

[00158] Uma segunda camada de convolução opera 128 filtros de tamanho 5 na saída normalizada. O exemplo de CNN mostrado na **FIGURA2**, inclui uma camada de nivelamento que nivela a saída da segunda camada de convolução para uma matriz unidimensional que é passada através de um segundo conjunto de normalização de lote e camadas de ativação de ReLU. A saída normalizada da segunda camada de convolução é alimentada à rede neural totalmente conectada (FCNN). A rede neural totalmente conectada compreende camadas totalmente conectadas - cada neurônio recebe entrada de todos os neurônios da camada anterior e envia sua saída para todos os neurônios da próxima camada. Isso contrasta com o funcionamento das camadas convolucionais, onde os neurônios enviam sua saída para apenas alguns neurônios da próxima camada. Os neurônios das camadas totalmente conectadas são otimizados em várias iterações de atualização de gradiente durante o treinamento. Detalhes adicionais sobre a rede neural convolucional podem ser encontrados em I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio, “CONVOLUTIONAL NETWORKS,” Deep Learning, MIT Press, 2016; J. Wu, “INTRODUCTION TO CONVOLUTIONAL NEURAL NETWORKS,” Nanjing University, 2017; e N. ten DIJKE, “Convolutional Neural Networks for Regulatory Genomics,” Tese de Mestrado, Universiteit Leiden Opleiding Informatica, 17 de junho de

2017, cujo assunto completo é expressamento incorporado neste documento por referência em sua totalidade. Uma camada de classificação (por exemplo, camada softmax) após as camadas totalmente conectadas produz pontuações de classificação para a probabilidade de que cada variante candidata na posição de nucleotídeo alvo seja uma variante verdadeira ou uma variante falsa. A camada de classificação pode ser uma camada softmax ou uma camada sigmóide. O número de classes e seu tipo pode ser modificado, dependendo da implementação.

[00159] A **FIGURA 3** mostra uma implementação do pipeline de processamento **300** do filtro de variante **111**. Na implementação ilustrada, a rede neural de convolução (CNN) possui duas camadas de convolução e a rede neural totalmente conectada (FCNN) possui duas camadas totalmente conectadas. Em outras implementações, o filtro de variante **111**e sua rede neural de convolução e rede neural totalmente conectada podem ter mais, menos ou diferentes parâmetros e hiperparâmetros. Alguns exemplos de parâmetros são número de camadas de convolução, número de normalização de lotes e camadas ReLU, número de camadas totalmente conectadas, número de filtros de convolução nas respectivas camadas de convolução, número de neurônios nas respectivas camadas totalmente conectadas, número de saídas produzidas pela camada de classificação final e conectividade residual. Alguns exemplos de hiperparâmetros são o tamanho da janela dos filtros de convolução, comprimento da passada dos filtros de convolução, preenchimento e dilatação. Na discussão abaixo, o termo "camada" refere-se a um algoritmo implementado no código como uma lógica ou módulo de software. Alguns exemplos de camadas podem ser encontrados na documentação do Keras ™ disponível em https://keras.io/layers/about-keras-layers/, cujo assunto completo é expressamente incorporado neste documento por referência em sua totalidade.

[00160] Uma sequência de entrada codificada one-hot **302** é

alimentada a uma primeira camada de convolução **304** da rede neural convolucional (CNN). A dimensionalidade da sequência de entrada **302** é 101, 5, em que 101 representa os 101 nucleotídeos na sequência de entrada **302** com uma variante específica na posição alvo central, flanqueada por 50 nucleotídeos em cada lado e 5 representa os 5 canais A, T, C, G, N usados para codificar a sequência de entrada **302**. A preparação das sequências de entrada **302** é descrita com referência à **FIGURA 5**.

[00161] A primeira camada de convolução **304** tem 64 filtros, cada um dos quais convolui sobre a sequência de entrada **302** com um tamanho de janela de 3 e comprimento de passada de 1. A convolução é seguida por normalização de lote e camadas de não linearidade ReLU **306**. O que resulta é uma saída (mapa de recursos) **308** da dimensionalidade 101, 64. A saída **308** pode ser considerada como o primeiro recurso convoluído intermediário.

[00162] A saída **308** é alimentada como entrada para uma segunda camada de convolução **310** da rede neural convolucional. A segunda camada de convolução **310** tem 128 filtros, cada um dos quais convolui sobre a saída **308** com um tamanho de janela de 5 e comprimento de passada de 1. A convolução é seguida por normalização de lote e camadas de não linearidade ReLU **312**. O que resulta é uma saída (mapa de recursos) **314** da dimensionalidade 101, 128. A saída **314** pode ser considerada como o segundo recurso convoluído intermediário e também a saída final da rede neural convolucional.

[00163] O dropout é uma técnica eficaz para impedir que uma rede neural sobreajuste. Ele funciona retirando aleatoriamente uma fração de neurônios da rede em cada iteração do treinamento. Isso significa que a saída e os gradientes dos neurônios selecionados são definidos como zero, para que não tenham impacto nos passos para frente e para trás. Na **FIGURA3**, o dropout é realizado na camada de dropout**316** usando uma probabilidade de 0,5.

[00164] Depois de processar a saída através da camada dropout,

a saída é nivelada por uma camada niveladora **318** para permitir o processamento a jusante pela rede neural totalmente conectada. O nivelamento inclui a vetorização da saída **314** para ter uma linha ou uma coluna. Isto é, a título de exemplo, converter a saída **314** da dimensionalidade 101, 128 em um vetor nivelado da dimensionalidade 1, 12928 (1 linha e 101x128 = 12928 colunas).

[00165] A saída nivelada da dimensionalidade 1, 12928 da camada niveladora**318** é então alimentada como entrada para a rede neural totalmente conectada (FCNN). A rede neural totalmente conectada possui duas camadas totalmente conectadas **320** e **328**. A primeira camada totalmente conectada **320** tem 128 neurônios, que são totalmente conectados a 2 neurônios na segunda camada totalmente conectada **328**. A primeira camada totalmente conectada **320** é seguida por camadas de normalização de lote, não linearidade ReLU e de dropout de **322** e **326**. A segunda camada totalmente conectada **328** é seguida por uma camada de normalização de lote **330**. A camada de classificação **332** (por exemplo, softmax) possui 2 neurônios que geram as 2 pontuações ou probabilidades de classificação **334** para a variante em particular, sendo uma variante verdadeira ou uma variante falsa.

Desempenho do Chamador de Variante em Dados Retidos

[00166] A **FIGURA 4A** mostra plotagens positivas verdadeiras e falsas que ilustram graficamente o desempenho do filtro de variante em dados retidos. Existem 28,000 exemplos de validação no conjunto de dados retido, com cerca de 14,000 exemplos de validação de variantes verdadeiras (exemplos positivos) e 14,000 exemplos de validação de variantes falsas (exemplos negativos). Os dois gráficos **410** e **416** mostram o desempenho do filtro de variante **111** quando 28,000 exemplos de validação são alimentados como entrada durante o estágio de validação. Os gráficos **410** e **416** plotam as pontuações de classificação ao longo do eixo x, indicando a confiança do modelo treinado em prever as variantes verdadeiras e as

variantes falsas como positivas verdadeiras. Assim, espera-se que o modelo treinado produza altas pontuações de classificação para as variantes verdadeiras e baixas pontuações de classificação para as variantes falsas. A altura das barras verticais indica a contagem de exemplos de validação com as respectivas pontuações de classificação ao longo do eixo x.

[00167] O gráfico **416** mostra que o filtro de variante **111** classificou mais de 7,000 exemplos de validação de variantes falsas como "verdadeiros positivos de baixa confiança" (ou seja, pontuação de classificação <0,5 (por exemplo, **426**)), confirmando que o modelo aprendeu com êxito a classificar exemplos negativos como variantes falsas. O filtro de variantes **111** classificou alguns exemplos de validação de variantes falsas como "positivos verdadeiros de alta confiança" (por exemplo, **468**). Isso ocorreu porque, nos dados de treinamento e/ou nos dados retidos, algumas variantes de novo observadas em apenas uma criança foram classificadas incorretamente como variantes falsas quando na verdade eram variantes verdadeiras.

[00168] O gráfico **410** mostra que o filtro de variante **111** classificou mais de 11,000 exemplos de validação de variantes verdadeiras como "positivos verdadeiros de alta confiança" (isto é, pontuação de classificação > 0,5), confirmando que o modelo aprendeu com êxito a classificar exemplos positivos como variantes verdadeiras.

[00169] Na **FIGURA 4B**, os resultados da classificação do filtro de variante **111** são comparados com a análise derivada de uma imagem acumulada que alinha leituras produzidas por um sequenciador a uma sequência de referência **498**. A sequência de referência **498** compreende um padrão de repetição de homopolímero de comprimento 18 de uma única base "T", conforme mostrado pelo marcador **494** na **FIGURA4B**. A imagem acumulada mostra que pelo menos sete leituras (indicadas pelo rótulo de referência **455**) relataram uma base "T" na posição de um nucleotídeo "G" em relação ao genoma de referência **498**. Portanto, existem duas chamadas resultantes possíveis para chamar a base nesta posição na sequência: “G”

ou “T”. A verdade fundamental da "árvore genealógica dos genomas de platina" mostra que nenhum dos pais e avós tem um nucleotídeo variante nessa posição em suas respectivas sequências de referência. Portanto, a chamada base "T" é determinada como "falso positivo" que ocorreu devido a um erro de sequenciamento. Além disso, a imagem acumulada mostra que os “Ts” aparecem apenas no final da leitura 1, o que confirma ainda mais que a variante é falsa.

[00170] O desempenho do filtro de variante **111** é consistente com a análise acima porque o filtro de variante **111** classificou o nucleotídeo nesta posição como uma variante falsa com uma alta confiança, conforme ilustrado na **FIGURA 4B** por "P (X é Falso) = 0,974398".

[00171] A **FIGURA4C** mostra a imagem acumulada **412** de leituras de sequenciamento para um exemplo que contém uma variante verdadeira. As leituras de sequenciamento para a criança (rotuladas como "NA12881") têm pelo menos três nucleotídeos "T" identificados por um marcador **495**. A sequência de referência tem um nucleotídeo "C" nessa posição, conforme identificado por um marcador **496**. No entanto, as leituras de sequenciamento da mãe indicam pelo menos sete nucleotídeos "T" na mesma posição. Portanto, esta é uma instância de um exemplo com uma variante verdadeira, conforme mostrado pelo gráfico **410** no canto superior esquerdo. O filtro de variante **111** classificou este exemplo como um verdadeiro positivo com um baixo índice de confiança ("P (X é verdadeiro) = 0,304499"). Ou seja, o filtro de variante **111** classificou o nucleotídeo alvo como uma variante falsa (ou fracamente classificada como uma variante verdadeira) devido à presença de um padrão de repetição do copolímero "AC" antes da posição do nucleotídeo alvo. A sequência treinada considera o padrão de repetição como um potencial erro específico da sequência (SSE) e, portanto, classificou a variante “T” com uma baixa pontuação de confiança.

[00172] A **FIGURA5** mostra um exemplo de preparação de entrada pelo preparador de entrada **161** usando codificação one-hot para codificar as

sequências de nucleotídeos sobrepostas que têm um nucleotídeo variante em uma posição alvo para entrada no filtro de variante **111**. Uma sequência nucleotídica **514** compreendendo pelo menos 50 nucleotídeos em ambos os lados (esquerdo e direito) de um nucleotídeo variante na posição alvo é usada para preparar a entrada. Observe que a sequência de nucleotídeos **514** é uma porção do genoma de referência. Na codificação one-hot, cada par de bases em uma sequência é codificado com um vetor binário de quatro bits, com um dos bits sendo hot (ou seja, 1) enquanto o outro é 0. Por exemplo, T = (1, 0, 0, 0), G = (0, 1, 0, 0), C = (0, 0, 1, 0) e A = (0, 0, 0, 1). Em algumas implementações, um nucleotídeo desconhecido é codificado como N = (0, 0, 0, 0). A figura mostra um exemplo de sequência nucleotídica de 101 nucleotídeos representados usando um codificado one-hot.

[00173] A **FIGURA6** ilustra preparações de amostras sobrepostas produzidas pelo preparador de entrada sobrepondo os padrões de repetição nas sequências de nucleotídeos. As amostras sobrepostas são armazenadas no banco de dados de amostras sobrepostas **119**. O exemplo mostra uma amostra sobreposta 1 que é gerada sobrepondo um padrão de repetição de homopolímero de 7 "A"s à esquerda de um nucleotídeo central na posição alvo na amostra sobreposta. Uma amostra sobreposta 2 é criada sobrepondo o mesmo padrão de repetição de 7 "A"s na sequência de nucleotídeos para incluir um nucleotídeo central. Uma terceira amostra sobreposta n é gerada sobrepondo o padrão de repetição de 7 "A"s à direita de um nucleotídeo central nas amostras sobrepostas.

[00174] O subsistema de filtro de variante traduz a análise pelo filtro de variante **111** em pontuações de classificação para a probabilidade de que o nucleotídeo variante em cada uma das amostras sobrepostas seja uma variante verdadeira ou uma variante falsa. O subsistema de filtro de variante é seguido por um subsistema de análise no qual o analisador **194** causa a exibição das pontuações de classificação como uma distribuição para cada um dos fatores de repetição para apoiar a avaliação da causa de erros

específicos da sequência pelos padrões de repetição. As **FIGURAS7A** a **7C** apresentam exemplos de tal exibição a partir do analisador **194**. A **FIGURA7A** usa um diagrama de caixa e bigode para identificar a causa do erro específico da sequência pelo padrão de repetição sobreposto à esquerda de um nucleotídeo central nas amostras sobrepostas.

[00175] O eixo y do diagrama gráfico mostra a distribuição das pontuações de classificação geradas pelo filtro de variante quando as amostras sobrepostas contendo diferentes padrões de repetição foram alimentadas ao filtro variante como entrada. O eixo x mostra os fatores de repetição (m) aplicados ao padrão que produziu o padrão de repetição alimentado como entrada. Os padrões de repetição considerados aqui são homopolímeros gerados usando fatores de repetição indicados no eixo x. O exemplo mostra quatro diagramas de caixa e bigode por valor único do fator de repetição. Os quatro diagramas correspondem aos padrões de repetição do homopolímero dos quatro tipos de nucleotídeos (G, A, T e C). Cada padrão de repetição é colocado em pelo menos 100 sequências de nucleotídeos para gerar 100 amostras sobrepostas alimentadas como entrada no CNN do filtro de variante **111**. Em outra implementação, pelo menos 200 sequências de nucleotídeos são usadas para gerar pelo menos 200 amostras sobrepostas por padrão de repetição. O mesmo processo é repetido para gerar padrões de repetição de homopolímeros para todos os fatores de repetição mostrados ao longo do eixo x.

[00176] O diagrama gráfico na **FIGURA7A** mostra que padrões de repetição mais curtos (comprimento menor que 10 nucleotídeos) de uma única base "G" podem introduzir erros específicos de sequência na identificação de variantes. Da mesma forma, os padrões de repetição mais curtos de uma única base "C" também podem introduzir alguns erros, enquanto os padrões de repetição das bases de nucleotídeos "A" e "T" têm menos probabilidade de causar erros específicos de sequência quando os padrões de repetição são curtos. No entanto, padrões de repetição mais

longos (comprimento maior que 10 nucleotídeos) dos quatro tipos de nucleotídeos causam mais erros específicos de sequência.

[00177] A **FIGURA7B** é um diagrama de caixa e bigode que exibe pontuações de classificação como uma distribuição para a probabilidade de que um nucleotídeo variante seja uma variante verdadeira ou uma variante falsa quando padrões de repetição são sobrepostos em uma sequência de nucleotídeos à direita de um nucleotídeo central nas amostras sobrepostas. Em comparação com a **FIGURA7A**, os padrões mais curtos de homopolímeros de um único nucleotídeo "C" têm maior probabilidade de causar um erro na identificação de uma variante verdadeira. A **FIGURA7C** é um diagrama de caixa e bigode que exibe pontuações de classificação como uma distribuição para a probabilidade de que um nucleotídeo variante seja uma variante verdadeira ou uma variante falsa quando os padrões de repetição incluem um nucleotídeo central (na posição alvo) nas amostras sobrepostas. Em comparação com as **FIGURAS7A** e **7B**, a **FIGURA7C** mostra que padrões de repetição mais curtos de todos os quatro tipos de nucleotídeos têm menos probabilidade de causar um erro específico de sequência na identificação de variantes.

[00178] As **FIGURAS8A** a **8C** apresentam gráficos para identificar a causa de erros específicos de sequência quando os padrões de repetição de homopolímeros de uma única base (A, C, G ou T) são sobrepostos em deslocamentos variados nas sequências de nucleotídeos para produzir amostras sobrepostas. Os deslocamentos variados variam uma posição na qual os padrões de repetição são sobrepostos nas sequências nucleotídicas. O deslocamento variado é mensurável como um deslocamento entre uma posição de origem dos padrões de repetição e uma posição de origem das sequências nucleotídicas. Em uma implementação, pelo menos dez deslocamentos são usados para produzir as amostras sobrepostas. Dez é um piso razoável para gerar amostras sobrepostas com padrões de repetição em uma variedade de deslocamentos para analisar a causa dos erros

específicos da sequência.

[00179] A **FIGURA8A** usa um diagrama de caixa e bigode para identificar a causa de erros específicos de sequência por padrões de repetição de homopolímeros de uma única base "C" sobrepostos a deslocamentos variados nas sequências de nucleotídeos. O fator de repetição m=15, o que significa que o padrão de repetição é um homopolímero de comprimento 15 de uma única base “C”. Este padrão de repetição é sobreposto em sequências de nucleotídeos que consistem em 101 nucleotídeos para gerar amostras sobrepostas em deslocamentos variados. Para cada valor de deslocamento, combinações de padrões de repetição sobrepostas em pelo menos 100 sequências de nucleotídeos em pelo menos 100 amostras sobrepostas são alimentadas ao CNN do filtro de variante da **FIGURA1**. A **FIGURA8A** mostra diagramas de caixa e bigode para posições de deslocamento em 0, 2, 4, até 84 quando o padrão de repetição de 15 bases únicas "C" é sobreposto nas sequências de nucleotídeos. Por exemplo, quando o deslocamento é "0", a posição de origem do padrão de repetição coincide com a posição de origem das sequências de nucleotídeos. No deslocamento "2", a posição de origem do padrão de repetição é alinhada à terceira base (em um índice de 2) para sobrepor o padrão de repetição nas sequências de nucleotídeos. À medida que o deslocamento aumenta, o padrão de repetição sobreposto fica mais próximo do nucleotídeo variante em uma sequência nucleotídica na posição alvo. No exemplo usado para fins de ilustração na **FIGURA8A**, o nucleotídeo alvo está na posição de índice de "50", que é o centro da sequência de nucleotídeos que compreende 101 nucleotídeos. À medida que o valor de deslocamento aumenta acima de 50, o padrão de repetição passa além do nucleotídeo variante e é posicionado no lado direito do nucleotídeo variante na posição alvo.

[00180] As **FIGURAS8B**, **8C**e **8D** são diagramas de caixa e bigode semelhantes aos descritos acima para identificar a causa específica de erros

de sequência por padrões de repetição de homopolímeros de bases únicas "G", "A" e "T", respectivamente, sobrepostos em deslocamentos variáveis nas sequências nucleotídicas. O fator de repetição m=15 para cada um dos três padrões de repetição.

[00181] A **FIGURA9** mostra a exibição de pontuações de classificação como uma distribuição para a probabilidade de que um nucleotídeo variante seja uma variante verdadeira ou uma variante falsa quando padrões de repetição de homopolímeros de uma única base são sobrepostos "antes" e "depois" de um nucleotídeo variante. Os padrões de repetição do homopolímero são sobrepostos um a um antes e depois dos nucleotídeos de variantes na posição alvo para produzir amostras sobrepostas. Um diagrama de caixa e bigode **932** mostra pontuações de classificação quando um padrão de repetição de homopolímero de uma única base "G" é sobreposto à esquerda de um nucleotídeo central em uma sequência de nucleotídeos. Os resultados são gerados para quatro tipos de nucleotídeos (A, C, G e T) como o nucleotídeo variante na posição alvo, seguido pelo padrão de repetição do homopolímero. Os resultados mostram que as pontuações de classificação variam de acordo com uma propagação maior se o nucleotídeo alvo for do tipo "A" e "C".

[00182] Um gráfico **935** mostra uma visualização semelhante, mas para um padrão de repetição de homopolímero de uma única base "C" sobreposta à direita de um nucleotídeo central em uma sequência de nucleotídeos **912**. A comparação dos diagramas de caixa e bigode mostra uma maior propagação de pontuações de classificação quando um nucleotídeo alvo é do tipo "G".

[00183] As **FIGURAS 10A** a **10C** exibem uma exibição de padrões de repetição de copolímeros que ocorrem naturalmente em cada uma das sequências nucleotídicas da amostra que contribuem para uma classificação falsa de variantes. As visualizações gráficas apresentadas nas **FIGURAS10A** a **10C** são gerados usando o DeepLIFT apresentado por

Shrikumar et. el., em seu artigo, “Not Just a Black Box: Learning Important Features Through Propagating Activation Differences” disponível em https://arxiv.org/pdf/1605.01713.pdf (referência 1). A implementação do modelo DeepLIFT é apresentada em http://github.com/kundajelab/deeplift (referência 2) e mais detalhes sobre a implementação do DeepLIFT são apresentados em https://www.biorxiv.org/content/biorxiv/suppl/2017/10/05/105957.DC1/105957-6.pdf (referência 3). Um ou mais padrões de repetição de copolímeros que ocorrem naturalmente, incluindo um nucleotídeo variante na posição alvo, são dados como entrada para o modelo DeepLIFT para gerar as visualizações mostradas nas FIGURAS 10A a 10C. A saída do modelo DeepLIFT são as matrizes de contribuições de entrada para a classificação de variantes de um nucleotídeo variante na posição alvo.

[00184] Por exemplo, considere a sequência de entrada mostrada na visualização gráfica **911**. O nucleotídeo variante **916** está na posição 50 na sequência de nucleotídeos de amostra que compreende 101 nucleotídeos. O nucleotídeo variante na posição alvo é flanqueado por 50 nucleotídeos de cada lado nas posições 0 a 49 e 51 a 100 na sequência nucleotídica da amostra. O filtro de variante 111 da **FIGURA2**, classificou o nucleotídeo variante ("C") na posição alvo como uma variante falsa. A saída do DeepLIFT é a visualização **911**, mostrando que o padrão de repetição de ocorrência natural **917** contribuiu mais para a classificação do nucleotídeo variante **916**. As alturas dos nucleotídeos indicam suas respectivas contribuições para a classificação do nucleotídeo variante. Conforme mostrado na visualização gráfica **911**, a maior contribuição é de uma sequência de nucleotídeos **917** que é um padrão de repetição compreendendo uma única base "A".

[00185] As matrizes de contribuição DeepLIFT têm o mesmo formato da entrada, isto é, a sequência de entrada de nucleotídeos multiplicada por 4 para a codificação one-hot padrão (apresentada na

FIGURA5). Portanto, o DeepLIFT atribui pontuações a cada posição de sequência somando as contribuições dos neurônios de entrada associados a uma posição fixa da sequência e associa essas contribuições somadas ao nucleotídeo presente nessa posição na sequência de nucleotídeos da amostra de entrada. As contribuições somadas são chamadas de "pontuação da interpretação do DeepLIFT". As seguintes práticas melhor recomendadas (como apresentadas na referência 3 acima) são seguidas no pedido do modelo DeepLIFT. As contribuições dos neurônios de entrada para a pré-ativação (ativação antes de aplicar a não linearidade final) de um neurônio de saída são calculadas. Quando uma camada de saída usa uma não linearidade softmax, os pesos que conectam um neurônio de penúltims camada fixa ao conjunto de neurônios de saída são centralizados na média. Como as sequências nucleotídicas da amostra são codificadas one-hot, conforme mostrado na **FIGURA5**, o método de "normalização de peso para entradas restritas" é usado antes da conversão de Keras para DeepLIFT, conforme descrito na referência 3 acima.

[00186] As visualizações gráficas **921**, **931**e **941** mostram padrões de repetição **927**, **934** e **946**, respectivamente, contribuindo mais para a classificação do nucleotídeo variante nas sequências de nucleotídeos da amostra. A **FIGURA10B** inclui visualizações gráficas **921**, **931**, **941** e **951**. Observe que nessas visualizações gráficas os padrões de repetição de copolímeros contêm padrões de dois ou mais nucleotídeos. Da mesma forma, a **FIGURA10C** apresenta mais exemplos de visualizações gráficas **931**, **932**, **933** e **934**, ilustrando uma variedade de padrões de repetição que contribuem para a classificação do nucleotídeo variante na posição alvo nas respectivas sequências de nucleotídeos de entrada.

Sistema Computadorizado

[00187] A **FIGURA 11** é um diagrama de blocos simplificado de um sistema de computador **1100** que pode ser usado para implementar o filtro de variante **111** da **FIGURA 1** para identificar padrões de repetição que

causam erros específicos de sequência. O sistema de computador **1100** inclui pelo menos uma unidade central de processamento (CPU) **1172** que se comunica com um número de dispositivos periféricos via subsistema de barramento **1155**. Estes dispositivos periféricos podem incluir um subsistema de armazenamento **1110**, incluindo, por exemplo, dispositivos de memória e um subsistema de armazenamento de arquivos **1136**, dispositivos de entrada da interface de usuário **1138**, dispositivos de saída da interface de usuário **1176** e um subsistema da interface de rede **1174**. Os dispositivos de entrada e saída permitem a interação do usuário com o sistema computadorizado **1100**. O subsistema da interface de rede **1174** fornece uma interface para redes externas, incluindo uma interface para os dispositivos de interface correspondentes em outros sistemas computadorizados.

[00188] Em uma implementação, o filtro de variante **111** da **FIGURA 1** está comunicativamente ligado ao subsistema de armazenamento **1110** e aos dispositivos de entrada da interface do usuário **1138**.

[00189] Os dispositivos de entrada da interface de usuário **1138** podem incluir um teclado; dispositivos apontadores, como mouse, trackball, touchpad ou mesa digitalizadora; um scanner; uma tela touch incorporada no visor; dispositivos de entrada de áudio, como sistemas de reconhecimento de voz e microfones; e outros tipos de dispositivos de entrada. Em geral, o uso do termo "dispositivo de entrada" deve incluir todos os tipos possíveis de dispositivos e maneiras de inserir informações no sistema computadorizado **1100**.

[00190] Os dispositivos de saída da interface de usuário **1176** podem incluir um subsistema de exibição, uma impressora, uma máquina de fax ou visores sem exibição de imagens, tais como os de dispositivos de saída de áudio. O subsistema de exibição pode incluir um visor de LED, um tubo de raios catódicos (CRT), um dispositivo de tela plana como um monitor de cristal líquido (LCD), um dispositivo de projeção ou algum outro mecanismo para criar uma imagem visível. O subsistema de exibição

também pode fornecer um visor sem exibição de imagens, como dispositivos de saída de áudio. Em geral, o uso do termo "dispositivo de saída" visa incluir todos os tipos possíveis de dispositivos e maneiras de enviar informações do sistema computadorizado **1100** para o usuário ou para outra máquina ou sistema computadorizado.

[00191] O subsistema de armazenamento **1110** armazena construtos de dados e programação que fornecem a funcionalidade de alguns ou todos os módulos e métodos descritos neste documento. O subsistema **1178** pode ser unidades de processamento gráfico (GPUs) ou matrizes de portas programáveis em campo (FPGAs).

[00192] O subsbsistema de memória **1122** usado no subsistema de armazenamento **1110** pode incluir várias memórias, incluindo uma memória de acesso aleatório (RAM) principal **1132** para armazenamento de instruções e dados durante a execução do programa e uma memória somente leitura (ROM) **1134** na qual as instruções fixas são armazenadas. Um subsistema de armazenamento de arquivos **1136** pode fornecer armazenamento persistente para arquivos de programa e dados e pode incluir uma unidade de disco rígido, uma unidade de disquete junto com a mídia removível associada, uma unidade de CD-ROM, uma unidade óptica ou cartuchos de mídia removíveis. Os módulos que implementam a funcionalidade de determinadas implementações podem ser armazenados pelo subsistema de armazenamento de arquivo **1136** no subsistema de armazenamento **1110** ou em outras máquinas acessíveis pelo processador.

[00193] O subsistema de barramento **1155** fornece um mecanismo para permitir que os vários componentes e subsistemas do sistema computadorizado **1100** se comuniquem entre si, conforme o pretendido. Embora o subsistema de barramento **1155** seja mostrado esquematicamente como um barramento único, implementações alternativas do subsistema de barramento podem usar vários barramentos.

[00194] O sistema computadorizado **1100** em si pode ser de vários

tipos, incluindo um computador pessoal, um computador portátil, uma estação de trabalho, um terminal de computador, um computador em rede, uma televisão, um mainframe, um farm de servidores, um conjunto amplamente distribuído de computadores de rede vagamente distribuídos ou qualquer outro sistema de processamento de dados ou dispositivo do usuário. Devido à natureza em constante mudança de computadores e redes, a descrição do sistema de computador **1100** representada na **FIGURA 11** destina-se apenas como um exemplo específico para fins de ilustrar as modalidades preferidas da presente invenção. Muitas outras configurações do sistema computadorizado **1100** são possíveis com mais ou menos componentes do que o sistema computadorizado representado na **FIGURA 11**.

Correlação de Erro Específico da Sequência (SSE)

[00195] A **FIGURA 12** ilustra uma implementação de como os erros específicos de sequência (SSEs) são correlacionados para repetir padrões com base em classificações de variantes falsas.

[00196] O subsistema de preparação de entrada **161** sobrepõe computacionalmente os padrões de repetição sob teste em inúmeras sequências de nucleotídeos e produz as amostras sobrepostas **119**. Cada padrão de repetição representa uma composição nucleotídica específica que tem um comprimento específico e aparece em uma amostra sobreposta em uma posição de deslocamento específica. Cada amostra sobreposta tem uma posição alvo considerada como um nucleotídeo variante. Para cada combinação da composição nucleotídica específica, o comprimento específico e a posição de deslocamento específica, um conjunto de amostras sobrepostas é gerado computacionalmente.

[00197] O subsistema de filtro de variante pré-treinado **111** processa as amostras sobrepostas **119** através da rede neural convolucional **200** e, com base na detecção de padrões de nucleotídeos nas amostras sobrepostas **119** por filtros de convolução da rede neural convolucional **200**,

gera pontuações de classificação **334** para a probabilidade de que o nucleotídeo variante em cada uma das amostras sobrepostas é uma variante verdadeira ou uma variante falsa.

[00198] O subsistema de saída de padrão de repetição **1202** gera distribuições **1212** das pontuações de classificação **334** que indicam suscetibilidade do subsistema de filtro de variantes **pré-treinado 111** a classificações de variantes falsas resultantes da presença dos padrões de repetição.

[00199] O subsistema de correlação de erro específico da sequência **199** especifica, com base em um limiar **1222**, um subconjunto das pontuações de classificação como indicativo das classificações de variantes falsas e classifica os padrões de repetição **1232** que estão associados ao subconjunto das pontuações de classificação que são indicativos das classificações de variantes falsas como causadoras dos erros específicos da sequência. O subsistema de correlação de erro específico da sequência **199** classifica comprimentos particulares e posições de deslocamento particulares dos padrões de repetição **1232** classificados como causadores dos erros específicos da sequência, como também causando os erros específicos da sequência.

[00200] As **Figuras 7A**, **7B**e **7C** mostram um limiar de exemplo **702** (por exemplo, 0,6) que é aplicado às distribuições de saída **1212** das pontuações de classificação **334** para identificar o subconjunto das pontuações de classificação que estão acima do limiar **702**. Tais pontuações de classificação são indicativas das classificações de falsas variantes e os padrões de repetição associados a elas são classificados como causadores dos erros específicos da sequência.

Implementações Específicas

[00201] A tecnologia divulgada refere-se à identificação de padrões de repetição que causam erros específicos de sequência.

[00202] A tecnologia divulgada pode ser praticada como um

sistema, método, dispositivo, produto, mídia legível por computador ou artigo de fabricação. Um ou mais recursos de uma implementação podem ser combinados com a implementação base. Implementações que não são mutuamente exclusivas são ensinadas a serem combináveis. Um ou mais recursos de uma implementação podem ser combinados com outras implementações. Esta divulgação lembra periodicamente o usuário dessas opções. A omissão de algumas implementações de recitações que repetem essas opções não deve ser tomada como limitativa das combinações ensinadas nas seções anteriores - essas recitações são incorporadas neste documento adiante por referência em cada uma das implementações a seguir.

[00203] Uma primeira implementação de sistema da tecnologia divulgada inclui um ou mais processadores acoplados à memória. A memória é carregada com instruções do computador para identificar padrões de repetição que causam erros específicos de sequência. O sistema inclui um subsistema de preparação de entrada em execução em vários processadores operando em paralelo e acoplados a memória. O subsistema de preparação de entrada sobrepõe padrões de repetição em teste em sequências nucleotídicas para produzir amostras sobrepostas. Cada uma das amostras sobrepostas possui um nucleotídeo variante na posição alvo flanqueado por pelo menos 20 nucleotídeos em cada lado. Os padrões de repetição são homopolímeros de uma base única (A, C, G ou T) com pelo menos 6 fatores de repetição que especificam várias repetições da base única nos padrões de repetição. O sistema inclui um subsistema de simulação alimenta cada combinação dos padrões de repetição sobrepostos em pelo menos 100 sequências nucleotídicas em pelo menos 100 amostras sobrepostas a um filtro de variante para análise. O sistema inclui um subsistema de filtro de variante que traduz a análise pelo filtro de variante em pontuações de classificação para a probabilidade de que o nucleotídeo variante em cada uma das amostras sobrepostas seja uma variante

verdadeira ou uma variante falsa. Finalmente, o sistema inclui um subsistema de análise que causa a exibição das pontuações de classificação como uma distribuição para cada um dos fatores de repetição para apoiar a avaliação da causa de erro específica de sequência pelos padrões de repetição.

[00204] Esta implementação do sistema e outros sistemas divulgados incluem opcionalmente um ou mais dos seguintes recursos.

[00205] O sistema também pode incluir recursos descritos em conexão com os métodos divulgados. A interesse de concisão, as combinações alternativas de recursos do sistema não são enumeradas individualmente. Os recursos aplicáveis aos sistemas, métodos e artigos de fabricação não são repetidos para cada conjunto de classes estatutárias de recursos básicos. O leitor entenderá como os recursos identificados nesta seção podem ser facilmente combinados com os recursos básicos de outras classes estatutárias.

[00206] Em uma implementação, os padrões de repetição estão à direita de um nucleotídeo central nas amostras sobrepostas e não se sobrepõem ao nucleotídeo central. Em outra implementação, os padrões de repetição estão à esquerda de um nucleotídeo central nas amostras sobrepostas e não se sobrepõem ao nucleotídeo central. Em outra implementação, os padrões de repetição incluem um nucleotídeo central nas amostras sobrepostas.

[00207] Os fatores de repetição são números inteiros no intervalo de 5 a um quarto de uma contagem de nucleotídeos nas amostras sobrepostas. O sistema é configurado ainda para aplicar a padrões repetidos que são os homopolímeros da base única para cada uma das quatro bases (A, C, G e T).

[00208] O subsistema de preparação de entrada é ainda configurado para produzir os padrões de repetição e as amostras sobrepostas para os homopolímeros para cada uma das quatro bases e o subsistema de análise é ainda configurado para causar exibição da

distribuição da pontuação de classificação para cada um dos homopolímeros em justaposição.

[00209] Os padrões de repetição estão à direita de um nucleotídeo central nas amostras sobrepostas e a justaposição se aplica aos homopolímeros sobrepostos diretamente ao nucleotídeo central. Os padrões de repetição estão à direita de um nucleotídeo central nas amostras sobrepostas e a justaposição se aplica aos homopolímeros sobrepostos à esquerda do nucleotídeo central. As sequências nucleotídicas nas quais os padrões de repetição são sobrepostos são geradas aleatoriamente. As sequências nucleotídicas nas quais os padrões de repetição são sobrepostos são selecionadas aleatoriamente a partir de sequências nucleotídicas de DNA que ocorrem naturalmente. O subsistema de análise é ainda configurado para causar a exibição da distribuição da pontuação de classificação para cada um dos fatores de repetição usando diagramas de caixa e bigode.

[00210] O filtro de variantes é treinado em pelo menos 500000 exemplos de treinamento de variantes verdadeiras e pelo menos 50000 exemplos de treinamento de variantes falsas. Cada exemplo de treinamento é uma sequência nucleotídica com um nucleotídeo variante na posição alvo flanqueado por pelo menos 20 nucleotídeos de cada lado. O filtro de variante é uma rede neural convolucional (CNN) com duas camadas convolucionais e uma camada totalmente conectada.

[00211] Outras implementações podem incluir um meio de armazenamento legível por computador não transitório armazenando instruções executáveis por um processador para realizar funções do sistema descrito acima. Ainda outra implementação pode incluir um método que realiza as funções do sistema descrito acima.

[00212] Uma primeira implementação de método implementado por computador da tecnologia divulgada inclui a identificação de padrões de repetição que causam erros específicos de sequência. O método

implementado por computador inclui a preparação de entrada sobrepondo padrões de repetição em teste em sequências nucleotídicas para produzir amostras sobrepostas. Cada uma das amostras sobrepostas possui um nucleotídeo variante na posição alvo flanqueado por pelo menos 20 nucleotídeos em cada lado. Os padrões de repetição são homopolímeros de uma base única (A, C, G ou T) com pelo menos 6 fatores de repetição que especificam várias repetições da base única nos padrões de repetição. O método implementado por computador inclui alimentar cada combinação dos padrões de repetição sobrepostos em pelo menos 100 sequências nucleotídicas em pelo menos 100 amostras sobrepostas a um filtro de variante para análise. O método implementado por computador inclui a tradução de análise pelo filtro de variante em pontuações de classificação para a probabilidade de que o nucleotídeo variante em cada uma das amostras sobrepostas seja uma variante verdadeira ou uma variante falsa em uma saída. Finalmente, o método implementado por computador inclui causar exibição das pontuações de classificação como uma distribuição para cada um dos fatores de repetição para apoiar a avaliação da causalidade de erro específica de sequência pelos padrões de repetição.

[00213] Cada um dos recursos discutidos nesta seção de implementação específica para a primeira implementação do sistema se aplica igualmente a esta implementação do método implementado por computador. Conforme indicado acima, todos os recursos do sistema não são repetidos neste documento e devem ser considerados repetidos por referência.

[00214] Uma implementação de mídia legível por computador (CRM) inclui um meio de armazenamento legível por computador não transitório que armazena instruções executáveis por um processador para executar um método implementado por computador, conforme descrito acima. Outra implementação CRM pode incluir um sistema incluindo memória e um ou mais processadores operáveis para executar instruções,

armazenadas na memória, para executar o método implementado por computador como descrito acima.

[00215] Cada um dos recursos discutidos nesta seção de implementação específica para a implementação do sistema se aplica igualmente a esta implementação CRM. Conforme indicado acima, todos os recursos do sistema não são repetidos neste documento e devem ser considerados repetidos por referência.

[00216] Uma segunda implementação de sistema da tecnologia divulgada inclui um ou mais processadores acoplados a memória. A memória é carregada com instruções do computador para identificar padrões de repetição que causam erros específicos de sequência. O sistema inclui um subsistema de preparação de entrada que sobrepõe os padrões de repetição em teste em deslocamentos variados nas sequências de nucleotídeos para produzir amostras sobrepostas. Cada uma das amostras sobrepostas possui um nucleotídeo variante na posição alvo flanqueado por pelo menos 20 nucleotídeos em cada lado. Os padrões de repetição são homopolímeros de uma base única (A, C, G ou T) com pelo menos 6 fatores de repetição que especificam várias repetições da base única nos padrões de repetição. Os deslocamentos variados variam uma posição na qual os padrões de repetição são sobrepostos nas sequências nucleotídicas. Os deslocamentos variáveis são mensuráveis como um deslocamento entre uma posição de origem dos padrões de repetição e uma posição de origem das sequências nucleotídicas. Em uma implementação, pelo menos dez deslocamentos são usados para produzir as amostras sobrepostas.

[00217] O sistema compreende ainda um subsistema de simulação que alimenta cada combinação dos padrões de repetição sobrepostos em pelo menos 100 sequências nucleotídicas em pelo menos 100 amostras sobrepostas a um filtro de variante para análise. O sistema inclui um subsistema de filtro de variante que traduz a análise pelo filtro de variante em pontuações de classificação para a probabilidade de que o nucleotídeo

variante em cada uma das amostras sobrepostas seja uma variante verdadeira ou uma variante falsa. Finalmente, o sistema inclui um subsistema de análise que causa a exibição das pontuações de classificação como uma distribuição para cada um dos fatores de repetição para apoiar a avaliação da causalidade de erro específica de sequência pela presença dos padrões de repetição nos deslocamentos variáveis.

[00218] Outras implementações podem incluir um meio de armazenamento legível por computador não transitório armazenando instruções executáveis por um processador para realizar funções do sistema descrito acima. Ainda outra implementação pode incluir um método que realiza as funções do sistema descrito acima.

[00219] Uma segunda implementação de método implementado por computador da tecnologia divulgada inclui a identificação de padrões de repetição que causam erros específicos de sequência. O método inclui a sobreposição de padrões de repetição em teste em diferentes deslocamentos nas sequências nucleotídicas para produzir amostras sobrepostas. Cada uma das amostras sobrepostas possui um nucleotídeo variante na posição alvo flanqueado por pelo menos 20 nucleotídeos em cada lado. Os padrões de repetição são homopolímeros de uma base única (A, C, G ou T) com pelo menos 6 fatores de repetição que especificam várias repetições da base única nos padrões de repetição. Os deslocamentos variados variam uma posição na qual os padrões de repetição são sobrepostos nas sequências nucleotídicas. Os deslocamento é mensurável como um deslocamento entre uma posição de origem dos padrões de repetição e uma posição de origem das sequências nucleotídicas. Em uma implementação, pelo menos dez deslocamentos são usados para produzir as amostras sobrepostas.

[00220] O método implementado por computador inclui alimentar cada combinação dos padrões de repetição sobrepostos em pelo menos 100 sequências nucleotídicas em pelo menos 100 amostras sobrepostas a um

filtro de variante para análise. Isto é seguido pela tradução de análise pelo filtro de variante em pontuações de classificação para a probabilidade de que o nucleotídeo variante em cada uma das amostras sobrepostas seja uma variante verdadeira ou uma variante falsa. Finalmente, o método implementado por computador causa a exibição das pontuações de classificação como uma distribuição para cada um dos fatores de repetição para apoiar a avaliação da causalidade de erro específica de sequência pela presença dos padrões de repetição nos deslocamentos variantes.

[00221] Uma implementação de mídia legível por computador (CRM) inclui um meio de armazenamento legível por computador não transitório que armazena instruções executáveis por um processador para executar um método implementado por computador, conforme descrito acima. Outra implementação CRM pode incluir um sistema incluindo memória e um ou mais processadores operáveis para executar instruções, armazenadas na memória, para executar o método implementado por computador como descrito acima.

[00222] Uma terceira implementação de sistema da tecnologia divulgada inclui um ou mais processadores acoplados a memória. A memória é carregada com instruções do computador para identificar padrões de repetição que causam erros específicos de sequência. O sistema inclui um subsistema de preparação de entrada, em execução em vários processadores operando em paralelo e acoplados a memória, que sobrepõe padrões de repetição em teste em sequências nucleotídicas para produzir amostras sobrepostas. Cada uma das amostras sobrepostas possui um nucleotídeo variante na posição alvo flanqueado por pelo menos 20 nucleotídeos em cada lado. Os padrões de repetição são copolímeros de pelo menos duas bases dentre quatro bases (A, C, G e T) com pelo menos 6 fatores de repetição que especificam um número de repetições das pelo menos duas bases nos padrões de repetição. O sistema inclui um subsistema de simulação, em execução nos inúmeros processadores

operando em paralelo e acoplados à memória, que alimenta cada combinação dos padrões de repetição sobrepostos em pelo menos 100 sequências de nucleotídeos em pelo menos 100 amostras sobrepostas a um filtro de variante para análise. O sistema inclui um subsistema de filtros de variante, em execução nos vários processadores que operam em paralelo e acoplados à memória. O subsistema de filtro de variante traduz a análise pelo filtro de variante em pontuações de classificação para a probabilidade de que o nucleotídeo variante em cada uma das amostras sobrepostas seja uma variante verdadeira ou uma variante falsa. Finalmente, o sistema inclui um subsistema de análise, em execução nos inúmeros processadores operando em paralelo e acoplados à memória, que causa a exibição das pontuações de classificação como uma distribuição para cada um dos fatores de repetição para apoiar a avaliação da causalidade de erro específica de sequência pelos padrões de repetição.

[00223] Esta implementação do sistema e outros sistemas divulgados incluem opcionalmente um ou mais dos seguintes recursos. O sistema também pode incluir recursos descritos em conexão com os métodos divulgados. A interesse de concisão, as combinações alternativas de recursos do sistema não são enumeradas individualmente. Os recursos aplicáveis aos sistemas, métodos e artigos de fabricação não são repetidos para cada conjunto de classes estatutárias de recursos básicos. O leitor entenderá como os recursos identificados nesta seção podem ser facilmente combinados com os recursos básicos de outras classes estatutárias.

[00224] Os padrões de repetição são uma enumeração combinatória de copadrões de fatores de repetição variados e comprimentos de padrão variados.

[00225] Outras implementações podem incluir um meio de armazenamento legível por computador não transitório armazenando instruções executáveis por um processador para realizar funções do sistema descrito acima. Ainda outra implementação pode incluir um método que

realiza as funções do sistema descrito acima.

[00226] Uma terceira implementação de método implementado por computador da tecnologia divulgada inclui a identificação de padrões de repetição que causam erros específicos de sequência. O método inclui a sobreposição de padrões de repetição em teste em sequências nucleotídicas para produzir amostras sobrepostas. Cada uma das amostras sobrepostas possui um nucleotídeo variante na posição alvo flanqueado por pelo menos 20 nucleotídeos em cada lado. Os padrões de repetição são copolímeros de pelo menos duas bases dentre quatro bases (A, C, G e T) com pelo menos 6 fatores de repetição que especificam um número de repetições das pelo menos duas bases nos padrões de repetição. O método inclui alimentar cada combinação dos padrões de repetição sobrepostos em pelo menos 100 sequências nucleotídicas em pelo menos 100 amostras sobrepostas a um filtro de variante para análise. O método inclui a tradução de análise pelo filtro de variante em pontuações de classificação para a probabilidade de que o nucleotídeo variante em cada uma das amostras sobrepostas seja uma variante verdadeira ou uma variante falsa. Finalmente, o método inclui causar exibição das pontuações de classificação como uma distribuição para cada um dos fatores de repetição para apoiar a avaliação da causalidade de erro específica de sequência pelo padrão de repetição.

[00227] Cada um dos recursos discutidos nesta seção de implementação específica para a terceira implementação do sistema se aplica igualmente a esta implementação do método implementado por computador. Conforme indicado acima, todos os recursos do sistema não são repetidos neste documento e devem ser considerados repetidos por referência.

[00228] Uma implementação de mídia legível por computador (CRM) inclui um meio de armazenamento legível por computador não transitório que armazena instruções executáveis por um processador para executar um método implementado por computador, conforme descrito

acima. Outra implementação CRM pode incluir um sistema incluindo memória e um ou mais processadores operáveis para executar instruções, armazenadas na memória, para executar o método implementado por computador como descrito acima.

[00229] Cada um dos recursos discutidos nesta seção de implementação específica para a terceira implementação do sistema se aplica igualmente a esta implementação CRM. Conforme indicado acima, todos os recursos do sistema não são repetidos neste documento e devem ser considerados repetidos por referência.

[00230] Uma quarta implementação de sistema da tecnologia divulgada inclui um ou mais processadores acoplados a memória. A memória é carregada com instruções do computador para identificar padrões de repetição que causam erros específicos de sequência. O sistema inclui um subsistema de preparação de entrada, em execução em vários processadores operando em paralelo e acoplados a memória, que sobrepõe padrões de repetição em teste em deslocamentos variados em sequências nucleotídicas para produzir amostras sobrepostas. Cada uma das amostras sobrepostas possui um nucleotídeo variante na posição alvo flanqueado por pelo menos 20 nucleotídeos em cada lado. Os padrões de repetição são copolímeros de pelo menos duas bases dentre quatro bases (A, C, G e T) com pelo menos 6 fatores de repetição que especificam um número de repetições das pelo menos duas bases nos padrões de repetição. Os deslocamentos variados variam uma posição na qual os padrões de repetição são sobrepostos nas sequências nucleotídicas. Os deslocamentos variáveis são mensuráveis como um deslocamento entre uma posição de origem dos padrões de repetição e uma posição de origem das sequências nucleotídicas. Em uma implementação, pelo menos dez deslocamentos são usados para produzir as amostras sobrepostas.

[00231] O sistema inclui um subsistema de simulação, em execução nos inúmeros processadores que operam em paralelo e acoplados

à memória, que alimenta cada combinação dos padrões de repetição. Os padrões de repetição são sobrepostos em pelo menos 100 sequências nucleotídicas em pelo menos 100 amostras sobrepostas a um filtro de variante para análise. O sistema também inclui um subsistema de filtro de variante, em execução nos inúmeros processadores operando em paralelo e acoplados à memória, que traduz a análise pelo filtro de variante em pontuações de classificação para a probabilidade de que o nucleotídeo variante em cada uma das amostras sobrepostas seja uma variante verdadeira ou uma variante falsa. Por fim, o sistema inclui um subsistema de análise em execução nos inúmeros processadores operando em paralelo e acoplados à memória. O subsistema de análise causa a exibição das pontuações de classificação como uma distribuição para cada um dos fatores de repetição para apoiar a avaliação da causalidade de erro específica de sequência pela presença dos padrões de repetição nos deslocamentos variáveis.

[00232] Outras implementações podem incluir um meio de armazenamento legível por computador não transitório armazenando instruções executáveis por um processador para realizar funções do sistema descrito acima. Ainda outra implementação pode incluir um método que realiza as funções do sistema descrito acima.

[00233] Uma quarta implementação de método implementado por computador da tecnologia divulgada inclui a identificação de padrões de repetição que causam erros específicos de sequência. O método implementado por computador inclui a sobreposição de padrões de repetição em teste em sequências nucleotídicas para produzir amostras sobrepostas. Cada uma das amostras sobrepostas possui um nucleotídeo variante na posição alvo flanqueado por pelo menos 20 nucleotídeos em cada lado. Os padrões de repetição são copolímeros de pelo menos duas bases dentre quatro bases (A, C, G e T) com pelo menos 6 fatores de repetição. Os fatores de repetição especificam um número de repetições de pelo menos duas

bases nos padrões de repetição. Os deslocamentos variados variam uma posição na qual os padrões de repetição são sobrepostos nas sequências nucleotídicas. Os fatores de repetição são mensuráveis como um deslocamento entre uma posição de origem dos padrões de repetição e uma posição de origem das sequências nucleotídicas. Em uma implementação, pelo menos dez deslocamentos são usados para produzir as amostras sobrepostas. O método implementado por computador inclui alimentar cada combinação dos padrões de repetição sobrepostos em pelo menos 100 sequências nucleotídicas em pelo menos 100 amostras sobrepostas a um filtro de variante para análise. O método implementado por computador inclui ainda a tradução de análise pelo filtro de variante em pontuações de classificação para a probabilidade de que o nucleotídeo variante em cada uma das amostras sobrepostas seja uma variante verdadeira ou uma variante falsa em uma saída. Finalmente, o método implementado por computador inclui causar a exibição das pontuações de classificação como uma distribuição para cada um dos fatores de repetição para apoiar a avaliação da causalidade de erro específica de sequência pela presença dos padrões de repetição nos deslocamentos variantes.

[00234] Uma implementação de mídia legível por computador (CRM) inclui um meio de armazenamento legível por computador não transitório que armazena instruções executáveis por um processador para executar um método implementado por computador, conforme descrito acima. Outra implementação CRM pode incluir um sistema incluindo memória e um ou mais processadores operáveis para executar instruções, armazenadas na memória, para executar o método implementado por computador como descrito acima.

[00235] Uma quinta implementação de sistema da tecnologia divulgada inclui um ou mais processadores acoplados a memória. A memória é carregada com instruções do computador para identificar padrões de repetição que causam erros específicos de sequência. O sistema inclui um

subsistema de preparação de entrada em execução em vários processadores operando em paralelo e acoplados a memória. O subsistema de preparação de entrada seleciona sequências nucleotídicas de amostra de sequências nucleotídicas de DNA natural. Cada uma das sequências nucleotídicas da amostra possui um ou mais padrões de repetição de copolímeros que ocorrem naturalmente e um nucleotídeo variante em uma posição alvo flanqueado por pelo menos 20 nucleotídeos de cada lado. O sistema inclui um subsistema de simulação em execução nos vários processadores que operam em paralelo e acoplados à memória. O subsistema de simulação alimenta cada uma das sequências nucleotídicas da amostra a um filtro de variante para análise.

[00236] O sistema inclui um subsistema de filtro de variante em execução nos vários processadores que operam em paralelo e acoplados à memória. O subsistema de filtro de variante traduz a análise pelo filtro de variante em pontuações de classificação para a probabilidade de que o nucleotídeo variante em cada uma das sequências nucleotídicas da amostra seja uma variante verdadeira ou uma variante falsa e disponibiliza ativações de parâmetros do filtro de variante que respondem à análise. Por fim, o sistema inclui um subsistema de análise em execução nos inúmeros processadores operando em paralelo e acoplados à memória. O subsistema de análise analisa as ativações dos parâmetros do filtro de variante e causa a exibição de uma representação de padrões de repetição de copolímeros que ocorrem naturalmente em cada uma das sequências nucleotídicas da amostra que contribuem para uma classificação de variantes falsas.

[00237] Outras implementações podem incluir um meio de armazenamento legível por computador não transitório armazenando instruções executáveis por um processador para realizar funções do sistema descrito acima. Ainda outra implementação pode incluir um método que realiza as funções do sistema descrito acima.

[00238] Uma quinta implementação de método implementado por

computador da tecnologia divulgada inclui a identificação de padrões de repetição que causam erros específicos de sequência. O método implementado por computador inclui a seleção de sequências nucleotídicas da amostra a partir de sequências nucleotídicas de DNA natural. Cada uma das sequências nucleotídicas da amostra possui um ou mais padrões de repetição de copolímeros que ocorrem naturalmente e um nucleotídeo variante em uma posição alvo flanqueado por pelo menos 20 nucleotídeos de cada lado. O método implementado por computador inclui alimentar cada uma das sequências nucleotídicas da amostra com um filtro de variante para análise. O método inclui a tradução de análise pelo filtro de variante em pontuações de classificação para a probabilidade de que o nucleotídeo variante em cada uma das sequências nucleotídicas da amostra seja uma variante verdadeira ou uma variante falsa. O método implementado por computador disponibiliza ativações de parâmetros do filtro de variante que respondem à análise. Finalmente, o método implementado por computador inclui a análise das ativações dos parâmetros do filtro de variante e a exibição de uma representação de padrões de repetição de copolímeros que ocorrem naturalmente em cada uma das sequências nucleotídicas de amostra que contribuem para uma classificação de variantes falsas.

[00239] Uma implementação de mídia legível por computador (CRM) inclui um meio de armazenamento legível por computador não transitório que armazena instruções executáveis por um processador para executar um método implementado por computador, conforme descrito acima. Outra implementação CRM pode incluir um sistema incluindo memória e um ou mais processadores operáveis para executar instruções, armazenadas na memória, para executar o método implementado por computador como descrito acima.

[00240] A tecnologia divulgada apresenta um sistema para identificar padrões de repetição que causam erros específicos de sequência.

[00241] O sistema compreende um subsistema de preparação de

entrada que é executado em vários processadores operando em paralelo e acoplados a memória. O subsistema de preparação de entrada sobrepõe padrões de repetição em teste em sequências nucleotídicas para produzir amostras sobrepostas. Cada uma das amostras sobrepostas possui um nucleotídeo variante na posição alvo flanqueado por pelo menos 20 nucleotídeos em cada lado. Os padrões de repetição incluem pelo menos uma base dentre quatro bases (A, C, G e T) com pelo menos 6 fatores de repetição.

[00242] O sistema compreende um subsistema de simulação que é executado nos vários processadores que operam em paralelo e acoplados à memória. O subsistema de simulação alimenta cada combinação dos padrões de repetição sobrepostos em pelo menos 100 sequências nucleotídicas em pelo menos 100 amostras sobrepostas a um filtro de variante para análise.

[00243] O sistema compreende um subsistema de filtro de variante que é executado nos vários processadores que operam em paralelo e acoplados à memória. O subsistema de filtro de variante traduz a análise pelo filtro de variante em pontuações de classificação para a probabilidade de que o nucleotídeo variante em cada uma das amostras sobrepostas seja uma variante verdadeira ou uma variante falsa.

[00244] O sistema compreende um subsistema de análise que é executado nos vários processadores que operam em paralelo e acoplados à memória. O subsistema de análise causa a exibição das pontuações de classificação como uma distribuição para cada um dos fatores de repetição para apoiar a avaliação da causalidade de erro específica de sequência pelos padrões de repetição.

[00245] Cada um dos recursos discutidos nesta seção de implementação específica para a primeira implementação do sistema se aplica igualmente a essa implementação do sistema. Conforme indicado acima, todos os recursos do sistema não são repetidos neste documento e

devem ser considerados repetidos por referência.

[00246] Em uma implementação, os padrões de repetição são homopolímeros de uma base única (A, C, G ou T) com os pelo menos 6 fatores de repetição que especificam várias repetições da base única nos padrões de repetição.

[00247] Em outra implementação, os padrões de repetição são copolímeros de pelo menos duas bases dentre quatro bases (A, C, G e T) com os pelo menos 6 fatores de repetição que especificam um número de repetições das pelo menos duas bases nos padrões de repetição.

[00248] Em algumas implementações, o subsistema de preparação de entrada é ainda configurado para sobrepor os padrões de repetição em teste em deslocamentos variados nas sequências nucleotídicas para produzir as amostras sobrepostas. Os deslocamentos variados variam uma posição na qual os padrões de repetição são sobrepostos nas sequências de nucleotídeos, mensuráveis como um deslocamento entre uma posição de origem dos padrões de repetição e uma posição de origem das sequências nucleotídicas, e pelo menos dez deslocamentos são usados para produzir as amostras sobrepostas. Em algumas implementações, o subsistema de análise é ainda configurado para causar a exibição das pontuações de classificação como uma distribuição para cada um dos fatores de repetição para apoiar a avaliação da causalidade de erro específica de sequência pela presença dos padrões de repetição nos deslocamentos variáveis.

[00249] Em uma implementação, os padrões de repetição estão à direita de um nucleotídeo central nas amostras sobrepostas e não se sobrepõem ao nucleotídeo central. Em outra implementação, os padrões de repetição estão à esquerda de um nucleotídeo central nas amostras sobrepostas e não se sobrepõem ao nucleotídeo central. Em outra implementação, os padrões de repetição incluem um nucleotídeo central nas amostras sobrepostas.

[00250] Os fatores de repetição são números inteiros no intervalo

de 5 a um quarto de uma contagem de nucleotídeos nas amostras sobrepostas. O sistema é configurado ainda para aplicar a padrões repetidos que são os homopolímeros da base única para cada uma das quatro bases (A, C, G e T).

[00251] O subsistema de preparação de entrada é ainda configurado para produzir os padrões de repetição e as amostras sobrepostas para os homopolímeros para cada uma das quatro bases e o subsistema de análise é ainda configurado para causar exibição da distribuição da pontuação de classificação para cada um dos homopolímeros em justaposição.

[00252] Os padrões de repetição estão à direita de um nucleotídeo central nas amostras sobrepostas e a justaposição se aplica aos homopolímeros sobrepostos diretamente ao nucleotídeo central. Os padrões de repetição estão à direita de um nucleotídeo central nas amostras sobrepostas e a justaposição se aplica aos homopolímeros sobrepostos à esquerda do nucleotídeo central. As sequências nucleotídicas nas quais os padrões de repetição são sobrepostos são geradas aleatoriamente. As sequências nucleotídicas nas quais os padrões de repetição são sobrepostos são selecionadas aleatoriamente a partir de sequências nucleotídicas de DNA que ocorrem naturalmente. O subsistema de análise é ainda configurado para causar a exibição da distribuição da pontuação de classificação para cada um dos fatores de repetição usando diagramas de caixa e bigode.

[00253] O filtro de variantes é treinado em pelo menos 500000 exemplos de treinamento de variantes verdadeiras e pelo menos 50000 exemplos de treinamento de variantes falsas. Cada exemplo de treinamento é uma sequência nucleotídica com um nucleotídeo variante na posição alvo flanqueada por pelo menos 20 nucleotídeos de cada lado. O filtro de variante é uma rede neural convolucional (CNN) com duas camadas convolucionais e uma camada totalmente conectada.

[00254] A tecnologia divulgada apresenta um método implementado por computador para identificar padrões de repetição que causam erros específicos de sequência.

[00255] O método implementado por computador inclui a sobreposição de padrões de repetição em teste em sequências nucleotídicas para produzir amostras sobrepostas.

[00256] O método implementado por computador inclui alimentar cada combinação dos padrões de repetição sobrepostos em pelo menos 100 sequências nucleotídicas em pelo menos 100 amostras sobrepostas a um filtro de variante para análise.

[00257] O método implementado por computador inclui ainda a tradução de análise pelo filtro de variante em pontuações de classificação para a probabilidade de que o nucleotídeo variante em cada uma das amostras sobrepostas seja uma variante verdadeira ou uma variante falsa em uma saída.

[00258] O método implementado por computador inclui causar exibição das pontuações de classificação como uma distribuição para cada um dos fatores de repetição para apoiar a avaliação da causalidade de erro específica de sequência pelos padrões de repetição.

[00259] Cada um dos recursos discutidos nesta seção de implementação específica para a primeira implementação do sistema se aplica igualmente a esta implementação do método implementado por computador. Conforme indicado acima, todos os recursos do sistema não são repetidos neste documento e devem ser considerados repetidos por referência.

[00260] A tecnologia divulgada apresenta outro sistema para identificar padrões de repetição que causam erros específicos de sequência em dados de sequenciamento de nucleotídeo. O sistema compreende um ou mais processadores e um ou mais dispositivos de armazenamento que armazenam instruções que, quando executadas em um ou mais

processadores, fazem com que um ou mais processadores implementem um subsistema de preparação de entrada, um subsistema de filtro variante e um subsistema de saída de padrão de repetição.

[00261] O subsistema de preparação de entrada é configurado par sobrepor padrões de repetição em teste em sequências nucleotídicas para produzir amostras sobrepostas. Cada uma das amostras sobrepostas possui um nucleotídeo variante e os padrões de repetição incluem pelo menos uma base dentre quatro bases (A, C, G e T).

[00262] O subsistema de filtro de variante é configurado para processar cada combinação dos padrões de repetição sobrepostos nas sequências de nucleotídeos nas amostras sobrepostas para gerar pontuações de classificação para a probabilidade de que o nucleotídeo variante em cada uma das amostras sobrepostas seja uma variante verdadeira ou uma variante falsa.

[00263] O subsistema de saída do padrão de repetição está configurado para gerar determinados padrões dos padrões de repetição que causam erros específicos de sequência nos dados de sequenciamento de nucleotídeo com base nas pontuações de classificação.

[00264] Cada um dos recursos discutidos nesta seção de implementação específica para a primeira implementação do sistema se aplica igualmente a essa implementação do sistema. Conforme indicado acima, todos os recursos do sistema não são repetidos neste documento e devem ser considerados repetidos por referência.

[00265] O sistema é ainda configurado para compreender um subsistema de análise que está configurado para causar a exibição das pontuações de classificação como uma distribuição para cada um dos fatores de repetição para apoiar a avaliação da causalidade de erro específica de sequência pelos padrões de repetição.

[00266] Uma implementação de mídia legível por computador (CRM) inclui um meio de armazenamento legível por computador não

transitório que armazena instruções executáveis por um processador para executar um método implementado por computador, conforme descrito acima. Outra implementação CRM pode incluir um sistema incluindo memória e um ou mais processadores operáveis para executar instruções, armazenadas na memória, para executar o método implementado por computador como descrito acima.

[00267] A tecnologia divulgada apresenta outro sistema para identificar padrões de repetição que causam erros específicos de sequência em dados de sequenciamento de nucleotídeo. O sistema compreende um ou mais processadores e um ou mais dispositivos de armazenamento que armazenam instruções que, quando executadas em um ou mais processadores, fazem com que um ou mais processadores implementem um subsistema de preparação de entrada, um subsistema de filtro variante e um subsistema de saída de padrão de repetição.

[00268] O subsistema de preparação de entrada é configurado par sobrepor padrões de repetição em teste em sequências nucleotídicas para produzir amostras sobrepostas. Cada uma das amostras sobrepostas possui um nucleotídeo variante e os padrões de repetição incluem pelo menos uma base dentre quatro bases (A, C, G e T).

[00269] O subsistema de filtro de variante é configurado para processar cada combinação dos padrões de repetição sobrepostos nas sequências de nucleotídeos nas amostras sobrepostas para gerar pontuações de classificação para a probabilidade de que o nucleotídeo variante em cada uma das amostras sobrepostas seja uma variante verdadeira ou uma variante falsa.

[00270] O subsistema de saída do padrão de repetição está configurado para gerar determinados padrões dos padrões de repetição que causam erros específicos de sequência nos dados de sequenciamento de nucleotídeo com base nas pontuações de classificação.

[00271] Cada um dos recursos discutidos nesta seção de

implementação específica para a primeira implementação do sistema se aplica igualmente a essa implementação do sistema. Conforme indicado acima, todos os recursos do sistema não são repetidos neste documento e devem ser considerados repetidos por referência.

[00272] O sistema é ainda configurado para compreender um subsistema de análise que está configurado para causar a exibição das pontuações de classificação como uma distribuição para cada um dos fatores de repetição para apoiar a avaliação da causalidade de erro específica de sequência pelos padrões de repetição.

[00273] A tecnologia divulgada apresenta um método implementado por computador para identificar padrões de repetição que causam erros específicos de sequência em dados de sequenciamento de nucleotídeo.

[00274] O método implementado por computador inclui a sobreposição de padrões de repetição em teste em sequências nucleotídicas para produzir amostras sobrepostas. Cada uma das amostras sobrepostas possui um nucleotídeo variante e os padrões de repetição incluem pelo menos uma base dentre quatro bases (A, C, G e T).

[00275] O método implementado por computador inclui o processamento de cada combinação dos padrões de repetição sobrepostos nas sequências nucleotídicas nas amostras sobrepostas por meio de um subsistema de filtro de variante para gerar pontuações de classificação para a probabilidade de que o nucleotídeo variante em cada uma das amostras sobrepostas seja uma variante verdadeira ou uma variante falsa.

[00276] O método implementado por computador inclui ainda a tradução de análise pelo filtro de variante em pontuações de classificação para a probabilidade de que o nucleotídeo variante em cada uma das amostras sobrepostas seja uma variante verdadeira ou uma variante falsa em uma saída.

[00277] O método implementado por computador inclui a saída de

determinados padrões de repetição que causam erros específicos de sequência nos dados de sequenciamento de nucleotídeos com base nas pontuações de classificação.

[00278] Cada um dos recursos discutidos nesta seção de implementação específica para a primeira implementação do sistema se aplica igualmente a esta implementação do método implementado por computador. Conforme indicado acima, todos os recursos do sistema não são repetidos neste documento e devem ser considerados repetidos por referência.

[00279] Uma implementação de mídia legível por computador (CRM) inclui um meio de armazenamento legível por computador não transitório que armazena instruções executáveis por um processador para executar um método implementado por computador, conforme descrito acima. Outra implementação CRM pode incluir um sistema incluindo memória e um ou mais processadores operáveis para executar instruções, armazenadas na memória, para executar o método implementado por computador como descrito acima.

[00280] A tecnologia divulgada apresenta outro sistema para identificar padrões de repetição que causam erros específicos de sequência em dados de sequenciamento de nucleotídeo. O sistema compreende um ou mais processadores e um ou mais dispositivos de armazenamento que armazenam instruções que, quando executadas em um ou mais processadores, fazem com que um ou mais processadores implementem um subsistema de preparação de entrada, um subsistema de filtro variante e um subsistema de saída de padrão de repetição.

[00281] O subsistema de preparação de entrada é configurado para selecionar sequências nucleotídicas da amostra a partir de sequências nucleotídicas de DNA natural. Cada uma das sequências nucleotídicas da amostra possui um ou mais padrões de repetição de copolímeros que ocorrem naturalmente e um nucleotídeo variante.

[00282] O subsistema de filtro de variante é configurado para processar cada uma das sequências nucleotídicas da amostra para gerar pontuações de classificação para a probabilidade de que o nucleotídeo variante em cada uma das sequências nucleotídicas da amostra seja uma variante verdadeira ou uma variante falsa.

[00283] O subsistema de saída de padrão de repetição é configurado para disponibilizar ativações de parâmetros do subsistema de filtro de variante que respondem à análise e produzem determinados padrões dos padrões de repetição que causam erros específicos de sequência nos dados de sequenciamento de nucleotídeos com base nas pontuações de classificação.

[00284] Cada um dos recursos discutidos nesta seção de implementação específica para a primeira implementação do sistema se aplica igualmente a essa implementação do sistema. Conforme indicado acima, todos os recursos do sistema não são repetidos neste documento e devem ser considerados repetidos por referência.

[00285] O sistema é ainda configurado para compreender um subsistema de análise que está configurado para analisar as ativações dos parâmetros do subsistema de filtro de variante e causar a exibição de uma representação de padrões de repetição de copolímeros que ocorrem naturalmente em cada uma das sequências de nucleotídeos da amostra que contribuem para uma falsa classificação de variantes.

[00286] A tecnologia divulgada apresenta um método implementado por computador para identificar padrões de repetição que causam erros específicos de sequência em dados de sequenciamento de nucleotídeo.

[00287] O método implementado por computador inclui a seleção de sequências nucleotídicas da amostra a partir de sequências nucleotídicas de DNA natural. Cada uma das sequências nucleotídicas da amostra possui um ou mais padrões de repetição de copolímeros que ocorrem naturalmente

e um nucleotídeo variante.

[00288] O método implementado por computador inclui o processamento de cada uma das sequências nucleotídicas da amostra através de um subsistema de filtro de variante para gerar pontuações de classificação para a probabilidade de que o nucleotídeo variante em cada uma das sequências nucleotídicas da amostra seja uma variante verdadeira ou uma variante falsa.

[00289] O método implementado por computador inclui disponibilizar ativações de parâmetros do subsistema de filtro de variante que respondem à análise.

[00290] O método implementado por computador inclui a saída de determinados padrões de repetição que causam erros específicos de sequência nos dados de sequenciamento de nucleotídeos com base nas pontuações de classificação.

[00291] Cada um dos recursos discutidos nesta seção de implementação específica para a primeira implementação do sistema se aplica igualmente a esta implementação do método implementado por computador. Conforme indicado acima, todos os recursos do sistema não são repetidos neste documento e devem ser considerados repetidos por referência.

[00292] Uma implementação de mídia legível por computador (CRM) inclui um meio de armazenamento legível por computador não transitório que armazena instruções executáveis por um processador para executar um método implementado por computador, conforme descrito acima. Outra implementação CRM pode incluir um sistema incluindo memória e um ou mais processadores operáveis para executar instruções, armazenadas na memória, para executar o método implementado por computador como descrito acima.

[00293] Quaisquer estruturas e códigos de dados descritos ou mencionados acima são armazenados de acordo com muitas

implementações em uma mídia de armazenamento legível por computador, que pode ser qualquer dispositivo ou mídia que pode armazenar código e/ou dados para uso por um sistema de computador. Isso inclui, entre outros, memória volátil, memória não volátil, circuitos integrados específicos para aplicativos (ASICs), arranjo de portas programáveis em campo (FPGAs), dispositivos de armazenamento magnético e óptico, como unidades de disco, fita magnética, CDs (discos compactos), DVDs (discos versáteis digitais ou discos de vídeo digital) ou outra mídia capaz de armazenar mídias legíveis por computador agora conhecidas ou desenvolvidas posteriormente.

[00294] A descrição anterior é apresentada para permitir a criação e o uso da tecnologia divulgada. Várias modificações às implementações divulgadas serão evidentes e os princípios gerais definidos neste documento podem ser aplicados a outras implementações e pedidos sem se afastar do espírito e âmbito da tecnologia divulgada. Assim, a tecnologia divulgada não se destina a ser limitada às implementações apresentadas, mas deve receber o escopo mais amplo consistente com os princípios e características divulgados neste documento. O escopo da tecnologia divulgada é definido pelas reivindicações anexas.

CLÁUSULAS

[00295] A divulgação também inclui as seguintes cláusulas:

[00296] Um sistema para identificar padrões de repetição que causam erros específicos de sequência nos dados de sequenciamento de nucleotídeo, compreendendo:

[00297] um ou mais processadores e um ou mais dispositivos de armazenamento que armazenam instruções que, quando executadas em um ou mais processadores, fazem com que os um ou mais processadores implementem:

[00298] um subsistema de preparação de entrada configurado para:

[00299] sobrepor computacionalmente padrões de repetição em teste em várias sequências de nucleotídeos e produzir amostras sobrepostas,

[00300] em que cada padrão de repetição representa uma composição nucleotídica específica que tem um comprimento específico e aparece em uma amostra sobreposta em uma posição de deslocamento específica,

[00301] em que cada amostra sobreposta tem uma posição de destino considerada um nucleotídeo variante, e

[00302] em que para cada combinação da composição nucleotídica específica, o comprimento específico e a posição de deslocamento específica, um conjunto de amostras sobrepostas é gerado computacionalmente;

[00303] um subsistema de filtro de variante pré-treinado configurado para:

[00304] processar as amostras sobrepostas por meio de uma rede neural convolucional e, com base na detecção de padrões de nucleotídeos nas amostras sobrepostas por filtros de convolução da rede neural convolucional, gerar pontuações de classificação para a probabilidade de que o nucleotídeo variante em cada uma das amostras sobrepostas seja uma variante verdadeira ou uma variante falsa;

[00305] um subsistema de saída de padrão de repetição configurado para:

[00306] gerar distribuições das pontuações de classificação que indicam suscetibilidade do subsistema de filtro de variante pré-treinado a classificações de variantes falsas resultantes da presença dos padrões de repetição; e

[00307] um subsistema de correlação de erro específico da sequência configurado para:

[00308] especificar, com base em um limite, um subconjunto das pontuações de classificação como indicativo das classificações de falsas variantes e

[00309] classificar os padrões de repetição associados ao subconjunto das pontuações de classificação que são indicativos das classificações de falsas variantes como causadores dos erros específicos de sequência.2. O sistema da cláusula 1, em que o subsistema de correlação de erro específico de sequência é ainda configurado para:

[00311] classificar comprimentos particulares e posições de deslocamento específicas dos padrões de repetição classificados como causando os erros específicos de sequência como também causando os erros específicos de sequência.

[00312] 3. O sistema de qualquer uma das cláusulas 1-2, em que o nucleotídeo variante está na posição alvo flanqueado por pelo menos 20 nucleotídeos em cada lado.

[00313] 4. O sistema de qualquer uma das cláusulas 1-3, em que o subsistema de filtro de variante pré-treinado é configurado para processar cada combinação dos padrões de repetição sobrepostos em pelo menos 100 sequências de nucleotídeos em pelo menos 100 amostras sobrepostas.

[00314] 5. O sistema de qualquer uma das cláusulas 1-5, em que os padrões de repetição incluem a pelo menos uma base dentre quatro bases (A, C, G e T) com pelo menos 6 fatores de repetição.

[00315] 6. O sistema da cláusula 5, em que os padrões de repetição são homopolímeros de uma única base (A, C, G ou T) com os pelo menos 6 fatores de repetição; e

[00316] em que os pelo menos 6 fatores de repetição especificam um número de repetições da base única nos padrões de repetição.

[00317] 7. O sistema de qualquer uma das cláusulas 1-6, em que os padrões de repetição são copolímeros de pelo menos duas bases

dentre quatro bases (A, C, G e T) com os pelo menos 6 fatores de repetição; e

[00318] em que os pelo menos 6 repetição especificam um número de repetições de pelo menos duas bases nos padrões de repetição.

[00319] 8. O sistema de qualquer uma das cláusulas 1-7, em que as posições de deslocamento variam em termos de uma posição na qual os padrões de repetição são sobrepostos nas sequências de nucleotídeos, mensuráveis como um deslocamento entre uma posição de origem dos padrões de repetição e uma posição de origem das sequências nucleotídicas e pelo menos dez deslocamentos são usados para produzir as amostras sobrepostas.

[00320] 9. O sistema de qualquer uma das cláusulas 1-8, em que os padrões de repetição estão à direita de um nucleotídeo central nas amostras sobrepostas e não se sobrepõem ao nucleotídeo central.

[00321] 10. O sistema de qualquer uma das cláusulas 1-9, em que os padrões de repetição estão à esquerda de um nucleotídeo central nas amostras sobrepostas e não se sobrepõem ao nucleotídeo central.

[00322] 11. O sistema de qualquer uma das cláusulas 1-10, em que os padrões de repetição incluem um nucleotídeo central nas amostras sobrepostas.

[00323] 12. O sistema de qualquer uma das cláusulas 1-11, em que os fatores de repetição são números inteiros na faixa de 5 a um quarto de uma contagem de nucleotídeos nas amostras sobrepostas.

[00324] 13. O sistema da cláusula 6, configurado ainda para aplicar a padrões repetidos que são os homopolímeros da base única para cada uma das quatro bases (A, C, G e T).

[00325] 14. O sistema da cláusula 13, em que o subsistema de preparação de entrada é ainda configurado para produzir os padrões de repetição e as amostras sobrepostas para os homopolímeros para cada uma das quatro bases.

[00326] 15. O sistema da cláusula 14, em que os padrões de repetição estão à direita de um nucleotídeo central nas amostras sobrepostas e a justaposição se aplica aos homopolímeros sobrepostos diretamente ao nucleotídeo central.

[00327] 16. O sistema da cláusula 14, em que padrões de repetição estão à direita de um nucleotídeo central nas amostras sobrepostas e a justaposição se aplica aos homopolímeros sobrepostos à esquerda do nucleotídeo central.

[00328] 17. O sistema de qualquer uma das cláusulas 1-16, em que as sequências de nucleotídeos nas quais os padrões de repetição são sobrepostos são geradas aleatoriamente.

[00329] 18. O sistema de qualquer uma das cláusulas 1-17, em que as sequências nucleotídicas nas quais os padrões de repetição são sobrepostos são selecionadas aleatoriamente a partir de sequências nucleotídicas de DNA que ocorrem naturalmente.

[00330] 19. O sistema de qualquer uma das cláusulas 1-18, em que um subsistema de análise está configurado para causar a exibição das distribuições das pontuações de classificação para cada um dos fatores de repetição.

[00331] 20. O sistema de qualquer uma das cláusulas 1-19, em que o subsistema de filtro de variantes pré-treinado é treinado em pelo menos 500000 exemplos de treinamento de variantes verdadeiras e pelo menos 50000 exemplos de treinamento de variantes falsas; e

[00332] em que cada exemplo de treinamento é uma sequência nucleotídica com um nucleotídeo variante na posição alvo flanqueado por pelo menos 20 nucleotídeos de cada lado.

[00333] 21. O sistema de qualquer uma das cláusulas 1-20, em que o subsistema de filtros variantes pré-treinado possui camadas convolucionais, uma camada totalmente conectada e uma camada de classificação.

[00334] 22. Um método implementado por computador para identificar padrões de repetição que causam erros específicos de sequência nos dados de sequenciamento de nucleotídeos, incluindo:

[00335] sobrepor computacionalmente padrões de repetição em teste em várias sequências de nucleotídeos e produzir amostras sobrepostas, em que cada padrão de repetição representa uma composição nucleotídica específica que possui um comprimento específico e aparece em uma amostra sobreposta em uma posição de deslocamento específica, em que cada amostra sobreposta tem uma posição de destino considerada como sendo um nucleotídeo variante e em que para cada combinação da composição nucleotídica específica, o comprimento específico e a posição de deslocamento específica, um conjunto de amostras sobrepostas é gerado computacionalmente;

[00336] processar as amostras sobrepostas por meio de uma rede neural convolucional e, com base na detecção de padrões de nucleotídeos nas amostras sobrepostas por filtros de convolução da rede neural convolucional, gerar pontuações de classificação para a probabilidade de que o nucleotídeo variante em cada uma das amostras sobrepostas seja uma variante verdadeira ou uma variante falsa;

[00337] gerar distribuições das pontuações de classificação que indicam suscetibilidade do subsistema de filtro de variante pré-treinado a classificações de variantes falsas resultantes da presença dos padrões de repetição; e

[00338] especificar, com base em um limite, um subconjunto das pontuações de classificação como indicativo das classificações de variantes falsas e classificar os padrões de repetição que estão associados ao subconjunto das pontuações de classificação que são indicativos das classificações de variantes falsas como causadoras dos erros específicos de sequência.

[00339] 23. O método implementado por computador da

cláusula 22, implementando cada uma das cláusulas que, em última análise, dependem da cláusula 1.

[00340] 24. Um meio de armazenamento legível por computador não transitório, impresso com instruções de programa de computador para identificar padrões de repetição que causam erros específicos de sequência nos dados de sequenciamento de nucleotídeos; as instruções, quando executadas em um processador, implementam um método implementado por computador, compreendendo:

[00341] sobrepor computacionalmente padrões de repetição em teste em várias sequências de nucleotídeos e produzir amostras sobrepostas, em que cada padrão de repetição representa uma composição nucleotídica específica que possui um comprimento específico e aparece em uma amostra sobreposta em uma posição de deslocamento específica, em que cada amostra sobreposta tem uma posição de destino considerada como sendo um nucleotídeo variante e em que para cada combinação da composição nucleotídica específica, o comprimento específico e a posição de deslocamento específica, um conjunto de amostras sobrepostas é gerado computacionalmente;

[00342] processar as amostras sobrepostas por meio de uma rede neural convolucional e, com base na detecção de padrões de nucleotídeos nas amostras sobrepostas por filtros de convolução da rede neural convolucional, gerar pontuações de classificação para a probabilidade de que o nucleotídeo variante em cada uma das amostras sobrepostas seja uma variante verdadeira ou uma variante falsa;

[00343] gerar distribuições das pontuações de classificação que indicam suscetibilidade do subsistema de filtro de variante pré-treinado a classificações de variantes falsas resultantes da presença dos padrões de repetição; e

[00344] especificar, com base em um limite, um subconjunto das pontuações de classificação como indicativo das classificações de variantes

falsas e classificar os padrões de repetição que estão associados ao subconjunto das pontuações de classificação que são indicativos das classificações de variantes falsas como causadoras dos erros específicos de sequência.

[00345] 25. O meio de armazenamento legível por computador não transitório da cláusula 24, implementando cada uma das cláusulas que, em última análise, dependem da cláusula 1.

[00346]

REIVINDICAÇÕES

1. Sistema para identificar padrões de repetição que causam erros específicos de sequência em dados de sequenciamento de nucleotídeos, **caracterizado** pelo fato de que compreende:

um ou mais processadores e um ou mais dispositivos de armazenamento que armazenam instruções que, quando executadas em um ou mais processadores, fazem com que um ou mais processadores implementem:

um subsistema de preparação de entrada configurado para:

sobrepor computacionalmente padrões de repetição em teste em inúmeras sequências de nucleotídeos e produzir amostras sobrepostas,

em que cada padrão de repetição representa uma composição de nucleotídeo particular que tem um comprimento particular e aparece em uma amostra sobreposta em uma posição de deslocamento particular,

em que cada amostra sobreposta tem uma posição de destino considerada um nucleotídeo variante, e

em que para cada combinação da composição de nucleotídeo particular, do comprimento particular e da posição de deslocamento particular, um conjunto das amostras sobrepostas é gerado computacionalmente;

um subsistema de filtro variante pré-treinado configurado para:

processar amostras sobrepostas por uma rede neural convolucional e, baseada na detecção de padrões de nucleotídeo nas amostras sobrepostas por filtros de convolução da rede neural convolucional, gerar pontuações de classificação para a probabilidade de que o nucleotídeo variante em cada uma das amostras sobrepostas seja uma variante verdadeira ou uma variante falsa;

um subsistema de saída de padrão de repetição configurado para:

distribuições de saída das pontuações de classificação geradas pelo subsistema de filtro de variante pré treinado para fatores de repetição para

respectivos padrões de repetição, em que as distribuições indicam impacto no subsistema de filtro de variante pré-treinado para classificações de variante falsa resultantes da presença de padrões de repetição; e

um subsistema de correlação de erro específico de sequência configurado para:

especificar, baseado em um limite, um subconjunto das pontuações de classificação nas distribuições como indicativo de classificações de variantes falsas, e

classificar aqueles padrões de repetição que estão associados com o subconjunto de pontuações de classificação que são indicativos das classificações de variante falsa como causadores dos erros específicos de sequência.

2. Sistema, de acordo com a reivindicação 1, **caracterizado** pelo fato de que o subsistema de correlação de erros específicos de sequência é configurado adicionalmente para:

classificar comprimentos e posições de deslocamento dos padrões de repetição que são associados com o subconjunto das pontuações de classificação como causadores dos erros específicos de sequência.

3. Sistema, de acordo com qualquer uma das reivindicações 1 ou 2, **caracterizado** pelo fato de que o nucleotídeo variante está na posição de destino flanqueada por pelo menos 20 nucleotídeos em cada lado.

4. Sistema, de acordo com qualquer uma das reivindicações de 1 a 3, **caracterizado** pelo fato de que o subsistema de filtro de variante pré-treinado é configurado para processar cada combinação dos padrões de repetição sobrepostos em pelo menos 100 sequências de nucleotídeos em pelo menos 100 amostras sobrepostas.

5. Sistema, de acordo com qualquer uma das reivindicações de 1 a 4, **caracterizado** pelo fato de que os padrões de repetição incluem pelo menos uma base das quatro bases (A, C, G e T) com pelo menos seis variações nos fatores de repetição dos respectivos padrões de repetição.

6. Sistema, de acordo com a reivindicação 5, **caracterizado** pelo fato de que os padrões de repetição são homopolímeros de uma base única (A, C, G ou T) com pelo menos seis variações nos fatores de repetição dos respectivos padrões de repetição; e

em que os pelo menos seis variações nos fatores de repetição especificam um número de repetições de uma base única nos padrões de repetição.

7. Sistema, de acordo com qualquer uma das reivindicações de 1 a 6, **caracterizado** pelo fato de que os padrões de repetição são copolímeros de pelo menos duas bases de quatro bases (A, C, G e T) com os pelo menos seis variações nos fatores de repetição; e

em que os pelo menos seis variações nos fatores de repetição especificam um número de repetições das pelo menos duas bases nos padrões de repetição.

8. Sistema, de acordo com qualquer uma das reivindicações de 1 a 7, **caracterizado** pelo fato de que as posições de deslocamento variam em termos de uma posição na qual os padrões repetição são sobrepostos nas sequências de nucleotídeos, mensuráveis como um deslocamento entre uma posição de origem dos padrões de repetição e uma posição de origem das sequências de nucleotídeos, e pelo menos dez deslocamentos são usados para produzir as amostras sobrepostas.

9. Sistema, de acordo com qualquer uma das reivindicações de 1 a 8, **caracterizado** pelo fato de que os padrões de repetição estão à direita de um nucleotídeo central nas amostras sobrepostas e não se sobreporem ao nucleotídeo central.

10. Sistema, de acordo com qualquer uma das reivindicações de 1 a 9, **caracterizado** pelo fato de que os padrões de repetição estão à esquerda de um nucleotídeo central nas amostras sobrepostas e não se sobreporem ao nucleotídeo central.

11. Sistema, de acordo com qualquer uma das reivindicações de 1 a 10, **caracterizado** pelo fato de que os padrões de repetição estão sobrepostos em um nucleotídeo central nas amostras sobrepostas.

12. Sistema, de acordo com qualquer uma das reivindicações de 1 a 11, **caracterizado** pelo fato de que os fatores de repetição são números inteiros na faixa de cinco a um quarto de uma contagem de nucleotídeos nas amostras sobrepostas.

13. Sistema, de acordo com a reivindicação 6, **caracterizado** pelo fato de que é configurado adicionalmente para aplicar padrões de repetição que são os homopolímeros da base única para cada uma das quatro bases (A, C, G e T).

14. Sistema, de acordo com a reivindicação 13, **caracterizado** pelo fato de que o subsistema de preparação de entrada é configurado adicionalmente para produzir os padrões de repetição e as amostras sobrepostas para os homopolímeros para cada uma das quatro bases.

15. Sistema, de acordo com a reivindicação 14, **caracterizado** pelo fato de que os padrões de repetição estão posicionados à direita de um nucleotídeo central nas amostras sobrepostas e uma exibição das justaposições de distribuições aos homopolímeros para as quatro bases.

16. Sistema, de acordo com a reivindicação 14, **caracterizado** pelo fato de que os padrões de repetição estão posicionados à esquerda de um nucleotídeo central nas amostras sobrepostas e uma exibição das justaposições de distribuições aos homopolímeros para as quatro bases.

17. Sistema, de acordo com qualquer uma das reivindicações de 1 a 16, **caracterizado** pelo fato de que as sequências de nucleotídeos nas quais os padrões de repetição são sobrepostos são geradas aleatoriamente.

18. Sistema, de acordo com qualquer uma das reivindicações de 1 a 17, **caracterizado** pelo fato de que as sequências de nucleotídeos nas quais os padrões de repetição são sobrepostos são selecionadas

aleatoriamente a partir de sequências de nucleotídeos de DNA que ocorrem naturalmente.

19. Sistema, de acordo com qualquer uma das reivindicações de 1 a 18, **caracterizado** pelo fato de que um subsistema de análise é configurado para causar a exibição das distribuições das pontuações de classificação para cada um dos fatores de repetição.

20. Sistema, de acordo com qualquer uma das reivindicações de 1 a 19, **caracterizado** pelo fato de que o subsistema de filtro de variantes pré-treinado é treinado em pelo menos 500.000 exemplos de treinamento de variantes verdadeiras e pelo menos 50.000 exemplos de variantes falsas; e

em que cada exemplo de treinamento é uma sequência de nucleotídeo com um nucleotídeo variante em uma posição de destino flanqueada por pelo menos 20 nucleotídeos em cada lado.

21. Sistema, de acordo com qualquer uma das reivindicações de 1 a 20, **caracterizado** pelo fato de que o subsistema de filtro de variante pré-treinado tem camadas convolucionais, uma camada totalmente conectada e uma camada de classificação.

22. Método implementado por computador para identificar padrões de repetição que causam erros específicos de sequência em dados de sequenciamento de nucleotídeos, **caracterizado** pelo fato de que inclui:

sobrepor computacionalmente padrões de repetição em teste em inúmeras sequências de nucleotídeos e produzir amostras sobrepostas, em que cada padrão de repetição representa uma composição de nucleotídeo particular que tem um comprimento particular e aparece em uma amostra sobreposta na posição de deslocamento particular, em que cada amostra sobreposta tem uma posição de destino considerada um nucleotídeo variante, e em que, para cada combinação da composição de nucleotídeo particular, do comprimento particular e da posição de deslocamento particular, um conjunto de amostras sobrepostas é gerado

computacionalmente;

processar as amostras sobrepostas por uma rede neural convolucional e, com base na detecção de padrões de nucleotídeo nas amostras sobrepostas por filtros de convolução da rede neural convolucional, gerar pontuações de classificação para a probabilidade de que o nucleotídeo variante em cada uma das amostras sobrepostas seja uma variante verdadeira ou uma variante falsa;

emitir distribuições das pontuações de classificação gerados pelo subsistema de filtro de variante pré treinado para fatores de repetição de respectivos padrões de repetição, em que as distribuições indicam impacto no subsistema de filtro de variante pré-treinado dos padrões de repetição; e

especificar, baseado em um limite, um subconjunto das pontuações de classificação nas distribuições como indicativo de classificações de variante falsa, e classificar aqueles padrões de repetição que estão associados com o subconjunto de pontuações de classificação que são indicativos das classificações de variante falsa como causadores dos erros específicos de sequência.

23. Método implementado por computador, de acordo com a reivindicação 22, **caracterizado** pelo fato de que implementa ainda qualquer uma das reivindicações 1 a 21 pelo exercício dos sistemas reivindicados.

24. Meio de armazenamento legível por computador não transitório, impresso com instruções de programa de computador para identificar padrões de repetição que causam erros específicos de sequência em dados de sequenciamento de nucleotídeo, **caracterizado** pelo fato de que as instruções, quando executadas em um processador, implementam um método implementado por computador que compreende:

sobrepor computacionalmente padrões de repetição em teste em inúmeras sequências de nucleotídeos e produzir amostras sobrepostas, em que cada padrão de repetição representa uma composição de nucleotídeo

particular que tem um comprimento particular e aparece em uma amostra sobreposta em uma posição de deslocamento particular, em que cada amostra sobreposta tem uma posição de destino considerada um nucleotídeo variante, e em que, para cada combinação da composição de nucleotídeo particular, do comprimento particular e da posição de deslocamento particular, um conjunto das amostras sobrepostas é gerado computacionalmente;

processar amostras sobrepostas por uma rede neural convolucional e, com base na detecção de padrões de nucleotídeo nas amostras sobrepostas por filtros de convolução da rede neural convolucional, gerar pontuações de classificação para a probabilidade de que o nucleotídeo variante em cada uma das amostras sobrepostas seja uma variante verdadeira ou uma variante falsa;

emitir distribuições das pontuações de classificação gerados pelo subsistema de filtro de variante pré treinado para fatores de repetição de respectivos padrões de repetição, em que as distribuições indicam impacto no subsistema de filtro de variante pré-treinado dos padrões de repetição; e

especificar, baseado em um limite, um subconjunto das pontuações de classificação nas distribuições como indicativo de classificações de variante falsa e classificar aqueles padrões de repetição que estão associados com o subconjunto das pontuações de classificação que são indicativos das classificações de variante falsa como causadores dos erros específicos de sequência.

25. Meio de armazenamento legível por computador não transitório, de acordo com a reivindicação 24, **caracterizado** pelo fato de que inclui ainda características que, quando combinadas com hardware, implementa qualquer uma das reivindicações 1 a 21.

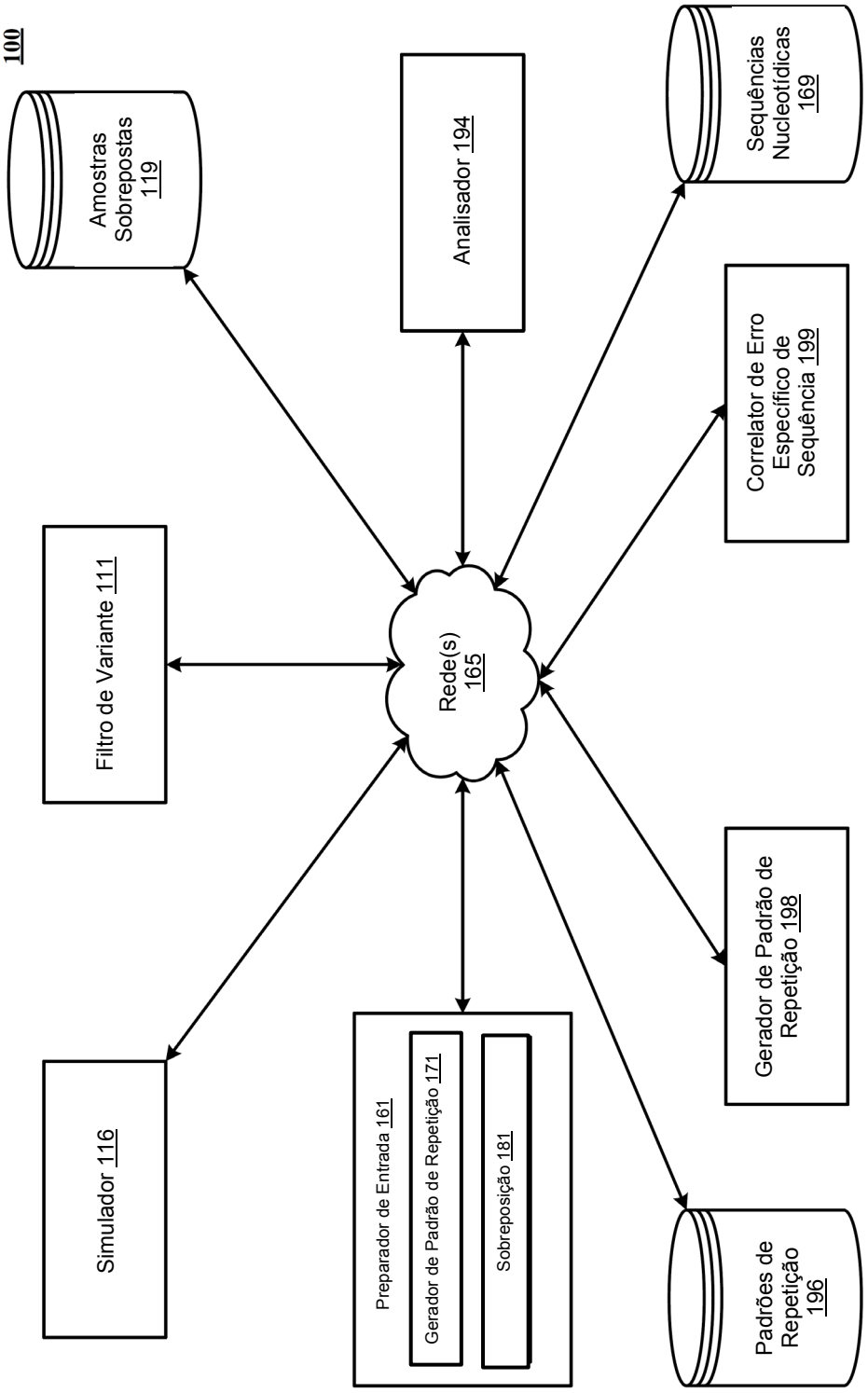

100
Amostras Sobrepostas 119
Analisador 194
Sequências Nucleotídicas 169
Correlator de Erro Específico de Sequência 199
Filtro de Variante 111
Rede(s) 165
Gerador de Padrão de Repetição 198
Simulador 116
Preparador de Entrada 161
Gerador de Padrão de Repetição 171
Sobreposição 181
Padrões de Repetição 196

FIG. 1

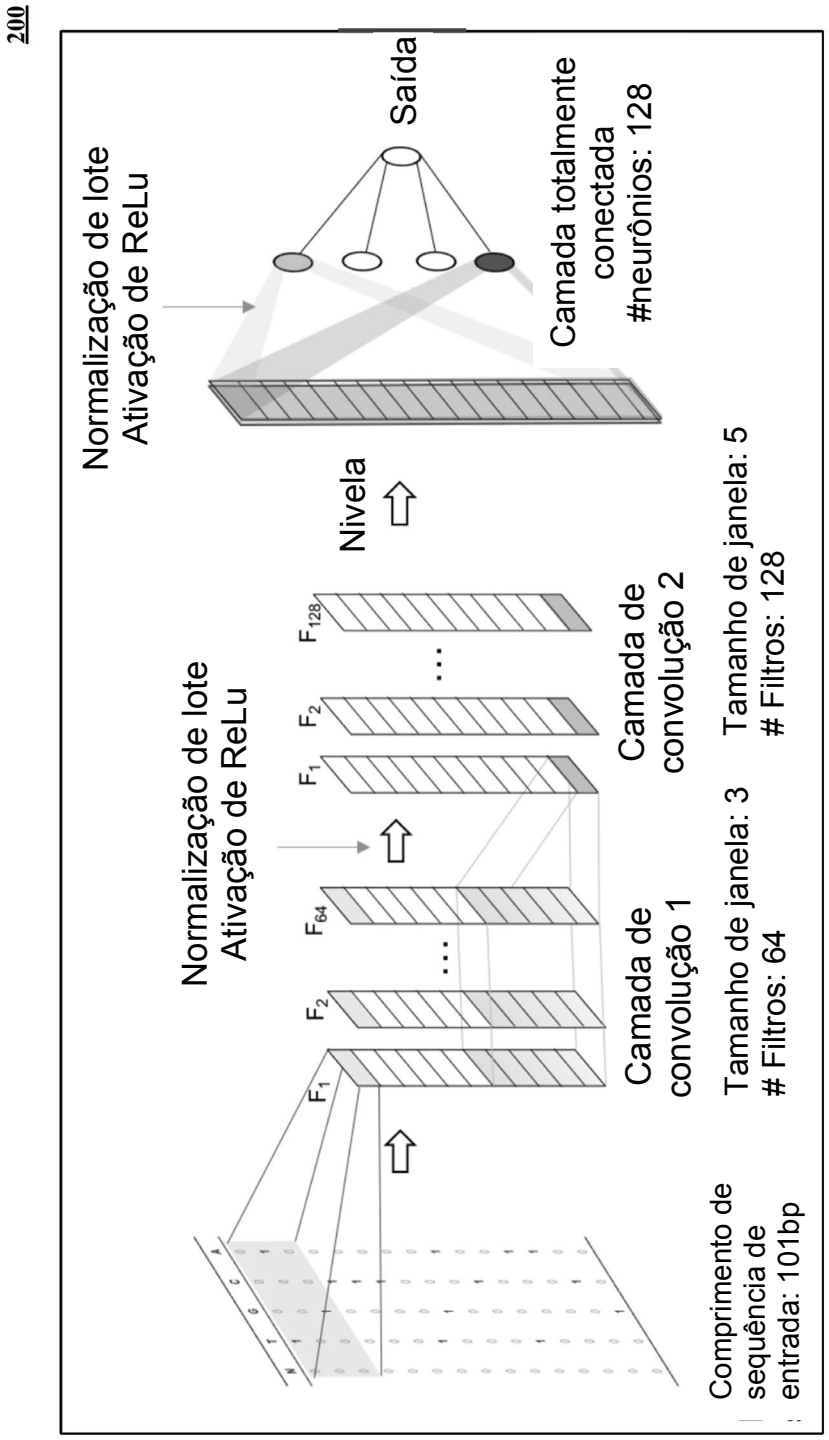

200
Comprimento de sequência de entrada: 101bp
Camada de convolução 1
Tamanho de janela: 3
Filtros: 64
Camada de convolução 2
Tamanho de janela: 5
Filtros: 128
Normalização de lote
Ativação de ReLu
Nivela
Camada totalmente conectada
#neurônios: 128
Saída

FIG. 2

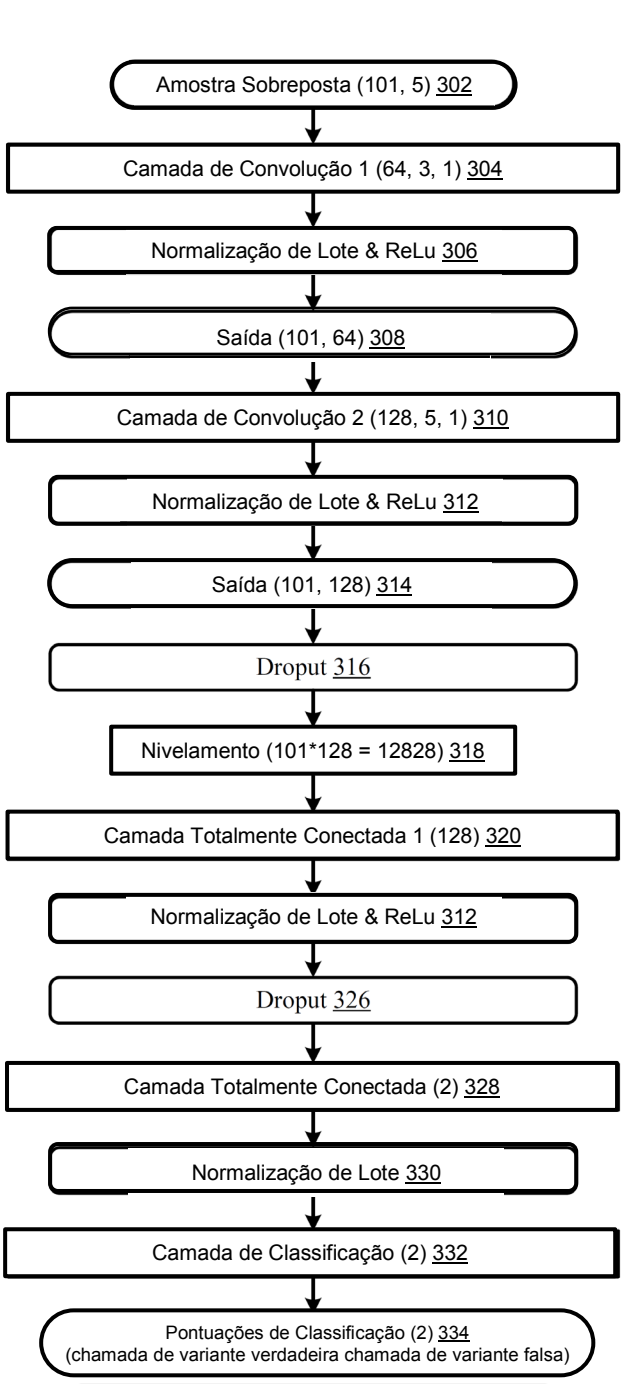

300
Amostra Sobreposta (101, 5) 302
Camada de Convolução 1 (64, 3, 1) 304
Normalização de Lote & ReLu 306
Saída (101, 64) 308
Camada de Convolução 2 (128, 5, 1) 310
Normalização de Lote & ReLu 312
Saída (101, 128) 314
Droput 316
Nivelamento (101*128 = 12828) 318
Camada Totalmente Conectada 1 (128) 320
Normalização de Lote & ReLu 312
Droput 326
Camada Totalmente Conectada (2) 328
Normalização de Lote 330
Camada de Classificação (2) 332
Pontuações de Classificação (2) 334
(chamada de variante verdadeira chamada de variante falsa)

FIG. 3

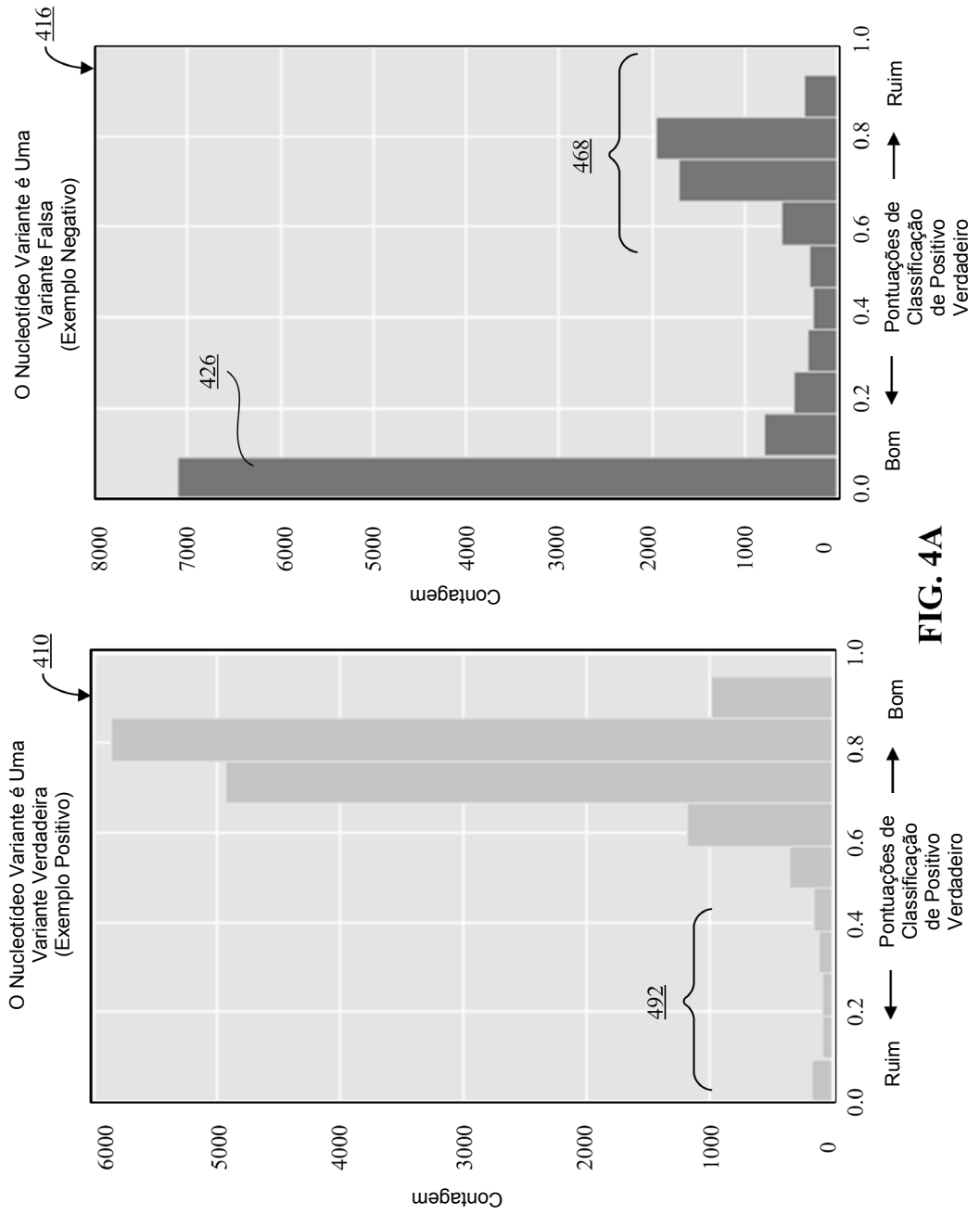

416
O Nucleotídeo Variante é Uma Variante Falsa (Exemplo Negativo)
426
468
Contagem
8000
7000
6000
5000
4000
3000
2000
1000
0
0.0
0.2
0.4
0.6
0.8
1.0
Bom
Ruim
Pontuações de Classificação de Positivo Verdadeiro
410
O Nucleotídeo Variante é Uma Variante Verdadeira (Exemplo Positivo)
492
Contagem
6000
5000
4000
3000
2000
1000
0
0.0
0.2
0.4
0.6
0.8
1.0
Ruim
Bom
Pontuações de Classificação de Positivo Verdadeiro

FIG. 4A

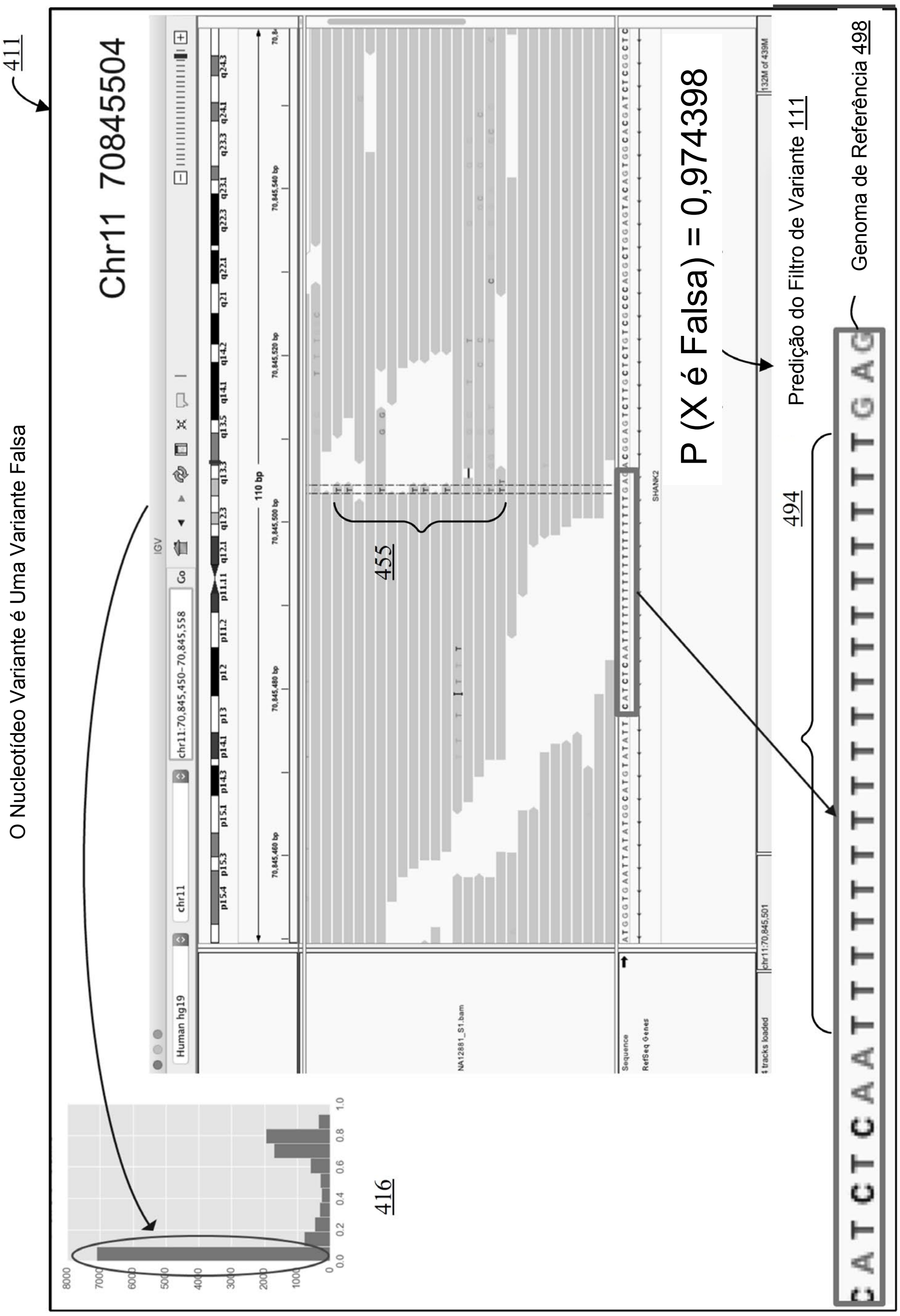

411
O Nucleotídeo Variante é Uma Variante Falsa
Chr11 70845504
P (X é Falsa) = 0,974398
Predição do Filtro de Variante 111
Genoma de Referência 498
455
494
416

FIG. 4B

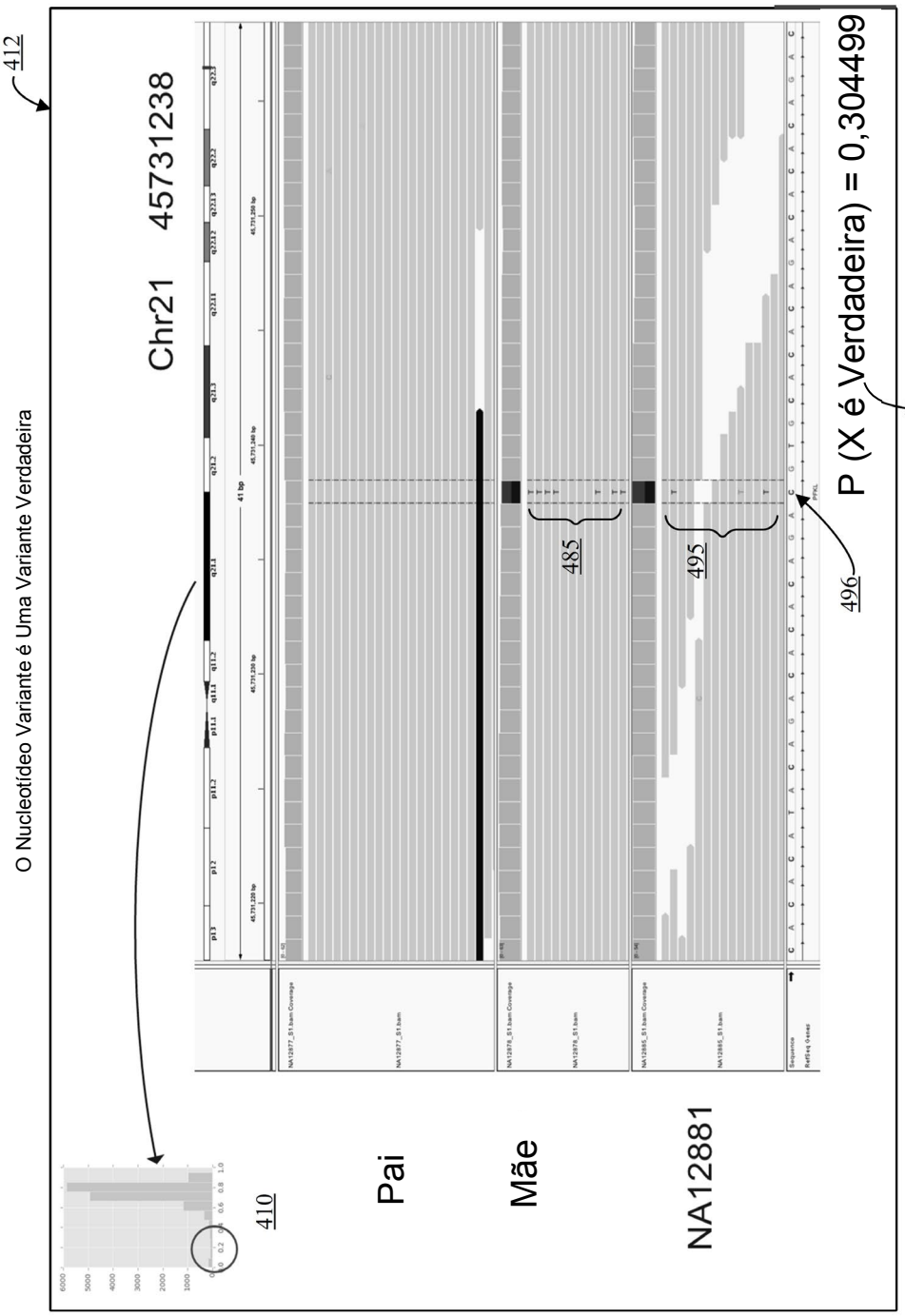

412
O Nucleotideo Variante é Uma Variante Verdadeira
Chr21 45731238
Pai
Mãe
NA12881
485
495
496
P (X é Verdadeira) = 0,304499
Predição do Filtro de Variante 111
410

FIG. 4C

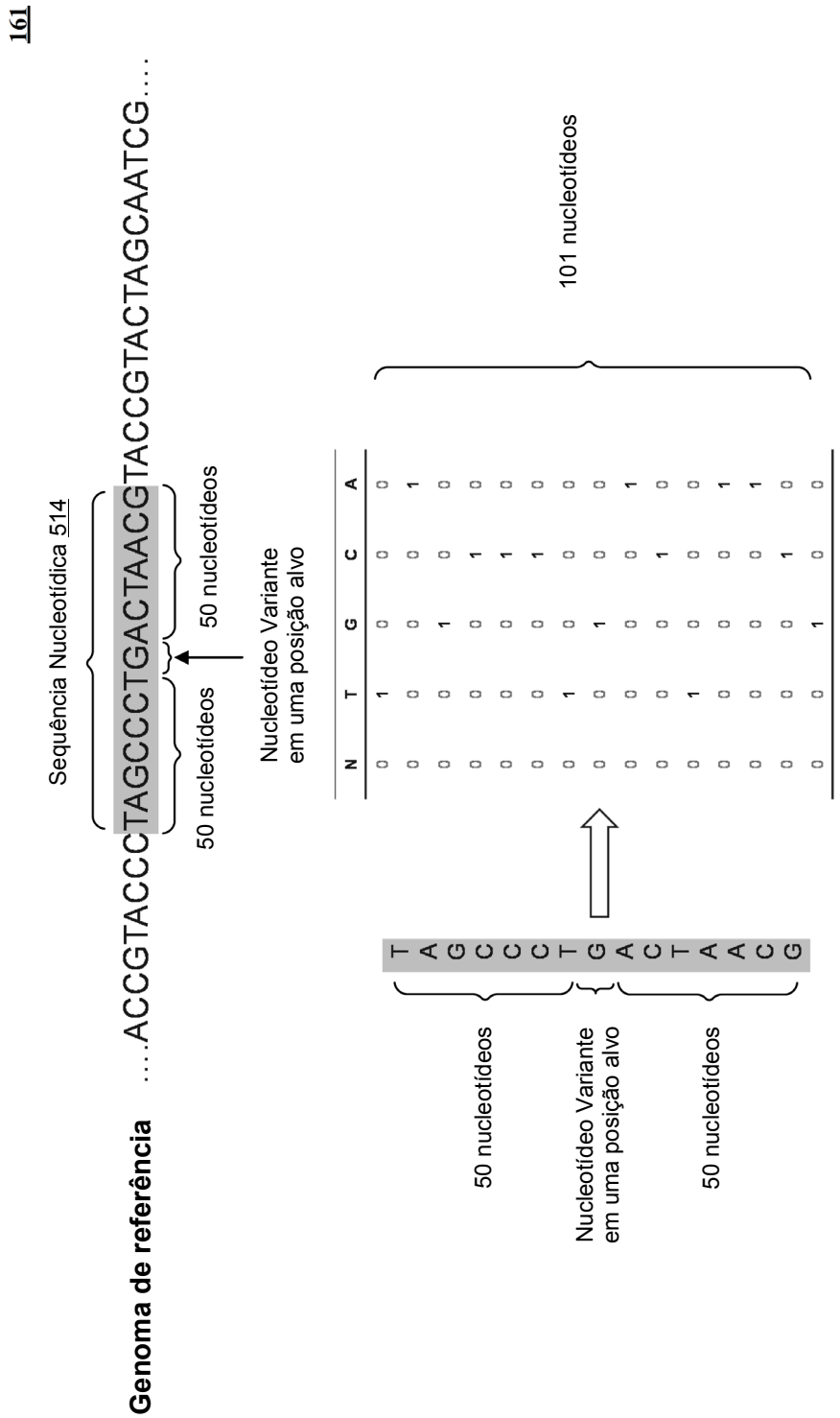

161
Sequência Nucleotídica 514
Genoma de referência
....ACCGTACCCTAGCCCTGACTAACGTACCGTACTAGCAATCG....
50 nucleotídeos
50 nucleotídeos
Nucleotídeo Variante em uma posição alvo
N T G C A
101 nucleotídeos
T A G C C C T G A C T A A C G
50 nucleotídeos
Nucleotídeo Variante em uma posição alvo
50 nucleotídeos

FIG. 5

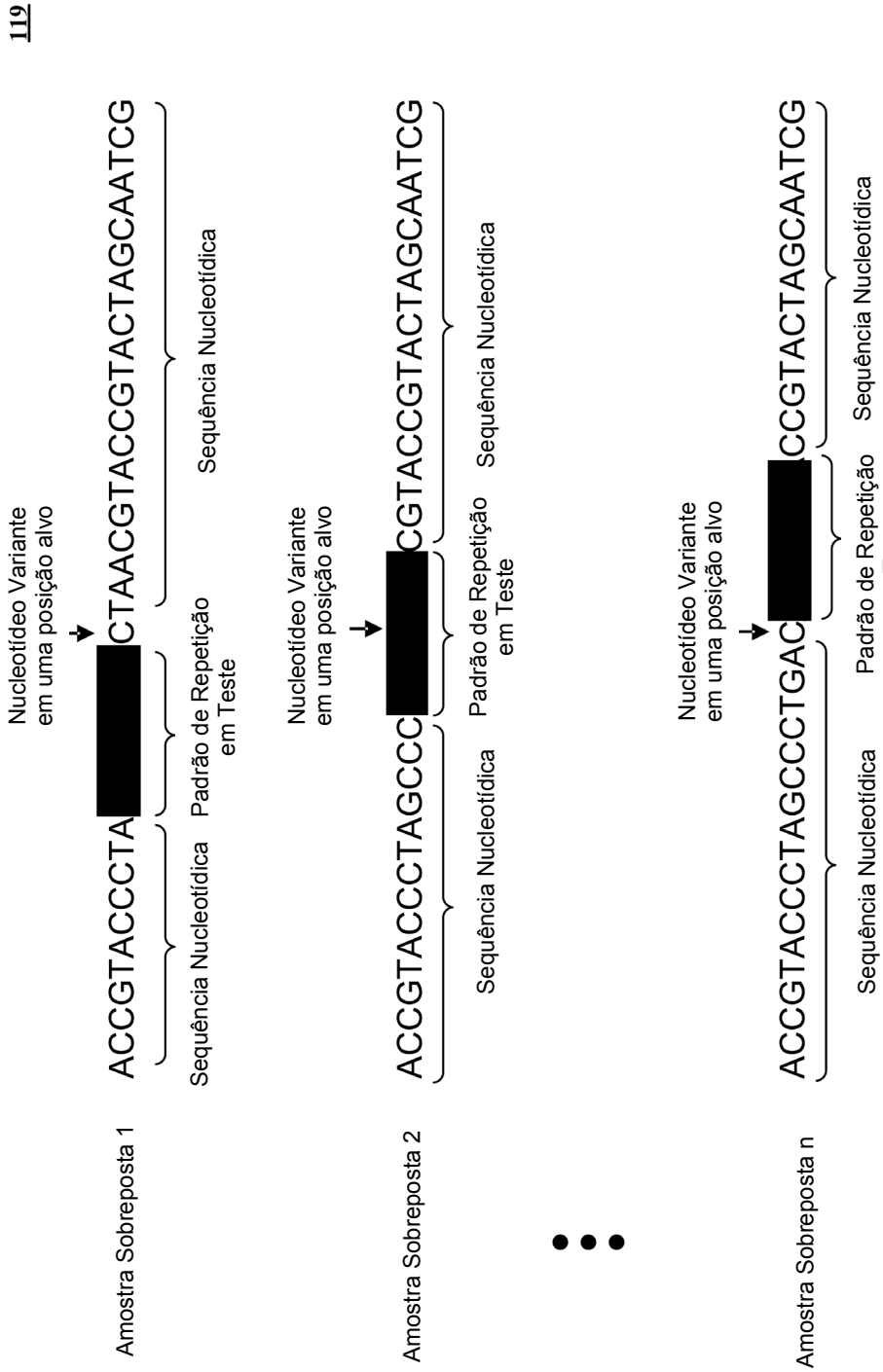

119
Amostra Sobreposta 1
ACCGTACCCTA
CTAACGTACCGTACTAGCAATCG
Nucleotídeo Variante em uma posição alvo
Sequência Nucleotídica
Padrão de Repetição em Teste
Sequência Nucleotídica
Amostra Sobreposta 2
ACCGTACCCTAGCCC
CGTACCGTACTAGCAATCG
Nucleotídeo Variante em uma posição alvo
Sequência Nucleotídica
Padrão de Repetição em Teste
Sequência Nucleotídica
Amostra Sobreposta n
ACCGTACCCTAGCCCTGAC
CCGTACTAGCAATCG
Nucleotídeo Variante em uma posição alvo
Sequência Nucleotídica
Padrão de Repetição em Teste
Sequência Nucleotídica

FIG. 6

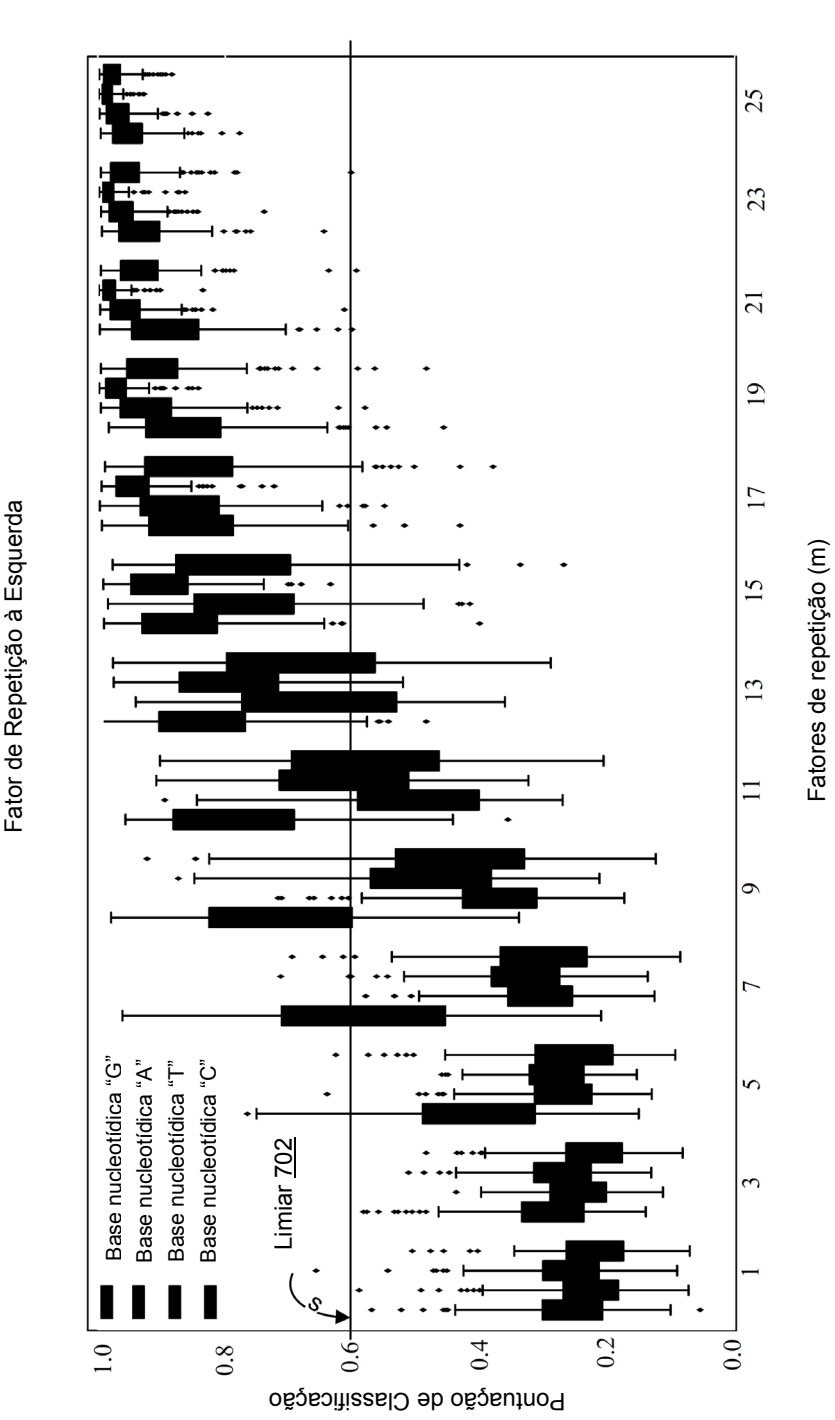

194
Fator de Repetição à Esquerda
Base nucleotídica "G"
Base nucleotídica "A"
Base nucleotídica "T"
Base nucleotídica "C"
Limiar 702
Pontuação de Classificação
1.0
0.8
0.6
0.4
0.2
0.0
1
3
5
7
9
11
13
15
17
19
21
23
25
Fatores de repetição (m)

FIG. 7A

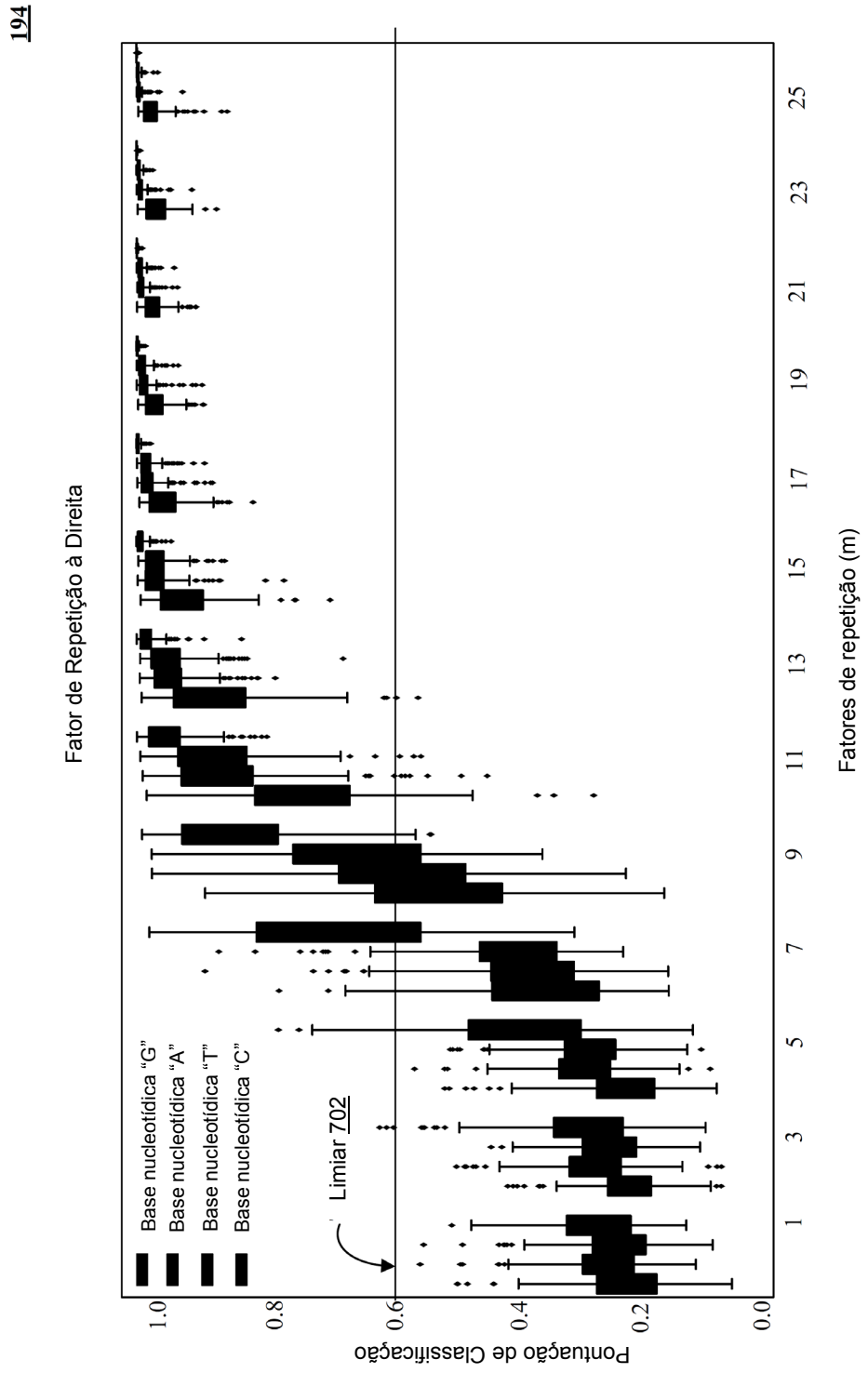

194
Fator de Repetição à Direita
Base nucleotídica "G"
Base nucleotídica "A"
Base nucleotídica "T"
Base nucleotídica "C"
Limiar 702
Pontuação de Classificação
1.0
0.8
0.6
0.4
0.2
0.0
1
3
5
7
9
11
13
15
17
19
21
23
25
Fatores de repetição (m)

FIG. 7B

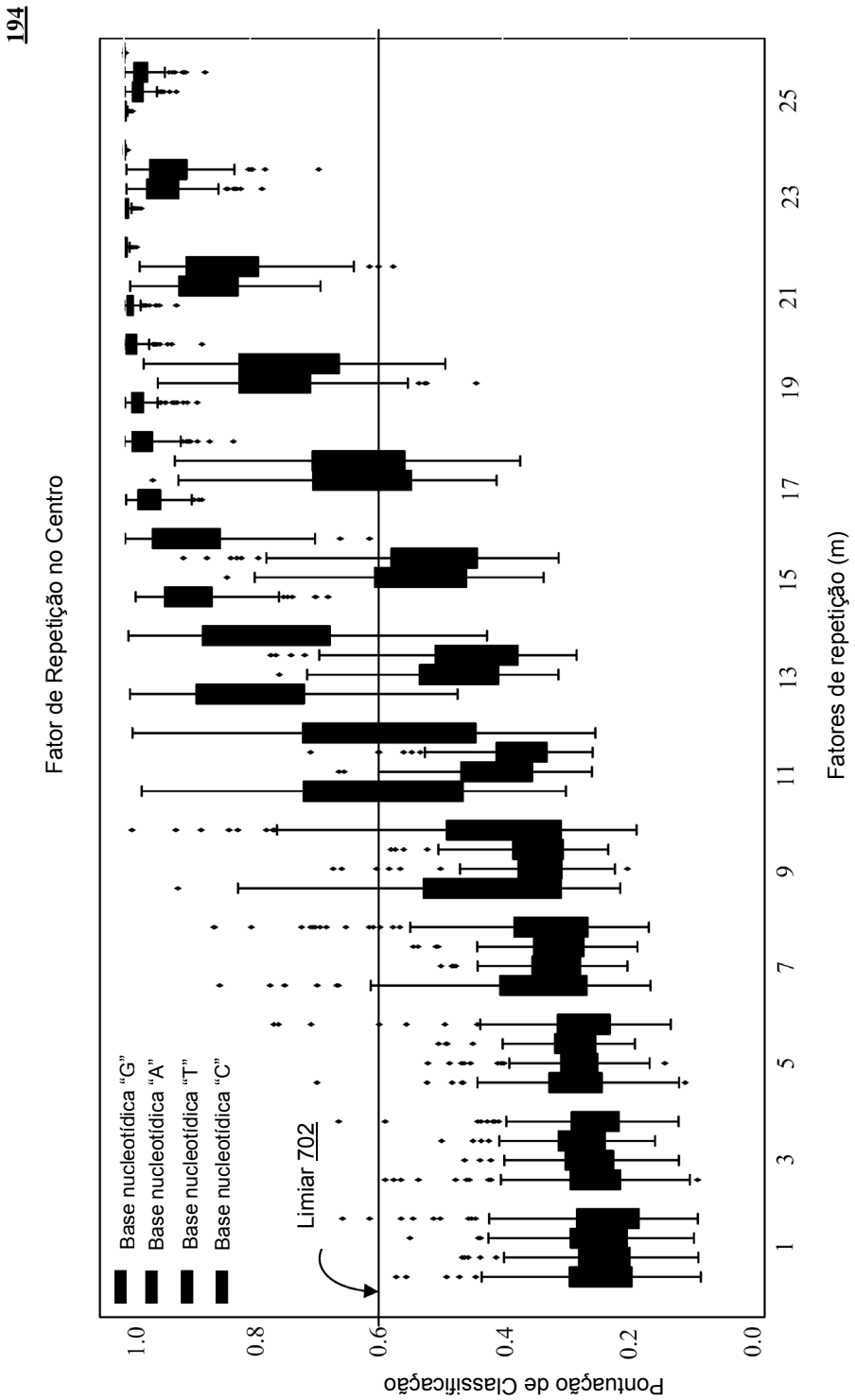

194
Fator de Repetição no Centro
Pontuação de Classificação
Fatores de repetição (m)
Base nucleotídica "G"
Base nucleotídica "A"
Base nucleotídica "T"
Base nucleotídica "C"
Limiar 702
1.0
0.8
0.6
0.4
0.2
0.0
1
3
5
7
9
11
13
15
17
19
21
23
25

FIG. 7C

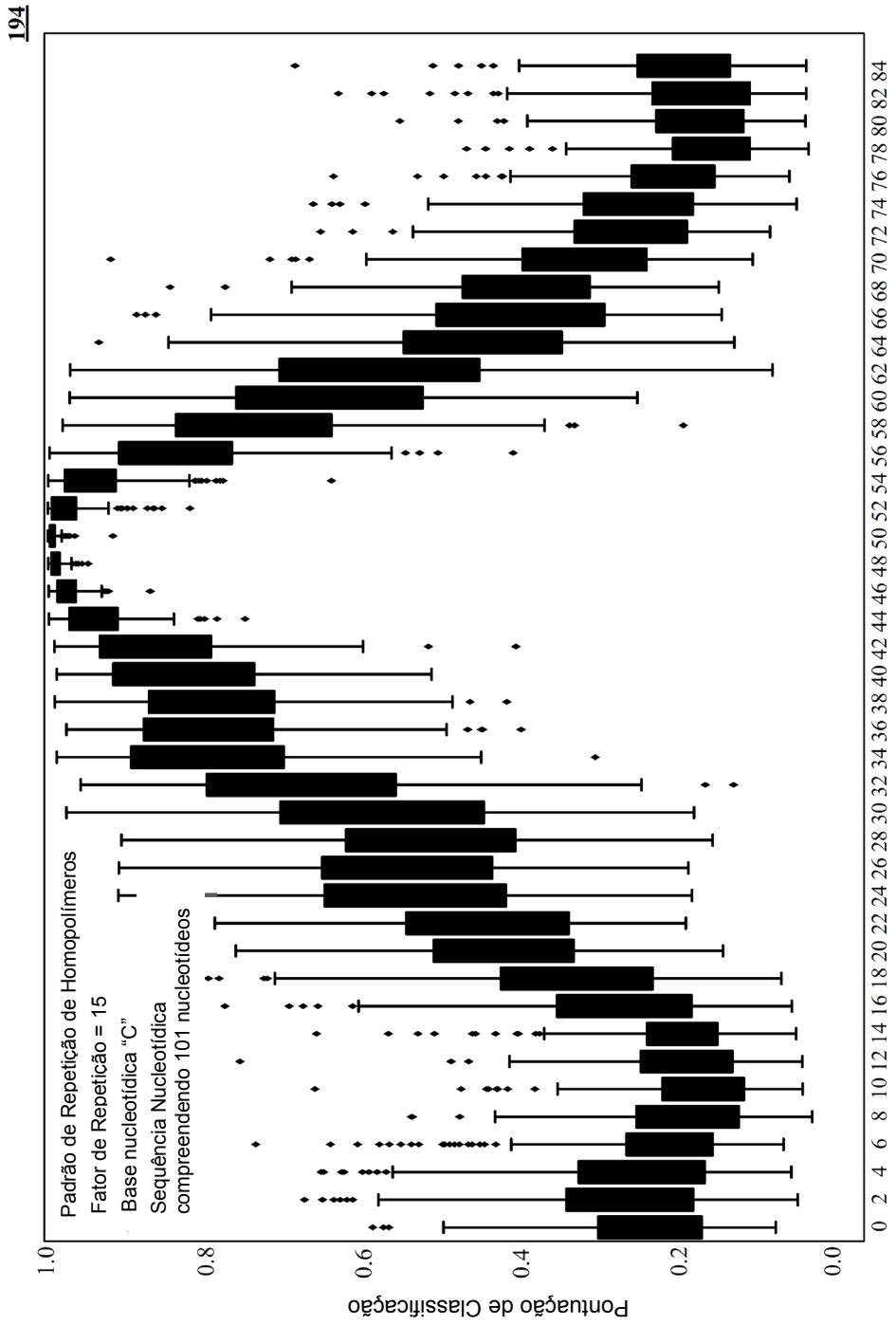

194
Padrão de Repetição de Homopolímeros
Fator de Repetição = 15
Base nucleotídica "C"
Sequência Nucleotídica
compreendendo 101 nucleotídeos
Pontuação de Classificação
1.0
0.8
0.6
0.4
0.2
0.0
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48 50 52 54 56 58 60 62 64 66 68 70 72 74 76 78 80 82 84
Posição de deslocamento

FIG. 8A

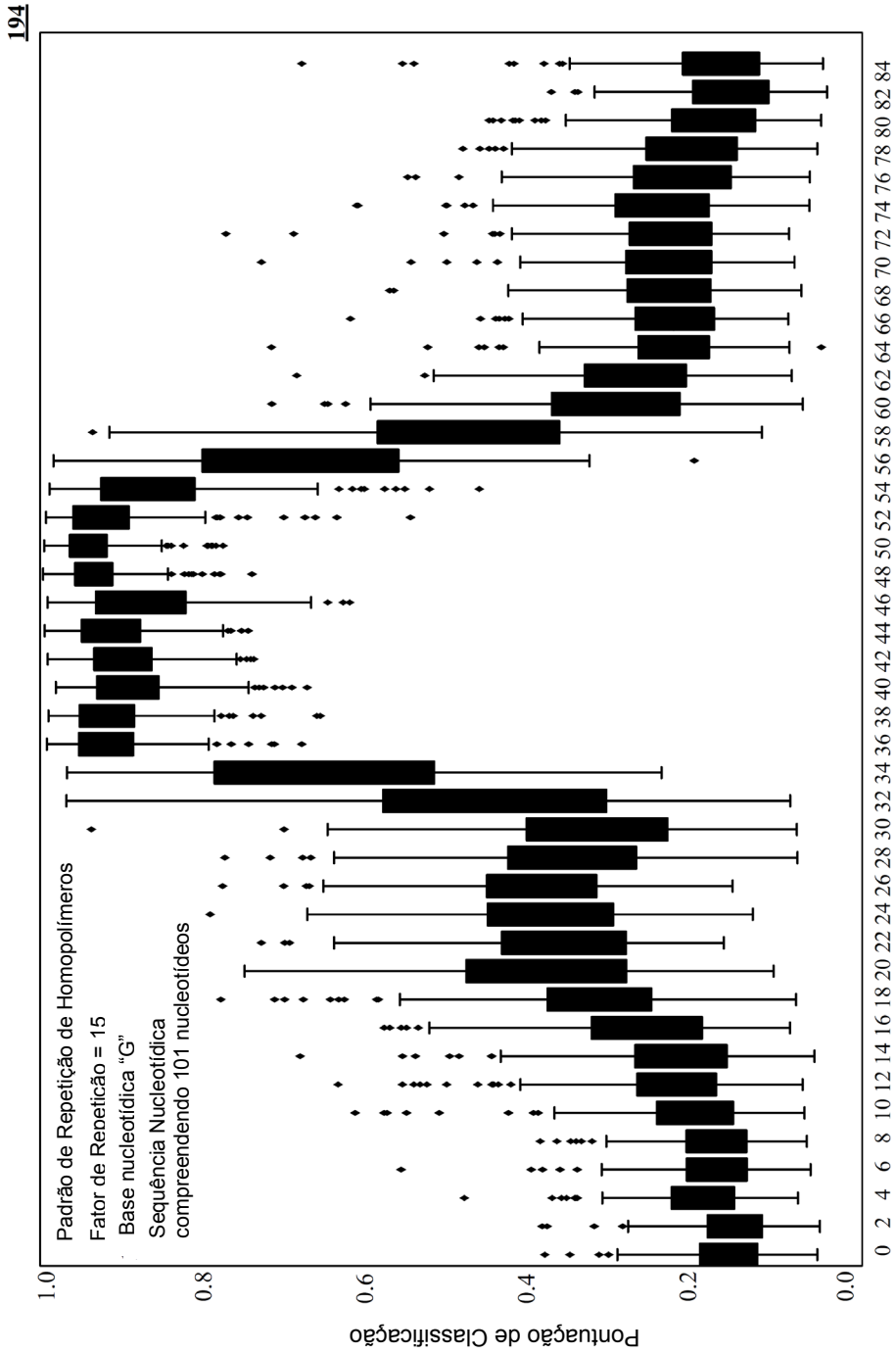

194
Padrão de Repetição de Homopolímeros
Fator de Repetição = 15
Base nucleotídica "G"
Sequência Nucleotídica
compreendendo 101 nucleotídeos
Pontuação de Classificação
1.0
0.8
0.6
0.4
0.2
0.0
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48 50 52 54 56 58 60 62 64 66 68 70 72 74 76 78 80 82 84
Posição de deslocamento

FIG. 8B

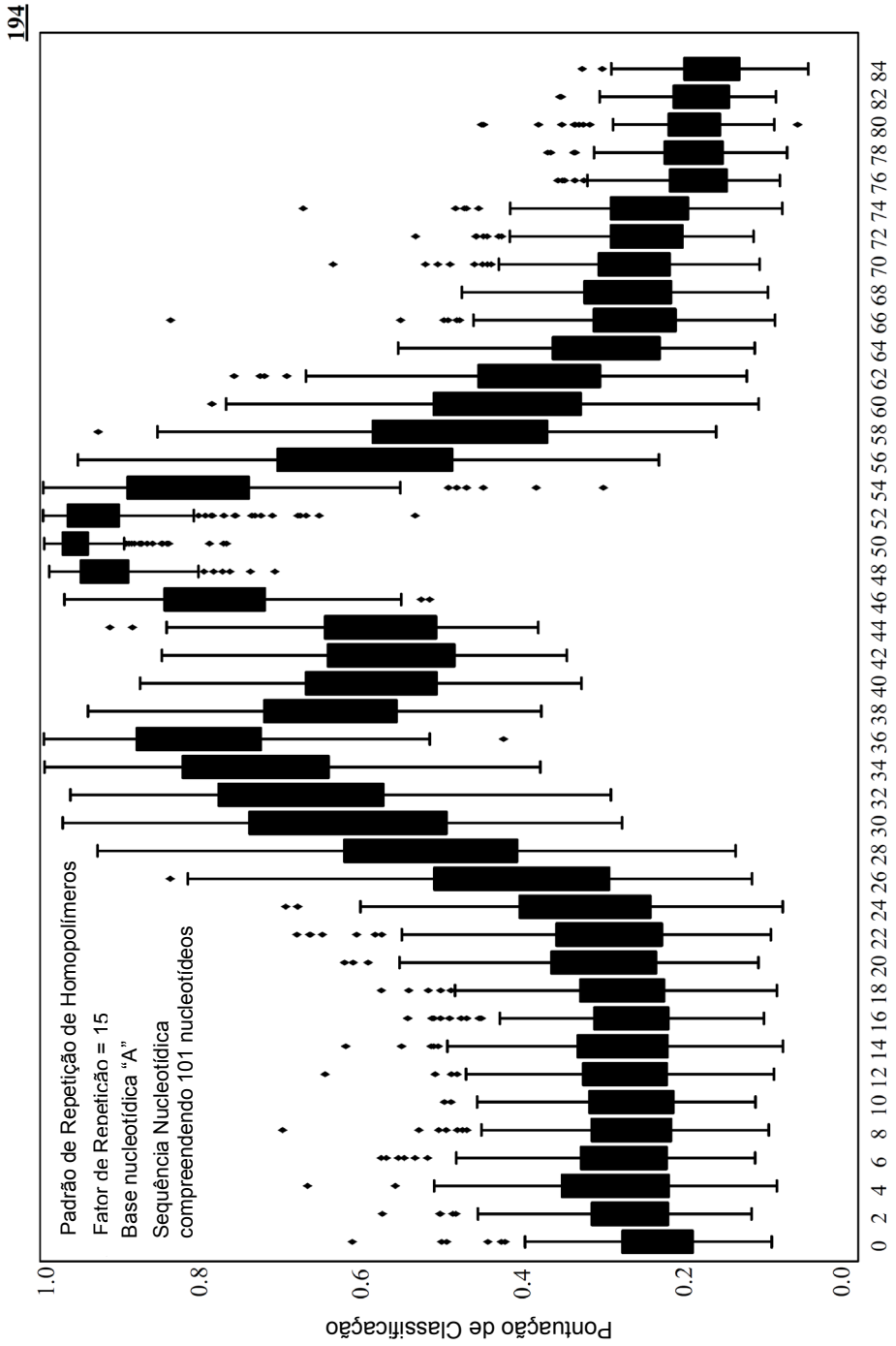

194
Padrão de Repetição de Homopolímeros
Fator de Repetição = 15
Base nucleotídica "A"
Sequência Nucleotídica
compreendendo 101 nucleotídeos
Pontuação de Classificação
1.0
0.8
0.6
0.4
0.2
0.0
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48 50 52 54 56 58 60 62 64 66 68 70 72 74 76 78 80 82 84
Posição de deslocamento

FIG. 8C

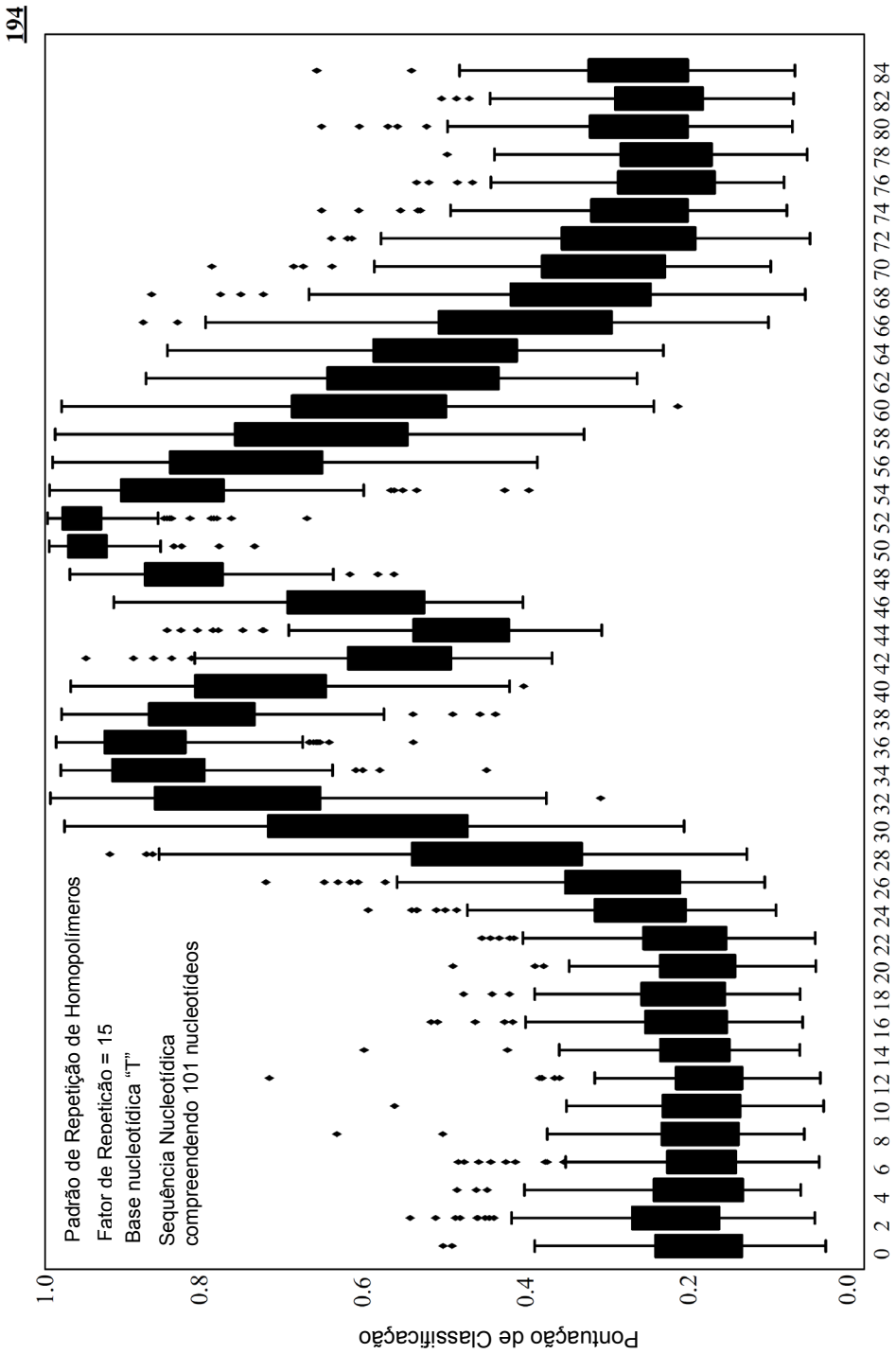

194
Padrão de Repetição de Homopolímeros
Fator de Repeticão = 15
Base nucleotídica "T"
Sequência Nucleotídica
compreendendo 101 nucleotídeos
Pontuação de Classificação
1.0
0.8
0.6
0.4
0.2
0.0
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48 50 52 54 56 58 60 62 64 66 68 70 72 74 76 78 80 82 84
Posição de deslocamento

FIG. 8D

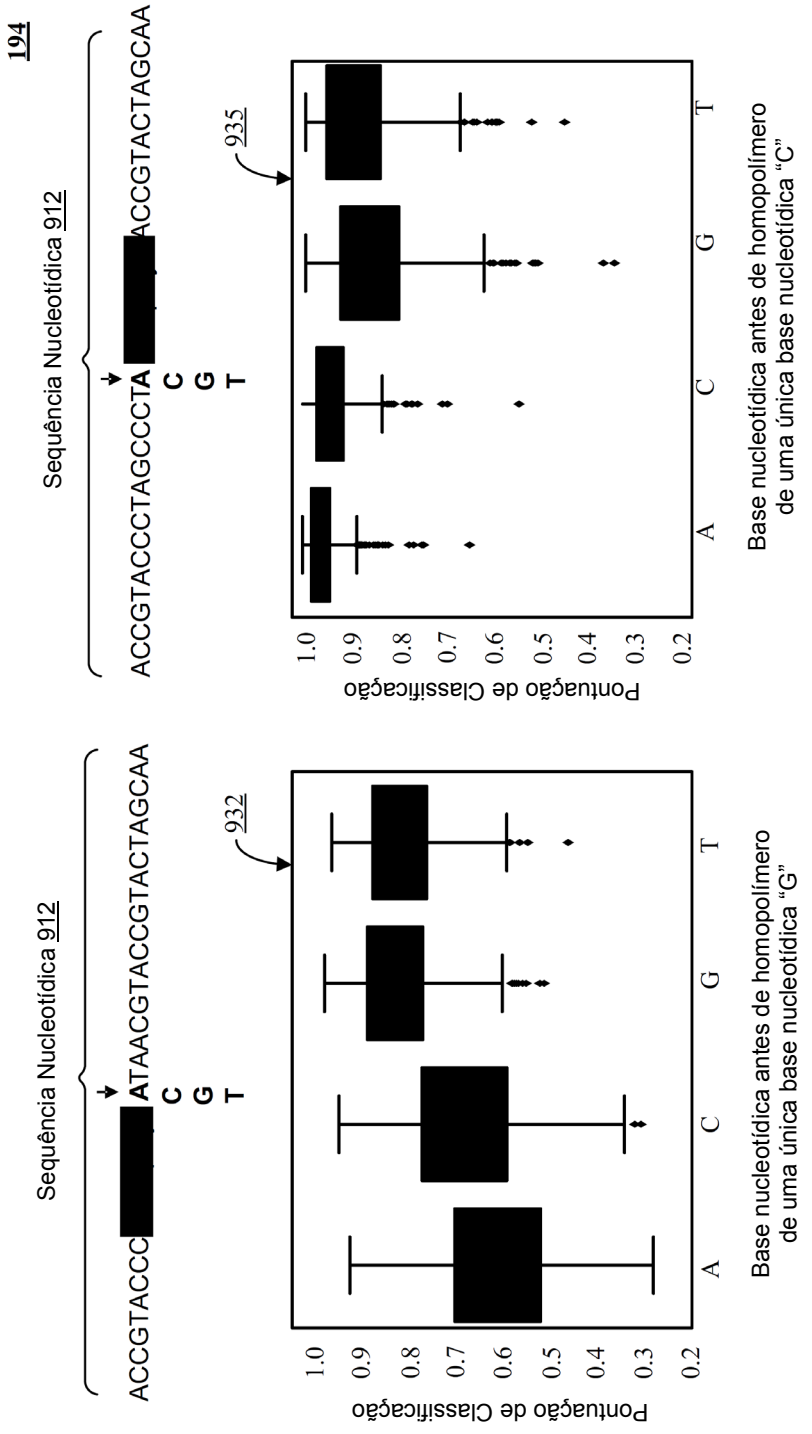

194
Sequência Nucleotídica 912
ACCGTACCC
ATAACGTACCGTACTAGCAA
C
G
T
932
Pontuação de Classificação
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
A
C
G
T
Base nucleotídica antes de homopolímero de uma única base nucleotídica "G"
Sequência Nucleotídica 912
ACCGTACCCTAGCCCTA
ACCGTACTAGCAA
C
G
T
935
Pontuação de Classificação
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
A
C
G
T
Base nucleotídica antes de homopolímero de uma única base nucleotídica "C"

FIG. 9

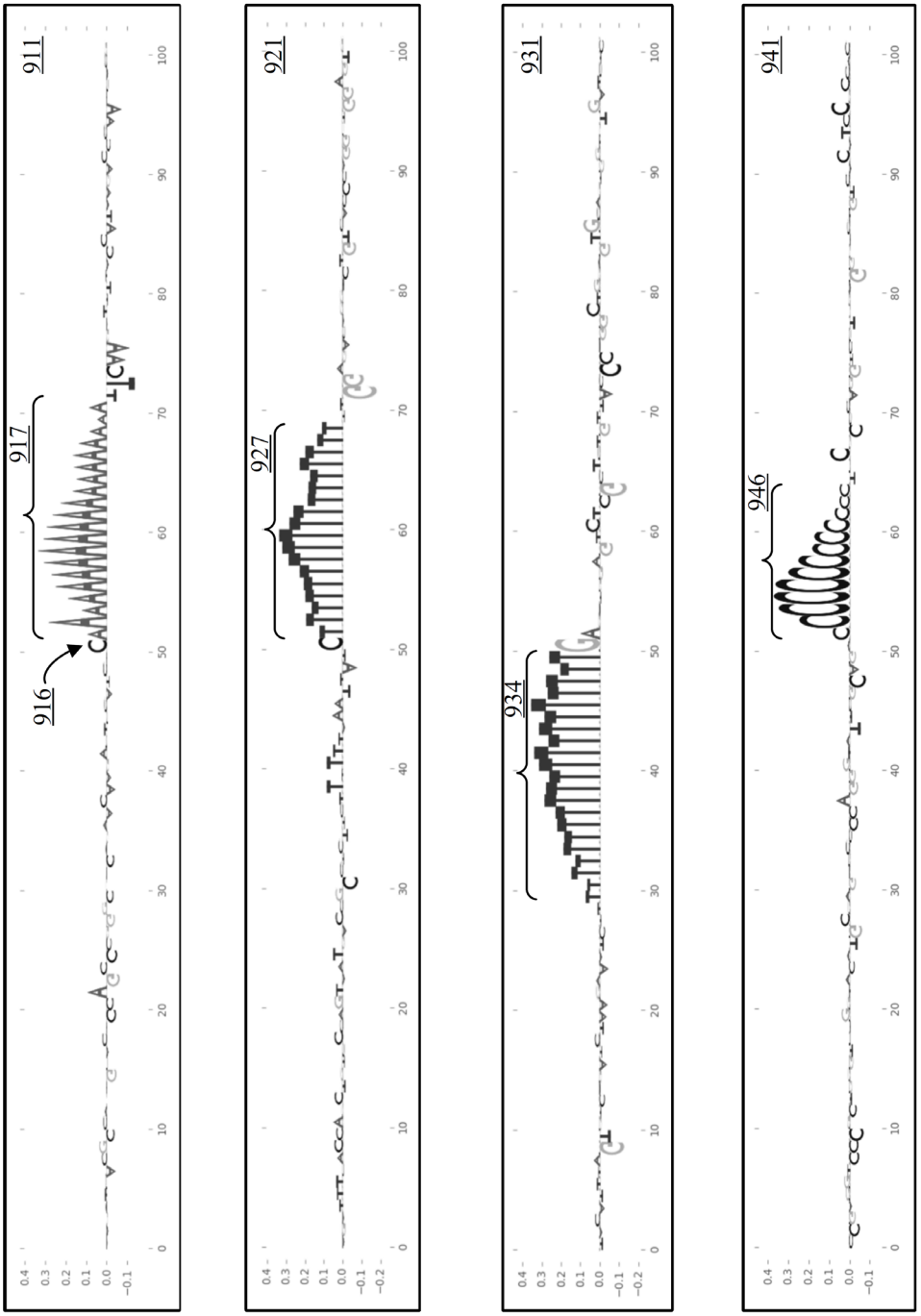


FIG. 10A

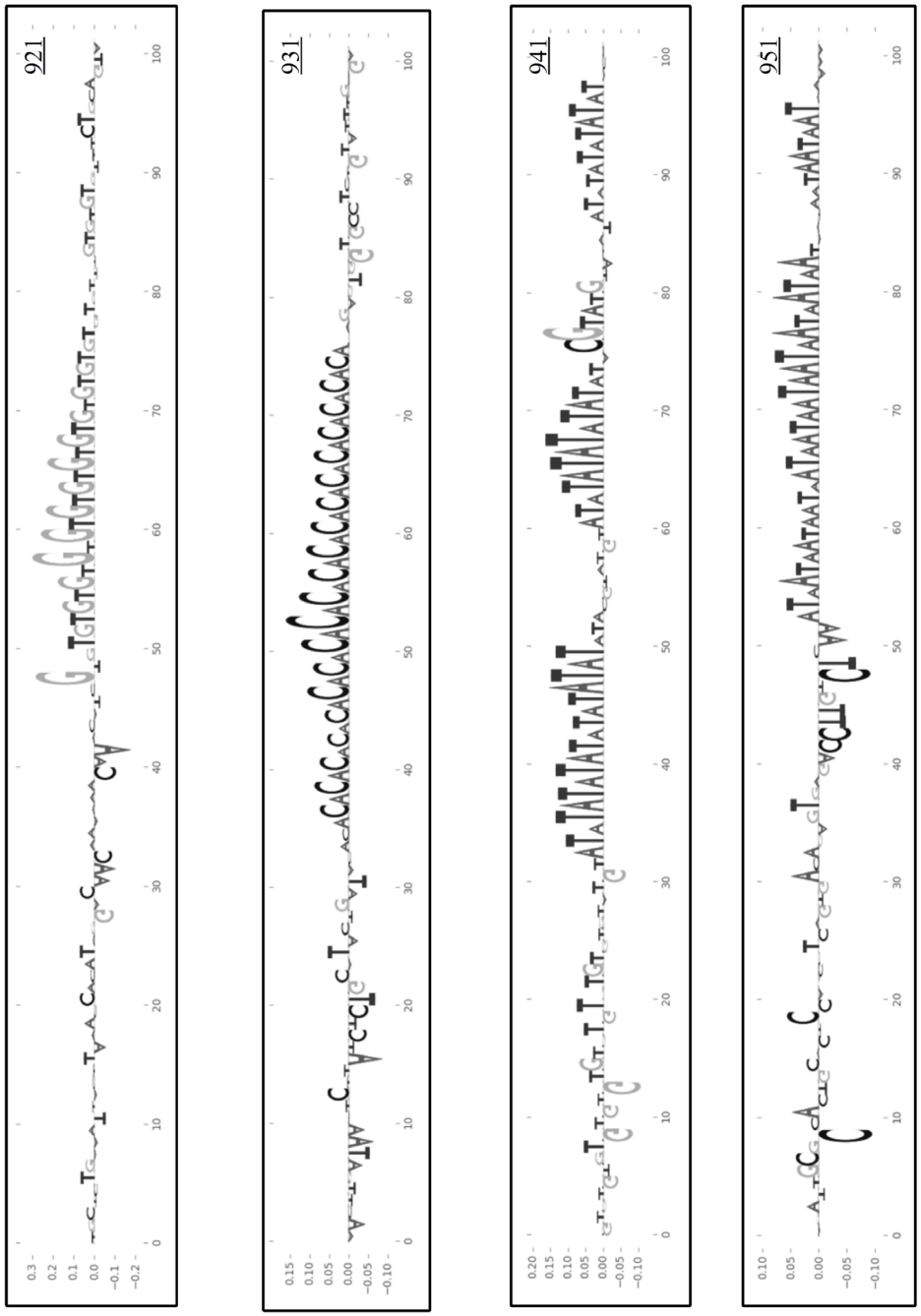


FIG. 10B

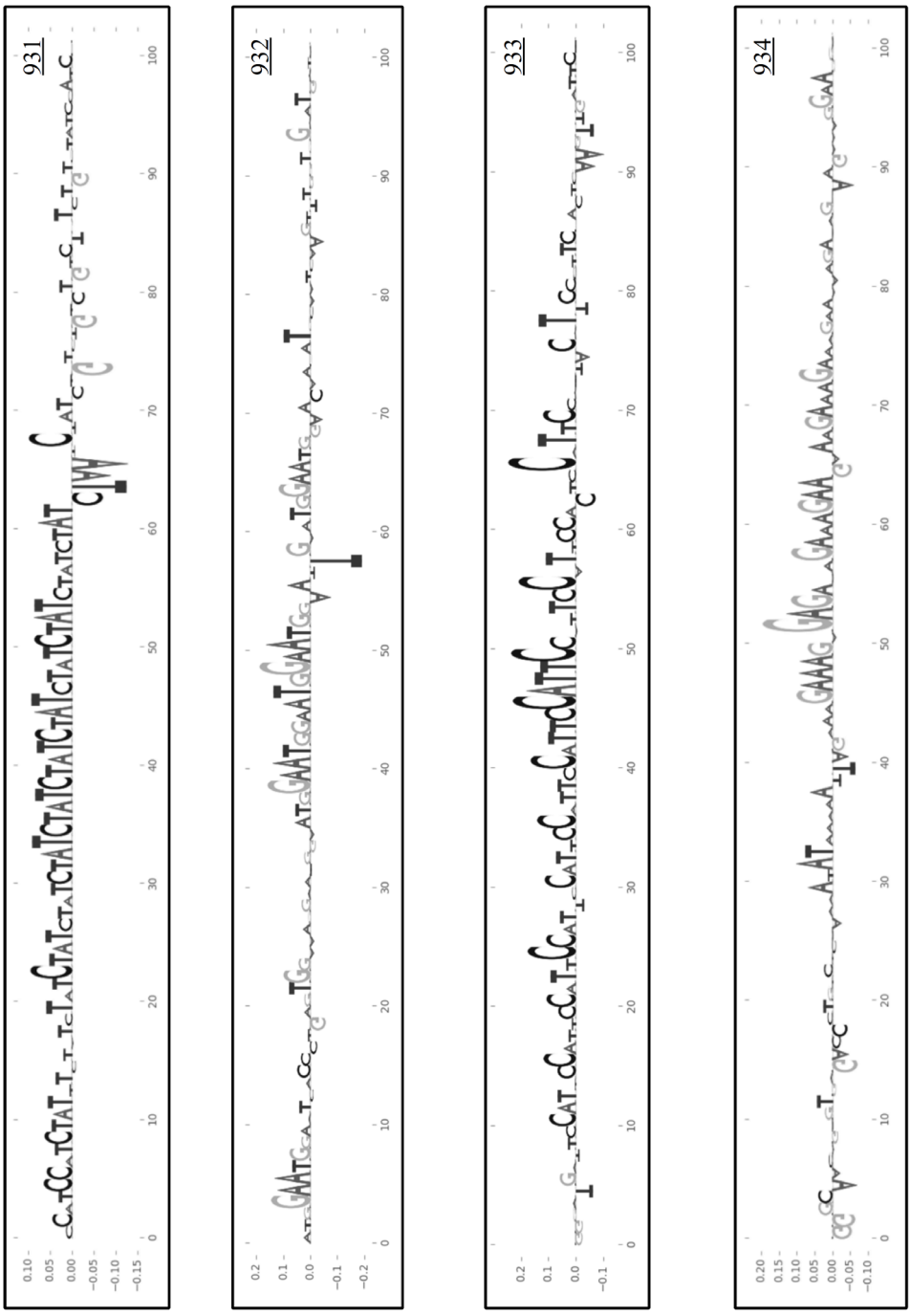


FIG. 10C

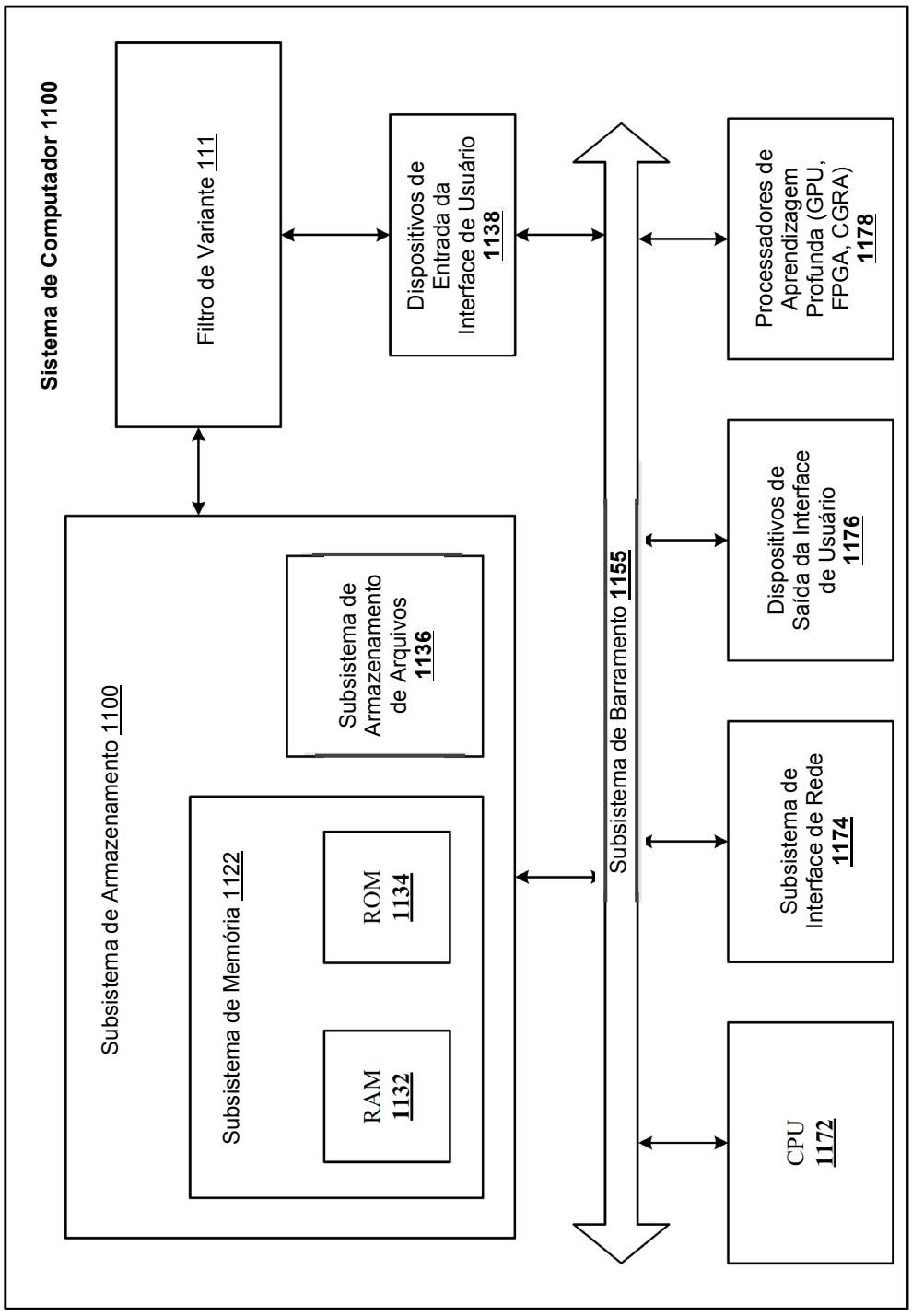

Sistema de Computador 1100
Filtro de Variante 111
Dispositivos de Entrada da Interface de Usuário 1138
Processadores de Aprendizagem Profunda (GPU, FPGA, CGRA) 1178
Subsistema de Armazenamento 1100
Subsistema de Armazenamento de Arquivos 1136
Subsistema de Memória 1122
ROM 1134
RAM 1132
Subsistema de Barramento 1155
Dispositivos de Saída da Interface de Usuário 1176
Subsistema de Interface de Rede 1174
CPU 1172

FIG. 11

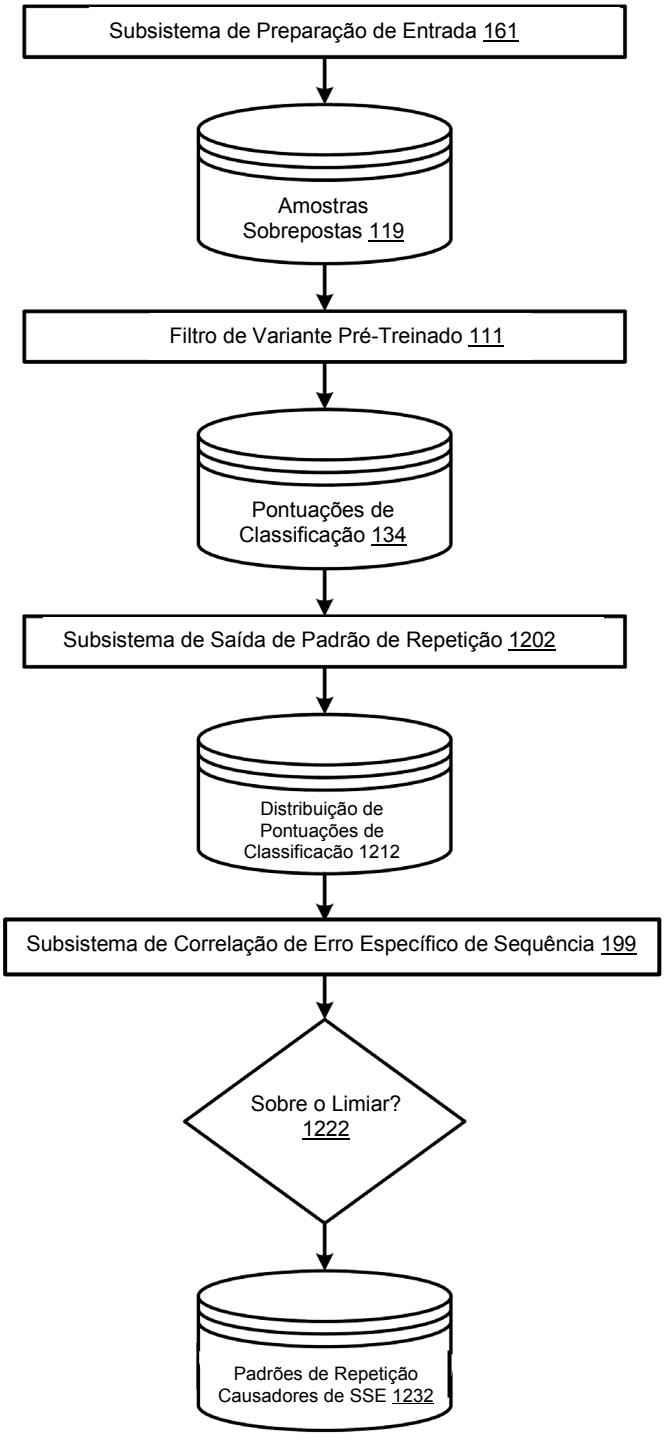

Subsistema de Preparação de Entrada 161
Amostras Sobrepostas 119
Filtro de Variante Pré-Treinado 111
Pontuações de Classificação 134
Subsistema de Saída de Padrão de Repetição 1202
Distribuição de Pontuações de Classificacão 1212
Subsistema de Correlação de Erro Específico de Sequência 199
Sobre o Limiar? 1222
Padrões de Repetição Causadores de SSE 1232

FIG. 12

RESUMO

SISTEMA PARA IDENTIFICAR PADRÕES DE REPETIÇÃO QUE CAUSAM ERROS ESPECÍFICOS DE SEQUÊNCIA, MÉTODO IMPLEMENTADO POR COMPUTADOR E MEIO DE ARMAZENAMENTO LEGÍVEL POR COMPUTADOR NÃO TRANSITÓRIO

A tecnologia divulgada apresenta uma estrutura baseada em aprendizado profundo, que identifica padrões de sequência que causam erros específicos de sequência (SSEs). Os sistemas e métodos treinam um filtro de variantes em dados de variante em larga escala para aprender dependências causais entre padrões de sequência e chamadas de variante falsa. O filtro variante possui uma estrutura hierárquica construída em redes neurais profundas, como redes neurais convolucionais e redes neurais totalmente conectadas. Os sistemas e métodos implementam uma simulação que usa o filtro de variante para testar padrões de sequência conhecidos quanto a seus efeitos na filtragem de variante. A premissa da simulação é a seguinte: quando um par de um padrão de repetição em teste e uma chamada variante são alimentados para o filtro de variante como parte de uma sequência de entrada simulada e o filtro de variante classifica a chamada variante como uma chamada de variante falsa, então o padrão de repetição é considerado como causador da chamada de variante falsa e identificado como causador de SSE.