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

(21) BR 112019009949-2 A2

(22) Data do Depósito: 14/11/2017

(43) Data da Publicação Nacional: 24/05/2018

(54) Título: MÉTODO IMPLANTADO POR COMPUTADOR PARA VALIDAR CHAMADAS DE VARIANTE E SISTEMA PARA VALIDAR CHAMADAS DE VARIANTE

(51) Int. Cl.: G06F 19/18.

(30) Prioridade Unionista: 16/11/2016 US 62/422,841; 17/01/2017 US 62/447,076.

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

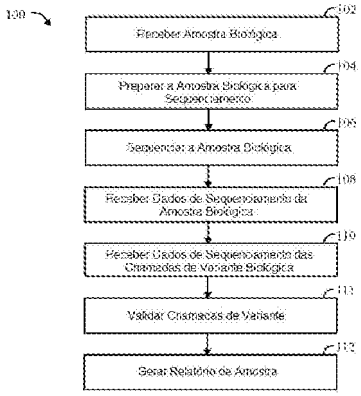
(72) Inventor(es): XIAO CHEN; CHEN ZHAO; JESSICA GORDON; SHILE ZHANG; TINGTING JIANG; GWENN BERRY; YESHA SHAH; HAN-YU CHUANG.

(86) Pedido PCT: PCT US2017061554 de 14/11/2017

(87) Publicação PCT: WO 2018/093780 de 24/05/2018

(85) Data da Fase Nacional: 15/05/2019

(57) Resumo: Trata-se de métodos e sistemas para validar chamadas de variante. Os dados de sequenciamento são recebidos para uma leitura simples junto com a sequência genômica de interesse. Uma indicação é recebida de uma chamada de variante potencial em uma posição designada dentro da sequência de nucleotídeos junto com a sequência genômica de interesse. Os métodos e sistemas obtêm frequências de variante de linha de base na posição designada dentro de uma ou mais sequências genômicas de linha de base, determinam uma frequência de variante de amostra na posição designada para a sequência genômica de interesse, analisam as frequências de variante de amostra e linha de base na posição designada para obter uma classificação de qualidade, e validam a chamada de variante potencial para a sequência genômica de interesse com base na classificação de qualidade.



Relatório Descritivo da Patente de Invenção para **"MÉTODO IMPLANTADO POR COMPUTADOR PARA VALIDAR CHAMADAS DE VARIANTE E SISTEMA PARA VALIDAR CHAMADAS DE VARIANTE"**

PEDIDO RELACIONADO

[001] O presente pedido refere-se a e reivindica o benefício de prioridade do Pedido de Patente Provisório nº U.S. 62/447.076, intitulado "Validation Methods and Systems for Sequence Variant Calls", depositado em 17 de janeiro de 2017 e Pedido de Patente Provisório nº U.S. 62/422.841, intitulado "Methods and Systems to Improve Accuracy in Variant Calling", depositado em 16 de novembro de 2016, cuja matéria completa é incorporada integralmente a título de referência.

ANTECEDENTES

[002] A matéria no presente documento refere-se, em geral, a sistemas e métodos para analisar operações de sequenciamento e, mais particularmente, para validar chamadas de variante feitas em conexão com operações de sequenciamento.

[003] Atualmente, técnicas de tratamento de câncer estão se afastando de uma abordagem do tipo um fármaco para tudo. Abordagens de diagnóstico abrangentes suscitam desafios para métodos que detectam variantes de DNA. Exemplos não limitantes de métodos de detecção de DNA incluem sequenciamento de Sanger, imuno-histoquímica (IHC), hibridização in situ por fluorescência (FISH) e PCR quantitativa (qPCR). Alguns métodos de detecção analisam o DNA para uma ou poucas mudanças genéticas em um tempo. Entretanto, como o número de DNA alterações conhecidas por impulsionar a progressão de câncer continua a crescer, novos métodos são propostos para analisar DNA para mais e mais mudanças genéticas durante uma operação de sequenciamento.

[004] Sequenciamento de próxima geração (NGS) oferece uma capacidade para avaliar variantes em múltiplos genes com o uso de uma amostra. Para este fim, foi proposto um sistema que oferece um teste (ensaio) híbrido de NGS com base em captura que visa múltiplos genes de câncer (por exemplo, acima de 100 genes de câncer) para sequenciamento nas plataformas NextSeq™ e HiSeq™. O teste de NGS

inclui um DNA fluxo de trabalho para a identificação de variantes de nucleotídeo únicas (SNVs), pequenas inserções e deleções (indels), variantes de nucleotídeo múltiplas (MNVs), amplificações de gene (CNVs). O teste de NGS também inclui um RNA fluxo de trabalho para a identificação de variantes de junção e fusões de gene. Uma variante é identificada quando uma sequência de ácido nucléico de amostra é determinada como diferente de uma sequência de referência em uma ou mais posições de par de bases junto com a sequência. Alterações tanto de DNA como RNA de genes diferentes podem ser avaliadas em um ponto comum no tempo enquanto se sequencia uma amostra única.

[005] Estão sendo desenvolvidas técnicas bioinformáticas de análise que entregam chamadas de variante em um tempo de resposta rápido. As técnicas incluem, entre outras coisas, um alinhador e chamador de variante. O processo de alinhamento é capaz de alinhar indels longas para superar os desafios a partir de sequenciamento sequencial com comprimento de leitura curto. Ademais, testes de NGS utilizam várias técnicas de preparação de amostra, um exemplo das quais é fixada em formalina, incorporada em parafina (FFPE). Amostras FFPE apresentam outra dificuldade para detectar variantes de DNA quando o DNA de interesse é de baixa abundância e/ou pode ter degradado.

[006] Entretanto, técnicas de sequenciamento propostas atuais vivenciam falsos positivos em conexão com chamada de variante. Por exemplo, a técnica pode determinar incorretamente que uma variante está presente em uma sequência de amostra em uma localização particular (par de bases) e/ou identificar incorretamente o tipo de variante (em geral, denominado falsos positivos). Erros sistemáticos podem provocar falsos positivos, tal como devido a artefatos de FFPE, erros de sequenciamento ou erros de PCR.

[007] Perdura uma necessidade por métodos e sistemas de sequenciamento que possam simultaneamente interrogar múltiplos tipos de variantes de DNA, sem declarar um número indevidamente grande de falsos positivos, em uma amostra única de uma maneira simples e econômica.

DEFINIÇÕES

[008] Toda a literatura e material similar citado nesse pedido, que inclui, porém, sem limitações, patentes, pedidos de patente, artigos, livros, tratados e páginas de web, independentemente do formato de tal literatura e materiais similares, são expressamente incorporados a título de referência em sua totalidade. No caso de uma ou mais dentre a literatura e materiais similares incorporados diferir desse pedido, ou contradizê-lo, incluindo, porém, sem limitações, termos definidos, uso de termo, técnicas descritas, ou similares, esse pedido controla.

[009] Como usado no presente documento, os termos a seguir têm os significados indicados.

[010] O termo "cromossoma" se refere ao portador de gene de hereditariedade de uma célula viva, que é derivado de cadeias de cromatina que compreendem DNA e componentes de proteína (especialmente histonas). O sistema de numeração de cromossoma de genoma de indivíduo humano reconhecido internacionalmente convencional é empregado no presente documento.

[011] O termo "site" se refere a uma posição única (por exemplo, ID de cromossoma, posição e orientação de cromossoma) em um genoma de referência. Em algumas modalidades, um site pode ser um resíduo, um marcador de sequência, ou uma posição do segmento em uma sequência. O termo "local" pode ser usado para se referir à localização específica de uma sequência de ácido nucléico ou polimorfismo em um cromossoma de referência.

[012] O termo "amostra" no presente documento se refere a uma amostra, tipicamente derivada de um fluido, célula, tecido, órgão ou organismo biológico que contém um ácido nucléico ou uma mistura de ácidos nucléicos que contém pelo menos uma sequência de ácido nucléico que deve ser sequenciada e/ou faseada. Tais amostras incluem, porém, sem limitações, expectoração/fluido oral, fluido amniótico, sangue, uma fração de sangue, amostras de biópsia com agulha fina (por exemplo, biópsia cirúrgica, biópsia com agulha fina, etc.), urina, fluido peritoneal, fluido pleural, explante de tecido, cultura de órgão e qualquer outro tecido ou preparação de célula,

ou fração ou derivativo dos mesmos ou isolado a partir dos mesmos. Embora a amostra seja frequentemente tirada de um sujeito humano (por exemplo, paciente), amostras podem ser tiradas de qualquer organismo que tenha cromossomas, que inclui, porém, sem limitações, cães, gatos, cavalos, caprinos, ovelha, gado, porcos, etc. A amostra pode ser usada diretamente como obtida a partir da fonte biológica ou seguindo um pré-tratamento para modificar o caráter da amostra. Por exemplo, tal pré-tratamento pode incluir preparar plasma de sangue, diluir fluidos viscosos e assim por diante. Métodos de pré-tratamento também podem envolver, porém, sem limitações, filtragem, precipitação, diluição, destilação, mistura, centrifugação, congelamento, liofilização, concentração, amplificação, fragmentação de ácido nucléico, inativação de componentes interferentes, a adição de reagentes, lise, etc.

[013] O termo "sequência" inclui ou representa uma cadeia de nucleotídeos acoplados entre si. Os nucleotídeos podem ter como base DNA ou RNA. Deve ser compreendido que uma sequência pode incluir múltiplas subsequências. Por exemplo, uma única sequência (por exemplo, de um fragmento amplificado por PCR) pode ter 350 nucleotídeos. A leitura de amostra pode incluir múltiplas subsequências dentro desses 350 nucleotídeos. Por exemplo, a leitura de amostra pode incluir primeira e segunda subsequências de acompanhamento que têm, por exemplo, 20 a 50 nucleotídeos. A primeira e segunda subsequências de acompanhamento podem estar localizadas em qualquer lado de um segmento repetitivo que tem uma subsequência correspondente (por exemplo, 40 a 100 nucleotídeos). Cada uma das subsequências de acompanhamento pode incluir (ou incluir porções de) uma subsequência de iniciador (por exemplo, 10 a 30 nucleotídeos). Para facilidade de leitura, o termo "subsequência" será denominado "sequência", mas é compreendido que duas sequências não são necessariamente separadas uma da outra em uma cadeia comum. Para diferenciar as várias sequências descritas no presente documento, as sequências podem receber identificadores diferentes (por exemplo, sequência alvo, sequência de iniciador, sequência de acompanhamento, sequência de referência, e similares). Outros termos, tais como "alelo", podem receber diferentes identificadores

para diferenciar entre objetos semelhantes.

[014] O termo “sequenciamento de extremidade pareada” se refere a métodos de sequenciamento que sequenciam ambas as extremidades de um fragmento alvo. Sequenciamento de extremidade pareada pode facilitar a detecção de rearranjos genômicos e segmentos repetitivos, bem como fusões de gene e transcrições inovadoras. Metodologia para sequenciamento de extremidade pareada é descrita na Publicação PCT WO07010252, o Pedido PCT nº serial PCTGB2007/003798 e Publicação de Pedido de Patente no U.S. 2009/0088327, cada um dos quais é incorporado ao presente documento a título de referência. Em um exemplo, uma série de operações pode ser realizada como segue; (a) gerar agrupamentos de ácidos nucléicos; (b) linearizar os ácidos nucléicos; (c) hibridizar um primeiro iniciador de sequenciamento e realizar ciclos repetidos de extensão, varredura e desbloqueio, como apresentado acima; (d) “inverter" os ácidos nucléicos alvos na superfície de célula de fluxo sintetizando-se uma cópia complementar; (e) linearizar a cadeia ressintetizada; e (f) hibridizar um segundo iniciador de sequenciamento e realizar ciclos repetidos de extensão, varredura e desbloqueio, como apresentado acima. A operação de inversão pode ser realizada entregando-se reagentes como apresentado acima para um ciclo único de amplificação de ponte.

[015] O termo "genoma de referência" ou "sequência de referência" se refere a qualquer sequência de genoma conhecida particular, seja parcial ou completa, de qualquer organismo que pode ser usado para referenciar sequências identificadas e um sujeito. Por exemplo, um genoma de referência usado para sujeitos humanos bem como muitos outros organismos é encontrado no Centro Nacional para Informações de Biotecnologia em ncbi.nlm.nih.gov. Um "genoma" se refere às informações genéticas completas de um organismo ou vírus, expressas em sequências de ácido nucléico. Um genoma inclui tanto os genes quanto as sequências sem codificação do DNA. A sequência de referência pode ser maior do que as leituras que são alinhadas à mesma. Por exemplo, a mesma pode ser pelo menos cerca de 100 vezes maior, ou pelo menos cerca de 1.000 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 10^7 vezes maior. Em um exemplo, o genoma de referência sequência é aquele de um genoma humano de comprimento total. Em outro exemplo, o genoma de referência sequência é limitado a um cromossoma humano específico tal como cromossoma 13. Em algumas modalidades, um cromossoma de referência é uma sequência de cromossoma da versão de genoma humano hg19. Tais sequências podem ser denominadas sequências de referência de cromossomas, embora o termo genoma de referência seja destinado a cobrir tais sequências. Outros exemplos de sequências de referência incluem genomas de outras espécies, bem como cromossomas, regiões subcromossômicas (tais como cadeias), etc., de quaisquer espécies. Em várias modalidades, o genoma de referência é uma sequência de consenso ou outra combinação derivada de múltiplos indivíduos. Entretanto, em certas aplicações, a sequência de referência pode ser tomada a partir de um indivíduo particular.

[016] O termo “ler” se refere a uma coleção de dados de sequência que descreve um fragmento de uma amostra ou referência de nucleotídeo. O termo “leitura” pode se referir a uma leitura de amostra e/ou uma leitura de referência. Típica, embora não necessariamente, uma leitura representa uma sequência curta de pares de bases contíguos na amostra ou referência. A leitura pode ser representada simbolicamente pelo par de bases sequência (em ATCG) do fragmento de amostra ou referência. A mesma pode ser armazenada em um dispositivo de memória e processada como apropriado para determinar se a leitura confirma uma sequência de referência ou atende outros critérios. Uma leitura pode ser obtida diretamente a partir 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 cromossoma ou região genômica ou gene.

[017] Métodos de sequenciamento de próxima geração incluem, por exemplo,

sequenciamento por tecnologia de síntese (Illumina), pirosequenciamento (454), tecnologia de semicondutor de íons (Ion Torrent sequencing), sequenciamento em tempo real de molécula única (Pacific Biosciences) e sequenciamento por ligadura (SOLiD sequencing). Dependendo dos métodos de sequenciamento, o comprimento de cada leitura pode variar de cerca de 30 bp a mais do que 10.000 bp. Por exemplo, método de sequenciamento Illumina com o uso de sequenciador SOLiD gera leituras de ácido nucléico de cerca de 50 bp. Como outro exemplo, Ion Torrent Sequencing gera leituras de ácido nucléico de até 400 bp e pirosequenciamento 454 gera leituras de ácido nucléico de cerca de 700 bp. Como ainda outro exemplo, métodos de sequenciamento em tempo real de molécula única podem gerar leituras de 10.000 bp a 15.000 bp. Portanto, em certas modalidades, a sequência de leituras de ácido nucléico tem um comprimento de 30 a 100 bp, 50 a 200 bp ou 50 a 400 bp.

[018] Os termos "leitura de amostra", “sequência de amostra” ou “fragmento de amostra” se referem a dados de sequência para uma sequência genômica de interesse de uma amostra. Por exemplo, a leitura de amostra compreende dados de sequência a partir de um fragmento amplificado por PCR que tem uma sequência de iniciador para a frente e inversa. Os dados de sequência podem ser obtidos a partir de qualquer metodologia de sequência selecionada. A leitura de amostra pode ser, por exemplo, a partir de uma reação de sequenciamento por síntese (SBS), uma reação de sequenciamento por ligadura, ou qualquer outra metodologia de sequenciamento adequada pela qual é desejado determinar o comprimento e/ou identidade de um elemento repetitivo. A leitura de amostra pode ser uma sequência consenso (por exemplo, calculada em média ou ponderada) derivada de múltiplas leituras de amostra. Em certas modalidades, fornecer uma sequência de referência compreende identificar um local de interesse com base na sequência de iniciador do fragmento amplificado por PCR.

[019] O termo “fragmento bruto” se refere a dados de sequência para uma porção de uma sequência genômica de interesse que sobrepõe pelo menos parcialmente uma posição designada ou posição secundária de interesse dentro de uma leitura de

amostra ou fragmento de amostra. Exemplos não limitantes de fragmentos brutos incluem um fragmento remendado duplex, um fragmento remendado simples, um fragmento não remendado duplex e um fragmento não remendado simples. O termo “bruto” é usado para indicar que o fragmento bruto inclui dados de sequência que tem alguma relação com os dados de sequência em uma leitura de amostra, independentemente de o fragmento bruto exibir uma variante de apoio que corresponde e autentica ou confirma uma variante potencial em uma leitura de amostra. O termo “fragmento bruto” não indica que o fragmento necessariamente inclui uma variante de apoio 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 carecem de um tipo correspondente de variante “de apoio” que possa de outra forma ser esperado ocorrer dada a variante na leitura de amostra.

[020] Os termos "alinhado", "alinhamento", ou "alinhar" se referem ao processo de comparar uma leitura ou marcador a uma sequência de referência e, desse modo, determinar se a sequência de referência contém a sequência de leitura. Caso a sequência de referência contenha a leitura, a leitura pode ser mapeada para a sequência de referência ou, em certas modalidades, para uma localização particular na sequência de referência. Em alguns casos, alinhamento simplesmente informa se uma leitura é ou não um membro de uma sequência de referência particular (isto é, se a leitura está presente ou ausente na sequência de referência). Por exemplo, o alinhamento de uma leitura para a sequência de referência para cromossoma humano 13 informará se a leitura está presente na sequência de referência para cromossoma 13. Uma ferramenta que fornece essas informações pode ser chamada de um testador de associação a conjunto. Em alguns casos, um alinhamento indica adicionalmente uma localização na sequência de referência para onde a leitura ou marcador mapeia. Por exemplo, se a sequência de referência é a sequência de genoma humano inteira, um alinhamento pode indicar que uma leitura está presente no cromossoma 13, e

pode adicionalmente indicar que a leitura está em um cadeia e/ou site particular do cromossoma 13.

[021] O termo “indel” se refere à inserção e/ou à deleção de bases no DNA de um organismo. Um microindel representa um indel que resulta em uma mudança líquida de 1 a 50 nucleotídeos. Em regiões de codificação do genoma, a menos que o comprimento de um indel seja um múltiplo de 3, o mesmo 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 mudar o número total no DNA. Indels também podem ser contrastados com uma Mutação de Base em Tandem (TBM), que pode ser definida como substituição em nucleotídeos adjacentes (primariamente substituições em dois nucleotídeos adjacentes, mas substituições em três nucleotídeos adjacentes têm sido observadas.

[022] O termo "variante" se refere a uma sequência de ácido nucléico que é diferente de uma referência de ácido nucléico. Variante de sequência de ácido nucléico típica inclui, sem limitação, polimorfismo de nucleotídeo único (SNP), polimorfismos de deleção e inserção curtos (Indel), variação de número de cópia (CNV), marcadores de microssatélite ou repetições curtas em tandem e variação estrutural. Chamada de variante somática é o esforço para identificar variantes presentes em baixa frequência na amostra de DNA. Chamada de variante somática é de interesse no contexto de tratamento de câncer. Câncer é provocado por um acúmulo de mutações no DNA. Uma amostra de DNA a partir de um tumor é, em geral, heterogênea, que inclui algumas células normais, algumas células em um estágio inicial de progressão de câncer (com menos mutações) e algumas células em estágio avançado (com mais mutações). Devido a essa heterogeneidade, quando se sequencia um tumor (por exemplo, a partir de uma amostra FFPE), mutações somáticas frequentemente aparecerão em uma baixa frequência. Por exemplo, um SNV poderia ser visto em apenas 10% das leituras que cobrem uma dada base.

[023] O termo “frequência de variante” representa a frequência relativa de um

alelo (variante de um gene) em um local particular em uma população, expressa como uma fração ou percentual. Por exemplo, a fração ou percentual pode ser a fração de todos os cromossomas na população que porta aquele alelo. A título de exemplo, frequência de variante de amostra representa a frequência relativa de um alelo/variante em um local/posição particular junto com uma sequência genômica de interesse sobre uma “população” que corresponde ao número de leituras e/ou amostras obtidas para a sequência genômica de interesse a partir de um indivíduo. Como outro exemplo, uma frequência de variante de linha de base representa a frequência relativa de um alelo/variante em um local/posição particular junto com uma ou mais sequências genômicas de linha de base em que a “população” corresponde ao número de leituras e/ou amostras obtidas para a uma ou mais sequências genômicas de linha de base a partir de uma população de indivíduos normais.

[024] Os termos “posição”, “posição designada” e “local” se referem a uma localização ou coordenada de um ou mais nucleotídeos dentro de uma sequência de nucleotídeos. Os termos “posição”, “posição designada” e “local” também se referem a uma localização ou coordenada de um ou mais pares de bases em uma sequência de nucleotídeos.

[025] O termo "haplótipo" se refere a uma combinação de alelos em sites adjacentes em um cromossoma que são herdados juntos. Um haplótipo pode ser um local, diversos locais ou um cromossoma inteiro dependendo do número de eventos de recombinação que tenham ocorrido entre um dado conjunto de locais, caso tenha ocorrido algum.

[026] O termo "limiar" no presente documento se refere a um valor numérico ou não numérico que é usado como um ponto de corte para caracterizar uma amostra, um ácido nucléico ou porção dos mesmos (por exemplo, uma leitura). Um limiar pode ser variado com base em análise empírica. O limiar pode ser comparado a um valor medido ou calculado para determinar se a fonte que dá origem a tal sugestão de valor deve ser classificada de uma maneira particular. Valores limiares podem ser identificados empírica ou analiticamente. A escolha de um limiar é dependente do

nível de confiança que o usuário quer ter para fazer a classificação. O limiar pode ser escolhido para um propósito particular (por exemplo, para equilibrar sensibilidade e seletividade). Como usado no presente documento, o termo "limiar" indica um ponto em que um curso de análise pode ser mudado e/ou um ponto em que uma ação pode ser desencadeada. Não é exigido que um limiar seja um número predeterminado. Em vez disso, o limiar pode ser, por exemplo, uma função que tem como base uma pluralidade de fatores. O limiar pode ser adaptativo às circunstâncias. Além disso, um limiar pode indicar um limite superior, um limite inferior ou uma faixa entre limites.

[027] Em algumas modalidades, uma métrica ou classificação, que tem como base os dados de sequenciamento, pode ser comparada ao limiar. Como usados no presente documento, os termos "métrica" ou "classificação" podem incluir valores ou resultados que foram determinados a partir dos dados de sequenciamento ou podem incluir funções que tem como base os valores ou resultados que foram determinados a partir dos dados de sequenciamento. Como um limiar, a métrica ou classificação pode ser adaptativa às circunstâncias. Por exemplo, a métrica ou classificação pode ser um valor normalizado. Como um exemplo de uma classificação ou métrica, uma ou mais modalidades podem usar classificações de contagem quando analisam os dados. Uma classificação de contagem pode ter como base 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 que são usadas para determinar uma classificação de contagem pode ter sido alinhada com uma sequência de referência ou pode ser atribuída como um alelo potencial. O número de leituras de amostra que têm uma característica comum pode ser contado para determinar uma contagem de leitura. Classificações de contagem pode ter como base a contagem de leitura. Em algumas modalidades, a classificação de contagem pode ser um valor que é igual à contagem de leitura. Em outras modalidades, a classificação de contagem pode ter como base a contagem de leitura e outras informações. Por exemplo, uma classificação de contagem pode ter como base a contagem de leitura

para um alelo particular de um local genético e um número total de leituras para o local genético. Em algumas modalidades, a classificação de contagem pode ter como base a contagem de leitura e dados obtidos previamente para o local genético. Em algumas modalidades, as classificações de contagem podem ser classificações normalizadas entre valores predeterminados. A classificação de contagem também pode ser uma função de contagens de leitura de outros locais de uma amostra ou uma função de contagens de leitura de outras amostras que foram realizadas concorrentemente com a amostra-de-interesse. Por exemplo, a classificação de contagem pode ser uma função da contagem de leitura de um alelo particular e as contagens de leitura de outros locais na amostra e/ou as contagens de leitura de outras amostras. Como um exemplo, as contagens de leitura de outros locais e/ou as contagens de leitura de outras amostras podem ser usadas para normalizar a classificação de contagem para o alelo particular.

[028] Os termos “cobertura” ou “cobertura de fragmento” se refere 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. Alternativamente, a cobertura pode ser determinada multiplicando-se a contagem de leitura por um fator designado que tem como base conhecimento histórico, conhecimento da amostra, conhecimento do local, etc.

SUMÁRIO

[029] De acordo com modalidades no presente documento, é fornecido um método implantado por computador para validar chamadas de variante. O método opera sob controle de um ou mais processadores que executam instruções de programa para, receber dados de sequenciamento que incluem uma leitura de amostra que tem uma sequência correspondente de nucleotídeos junto com a sequência genômica de interesse, receber uma indicação de uma chamada de variante potencial em uma posição designada dentro da sequência de nucleotídeos junto com a sequência genômica de interesse, e obter frequências de variante de linha

de base na posição designada dentro de uma ou mais sequências genômicas de linha de base. O método obtém uma frequência de variante de amostra na posição designada para a sequência genômica de interesse. O método analisa as frequências de variante de linha de base e amostra na posição designada para obter uma classificação de qualidade; e valida a chamada de variante potencial para a sequência genômica de interesse com base na classificação de qualidade.

[030] Opcionalmente, a operação de análise inclui obter uma relação entre a frequência de variante de amostra e uma distribuição das frequências de variante de linha de base, a classificação de qualidade com base na relação. Opcionalmente, a operação de análise compreende indexar a frequência de variante de amostra em relação a uma distribuição das frequências de variante de linha de base. A relação pode ser com base em um teste de soma de classificação de Wilcoxon não paramétrico. As frequências de variante de linha de base indicam um grau de ruído de fundo em posições correspondentes junto com a sequência genômica de linha de base.

[031] Opcionalmente, a validação compreende adicionalmente comparar a classificação de qualidade a um limiar; e declarar a chamada de variante potencial como sendo uma chamada de variante válida quando a classificação de qualidade excede o limiar. As frequências de variante de linha de base podem ser derivadas de múltiplas sequências genômicas de linha de base que são associadas com mais do que um tipo de alelo. Opcionalmente, o método compreende adicionalmente receber dados de sequenciamento que incluem uma pluralidade de leituras de referência de uma sequência de nucleotídeos junto com a sequência genômica de linha de base, e determinar as frequências de variante de linha de base para as leituras de referência nas posições designadas. A determinação das frequências de variante de linha de base pode compreender adicionalmente receber os dados de sequenciamento das leituras de referência para um conjunto de posições dentro de uma janela de par de bases atual; identificar uma frequência de variante candidata para uma ou mais posições no conjunto de posições dentro da janela de par de bases atual; selecionar

uma das frequências de variante candidatas como a frequência de variante de linha de base para uma posição designada dentro da leitura de referência; e deslocar a janela de par de bases junto com a sequência genômica de linha de base e repetir as operações

[032] De acordo com uma modalidade, é fornecido um método implantado por computador para validar chamadas de variante. As operações do método sob controle de um ou mais processadores que executam instruções de programa para, receber uma indicação de uma chamada de variante potencial para uma sequência genômica de interesse; e receber dados de sequenciamento que incluem leituras para fragmentos brutos e de amostra de nucleotídeos junto com a sequência genômica de interesse, sendo que as leituras incluem leituras de amostra para o fragmento de amostra que corresponde a uma sequência de nucleotídeos na posição designada junto com a sequência genômica de interesse. O método analisa uma frequência de variante de amostra na posição designada para a sequência genômica de interesse em relação a frequências de variante de linha de base na posição designada para uma sequência genômica de linha de base para obter uma classificação de qualidade; e determinar pelo menos um dentre os seguintes: A) se os fragmentos brutos confirmam a chamada de variante potencial; B) se as leituras de amostra fornecem uma quantidade predeterminada de cobertura para o fragmento de amostra; e C) se a chamada de variante potencial confirma uma chamada de variante anterior exibida através de uma população predefinida. O método valida a chamada de variante potencial com base na classificação de qualidade e na operação de determinação.

[033] Opcionalmente, a determinação pode incluir operação A) e operação A) compreender: identificar os fragmentos brutos que indicam uma chamada de variante de apoio; obter uma classificação de fragmento ponderada para os fragmentos brutos que indicam a chamada de variante de apoio; e comparar a classificação de fragmento ponderada a um limiar de fragmento bruto para determinar se confirma a chamada de variante potencial.

[034] Opcionalmente, a determinação pode incluir operação A), em que os

fragmentos brutos correspondem a pelo menos um dentre um fragmento remendado duplex, um fragmento remendado simples, um fragmento não remendado duplex ou um fragmento não remendado simples. Opcionalmente, a determinação pode incluir operação A) e operação A) pode compreender: identificar um fragmento remendado duplex como o fragmento bruto; e determinar se o fragmento remendado duplex indica uma chamada de variante de apoio para determinar se confirma a chamada de variante potencial. Opcionalmente, a operação de determinação pode incluir operação B), em que a chamada de variante potencial é declarada uma Sem chamada quando as leituras de amostra caem abaixo da quantidade predeterminada de cobertura. Opcionalmente, a operação de determinação pode incluir operação C). Quando uma correlação é determinada entre as chamadas de variante potencial e anterior, a operação de determinação realiza pelo menos uma dentre operação A) ou operação B) com a utilização de um primeiro limiar. Quando uma correlação não existe entre as chamadas de variante potencial e anterior, a operação de determinação realiza pelo menos uma dentre operação A) ou operação B) com a utilização de um segundo limiar.

[035] Opcionalmente, o método pode realizar a operação de análise e determinar operações a), B) e C) em uma hierarquia.

[036] De acordo com modalidades no presente documento, é fornecido um sistema para validar chamadas de variante. O sistema compreende memória que armazena instruções de programa e dados de sequenciamento que incluem uma leitura de amostra que tem uma sequência correspondente de nucleotídeos junto com a sequência genômica de interesse e um ou mais processadores que executam as instruções de programa. Os processadores recebem uma indicação de uma chamada de variante potencial em uma posição designada dentro da sequência de nucleotídeos junto com a sequência genômica de interesse; e obtêm frequências de variante de linha de base na posição designada dentro de uma ou mais sequências genômicas de linha de base. Os processadores determinam uma frequência de variante de amostra na posição designada para a sequência genômica de interesse; e analisam as frequências de variante de linha de base e amostra na posição designada para

obter uma classificação de qualidade. Os processadores validam a chamada de variante potencial para a sequência genômica de interesse com base na classificação de qualidade.

[037] Opcionalmente, os processadores podem obter uma relação entre a frequência de variante de amostra e uma distribuição das frequências de variante de linha de base, a classificação de qualidade com base na relação. Opcionalmente, os processadores indexam a frequência de variante de amostra em relação a uma distribuição das frequências de variante de linha de base. Opcionalmente, as frequências de variante de linha de base indicam um grau de ruído de fundo em um local correspondente junto com a sequência genômica de linha de base. Opcionalmente, os processadores comparam a classificação de qualidade a um limiar; e emitem uma indicação que declara a chamada de variante potencial como sendo inválida com base na comparação. Opcionalmente, o sistema compreende adicionalmente memória que armazena dados de sequenciamento que incluem uma pluralidade de leituras de referência de nucleotídeos junto com a sequência genômica de linha de base, o um ou mais processadores para determinar as frequências de variante de linha de base para as leituras de referência na posição designada.

[038] De acordo com modalidades no presente documento, é fornecido um sistema para validar chamadas de variante. O sistema compreende memória que armazena instruções de programa e dados de sequenciamento que incluem leituras para fragmentos brutos e de amostra de nucleotídeos junto com a sequência genômica de interesse, sendo que as leituras incluem leituras de amostra para o fragmento de amostra que correspondem a uma sequência de nucleotídeos na posição designada junto com a sequência genômica de interesse; e um ou mais processadores que executam as instruções de programa. Os processadores recebem uma indicação de uma chamada de variante potencial para uma sequência genômica de interesse; e analisam uma frequência de variante de amostra na posição designada para a sequência genômica de interesse em relação a frequências de variante de linha de base na posição designada para uma sequência genômica de linha de base para

obter uma classificação de qualidade. Os processadores determinam pelo menos um dentre os seguintes: A) se os fragmentos brutos confirmam a chamada de variante potencial; B) se as leituras de amostra fornecem uma quantidade predeterminada de cobertura para o fragmento de amostra; e C) se a chamada de variante potencial confirma uma chamada de variante anterior exibida através de uma população predefinida. Os processadores validam a chamada de variante potencial com base na classificação de qualidade e na operação de determinação.

[039] Opcionalmente, os processadores podem determinar se os fragmentos brutos confirmam a chamada de variante potencial identificando-se os fragmentos brutos que indicam uma chamada de variante de apoio; obter uma classificação de fragmento ponderada para os fragmentos brutos que indicam a chamada de variante de apoio; e comparar a classificação de fragmento ponderada a um limiar de fragmento bruto para determinar se confirma a chamada de variante potencial. Os fragmentos brutos podem corresponder a pelo menos um dentre um fragmento remendado duplex, um fragmento remendado simples, um fragmento não remendado duplex ou um fragmento não remendado simples.

[040] De acordo com modalidades no presente documento, são descritos sistemas e métodos para reduzir chamada de variante de falso positivo a partir de erros sistemáticos. Erros sistemáticos podem surgir devido a vários fatores tais como artefatos de FFPE, erros de sequenciamento, erros de preparação de biblioteca, erros de PCR e similares. Chamadas de variante são submetidas estaticamente a uma distribuição de erros antecedentes específicos do local que podem ser compilados partir de um painel de amostras normais FFPE com qualidade de DNA variada de vários tecidos sequenciadas pelo ensaio à base de NGS. Os mesmos dados de sequenciamento das amostras normais FFPE também podem ser utilizados para normalizar tendência sistemática em cobertura de leitura provocada por PCR, qualidade de DNA, eficiência de tração da sonda, ou conteúdo de GC de sequência para revelar tais alterações de número de cópias reais em uma amostra de teste. Para aumentar adicionalmente a razão sinal/ruído em chamada de CNV, sondas

melhoradoras adicionais podem ser adicionadas na captura híbrida para fornecer estimativa robusta de amplificação de gene. De acordo com modalidades no presente documento, métodos e sistemas são descritos que endereçam problemas de ruído e impedem que erros sistemáticos contribuam para chamadas de variante falso positivas. Em conexão com isso, um conjunto de amostras normais é usado para identificar tendência sistemática a fim de que o sistema aumente o rigor de chamada em amostras de tumor em regiões com alto ruído de fundo. Para amostras FFPE, amostras FFPE normais podem ser usadas para construir a linha de base. Para amostras ctDNA, dados de DNA genômico normal podem ser usados para construir a linha de base. Os problemas de abordagens convencionais que são descritos no presente documento se tornam particularmente problemáticos em amostras que exibem variância de frequência baixa, tal como quando se tenta fazer chamada de variante de uma amostra de tumor.

[041] De acordo com modalidades no presente documento, um método e sistema são fornecidos para processamento posterior de variante somática que endereça um problema de reduzir chamadas de variante somática falso positiva. O método e sistema de construir uma linha de base de frequência de variante a partir de amostras normais para ajustar confiança de chamada de variante em regiões genômicas com níveis de ruído de fundo diferentes. Separar sinal de ruído é um desafio tanto para detectar variantes somáticas em dados de FFPE com ruído e para descoberta de variante de tumor em frequência ultrabaixa em dados de DNA livre de célula. Falsos positivos sistemáticos são encontrados comumente em chamada de variante pequena (Figura 1). Falsos positivos sistemáticos podem ser introduzidos durante etapas experimentais que incluem preparação e sequenciamento de biblioteca. Além disso, há regiões no genoma humano que são sensíveis a erros de alinhamento, tais como regiões repetitivas ou regiões com complexidade de sequência baixa.

BREVES DESCRIÇÕES DOS DESENHOS

[042] O presente pedido descreve vários métodos e sistemas para realizar os métodos. Pelo menos alguns dos métodos são ilustrados nas Figuras como uma

pluralidade de operações. Entretanto, deve ser compreendido que modalidades não são limitadas às operações ilustradas nas Figuras. Operações podem ser omitidas, operações podem ser modificadas e/ou outras operações podem ser adicionadas. Além disso, operações descritas no presente documento podem ser combinadas, operações podem ser realizadas simultaneamente, operações podem ser realizadas concorrentemente, operações podem ser divididas em múltiplas suboperações, operações podem ser realizadas em uma ordem diferente, ou operações (ou uma série de operações) podem ser realizadas novamente em uma forma iterativa. Além disso, embora métodos diferentes sejam apresentados no presente documento, deve ser compreendido que os métodos diferentes (ou operações dos métodos diferentes) podem ser combinados em outras modalidades.

[043] A Figura 1A ilustra um método para analisar amostras de acordo com uma modalidade no presente documento.

[044] A Figura 1B ilustra um fluxograma para um método analisar dados de sequenciamento para identificar chamadas de variante potenciais de acordo com uma modalidade no presente documento.

[045] A Figura 1C ilustra um método implantado por computador para validar chamadas de variante de acordo com modalidades no presente documento.

[046] A Figura 1D ilustra um processo para obter frequências de variante de linha de base de acordo com modalidades no presente documento.

[047] A Figura 2A ilustra painéis exemplificativos para indivíduos diferentes dentro de uma população de linha de base utilizada de acordo com modalidades no presente documento.

[048] A Figura 2B ilustra um painel de par de bases que representa uma ampliação para uma porção pequena de um dos painéis da Figura 2A.

[049] A Figura 3 ilustra um gráfico em conexão com o princípio de correção de fundo com a utilização de sequências genômicas de linha de base normais de acordo com modalidades no presente documento.

[050] A Figura 4 ilustra uma árvore de decisão com base em hierarquia para filtrar

chamadas de variante implantadas de acordo com modalidades no presente documento.

[051] A Figura 5A ilustra uma árvore de decisão com base em hierarquia para filtrar chamadas de variante implantadas de acordo com modalidades alternativas no presente documento.

[052] A Figura 5B ilustra um processo realizado em conexão com a utilização de uma classificação ponderada de acordo com modalidades no presente documento.

[053] A Figura 6 ilustra um diagrama de blocos de um ambiente de sistema de computação formado de acordo com modalidades no presente documento.

[054] A Figura 7 ilustra um diagrama de blocos de um sistema alternativo formado de acordo com modalidades no presente documento.

DESCRIÇÃO DETALHADA

[055] A descrição detalhada de várias modalidades será mais bem compreendida quando lida em conjunto com os desenhos anexos. Na medida em que as Figuras ilustram diagramas dos blocos funcionais das várias modalidades, os blocos funcionais não são necessariamente indicativos da divisão entre conjunto de circuitos de hardware. Assim, por exemplo, um ou mais dos blocos funcionais (por exemplo, módulos, processadores ou memórias) podem ser implantados em uma peça única de hardware (por exemplo, um processador de sinal de propósito geral ou um bloco de memória de acesso aleatório, disco rígido, ou similares) ou múltiplas peças de hardware. De maneira similar, os programas podem ser programas autônomos, podem ser incorporados como sub-rotinas em um sistema operacional, podem ser funções em um pacote de software instalado, e similares. Deve ser compreendido que as várias modalidades não são limitadas às disposições e instrumentalidade mostradas nos desenhos.

[056] Modalidades apresentadas no presente documento podem ser aplicáveis para analisar sequências de ácido nucléico para identificar variações de sequência. Modalidades podem ser usadas para analisar variantes/alelos potenciais de uma posição/local genético e determinar um genótipo do local genético ou, em outras

palavras, fornecer uma chamada de genótipo para o local. A título de exemplo, sequências de ácido nucléico podem ser analisadas de acordo com os métodos e sistemas descritas na Publicação de Pedido de Patente nº U.S. 2016/0085910 e Publicação de Pedido de Patente nº U.S. 2013/0296175, cuja matéria completa é expressamente incorporada em sua totalidade ao presente documento a título de referência.

[057] O método descrito no presente documento pode ser implantado por um ou mais processadores de vários sistemas e dispositivos que executam instruções de programa armazenadas em memória como descrito no presente documento. A menos que especificado de outra forma, a ordem das operações nos métodos no presente documento pode ser variada. Além disso, uma ou mais das operações nos métodos descritos no presente documento podem ser omitidas inteiramente, complementadas com operações adicionais.

PROCESSO DE SEQUENCIAMENTO

[058] A Figura 1A ilustra um método 100 de acordo com uma modalidade. O método 100 inclui receber, em 102, uma amostra que inclui ou é suspeita de incluir ácidos nucléicos, tais como DNA. A amostra pode ser de uma fonte conhecida ou desconhecida, tal como um animal (por exemplo, ser humano), planta, bactéria ou fungo. A amostra pode ser tomada diretamente da fonte. Por exemplo, sangue ou saliva podem ser tomados diretamente de um indivíduo. Alternativamente, a amostra pode não ser obtida diretamente da fonte. Em 104, um ou mais processadores dirigem o sistema para preparar a amostra para sequenciamento. A preparação 104 pode incluir remover material estranho e/ou isolar certo material (por exemplo, DNA). A amostra biológica pode ser preparada para incluir recursos para um ensaio particular. Por exemplo, a amostra biológica pode ser preparada para sequenciamento por síntese (SBS). Em certas modalidades, a preparação pode incluir amplificação de certas regiões de um genoma. Por exemplo, a preparação, em 104, pode incluir amplificar locais genéticos predeterminados que são conhecidos por incluírem STRs e/ou SNPs. Os locais genéticos podem ser amplificados com o uso de sequências de

iniciador predeterminadas.

[059] Em 106, o um ou mais processadores dirigem o sistema para sequenciar a amostra. O sequenciamento pode ser realizado através de uma variedade de protocolos de sequenciamento conhecidos. Em modalidades particulares, o sequenciamento inclui SBS. Em SBS, uma pluralidade de nucleotídeos identificados de modo fluorescente é usada para sequenciar uma pluralidade de agrupamentos de DNA amplificado (possivelmente milhões de agrupamentos) presente na superfície de um substrato óptico (por exemplo, uma superfície que define pelo menos parcialmente um canal em uma célula de fluxo). As células de fluxo podem conter amostras de ácido nucléico para sequenciamento em que as células de fluxo são colocadas dentro dos portadores de célula de fluxo apropriados.

[060] Os ácidos nucléicos podem ser preparados de modo que os mesmos compreendam uma sequência de iniciador conhecida que seja adjacente a uma sequência alvo desconhecida. Para iniciar o primeiro ciclo de sequenciamento SBS, um ou mais nucleotídeos identificados de forma diferente, e polimerase de DNA, etc., podem ser fluídos para/através da célula de fluxo por um subsistema de fluxo de fluido (não mostrado). Um tipo único de nucleotídeo pode ser adicionado em um tempo, ou os nucleotídeos usados no procedimento de sequenciamento podem ser projetados especialmente para possuir uma propriedade de terminação reversível permitindo, desse modo, que cada ciclo da reação de sequenciamento ocorra simultaneamente na presença de diversos tipos de nucleotídeos identificados (por exemplo, A, C, T, G). Os nucleotídeos podem incluir identificação porções químicas de identificação detectáveis tais como fluoróforos. Onde os quatro nucleotídeos são misturados, a polimerase é capaz de selecionar a base correta para incorporar e cada sequência é estendida por uma única base. Nucleotídeos não incorporados podem ser lavados escoando-se uma solução de lavagem através da célula de fluxo. Um ou mais lasers podem excitar os ácidos nucléicos e induzir fluorescência. A fluorescência emitida dos ácidos nucléicos tem como base os fluoróforos da base incorporada, e fluoróforos diferentes podem emitir comprimentos de onda diferentes de luz de emissão. Um

reagente de desbloqueio pode ser adicionado à célula de fluxo para remover grupos terminador reversíveis das cadeias de DNA que foram estendidas e detectadas. O reagente de desbloqueio pode, então, ser lavado escoando-se uma solução de lavagem através da célula de fluxo. A célula de fluxo está, então, pronta para um ciclo adicional de sequenciamento que inicia com a introdução de um nucleotídeo identificado como apresentado acima. As operações fluídica e de detecção podem ser repetidas diversas vezes para completar uma execução de sequenciamento. Métodos de sequenciamento exemplificativos são descritos, por exemplo, em Bentley et al., Nature 456:53 a 59 (2008), Publicação Internacional nº WO 04/018497; Patente nº U.S. 7.057.026; Publicação Internacional no WO 91/06678; Publicação Internacional no WO 07/123744; Patente nº U.S. 7.329.492, a Patente nº U.S. 7.211.414, a Patente nº U.S. 7.315.019, a Patente nº U.S. 7.405.281, e Publicação nº U.S. 2008/0108082, cada uma das quais é incorporada ao presente documento a título de referência.

[061] Em algumas modalidades, ácidos nucléicos podem ser fixados a uma superfície e amplificados anteriormente a ou durante o sequenciamento. Por exemplo, a amplificação pode ser realizada com o uso de amplificação de ponte para formar agrupamentos de ácido nucléico em uma superfície. Métodos de amplificação de ponte úteis são descritos, por exemplo, na Patente nº U.S. 5.641.658; Publicação de Patente nº U.S. 2002/0055100; Patente nº U.S. 7.115.400; Publicação de Patente nº U.S. 2004/0096853; Publicação de Patente nº U.S. 2004/0002090; Publicação de Patente nº U.S. 2007/0128624 e Publicação de Patente nº U.S. 2008/0009420, cada uma das quais é incorporada integralmente ao presente documento a título de referência. Outro método útil para amplificar ácidos nucléicos em uma superfície é amplificação em círculo rolante (RCA), por exemplo, como descrito em Lizardi et al., Nat. Genet. 19:225 a 232 (1998) e Publicação de Patente nº U.S. 2007/0099208 A1, cada uma das quais é incorporada ao presente documento a título de referência.

[062] Um protocolo de SBS exemplificativo explora nucleotídeos modificados que têm blocos 3' removíveis, por exemplo, como descrito na Publicação Internacional nº WO 04/018497, Publicação de Patente nº U.S. 2007/0166705A1, e Patente nº U.S.

7.057.026, cada uma das quais é incorporada ao presente documento a título de referência. Por exemplo, ciclos repetidos de reagentes de SBS podem ser entregues para uma célula de fluxo que tem ácidos nucléicos alvos fixados à mesma, por exemplo, como um resultado do protocolo de amplificação de ponte. Os agrupamentos de ácido nucléico podem ser convertidos para forma de cadeia única com o uso de uma solução de linearização. A solução de linearização pode conter, por exemplo, uma restrição de endonuclease capaz de clivagem de uma cadeia de cada agrupamento. Outros métodos de clivagem podem ser usados como uma alternativa à enzimas de restrição ou enzimas de corte, que incluem, entre outras coisas, clivagem química (por exemplo, clivagem de um ligação diol com periodato), clivagem de sites abásicos por clivagem com endonuclease (por exemplo `USER`, como fornecido por NEB, Ipswich, Mass., EUA, parte número M5505S), por exposição a calor ou alcalinos, clivagem de ribonucleotídeos incorporados em produtos de amplificação de outra forma compreendidos de desoxiribonucleotídeos, clivagem fotoquímica ou clivagem de um ligante peptídeo. Após a operação de linearização um iniciador de sequenciamento pode ser entregue para a célula de fluxo sob condições para hibridização do iniciador de sequenciamento para os ácidos nucléicos alvos que devem ser sequenciados.

[063] Uma célula de fluxo pode, então, ser contatada com um reagente de extensão de SBS que tem nucleotídeos modificados com blocos 3' removíveis e identificadores fluorescentes sob condições para estender um iniciador hibridizado para cada ácido nucléico alvo por uma adição de nucleotídeo único. Apenas um nucleotídeo único é adicionado a cada iniciador devido ao fato de que uma vez que o nucleotídeo modificado tenha sido incorporado à cadeia de polinucleotídeo crescente de forma complementar à região do modelo que é sequenciado hão há grupo 3'-OH livre disponível para dirigir extensão de sequência adicional e, portanto, a polimerase não pode adicionar nucleotídeos adicionais. O reagente de extensão de SBS pode ser removido e substituído por reagente de varredura que contém componentes que protegem a amostra sob excitação com radiação. Componentes exemplificativos para

reagente de varredura são descritos na Publicação US 2008/0280773 A1 e Ser. nº U.S. 13/018.255, cada uma das quais é incorporada ao presente documento a título de referência. Os ácidos nucléicos estendidos podem, então, ser detectados de modo fluorescente na presença de reagente de varredura. Uma vez que a fluorescência tenha sido detectada, o bloco 3' pode ser removido com o uso de um reagente de desbloqueio que é apropriado para o grupo de bloqueio usado. Reagentes de desbloqueio exemplificativos que são úteis para respectivos grupos de bloqueio são descritos em WO004018497, US 2007/0166705A1 e Patente nº U.S. 7.057.026, cada uma das quais é incorporada ao presente documento a título de referência. O reagente de desbloqueio pode ser lavado deixando ácidos nucléicos alvos hibridizados para iniciadores estendidos que têm grupos 3' OH que são agora competentes para a adição de um nucleotídeo adicional. Consequentemente, os ciclos de adicionar 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 com o uso de uma operação de entrega de reagente de extensão única por ciclo quando cada um dos nucleotídeos modificados tem uma identificação diferente fixada ao mesmo, conhecida por corresponder à base particular. Os identificadores diferentes facilitam a discriminação entre os nucleotídeos adicionados durante cada operação de incorporação. Alternativamente, cada ciclo pode incluir operações separadas de entrega de reagente de extensão seguida por operações separadas de entrega e detecção de reagente de varredura, neste caso dois ou mais dos nucleotídeos podem ter a mesma identificação e podem ser distinguidos com base na ordem de entrega conhecida.

[064] Embora A operação de sequenciamento, em 106, tenha sido exemplificada acima em relação a um protocolo de SBS particular, será compreendido que outros protocolos para sequenciamento de qualquer uma de uma variedade de outras análises moleculares podem ser realizados como desejado.

[065] Em 108, o um ou mais processadores do sistema recebem os dados de

sequenciamento para análise subsequente em 110. Os dados de sequenciamento podem ser formatados de diversas maneiras, tais como em um arquivo. BAMS. 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 têm sequências de amostra correspondentes dos nucleotídeos. Embora apenas uma leitura de amostra seja discutida, deve ser compreendido que os dados de sequenciamento podem incluir, por exemplo, centenas, milhares, centenas de milhares, ou milhões de leituras de amostra. Leituras de amostra diferentes podem ter números diferentes de nucleotídeos. Por exemplo, uma leitura de amostra pode variar entre 10 nucleotídeos a cerca de 500 nucleotídeos ou mais. As leituras de amostra podem abranger o genoma inteiro da fonte (ou fontes). Como um exemplo, as leituras de amostra são dirigidas a locais genéticos predeterminados, tais como aqueles locais genéticos que têm STRs suspeitos ou SNPs suspeitos.

[066] Cada leitura de amostra pode incluir uma sequência de nucleotídeos, que pode ser denominada uma sequência de amostra, fragmento de amostra ou uma sequência alvo. A sequência de amostra pode incluir, por exemplo, sequências de iniciador, sequências de acompanhamento 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 modalidades, uma ou mais das leituras de amostra (ou sequências de amostra) inclui pelo menos 150 nucleotídeos, 200 nucleotídeos, 300 nucleotídeos, 400 nucleotídeos, 500 nucleotídeos, ou mais. Em algumas modalidades, as leituras de amostra podem incluir mais do que 1.000 nucleotídeos, 2.000 nucleotídeos, ou mais. As leituras de amostra (ou as sequências de amostra) podem incluir sequências de iniciador em uma ou ambas as extremidades.

[067] Em 110, o um ou mais processadores analisam os dados de sequenciamento para obter chamada (ou chamadas) de variante potencial e uma frequência de variante de amostra da chamada (ou chamadas) de variante de amostra. A operação em 110 também pode ser denominada um aplicativo de chamada de variante ou chamador de variante. Chamadores de variante alternativos

podem ser utilizados de acordo com modalidades no presente documento, em que chamadores de variante diferentes podem ser usados com base no tipo de operação de sequenciamento que é realizada, com base em recursos da amostra que são de interesse e similares. Um exemplo não limitante de um aplicativo de chamada de variante é o aplicativo Pisces™ por Illumina Inc. (San Diego, CA). Adicional ou alternativamente, a operação em 110 pode utilizar o aplicativo de chamada de variante descrito no artigo "Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs" por Sanders et al., Bioformatics, 15 de julho de 2012, Volume 28, No. 14, páginas 1811 a 1817, cuja matéria completa é expressamente incorporada em sua totalidade ao presente documento a título de referência.

[068] De acordo com pelo menos algumas modalidades, o aplicativo de chamada de variante em 110 fornece chamadas para variantes de baixa frequência, chamada de linha germinativa e similares. Como exemplo não limitante, o aplicativo de chamada de variante em 110 pode executar em amostras apenas de tumor e/ou amostras pareadas tumor-normal. O aplicativo de chamada de variante em 110 pode procurar por variações de nucleotídeo único (SNV), variações de múltiplos nucleotídeos (MNV), indels e similares. O aplicativo de chamada de variante identifica variantes, enquanto filtragem para incompatibilidades devido a erros de preparação de sequenciamento ou amostra. Para cada variante, o chamador de variante identifica a sequência de referência, uma posição da variante, e a sequência (ou sequências) de variante potencial (por exemplo, SNV de A a C, ou deleção de AG a A). O aplicativo de chamada de variante identifica a sequência de amostra (ou fragmento de amostra), uma sequência de referência/fragmento e uma chamada de variante como uma indicação de que uma variante está presente. Como explicado doravante em conexão com A Figura 1B, o aplicativo de chamada de variante em 110 pode identificar fragmentos brutos, e emitir uma designação dos fragmentos brutos, uma contagem do número de fragmentos brutos que verifica a chamada de variante potencial, a posição dentro do fragmento bruto na qual uma variante de apoio ocorreu e outras informações relevantes. Exemplos não limitantes de fragmentos brutos incluem um

fragmento remendado duplex, um fragmento remendado simples, um fragmento não remendado duplex e um fragmento não remendado simples.

[069] O aplicativo de chamada de variante pode emitir as chamadas em vários formatos, tais como em um arquivo .VCF ou .GVCF. A título de exemplo apenas, o aplicativo de chamada de variante pode ser incluído em uma sequência de MiSeqReporter (por exemplo, quando implantado no instrumento sequenciador MiSeq®). Opcionalmente, o aplicativo pode ser implantado com vários fluxos de trabalho. A análise, em 110, pode incluir um protocolo único ou uma combinação de protocolos que analisam as leituras de amostra em uma maneira designada para obter informações desejadas.

[070] Em 111, o um ou mais processadores realiza uma operação de validação em conexão com a chamada de variante potencial. A operação de validação como descrita abaixo em mais detalhes em conexão com as Figuras 1C e 5B. A operação de validação pode ter como base uma classificação de qualidade, e/ou uma hierarquia de testes escalonados, como explicado doravante. Quando a operação de validação autentica ou verifica que a chamada de variante potencial, a operação de validação passa as informações de chamada de variante (do aplicativo de chamada de variante em 110) para o gerador de relatório de amostra em 112. Alternativamente, 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 sem chamada, um indicador de chamada inválida) para o gerador de relatório de amostra em 112. A operação de validação em 111 também pode passar uma classificação de confiança relacionada a um grau de confiança de que a chamada de variante é correta ou a designação chamada inválida é correta.

[071] Em 112, o um ou mais processadores gera e armazena um relatório de amostra. O relatório de amostra pode incluir, por exemplo, informações a respeito de uma pluralidade de locais genéticos em relação à amostra. Por exemplo, para cada local genético de um conjunto predeterminado de locais genéticos, o relatório de

amostra pode realizar pelo menos um dentre fornecer uma chamada de genótipo; indicar que uma chamada de genótipo não pode ser feita; fornecer uma classificação de confiança sobre uma certeza da chamada de genótipo; ou indicar problemas potenciais com um ensaio a respeito de um ou mais locais genéticos. O relatório de amostra também pode indicar um gênero de um indivíduo que forneceu uma amostra e/ou indicar que a amostra inclui múltiplas fontes. Como usado no presente documento, um "relatório de amostra" pode incluir dados digitais (por exemplo, um arquivo de dados) de um local genético ou conjunto predeterminado de local genético e/ou um relatório impresso do local genético ou do conjunto de locais genéticos. Assim, gerar ou fornecer, em 112, pode incluir criar um arquivo de dados e/ou imprimir o relatório de amostra ou exibir o relatório de amostra.

[072] O relatório de amostra pode indicar que uma chamada de variante foi determinada, mas foi não validada. Quando uma chamada de variante é determinada inválida, o relatório de amostra pode indicar informações adicionais a respeito da base para a determinação para 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) em que os fragmentos brutos suportam ou contradizem a chamada de variante. Adicional ou alternativamente, as informações adicionais no relatório podem incluir a classificação de qualidade obtida de acordo com modalidades descritas no presente documento.

APLICATIVO DE CHAMADA DE VARIANTE

[073] A Figura 1B ilustra um fluxograma para um método 114 para analisar dados de sequenciamento para identificar chamadas de variante potenciais de acordo com uma modalidade no presente documento. A título de exemplo, as operações da Figura 1B podem ser realizadas durante 110 na Figura 1A. É compreendido que as operações da Figura 1B podem operar sobre dados armazenados por uma operação de sequenciamento realizada previamente. Adicional ou alternativamente, a operação da Figura 1B pode ser realizada em tempo real enquanto uma operação de sequenciamento está sendo realizada. Em 118, cada uma das leituras de amostra é

atribuída a locais genéticos correspondentes. As leituras de amostra podem ser atribuídas a locais genéticos correspondentes com base na sequência dos nucleotídeos da leitura de amostra ou, em outras palavras, na ordem de nucleotídeos dentro da leitura de amostra (por exemplo, A, C, G, T). Com base nessa análise, a leitura de amostra pode ser designada como incluindo uma variante/alelo possível de um local genético particular. A leitura de amostra pode ser coletada (ou agregada ou guardada) com outras leituras de amostra que tenham sido designadas como incluindo variantes/alelos possíveis do local genético. A operação de atribuição, em 118, também pode ser denominada uma operação de chamada na qual a leitura de amostra é identificada como possivelmente sendo associada a uma posição/local genético particular. As leituras de amostra podem ser analisadas para localizar uma ou mais sequências de identificação (por exemplo, sequências de iniciador) de nucleotídeos que diferenciam a leitura de amostra de outras leituras de amostra. Mais especificamente, a sequência (ou sequências) de identificação pode identificar a leitura de amostra a partir de outras leituras de amostra como sendo associada com um local genético particular.

[074] A operação de atribuição, em 118, pode incluir analisar a série de n nucleotídeos da sequência de identificação para determinar se a série de n nucleotídeos da sequência de identificação correspondem efetivamente a uma ou mais das sequências de seleção. Em modalidades particulares, a operação de atribuição, em 118, pode incluir analisar os 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 de seleção. 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 sequência de seleção mais curta 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. Entretanto, menos ou mais nucleotídeos podem ser usados

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

[075] Em alguns casos, a série de n nucleotídeos de uma sequência de identificação pode não correlacionar com precisão os nucleotídeos da sequência de seleção. Ainda assim, a sequência de identificação pode correlacionar efetivamente a sequência de seleção se a sequência de identificação for quase idêntica à sequência de seleção. Por exemplo, a leitura de amostra pode ser chamada para um local genético se a série de n nucleotídeos (por exemplo, os primeiros n nucleotídeos) da sequência de identificação correlacionar uma sequência de seleção com não mais do 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 inadequação ou deslocamento possa contar como uma diferença entre a leitura de amostra e a sequência de iniciador. Caso o número de diferenças seja menor do que um número designado, então, a leitura de amostra pode ser chamada para o local genético correspondente (isto é, atribuído para o local genético correspondente). Em algumas modalidades, pode ser determinado que uma classificação de correlação tem como base o número de diferenças entre a sequência de identificação da leitura de amostra e a sequência de seleção associada com um local genético. Caso a classificação de correlação passe um designado limiar de correlação, então, o local genético que corresponde à sequência de seleção pode ser designado como um local potencial para a leitura de amostra. Em algumas modalidades, análise subsequente pode ser realizada para determinar se a leitura de amostra é chamada para o local genético.

[076] Caso a leitura de amostra confirme efetivamente uma das sequências de seleção no banco de dados (isto é, confirme exatamente ou quase confirme como descrito acima), então, a leitura de amostra é atribuída ou designada para o local genético que correlaciona à sequência de seleção. Essa pode ser denominada chamada de local ou chamada de local provisório, em que a leitura de amostra é chamada para o local genético que correlaciona à sequência de seleção. Entretanto, como discutido acima, uma leitura de amostra pode ser chamada para mais do que um local genético. Em tais modalidades, análise adicional pode ser realizada para chamar ou atribuir a leitura de amostra a apenas um dos locais genéticos potenciais. Em algumas modalidades, a leitura de amostra que é comparada ao banco de dados de sequências de referência é a primeira leitura de sequenciamento de extremidade pareada. Quando realiza sequenciamento de extremidade pareada, é obtida uma segunda leitura (que representa um fragmento bruto) que correlaciona à leitura de amostra. Após atribuir, em 118, sendo que a análise subsequente é realizada com as leituras atribuídas pode ter como base o tipo de local genético que tenha sido chamado para a leitura atribuída.

[077] Em 120, as leituras de amostra são analisadas para identificar chamadas de variante potenciais. Entre outras coisas, os resultados da análise identificam a chamada de variante potencial, uma frequência de variante de amostra, uma sequência de referência e uma posição dentro da sequência genômica de interesse em que a variante ocorreu. Por exemplo, se um local genético é conhecido por incluir SNPs, então, as leituras atribuídas que foram chamadas para o local genético podem passar por análise, em 120, para identificar os SNPs das leituras atribuídas. Se o local genético é conhecido por incluir elementos de DNA polimórfico repetitivo, então, as leituras atribuídas podem ser analisadas, em 120, para identificar ou caracterizar os elementos de DNA polimórfico repetitivo dentro das leituras de amostra. Em algumas modalidades, se uma leitura atribuída correlaciona efetivamente com um STR local e um SNP local, um alerta ou sinalizador pode ser atribuído à leitura de amostra. A leitura de amostra pode ser designada tanto como um STR local quanto um SNP local.

A análise, em 1202, pode incluir alinhar as 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 Internacional nº PCT/US2013/030867 (Publicação nº WO 2014/142831), depositado em 15 de março de 2013, o qual é incorporado em sua totalidade ao presente documento a título de referência.

[078] Em 122, o um ou mais processadores analisam fragmentos brutos para determinar se existem variantes de apoio em posições correspondentes dentro dos 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 chamada de variante original. Por exemplo, o tipo de fragmento bruto pode representar um fragmento remendado duplex, um fragmento remendado simples, um fragmento não remendado duplex ou um fragmento não remendado simples. Opcionalmente outros fragmentos brutos podem ser identificados em vez de ou além dos exemplos supramencionados. Em conexão com identificar cada tipo de fragmento bruto, o chamador de variante também identifica a posição, dentro do fragmento bruto, em que a variante de apoio ocorreu, bem como uma contagem do número de fragmentos brutos que exibiram a variante de apoio. Por exemplo, o chamador de variante pode emitir uma indicação de que 10 leituras de fragmentos brutos foram identificados para representar fragmentos remendados duplex que têm uma variante de apoio em uma posição particular X. O chamador de variante também pode emitir indicação de que cinco leituras de fragmentos brutos foram identificadas para representar fragmentos não remendados simplex que têm uma variante de apoio em uma posição particular Y. O chamador de variante também pode emitir um número de fragmentos brutos que correspondem às sequências de referência e, desse modo, não incluir uma variante de apoio que de outra forma forneceria evidência para validar a chamada de variante potencial na sequência genômica de interesse.

[079] Em 122, é mantida uma contagem dos fragmentos brutos que incluem

variantes de apoio, bem como a posição em que a variante de apoio ocorreu. Adicional ou alternativamente, pode ser mantida uma contagem dos fragmentos brutos que não incluem variantes de apoio na posição de interesse (relativa à posição da chamada de variante potencial na leitura de amostra ou fragmento de amostra). Adicional ou alternativamente, pode ser mantida uma contagem de fragmentos brutos que correspondem a uma sequência de referência e não autenticam ou confirmam a chamada de variante potencial. As informações determinadas em 122 são emitidas para o aplicativo de validação de chamada de variante, que incluem uma contagem e tipo dos 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 similares.

[080] Quando uma chamada de variante potencial é identificada, o processo da Figura 1B emite uma indicação da chamada de variante potencial, da sequência de variante, da posição de variante e uma sequência de referência associada às mesmas. A chamada de variante é designada para representar uma variante "potencial" uma vez que erros podem fazer com que o processo de chamada identifique uma variante falsa. De acordo com modalidades no presente documento, a chamada de variante potencial é analisada para reduzir e eliminar variantes falsas ou falsos positivos. Adicional ou alternativamente, o processo da Figura 1B analisa um ou mais fragmentos brutos associados com uma leitura de amostra e emite uma chamada de variante correspondente associada com os fragmentos brutos.

APLICATIVO DE VALIDAÇÃO DE CHAMADA DE VARIANTE

[081] A Figura 1C ilustra um método implantado por computador para validar chamadas de variante de acordo com modalidades no presente documento. Em 130, o um ou mais processadores recebe dados do aplicativo de chamada de variante (Figura 1B) que incluem uma indicação da chamada de variante potencial em uma posição designada, uma posição da chamada de variante potencial dentro da sequência genômica de interesse. Em 130, o um ou mais processadores também recebem uma contagem e tipo dos fragmentos brutos que suportam a chamada de

variante potencial, posições das variantes de apoio nos fragmentos brutos, uma contagem dos fragmentos brutos que não suportam a chamada de variante potencial e outras informações relevantes. É compreendido que a operação em 130 pode corresponder a uma ou mais operações para acessar dados armazenados a uma sequência realizada previamente. Adicional ou alternativamente, a operação em 130 pode ser realizada em tempo real enquanto uma operação de sequenciamento está sendo realizada.

[082] Opcionalmente, em 130, o um ou mais processadores podem receber uma indicação de uma sequência de referência, em relação à qual a chamada de variante potencial foi feita. A sequência de referência corresponde a uma sequência de nucleotídeos dentro de uma ou mais sequência (ou sequências) genômica de linha de base.

[083] Em 132, o um ou mais processadores obtêm uma ou mais frequências de variante de linha de base na posição designada dentro de uma ou mais sequências genômicas de linha de base (por exemplo, de acordo com o processo da Figura 1D). Como um exemplo, as frequências de variante de linha de base são derivadas de múltiplas sequências genômicas de linha de base que são associadas com mais do que um tipo de alelo. Opcionalmente, as frequências de variante de linha de base podem ser associadas com um tipo de alelo, mas derivadas de amostras de múltiplos indivíduos. As frequências de variante de linha de base podem ser determinadas ao mesmo tempo em que as outras operações da Figura 1C. Alternativamente, as frequências de variante de linha de base podem ser predeterminadas antes da implantação de uma operação de sequenciamento para uma amostra de interesse ou do processo de validação da Figura 1C. As frequências de variante de linha de base podem ser armazenadas (por exemplo, em uma biblioteca ou banco de dados) em conexão com várias posições junto com várias sequências de referência genômicas de linha de base. Por exemplo, uma biblioteca ou banco de dados pode ser mantida para múltiplas sequências genômicas de linha de base. A biblioteca ou banco de dados (por exemplo, 784 na Figura 7) também pode reter frequências de variante de

linha de base associadas com cada posição junto com a sequência genômica de linha de base. É reconhecido que uma sequência genômica de linha de base pode exibir diferentes tipos de frequências de variante de linha de base, tal como em conexão com diferentes tipos de métodos de sequenciamento, métodos de preparação, equipamento de sequenciamento e similares. Consequentemente, a biblioteca ou banco de dados pode manter um conjunto de frequências de variante de linha de base para qualquer dada posição junto com uma sequência genômica de linha de base. A operação em 132 pode incluir enviar uma solicitação para um servidor de rede ou computador remoto (por exemplo, 780) que gerencia uma biblioteca ou banco de dados. A solicitação pode designar a sequência genômica de linha de base de interesse, a posição designada junto, bem como informações adicionais (por exemplo, o tipo de operação de sequenciamento, o tipo de equipamento que é usado, ou o protocolo de preparação de biblioteca).

[084] As frequências de variante de linha de base podem ser determinadas em maneiras diferentes para identificar características diferentes de interesse que descrevem variação em uma ou mais posições junto com a sequência genômica de linha de base. A título de exemplo, as frequências de variante de linha de base indicam um grau (ou uma extensão) de ruído de fundo em posições correspondentes junto com a sequência genômica de linha de base. O grau/extensão de ruído de fundo pode variar, por exemplo dependendo do tipo de operação de sequenciamento, tipo de equipamento, processo de preparação de biblioteca e similares.

[085] Em 134, o um ou mais processadores obtém a frequência de variante de amostra na posição designada para a sequência genômica de interesse. Por exemplo, os processadores obtêm a frequência de variante de amostra a partir do aplicativo de chamada de variante, em que a frequência de variante de amostra representa uma frequência relativa de um alelo (variante de um gene) em uma posição particular/local em uma sequência genômica de interesse através de uma população de amostras. Por exemplo, a frequência de variante de amostra pode ser expressa como uma fração ou percentual de todas as amostras de interesse (por exemplo, cromossomas) do

indivíduo que foram analisadas.

[086] Em 136, o um ou mais processadores analisa as frequências de variante de linha de base e amostra na posição designada para obter uma classificação de qualidade. A título de exemplo, a análise pode incluir obter uma relação entre a frequência de variante de amostra e uma distribuição das frequências de variante de linha de base, sendo que a classificação de qualidade é determinada com base na relação. Por exemplo, múltiplas frequências de variante de linha de base podem ser armazenadas para uma população de sequências genômicas de linha de base. As múltiplas frequências de variante de linha de base são organizadas em uma distribuição. Os processadores podem indexar a frequência de variante de amostra em relação à distribuição das frequências de variante de linha de base. Como um exemplo, a relação pode ter como base um teste não paramétrico, tal como um teste de soma de classificação de Wilcoxon. O teste de soma de classificação de Wilcoxon produz um valor p que representa uma relação entre a frequência de variante de amostra e a distribuição das frequências de variante de linha de base. O valor p representa uma indicação numérica de um grau de confiança que uma chamada de variante de amostra é devido a ruído ou não devido a ruído de fundo.

[087] Os processadores convertem o valor numérico p para uma classificação de qualidade (por exemplo, denominada uma classificação Q). Por exemplo, a classificação de qualidade pode simplesmente ser igual ao valor p. Alternativamente, a classificação de qualidade pode ser formada aplicando-se um operador matemático predeterminado ao valor p (por exemplo, normalizar o valor p, converter o valor p para um número inteiro). Opcionalmente, o valor p pode ser combinado com outras informações/fatores para formar a classificação de qualidade. Por exemplo, o valor p pode ser modificado com base em um número de amostras e/ou leituras de amostra obtidas. Opcionalmente, testes alternativos podem ser aplicados para determinar uma relação entre as frequências de variante de linha de base e a frequência de variante de amostra, em que tais testes alternativos proporcionam uma indicação de um grau de confiança em uma chamada de variante de amostra. Por exemplo, o grau de

confiança pode indicar que a chamada de variante de amostra é devido a ruído de fundo ou não devido a ruído de fundo. Alternativamente, o grau de confiança pode indicar que a chamada de variante de amostra inclui vários tipos de erros sistemáticos que provocam variantes falsas, tal como devido a artefatos de FFPE, erros de sequenciamento ou erros de PCR.

[088] Em 138, o um ou mais processadores validam a chamada de variante potencial para a sequência genômica de interesse com base na classificação de qualidade. A título de exemplo, a operação de validação pode compreender adicionalmente comparar a classificação de qualidade a um limiar e declarar a chamada de variante potencial como sendo uma chamada de variante válida quando a classificação de qualidade excede o limiar. O limiar representa um ponto de corte de classificação Q para filtrar variantes falsas potenciais. Uma chamada de variante válida pode representar uma chamada de variante para a qual existe uma confiança "alta" de que a chamada é precisa. A medida do que representa confiança alta, versus confiança baixa é reconhecido que representa um termo relativo que pode variar com base nos fatos e circunstâncias particulares da implantação. Como um exemplo, o limiar pode ser variado com base em análise empírica.

[089] A Figura 1D ilustra um processo para obter frequências de variante de linha de base de acordo com modalidades no presente documento. O processo da Figura 1D pode ser realizado por um ou mais processadores de vários sistemas e dispositivos. É compreendido que a ordem das operações da Figura 1D representam um exemplo, enquanto as operações podem ser realizadas em ordens alternativas. Ademais, uma ou mais das operações da Figura 1D pode ser omitida inteiramente. As operações da Figura 1D serão descritas em conexão com A Figura 2.

[090] O método da Figura 1D pode ser repetido periodicamente para compilar uma distribuição de erros antecedentes específicos do local de uma pluralidade de painéis para amostras FFPE normais com qualidade de DNA variada de vários tecidos sequenciados por um ensaio à base de NGS. O método da Figura 1D pode utilizar os mesmos dados de sequenciamento da pluralidade de painéis das amostras FFPE

normais para normalizar tendência sistemática em cobertura de leitura provocada por PCR, qualidade de DNA, eficiência de tração da sonda, ou conteúdo GC de sequência para revelar um número preciso de variantes em uma amostra de teste.

[091] A Figura 2A ilustra painéis exemplificativos 202 a 206 como indivíduos diferentes dentro de uma população de linha de base. No presente exemplo, a população de linha de base representa indivíduos normais/saudáveis, embora, é reconhecido que a população de linha de base pode representar indivíduos que exibem uma mutação de interesse particular, tal como um tipo particular de câncer e similares. O eixo geométrico vertical dos painéis 202 a 206 representa o número de falsos positivos ou chamadas de variante falsas encontrado por 1.000 bases analisadas. É compreendido que cada painel 202 a 206 corresponda a um acúmulo de um grande número de leituras de referência, por exemplo 1.000, 5.000, 10.000 e similares. O eixo geométrico horizontal indica toda ou uma porção dos cromossomas dentro de uma sequência genômica de linha de base. No exemplo da Figura 2, 19 cromossomas são ilustrados junto com a sequência genômica de linha de base, embora seja compreendido que a sequência genômica de linha de base pode ser mantida para mais ou menos cromossomas. Uma série de barras verticais é designada junto com o eixo geométrico horizontal, sendo que cada barra se estende para cima por uma quantidade que corresponde ao número de variantes falsas na posição ou site correspondente junto com o cromossoma correspondente. As barras verticais são posicionadas junto com o eixo geométrico horizontal em localizações aproximadas da posição ou coordenada correspondente dentro do cromossoma associado. Por exemplo, uma barra vertical 221 é apresentada em conexão com o 10º cromossoma para indicar que o 10º cromossoma registrou aproximadamente 25 variantes falsas por 1.000 Bases, em que a posição ou coordenada dentro do cromossoma é levemente abaixo do centro do cromossoma. Como outro exemplo, uma barra vertical 223 é apresentada em conexão com o primeiro cromossoma para indicar que o primeiro cromossoma registrou aproximadamente 7 variantes falsas por 1.000 Bases em uma posição ou coordenada levemente acima do centro do primeiro

cromossoma.

[092] A Figura 2B ilustra um painel de par de bases que representa uma ampliação para uma porção pequena de um dos painéis 202 a 206 da Figura 2A. A título de exemplo, a Figura 2B pode corresponder a leituras de referência para pares de bases 1 a 9 no cromossoma 16 do painel 202 (Figura 2A). O painel de par de bases 208 inclui um eixo geométrico vertical que representa o número de chamadas de variante falsas encontrado por 1.000 bases analisadas. Como explicado acima em conexão com A Figura 2A, o painel de par de bases 208 corresponde a um acúmulo de um grande número de leituras de referência, por exemplo 1.000, 5.000, 10.000 e similares. O eixo geométrico horizontal indica uma sequência curta de pares de bases associada com um único cromossoma dentro de uma sequência genômica de linha de base. No exemplo da Figura 2B, 9 pares de bases são ilustrados junto com a sequência genômica de linha de base, embora seja compreendido que a sequência genômica de linha de base é mantida para mais pares de bases para os cromossomas presentes e outros. Uma série de barras verticais é designada junto com o eixo geométrico horizontal, sendo que cada barra se estende para cima por uma quantidade correspondente ao número de variantes falsas na posição ou site de par de bases correspondente junto com o cromossoma correspondente. As barras verticais são posicionadas junto com o eixo geométrico horizontal em localizações do par de bases correspondente dentro do cromossoma associado. Por exemplo, uma barra vertical é apresentada no par de bases #2 para indicar que o par de bases #1 registrou aproximadamente 10 variantes falsas por 1.000 Bases. Como outro exemplo, uma barra vertical é apresentada em conexão com o par de bases #5 para indicar que o par de bases #5 registrou aproximadamente 25 variantes falsas por 1.000 Bases.

[093] Voltando para a Figura 1D, em 150, o um ou mais processadores recebem dados de sequência para uma pluralidade de leituras de referência para uma ou mais sequências genômicas de linha de base. Em relação à Figura 2A, caso seja assumido que cada um dos painéis 202 a 206 representa um acúmulo de 1.000 leituras de referência, então, o sistema obteria cerca de 3.000 leituras de referência. Em 151, o

um ou mais processadores selecionam um conjunto das leituras de referência a partir da pluralidade de leituras de referência. Em relação à Figura 2A, o conjunto de leituras de referência pode ser selecionado correspondente aos pares de bases 1 a 9 no cromossoma 16 para o indivíduo designado como WGDNA04 (painel 202). A Figura 2B ilustra um exemplo para os pares de bases 1 a 9 no cromossoma 16. Em 152, o um ou mais processadores do sistema selecionam uma posição atual no conjunto de leituras de referência atual. Por exemplo, a posição atual pode ser centralizada na posição #2 junto com o painel 208. Em 154, o um ou mais processadores do sistema sobrepõe a posição atual no conjunto de leituras de referência com uma janela de par de bases.

[094] Na Figura 2B, um suporte 210 é apresentado para ilustrar um exemplo de uma janela de par de bases. O comprimento da janela de par de bases pode ser variado, por exemplo variando de um par de bases a qualquer número desejado de múltiplos pares de bases. No exemplo da Figura 2B, a janela de par de bases foi designada para corresponder a três pares de bases, embora outros comprimentos possam ser utilizados. Consequentemente, na presente iteração através das operações da Figura 1D, a janela de par de bases 210 abrange as posições #1 a #3. Utilizar uma janela de par de bases que cobre mais do que um par de bases ser responsável pela circunstância em que posições ruidosas podem ocorrer juntas. Por exemplo, quando preparação de biblioteca ou outros fatores fazem com que ocorra ruído em uma posição em uma sequência de nucleotídeo, os mesmos fatores podem fazer com que ruído ocorra em uma ou mais posições adjacentes àquela posição. O comprimento da janela de par de bases pode ser definido em parte com base na extensão que é esperada que ruído abranja através de múltiplas posições de par de bases.

[095] Em 156, o um ou mais processadores identifica uma frequência de variante candidata para cada posição das leituras de referência dentro da janela de par de bases. Continuando com o exemplo supramencionado, uma frequência de variante candidata seria identificada para cada uma das posições #1 a #3 dentro da janela de

par de bases 210 (por exemplo, cerca de 2 variantes falsas em #1, cerca de 10 variantes falsas na posição #2, e cerca de 0 variantes falsas na posição #3).

[096] Em 158, o um ou mais processadores selecionam uma das frequências de variante candidatas para representar a frequência de variante de linha de base (também denominada a frequência de variante resultante) para a posição atual dentro da sequência genômica de linha de base. Por exemplo, a posição atual pode corresponder ao centro de janela de par de bases 210. Alternativamente, a posição atual pode corresponder à borda dianteira ou borda traseira da janela de par de bases 210. A seleção para a linha de base/frequência de variante resultante pode ser com base em vários critérios. Por exemplo, os processadores podem selecionar a frequência de variante candidata mais alta dentro da janela de par de bases. Alternativamente, os processadores podem formar uma combinação matemática das frequências de variante candidatas, tal como através de ponderação, uma soma ponderada ou similares, para formar a frequência de variante de linha de base/resultante. A frequência de variante de linha de base/resultante é armazenada em conexão com a posição de par de bases atual. Por exemplo, a linha de base/frequência de variante resultante pode ser armazenada no banco de dados de sequência genômica de linha de base (784 na Figura 7). Em referência à Figura 2B, a janela de par de bases 210 pode ser atribuída à frequência de variante de linha de base de 10 (correspondente à frequência de variante mais alta dentro da janela).

[097] Em 162, o um ou mais processadores do sistema armazenam a frequência de variante de linha de base em conexão com a posição atual. A frequência de variante de linha de base pode ser gravada, em relação à posição atual, de maneiras diferentes. A frequência de variante para o par de bases #1 e um par de bases #3 (ambos dentro da janela 210) pode ser deixada inalterada. Em relação à Figura 2B, o valor de frequência de variante de linha de base de 10 pode ser gravado em conexão com o par de bases posicionado #2. Alternativamente, a frequência de variante de linha de base pode ser atribuída para a posição dianteira #1 ou posição traseira #3 dentro da janela de par de bases, enquanto às posições adjacentes é atribuído um

valor nulo ou são deixadas inalteradas. Opcionalmente, a frequência de variante de linha de base pode ser atribuída para cada posição dentro da janela de par de bases.

[098] Opcionalmente, a frequência de variante de linha de base pode ser definida para um nível predefinido, em que níveis predefinidos diferentes são utilizados em conexão com frequências de variante candidatas diferentes. Por exemplo, quando as frequências de variante candidatas são relativamente altas (por exemplo, acima cerca de 30%), a frequência de variante de linha de base pode ser definida para zero ou outro nível inferior predefinido. A título de exemplo, a frequência de variante de linha de base pode ser definida para zero ou um nível predefinido, quando frequências de variante candidatas excedem um limiar para evitar contar variante semelhante a germe como ruído de fundo.

[099] Em 164, o um ou mais processadores determinam se deslocam a janela de par de bases para outra posição junto com o conjunto de leituras de referência atual. A determinação para deslocar a janela de par de bases pode ser com base em se uma porção adicional da leitura de referência permanece para ser analisada. Quando a decisão em 164 é para deslocar a janela de par de bases, o fluxo se move para 166. Caso contrário, o fluxo continua para 168.

[0100] Em 166, o um ou mais processadores do sistema deslocam a janela de par de bases para uma próxima posição junto com o conjunto de leituras de referência, e o fluxo retorna para 154. Depois disso, as operações em 154 e 156 são repetidas. A título de exemplo, o deslocamento em 166 pode deslocar a janela de par de bases por um único par de bases, ou por múltiplos pares de bases. Por exemplo, com referência à Figura 2B, o deslocamento pode corresponder a um único par de bases. Assim, a janela de par de bases é deslocada da posição designada no suporte 210 para a posição designada no suporte 212 (cobrindo as posições de pares de bases #2 a #4). Durante uma iteração subsequente através das operações em 164 a 166, a janela de par de bases pode ser deslocada da posição designada em 212 para a posição designada em 214 (cobrindo as posições de par de bases #3 a #5).

[0101] No presente exemplo, a janela de par de bases é deslocada para posições

de sobreposição sucessivas. Quando a janela de par de bases inclui mais do que 3 pares de bases, o deslocamento em 166 pode de modo similar ser maior do que 1 par de bases. Por exemplo, caso a janela de par de bases seja de 5 pares de bases de comprimento e a posição atual corresponda a janela (posições #1 a #5), em 166, a janela de par de bases pode ser deslocada para a frente 2 posições para sobrepor as posições #3 a #7, de modo que posições sucessivas da janela de par de bases sejam sobrepostas, mas avança por múltiplos pares de bases entre janelas de par de bases sucessivas.

[0102] Alternativamente, a janela de par de bases pode ser deslocada para posições não de sobreposição sucessivas junto com a leitura de referência. Por exemplo, se a janela de par de bases for de 3 pares de bases de comprimento e a posição atual corresponder às posições #1, #2 e #3 no conjunto de leituras de referência, em 166, a janela de par de bases pode ser deslocada para a frente 3 posições para sobrepor as posições #4, #5 e #6, de modo que posições sucessivas da janela de pares de bases não sejam sobrepostas.

[0103] Em seguida, as operações em 154 a 162 são repetidas para a próxima posição. Quando o fluxo avança de 164 a 168, o um ou mais processadores do sistema determina se conjuntos adicionais de leituras de referência devem ser analisados a partir da pluralidade de leituras de referência. Quando leituras de referência adicionais devem ser analisadas, o fluxo retorna para 170. Caso contrário, o processo da Figura 1D termina e o fluxo retorna para A Figura 1C. Em 170, o um ou mais processadores seleciona o próximo conjunto de leituras de referência a ser analisado. Depois disso, o fluxo retorna para 152. As operações em 152 a 166 são, então, repetidas.

[0104] A Figura 3 ilustra um gráfico em conexão com o princípio de correção de fundo com a utilização de sequências genômicas de linha de base normais. O eixo geométrico horizontal representa frequências de alelo variantes (VAF) que correspondem à frequência de variação em uma amostra em alguma posição designada junto com uma sequência de nucleotídeo. Cada círculo 302 representa uma

variante associada com uma amostra, enquanto o agrupamento 304 de variantes 302 corresponde a uma posição junto com uma sequência genômica para um agrupamento de amostras. O agrupamento 306 exibe uma quantidade relativamente grande de ruído (também denominado um local com ruído) através da população de linha de base. Os círculos 306 também representam uma variante associada com uma amostra diferente, enquanto que o agrupamento 308 corresponde a uma posição junto com uma sequência genômica para um agrupamento de amostras. O agrupamento 308 exibe uma quantidade relativamente pequena de ruído (também denominado um local limpo) através da população de linha de base.

[0105] O método da Figura 1D desenvolve um perfil para o ruído de fundo para uma população de amostras normais coletando-se as frequências de variante em posições de interesse junto com a sequência genômica de interesse. A população de fundo exibe um "local limpo" em que frequência de alelo é relativamente distribuída de forma apertada/próxima em uma área próxima a zero, tal como no agrupamento 308. Consequentemente, uma chamada de variante de tumor (com um indivíduo que é testado) pode ser facilmente diferenciada de ruído mesmo quando a frequência de alelo para a sequência genômica de amostra de interesse é baixa (por exemplo, na amostra de tumor número 1). Alternativamente, a população de fundo pode exibir um "local com ruído" em que a frequência de alelo é distribuída/espalhada mais amplamente em vários pontos de interesse junto com a sequência genômica de interesse, tal como no agrupamento 304. Consequentemente, pode ser mais difícil para diferenciar uma chamada de variante de tumor (para uma amostra de interesse) que exibe uma baixa frequência de variante (por exemplo, tal como a amostra de tumor 1) de ruído de fundo. Embora, uma chamada de variante de alta frequência (por exemplo, tal como amostra de tumor 2) ainda possa ser chamada com confiança, independentemente de se a posição de interesse exibe características de fundo limpas ou com ruído. Modalidades no presente documento utilizam o nível de ruído de fundo de amostras normais para ajustar rigor de chamada em conexão com amostras de tumor.

[0106] A seguir, aspectos de modalidades no presente documento são descritos em conexão com a utilização de informações dos processos supramencionados e informações de fragmentos brutos para melhorar sensibilidade e especificidade de chamada de variante.

[0107] A Figura 4 ilustra uma árvore de decisão com base em hierarquia implantada de acordo com modalidades no presente documento para filtrar chamadas de variante. Em 402, um ou mais processadores determinam se uma chamada entrante indica que uma variante (alelo) foi observada. Quando nenhuma variante é observada, isso pode representar que foi identificado que a chamada corresponde a uma sequência de referência sem qualquer variante. Quando é indicado que uma chamada entrante corresponde a uma sequência de referência, o fluxo ramifica para 404. Caso contrário, quando é indicado que uma chamada entrante inclui uma variante/alelo, o fluxo ramifica para 406.

[0108] Em 404, o um ou mais processadores determinam se uma quantidade predeterminada de cobertura foi obtida para a variante/alelo atual. A cobertura pode ser determinada com base em uma indicação de um número de leituras que foi coletado que cobre um fragmento de uma sequência de genoma de interesse que inclui uma posição/local atual. A cobertura pode ser definida de diversas maneiras. Por exemplo, a cobertura pode ser definida com base em um número de moléculas que foram analisadas para obter leituras para a posição/local atual. Adicional ou alternativamente, a cobertura pode ser definida com base em um número de amostras, a partir das quais leituras foram obtidas para a posição/local atual. Em 404, um limiar de cobertura é definido (por exemplo, 10X, 100X, etc.). Quando o número de leituras, que inclui o local atual, excede o limiar de cobertura, o fluxo avança para 408 onde uma “passagem negativa” é declarada. A “passagem negativa” indica que o processo valida a chamada entrante na forma original (por exemplo, verifica a chamada negativa). Por exemplo, quando a chamada entrante identifica uma sequência de amostra para corresponder a uma sequência de referência particular, em 408, o processo passa a chamada sem modificação.

[0109] Caso contrário, quando os processadores determinam que a cobertura não iguala ou excede o limiar de cobertura, o fluxo avança para 410 onde uma condição "sem chamada" é designada. Uma condição sem chamada é designada quando a cobertura do local presente é baixa demais para exibir confiança suficiente de que uma "sem chamada de variante" é precisa. Portanto, a chamada entrante é mudada para uma "sem chamada", tal como para não fornecer indicação de qualquer correlação da sequência de amostra a qualquer sequência de referência. Chamadas de referência abaixo de um certo ponto de corte de cobertura são filtradas para evitar declarar nenhuma variante devido a uma leitura que porta variante em baixa profundidade.

[0110] Em 406, o um ou mais processadores determinam se a variante atual corresponde a uma variante anterior. Por exemplo, a variante atual pode ser comparada a uma coleção de variantes (no banco de dados anterior 782 na Figura 7) que armazena informações a respeito de variantes comuns exibidas através de uma população. Por exemplo, informações de bancos de dados conhecidos (por exemplo, 1000G phase 3 e banco de dados Cosmic) podem ser usados para aumentar a confiança em chamadas de variante que se sabe que ocorrem na população. Quando a variante atual é similar a uma variante anterior no banco de dados de variantes anteriores, o fluxo se move ao longo da ramificação 412, em que uma série de testes escalonados com base em hierarquia são realizados. Em 406, quando nenhuma variante anterior é identificada a partir de um banco de dados, o fluxo se move ao longo da ramificação 414, em que uma série diferente de testes escalonados com base em hierarquia é realizada. Um ou mais dos testes escalonados podem ser realizados ao longo de cada uma das ramificações 412 e 414. Ademais, os vários testes podem ser realizados em ordens alternativas e em combinações alternativas além da disposição particular ilustrada na Figura 4. Cada um dos testes na Figura 4 fornece um grau/nível correspondente de evidência secundária de apoio de que uma variante ocorreu ou não, sendo que os testes exibem graus de confiança diferentes como evidência secundária de apoio.

[0111] Seguindo a ramificação 414, em 416, o um ou mais processadores determinam se uma classificação de qualidade de variante para a variante atual excede um limiar de classificação de qualidade. A classificação de qualidade de variante é determinada de acordo com modalidades no presente documento, tal como descrito em conexão com as Figuras 1C e 1D. Como explicado no presente documento, a classificação de qualidade de variante representa uma relação entre uma frequência de variante de amostra e múltiplas frequências de variante de linha de base. Quando a classificação de qualidade de variante é menor do que o limiar de classificação de qualidade, o fluxo se move para 418 onde a variante atual é invalidada, tal como designada como uma variante falsa (por exemplo, declarar que a chamada representar um "negativo" devido a uma classificação de qualidade baixa). Quando a classificação de qualidade de variante é igual ou maior do que o limiar de classificação de qualidade, o fluxo se move para 420.

[0112] Em 420, o um ou mais processadores obtêm e revisam uma contagem de fragmentos duplex (por exemplo, remendado ou não remendado duplex) que suportam uma chamada de variante. A contagem indica um número de fragmentos brutos duplex que foram identificados pelo aplicativo de chamada de variante para obter uma variante de apoio que corresponde à variante potencial na leitura de amostra. Durante o sequenciamento, ambas as cadeias de DNA são analisadas e leituras são obtidas para fragmentos de cada cadeia. A título de exemplo, pelo menos primeira e segunda leituras podem ser obtidas para sobrepor fragmentos da primeira cadeia, enquanto terceira e quarta leituras são obtidas para os fragmentos da segunda cadeia. O conjunto de 1 a 4 leituras é revisado para determinar quantas das leituras indicam uma variante na posição/local atual.

[0113] Quando uma variante está presente em uma leitura de amostra, seria esperado que todas as leituras para o local atual em cada uma da primeira e da segunda cadeias exibisse uma variante relacionada ou "de apoio". Por exemplo, quando a leitura de amostra corresponde a uma primeira cadeia e é chamada para uma variante potencial (por exemplo, um local exibe um "A"), seria esperado que o

local correspondente na segunda cadeia (correspondente ao fragmento bruto) fosse um “T”. Quando o local correspondente na segunda cadeia (fragmento bruto) inclui uma variante que é um “T”, a leitura atual do fragmento bruto seria contada como tendo uma variante de apoio. Caso uma variante de apoio de correlação ocorra em cada uma das cadeias (leitura de fragmento de amostra e leitura de fragmento bruto) dentro de um DNA de cadeia dupla, isso pode ser considerado um bom indicador de que uma variante ocorreu na posição designada na leitura de amostra.

[0114] Quando a análise de sequenciamento (pelo aplicativo de chamada de variante) produz uma contagem alta de leituras de fragmento bruto que exibem uma variante de apoio (como denotado por “DS>=Limiar” no bloco em 420), o fluxo se move para 422. Em 422, a chamada entrante é validada, a saber, passada como uma variante ou designada como uma chamada de variante positiva. Quando a análise de sequenciamento produz uma contagem baixa de leituras de fragmento bruto que exibem uma variante de apoio, essa circunstância é interpretada como uma indicação de que um erro ocorreu na chamada de variante potencial e que o aplicativo de validação de chamada de variante indica um nível de confiança baixo de que a chamada entrante representa uma chamada de variante. Consequentemente, o fluxo se move para 424 onde a chamada de variante entrante é invalidada, a saber, designada como uma “passagem negativa”.

[0115] Os testes de variante em 416 e 420 representam testes que exibem um alto nível de confiança (em relação à confiança em outros tipos de teste) nos resultados a partir dos mesmos. É reconhecido que testes adicionais ou alternativos podem ser aplicados, bem como inverter a ordem dos testes em 416 e 420.

[0116] Voltando para 406, a seguir a árvore de decisão com base em hierarquia é descrita em conexão com a ramificação 412. Quando o fluxo se move ao longo da ramificação 412, em 426, o um ou mais processadores revisam a classificação de qualidade de variante atribuída para a chamada de variante entrante. Quando a classificação de qualidade de variante é menor do que um limiar de classificação de qualidade predeterminado, o fluxo se move para 428, em que a chamada de variante

entrante é declarada uma variante falsa. Por exemplo, passagem é negada para a chamada de variante entrante ou a mesma é designada como uma “negativa” (devido à baixa-qualidade). Quando a classificação de qualidade de variante é maior ou igual ao limiar de classificação de qualidade predeterminado, o fluxo se move para 430. O limiar de classificação de qualidade utilizado em 426 pode ser igual ou diferente do limiar de classificação de qualidade utilizado em 416.

[0117] Em 430, o um ou mais processadores determinam uma classificação ponderada (WS) associada com as contagens para um ou mais tipos de fragmentos brutos. Por exemplo, o aplicativo de chamada de variante pode emitir contagens para múltiplos tipos de fragmento bruto diferentes, tais como fragmento remendado duplex, fragmento remendado simples, fragmento não remendado duplex e fragmento não remendado simples. As contagens indicam o número de leituras dos fragmentos brutos, do tipo correspondente, que incluíram uma variante de apoio. Em 430, os processadores aplicam pesos às contagens e somam o resultado para obter uma classificação ponderada que combina os tipos de fragmentos brutos diferentes. A soma ponderada é comparada a um limiar de classificação ponderada (WS). Quando a soma das classificações ponderadas excedem o limiar de WS, o fluxo avança para 432. Caso contrário, o fluxo se move para 434 onde a chamada de variante entrante não é confirmada e declarada inválida, denotada como uma “passagem negativa”.

[0118] Em 432, o um ou mais processadores realizam uma contagem de soma dos fragmentos brutos que validaram a chamada de variante entrante. Por exemplo, os processadores podem somar as contagens que indicam o número de leituras dos fragmentos brutos, a partir do tipo correspondente, que incluía uma variante de apoio. A contagem de soma é comparada a um limiar de contagem de soma (SC). Quando a contagem de soma cai abaixo do limiar de SC, o fluxo se move para 436. Caso contrário, o fluxo se move para 438, onde a chamada de variante entrante é validada, a saber, declarada uma “passagem positiva”.

[0119] Em 436, o um ou mais processadores revisam a contagem do fragmento (ou fragmentos) remendado duplex que inclui variantes de apoio (similar ao processo

em 420) para determinar se o fragmento (ou fragmentos) remendado duplex autentica/confirma a chamada de variante entrante. A determinação de variante associada com o fragmento (ou fragmentos) remendado duplex é comparada a um limiar remendado duplex (DS). Quando a contagem de variantes de apoio de DS cai abaixo do limiar, o fluxo se move para 440 com a chamada de variante entrante invalidada, a saber, declarada uma "passagem negativa". Quando a contagem de fragmentos brutos remendados duplex iguala ou excede o limiar de DS, em 436, o fluxo se move para 442 onde a chamada de variante entrante é validada, a saber, declarada como representando uma "passagem positiva".

[0120] É reconhecido que a árvore de decisão com base em hierarquia supramencionada representa um exemplo de uma maneira na qual os fatores relacionados à variante podem ser analisados. Por exemplo, em certos casos, os fatores relacionados à variante podem ser rearrumados para serem considerados em uma ordem diferente dentro da árvore de decisão. Em conexão com A Figura 4, várias decisões são feitas com base em informações relativas a fragmentos brutos (por exemplo, fragmentos brutos remendados duplex, classificações ponderadas para fragmentos brutos, contagens de fragmento bruto de variante e similares). Além disso, a classificação de qualidade de variante é usada como outros critérios filtro. A árvore de decisão com base em hierarquia usa as informações supramencionadas para melhorar sensibilidade e especificidade de chamada de variante, tal como com dados de ctDNA. A árvore de decisão com base em hierarquia é usada para filtrar primeiras variantes chamadas por um chamador de variante. Na modalidade da Figura 4, apoio de fragmento ponderado é usado para ajudar a derivar confiança em chamadas de variante, desde que a chamada de variante entrante corresponda a uma variante anterior e a classificação de qualidade exceda um limiar de classificação de qualidade. A classificação dá pesos a tipos de fragmento bruto diferentes para cada variante, que incluem fragmentos duplex/simplex e remendados/não remendados. Os pesos podem ser treinados a partir de dados com resultados conhecidos. Uma chamada de variante de passagem tem que ter uma classificação Q de variante suficientemente alta

calculada a partir de comparação à linha de base normal (consultar seção anterior) bem como uma classificação ponderada suficiente. Entretanto, é reconhecido que o apoio de fragmento ponderado pode ser realizado em um ponto diferente na árvore de decisão, ou omitido inteiramente.

[0121] Na modalidade da Figura 4, a árvore de decisão com base em hierarquia inclui pelo menos três camadas de testes de fragmento de apoio que são aplicadas antes de validar uma chamada de variante entrante. Por exemplo, antes de validar uma chamada de variante entrante em 422, testes de fragmento bruto proporcionam resultados positivos em três camadas, a saber, em 406, 416 e 420. Antes de validar uma chamada de variante entrante em 438, testes de fragmento bruto proporcionam resultados positivos em quatro camadas, a saber, em 406, 426 430 e 432. Antes de validar uma chamada de variante entrante em 442, testes de fragmento bruto proporcionam resultados positivos em cinco camadas, a saber, em 406, 426, 430, 432 e 436. O número de camadas de testes que são aplicadas aos fragmentos brutos pode ser variado com base no grau de confiança proporcionado com qualquer teste particular.

[0122] A Figura 5A ilustra uma árvore de decisão com base em hierarquia implantada de acordo com uma modalidade alternativa no presente documento para filtrar chamadas de variante. A Figura 5A é um processo generalizado, uma implantação do qual é apresentada na Figura 4. Em 550, um ou mais processadores recebem uma indicação de uma chamada de variante potencial para uma sequência genômica de interesse. Em 552, o um ou mais processadores recebem dados de sequenciamento que incluem leituras para fragmentos brutos e de amostra de nucleotídeos junto com a sequência genômica de interesse. As leituras incluem leituras de amostra para o fragmento de amostra que correspondem a uma sequência de nucleotídeos na posição designada junto com a sequência genômica de interesse.

[0123] Em 554, o um ou mais processadores analisam uma frequência de variante de amostra na posição designada para a sequência genômica de interesse em relação a frequências de variante de linha de base na posição designada para uma sequência

genômica de linha de base para obter uma classificação de qualidade. Em seguida, o um ou mais processadores realizam um ou mais de múltiplos testes relativos a fragmentos brutos. No exemplo da Figura 5A, três determinações são ilustradas em 556, 558 e 560. Opcionalmente, as determinações em 556, 558 e 560 podem ser repetidas (como denotado na ramificação 564). Adicional ou alternativamente, as determinações em 556, 558 e 560 podem ser repetidas mais do que uma vez em conexão com informações diferentes, podem ser realizadas em qualquer ordem e podem ser proporcionados pesos diferentes em conexão com validação de uma chamada de variante potencial. Opcionalmente, uma ou mais das determinações em 556, 558 e 560 podem ser omitidas inteiramente.

[0124] Em 556, o um ou mais processadores determinam se os fragmentos brutos confirmam a chamada de variante potencial. Por exemplo, os fragmentos brutos podem corresponder a pelo menos um dentre um fragmento remendado duplex, um fragmento remendado simples, um fragmento não remendado duplex ou um fragmento não remendado simples. Como um exemplo, a determinação pode incluir identificar uma contagem de um número de fragmentos remendados duplex como o fragmento bruto que incluía variantes de apoio. Os processadores determinarão se a contagem de fragmentos remendados duplex que incluíam variantes de apoio excedem um limiar de DS confirmando, desse modo, a chamada de variante potencial. Opcionalmente, em 556, os processadores podem realizar a classificação ponderada duplex e determinações de contagem de soma como discutido acima em conexão com as operações em 420, 430, 432 e 436 na Figura 4.

[0125] Em 558, o um ou mais processadores determina se as leituras de amostra fornecem uma quantidade predeterminada de cobertura para o fragmento de amostra. A chamada de variante potencial pode ser declarada uma Sem chamada quando as leituras de amostra caem abaixo da quantidade predeterminada de cobertura.

[0126] Em 560, o um ou mais processadores determinam se a chamada de variante potencial confirma uma chamada de variante anterior exibida através de uma população predefinida. A correlação para uma variante anterior pode ser utilizada

como informações para apenas um ou mais limiares utilizados em outros testes. Por exemplo, quando uma correlação é determinada entre as chamadas de variante potencial e anterior, a operação de determinação realiza pelo menos uma dentre a operação 556 ou a operação 558 com a utilização de um primeiro limiar. Além disso, quando uma correlação não existe entre as chamadas de variante potencial e anterior, a operação de determinação realiza pelo menos uma dentre a operação 556 ou a operação 558 com a utilização de um segundo limiar.

[0127] Em 562, o um ou mais processadores validam a chamada de variante potencial com base na classificação de qualidade e na operação (ou operações) de determinação. Por exemplo, uma chamada de variante potencial pode ser validada quando a classificação de qualidade excede um limiar de classificação de qualidade e pelo menos uma dentre as determinações em 556 a 560 suporta a chamada de variante potencial. Adicional ou alternativamente, as determinações em 556 a 560 podem ser combinadas, tal como através de uma soma ponderada em que o peso de cada determinação pode ser igual ou diferente. A soma ponderada das determinações em 556 a 560 pode ser comparada a um limiar e a chamada de variante potencial validada apenas quando a soma ponderada excede o limiar. Opcionalmente, a importância e/ou peso proporcionado às determinações em 556 a 560 pode ser variada com base no nível da classificação de qualidade. Por exemplo, quando uma classificação de qualidade muito alta é determinada, para validar uma chamada de variante potencial, pode ser determinado que apenas uma das determinações em 556 a 560 precisa apoiar a chamada de variante potencial. Alternativamente, quando uma classificação de qualidade de nível médio é determinada, para uma ou mais das determinações em 556 a 560 pode ser necessário suportar uma chamada de variante potencial, antes de validar a chamada de variante potencial. Como outro exemplo, os limiares aplicados às determinações em 556 a 560 podem ser variados com base em um nível da classificação de qualidade.

[0128] A Figura 5B ilustra um processo realizado em conexão com a utilização de uma classificação ponderada (por exemplo, a operação em 430 na Figura 4 e/ou 556

na Figura 5A) de acordo com modalidades no presente documento. Em 570, o um ou mais processadores identificam uma contagem dos fragmentos brutos que incluem uma variante de apoio relacionada à chamada de variante potencial. Em 572, o um ou mais processadores obtêm uma classificação de fragmento ponderada para os fragmentos brutos que indicam a chamada de variante de apoio. Por exemplo, a contagem associada com os fragmentos brutos remendados duplex podem ser multiplicadas por um primeiro fator de ponderação, a contagem associada com os fragmentos brutos remendados simplex pode ser multiplicada por um segundo fator de ponderação, a contagem associada com os fragmentos não remendados duplex pode ser multiplicada por um terceiro fator de ponderação e a contagem associada com os fragmentos não remendados simplex pode ser multiplicada por um quarto fator de ponderação. O primeiro ao quarto fatores de ponderação podem, então, ser somados ou combinados de alguma outra maneira para obter uma classificação de fragmento ponderada. Adicional ou alternativamente, a ponderação de classificação de fragmento pode ser modificada com base no número de fragmentos brutos que não incluem variantes de apoio. Por exemplo, quando um grande número de fragmentos de apoio corresponde a uma sequência de referência, a contagem dos fragmentos brutos (que correspondem à sequência de referência) pode ser utilizada como um fator para diminuir a classificação de fragmento ponderada. Em 574, o um ou mais processadores comparam a classificação de fragmento ponderada a um limiar de fragmento bruto para determinar se confirmam a chamada de variante potencial.

ASPECTOS IMPLANTADOS POR COMPUTADOR

[0129] Como compreendido pelas pessoas de habilidade comum na técnica, os métodos e informações descritos no presente documento podem ser implantados, no todo ou em parte, como instruções executáveis por computador em mídias legíveis por computador conhecidas. Por exemplo, os métodos descritos no presente documento podem ser implantados em hardware. Alternativamente, o método pode ser implantado em software armazenado em, por exemplo, uma ou mais memórias ou outra mídia legível por computador e implantado em um ou mais processadores.

Como é conhecido, os processadores podem ser associados com um ou mais controladores, unidades de cálculo e/ou outras unidades de um sistema computador, ou implantadas em firmware como desejado. Caso implantadas em software, as rotinas podem ser armazenadas em qualquer memória legível por computador tal como em RAM, ROM, memória flash, um disco magnético, um disco laser, ou outra mídia de armazenamento, como também é conhecido. Da mesma forma, esse software pode ser entregue para um dispositivo de computação por meio de qualquer método de entrega conhecido que inclui, por exemplo, por um canal de comunicação tal como uma linha de telefone, a Internet, uma conexão sem fio, etc., ou por meio de uma mídia transportável, tal como um disco legível por computador, unidade de memória flash, etc.

[0130] De forma mais geral, e como compreendido pelas pessoas de habilidade comum na técnica, as várias etapas descritas acima podem ser implantadas como vários blocos, operações, ferramentas, módulos e técnicas que, por sua vez, podem ser implantados em hardware, firmware, software, ou qualquer combinação de hardware, firmware e/ou software. Quando implantados em hardware, alguns ou todos os blocos, operações, técnicas, etc. podem ser implantados em, por exemplo, um circuito integrado personalizado (IC), um circuito integrado de aplicação específica (ASIC), uma matriz lógica programável em campo (FPGA), uma matriz lógica programável (PLA), etc.

[0131] Quando implantado em software, o software pode ser armazenado em qualquer mídia legível por computador conhecida tal como em um disco magnético, um disco óptico, ou outra mídia de armazenamento, em uma RAM ou ROM ou memória flash de um computador, processador, controlador de disco rígido, controlador de disco óptico, controlador de fita, etc. Da mesma forma, o software pode ser entregue para um usuário ou um sistema de computação por meio de qualquer método de entrega conhecido que inclui, por exemplo, em um disco legível por computador ou outro mecanismo de armazenamento de computador transportável.

[0132] A Figura 6 ilustra um diagrama de blocos de um ambiente de sistema de

computação 600 formado de acordo com uma modalidade no presente documento. O ambiente de sistema de computação 600 é apenas um exemplo de um ambiente de computação adequado e não é destinado a sugerir qualquer limitação ao escopo de uso ou funcionalidade no presente documento. Tampouco o ambiente de computação 600 deve ser interpretado como tendo qualquer dependência ou exigência relacionada a qualquer um ou combinação de componentes ilustrados no ambiente operacional exemplificativo 600. Os métodos e sistemas são operacionais com inúmeros ambientes ou configurações de sistemas de computação de propósito geral ou propósito especial diferentes. Exemplos de sistemas, ambientes, e/ou configurações de computação que podem ser adequados para usar com os métodos ou sistema das reivindicações incluem, porém, sem limitações, computadores pessoais, computadores servidores, dispositivos portáteis ou do tipo laptop, sistemas multiprocessadores, sistemas com base em microprocessador, set top boxes, produtos eletrônicos de consumo programáveis, PCs de rede, minicomputadores, computadores do tipo mainframe, ambientes computação distribuídos que incluem qualquer dos sistemas ou dispositivos acima, e similares.

[0133] As operações dos métodos e sistemas podem ser descritas no contexto geral de instruções executáveis por computador, tal como módulos de programa, que são executados por um computador. De modo geral, módulos de programa incluem rotinas, programas, objetos, componentes, estruturas de dados, etc. que realizam tarefas particulares ou implantam tipos de dados abstratos particulares. Os métodos e aparelho também podem ser praticados em ambientes de computação distribuídos em que tarefas são realizadas por dispositivos de processamento remotos que são ligados através de uma rede de comunicações. Em ambientes de computação tanto integrados quanto distribuídos, módulos de programa podem estar localizados em mídias de armazenamento de computador tanto locais quanto remotas que incluem dispositivos de armazenamento em memória.

[0134] Em referência à Figura 6, componentes de computador 610 podem incluir, porém, sem limitações, uma unidade de processamento 620, uma memória de

sistema 630 e um barramento de sistema 621 que acopla vários componentes de sistema que incluem a memória de sistema à unidade de processamento 620. O barramento de sistema 621 pode ser qualquer de diversos tipos de barramento estruturas que incluem um barramento de memória ou controlador de memória, um barramento de periféricos e um barramento local com o uso de qualquer um dentre uma variedade de arquiteturas de barramento. A título de exemplo, e não limitação, tais arquiteturas incluem barramento Arquitetura Padrão de Indústria (EUA), barramento de Arquitetura de Canal Micro (MCA), barramento ISA Aprimorado (EISA), barramento local de Associação de Padrões de Componentes Eletrônicos de Vídeo (VESA) e barramento de Interconexão de Componente Periférico (PCI) também conhecido como barramento Mezanino.

[0135] O computador 610 inclui, tipicamente, uma variedade de mídias legíveis por computador. As mídias legíveis por computador podem ser quaisquer mídias disponíveis que possam ser acessadas por computador 610 e incluem tanto mídias voláteis quanto não voláteis, mídias removíveis e não removíveis. A título de exemplo, e não limitação, mídias legíveis por computador podem compreender mídias de armazenamento de computador e mídias de comunicação. As mídias de armazenamento computador incluem tanto mídias voláteis quanto não voláteis, removíveis quanto não removíveis implantadas em qualquer método ou tecnologia para armazenamento de informações tal como instruções legíveis por computador, estruturas de dados, módulos de programa ou outros dados. As mídias de armazenamento computador incluem, porém, sem limitações, memória RAM, ROM, EEPROM, flash ou outra tecnologia de memória, CD-ROM, discos versáteis digitais (DVD) ou outro armazenamento de disco óptico, cassetes magnéticos, fita magnética, armazenamento de disco magnético ou outros dispositivos de armazenamento magnéticos, ou qualquer outra mídia que possa ser usada para armazenar as informações desejadas e que possa ser acessada pelo computador 610. As mídias de comunicação, tipicamente, incorporam instruções legíveis por computador, estruturas de dados, módulos de programa ou outros dados em um sinal de dados modulado tal

como uma onda portadora ou outro mecanismo de transporte e incluem quaisquer mídias de entrega de informações. O termo "sinal de dados modulado" significa um sinal que tem uma ou mais de suas características definidas ou mudadas de uma maneira tal que codifique informações no sinal. A título de exemplo, e não limitação, mídias de comunicação incluem mídias por fio tais como uma rede por fio ou conexão direta por fio, e mídias sem fio tais como mídias acústicas, RF, infravermelhas e outras mídias sem fio. Combinações de quaisquer das acima também devem ser incluídas dentro do escopo de mídias legíveis por computador.

[0136] A memória de sistema 630 inclui mídias de armazenamento de computador na forma de memória volátil e/ou não volátil tal como memória apenas de leitura (ROM) 631 e memória de acesso aleatório (RAM) 632. Um sistema de entrada/saída básico 633 (BIOS), que contém as rotinas básicas que ajudam a transferir informações entre elementos dentro do computador 610, tal como durante inicialização, é armazenado, tipicamente, na ROM 631. A RAM 632 tipicamente contém dados e/ou módulos de programa que são acessíveis imediatamente para e/ou que atualmente são operados pela unidade de processamento 620. A título de exemplo, e não limitação, a Figura 6 ilustra sistema operacional 634, programas aplicativos 635, outros módulos de programa 636 e dados de programa 637.

[0137] O computador 610 também pode incluir outras mídias de armazenamento de computador removíveis/não removíveis, voláteis/não voláteis. A título de exemplo apenas, a Figura 6 ilustra um controlador de disco rígido 640 que lê de ou grava em mídias magnéticas não removíveis, não voláteis, um controlador de disco magnético 651 que lê de ou grava em um disco magnético removível não volátil 652, e um controlador de disco óptico 655 que lê de ou grava em um disco óptico removível não volátil 656 tal como um CD ROM ou outras mídias ópticas. Outras mídias de armazenamento de computador removíveis/não removíveis, voláteis/não voláteis que podem ser usadas no ambiente operacional exemplificativo incluem, porém, sem limitações, cassetes de fita magnética, cartões de memória flash, discos versáteis digitais, fita de vídeo digital, RAM de estado sólido, ROM de estado sólido, e similares.

O controlador de disco rígido 641 é tipicamente conectado ao barramento de sistema 621 através de uma interface de memória não removível tal como a interface 640, e controlador de disco magnético 651 e controlador de disco óptico 655 são, tipicamente, conectados ao barramento de sistema 621 por uma interface de memória removível, tal como a interface 650.

[0138] Os controladores e as mídias de armazenamento de computador associadas discutidos acima e ilustrados na Figura 6, fornecem armazenamento de instruções legíveis por computador, estruturas de dados, módulos de programa e outros dados para o computador 610. Na Figura 6, por exemplo, o controlador de disco rígido 641 é ilustrado como armazenando o sistema operacional 644, programas aplicativos 645, outros módulos de programa 646, e dados de programa 647. Note-se que esses componentes podem ser os mesmos que ou diferentes de sistema operacional 634, programas aplicativos 635, outros módulos de programa 636 e dados de programa 637. O sistema operacional 644, programas aplicativos 645, outros módulos de programa 646 e dados de programa 647 recebem números diferentes aqui para ilustrar que, no mínimo, os mesmos são cópias diferentes. Um usuário pode inserir comandos e informações no computador através de dispositivos de entrada tais como um teclado 662 e dispositivo de apontamento 661, comumente denominado um mouse, trackball ou painel sensível ao toque. Outros dispositivos de entrada (não mostrados) podem incluir um microfone, joystick, controle de jogos, antena parabólica, digitalizador ou similares. Esses e outros dispositivos de entrada são frequentemente conectados à unidade de processamento 620 através de uma interface de entrada de usuário 660 que é acoplada ao barramento de sistema, mas pode ser conectada por outras estruturas de interface e barramento, tais como uma porta paralela, porta de jogos ou um barramento serial universal (USB). Um monitor 691 ou outro tipo de dispositivo de exibição também é conectado ao barramento de sistema 621 por meio de uma interface, tal como uma interface de vídeo 690. Além do monitor, computadores também podem incluir outros dispositivos de saída periféricos tais como alto-falantes 697 e impressora 696, que podem ser conectados através de uma

interface periférica de saída 690.

[0139] O computador 610 pode operar em um ambiente em rede com o uso de conexões lógicas para um ou mais computadores remotos, tal como um computador remoto 680. O computador remoto 680 pode ser um computador pessoal, um servidor, um roteador, um PC de rede, um dispositivo par ou outro nó de rede comum, e tipicamente inclui muitos ou todos os elementos descritos acima em relação ao computador 610, embora apenas um dispositivo de armazenamento de memória 681 foi ilustrado na Figura 6. As conexões lógicas retratadas na Figura 6 incluem uma rede de área local (LAN) 671 e uma rede de área ampla (WAN) 673, mas também pode incluir outras redes. Tais ambientes de rede são lugar comum em escritórios, redes de computadores para toda a empresa, intranets e a Internet.

[0140] Quando usado em um ambiente de rede LAN, o computador 610 é conectado à LAN 671 através de uma interface ou adaptador de rede 670. Quando usado em um ambiente de rede WAN, o computador 610 tipicamente inclui um modem 672 ou outros meios para estabelecer comunicações pela WAN 673, tal como a Internet. O modem 672, que pode ser interno ou externo, pode ser conectado ao barramento de sistema 621 por meio da interface de entrada de usuário 660, ou outro mecanismo apropriado. Em um ambiente em rede, módulos de programa retratados em relação ao computador 610, ou porções dos mesmos, podem ser armazenados no dispositivo de armazenamento de memória remoto. A título de exemplo, e não limitação, a Figura 6 ilustra programas aplicativos remotos 685 como residindo no dispositivo de memória 681. Será reconhecido que as conexões de rede mostradas são exemplificativas e outros meios de estabelecer um enlace de comunicações entre os computadores podem ser usados.

[0141] Embora o sistema e método de avaliação de risco, e outros elementos, tenham sido descritos como sendo preferencialmente implantados em software, os mesmos podem ser implantados em hardware, firmware, etc., e podem ser implantados por qualquer outro processador. Assim, os elementos descritos no presente documento podem ser implantados em uma de CPU de múltiplos fins padrão

ou em hardware ou firmware projetados especificamente tais como um circuito integrado de aplicação específica (ASIC) ou outro dispositivo conectado por cabo, como desejado, que incluem, porém, sem limitações, o computador 610 da Figura 6. Quando implantado em software, a rotina de software pode ser armazenada em qualquer memória legível por computador tal como em um disco magnético, um disco laser, ou outra mídia de armazenamento, em uma RAM ou ROM de um computador ou processador, em qualquer banco de dados, etc. Da mesma forma, esse software pode ser entregue para um usuário ou um sistema de diagnóstico por meio de qualquer método de entrega conhecido ou desejado que inclui, por exemplo, em um disco legível por computador ou outro mecanismo de armazenamento de computador transportável ou por um canal de comunicação tal como uma linha de telefone, a internet, comunicação sem fio, etc. (os qual é visto como sendo o mesmo que ou intercambiável com fornecer tal software por meio de uma mídia de armazenamento transportável). Assim, muitas modificações e variações podem ser feitas nas técnicas e estruturas descritas e ilustradas no presente documento sem que se afaste do espírito e escopo da presente invenção. Assim, deve ser compreendido que os métodos e aparelho descritos no presente documento são apenas ilustrativos e não são limitantes ao escopo da invenção.

[0142] A Figura 7 ilustra um diagrama de blocos de um sistema alternativo formado de acordo com uma modalidade no presente documento. Componentes mostrados em contorno tracejado não são tecnicamente parte do computador 710, mas são usados para ilustrar a modalidade exemplificativa da Figura 7. Componentes de computador 710 podem incluir, porém, sem limitações, um processador 720, uma memória de sistema 730, uma interface de memória/gráfica 721, também conhecida como um chip Northbridge, e uma interface de I/O 722, também conhecida como um chip Southbridge. A memória de sistema 730 e um processador gráfico 790 podem ser acoplados à interface de memória/gráfica 721. Um monitor 791 ou outro dispositivo de saída gráfico pode ser acoplado ao processador gráfico 790.

[0143] Uma série de barramentos de sistema pode acoplar vários componentes

de sistema que incluem um barramento de sistema de alta velocidade 723 entre o processador 720, a interface de memória/gráfica 721 e a interface de I/O 722, um barramento frontal 724 entre a interface de memória/gráfica 721 e a memória de sistema 730, e um barramento de processamento gráfico avançado (AGP) 725 entre a interface de memória/gráfica 721 e o processador gráfico 790. O barramento de sistema 723 pode ser qualquer de diversos tipos de estruturas de barramento que incluem, a título de exemplo, e não limitação, arquiteturas que incluem barramento de Arquitetura Padrão de Indústria (EUA), barramento de Arquitetura de Canal Micro (MCA), barramento ISA Aprimorado (EISA). Como arquiteturas de sistema evoluem, outras arquiteturas de barramento e chipsets podem ser usadas, mas frequentemente, em geral, seguem esse padrão. Por exemplo, companhias tais como Intel e AMD suportam a Intel Hub Architecture (IHA) e a arquitetura Hypertransport™, respectivamente.

[0144] O computador 710 inclui, tipicamente, uma variedade de mídias legíveis por computador. Mídias legíveis por computador podem ser quaisquer mídias disponíveis que posam ser acessadas pelo computador 710 e incluem mídias tanto voláteis e não voláteis, mídias removíveis e não removíveis. A título de exemplo, e não limitação, mídias legíveis por computador podem compreender mídias de armazenamento de computador. As mídias de armazenamento computador incluem tanto mídias voláteis quanto não voláteis, removíveis quanto não removíveis implantadas em qualquer método ou tecnologia para armazenamento de informações tal como instruções legíveis por computador, estruturas de dados, módulos de programa ou outros dados. As mídias de armazenamento computador incluem, porém, sem limitações, memória RAM, ROM, EEPROM, flash ou outra tecnologia de memória, CD-ROM, discos versáteis digitais (DVD) ou outro armazenamento de disco óptico, cassetes magnéticos, fita magnética, armazenamento de disco magnético ou outros dispositivos de armazenamento magnéticos, ou qualquer outra mídia física que possa ser usada para armazenar as informações desejadas e que possa ser acessada pelo computador 710.

[0145] A memória de sistema 730 inclui mídias de armazenamento de computador na forma de memória volátil e/ou não volátil tal como memória apenas de leitura (ROM) 731 e memória de acesso aleatório (RAM) 732. A ROM de sistema 731 pode conter dados de sistema permanentes 743, tal como informações de identificação e fabricação. Em algumas modalidades, um sistema de entrada/saída básico (BIOS) também pode ser armazenado na ROM do sistema 731. A RAM 732 tipicamente contém dados e/ou módulos de programa que são acessíveis imediatamente para e/ou que atualmente são operados pelo processador 720. A título de exemplo, e não limitação, a Figura 5 ilustra sistema operacional 734, programas aplicativos 735, outros módulos de programa 736 e dados de programa 737.

[0146] A interface de I/O 722 pode acoplar o barramento de sistema 723 com um número de outros barramentos 726, 727 e 728 que acoplam uma variedade de dispositivos internos e externos ao computador 710. Um barramento de interface periférica serial (SPI) 726 pode conectar a uma memória de sistema de entrada/saída básico (BIOS) 733 que contém as rotinas básicas que ajudam a transferir informações entre elementos dentro do computador 710, tal como durante a inicialização.

[0147] Um superchip de entrada/saída 760 pode ser usado para conectar a um número de periféricos “legados”, tais como disquete 752, teclado/mouse 762 e impressora 796, como exemplos. O superchip de I/O 760 pode ser conectado à interface de I/O 722 com um barramento 727, tal como um barramento de pinagem reduzida (Low Pin Count, LPC), em algumas modalidades. Várias modalidades do superchip de I/O 760 são amplamente disponíveis no mercado comercial. Em uma modalidade, o barramento 728 pode ser um barramento de Interconexão de Componente Periférico (PCI), ou uma variação do mesmo, pode ser usada para conectar periféricos de velocidade superior à interface de I/O 722. Um barramento PCI também pode ser conhecido como um barramento Mezanino. Variações do barramento PCI incluem os barramentos de Interconexão de Componente Periférico-Expressa (PCI-E) e o Interconexão de Componente Periférico-Estendida (PCI-X), sendo que o anterior tem uma interface serial e o último é uma interface paralela

compatível para trás. Em outras modalidades, o barramento 728 pode ser um barramento de fixação de tecnologia avançada (ATA), na forma de um barramento ATA serial (SATA) ou ATA paralelo (PATA).

[0148] O computador 710 também pode incluir outras mídias de armazenamento de computador removíveis/não removíveis, voláteis/não voláteis. A título de exemplo apenas, a Figura 7 ilustra um controlador de disco rígido 740 que lê de ou grava para mídias magnéticas não voláteis não removíveis. O controlador de disco rígido 740 pode ser um controlador de disco rígido convencional. Mídias removíveis, tais como uma memória de barramento serial universal (USB) 753, firewire (IEEE 7394), ou controlador de CD/DVD 756 podem ser conectadas ao barramento PCI 728 diretamente ou através de uma interface 750. Uma mídia de armazenamento 754 pode ser acoplada através da interface 750. Outras mídias de armazenamento de computador removíveis/não removíveis, voláteis/não voláteis que podem ser usadas no ambiente operacional exemplificativo incluem, porém, sem limitações, cassetes de fita magnética, cartões de memória flash, discos versáteis digitais, fita de vídeo digital, RAM de estado sólido, ROM de estado sólido, e similares.

[0149] Os controladores e suas mídias de armazenamento de computador associadas discutidos acima e ilustrados na Figura 7, fornecem armazenamento de instruções legíveis por computador, estruturas de dados, módulos de programa e outros dados para o computador 710. Na Figura 7, por exemplo, o controlador de disco rígido 740 é ilustrado como armazenando o sistema operacional 744, programas aplicativos 745, outros módulos de programa 746, e dados de programa 747. Note-se que esses componentes podem ser os mesmos que ou diferentes de sistema operacional 734, programas aplicativos 735, outros módulos de programa 736 e dados de programa 737. O sistema operacional 744, programas aplicativos 745, outros módulos de programa 746 e dados de programa 747 recebem números diferentes aqui para ilustrar que, no mínimo, os mesmos são cópias diferentes. Um usuário pode inserir comandos e informações no computador através de dispositivos de entrada tais como um mouse/teclado 762 ou outra combinação de dispositivo de entrada. Outros

dispositivos de entrada (não mostrados) podem incluir um microfone, joystick, controle de jogos, antena parabólica, digitalizador ou similares. Esses e outros dispositivos de entrada são frequentemente conectados ao processador 720 através de um dos barramentos de interface de I/O, tais como o SPI 726, o LPC 727 ou o PCI 728, mas outros barramentos podem ser usados. Em algumas modalidades, outros dispositivos podem ser acoplados a portas paralelas, interfaces infravermelhas, portas de jogos e similares (não retratados), por meio do superchip de I/O 760. O computador 710 pode operar em um ambiente em rede com o uso de conexões lógicas para um ou mais computadores remotos, tal como um computador remoto 780 por meio de um controlador de interface de rede (NIC) 770. O computador remoto 780 pode ser um computador pessoal, um servidor, um roteador, um PC de rede, um dispositivo par ou outro nó de rede comum, e tipicamente inclui muitos ou todos os elementos descritos acima em relação ao computador 710. A lógica conexão entre o NIC 770 e o computador remoto 780 retratada na Figura 7 pode incluir uma rede de área local (LAN), uma rede de área ampla (WAN), ou ambas, mas também pode incluir outras redes. Tais ambientes de rede são lugar comum em escritórios, redes de computadores para toda a empresa, intranets e a Internet. O computador remoto 780 também pode representar um servidor web que suporta sessões interativas com o computador 710, ou no caso específico de aplicações com base em localização, pode ser um servidor de localização ou um servidor de aplicativo. O computador remoto 780 pode representar um servidor que gerencia um banco de dados de sequência genômica de linha de base 784, um banco de dados anterior 782 e similares. Por exemplo, como explicado no presente documento, o banco de dados de sequência genômica de linha de base 784 é atualizado periodicamente com informações linha de base novas. O banco de dados 784 é acessado em conexão com a validação de chamadas de variante potenciais para amostras (por exemplo, para obter frequências de variante de linha de base em posições designadas). Como outro exemplo, o banco de dados anterior 782 também pode ser acessado em conexão com a validação de chamadas de variante potenciais para amostras (por exemplo, para determinar se

uma chamada de variante potencial corresponde a uma variante anterior).

[0150] Em algumas modalidades, a interface de rede pode usar um modem (não retratado) quando uma conexão de banda larga não estiver disponível ou não for usada. Será reconhecido que a conexão de rede mostrada é exemplificativa e outros meios de estabelecer um enlace de comunicações entre os computadores podem ser usados. Processadores (unidades de processamento) exemplificativos incluem todas as variedades de microprocessadores e outras unidades de processamento usadas em dispositivos de computação. Mídias legíveis por computador exemplificativas são descritas acima. Quando dois ou mais componentes do sistema envolvem um processador ou uma mídia legível por computador, pode ser criado, em geral, o sistema em que um único processador e/ou mídia legível por computador é dedicada a um único componente do sistema; ou em que duas ou mais funções compartilham um único processador e/ou compartilham uma única mídia legível por computador, de modo que o sistema contenha tão pouco quanto um processador e/ou uma mídia legível por computador. Em algumas variações, é vantajoso usar múltiplos processadores ou mídias, por exemplo, em que é conveniente ter componentes do sistema em localizações diferentes. Por exemplo, alguns componentes de um sistema podem estar localizados em um laboratório de testes dedicado ao laboratório ou análise de dados, enquanto que outros componentes, que incluem componentes (opcionais) para fornecer informações de entrada ou obter uma comunicação de saída, podem estar localizados em uma instalação de tratamento médico ou consultoria (por exemplo, consultório médico, clínica de saúde, HMO, farmacêutico, geneticista, hospital) e/ou na residência ou empresa do sujeito humano (paciente) para quem o serviço de teste é realizado.

[0151] O computador remoto 780 pode representar um servidor ou outro recurso rede que é acoplado comunicativamente a um banco de dados anterior 782 e que contém informações de população que correlacionam a presença ou ausência de variantes/alelos em uma população de seres humanos. Por exemplo, o um ou mais variantes/alelos incluem alelos mutantes que provocam, ou são indicativos, defeitos

de seleção. Em uma variação simples, o banco de dados anterior 782 contém dados relacionados à frequência que alelos selecionados foram observados em uma população de seres humanos, por exemplo com câncer de bexiga, e uma população de seres humanos, por exemplo livre de câncer de bexiga. Adicional ou alternativamente, o banco de dados anterior pode incluir dados similares em relação a dois ou mais alelos fornecendo, desse modo, uma referência útil se o sujeito humano tem qualquer um dos dois ou mais alelos. Adicional ou alternativamente, o banco de dados anterior pode incluir informações pessoais, médicas ou genéticas quantitativas adicionais sobre os indivíduos no banco de dados diagnosticados com ou sem uma condição. Tais informações incluem, porém, sem limitações, informações sobre parâmetros tais como idade, sexo, etnia, raça, histórico médico, peso, situação de diabetes, pressão sanguínea, histórico familiar de câncer de bexiga, histórico de tabagismo e uso de álcool em seres humanos e impacto do pelo menos um parâmetro em relação a um anterior. O anterior também pode incluir informações sobre outros fatores de risco genéticos para alelos.

[0152] Um banco de dados ou biblioteca de sequência genômica de linha de base 784 pode ser fornecido para armazenar sequências genômicas de linha de base que são utilizadas, de acordo com modalidades no presente documento, para identificar frequências de variante de linha de base e outras informações.

REIVINDICAÇÕES

1. Método implantado por computador para validar chamadas de variante, em que o método é **caracterizado por** compreender:

sob controle de um ou mais processadores, executar instruções de programa para,

receber dados de sequenciamento que incluem uma leitura simples que tem uma sequência de nucleotídeos correspondente junto com a sequência genômica de interesse;

receber uma indicação de uma chamada de variante potencial em uma posição designada dentro da sequência de nucleotídeos junto com a sequência genômica de interesse;

obter frequências de variante de linha de base na posição designada dentro de uma ou mais sequências genômicas de linha de base;

obter uma frequência de variante de amostra na posição designada para a sequência genômica de interesse;

analisar as frequências de variante de amostra e linha de base na posição designada para obter uma classificação de qualidade; e

validar a chamada de variante potencial para a sequência genômica de interesse com base na classificação de qualidade.

2. Método, de acordo com a reivindicação 1, **caracterizado por** a operação de análise incluir obter uma relação entre a frequência de variante de amostra e uma distribuição das frequências de variante de linha de base, a classificação de qualidade com base na relação.

3. Método, de acordo com a reivindicação 1, **caracterizado por** a operação de análise compreender indexar a frequência de variante de amostra em relação a uma distribuição das frequências de variante de linha de base.

4. Método, de acordo com a reivindicação 3, **caracterizado por** a relação ter como base um teste de soma de classificação de Wilcoxon não paramétrico.

5. Método, de acordo com a reivindicação 1, **caracterizado por** as frequências de variante de linha de base indicarem um grau de ruído de fundo em posições correspondentes junto com a sequência genômica de linha de base.

6. Método, de acordo com a reivindicação 1, **caracterizado por** a validação compreender ainda comparar a classificação de qualidade a um limiar; e declarar a chamada de variante potencial como sendo uma chamada de variante válida quando a classificação de qualidade excede o limiar.

7. Método, de acordo com a reivindicação 1, **caracterizado por** as frequências de variante de linha de base serem derivadas de múltiplas sequências genômicas de linha de base que são associadas a mais de um tipo de alelo.

8. Método, de acordo com a reivindicação 1, **caracterizado por** compreender ainda receber dados de sequenciamento que incluem uma pluralidade de leituras de referência de uma sequência de nucleotídeos junto com a sequência genômica de linha de base, e determinar as frequências de variante de linha de base para as leituras de referência nas posições designadas.

9. Método, de acordo com a reivindicação 8, **caracterizado por** a determinação das frequências de variante de linha de base compreender ainda:

a. receber os dados de sequenciamento das leituras de referência para um conjunto de posições dentro de uma janela de par de bases atual;

b. identificar uma frequência de variante candidata para uma ou mais posições no conjunto de posições dentro da janela de par de bases atual;

c. selecionar uma das frequências de variante candidatas como a frequência de variante de linha de base para uma posição designada dentro da leitura de referência; e

d. deslocar a janela de par de bases junto com a sequência genômica de linha de base e repetir as operações em a, b e c.

10. Método implantado por computador para validar chamadas de variante, em que o método é **caracterizado por** compreender:

sob controle de um ou mais processadores executar instruções de programa para,

receber uma indicação de uma chamada de variante potencial para uma sequência genômica de interesse;

receber dados de sequenciamento que inclui leituras para fragmentos brutos e de amostra de nucleotídeos junto com a sequência genômica de interesse, em que as leituras incluem leituras de amostra para o fragmento de amostra que corresponde a uma sequência de nucleotídeos na posição designada junto com a sequência genômica de interesse;

analisar uma frequência de variante de amostra na posição designada para a sequência genômica de interesse em relação às frequências de variante de linha de base na posição designada para uma sequência genômica de linha de base para obter uma classificação de qualidade; e

determinar pelo menos um dentre os seguintes:

A) se os fragmentos brutos confirmarem a chamada de variante potencial;

B) se as leituras de amostra fornecerem uma quantidade predeterminada de cobertura para o fragmento de amostra; e

C) se a chamada de variante potencial corresponder a uma chamada de variante anterior exibida através de uma população predefinida; e

validar a chamada de variante potencial com base na classificação de qualidade e na operação de determinação.

11. Método, de acordo com a reivindicação 10, **caracterizado por** a determinação incluir operação A) e operação A) compreende:

identificar os fragmentos brutos que indicam uma chamada de variante de suporte;

obter uma classificação de fragmento ponderada para os fragmentos brutos que indicam a chamada de variante de suporte; e

comparar a classificação de fragmento ponderada a um limiar de fragmento bruto para determinar se confirma a chamada de variante potencial.

12. Método, de acordo com a reivindicação 10, **caracterizado por** a determinação incluir operação A) e em que os fragmentos brutos correspondem a pelo menos um dentre um fragmento remendado duplex, um fragmento remendado simples, um fragmento não remendado duplex ou um fragmento não remendado simples.

13. Método, de acordo com a reivindicação 10, **caracterizado por** a determinação incluir operação A) e operação A) compreende: identificar um fragmento remendado duplex como o fragmento bruto; e determinar se o fragmento remendado duplex indica uma chamada de variante de suporte para determinar se confirma a chamada de variante potencial.

14. Método, de acordo com a reivindicação 10, **caracterizado por** a determinação incluir operação B) e em que a chamada de variante potencial é declarada uma Sem chamada quando as leituras de amostra caem para abaixo da quantidade predeterminada de cobertura.

15. Método, de acordo com a reivindicação 10, **caracterização por** a determinação incluir operação C) e:

i. quando uma correlação é determinada entre as chamadas de variante potenciais e anteriores, a operação de determinação realiza pelo menos uma dentre operação A) ou operação B) que utiliza um primeiro limiar; e

ii. quando uma correlação não existe entre as chamadas de variante potenciais e anteriores, a operação de determinação realiza pelo menos uma dentre operação A) ou operação B) que utiliza um segundo limiar.

16. Método, de acordo com a reivindicação 10, **caracterizado por** a operação de análise e operações de determinação A), B) e C) serem realizadas em uma hierarquia.

17. Sistema para validar chamadas de variante, em que o sistema é **caracterizado por** compreender:

memória que armazena instruções de programa e dados de sequenciamento que incluem uma leitura simples que tem uma sequência de nucleotídeos correspondente junto com a sequência genômica de interesse;

um ou mais processadores que, quando executam as instruções de programa:

recebem uma indicação de uma chamada de variante potencial em uma posição designada dentro da sequência de nucleotídeos junto com a sequência genômica de interesse;

obtêm frequências de variante de linha de base na posição designada dentro de uma ou mais sequências genômicas de linha de base;

obtêm uma frequência de variante de amostra na posição designada para a sequência genômica de interesse;

analisam as frequências de variante de amostra e linha de base na posição designada para obter uma classificação de qualidade; e

validam a chamada de variante potencial para a sequência genômica de interesse com base na classificação de qualidade.

18. Sistema, de acordo com a reivindicação 17, **caracterizado por** o um ou mais processadores obterem uma relação entre a frequência de variante de amostra e uma distribuição das frequências de variante de linha de base, a classificação de qualidade com base na relação.

19. Sistema, de acordo com a reivindicação 17, **caracterizado por** o um ou mais processadores indexarem a frequência de variante de amostra em relação a uma distribuição das frequências de variante de linha de base.

20. Sistema, de acordo com a reivindicação 17, **caracterizado por** as frequências de variante de linha de base indicarem um grau de ruído de fundo em um local correspondente junto com a sequência genômica de linha de base.

21. Sistema, de acordo com a reivindicação 17, **caracterizado por** o um ou mais processadores compararem a classificação de qualidade a um limiar; e

emitirem uma indicação que declara a chamada de variante potencial como sendo inválida com base na comparação.

22. Sistema, de acordo com a reivindicação 17, **caracterizado por** compreender ainda memória que armazena dados de sequenciamento que incluem uma pluralidade de leituras de referência de nucleotídeos junto com a sequência genômica de linha de base, o um ou mais processadores para determinar as frequências de variante de linha de base para as leituras de referência na posição designada.

23. Sistema para validar chamadas de variante, em que o sistema é **caracterizado por** compreender:

memória que armazena instruções de programa e dados de sequenciamento que incluem leituras para fragmentos brutos e de amostra de nucleotídeos junto com a sequência genômica de interesse, em que as leituras incluem leituras de amostra para o fragmento de amostra que corresponde a uma sequência de nucleotídeos na posição designada junto com a sequência genômica de interesse;

um ou mais processadores que, quando executam as instruções de programa:

recebem uma indicação de uma chamada de variante potencial para uma sequência genômica de interesse;

analisam uma frequência de variante de amostra na posição designada para a sequência genômica de interesse em relação às frequências de variante de linha de base na posição designada para uma sequência genômica de linha de base para obter uma classificação de qualidade; e

determinam pelo menos um dentre os seguintes:

A) se os fragmentos brutos confirmarem a chamada de variante potencial;

B) se as leituras de amostra fornecerem uma quantidade predeterminada de cobertura para o fragmento de amostra; e

C) se a chamada de variante potencial corresponder a uma chamada de variante anterior exibida através de uma população predefinida; e

validar a chamada de variante potencial com base na classificação de qualidade e na operação de determinação.

24. Sistema, de acordo com a reivindicação 23, **caracterizado por** a determinação incluir operação A) e operação A) compreende:

identificar os fragmentos brutos que indicam uma chamada de variante de suporte;

obter uma classificação de fragmento ponderada para os fragmentos brutos que indicam a chamada de variante de suporte; e

comparar a classificação de fragmento ponderada a um limiar de fragmento bruto para determinar se confirma a chamada de variante potencial.

25. Sistema, de acordo com a reivindicação 23, **caracterizado por** a determinação incluir operação A) e em que os fragmentos brutos correspondem a pelo menos um dentre um fragmento remendado duplex, um fragmento remendado simples, um fragmento não remendado duplex ou um fragmento não remendado simples.

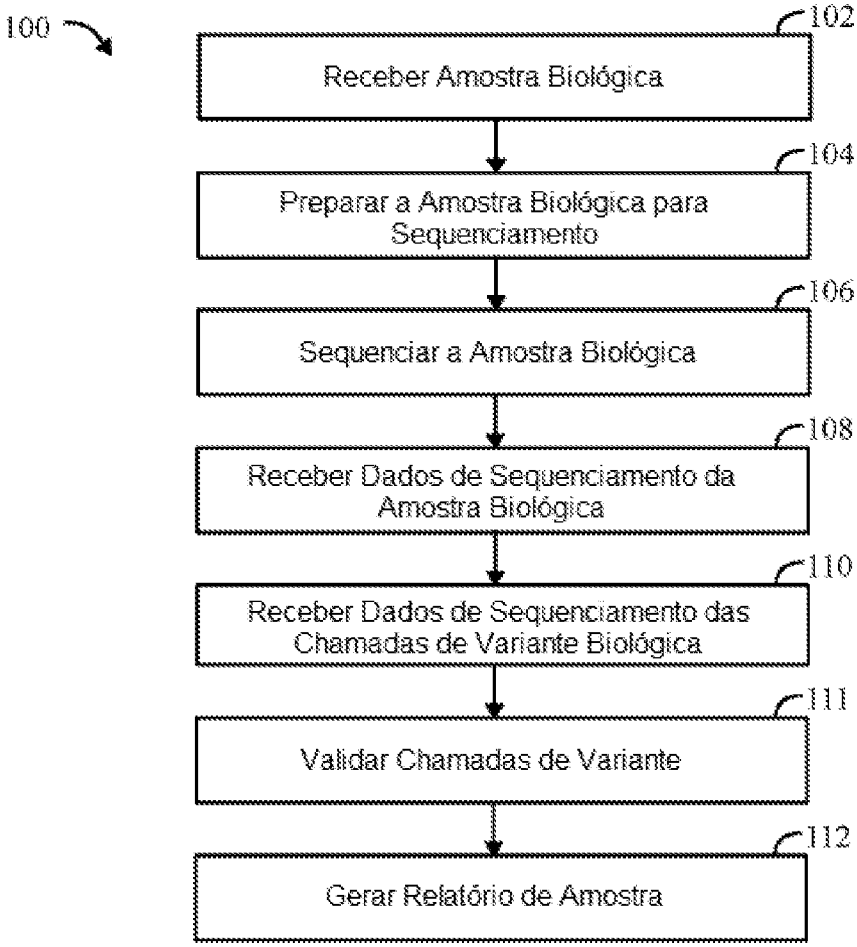

100
102
Receber Amostra Biológica
104
Preparar a Amostra Biológica para Sequenciamento
106
Sequenciar a Amostra Biológica
108
Receber Dados de Sequenciamento da Amostra Biológica
110
Receber Dados de Sequenciamento das Chamadas de Variante Biológica
111
Validar Chamadas de Variante
112
Gerar Relatório de Amostra

FIG. 1A

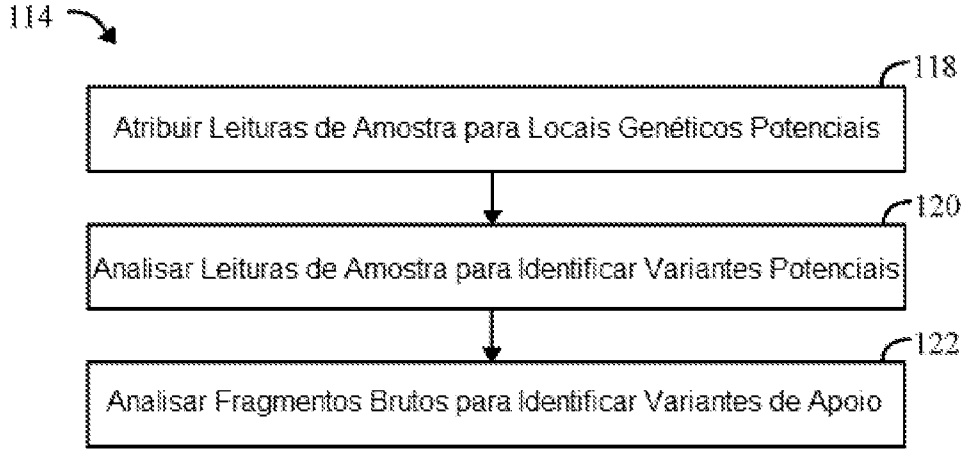

114
118
Atribuir Leituras de Amostra para Locais Genéticos Potenciais
120
Analisar Leituras de Amostra para Identificar Variantes Potenciais
122
Analisar Fragmentos Brutos para Identificar Variantes de Apoio

FIG. 1B

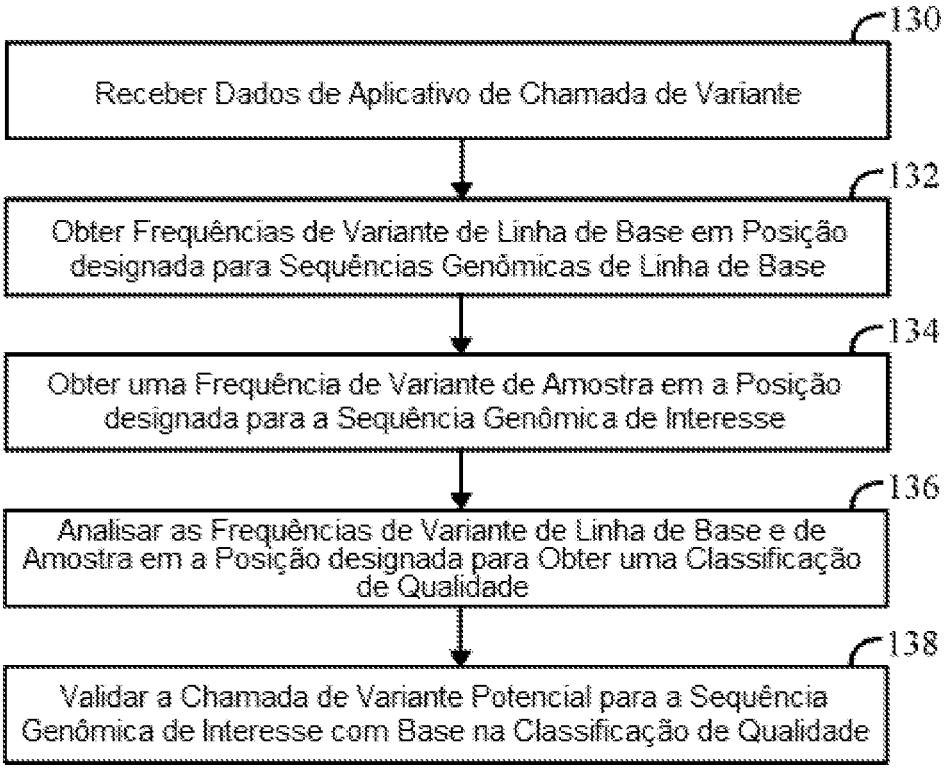

130
Receber Dados de Aplicativo de Chamada de Variante
132
Obter Frequências de Variante de Linha de Base em Posição designada para Sequências Genômicas de Linha de Base
134
Obter uma Frequência de Variante de Amostra em a Posição designada para a Sequência Genômica de Interesse
136
Analisar as Frequências de Variante de Linha de Base e de Amostra em a Posição designada para Obter uma Classificação de Qualidade
138
Validar a Chamada de Variante Potencial para a Sequência Genômica de Interesse com Base na Classificação de Qualidade

FIG. 1C

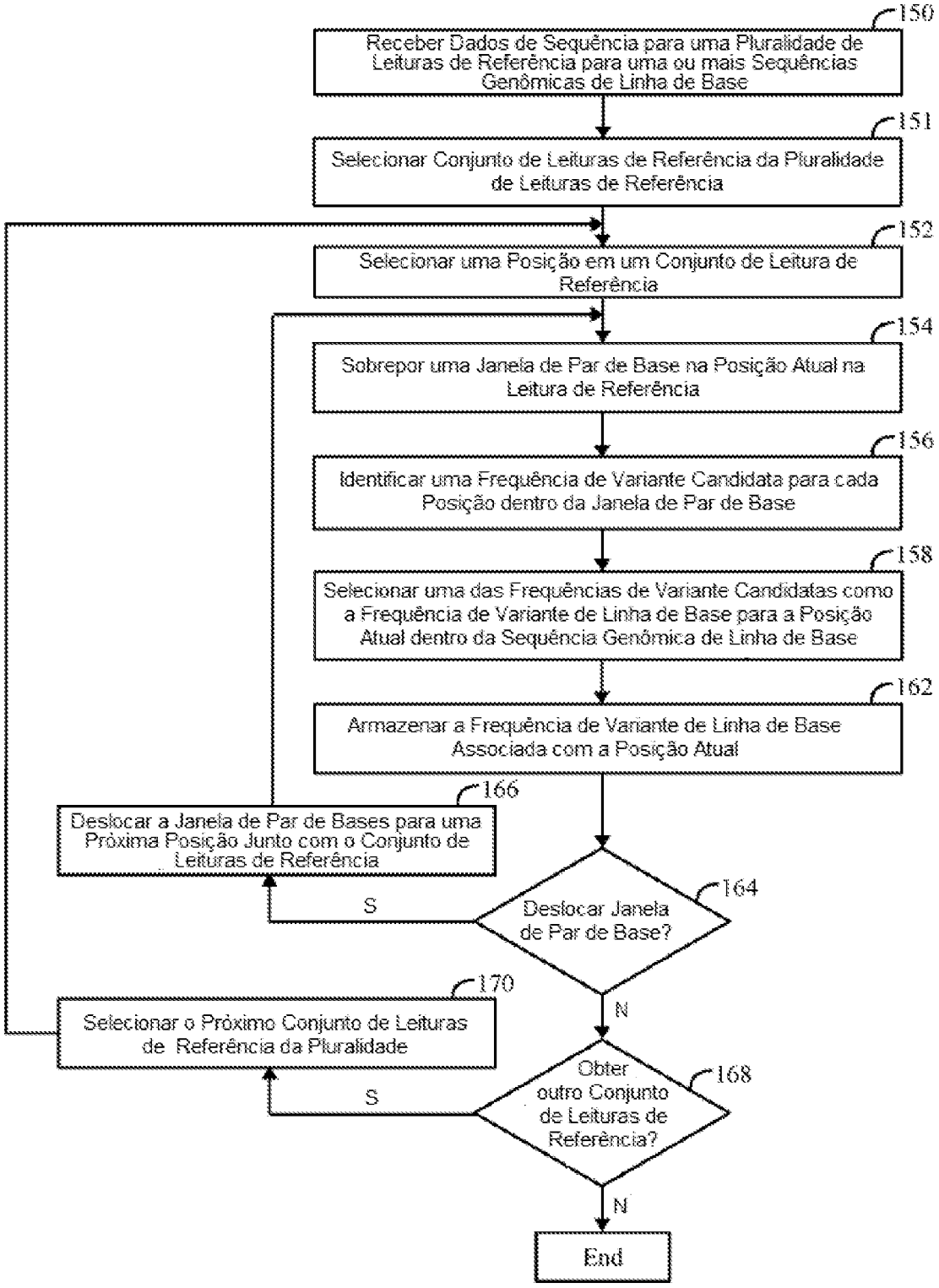

150
Receber Dados de Sequência para uma Pluralidade de Leituras de Referência para uma ou mais Sequências Genômicas de Linha de Base
151
Selecionar Conjunto de Leituras de Referência da Pluralidade de Leituras de Referência
152
Selecionar uma Posição em um Conjunto de Leitura de Referência
154
Sobrepor uma Janela de Par de Base na Posição Atual na Leitura de Referência
156
Identificar uma Frequência de Variante Candidata para cada Posição dentro da Janela de Par de Base
158
Selecionar uma das Frequências de Variante Candidatas como a Frequência de Variante de Linha de Base para a Posição Atual dentro da Sequência Genômica de Linha de Base
162
Armazenar a Frequência de Variante de Linha de Base Associada com a Posição Atual
166
Deslocar a Janela de Par de Bases para uma Próxima Posição Junto com o Conjunto de Leituras de Referência
S
164
Deslocar Janela de Par de Base?
N
170
Selecionar o Próximo Conjunto de Leituras de Referência da Pluralidade
S
168
Obter outro Conjunto de Leituras de Referência?
N
End

FIG. 1D

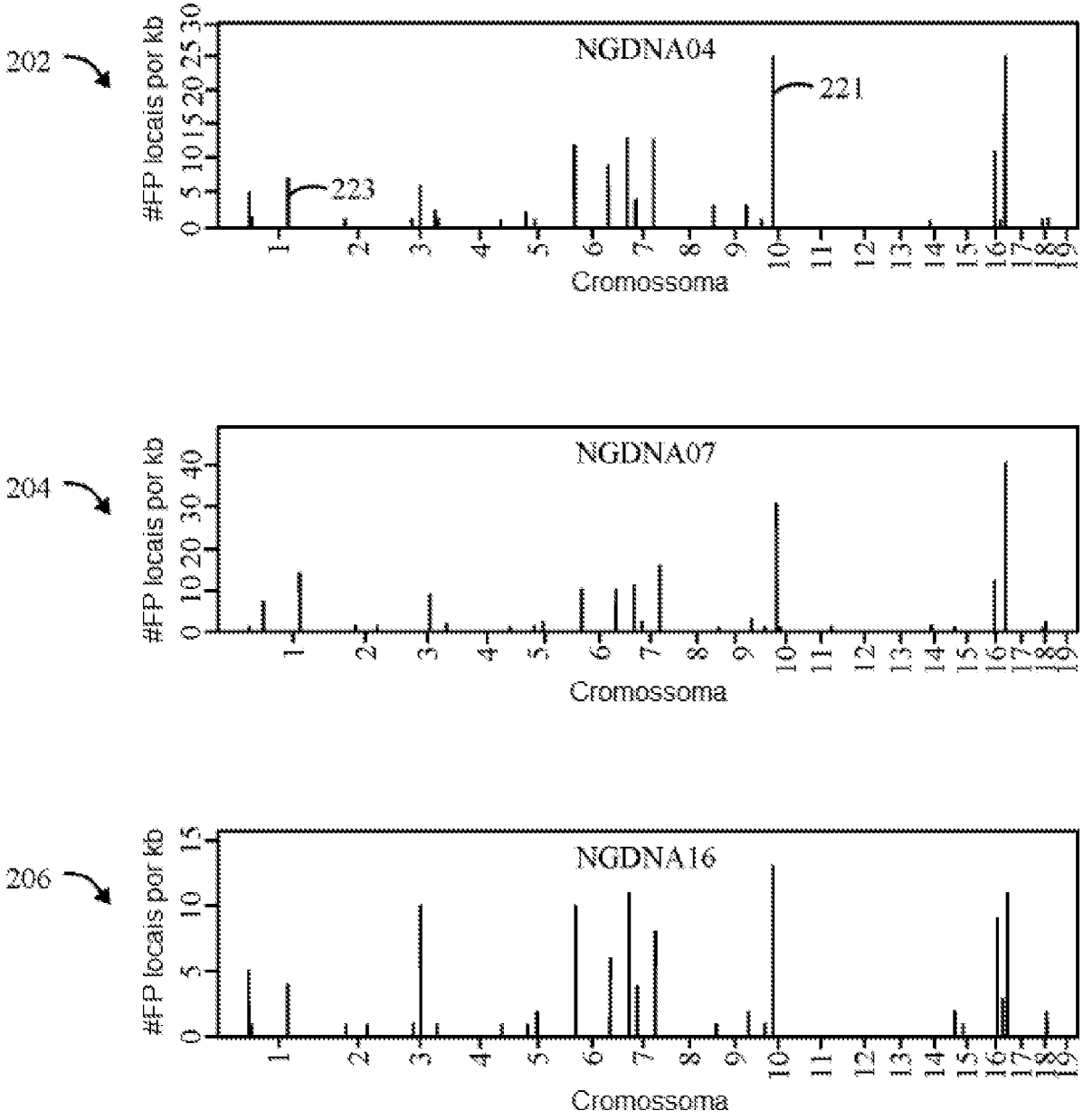

202
NGDNA04
#FP locais por kb
0 5 10 15 20 25 30
221
223
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
Cromossoma
204
NGDNA07
#FP locais por kb
0 10 20 30 40
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
Cromossoma
206
NGDNA16
#FP locais por kb
0 5 10 15
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
Cromossoma

FIG. 2A

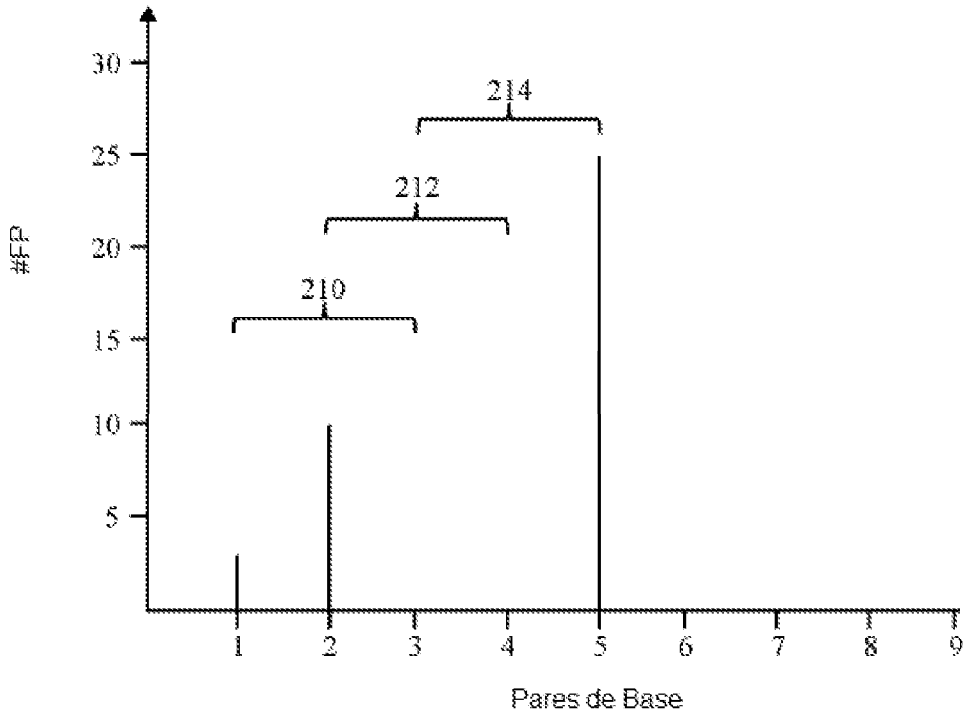

30
25
20
15
10
5
#FP
214
212
210
1
2
3
4
5
6
7
8
9
Pares de Base

FIG. 2B

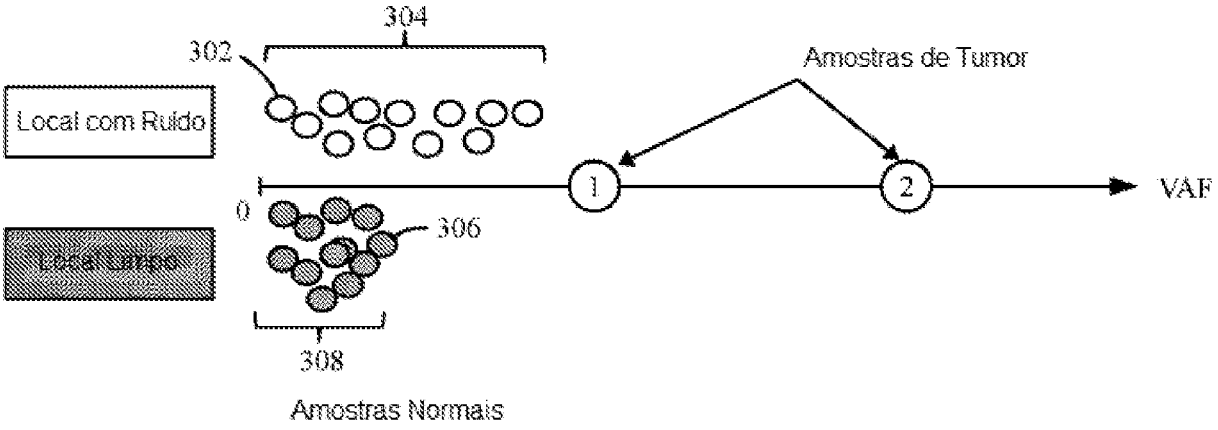

304
302
Local com Ruído
Local Limpo
Amostras de Tumor
1
2
VAF
0
306
308
Amostras Normais

FIG. 3

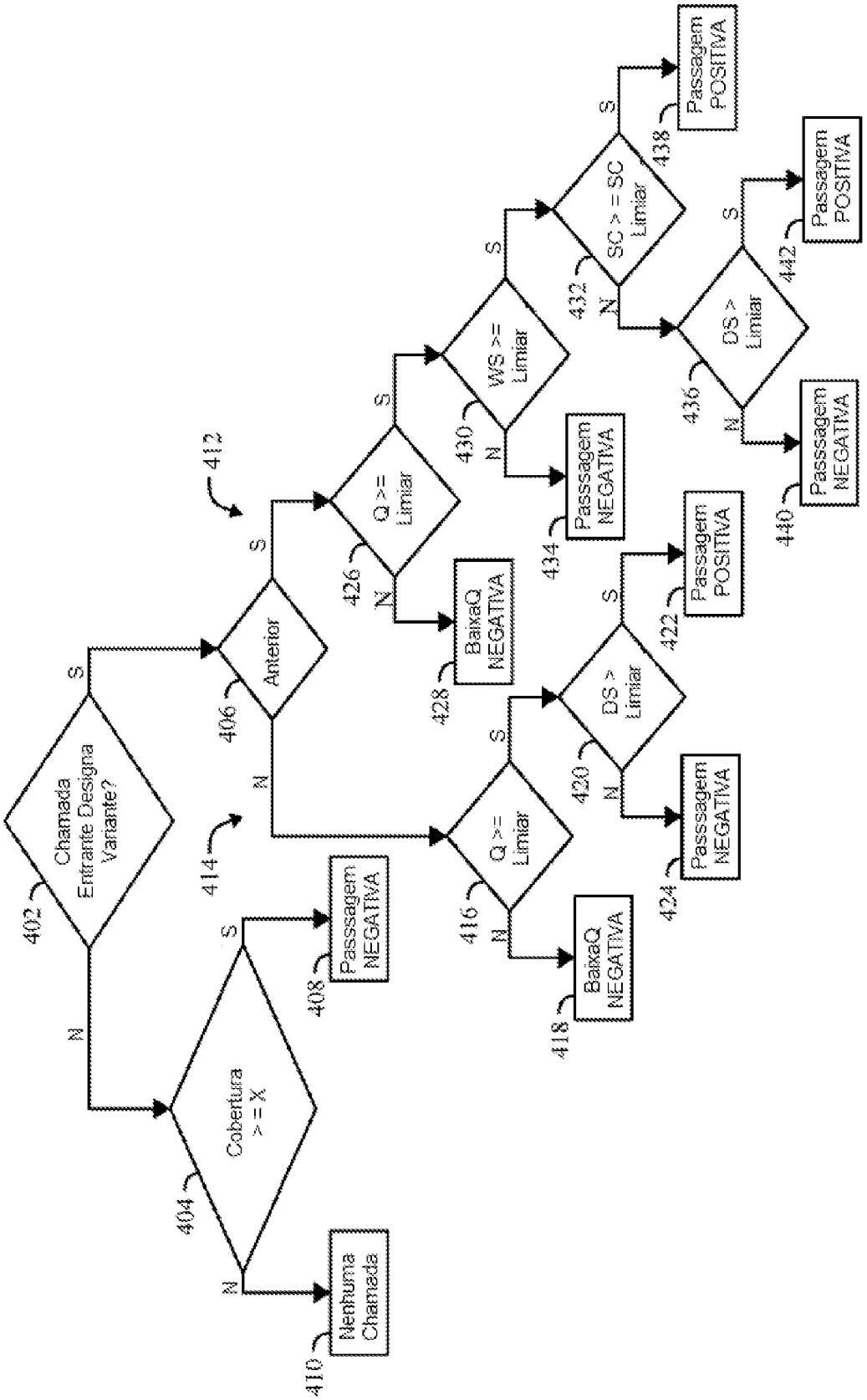

Chamada Entrante Designa Variante?
402
S
N
Cobertura >= X
404
S
N
Passagem NEGATIVA
408
Nenhuma Chamada
410
414
Anterior
406
S
N
412
Q >= Limiar
426
S
N
BaixaQ NEGATIVA
428
Q >= Limiar
416
S
N
BaixaQ NEGATIVA
418
WS >= Limiar
430
S
N
Passagem NEGATIVA
434
DS > Limiar
420
S
N
Passagem POSITIVA
422
Passagem MEGATIVA
424
SC >= SC Limiar
432
S
N
Passagem POSITIVA
438
DS > Limiar
436
S
N
Passagem POSITIVA
442
Passagem NEGATIVA
440

FIG. 4

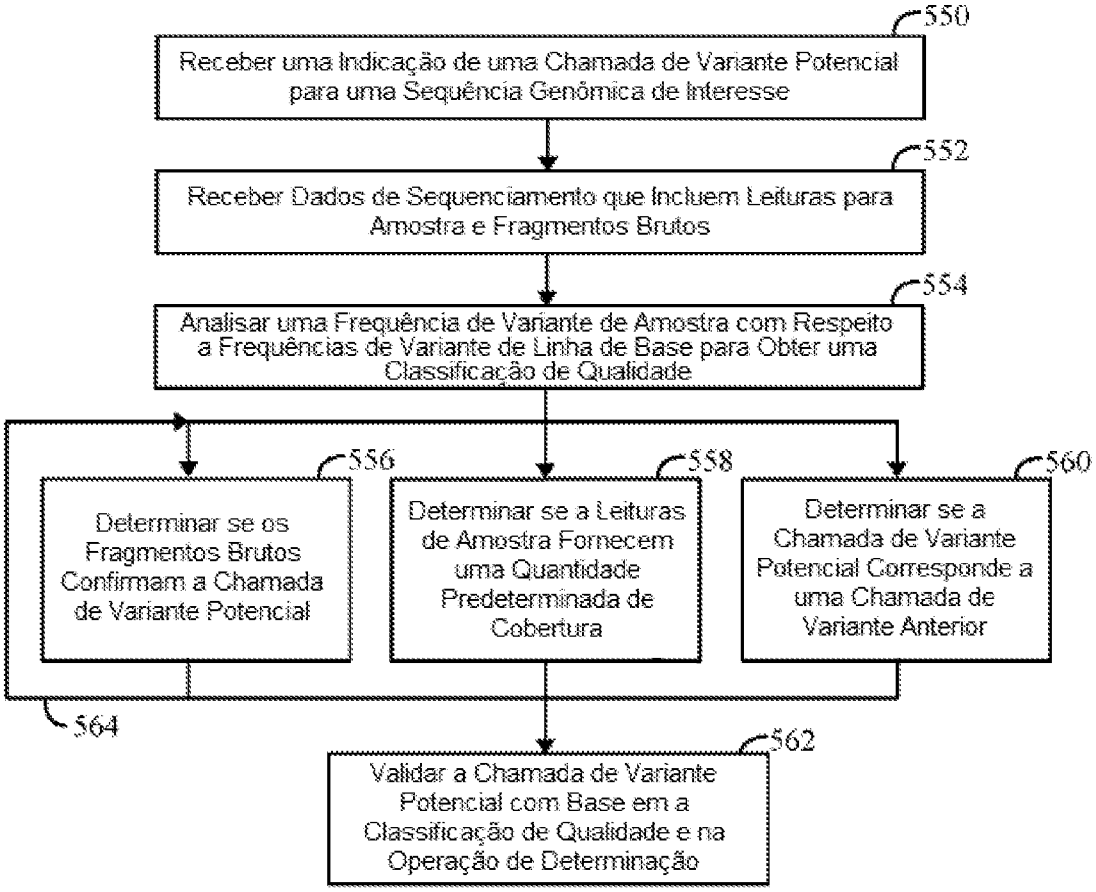

550
Receber uma Indicação de uma Chamada de Variante Potencial para uma Sequência Genômica de Interesse
552
Receber Dados de Sequenciamento que Incluem Leituras para Amostra e Fragmentos Brutos
554
Analisar uma Frequência de Variante de Amostra com Respeito a Frequências de Variante de Linha de Base para Obter uma Classificação de Qualidade
556
Determinar se os Fragmentos Brutos Confirmam a Chamada de Variante Potencial
558
Determinar se a Leituras de Amostra Fornecem uma Quantidade Predeterminada de Cobertura
560
Determinar se a Chamada de Variante Potencial Corresponde a uma Chamada de Variante Anterior
564
562
Validar a Chamada de Variante Potencial com Base em a Classificação de Qualidade e na Operação de Determinação

FIG. 5A

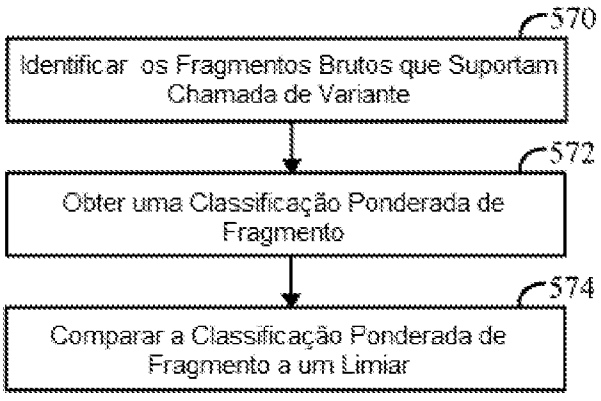

570
Identificar os Fragmentos Brutos que Suportam Chamada de Variante
572
Obter uma Classificação Ponderada de Fragmento
574
Comparar a Classificação Ponderada de Fragmento a um Limiar

FIG. 5B

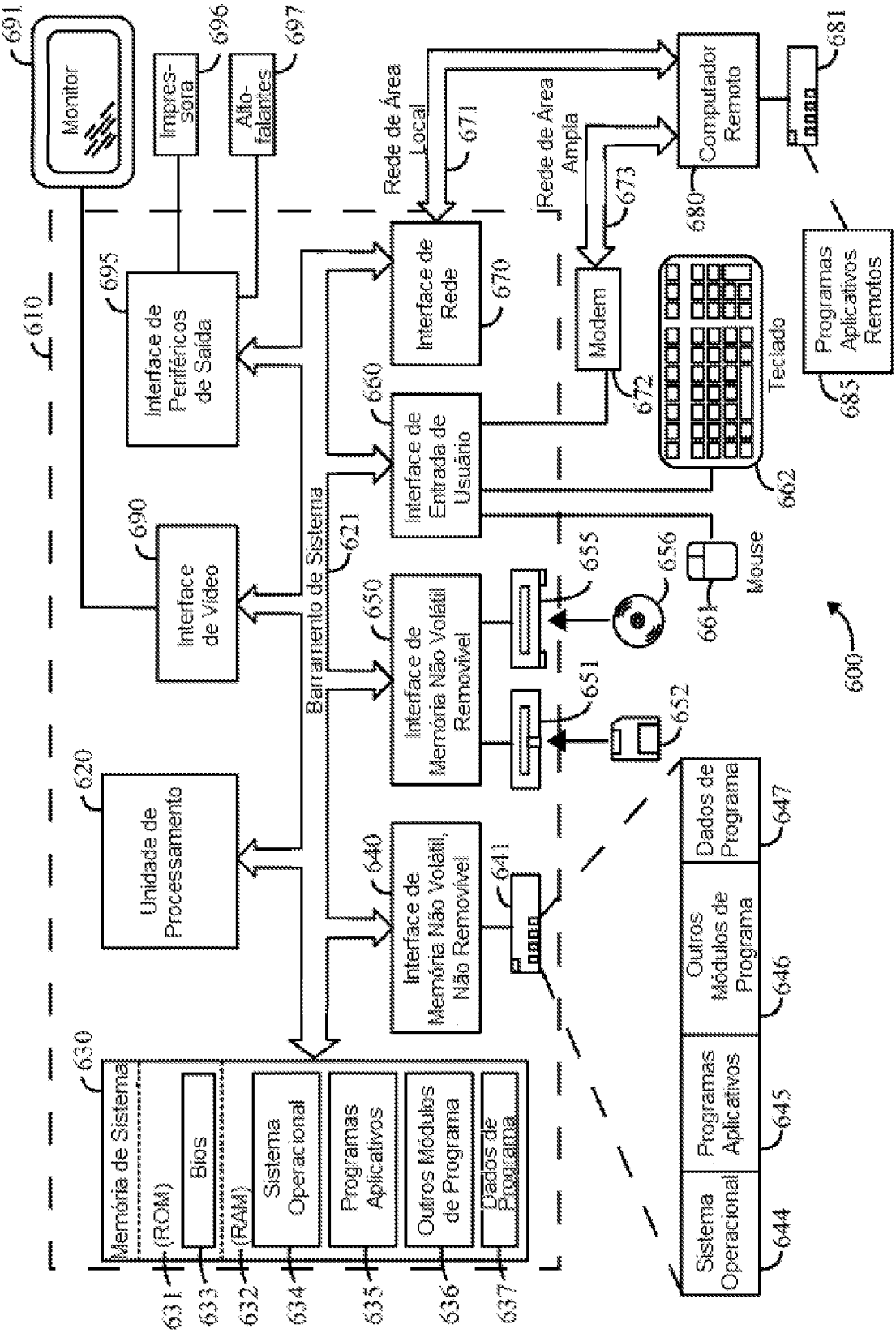

600
610
620
630
631
632
633
634
635
636
637
640
641
644
645
646
647
650
651
652
655
656
660
661
662
670
671
672
673
680
681
685
690
691
695
696
697
621
Monitor
Impressora
Alto-falantes
Interface de Periféricos de Saída
Interface de Vídeo
Unidade de Processamento
Memória de Sistema
(ROM)
Bios
(RAM)
Sistema Operacional
Programas Aplicativos
Outros Módulos de Programa
Dados de Programa
Interface de Memória Não Volátil, Não Removível
Interface de Memória Não Volátil Removível
Barramento de Sistema
Interface de Entrada de Usuário
Interface de Rede
Rede de Área Local
Rede de Área Ampla
Modem
Mouse
Teclado
Computador Remoto
Programas Aplicativos Remotos
Sistema Operacional
Programas Aplicativos
Outros Módulos de Programa
Dados de Programa

FIG. 6

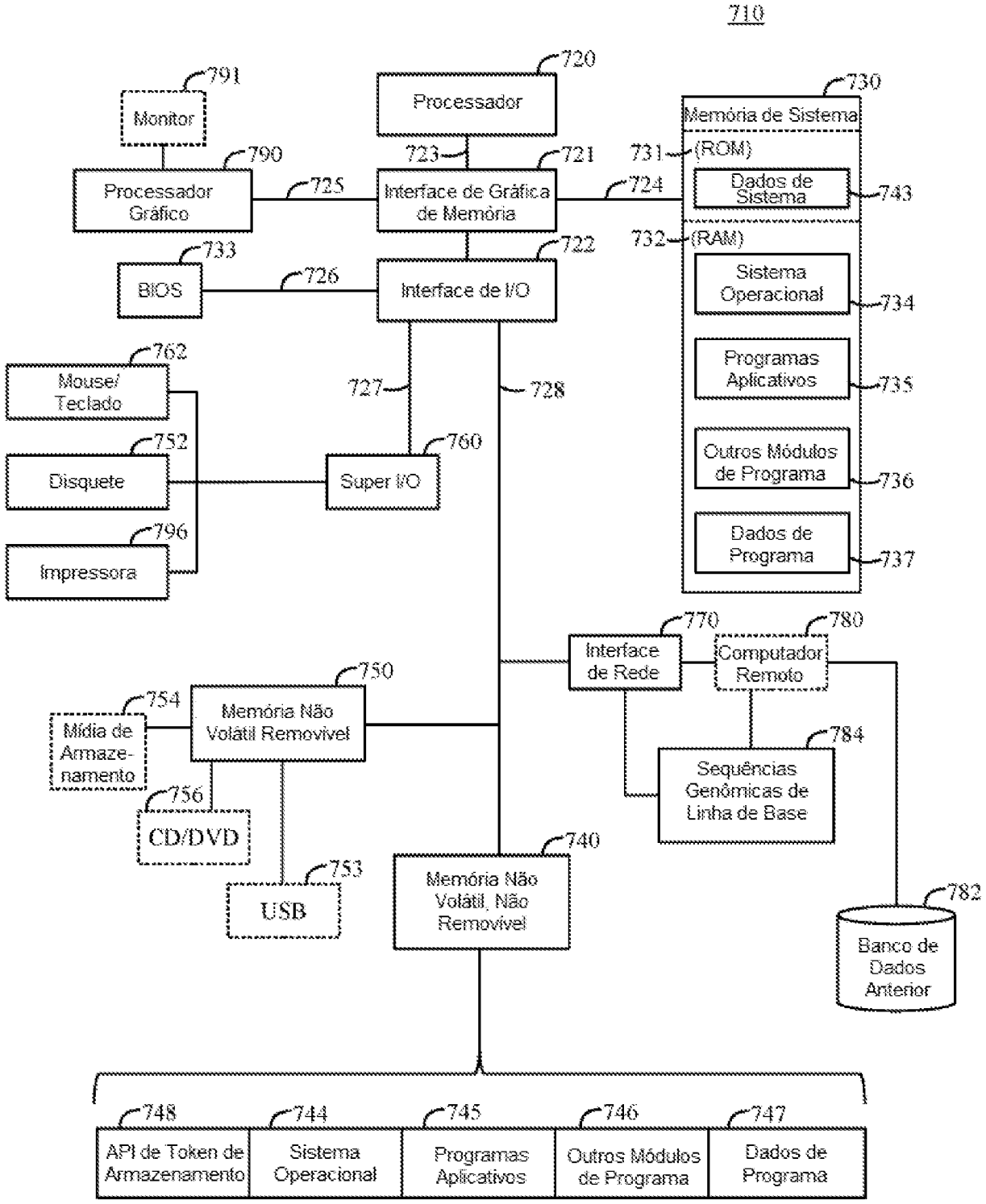

710
720
Processador
791
Monitor
730
Memória de Sistema
723
721
731
(ROM)
790
Processador Gráfico
725
Interface de Gráfica de Memória
724
Dados de Sistema
743
732
(RAM)
733
BIOS
726
722
Interface de I/O
Sistema Operacional
734
762
Mouse/ Teclado
Programas Aplicativos
735
727
728
752
Disquete
760
Super I/O
Outros Módulos de Programa
736
796
Impressora
Dados de Programa
737
770
780
Interface de Rede
Computador Remoto
750
754
Mídia de Armaze-namento
Memória Não Volátil Removível
784
Sequências Genômicas de Linha de Base
756
CD/DVD
740
753
USB
Memória Não Volátil, Não Removível
782
Banco de Dados Anterior
748
744
745
746
747
API de Token de Armazenamento
Sistema Operacional
Programas Aplicativos
Outros Módulos de Programa
Dados de Programa

FIG. 7

Resumo da Patente de Invenção para **"MÉTODO IMPLANTADO POR COMPUTADOR PARA VALIDAR CHAMADAS DE VARIANTE E SISTEMA PARA VALIDAR CHAMADAS DE VARIANTE"**

Trata-se de métodos e sistemas para validar chamadas de variante. Os dados de sequenciamento são recebidos para uma leitura simples junto com a sequência genômica de interesse. Uma indicação é recebida de uma chamada de variante potencial em uma posição designada dentro da sequência de nucleotídeos junto com a sequência genômica de interesse. Os métodos e sistemas obtêm frequências de variante de linha de base na posição designada dentro de uma ou mais sequências genômicas de linha de base, determinam uma frequência de variante de amostra na posição designada para a sequência genômica de interesse, analisam as frequências de variante de amostra e linha de base na posição designada para obter uma classificação de qualidade, e validam a chamada de variante potencial para a sequência genômica de interesse com base na classificação de qualidade.