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

(21) BR 112019013886-2 A2

(22) Data do Depósito: 05/01/2018

(43) Data da Publicação Nacional: 03/03/2020

(54) Título: MÉTODO PARA DETERMINAR OS VALORES DE COR CORRIGIDOS, E, SEQUENCIADOR DE ÁCIDOS NUCLEICOS

(51) Int. Cl.: G06F 19/20; C12Q 1/6869.

(30) Prioridade Unionista: 06/01/2017 US 62/443,294.

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

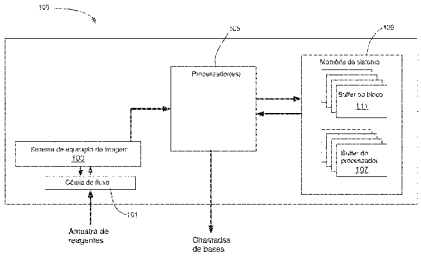
(72) Inventor(es): ROBERT LANGLOIS; PAUL BELITZ.

(86) Pedido PCT: PCT US2018012580 de 05/01/2018

(87) Publicação PCT: WO 2018/129314 de 12/07/2018

(85) Data da Fase Nacional: 04/07/2019

(57) Resumo: Métodos eficientes de memória determinam valores de cor corrigidos de dados de imagem adquiridos por um sequenciador de ácidos nucleicos durante um ciclo de chamada de bases. Tais métodos podem: (a) obter uma imagem de um substrato (por exemplo, de uma porção de uma célula de fluxo) incluindo uma pluralidade de sítios onde as bases de ácido nucleico são lidas; (b) medir valores de cor da pluralidade de sítios a partir da imagem do substrato; (c) armazenar os valores de cor em um buffer do processador do um ou mais processadores do sequenciador; (d) recuperar valores de cor parcialmente corrigidos na fase da pluralidade de sítios, onde os valores de cor parcialmente corrigidos na fase foram armazenados na memória do sequenciador durante um ciclo de chamada de bases imediatamente anterior; (e) determinar uma correção pré-faseamento; e (f) determinar os valores de cor corrigidos. Em várias implementações, essas operações são todas realizadas durante um único ciclo de chamada de bases. Em certas modalidades, os métodos adicionalmente incluem o uso dos valores de cor corrigidos para fazer chamadas de base para a pluralidade de sítios. Sequenciadores podem ser concebidos ou configurados para implementar tais métodos.



MÉTODO PARA DETERMINAR OS VALORES DE COR CORRIGIDOS, E, SEQUENCIADOR DE ÁCIDOS NUCLEICOS

REFERÊNCIA CRUZADA AOS PEDIDOS RELACIONADOS

[001] Este pedido reivindica os benefícios do Pedido de Patente Provisório Norte-Americano No. 62/443.294, depositado em 6 de janeiro de 2017 e intitulado "PHASING CORRECTION", o qual é aqui incorporado por referência em sua totalidade e para todos os efeitos.

ANTECEDENTES DA INVENÇÃO

[002] A revelação refere-se ao sequenciamento de ácidos nucleicos. Mais especificamente, a revelação refere-se a sistemas e métodos para sequenciamento em tempo real com correções de faseamento.

[003] Em um sítio particular em uma célula de fluxo ou em outro substrato, múltiplas cópias de uma molécula de ácido nucleico, todas tendo a mesma sequência (possivelmente com variações limitadas involuntariamente introduzidas por processamento de amostras), são analisadas ao mesmo tempo. São usadas cópias suficientes para garantir que um sinal suficiente seja produzido para permitir uma chamada de base confiável. A coleta de moléculas de ácido nucleico em um sítio é chamada de agrupamento.

[004] O faseamento representa um artefato não intencional que surge do sequenciamento de múltiplas moléculas de ácido nucleico dentro de um agrupamento. O faseamento é a taxa na qual sinais, como a fluorescência de moléculas únicas dentro de um agrupamento, perdem a sincronização entre si. Muitas vezes, o termo faseamento é reservado para contaminar o sinal de algumas moléculas que ficam para trás, e o termo pré-faseamento é usado para contaminar o sinal de outras moléculas que vão adiante. Juntos, o faseamento e o pré-faseamento descrevem o quão bem o aparelho de sequenciamento e a química estão se saindo.

SUMÁRIO

[005] Certos aspectos desta revelação referem-se a métodos para

determinar valores de cor corrigidos a partir de dados de imagem obtidos durante um ciclo de chamada de bases, onde o sequenciador inclui um sistema de aquisição de imagem, um ou mais processadores e memória. Tais métodos podem ser distinguidos pelas seguintes operações: (a) obter uma imagem de um substrato (por exemplo, de uma porção de uma célula de fluxo) incluindo uma pluralidade de sítios onde as bases de ácido nucleico são lidas; (b) medir valores de cor da pluralidade de sítios a partir da imagem do substrato; (c) armazenar os valores de cor em um buffer do processador do um ou mais processadores do sequenciador; (d) recuperar valores de cor parcialmente corrigidos na fase da pluralidade de sítios, onde os valores de cor parcialmente corrigidos na fase foram armazenados na memória do sequenciador durante um ciclo de chamada de bases imediatamente anterior; (e) determinar uma correção pré-faseamento; e (f) determinar os valores de cor corrigidos. Em várias implementações, essas operações são todas executadas durante um único ciclo de chamada de bases. Em certas modalidades, os métodos adicionalmente incluem o uso dos valores de cor corrigidos para fazer chamadas de base para a pluralidade de sítios.

[006] Durante o sequenciamento, os sítios exibem cores representando os tipos de base de ácido nucleico. Os valores de cor medidos e armazenados podem ser de intensidade ou outros valores de magnitude compreendidos em um determinado comprimento de onda ou faixa de comprimentos de onda. Em algumas implementações, os valores das cores são determinados a partir de apenas dois canais do sequenciador. Em algumas implementações, os valores das cores são obtidos a partir de quatro canais do sequenciador. Embora esta revelação se concentre na correção de faseamento dos sinais de cor, os conceitos aplicam-se a outros tipos de sinais gerados durante os agrupamentos de sequenciamento dos ácidos nucleicos com sequências idênticas. Exemplos desses outros sinais incluem radiação fora do espectro visível, concentração de íons e outros.

[007] Em certas modalidades, a determinação dos valores de cor corrigidos em (f) utiliza (i) os valores de cor no buffer do processador, (ii) os valores parcialmente corrigidos na fase armazenados durante o ciclo imediatamente anterior e (iii) a correção pré-faseamento. Em certas modalidades, a determinação da correção pré-faseamento em (e) utiliza (i) os valores de cor parcialmente corrigidos na fase armazenados durante o ciclo de chamada de bases imediatamente anterior e (ii) os valores de cor armazenados no buffer do processador.

[008] Em certas modalidades, a correção pré-faseamento inclui um peso. Nessas modalidades, a operação de determinar os valores de cor corrigidos pode incluir multiplicar o peso pelos valores de cor da pluralidade de sítios medidos a partir da imagem do substrato.

[009] Em certas implementações, os métodos incluem adicionalmente determinar uma correção de faseamento para o ciclo de chamada de bases imediatamente subsequente. Como exemplo, a determinação da correção de faseamento para o ciclo de chamada de bases imediatamente subsequente inclui a análise (i) dos valores de cor parcialmente corrigidos na fase armazenados na memória do sequenciador e (ii) os valores de cores armazenados no buffer do processador. Em certas modalidades incluindo a determinação de uma correção de faseamento para o ciclo de chamada de bases imediatamente subsequente, os métodos incluem adicionalmente (i) produzir valores de cor parcialmente corrigidos na fase para o chamada de bases imediatamente subsequente aplicando a correção de faseamento aos valores de cor da pluralidade de sítios armazenados na memória do sequenciador; e (ii) armazenar os valores de cor parcialmente corrigidos na fase para a chamada de bases imediatamente subsequente na memória do sequenciador. Em certas modalidades, produzir os valores de cor parcialmente corrigidos na fase para o ciclo de chamada de bases imediatamente subsequente inclui adicionalmente somar (i) os valores de cor

faseados corrigidos da pluralidade de sítios e (ii) os valores de cor da pluralidade de sítios da imagem do substrato medidos em (b). Em algumas implementações, o armazenamento dos valores de cor parcialmente corrigidos na fase para o ciclo de chamada de bases imediatamente subsequente armazena os valores de cor parcialmente corrigidos nos buffers do bloco da memória do sequenciador.

[0010] Em certas modalidades, os métodos são realizados em tempo real durante a aquisição de leituras das sequências pelo sequenciador de ácidos nucleicos. Em certas modalidades, o sequenciador de ácidos nucleicos sequencia, por síntese, ácidos nucleicos na pluralidade de sítios. Em certas modalidades em que o substrato inclui uma célula de fluxo, a célula de fluxo está logicamente dividida em blocos, e cada bloco representa uma região da célula de fluxo que compreende um subconjunto de sítios, cujo subconjunto é capturado em uma única imagem do sistema de aquisição de imagem.

[0011] Em algumas modalidades que utilizam esses sistemas, na operação (d) (recuperação dos valores de cor parcialmente corrigidos na fase da pluralidade de sítios), os valores de cor parcialmente corrigidos na fase foram previamente armazenados nos buffers do bloco da memória do sequenciador, onde os buffers do bloco são designados para armazenar dados representando as imagens dos blocos individuais no substrato. Em certas modalidades, a memória tem uma capacidade de armazenamento de cerca de 512 gigabytes ou menos, ou de cerca de 256 gigabytes ou menos. Em certas modalidades, por exemplo, a memória tem uma capacidade de armazenamento inferior a duas vezes a capacidade necessária para armazenar os dados contidos no número total de blocos nas duas células de fluxo. Em algumas modalidades, o processamento descrito na presente invenção economiza pelo menos cerca de 50 gigabytes; em algumas modalidades, ele economiza pelo menos cerca de 100 gigabytes.

[0012] Em algumas implementações, antes da operação (a) (obtenção

de uma imagem a partir de um substrato), os métodos incluem adicionalmente prover reagentes para a célula de fluxo e permitir que os reagentes interajam com os sítios para exibir as cores que representam os tipos de base de ácido nucleico durante o ciclo de chamada de bases. Em tais implementações, o método pode incluir adicionalmente, após a operação (f) (determinar os valores de cor corrigidos): (i) prover reagentes frescos à célula de fluxo e permitir que os reagentes frescos interajam com os sítios para exibir cores representando os tipos de base de ácido nucleico para um próximo ciclo de chamada de bases; e (ii) repetir as operações (a)-(e) para o próximo ciclo de chamada de bases. Tais métodos podem adicionalmente incluir a criação de um primeiro encadeamento de processador para executar as operações (a) a (f) para o ciclo de chamada de bases, e criar um segundo encadeamento de processador para executar as operações (a) a (f) para o ciclo de chamada de bases seguinte. Em certas modalidades, os métodos adicionalmente incluem alocar o buffer do processador e um segundo buffer do processador, sendo que o segundo buffer do processador é usado para determinar os valores de cor corrigidos em (f).

[0013] Certos outros aspectos da revelação referem-se aos sequenciadores de ácidos nucleicos que podem ser distinguidos pelos seguintes elementos: um sistema de aquisição de imagem; memória; e um ou mais processadores concebidos ou configurados para: (a) obter dados representando uma imagem de um substrato incluindo uma pluralidade de sítios onde as bases de ácido nucleico são lidas (os sítios exibem, por exemplo, cores representando tipos de bases de ácido nucleico); (b) obter valores das cores da pluralidade de sítios a partir da imagem do substrato; (c) armazenar os valores das cores em um buffer de processador; (d) recuperar os valores de cor parcialmente corrigidos na fase da pluralidade de sítios para um ciclo de chamada de bases (os valores de cor parcialmente corrigidos na fase foram armazenados na memória do sequenciador durante um ciclo de

chamada de bases imediatamente anterior); (e) determinar uma correção pré-faseamento; e (f) determinar os valores de cor corrigidos, por exemplo, (i) dos valores de cor no buffer do processador, (ii) dos valores de cor parcialmente corrigidos na fase armazenados durante o ciclo imediatamente anterior e (iii) da correção pré-faseamento.

[0014] As instruções ou outra configuração para determinar uma correção pré-faseamento podem incluir a configuração para determinar a correção pré-faseamento a partir dos (i) valores de cor parcialmente corrigidos na fase armazenados durante o ciclo de chamada de bases imediatamente anterior e dos (ii) valores de cores armazenados no buffer do processador.

[0015] Em certas modalidades, a memória é dividida em uma pluralidade de buffers do bloco, sendo cada um designado para armazenar dados que representam uma única imagem de um bloco no substrato. Em certas modalidades, a memória tem uma capacidade de armazenamento inferior a cerca de 550 gigabytes (em alguns exemplos, isto é menos do que duas vezes a capacidade necessária para armazenar os dados contidos no número total de blocos nas duas células de fluxo).

[0016] Os processadores podem ser configurados para executar as operações citadas de várias maneiras, tais como recebendo instruções executáveis legíveis por máquina. Em alguns casos, os processadores são programados com núcleos de processamento de firmware ou personalizados, como núcleos de processamento de sinal digital. Em várias modalidades, o(s) processador(es) são projetados ou configurados para executar (e/ou controlar) qualquer uma ou mais das operações do método descritas acima.

[0017] Em algumas implementações, as características de correção de faseamento aqui descritas reduzem substancialmente o custo de um instrumento de sequenciamento utilizando a memória de forma mais eficiente (por exemplo, a memória de acesso aleatório (RAM)). Algumas modalidades empregam estes atributos de correção de faseamento no contexto da análise

em tempo real (RTA) em plataformas de sequenciamento.

[0018] Estes e outros atributos da revelação serão apresentados em mais detalhes abaixo, com referência aos desenhos associados.

BREVE DESCRIÇÃO DOS DESENHOS

[0019] A figura 1 é um diagrama de blocos de um sequenciador com hardware para análise em tempo real dos dados de imagens tiradas dos agrupamentos de ácidos nucleicos.

[0020] A figura 2 é uma ilustração dos dados de sequenciamento de dois canais usados para ilustrar os conceitos de faseamento e pré-faseamento.

[0021] A figura 3 descreve uma arquitetura de célula de fluxo incluindo uma pluralidade de blocos, com cada um contendo muitos agrupamentos.

[0022] A figura 4 representa uma matriz de dados contendo dados de magnitude para agrupamentos em um bloco ou outra porção fotografada de uma célula de fluxo; os dados de magnitude podem ser valores de intensidade de luz para cada um dos dois ou mais canais de cor.

[0023] A figura 5 representa esquematicamente uma primeira configuração de processamento e a metodologia para realizar a correção de faseamento em tempo real.

[0024] A figura 6 apresenta um fluxograma de um processo de chamada bases que pode empregar a configuração do processador e da memória mostrada na figura 5.

[0025] A figura 7 representa esquematicamente uma segunda configuração de processamento e a metodologia para realizar a correção de faseamento em tempo real. Esta configuração reduz os requisitos de memória do sistema.

[0026] A figura 8 representa esquematicamente uma terceira configuração de processamento e a metodologia para realizar a correção de faseamento em tempo real. Esta configuração reduz ainda mais os requisitos

de memória do sistema.

[0027] A figura 9 representa um fluxograma de alto nível dos primeiros ciclos de processamento que podem ser utilizados com a configuração de processador e memória da figura 8 e, em algumas implementações, da figura 7.

[0028] A figura 10 representa um fluxograma de ciclos de processamento que realizam chamada de bases corrigidas por faseamento completo. Este ciclo pode ser realizado no terceiro e nos ciclos de processamento subsequentes ao sequenciar agrupamentos de um bloco.

[0029] A figura 11 representa dados comparativos para métodos de correção de faseamento, um usando um algoritmo de memória principal reduzido.

DESCRIÇÃO DETALHADA

DEFINIÇÕES

[0030] As faixas numéricas incluem os números que definem a faixa. Deve-se compreender que cada limite numérico máximo mencionado nesse relatório descritivo inclui cada um dos limites numéricos inferiores, como se tais limites numéricos inferiores estivessem expressamente registrados no presente documento. Cada limitação numérica mínima dada ao longo deste relatório descritivo incluirá todas as limitações numéricas mais altas, como se tais limitações numéricas mais altas estivessem expressamente escritas na presente invenção. Cada faixa numérica dada ao longo deste relatório descritivo incluirá todas as faixas numéricas mais restritas que se enquadrem dentro de tal faixa numérica mais ampla, como se tais faixas numéricas mais restritas estivessem todas aqui expressamente escritas.

[0031] Os títulos providos na presente invenção não se destinam a limitar a revelação.

[0032] A menos que seja definido de outra forma, todos os termos técnicos e científicos neste documento têm o mesmo significado como é

comumente entendido por uma pessoa normalmente versada na técnica. Vários dicionários científicos que incluem os termos aqui incluídos são bem conhecidos e estão disponíveis para as pessoas versadas na técnica. Embora métodos e materiais similares ou equivalentes àqueles descritos na presente invenção podem ser usados na prática ou teste das modalidades reveladas na presente invenção, alguns métodos e materiais são descritos.

[0033] Os termos definidos imediatamente abaixo são descritos mais detalhadamente por referência ao relatório descritivo como um todo. Deve ser compreendido que esta revelação não se limita à metodologia, protocolos e reagentes particulares descritos, uma vez que estes podem variar, dependendo do contexto em que são utilizados pelas pessoas versadas na técnica.

[0034] Conforme utilizado na presente invenção, os termos no singular "um", "uma" e "o/a" incluem as referências no plural, a menos que o contexto claramente especifique de outra forma. O termo "pluralidade" refere-se a mais de um elemento. Por exemplo, o termo é aqui utilizado em referência a um número de leituras para produzir ilhas faseadas usando os métodos revelados na presente invenção.

[0035] O termo "porção" é usado na presente invenção em referência à quantidade de informação de sequência de genoma, cromossomo ou haplótipo em uma amostra biológica que, somadas, é menor do que a informação de sequência de um genoma completo, um cromossomo completo ou um haplótipo completo, como fica evidente a partir do contexto.

[0036] O termo "amostra", na presente invenção, refere-se a uma amostra, tipicamente derivada de um fluido biológico, célula, tecido, órgão ou organismo contendo um ácido nucleico ou uma mistura de ácidos nucleicos contendo pelo menos uma sequência de ácido nucleico a ser sequenciada. Tais amostras incluem, mas não se limitam, a escarro/fluido oral, líquido amniótico, líquido cefalorraquidiano, sangue, uma fração do sangue (por exemplo, soro ou plasma), amostras de biópsia por agulha fina (por exemplo,

biópsia cirúrgica, biópsia por agulha fina, etc.), urina, saliva, sêmen, suor, lágrimas, fluido peritoneal, fluido pleural, explante de tecido fluido de lavagem, cultura de órgão e qualquer outro tecido ou preparação de células, ou fração ou derivado dos mesmos, ou que seja isolado a partir dos mesmos.

[0037] Embora a amostra seja frequentemente retirada de um indivíduo humano (por exemplo, paciente), as amostras podem ser coletadas de qualquer organismo que tenha cromossomos, incluindo, mas sem se limitar, a cães, gatos, cavalos, cabras, ovelhas, gado, porcos, etc. A amostra pode ser usada diretamente como obtida de uma fonte ou após um pré-tratamento para modificar o caráter da amostra. Por exemplo, tal pré-tratamento pode incluir a preparação do plasma do sangue, diluição de fluidos viscosos e similares. Os métodos de pré-tratamento também podem envolver, mas não se limitam, a filtração, precipitação, diluição, destilação, mistura, centrifugação, congelamento, liofilização, concentração, amplificação, fragmentação de ácido nucleico, inativação de componentes interferentes, adição de reagentes, lise, etc. Se tais métodos de pré-tratamento forem utilizados em relação à amostra, tais métodos de pré-tratamento são tipicamente tais que o(s) ácido(s) nucleico(s) de interesse permanecem na amostra de teste, por vezes numa concentração proporcional àquela em uma amostra de teste não tratada (por exemplo, em uma amostra de teste que não está sujeita a nenhum desses métodos de pré-tratamento). Essas amostras "tratadas" ou "processadas" ainda são consideradas como sendo amostras biológicas de "teste" com relação aos métodos descritos na presente invenção.

[0038] Os termos "polinucleotídeo", "ácido nucleico" e "moléculas de ácido nucleico" são usados indistintamente e referem-se a uma sequência covalentemente ligada de nucleotídeos (ou seja, de ribonucleotídeos para RNA e desoxirribonucleotídeos para DNA) em que a posição 3' da pentose de um nucleotídeo é unida por um grupo fosfodiéster à posição 5' da pentose da seguinte. Os nucleotídeos incluem sequências de qualquer forma de ácido

nucleico, incluindo, mas sem se limitar, moléculas de DNA e RNA. O termo "polinucleotídeo" inclui, sem limitação, polinucleotídeo de fita simples e dupla.

[0039] Moléculas polinucleotídicas de fita simples podem ter se originado na forma de fita simples, como DNA ou RNA, ou ter se originado na forma de DNA de fita dupla (dsDNA) (por exemplo, segmentos de DNA genômico, produtos de PCR e amplificação e similares). Assim, um polinucleotídeo de fita simples pode ser a fita de sentido ou antissentido de um polinucleotídeo duplex. Os métodos de preparação de moléculas de polinucleotídeo de fita simples adequadas para uso nos métodos descritos utilizando técnicas padrão são bem conhecidos na técnica. A sequência precisa das moléculas de polinucleotídeo primárias geralmente não é material às modalidades reveladas, e pode ser conhecida ou desconhecida. As moléculas de polinucleotídeo de fita simples podem representar moléculas de DNA genômico (por exemplo, DNA genômico humano) incluindo sequências de íntron e éxon (sequências codificantes), bem como sequências reguladoras não codificantes, tais como sequências promotoras e intensificadoras.

[0040] O ácido nucleico descrito na presente invenção pode ser de qualquer comprimento adequado para uso nos métodos providos. Por exemplo, os ácidos nucleicos alvo podem ter pelo menos 10, pelo menos 20, pelo menos 30, pelo menos 40, pelo menos 50, pelo menos 75, pelo menos 100, pelo menos 150, pelo menos 200, pelo menos 250, pelo menos 500, ou pelo menos 1000 kb de comprimento ou mais.

[0041] No contexto de uma célula de fluxo ou outro substrato para sequenciamento, o termo "sítio" refere-se a uma pequena região onde o sequenciamento ocorre. Em muitas modalidades, um sítio contém múltiplas, e tipicamente numerosas cópias de uma única sequência de ácidos nucleicos a partir da qual os dados de sequenciamento são obtidos. Os dados da sequência obtidos de um sítio podem ser uma "leitura".

[0042] O termo "polimorfismo" ou "polimorfismo genético" é usado na presente invenção em referência à ocorrência na mesma população de dois ou mais alelos em um *locus* genético. Várias formas de polimorfismo incluem polimorfismos de nucleotídeo único, repetições em tandem, microdeleções, inserções, inserções/eliminações e outros polimorfismos.

[0043] Uma "chamada de bases" é uma base atribuída (tipo de nucleotídeo) para sequenciar os dados de uma localização específica em uma sequência polinucleotídica. Uma chamada de bases pode ser produzida por um sequenciador para cada posição no ácido nucleico sendo sequenciado. Uma qualidade da chamada às vezes é atribuída a uma chamada de bases.

[0044] O termo "leitura" refere-se a uma sequência lida de uma porção de uma amostra de ácido nucleico. Normalmente, embora não necessariamente, uma leitura representa uma sequência curta de pares de bases contíguos na amostra. A leitura pode ser representada simbolicamente pela sequência de pares de bases (em ATCG) da porção da amostra. Ela pode ser armazenada em um dispositivo de memória e processada conforme apropriado para determinar se ela corresponde a uma sequência de referência ou se atende a outros critérios. Uma leitura pode ser obtida diretamente de um aparelho de sequenciamento, ou indiretamente a partir das informações de sequência armazenadas em relação à amostra. Em alguns casos, uma leitura é uma sequência de DNA de comprimento suficiente (por exemplo, de pelo menos 25 pb) que pode ser usada para identificar uma sequência ou região maior, por exemplo, que pode ser alinhada e especificamente atribuída a uma região cromossômica ou genômica, ou gene.

[0045] O termo "Sequenciação de Nova Geração (NGS)", usado na presente invenção, refere-se aos métodos de sequenciamento que permitem o sequenciamento massivamente paralelo de moléculas clonalmente amplificadas e de moléculas de ácido nucleico únicas. Exemplos não limitantes de NGS incluem sequenciamento por síntese usando terminadores

de corante reversíveis e sequenciamento por ligação.

[0046] O termo "parâmetro", na presente invenção, se refere a um valor numérico que caracteriza uma propriedade física ou uma representação dessa propriedade. Em algumas situações, um parâmetro caracteriza numericamente um conjunto de dados quantitativos e/ou uma relação numérica entre conjuntos de dados quantitativos. Por exemplo, a média e a variância de uma distribuição padrão ajustada a um histograma são parâmetros.

[0047] Os termos "limite", na presente invenção, referem-se a qualquer número que é usado como ponto de corte para caracterizar uma amostra, um ácido nucleico ou uma parte dele (por exemplo, uma leitura). O limite pode ser comparado a um valor medido ou calculado para determinar se a fonte que dá origem a tal valor sugere que deve ser classificada de uma maneira particular. Os valores limites podem ser identificados empiricamente ou analiticamente. A escolha de um limite depende do nível de confiança que o usuário deseja ter para fazer a classificação. Às vezes, eles são escolhidos para uma finalidade específica (por exemplo, para equilibrar a sensibilidade e a seletividade).

[0048] A análise em tempo real refere-se a um processo e sistema no qual o processamento e a análise de dados são realizados em segundo plano na aquisição de dados durante uma execução de sequenciamento de DNA. Um exemplo de um sistema de análise em tempo real é descrito na Patente Norte-Americana No. 8.965.076, que é incorporada na presente invenção por referência em sua totalidade.

CONTEXTO PARA FASEAMENTO

Aparelho de sequenciamento

[0049] A figura 1 mostra um diagrama de blocos de algumas características de um sequenciador de ácidos nucleicos típico 100 ou um sistema incluindo tal sequenciador. Notavelmente, o sistema 100 inclui uma

célula de fluxo 101 e o sistema de aquisição de imagem 103, um ou mais processadores 105 com um ou mais buffers 107 e a memória do sistema (às vezes referida como memória principal) 109 incluindo uma pluralidade de buffers do bloco 111. Tipicamente, a memória do sistema 109 é provida no dispositivo que não faz parte de um circuito integrado contendo qualquer um ou mais processador(es) 105. Em certas modalidades, a memória do sistema é memória volátil, como a memória de acesso aleatório ou RAM, por exemplo, DRAM, uma unidade de disco rígido de estado sólido ou uma unidade de disco rígido.

[0050] A célula de fluxo e o sistema de aquisição de imagem contêm componentes concebidos ou configurados de acordo com princípios compreendidos no campo do sequenciamento de ácido nucleico, e eles não serão aqui descritos em detalhes. Sistemas de análise de imagem adequados e as células de fluxo associadas são utilizados nos sequenciadores de ácidos nucleicos, como nas séries MiSeq e HiSeq disponíveis junto à Illumina, Inc. de São Diego, Califórnia. Para obter mais informações, consulte a Patente Norte-Americana No. 8.241.573, a Patente Norte-Americana No. 9.193.996 e a Patente Norte-Americana No. 8.951.781, cada uma das quais sendo aqui incorporada por referência em suas totalidades.

[0051] Em geral, as sequências de ácidos nucleicos adequadas para uso com os métodos revelados proporcionam uma detecção rápida e eficiente de uma pluralidade de ácidos nucleicos alvos em paralelo. Elas podem incluir componentes fluidos capazes de distribuir reagentes de amplificação e/ou reagentes de sequenciamento para um ou mais fragmentos de DNA imobilizados, com o sistema incluindo componentes tais como bombas, válvulas, reservatórios, linhas fluidas e similares. Uma célula de fluxo pode ser configurada e/ou usada em um sistema integrado para detecção de ácidos nucleicos alvo. Células de fluxo exemplificadoras são descritas, por exemplo, nos documentos US 2010/0111768 A1 e no documento US No. de série

13/276366, sendo cada uma deles aqui incorporado a título de referência, em suas totalidades. Como exemplificado para as células de fluxo, um ou mais dos componentes fluidos de um sistema integrado podem ser utilizados tanto para um método de amplificação como para um método de detecção. Por exemplo, um ou mais dos componentes fluidos de um sistema integrado podem ser utilizados para um método de amplificação e para a liberação dos reagentes de sequenciamento em um método de sequenciamento. Alternativamente, um sistema integrado pode incluir sistemas fluidos separados para realizar métodos de amplificação e para realizar métodos de detecção.

[0052] Para fins desta revelação, é suficiente compreender que a célula de fluxo primeiro recebe e imobiliza ou, de outro modo, captura uma amostra de ácido nucleico a ser sequenciada e, em seguida, expõe a vários reagentes associados ao processo de sequenciamento. Em certas modalidades, o processo de sequenciamento é um processo de sequência por síntese, embora outras tecnologias de sequenciamento possam ser utilizadas.

[0053] O sistema de aquisição de imagem 103 inclui componentes ópticos, tais como componentes de excitação de fluorescência (por exemplo, um laser e espelhos e lentes associados) para iluminar sítios na célula de fluxo onde o sequenciamento está ocorrendo e componentes de captura de imagem para capturar imagens de fluorescência nas porções da célula de fluxo com vários sítios. Os dados capturados pelo sistema de aquisição de imagens contêm informações adequadas para determinar qual nucleotídeo está sendo lido em qualquer local dado em qualquer dado ciclo de sequenciamento.

[0054] Para permitir uma análise em tempo real, o sequenciador 100 inclui tipicamente processadores e memória integrados que interpretam e armazenam dados de imagem a partir do sistema de aquisição de imagem 103. Exemplos de processadores adequados para o sequenciador incluem a classe Xeon E5 da Intel. Tipicamente, o processador 105 inclui múltiplos buffers

107 que armazenam temporariamente dados de imagens tiradas durante um único ciclo de aquisição de imagem. Na modalidade descrita, os buffers do processador são alocados na memória do sistema. Um determinado buffer do processador pode ser associado a um encadeamento de processador específico criado para analisar dados de imagem de uma região da célula de fluxo durante a análise em tempo real. Em certas modalidades, os dados de imagem analisados por um encadeamento são os de um único bloco (descrito abaixo), capturados durante um único ciclo de aquisição de imagem. Em certas modalidades, o buffer pode armazenar cerca de 400 gigabytes de dados. Conforme utilizado na presente invenção, um encadeamento é uma sequência ordenada de instruções que informa ao processador quais operações executar. As instruções configuram o processador usando código de máquina executável selecionado de um conjunto de instruções de linguagem de máquina específico, ou "instruções nativas", projetadas no processador de hardware.

[0055] O conjunto de instruções de linguagem de máquina, ou conjunto de instruções nativas, é conhecido e é essencialmente incorporado ao(s) processador(es) de hardware ou CPUs. Essa é a "linguagem" pela qual o sistema e o software do aplicativo se comunicam com os processadores de hardware. Cada instrução nativa é um código discreto que é reconhecido pela arquitetura de processamento e que pode especificar registros particulares para funções aritméticas, de endereçamento ou de controle; determinados locais ou deslocamentos de memória; e modos de endereçamento específicos usados para interpretar operandos. Operações mais complexas são construídas combinando-se essas instruções nativas simples, que são executadas sequencialmente ou, de outra forma, direcionadas por instruções de fluxo de controle.

[0056] A memória do sistema 109 inclui vários buffers do bloco 111, cada um configurado para armazenar uma parte dos dados de imagem

adquiridos da célula de fluxo durante um único ciclo de aquisição de imagem. Os buffers do bloco neste exemplo são referidos como tais porque estão configurados para conter o valor de dados de imagem de um único bloco. Conforme explicado mais detalhadamente abaixo, um bloco é uma região de uma célula de fluxo que pode ser capturada em uma única imagem tirada durante um único ciclo de aquisição de imagem. Os buffers do bloco 111 destinam-se a armazenar dados de imagem durante um período de tempo mais longo do que os buffers do processador 107. Em certas modalidades, os buffers do bloco 111 armazenam dados de imagem para pelo menos dois ciclos de aquisição de imagem. Embora este aplicativo descreva buffers que armazenam dados de um bloco de uma célula de fluxo, as modalidades reveladas não estão limitadas aos buffers que armazenam esta quantidade de dados. A menos que seja especificado de outra forma ou esteja claro a partir do contexto, as referências aos "buffers do bloco" incluem qualquer tipo de buffer que armazene dados de imagem de uma porção de uma célula de fluxo, cujos dados de imagem são processados como uma unidade, conforme descrito na presente invenção.

[0057] Para fazer chamadas de base, um ou mais processadores 105 atuam nos dados providos a partir da memória do sistema 109 e dos dados armazenados nos buffers do processador 107. Normalmente, uma única chamada de base é feita para um único sítio durante um único ciclo de aquisição de imagem.

[0058] Como mostrado, o um ou mais processadores 105 e a memória principal 109 compartilham dados bidirecionalmente. Além disso, o um ou mais processadores 105 recebem dados de imagem do sistema de aquisição de imagem 103. Em certas modalidades, o sistema de aquisição de imagem 103 obtém dados da célula de fluxo 101 estimulando os sítios de sequenciamento na célula de fluxo 101 e recebendo sinais ópticos desses sítios. Em certas modalidades, o sinal recebido pelo sistema de aquisição de imagem 103 é um

sinal de fluorescência criado quando o sistema 103 ilumina a célula de fluxo 101 com luz em comprimentos de onda apropriados. Em tais modalidades, o sinal de fluorescência é provido como valores de intensidade para uma pluralidade de cores.

[0059] O conceito de um ciclo é usado ao longo desta revelação. Um único ciclo de *sequenciamento* envolve a leitura de um único nucleotídeo de cada um dos sítios capturados em uma imagem. A leitura é referida como fazer uma chamada de bases. Em várias modalidades descritas na presente invenção, um único ciclo *computacional* - da perspectiva do(s) processador(es) e memória - realiza tanto a chamada de bases quanto a captura de imagem, mas para nucleotídeos diferentes, com a captura de imagens atrasando a chamada de bases na sequência de nucleotídeos sendo lidos ou chamados. Por exemplo, em um único ciclo computacional, um ou mais processadores realizam a chamada de bases para um nucleotídeo no ciclo de sequenciamento ***n*** e simultaneamente realizam a captura de imagem para o nucleotídeo no ciclo de sequenciamento ***n +1***. Assim, em um único ciclo computacional, o sequenciador (a) armazena e processa dados de imagem não modificados para os nucleotídeos no ciclo de sequenciamento ***n +1*** e (b) faz uma chamada de bases para os nucleotídeos no ciclo de sequenciamento ***n***. O uso dos buffers do processador e dos buffers do bloco neste processamento ciclo a ciclo será descrito em mais detalhes abaixo.

Faseamento em geral

[0060] Em um sítio particular em uma célula de fluxo ou em outro substrato, múltiplas cópias de uma molécula de ácido nucleico, todas tendo a mesma sequência (possivelmente com variações limitadas involuntariamente introduzidas por processamento de amostras), são analisadas ao mesmo tempo. São usadas cópias suficientes para garantir que um sinal suficiente seja produzido para permitir uma chamada de base confiável. A coleta de moléculas de ácido nucleico em um sítio é chamada de agrupamento. Em

alguns casos, um agrupamento não sequenciado contém apenas moléculas de ácido nucleico de fita simples.

[0061] O faseamento representa um artefato não intencional que surge do sequenciamento de múltiplas moléculas de ácido nucleico dentro de um agrupamento. O faseamento é a taxa na qual sinais, como a fluorescência de moléculas únicas dentro de um agrupamento, perdem a sincronização entre si. Muitas vezes, o termo faseamento é reservado para contaminar o sinal de algumas moléculas que ficam para trás, e o termo pré-faseamento é usado para contaminar o sinal de outras moléculas que vão adiante. Juntos, o faseamento e o pré-faseamento descrevem o quão bem o aparelho de sequenciamento e a química estão se saindo.

[0062] Números baixos são melhores. Valores de 0,10/0,10 significam que 0,10% das moléculas em um agrupamento estão ficando para trás e 0,10% está à frente em cada ciclo de chamada de bases. Em outras palavras, 0,20% do sinal verdadeiro é perdido em cada ciclo e, portanto, contribuirá para o ruído. Em outro exemplo, 0,20/0,20 significa que 0,4% do sinal verdadeiro é perdido por ciclo, em cujo caso após 250 ciclos (sem correção) o ruído seria igual ao sinal.

[0063] Um componente de análise em tempo real de um sequenciador pode determinar o escalonamento e o pré-faseamento, a fim de aplicar o nível correto de correção de faseamento, à medida que o sequenciamento prossegue. Isso funciona empurrando artificialmente o sinal para dentro ou para fora de cada canal do sequenciador com base em chamadas de base antes ou depois do ciclo atual.

[0064] Anteriormente, o faseamento e o pré-faseamento eram estimados ao longo de um número definido de ciclos (por exemplo, os primeiros 12 ciclos de cada leitura) e, depois, aplicados a todos os ciclos subsequentes. Alguns sequenciadores recentes empregam um algoritmo chamado de correção de faseamento empírica para otimizar a correção de

faseamento em cada ciclo, tentando uma faixa de correções e selecionando aquela que resulta na maior castidade (pureza do sinal). Embora a correção de faseamento empírica forneça melhor desempenho, ela requer mais recursos computacionais.

[0065] Nos sequenciadores convencionais, cada base possui uma única cor de corante fluorescente; por exemplo, verde para timina, vermelho para citosina, azul para guanina e amarelo para adenina. Para capturar informações para as chamadas de base, um sequenciador de quatro canais obtém quatro imagens de um bloco ou outra porção de uma célula de fluxo. Alguns sequenciadores agora têm apenas dois canais e, portanto, tiram apenas duas imagens da mesma parte da célula de fluxo. Um sequenciador de dois canais usa uma mistura de corantes para cada base e usa filtros vermelhos e verdes para as duas imagens. Em um exemplo de sequenciador de dois canais, os agrupamentos vistos em imagens vermelhas ou verdes são interpretados como bases C e T, respectivamente. Os agrupamentos observados em imagens vermelhas e verdes são sinalizados como bases A, enquanto os agrupamentos não marcados são identificados como bases G.

[0066] A figura 2 ilustra o faseamento durante o sequenciamento de um agrupamento nucleico com a sequência. . . ACGTAAG . . . . Como ilustrado, durante o ciclo de chamada de bases para o primeiro G, 98,4% do sinal de fluorescência se origina de sequências atualmente gerando sinal para G, enquanto 1,5% do sinal de fluorescência se origina de sequências atualmente produzindo sinal para a base anterior C e 1,1 % do sinal de fluorescência se origina de sequências atualmente produzindo sinal para a próxima base T. A contribuição do sinal para a base anterior C vem do faseamento e a contribuição do sinal da próxima base T vem do pré-faseamento

[0067] A correção de faseamento para esta chamada de bases G é refletida no gráfico no lado direito da figura 2. Como mostrado para um

sequenciador de dois canais, o sinal de fluorescência pode ser representado em um gráfico bidimensional, com o sinal de intensidade máxima em um "eixo verde" representando T, a intensidade máxima em um "eixo vermelho" representando C, a metade da intensidade máxima entre os eixos que representam A e a intensidade mínima em ambos os eixos representando G. Sem erro de faseamento, o sinal para G deve ter intensidade zero nos eixos vermelho e verde. Em vez disso, com o erro de faseamento discutido, o sinal de fluorescência tem certo grau de contribuição de intensidade nos eixos verde e vermelho. Neste exemplo, a correção pré-faseamento reduz a intensidade do sinal até zero no eixo verde e a correção de faseamento reduz a intensidade do sinal até zero no eixo vermelho. Correções semelhantes podem ser feitas nas chamadas de bases para as bases T, C e A.

Blocos e células de fluxo

[0068] Conforme explicado, uma célula de fluxo contém vários sítios em que as informações de sequenciamento são coletadas. Em certas modalidades, cada sítio de uma célula de fluxo contém um agrupamento de ácidos nucleicos de fita simples que partilham a mesma sequência. Uma única imagem usada no sequenciamento em tempo real pode conter milhões desses agrupamentos. Uma célula de fluxo típica é tão grande que requer centenas ou até milhares de imagens separadas para cobrir toda a sua área. Em certas modalidades, o processador e a memória associada empregados para a análise em tempo real processam todas essas imagens atualmente para fazer as chamadas de base para um único ciclo. Em algumas implementações, o processador e a memória processam simultaneamente todas as imagens adquiridas em duas ou mais células de fluxo durante um único ciclo de chamada de bases. A figura 3 descreve esquematicamente uma arquitetura de célula de fluxo usada em alguns sequenciadores da Illumina, Inc. No exemplo descrito, o sequenciador faz chamadas de bases concorrentes em duas células de fluxo, na célula de fluxo 1 e na célula de fluxo 2. Em certas modalidades,

cada célula de fluxo possui sítios de sequenciamento em cada uma das duas superfícies, em uma superfície superior na superfície inferior. Nesses casos, o sequenciador visualiza as superfícies superior e inferior durante cada ciclo de chamada de bases. Como representado na figura 3, cada superfície de célula de fluxo inclui quatro raias, L1, L2, L3 e L4; naturalmente, outros números são possíveis. Cada raia de cada superfície pode ter várias subdivisões referidas como faixas. Cada faixa é, por sua vez, dividida em vários blocos. Por exemplo, pode haver aproximadamente 120 blocos por faixa. Considerando duas células de fluxo, cada uma tendo duas superfícies, com cada superfície tendo quatro raias, e cada raia tendo seis faixas, e cada faixa tendo 120 blocos, vários milhares de blocos de dados precisam ser analisados por ciclo. Em várias modalidades, cada imagem de bloco (ou outra imagem de uma porção de uma célula de fluxo) é acionada por um único encadeamento de processador. Em certas modalidades, um sequenciador que emprega uma célula de fluxo tendo a arquitetura representada na figura 3 processa 8000 ou mais blocos de dados em cada ciclo de chamada de bases. Em tais casos, a lógica de processamento em tempo real empregaria 8000 ou mais encadeamentos de processador em cada ciclo de chamada de bases.

[0069] Os dados de um único bloco capturado durante um único ciclo podem ser armazenados na memória como uma matriz, com cada entrada na matriz representando um valor de cor para cada canal de um único agrupamento no bloco. Uma matriz para um arranjo de dois canais é mostrada na figura 4. Por exemplo, um detector de intensidade de cor pode gerar contagens de sinal entre cerca de 400 e 1500 para cada canal. Um buffer de bloco na memória do sistema é configurado para armazenar todas as informações na matriz, em outras palavras, os valores de cores de todos os agrupamentos em um bloco em um único ciclo de chamada de bases. Um buffer do processador pode ser configurado de forma semelhante para armazenar todas as informações na matriz.

Processo de Faseamento

[0070] Uma carga de memória significativa da análise em tempo real dos dados de sequência deriva do requisito na correção de faseamento de que dois ou três ciclos de intensidades de agrupamento devem ser salvos para cada bloco durante toda a análise. Em um HiSeqX da Illumina com uma célula de fluxo de 700nm, isso ocupa 73 gigabytes de memória. Essa carga é suficientemente grande para que a maioria dos dados (nesta plataforma) seja armazenada em cache em um disco rígido de estado sólido.

[0071] Como explicado, a correção de faseamento ajusta os valores de intensidade de uma imagem para resolver o sequenciamento de fase de algumas fitas de ácido nucleico em um agrupamento. A correção de faseamento faz isso começando com os valores medidos da intensidade de cor do agrupamento (ou outros sinais medidos pelo método de sequenciamento) para um ciclo de chamada de bases atual e adicionando ou subtraindo um valor de correção usando valores de intensidade medidos do ciclo de chamada de bases anterior e/ou usando valores de intensidade medidos a partir do ciclo de chamada de bases subsequente. Em várias implementações, um valor de intensidade de faseamento corrigido para fazer uma chamada de bases aplica uma expressão conforme é mostrada na parte inferior da figura 5. Como mostrado aqui, os valores de intensidade de faseamento corrigidos para um ciclo de chamada de bases atual em uma imagem são iguais aos valores de intensidade medidos para o ciclo de chamada de bases atual menos o produto de um primeiro coeficiente e os valores de intensidade medidos no ciclo de chamada de bases imediatamente anterior e menos o produto de um segundo coeficiente e valores de intensidade medidos no ciclo de chamada de bases imediatamente sucessivo:

$$\text{Intensidade Corrigida} = -a \cdot I_{n-1} + I_n - b \cdot I_{n+1}$$

onde I_{n-1} , I_n , e I_{n+1} são os valores de intensidade de agrupamentos em um bloco no ciclo de chamada de bases imediatamente

anterior, no ciclo de chamada de bases atual e no chamada de bases imediatamente subsequente, respectivamente. Os coeficientes *a* e *b* são os coeficientes de faseamento e de pré-faseamento (às vezes chamados de pesos), respectivamente. Estes podem ser calculados de novo para cada ciclo de chamada de bases de um bloco.

[0072] Voltando à figura 2, o valor da intensidade medida para a terceira base na sequência descrita (para um único agrupamento em uma imagem) é mostrado como um ponto no gráfico no lado direito da figura 2. A correção pré-faseamento para este valor de intensidade medido é refletida pela seta vertical do valor da intensidade medida até o eixo horizontal. Na expressão para os valores de intensidade de faseamento corrigidos, esta correção pré-faseamento é representada pelo produto do coeficiente *b* e o valor de intensidade medido para o próximo ciclo de chamada de bases sucessivo. Além disso, o valor da intensidade medida é corrigido por uma correção de faseamento representada pela seta horizontal no gráfico. Esta correção de faseamento é implementada subtraindo do valor da intensidade medida, o produto de um coeficiente *a* e o valor da intensidade medida para o ciclo de chamada de bases imediatamente anterior. Os coeficientes *a* e *b* podem ser determinados por vários métodos, mas em muitas implementações, eles são calculados de novo para cada ciclo de chamada de bases. Uma descrição dos métodos para determinar os coeficientes a serem utilizados na correção de faseamento está descrita no pedido de patente internacional com o número de publicação WO2015/084985, de Belitz et al. e publicado em 11 de junho de 2015, que é aqui incorporado por referência na presente invenção.

[0073] Em certas modalidades, o algoritmo de faseamento determina empiricamente os coeficientes de faseamento maximizando a castidade cumulativa (ou métrica similar) dos dados de intensidade do agrupamento durante um ciclo de chamada de bases. Uma implementação do algoritmo itera sobre todos ou muitos coeficientes de faseamento e determina quais

fornecem os melhores resultados. Por exemplo, o algoritmo de faseamento pode otimizar *a* e *b* em cada ciclo usando uma pesquisa de padrões que emprega uma função de custo que conta o número de agrupamentos que falham em um filtro de castidade. Assim, *a* e *b* são selecionados para maximizar a qualidade dos dados.

[0074] Em algumas modalidades, os coeficientes de faseamento são determinados como uma análise progressivo ao longo de uma análise de sequenciamento (por exemplo, durante a geração de uma leitura). Como resultado dessa abordagem, uma estimativa de faseamento imprecisa feita durante os ciclos iniciais não afetará desfavoravelmente os ciclos posteriores.

[0075] Alguns métodos determinam a castidade de um valor de intensidade de agrupamento como uma função de distâncias relativas a centroides gaussianos para os outros valores de intensidade de agrupamento determinados para o mesmo ciclo de chamada de bases. Os centroides idealmente se alinham com os locais esperados das intensidades A, T, C e G para dois canais (consulte a figura 2), assumindo que um sistema de dois canais é usado. Em certas modalidades, a castidade pode ser calculada usando a expressão:

castidade = 1 - Dl/(Dl + D2),

onde D1 é a distância até o centroide gaussiano mais próximo e D2 é a distância até o próximo centroide mais próximo. Utilizando essa abordagem, quando a castidade média (qualidade) dos valores de intensidade é maximizada, os valores corretos de *a* e *b* são escolhidos. Após esses valores serem identificados, uma correção pode ser aplicada a todos os valores de agrupamento e a chamada de bases pode ocorrer diretamente. Os métodos de ajuste das distribuições gaussianas a um conjunto de dados de dois canais estão descritos no pedido de patente internacional No. WO2015/084985, previamente incorporado por referência.

[0076] Em algumas modalidades, uma correção de faseamento é

calculada em quase todos os ciclos durante uma execução de sequenciamento. Em algumas modalidades, uma correção de faseamento é calculada em cada ciclo durante uma execução de sequenciamento. Em algumas modalidades, uma correção de faseamento separada é calculada para diferentes localizações de uma superfície imageada no mesmo ciclo. Por exemplo, em algumas modalidades, uma correção de faseamento separada é calculada para cada raia individual de uma superfície fotografada, tal como uma raia de célula de fluxo individual. Em algumas modalidades, uma correção de faseamento separada é calculada para cada subconjunto de uma raia, tal como uma faixa de imagem dentro de uma raia de células de fluxo. Em algumas modalidades, uma correção de faseamento separada é calculada para cada imagem individual, como, por exemplo, para cada bloco. Em certas modalidades, uma correção de faseamento separada é calculada para cada bloco em cada ciclo.

[0077] À medida que as leituras aumentam, termos de ordem mais alta podem se tornar mais importantes na correção de faseamento. Assim, em modalidades particulares, para corrigir isto, pode ser calculada uma correção de faseamento empírica de segunda ordem. Por exemplo, em algumas modalidades, o método compreende uma correção de faseamento de segunda ordem, conforme definido pela equação a seguir:

I(ciclo)= -a*I(ciclo-2)- A *I(ciclo-I)+ I(ciclo)- B*I(ciclo+ 1)- b*I(ciclo+2)

onde I representa a intensidade e a, A, B e b representam os termos de primeira e segunda ordem para a correção de faseamento. Em modalidades particulares, o cálculo é otimizado sobre a, A, B e b.

[0078] A figura 5 representa esquematicamente uma configuração de processamento e a metodologia para realizar a correção de faseamento em tempo real. Na modalidade representada, um processador 502 cria um novo encadeamento de processamento 503 quando o processador é chamado para fazer chamadas de bases a partir de agrupamentos em uma imagem, por

exemplo, em uma imagem de um bloco. Um novo encadeamento pode ser gerado para cada ciclo de chamada de bases para cada bloco. Na modalidade representada, o processador 502 disponibiliza um único buffer de processador 505 para cada ciclo de chamada de bases de um bloco (e o segmento de processamento designado). O buffer do processador armazena temporariamente os valores de intensidade que são manipulados computacionalmente pelo processador para realizar a correção de fases para um ciclo de chamada de bases atual *n*. Na modalidade representada, o processador faz interface com uma memória do sistema 507 contendo três buffers, cada um para armazenar dados de imagem capturados para um determinado ciclo de chamada de bases. No caso da arquitetura da célula de fluxo descrita na figura 3, cada buffer armazena dados de imagem para os agrupamentos de um único bloco; portanto, os buffers são chamados de buffers do bloco. Obviamente, para outras arquiteturas de célula de fluxo e/ou sistemas de aquisição de imagem, os buffers podem armazenar mais ou menos dados de agrupamento. Por conveniência, o relatório descritivo se referirá aos buffers do bloco. Cada buffer do bloco armazena dados para um único bloco (ou outra parte de uma célula de fluxo) capturado durante um único ciclo de chamada de bases. Os dados da imagem podem ser providos como uma matriz de dados, como mostrado na figura 4.

[0079] Como representado, a memória de sistema 507 inclui um buffer de bloco 509 que armazena temporariamente os valores de intensidade para o ciclo de chamada de bases imediatamente anterior (em comparação com o ciclo de chamada de bases atual tratado pelo processador), um buffer do bloco 511 que armazena os valores de intensidade medidos para o ciclo de chamada de bases atual e um buffer do bloco 513 que armazena valores de intensidade para o ciclo de chamada de bases imediatamente subsequente. Novamente, cada um dos buffers do bloco 509, 511 e 513 contém dados medidos de um único bloco para um único ciclo de chamada de bases *n*.

[0080] Como mostrado, o encadeamento 503 usa os valores de intensidade em cada um dos buffers do bloco 509, 511 e 513 durante um único ciclo de chamada de bases. Os valores de intensidade são sucessivamente carregados no buffer do processador 505 e manipulados para implementar a expressão de correção de faseamento apresentada na parte inferior da figura 5. Depois que o processo de chamada de bases é concluído, conforme representado na configuração do processador e da memória da figura 5, o buffer do processador mantém valores de intensidade ajustados usados para fazer uma chamada de bases corrigida por faseamento.

[0081] A figura 6 apresenta um fluxograma de um processo de chamada bases que pode empregar a configuração do processador e da memória mostrada na figura 5. Como mostrado na Figura 6, um processo 601 inicia um novo ciclo de chamada de bases criando um encadeamento de processador e alocando um buffer de processador a esse encadeamento. Veja o bloco de processo 603. Depois disso, o processador extrai dados de intensidade de uma imagem de um bloco de célula de fluxo (ou outra porção apropriada da célula de fluxo) tomada simultaneamente com o ciclo de processamento atual. Na implementação representada, a imagem capturada e os valores de intensidade associados são os valores de intensidade primários para o próximo ciclo de chamada de bases sucessivo, e não o ciclo de chamada de bases atual (a iteração de processamento atual). Em outras palavras, o ciclo de processamento atual executa uma chamada de bases para dados de imagem coletados em um ciclo de processamento imediatamente anterior. Assim, como representado em um bloco de processo 605 do processo 601, os valores de intensidade extraídos recebem a referência I_{n+1}, onde n representa o ciclo de chamada de bases atual. Dito de outra maneira, um ciclo de processamento tanto (i) chama bases para o ciclo de chamada de bases n e (ii) captura dados de imagem para o ciclo de chamada de bases $n+1$.

[0082] Os dados de intensidade recém-extraídos, que podem ser providos na forma de uma matriz, como representado na figura 4, são armazenados em um buffer do bloco disponível na memória do sistema (por exemplo, no buffer do bloco 513). Em certas modalidades, este buffer do bloco é aquele que armazenou dados de intensidade que foram usados anteriormente, mas que não são mais necessários para as chamadas de bases.

[0083] No ciclo de processamento atual, o processo 601 também recupera dados de intensidade armazenados durante um ciclo computacional anterior ao ciclo computacional atual. Veja o bloco de processo 607. Os dados de intensidade recuperados são para o ciclo de chamada de bases atual e recebem a referência I_n. Os dados de intensidade recuperados são obtidos a partir de um buffer do bloco apropriado, tal como o buffer do bloco 511 da memória do sistema, como mostrado na figura 5.

[0084] Além disso, o processo 601 recupera dados de intensidade que foram armazenados dois ciclos anteriores ao ciclo de chamada de bases atual. Veja o bloco de processo 609. Como exemplo, com referência à figura 5, tais dados de intensidade podem ser obtidos a partir de um buffer do bloco 509 da memória do sistema. A matriz de valores de intensidade recuperados na operação 609 é identificada por I_{n-1}.

[0085] Enquanto as operações 605, 607 e 609 são mostradas como ocorrendo sequencialmente, esta ordem de operações é flexível e o processo pode ser implementado de tal modo que qualquer ordem seja aceitável, desde que seja consistente com a chamada de bases que incorpora a correção de faseamento.

[0086] Ao recuperar os valores de intensidade para o ciclo de chamada de bases atual (bloco de processo 607) e os valores de intensidade para o ciclo de chamada de bases imediatamente anterior (bloco de processamento 609), o processador tem disponível todos os valores de intensidade necessários para executar uma correção de faseamento. Ele faz

isso primeiro determinando o peso da correção pré-faseamento *b* e o peso da correção de faseamento *a* para o ciclo de chamada de bases atual. Veja o bloco de processo 611, que ilustra que isto pode ser conseguido usando os valores de intensidade extraídos para o próximo ciclo de chamada de bases subsequente juntamente com os valores de intensidade para os ciclos de chamada de bases atual e imediatamente anterior. Em seguida, utilizando os pesos de correção de faseamento e pré-faseamento, o processador calcula os valores de intensidade de faseamento corrigidos para o ciclo de chamada de bases atual, conforme é representado no bloco de processo 613. Os valores corrigidos são para os agrupamentos no bloco em análise. O cálculo pode empregar a expressão representada no bloco 613. Utilizando os valores de intensidade de faseamento corrigidos, o processador faz chamadas para o ciclo de chamada de bases atual, como representado no bloco de processo 615.

[0087] Neste ponto, o processamento para o ciclo de chamada de bases atual é concluído e a próxima iteração da chamada de bases pode ser executada. A decisão de se realizar outro ciclo de chamada de bases é representada em um bloco 617 que determina se existem nucleotídeos adicionais a serem sequenciados nos agrupamentos do bloco sob análise. Se não houver nenhum, o processo é concluído, como mostrado no bloco 619. Se houver algum, o controle de processo passa para um bloco de processo 621, onde o processador incrementa uma contagem de ciclo. Isso efetivamente indexa os valores de intensidade para o ciclo de chamada de bases atual I_n aos valores de intensidade para o ciclo de chamada de bases imediatamente anterior I_{n-1} . Ao mesmo tempo, os valores de intensidade para o ciclo de chamada de bases imediatamente subsequente (I_{n+1}) se tornam os valores de intensidade para o novo ciclo de chamada de bases atual (I_n). Esses incrementos são feitos com relação aos índices aplicados aos dados de intensidade armazenados nos buffers do bloco.

Processo de Faseamento (Memória Principal Reduzida)

[0088] A abordagem das figuras 5 e 6 pode funcionar bem, desde que o sequenciador e seu sistema de análise em tempo real associado não sejam limitados pela memória. No entanto, dada a quantidade de dados que devem ser processados em certos sequenciadores modernos, tais como aqueles empregados para realizar o sequenciamento completo do genoma, pode não haver memória suficiente disponível, particularmente a um custo comercialmente viável. Portanto, armazenar três vezes a quantidade de dados necessária para visualizar totalmente a célula de fluxo (ou as células de fluxo) durante um ciclo de chamada de bases pode representar uma séria restrição.

[0089] Um algoritmo de faseamento, tal como o representado nas figuras 5 e 6, é uma contribuição importante para a análise em tempo real, na medida em que melhora significativamente os resultados do sequenciamento, particularmente em amostras não padronizadas, como por exemplo, em amostras de baixa diversidade. No entanto, a carga de memória imposta torna-se maior à medida que o rendimento dos sistemas de sequenciamento de nova geração aumenta. As modalidades a seguir reduzem a carga de memória usando os pesos de faseamento aprendidos de dados que já estavam parcialmente corrigidos por faseamento. Os pesos de faseamento e pré-faseamento podem ser aprendidos independentemente e ainda assim prover resultados de sequenciamento de alta qualidade. Em certas modalidades, por exemplo, a memória tem uma capacidade de armazenamento inferior a duas vezes a capacidade necessária para armazenar os dados contidos no número total de blocos nas duas células de fluxo.

[0090] Em certas formas de realização, a configuração do processador e da memória para a chamada de bases corrigida por faseamento é ajustada para reduzir os requisitos de memória do sistema. Um exemplo de como isso funciona é mostrado na figura 7. Os valores de intensidade são corrigidos como descrito acima, por exemplo, os pesos de faseamento e pré-faseamento

são calculados e aplicados aos ciclos imediatamente anterior e imediatamente posterior. No entanto, no exemplo da figura 7, a memória do sistema 707 emprega apenas dois buffers do bloco para a correção de faseamento: o buffer do bloco 709 e o buffer do bloco 711. Neste exemplo, um processador 702 emprega um encadeamento de processamento 703 que, ao contrário do exemplo da figura 5, tem dois buffers de processador associados: um buffer de processador 705 para armazenar e operar nos valores de intensidade recuperados da memória 707 e um buffer do processador 706 para armazenar e usar os valores de intensidade de imagem recém-capturados I_{n+1}. No exemplo descrito, os buffers do processador são alocados na memória principal, mas isso nem sempre é necessário. Em algumas modalidades, os buffers do processador são alocados em uma memória física diferente ou até mesmo no chip do processador.

[0091] A substituição de buffers do bloco com os buffers do processador reduz efetivamente os requisitos totais de memória. Ao usar vários processadores e/ou processamento multitarefa, alguns processadores manipulam muitos blocos. Como exemplo, o número de blocos em um sistema pode ser da ordem de 1000 a 2000, enquanto o número de processadores que manipulam todos esses blocos é de cerca de vinte. Em teoria, esse sistema pode realizar uma redução de memória na ordem de 50x. Em algumas implementações, a redução é da ordem de 20x.

[0092] Nesta implementação, os valores de intensidade capturados das imagens do bloco no ciclo de processamento atual (I_{n+1}) são armazenados localmente no processador e usados para calcular os pesos de faseamento e pré-faseamento e posteriormente fazer uma chamada de bases. Em algumas implementações, somente após a conclusão desse processo os valores de intensidade capturados mais recentemente (I_{n+1}) são armazenados em um buffer do bloco na memória do sistema 707.

[0093] Em algumas modalidades, um processador e uma memória do

sistema são configurados conforme é representado na figura 8. Tal como acontece com a configuração do processador/memória na figura 7, um processador 802 emprega encadeamentos de processamento 803, sendo cada um associado a dois buffers do processador: um buffer do processador 805 para armazenar temporariamente valores de intensidade de uma memória do sistema 807 (buffer do bloco 811) e um buffer do processador 806 para armazenar temporariamente valores de intensidade capturados durante o ciclo de processamento atual (I_{n+1}). Para permitir que esta configuração funcione de forma eficiente e eficaz, os valores de intensidade armazenados no buffer do bloco 811 devem ser parcialmente corrigidos por faseamento. Exemplos de mecanismos para realizar isso são descritos abaixo. O buffer do processador 705 na figura 7 e o buffer do processador 805 na figura 8 carregam intensidades da memória principal e, em seguida, manipulam essas intensidades para gerar as intensidades corrigidas que são empregadas para a chamada de bases. No exemplo descrito, os buffers do processador são alocados na memória principal, mas isso nem sempre é necessário. Em algumas modalidades, os buffers do processador são alocados em uma memória física diferente ou até mesmo no chip do processador.

[0094] A figura 9 representa uma vista de alto nível de um processo 901 que pode ser utilizado com a configuração de processador e memória da figura 8 e, em algumas implementações, da figura 7. Conforme ilustrado na figura 9, o primeiro e o segundo ciclos de processamento utilizam informações insuficientes para realizar a correção de faseamento completa nos agrupamentos imageados em um bloco. No entanto, o faseamento não é um problema significativo nos primeiros ciclos.

[0095] Para realizar a correção de faseamento completa, o sequenciador requer três ciclos consecutivos de dados de imagem. No primeiro ciclo de processamento, o sequenciador não faz uma chamada de bases; ele meramente armazena dados de intensidade para o próximo

processamento, ou seja, o ciclo no qual a primeira chamada de bases é feita.

[0096] Como representado, o processo 901 começa em um bloco de processo 903 onde um encadeamento é criado para o primeiro ciclo de processamento. As instruções neste encadeamento orientam a extração de dados de intensidade de uma imagem dos agrupamentos durante o primeiro ciclo de sequenciamento (I_1), ou seja, o ciclo durante o qual os primeiros nucleotídeos dos agrupamentos são lidos. Veja o bloco de processo 905. Os dados da imagem são armazenados em um buffer do bloco na memória do sistema. Neste ponto, o primeiro ciclo de processamento é efetivamente concluído.

[0097] O processo continua em um bloco de processo 907 onde um novo encadeamento é criado em preparação para o segundo ciclo de processamento. Nesse processo, o primeiro e o segundo buffers do processador são alocados para o segundo ciclo de processamento. Veja o bloco 907. Coletivamente, os blocos de processo 907, 909, 911, 913, 915, 917, 919, 921 e 923 são executados durante o segundo ciclo de processamento, que é executado utilizando o encadeamento e os buffers do processador gerados no bloco de processo 907.

[0098] Como descrito, o processador extrai dados de intensidade da imagem para o próximo ciclo de chamada de bases (I_2) e armazena esses dados em um primeiro buffer do processador. Veja o bloco de processo 909. Em seguida, durante o segundo ciclo de processamento, o processador recupera os dados de intensidade armazenados no buffer do bloco durante o primeiro ciclo de processamento, cujos dados de intensidade são para o ciclo de chamada de bases atual (I_1). Veja o bloco 911. Usando os dados de intensidade coletados durante o primeiro e o segundo ciclos de processamento, o processador pode calcular um peso pré-faseamento *b* para o ciclo de chamada de bases atual (ou seja, as primeiras chamadas de bases nas leituras). Veja o bloco de processo 913. Com os valores de intensidade para

os dois primeiros ciclos e o peso pré-faseamento, o processador calcula os valores de dados de intensidade corrigidos para o segundo ciclo de chamada de bases (I_2). Os valores dos dados de intensidade corrigidos podem ser armazenados no segundo buffer do processador. Veja o bloco de processo 915. Em seguida, o processador faz as chamadas de bases para o segundo ciclo de chamada de bases usando os valores de dados de intensidade corrigidos obtidos no bloco 915. Veja o bloco de processo 917.

[0099] Neste ponto, o processo de sequenciamento está pronto para começar a preparar o próximo ciclo de chamada de bases. Ele começa em um bloco de processo 919 determinando um peso de correção de faseamento *a* usando os próximos (ou segundos) dados de intensidade do ciclo de chamada de bases (I_2) e os dados do ciclo de chamada de bases atual (I_1), que foram armazenados no buffer do bloco. Usando o peso de correção de faseamento *a*, o processador em seguida calcula os valores dos dados de intensidade corrigidos por faseamento (mas não corrigidos por pré-faseamento) a partir dos dados de intensidade atualmente não corrigidos (I_2) extraídos durante este segundo ciclo de processamento e os valores de dados de intensidade para o primeiro ciclo de processamento (I_1) de acordo com a expressão apresentada no bloco de processo 921. Isso resulta em uma matriz de valor de intensidade parcialmente corrigida ($I_{2(\text{parcialmente corrigida})}$) para o segundo ciclo de chamada de bases. O sequenciador terá que aguardar o próximo ciclo de processamento antes de realizar a correção pré-faseamento. No entanto, nesse ponto, muito do cálculo é concluído e os dados da matriz para uma única imagem podem ser armazenados em um buffer do bloco para uso no próximo ciclo de chamada de bases. Para este fim, o processador armazena os dados de intensidade corrigidos por faseamento (mas não por pré-faseamento) no buffer do bloco (de modo que $I_{2(\text{parcialmente corrigido})}$ substitui I_1 no buffer do bloco). Veja o bloco de processo 923.

[00100] Neste ponto, o primeiro e o segundo ciclos de processamento

são concluídos e as chamadas de bases são feitas para o primeiro ciclo de chamada de bases, que é o segundo ciclo de processamento. Os ciclos subsequentes de chamadas de bases podem ser executados com a correção de faseamento completa, conforme descrito na figura 10. Veja o bloco de processo 925.

[00101] A figura 10 mostra uma sequência de operações que podem ser realizadas durante um ciclo de processamento que efetua chamada de bases corrigidas por faseamento completo. Este ciclo pode ser realizado no terceiro e nos ciclos de processamento subsequentes ao sequenciar agrupamentos de um bloco. Em certas modalidades, a sequência de operações representada na figura 10 corresponde ao bloco de processo 925 da figura 9.

[00102] Conforme descrito, o processo começa alocando um encadeamento e os buffers de primeiro e segundo processadores associados. Veja o bloco de processo 1003. Em seguida, o processador extrai os valores dos dados de intensidade de uma imagem para o próximo ciclo de chamada de bases (I_{n+1}) e armazena esses valores em um primeiro buffer do processador. Veja o bloco de processo 1005. Simultaneamente, o processador recupera os valores de dados de intensidade parcialmente corrigidos que foram armazenados durante o ciclo de chamada de bases anterior (como um exemplo não limitativo, $I_{2\ (\text{parcialmente corrigido})}$ na modalidade da figura 9, ou $I_n - a(I_{n-1})$). Esses valores agora representam os valores de intensidade para o ciclo de chamada de bases atual (I_n). Eles foram armazenados anteriormente no buffer do bloco da memória do sistema e agora são recuperados a partir dele. Veja o bloco de processo 1007. Com os valores de dados de intensidade parcialmente corrigidos para o ciclo de chamada de bases atual, que foram corrigidos por faseamento, o processador precisa apenas realizar a correção pré-faseamento para concluir a correção dos dados de intensidade e fazer as chamadas de bases necessárias para o ciclo de chamada de bases atual. Para este fim, o processador determina o peso de correção pré-faseamento *b* para o

ciclo de chamada de bases atual. Ele faz isso usando os dados de intensidade extraídos que acabaram de ser recuperados dos dados da imagem, para o próximo ciclo (I_{n+1}), juntamente com os dados de intensidade parcialmente corrigidos para o ciclo de chamada de bases atual. Lembre-se de que esses dados parcialmente corrigidos foram recém-recuperados do buffer do bloco. Os dados de intensidade parcialmente corrigidos podem ser representados pela expressão $I_n - a(I_{n-1})$. Veja o bloco de processo 1009.

[00103] Com o peso da correção pré-faseamento *b* calculado para o ciclo de chamada de bases atual, o processador tem tudo o que precisa para calcular uma matriz de dados de intensidade totalmente corrigida por faseamento para o ciclo de chamada de bases atual (I_n). O cálculo é realizado conforme representado no bloco de processo 1009. Os valores dos dados de intensidade totalmente corrigidos resultantes são armazenados no segundo buffer do processador. Veja o bloco de processo 1011. Em seguida, o processador faz as chamadas de bases para o ciclo de chamada de bases atual usando os valores de dados de intensidade corrigidos armazenados no segundo buffer do processador. Veja o bloco de processo 1013.

[00104] O ciclo de processamento atual pode começar a se preparar para o próximo ciclo de chamada de bases que será executado durante o próximo ciclo de processamento. Na modalidade representada, o processador determina o peso da correção de faseamento *a* para o próximo ciclo de chamada de bases usando os dados de intensidade disponíveis para o ciclo de chamada de bases atual. Veja o bloco de processo 1015. Lembre-se de que os dados de intensidade do ciclo de chamada de bases seguinte foram extraídos e armazenados no primeiro buffer do processador na operação do processo 1005. Os valores de intensidade parcialmente corrigidos para o ciclo de chamada de bases atual foram recuperados do buffer do bloco para fazer as chamadas de bases atuais. Os mesmos valores de intensidade parcialmente corrigidos são agora usados para calcular o peso de correção de faseamento *a*

para o próximo ciclo de chamada de bases. Com o peso de correção de faseamento para o próximo ciclo de chamada de bases agora calculado, o processador calcula os valores de dados de intensidade corrigidos por faseamento (mas não corrigidos pré-faseamento), conforme representado no bloco de processo 1017. Em seguida, o processador armazena esses valores de dados de intensidade corrigidos por faseamento para o próximo ciclo de chamada de bases no buffer do bloco. Veja o bloco de processo 1019.

[00105] Antes desta invenção, supunha-se que a exatidão da chamada de bases se deterioraria aprendendo os pesos pré-faseamento a partir das intensidades corrigidas por faseamento. No entanto, os resultados apresentados mostram que há pouca ou nenhuma imprecisão. Em algumas implementações, os dados de imagem são comprimidos (por exemplo, por compressão com perda de dados) e até mesmo os dados corrigidos parcialmente faseados são comprimidos. Em ambos os casos, foi demonstrado que a compressão poderia ser realizada sem perda de precisão. Como exemplo, sem compressão, uma implementação usa dois buffers de flutuação para cada bloco (um buffer de flutuação tem tamanho de 4 bytes). Com a compressão, uma implementação usa um buffer de um único byte, alcançando assim menos 4x de memória.

[00106] Neste ponto, o ciclo de processamento atual é efetivamente concluído, de modo que o processador determina se há mais ciclos que precisam ser realizados no sequenciamento dos agrupamentos do bloco atual. Veja o bloco de decisão 1021. Se nenhuma base adicional precisar ser lida dos agrupamentos, o processo estará completo e nenhum ciclo de processamento adicional será realizado. No entanto, se um ou mais ciclos de sequenciamento adicionais forem necessários, o controle do processo é direcionado para um bloco de processo 1023, onde o processador incrementa o ciclo atual no qual os valores de dados de intensidade parcialmente corrigidos armazenados no buffer do bloco se tornam atuais; ou seja, eles se tornam os valores para o

novo ciclo de chamada de bases. O controle do processo retorna então ao bloco de processo 1003, onde o próximo ciclo de processamento começa.

EXEMPLO

[00107] Como explicado, certas modalidades reduzem a carga de memória usando os pesos de faseamento aprendidos de dados que já estavam parcialmente corrigidos por faseamento. No entanto, não ficou claro que os pesos de faseamento e pré-faseamento podem ser aprendidos de forma independente e ainda prover resultados de sequenciamento de alta qualidade. O exemplo apresentado na figura 11 estabelece que eles podem.

[00108] Como mostrado, duas comparações foram feitas, cada uma usando um processo de linha de base (por exemplo, um processo das figuras 5 e 6) e um novo processo que foi otimizado para reduzir os requisitos de memória principal (por exemplo, um processo das figuras 8 e 10). Em cada comparação, o mesmo sequenciador e amostra foram utilizados. Especificamente, um instrumento Illumina HiSeqX foi convertido para usar a química de 2 corantes. As imagens produzidas pelo sequenciador foram salvas e os dois algoritmos de faseamento foram testados nas mesmas imagens de sequenciamento, fornecendo um teste completamente controlado. O "Agrupamentos PF" indica o rendimento provido pelo sequenciador; o %Aligned indica o número de agrupamentos alinhados com sucesso ao genoma de referência, e o "% Error Rate" indica a taxa média de erro das sequências chamadas pelo software em comparação com o genoma de referência.

[00109] Os resultados do sequenciamento demonstram que o algoritmo de faseamento com memória eficiente é comparável ao algoritmo da linha de base. Neste exemplo, o processo eficiente de memória produziu um aumento de aproximadamente 3% na taxa de erro, que é compensado por uma redução na memória principal (estimada de 420 gigabytes a 340 gigabytes em algumas implementações).

MÉTODOS DE SEQUENCIAMENTO

[00110] Como indicado acima, a revelação refere-se ao sequenciamento de amostras de ácido nucleico. Qualquer uma das várias tecnologias de sequenciamento que usam um ou mais canais de informação para chamadas de bases, e particularmente canais ópticos, pode ser usada. Técnicas particularmente aplicáveis são aquelas em que os ácidos nucleicos são ligados em locais fixos em uma matriz (por exemplo, como agrupamento) e onde a matriz é repetidamente imageada. As modalidades nas quais as imagens são obtidas em diferentes canais de cor, por exemplo, coincidindo com diferentes marcadores utilizados para distinguir um tipo de base de nucleotídeo de outro, são particularmente aplicáveis. Em algumas modalidades, o processo para determinar a sequência nucleotídica de um ácido nucleico alvo pode ser um processo automatizado. Certas modalidades incluem técnicas de sequenciamento por síntese ("SBS"). Embora as técnicas de sequenciamento por síntese sejam aqui enfatizadas, outras tecnologias de sequenciamento podem ser utilizadas.

[00111] Em muitas implementações, as técnicas de SBS envolvem a extensão enzimática de uma fita de ácido nucleico nascente através da adição iterativa de nucleotídeos contra uma fita molde. Nos métodos tradicionais de SBS, pode ser provido um monômero de nucleotídeo simples a um nucleotídeo alvo na presença de uma polimerase em cada liberação. Contudo, nos métodos aqui descritos, mais de um tipo de monômero de nucleotídeo pode ser provido a um ácido nucleico alvo na presença de uma polimerase em uma liberação.

[00112] O SBS pode utilizar monômeros de nucleotídeo que possuem uma porção terminadora ou aqueles que não possuem quaisquer porções terminadoras. Métodos utilizando monômeros de nucleotídeo sem terminadores incluem, por exemplo, pirossequenciamento e sequenciamento utilizando nucleotídeos marcados com fosfato-γ. Nos métodos que utilizam

monômeros de nucleotídeo sem terminadores, o número de nucleotídeos adicionados em cada ciclo é geralmente variável e dependente da sequência molde e do modo de distribuição do nucleotídeo. Para as técnicas de SBS que utilizam monômeros de nucleotídeo com uma porção terminadora, o terminador pode ser eficazmente irreversível nas condições de sequenciamento utilizadas, como é o caso do Sanger de sequenciamento tradicional que utiliza didesoxinucleotídeos, ou o terminador pode ser reversível como é o caso dos métodos de sequenciamento desenvolvidos por Solexa (atualmente Illumina, Inc.).

[00113] As técnicas de SBS podem utilizar monômeros de nucleotídeo que possuem uma porção marcadora ou aqueles que não possuem uma porção marcadora. Por conseguinte, eventos de incorporação podem ser detectados com base em uma característica do marcador, tal como a fluorescência do marcador; uma característica do monômero de nucleotídeo, tal como peso molecular ou carga; um subproduto da incorporação do nucleotídeo, como a liberação de pirofosfato; ou similar. Nas modalidades onde dois ou mais nucleotídeos diferentes estão presentes em um reagente de sequenciamento, diferentes nucleotídeos podem ser distinguíveis uns dos outros, ou alternativamente, os dois ou mais marcadores diferentes podem ser indistinguíveis sob as técnicas de detecção sendo utilizadas. Por exemplo, os diferentes nucleotídeos presentes em um reagente de sequenciamento podem ter diferentes marcadores e podem ser distinguidos utilizando óptica apropriada, como exemplificado pelos métodos de sequenciamento desenvolvidos pela Solexa (atualmente Illumina, Inc.)..

[00114] Algumas modalidades incluem técnicas de pirossequenciamento. O pirossequenciamento detecta a liberação de pirofosfato inorgânico (PPi) à medida que nucleotídeos particulares são incorporados à fita nascente (Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlen, M. e Nyren, P. (1996) “Real-time DNA sequencing using detection of

pyrophosphate release." Analytical Biochemistry 242(1), 84-9; Ronaghi, M. (2001) "Pyrosequencing sheds light on DNA sequencing." Genome Res. 11(1), 3-11; Ronaghi, M., Uhlen, M. e Nyren, P. (1998) "A sequencing method based on real-time pyrophosphate." Science 281 (5375), 363; Patente Norte-Americana No. 6.210.891; Patente Norte-Americana No. 6.258.,568 e Patente Norte-Americana No. 6.274.320, cujas revelações são aqui incorporadas por referência em suas totalidades). No pirossequenciamento, o PPi libertado pode ser detectado sendo imediatamente convertido em trifosfato de adenosina (ATP) pela ATP sulfurilase, e o nível de ATP gerado é detectado através de fótons produzidos por luciferase. Os ácidos nucleicos a serem sequenciados podem ser ligados a características em uma matriz e a matriz pode ser imageada para capturar os sinais quimiluminescentes que são produzidos devido à incorporação dos nucleotídeos nos atributos da matriz. Uma imagem pode ser obtida após a matriz ser tratada com um tipo de nucleotídeo específico (por exemplo, A, T, C ou G). As imagens obtidas após a adição de cada tipo de nucleotídeo serão diferentes em relação a quais recursos da matriz são detectados. Essas diferenças na imagem refletem o conteúdo de sequência diferente dos atributos na matriz. No entanto, os locais relativos de cada atributo permanecerão inalterados nas imagens. As imagens podem ser armazenadas, processadas e analisadas usando os métodos definidos na presente invenção. Por exemplo, as imagens obtidas após o tratamento da matriz com cada tipo de nucleotídeo diferente podem ser tratadas da mesma maneira como exemplificado aqui para imagens obtidas de diferentes canais de detecção para os métodos de sequenciamento baseados em terminadores reversíveis.]

[00115] Em outro tipo exemplar de SBS, o sequenciamento do ciclo é alcançado pela adição gradual de nucleotídeos terminadores reversíveis contendo, por exemplo, um marcador de corante clivável ou fotoalvejável, como descrito, por exemplo, no documento WO 04/018497 e na Patente

Norte-Americana No. 7.057.026, cujas revelações são incorporadas na presente invenção por referência. Esta abordagem está sendo comercializada pela Solexa (atualmente Illumina Inc.)e também está descrita nos documentos WO 91/06678 e WO 07/123.744, sendo cada um deles aqui incorporado por referência. A disponibilidade dos terminadores fluorescentemente marcados, nos quais tanto a terminação pode ser revertida como marcador fluorescente pode ser clivado facilita o eficiente sequenciamento da terminação reversível cíclica (CRT). As polimerases também podem ser coprojetadas para incorporar e se prolongar eficientemente a partir desses nucleotídeos modificados.

[00116] Nas modalidades de sequenciamento baseadas em terminador reversíveis, os marcadores podem não inibir substancialmente o prolongamento sob as condições reacionais de SBS. No entanto, os marcadores de detecção podem ser removíveis, por exemplo, por clivagem ou degradação. As imagens podem ser capturadas após a incorporação dos marcadores nos atributos do ácido nucleico na matriz. Em modalidades particulares, cada ciclo envolve a liberação simultânea de quatro tipos de nucleotídeos diferentes na matriz e cada tipo de nucleotídeo possui um marcador espectralmente distinto. Quatro imagens podem ser obtidas, cada uma usando um canal de detecção que é seletivo para um dos quatro marcadores diferentes. Alternativamente, diferentes tipos de nucleotídeos podem ser adicionados sequencialmente e uma imagem da matriz pode ser obtida entre cada etapa de adição. Em tais modalidades, cada imagem mostrará atributos de ácido nucleico que têm nucleotídeos incorporados de um tipo particular. Diferentes atributos estarão presentes ou ausentes nas diferentes imagens devido ao conteúdo de sequência diferente de cada atributo. No entanto, a posição relativa dos recursos permanecerá inalterada nas imagens. As imagens obtidas a partir de tais métodos de terminador-SBS reversíveis podem ser armazenadas, processadas e analisadas, como definido

na presente invenção. Após a etapa de captura da imagem, os marcadores podem ser removidos e as porções do terminador reversíveis podem ser removidas para os ciclos subsequentes de adição e detecção de nucleotídeos. A remoção dos marcadores depois de terem sido detectados em um ciclo específico e antes de um ciclo subsequente pode prover a vantagem de reduzir o sinal de fundo e a interferência entre os ciclos.

[00117] Em modalidades particulares, alguns ou todos os monômeros de nucleotídeo podem incluir terminadores reversíveis. Nessas modalidades, os terminadores reversíveis/átomos de flúor cliváveis podem incluir flúor ligado à porção de ribose através de uma ligação éster 3' (Metzker, Genome Res. 15: 1767-1776 (2005), que é aqui incorporada por referência). Outras abordagens separaram a química do terminador da clivagem do marcador de fluorescência (Ruparel et al., Proc Natl Acad Sci USA 102: 5932-7 (2005), que é aqui incorporado por referência em sua totalidade). Ruparel et al descreveram o desenvolvimento de terminadores reversíveis que usavam um pequeno grupo alila 3' para bloquear a extensão, mas poderiam ser facilmente desbloqueados por um tratamento curto com um catalisador de paládio. O fluoróforo foi ligado à base por meio de um ligante fotoclivável que poderia ser facilmente clivado por uma exposição de 30 segundos à luz UV de comprimento de onda longo. Assim, a redução do dissulfeto ou a fotoclivagem podem ser utilizadas como um ligante clivável. Outra abordagem para a terminação reversível é o uso da terminação natural que ocorre após a colocação de um corante volumoso em um dNTP. A presença de um corante volumoso carregado no dNTP pode agir como um terminador eficaz através do impedimento estereoquímico e/ou eletrostático. A presença de um evento de incorporação impede novas incorporações, a menos que o corante seja removido. A clivagem do corante remove o flúor e efetivamente inverte a terminação. Exemplos de nucleotídeos modificados também são descritos na Patente Norte-Americana No. 7.427.673 e na Patente Norte-

Americana No. 7.057.026, cujas revelações são aqui incorporadas por referência em suas totalidades.

[00118] Sistemas e métodos de SBS exemplares adicionais que podem ser utilizados com os métodos e sistemas aqui descritos são descritos na Publicação do Pedido de Patente Norte-Americano No. 2007/0166705, na Publicação do Pedido de Patente Norte-Americano No. 2006/0188901, na Patente Norte-Americana No. 7.057.026, na Publicação do Pedido de Patente Norte-Americano No. 2006/0240439, na Publicação do Pedido de Patente Norte-Americano No. 2006/0281109, na Publicação PCT No. WO 05/065814, na Publicação do Pedido de Patente Norte-Americano No. 2005/0100900, na Publicação PCT No. WO 06/064199, na Publicação PCT No. WO 07/010.251, na Publicação do Pedido de Patente Norte-Americano No. 2012/0270305 e na Publicação do Pedido de Patente Norte-Americano No. 2013/0260372, cujas revelações são aqui incorporadas por referência em suas totalidades.

[00119] Algumas modalidades podem utilizar a detecção de quatro nucleotídeos diferentes utilizando menos de quatro marcadores diferentes. Por exemplo, o SBS pode ser realizado utilizando métodos e sistemas descritos nos materiais incorporados da Publicação do Pedido de Patente Norte-Americano No. 2013/0079232. Como primeiro exemplo, um par de tipos de nucleotídeos pode ser detectado no mesmo comprimento de onda, mas distinguido com base em uma diferença de intensidade para um membro do par em comparação com o outro, ou baseado em uma mudança para um membro do par (por exemplo, através de modificação química, modificação fotoquímica ou modificação física) que faz com que o sinal aparente apareça ou desapareça em comparação com o sinal detectado para o outro membro do par. Como um segundo exemplo, três dos quatro tipos diferentes de nucleotídeos podem ser detectados sob condições particulares, enquanto um quarto tipo de nucleotídeo não possui um marcador detectável nessas

condições, ou é minimamente detectado sob essas condições (por exemplo, detecção mínima devido ao sinal de fluorescência de fundo, etc.). A incorporação dos primeiros três tipos de nucleotídeos em um ácido nucleico pode ser determinada com base na presença dos seus respectivos sinais e a incorporação do quarto tipo de nucleotídeo no ácido nucleico pode ser determinada com base na ausência ou detecção mínima de qualquer sinal. Como terceiro exemplo, um tipo de nucleotídeo pode incluir marcador(es) que são detectados em dois canais diferentes, enquanto outros tipos de nucleotídeos são detectados em não mais do que um dos canais. As três configurações exemplificativas acima mencionadas não são consideradas mutuamente exclusivas e podem ser usadas em várias combinações. Uma modalidade exemplificadora que combina todos os três exemplos, um método de SBS baseado em fluorescência que utiliza um primeiro tipo de nucleotídeo que é detectado em um primeiro canal (por exemplo, dATP tendo um marcador que é detectado no primeiro canal quando excitado por um primeiro comprimento de onda de excitação), um segundo tipo de nucleotídeo que é detectado em um segundo canal (por exemplo, dCTP tendo um marcador que é detectado no segundo canal quando excitado por um segundo comprimento de onda de excitação), um terceiro tipo de nucleotídeo que é detectado em um primeiro e em um segundo canal (por exemplo, dTTP tendo pelo menos um marcador que é detectado em ambos os canais quando excitado pelo primeiro e/ou pelo segundo comprimento de onda de excitação) e um quarto tipo de nucleotídeo que não possui um marcador que não é, ou é minimamente detectado em qualquer canal (por exemplo, dGTP sem marcador).

[00120] Além disso, conforme descrito nos materiais incorporados da Publicação do Pedido de Patente Norte-Americano No. 2013/0079232, os dados de sequenciamento podem ser obtidos usando um único canal. Em tais abordagens de sequenciamento com um corante, o primeiro tipo de nucleotídeo é marcado, mas o marcador é removido depois que a primeira

imagem é gerada, e o segundo tipo de nucleotídeo é marcado somente após uma primeira imagem ser gerada. O terceiro tipo de nucleotídeo retém o seu marcador tanto na primeira como na segunda imagem, e o quarto tipo de nucleotídeo permanece não marcado em ambas as imagens.

[00121] Algumas modalidades podem utilizar o sequenciamento por técnicas de ligação. Tais técnicas utilizam a DNA ligase para incorporar oligonucleotídeos e identificar a incorporação de tais oligonucleotídeos. Os oligonucleotídeos têm tipicamente marcadores diferentes que estão correlacionados com a identidade de um nucleotídeo particular em uma sequência com a qual os oligonucleotídeos hibridizam. Tal como com outros métodos de SBS, as imagens podem ser obtidas após o tratamento de uma série de atributos de ácido nucleico com os reagentes de sequenciamento marcados. Cada imagem mostrará atributos do ácido nucleico que incorporaram marcadores de um tipo particular. Diferentes atributos estarão presentes ou ausentes nas diferentes imagens devido ao conteúdo de sequência diferente de cada atributo, mas a posição relativa dos atributos permanecerá inalterada nas imagens. As imagens obtidas a partir de métodos de sequenciamento baseados em ligação podem ser armazenados, processados e analisados, como definido na presente invenção. Sistemas e métodos de SBS exemplares que podem ser utilizados com os métodos e sistemas aqui descritos são descritos na Patente Norte-Americana No. 6.969.488, na Patente Norte-Americana No. 6.172.218 e na Patente Norte-Americana No. 6.306.597, cujas revelações são aqui incorporadas por referência em suas totalidades.

[00122] Algumas modalidades podem utilizar sequenciamento por nanoporos (Deamer, D. W. & Akeson, M. "Nanopores and nucleic acids: prospects for ultrarapid sequencing." Trends Biotechnol. 18, 147-151 (2000); Deamer, D. e D. Branton, "Characterization of nucleic acids by nanopore analysis". Acc. Chem. Res. 35:817-825 (2002); Li, J., M. Gershow, D. Stein,

E. Brandin, e J. A. Golovchenko, "DNA molecules and configurations in a solid-state nanopore microscope" Nat. Mater. 2: 611-615 (2003), cujas revelações são aqui incorporadas por referência em suas totalidades). Em tais modalidades, o ácido nucleico alvo passa através de um nanoporo. O nanoporo pode ser um poro sintético ou uma proteína de membrana biológica, como α-hemolisina. À medida que o ácido nucleico alvo passa através do nanoporo, cada par de bases pode ser identificado medindo as flutuações na condutância elétrica do poro. Patente Norte-Americana No. 7.001.792; Soni, G. V. & Meller, "A. Progress toward ultrafast DNA sequencing using solid-state nanopores." Clin. Chem. 53, 1996-2001 (2007); Healy, K. "Nanopore-based single-molecule DNA analysis." Nanomed. 2, 459-481 (2007); Cockroft, S. L., Chu, J., Amorin, M. & Ghadiri, M. R. "A single-molecule nanopore device detects DNA polymerase activity with single-nucleotide resolution." J. Am. Chem. Soc. 130, 818-820 (2008), cujas revelações são aqui incorporadas por referência em suas totalidades). Os dados obtidos a partir do sequenciamento por nanoporos podem ser armazenados, processados e analisados, como definido na presente invenção. Em particular, os dados podem ser tratados como uma imagem de acordo com o tratamento exemplificativo de imagens ópticas e de outras imagens que são apresentadas na presente invenção.

[00123] Algumas modalidades podem utilizar métodos envolvendo o monitoramento em tempo real da atividade da DNA polimerase. As incorporações de nucleotídeos podem ser detectadas por meio de interações de transferência de energia de ressonância por fluorescência (FRET) entre uma polimerase portadora de fluoróforo e nucleotídeos marcados com fosfato γ, como descrito, por exemplo, na Patente Norte-Americana No. 7.329.492 e na Patente Norte-Americana No. 7.211.414 (cada uma das quais sendo aqui incorporada por referência) ou as incorporações de nucleotídeos podem ser detectadas com guias de onda de modo zero como descrito, por exemplo, na

Patente Norte-Americana No. 7.315.019 (que é aqui incorporada por referência) e utilizando análogos de nucleotídeo fluorescentes e polimerases manipuladas, como descrito, por exemplo, na Patente Norte-Americana No. 7.405.281 e na Publicação do Pedido de Patente Norte-Americano No. 2008/0108082 (cada um dos quais sendo aqui incorporado a título de referência). A iluminação pode ser restrita a um volume em escala de zeptólito ao redor de uma polimerase presa à superfície, de tal forma que a incorporação dos nucleotídeos fluorescentemente marcados pode ser observada com um baixo plano de fundo (Levene, M. J. et al. "Zero-mode waveguides for single-molecule analysis at high concentrations." Science 299, 682-686 (2003); Lundquist, P. M. et al. "Parallel confocal detection of single molecules in real time." Opt. Lett. 33, 1026-1028 (2008); Korlach, J. et al. "Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nano structures." Proc. Natl. Acad. Sci. USA 105, 1176-1181 (2008), cujas revelações são aqui incorporadas por referência em suas totalidades). As imagens obtidas a partir de tais métodos podem ser armazenadas, processadas e analisadas, como definido na presente invenção.

[00124] Algumas modalidades de SBS incluem a detecção de um próton liberado por meio da incorporação de um nucleotídeo em um produto de extensão. Por exemplo, o sequenciamento baseado na detecção de prótons liberados pode usar um detector elétrico e técnicas associadas que são comercialmente disponíveis junto à Ion Torrent (Guilford, CT, uma subsidiária da Life Technologies) ou métodos e sistemas de sequenciamento descritos nos documentos US 2009/0026082 A1; US 2009/0127589 A1; US 2010/0137143 A1; ou US 2010/0282617 A1, sendo cada um dos quais aqui incorporado por referência. Os métodos aqui apresentados para amplificação dos ácidos nucleicos alvo utilizando exclusão cinética podem ser facilmente aplicados a substratos utilizados para a detecção de prótons. Mais

especificamente, os métodos apresentados na presente invenção podem ser usados para produzir populações clonais de amplicons que são usados para detectar prótons.

[00125] Os métodos SBS acima podem ser vantajosamente realizados em formatos multiplex, de modo que múltiplos ácidos nucleicos alvo diferentes sejam manipulados simultaneamente. Em modalidades particulares, podem -se tratar diferentes ácidos nucleicos alvo em um recipiente reacional comum, ou em uma superfície de um substrato particular. Isto possibilita a liberação conveniente de reagentes de sequenciamento, a remoção de reagentes que não reagiram e a detecção de eventos de incorporação de maneira multiplex. Em modalidades utilizando ácidos nucleicos alvo ligados à superfície, os ácidos nucleicos alvo podem estar em um formato de matriz. Em um formato de matriz, os ácidos nucleicos alvo podem ser tipicamente ligados a uma superfície de uma maneira espacialmente distinguível. Os ácidos nucleicos alvo podem ser ligados por ligação covalente direta, ligação a uma esfera ou outra partícula, ou ligados a uma polimerase ou outra molécula que está ligada à superfície. A matriz pode incluir uma única cópia de um ácido nucleico alvo em cada sítio (também referido como um atributo) ou várias cópias com a mesma sequência podem estar presentes em cada sítio ou atributo. Múltiplas cópias podem ser produzidas por métodos de amplificação, como amplificação em ponte ou PCR em emulsão.

[00126] Os métodos definidos na presente invenção podem usar matrizes tendo atributos em qualquer uma dentre várias densidades incluindo, por exemplo, pelo menos cerca de 10 atributos/cm^2, 100 atributos/cm^2 , 500 atributos/cm^2, 1.000 atributos/cm^2 , 5.000 atributos/cm^2, 10.000 atributos/cm^2, 50.000 atributos/cm^2, 100.000 atributos/cm^2, 1.000.000 atributos/cm^2, 5.000.000 atributos/cm^2 ou mais.

[00127] Os métodos definidos na presente invenção podem proporcionar uma detecção rápida e eficiente de uma pluralidade de ácidos

nucleicos alvo simultaneamente. Consequentemente, a presente revelação provê sistemas integrados capazes de preparar e detectar ácidos nucleicos utilizando técnicas conhecidas na arte, como aquelas exemplificadas acima. Assim, um sistema integrado da presente revelação pode incluir componentes fluidos capazes de distribuir reagentes de amplificação e/ou reagentes de sequenciamento para um ou mais fragmentos de DNA imobilizados, com o sistema compreendendo componentes tais como bombas, válvulas, reservatórios, linhas fluidas e similares. Uma célula de fluxo pode ser configurada e/ou usada em um sistema integrado para detecção de ácidos nucleicos alvo. Células de fluxo exemplificadoras são descritas, por exemplo, nos documentos US 2010/0111768 A1 e no documento US No. de série 13/276366, sendo cada uma deles aqui incorporado a título de referência. Como exemplificado para as células de fluxo, um ou mais dos componentes fluidos de um sistema integrado podem ser utilizados tanto para um método de amplificação como para um método de detecção. Tomando-se como exemplo uma modalidade de sequenciamento de ácido nucleico, um ou mais dos componentes fluidos de um sistema integrado podem ser utilizados para um método de amplificação definido na presente invenção e para a liberação dos reagentes de sequenciamento em um método de sequenciamento, tal como os exemplificados acima. Alternativamente, um sistema integrado pode incluir sistemas fluidos separados para realizar métodos de amplificação e para realizar métodos de detecção. Exemplos de sistemas de sequenciamento integrados que são capazes de criar ácidos nucleicos amplificados e também determinar a sequência dos ácidos nucleicos incluem, sem limitação, a plataforma MiSeq™ (Illumina, Inc., São Diego, CA) e os dispositivos descrito no documento US No. de Série 13/276.366, que é incorporado na presente invenção por referência.

[00128] Em algumas modalidades dos métodos descritos na presente invenção, as etiquetas de sequência mapeadas compreendem leituras de

sequências de cerca de 20 pb, cerca de 25 pb, cerca de 30 pb, cerca de 35 pb, cerca de 40 pb, cerca de 45 pb, cerca de 50 pb, cerca de 55 pb, cerca de 60 pb, cerca de 65 pb, cerca de 70 pb, cerca de 75 pb, cerca de 80 pb, cerca de 85 pb, cerca de90 pb, cerca de 95 pb, cerca de 100 pb, cerca de 110 pb, cerca de 120 pb, cerca de 130, cerca de 140 pb, cerca de 150 pb, cerca de 200 pb, cerca de 250 pb, cerca de 300 pb, cerca de 350 pb, cerca de 400 pb, cerca de 450 pb ou cerca de 500 pb. Em alguns casos, leituras de extremidade única maiores que 500 pb são utilizadas para leituras de mais de 1.000 pb, quando leituras finais em pares são geradas. O mapeamento das etiquetas de sequência é conseguido comparando a sequência da etiqueta com a sequência da referência para determinar a origem cromossômica da molécula de ácido nucleico sequenciada, e não são necessárias as informações sobre a sequência genética específica. Um pequeno grau de incompatibilidade (0 a 2 incompatibilidades por etiqueta de sequência) pode ser permitido que representa polimorfismos menores que podem existir entre o genoma de referência e os genomas na amostra misturada.

SISTEMAS E APARELHOS PARA ANÁLISE EM TEMPO REAL DOS DADOS DE SEQUENCIAMENTO

[00129] A análise dos dados de sequenciamento é tipicamente realizada usando vários algoritmos e programas executados por computador. Portanto, certas modalidades empregam processos envolvendo dados armazenados ou transferidos através de um ou mais sistemas de computador ou outros sistemas de processamento. As modalidades reveladas na presente invenção também se referem aos aparelhos para realizar estas operações. Este aparelho pode ser especialmente construído para os fins requeridos, ou pode ser um computador (ou um grupo de computadores) de aplicação geral seletivamente ativado ou reconfigurado por um programa de computador e/ou estrutura de dados armazenada no computador. Em algumas modalidades, um grupo de processadores realiza algumas ou todas as operações analíticas

citadas de forma colaborativa (por exemplo, através de uma rede ou computação em nuvem) e/ou em paralelo. Um processador ou grupo de processadores para executar os métodos descritos na presente invenção pode ser de vários tipos, incluindo microcontroladores e microprocessadores, tais como dispositivos programáveis (por exemplo, CPLDs e FPGAs) e dispositivos não programáveis, como ASICs de porta matriz ou microprocessadores de uso geral.

[00130] Além disso, certas modalidades referem-se a mídias legíveis por computador não transitórias e/ou tangíveis ou a produtos de programa de computador que incluem instruções de programas e/ou dados (incluindo estruturas de dados) para executar várias operações implementadas por computador. Exemplos de mídias legíveis por computador incluem, mas não se limitam, aos dispositivos de memória semicondutores, mídia magnética, como unidades de disco, fita magnética, mídia óptica como CDs, mídia magneto-óptica e dispositivos de hardware que são especialmente configurados para armazenar e executar instruções do programa, como dispositivos de memória somente de leitura (ROM) e de memória de acesso aleatório (RAM). As mídias legíveis por computador podem ser controladas diretamente por um usuário final ou a mídia pode ser indiretamente controlada pelo usuário final. Exemplos de mídias controladas diretamente incluem a mídia localizada em uma instalação de usuário e/ou mídia que não é compartilhada com outras entidades. Exemplos de mídias controladas indiretamente incluem mídias que são indiretamente acessíveis ao usuário por meio de uma rede externa e/ou por meio de um serviço que provê recursos compartilhados, como a "nuvem". Exemplos de instruções de programa incluem tanto código de máquina, como aquele produzido por um compilador, quanto arquivos contendo código de nível mais alto que podem ser executados pelo computador usando um intérprete.

[00131] Em várias modalidades, os dados ou informações utilizados

nos métodos e aparelhos revelados são providos em um formato eletrônico. Tais dados ou informações podem incluir leituras derivadas de uma amostra de ácido nucleico, contagens ou densidades de tais etiquetas que se alinham com regiões específicas de uma sequência de referência (por exemplo, que se alinham a um cromossomo ou segmento cromossômico), distâncias de separação entre leituras ou fragmentos adjacentes, distribuições de tais distâncias de separação, diagnósticos e similares. Conforme utilizado na presente invenção, dados ou outras informações providas em formato eletrônico estão disponíveis para armazenamento em uma máquina e transmissão entre máquinas. Convencionalmente, os dados em formato eletrônico são providos digitalmente e podem ser armazenados como bits e/ou bytes em várias estruturas de dados, listas, bancos de dados, etc. Os dados podem ser incorporados eletronicamente, opticamente, etc.

[00132] Uma modalidade provê um produto de programa de computador para determinar os coeficientes de faseamento e de pré-faseamento, bem como valores de magnitude corrigidos por faseamento e as chamadas de bases associadas. O produto de computador pode conter instruções para realizar qualquer um ou mais dos métodos descritos acima para o faseamento e a chamada de bases. Como explicado, o produto de computador pode incluir uma mídia legível por computador não transitória e/ou tangível tendo uma lógica executável ou compilável por computador (por exemplo, instruções) nele registradas para permitir que um processador alinhe leituras, identifique fragmentos e/ou ilhas de leituras alinhadas, identifique alelos, incluindo alelos de inserção/eliminação, de polimorfismos heterozigotos, porções de fase dos cromossomos e cromossomos e genomas haplótipos. Em um exemplo, o produto de computador inclui (1) uma mídia legível por computador com uma lógica executável ou compilável por computador (por exemplo, instruções) armazenada nele para que um processador realiza correção de faseamento nos dados de magnitude (por

exemplo, dados de intensidade de cor de dois ou mais canais) em amostras de ácido nucleico; (2) lógica assistida por computador para fazer chamadas de bases das amostras de ácido nucleico; e (3) um procedimento de saída para gerar um resultado caracterizando as amostras de ácido nucleico.

[00133] Deve ser compreendido que não é prático, ou até mesmo possível na maioria dos casos, para um ser humano sem ajuda realizar as operações computacionais dos métodos reveladas na presente invenção. Por exemplo, gerar coeficientes de faseamento para até mesmo um único bloco durante um único ciclo de chamada de bases pode exigir anos de esforço sem a assistência de um aparelho computacional. Naturalmente, o problema é agravado porque o sequenciamento confiável de NGS geralmente requer correção de faseamento e chamada de bases para pelo menos milhares ou até milhões de leituras.

[00134] Os métodos revelados na presente invenção podem ser realizados utilizando um sistema para sequenciar amostras de ácido nucleico. O sistema pode incluir: (a) um sequenciador para receber ácidos nucleicos da amostra de teste, fornecendo informações da sequência de ácidos nucleicos da amostra; (b) um processador; e (c) uma ou mais mídias de armazenamento legíveis por computador, tendo armazenado nela instruções para execução no processador, para avaliar os dados do sequenciador. A mídia de armazenamento legível por computador também pode armazenar dados de magnitude de faseamento parcialmente corrigidos dos agrupamentos em uma célula de fluxo.

[00135] Em algumas modalidades, os métodos são instruídos por uma mídia legível por computador tendo armazenado na mesma instruções legíveis por computador para realizar um método de determinação da fase de uma sequência. Deste modo, uma modalidade proporciona um produto de programa de computador que inclui uma ou mais mídias de armazenamento legível(is) por computador não transitória(s), tendo armazenada(s) na(s)

mesma(s) instruções executáveis no computador que, quando executadas por um ou mais processadores de um sistema computacional, fazem com que o sistema computacional implemente um método para o sequenciamento de uma amostra de DNA. O método inclui: (a) obter dados representando uma imagem (por exemplo, a própria imagem) de um substrato compreendendo uma pluralidade de sítios onde as bases de ácido nucleico são lidas; (b) obter valores de cor (ou outros valores que representam bases/nucleotídeos individuais) da pluralidade de sítios da imagem do substrato; (c) armazenar os valores de cor em um buffer do processador; (d) recuperar valores de cor parcialmente corrigidos na fase da pluralidade de sítios para um ciclo de chamada de bases, onde os valores de cor parcialmente corrigidos na fase foram armazenados na memória do sequenciador durante um ciclo de chamada de bases imediatamente anterior; (e) determinar uma correção pré-faseamento a partir dos (i) valores de cor parcialmente corrigidos na fase armazenados durante o ciclo de chamada de bases imediatamente anterior, e (ii) dos valores de cor armazenados no buffer do processador; e (f) determinar valores de cor corrigidos dos (i) valores de cor no buffer do processador, dos (ii) valores parcialmente corrigidos da fase armazenados durante o ciclo imediatamente anterior e da (iii) correção pré-faseamento.

[00136] Dados de sequência ou de outra natureza podem ser inseridos em um computador ou armazenados em uma mídia legível por computador, direta ou indiretamente. Em várias modalidades, um sistema computacional está incorporado ou é diretamente acoplado a um dispositivo de sequenciamento que lê e/ou analisa sequências de ácidos nucleicos das amostras. As sequências ou outras informações de tais ferramentas são providas ao sistema computacional (ou simplesmente ao hardware de processamento integrado) por meio de uma interface de transmissão de dados. Além disso, o dispositivo de memória pode armazenar leituras, informações de qualidade de chamadas de bases, informações de coeficientes de

faseamento, etc. A memória também pode armazenar várias rotinas e/ou programas para analisar e apresentar os dados da sequência. Tais programas/rotinas podem incluir programas para realizar análises estatísticas, etc.

[00137] Em um exemplo, um usuário provê uma amostra em um aparelho de sequenciamento. Os dados são coletados e/ou analisados pelo aparelho de sequenciamento conectado a um computador. O software no computador permite a coleta e/ou a análise de dados. Os dados podem ser armazenados, exibidos (por meio de um monitor ou outro dispositivo similar) e/ou enviados para outro local. O computador pode estar conectado à internet, que é usada para transmitir os dados para um dispositivo portátil utilizado por um usuário remoto (por exemplo, um médico, cientista ou analista). Entende-se que os dados podem ser armazenados e/ou analisados antes da transmissão. Em algumas modalidades, os dados brutos são coletados e enviados para um usuário ou aparelho remoto que analisará e/ou armazenará os dados. Por exemplo, as leituras podem ser transmitidas à medida que são geradas, ou logo em seguida, e alinhadas, e outras podem ser analisadas remotamente. A transmissão pode ocorrer através da internet, mas também pode ocorrer via satélite ou outro tipo de conexão. Alternativamente, os dados podem ser armazenados em uma mídia legível por computador, e a mídia pode ser enviada para um usuário final (por exemplo, via correio). O usuário remoto pode estar no mesmo local geográfico ou em uma localização geográfica diferente, incluindo, sem limitação, um prédio, cidade, estado, país ou continente.

[00138] Em algumas modalidades, os métodos também incluem a coleta de dados em relação a uma pluralidade de sequências polinucleotídicas (por exemplo, leituras) e o envio dos dados para um computador ou outro sistema computacional. Por exemplo, o computador pode ser conectado ao equipamento laboratorial, por exemplo, a um aparelho de coleta de amostras,

a um aparelho de amplificação de polinucleotídeo ou a um aparelho de sequenciamento de nucleotídeos. Os dados coletados ou armazenados podem ser transmitidos do computador para um local remoto, por exemplo, através de uma rede local ou de uma rede de área ampla, como a internet. No local remoto, várias operações podem ser executadas nos dados transmitidos.

[00139] Em algumas modalidades de qualquer um dos sistemas aqui providos, o sequenciador é configurado para executar o sequenciamento de nova geração (NGS). Em algumas modalidades, o sequenciador é configurado para realizar o sequenciamento paralelo em massa utilizando sequenciamento por síntese com terminadores de corante reversíveis. Em outras modalidades, o sequenciador é configurado para realizar o sequenciamento de uma molécula única.

CONCLUSÃO

[00140] A presente revelação pode ser apresentada em outras formas específicas sem se afastar do seu espírito ou das suas características essenciais. As modalidades descritas devem ser consideradas em todos os aspectos apenas como ilustrativas e não de forma restritiva. O escopo da revelação, portanto, é indicado pelas reivindicações em anexo ao invés de pela descrição acima. Todas as alterações que estão dentro do significado e da faixa de equivalência das reivindicações devem ser abrangidas por seu escopo.

REIVINDICAÇÕES

1. Método para determinar os valores de cor corrigidos a partir de dados de imagem obtidos, durante um ciclo de chamada de bases, por um sequenciador de ácidos nucleicos compreendendo um sistema de aquisição de imagem, um ou mais processadores e memória, o método caracterizado pelo fato de que compreende:

(a) obter uma imagem de um substrato compreendendo uma pluralidade de sítios onde as bases de ácido nucleico são lidas, em que os sítios exibem cores representando os tipos de base de ácido nucleico;

(b) medir os valores de cor da pluralidade de sítios a partir da imagem do substrato;

(c) armazenar os valores de cor em um buffer do processador do um ou mais processadores do sequenciador;

(d) recuperar os valores de cor parcialmente corrigidos na fase da pluralidade de sítios, em que os valores de cor parcialmente corrigidos na fase foram armazenados na memória do sequenciador durante um ciclo de chamada bases imediatamente anterior;

(e) determinar uma correção pré-faseamento a partir dos

valores de cor parcialmente corrigidos na fase armazenados durante o ciclo de chamada de bases imediatamente anterior e

valores de cor armazenados no buffer do processador; e

(f) determinar os valores de cor corrigidos dos

valores de cor no buffer do processador,

valores cor parcialmente corrigidos na fase armazenados durante o ciclo imediatamente anterior, e

a correção pré-faseamento.

2. Método de acordo com a reivindicação 1, caracterizado pelo fato de que compreende adicionalmente usar os valores de cor corrigidos para fazer chamadas de base para a pluralidade de sítios.

3. Método de acordo com a reivindicação 1 ou 2, caracterizado pelo fato de que a correção pré-faseamento compreende um peso e em que a determinação dos valores de cor corrigidos compreende multiplicar o peso pelos valores de cor da pluralidade de sítios medidos a partir da imagem do substrato.

4. Método de acordo com qualquer uma das reivindicações anteriores, caracterizado pelo fato de que compreende adicionalmente determinar uma correção de faseamento para o ciclo de chamada de bases imediatamente subsequente.

5. Método de acordo com a reivindicação 4, caracterizado pelo fato de que a determinação da correção de faseamento para o ciclo de chamada de bases imediatamente subsequente compreende analisar

os valores de cor parcialmente corrigidos na fase armazenados na memória do sequenciador, e

os valores de cor armazenados no buffer do processador.

6. Método de acordo com a reivindicação 4, caracterizado pelo fato de que compreende adicionalmente:

produzir valores de cor parcialmente corrigidos na fase para o ciclo de chamada de bases imediatamente subsequente aplicando a correção de faseamento aos valores de cor da pluralidade de sítios armazenados na memória do sequenciador; e

armazenar os valores de cor parcialmente corrigidos na fase para o ciclo de chamada de bases imediatamente subsequente na memória do sequenciador.

7. Método de acordo com a reivindicação 6, caracterizado pelo fato de que os valores de cor parcialmente corrigidos na fase para o ciclo de chamada de bases imediatamente subsequente compreende adicionalmente somar

os valores de cor faseados corrigidos da pluralidade de sítios, e

os valores de cor da pluralidade de sítios a partir da imagem do substrato medido em (b).

8. Método de acordo com a reivindicação 6, caracterizado pelo fato de que o armazenamento dos valores de cor parcialmente corrigidos na fase para os ciclo de chamada de bases armazena os valores de cor parcialmente corrigidos nos buffers do bloco da memória do sequenciador.

9. Método de acordo com qualquer uma das reivindicações anteriores, caracterizado pelo fato de que o método é realizado em tempo real durante a aquisição das leituras de sequência pelo sequenciador de ácidos nucleicos.

10. Método de acordo com qualquer uma das reivindicações anteriores, caracterizado pelo fato de que o sequenciador de ácidos nucleicos sintetiza ácidos nucleicos na pluralidade de sítios.

11. Método de acordo com qualquer uma das reivindicações anteriores, caracterizado pelo fato de que os valores de cor são determinados a partir de apenas dois canais do sequenciador.

12. Método de acordo com qualquer uma das reivindicações 1 a 10, caracterizado pelo fato de que os valores de cor são obtidos a partir de quatro canais do sequenciador.

13. Método de acordo com qualquer uma das reivindicações anteriores, caracterizado pelo fato de que o substrato compreende uma célula de fluxo, sendo que a célula de fluxo está logicamente dividida em blocos, e em que cada bloco representa uma região da célula de fluxo que compreende um subconjunto de sítios, cujo subconjunto é capturado em uma única imagem do sistema de aquisição de imagem.

14. Método de acordo com a reivindicação 13, caracterizado pelo fato de que, na operação (d), os valores de cor parcialmente corrigidos na fase foram armazenados nos buffers do bloco da memória do sequenciador, e em que os buffers do bloco são designados para armazenar dados

representando imagens de blocos individuais no substrato.

15. Método de acordo com a reivindicação 14, caracterizado pelo fato de que a memória possui uma capacidade de memória de cerca de 512 gigabytes ou menos.

16. Método de acordo com a reivindicação 13, caracterizado pelo fato de que compreende, antes da operação (a), prover reagentes para a célula de fluxo e permitir que os reagentes interajam com os sítios para exibir as cores que representam os tipos de base de ácido nucleico durante o ciclo de chamada de bases.

17. Método de acordo com a reivindicação 16, caracterizado pelo fato de que compreende adicionalmente, após a operação (f):

prover reagentes frescos para a célula de fluxo e permitir que os reagentes frescos interajam com os sítios para exibir as cores que representam os tipos de base de ácido nucleico para um ciclo de chamada de bases seguinte; e

repetir as operações (a) a (e) para o próximo ciclo de chamada de bases.

18. Método de acordo com a reivindicação 17, caracterizado pelo fato de que compreende adicionalmente criar um primeiro encadeamento de processador para executar as operações (a) a (f) para o ciclo de chamada de bases, e criar um segundo encadeamento de processador para executar as operações (a) a (f) para o ciclo de chamada de bases seguinte.

19. Método de acordo com qualquer uma das reivindicações anteriores, caracterizado pelo fato de que compreende adicionalmente alocar o buffer do processador e um segundo buffer do processador, sendo que o segundo buffer do processador é usado para determinar os valores de cor corrigidos em (f).

20. Sequenciador de ácidos nucleicos, caracterizado pelo fato de que compreende:

um sistema de aquisição de imagem;

memória; e

um ou mais processadores projetados ou configurados para:

(a) obter dados que representam uma imagem de um substrato compreendendo uma pluralidade de sítios onde as bases de ácido nucleico são lidas, em que os sítios exibem cores representando os tipos de base de ácido nucleico;

(b) obter os valores de cor da pluralidade de sítios a partir da imagem do substrato;

(c) armazenar os valores das cores em um buffer do processador;

(d) recuperar os valores de cor parcialmente corrigidos na fase da pluralidade de sítios para um ciclo de chamada de bases, em que os valores de cor parcialmente corrigidos na fase foram armazenados na memória durante um ciclo de chamada bases imediatamente anterior;

(e) determinar uma correção pré-faseamento a partir dos

valores de cor parcialmente corrigidos na fase armazenados durante o ciclo de chamada de bases imediatamente anterior e

valores de cor armazenados no buffer do processador; e

(f) determinar os valores de cor corrigidos dos

valores de cor no buffer do processador,

valores cor parcialmente corrigidos na fase armazenados durante o ciclo imediatamente anterior, e

a correção pré-faseamento.

21. Sequenciador de ácidos nucleicos de acordo com a reivindicação 20, caracterizado pelo fato de que a memória é dividida em uma pluralidade de buffers do bloco, cada um designado para armazenar dados representando uma única imagem de um bloco no substrato.

22. Sequenciador de ácidos nucleicos de acordo com a

reivindicação 20 ou 21, caracterizado pelo fato de que a memória tem uma capacidade de armazenamento de cerca de 512 gigabytes ou menos.

23. Sequenciador de ácidos nucleicos de acordo com qualquer uma das reivindicações 20 a 22, caracterizado pelo fato de que o um ou mais processadores são adicionalmente concebidos ou configurados para utilizar os valores de cor corrigidos para fazer chamadas de base para a pluralidade de sítios.

24. Sequenciador de ácidos nucleicos de acordo com qualquer uma das reivindicações 20 a 23, caracterizado pelo fato de que a correção pré-faseamento compreende um peso e em que o um ou mais processadores são concebidos ou configurados para determinar os valores de cor corrigidos multiplicando o peso pelos valores de cor da pluralidade de sítios medidos a partir da imagem do substrato.

25. Sequenciador de ácidos nucleicos de acordo com qualquer uma das reivindicações 20 a 24, caracterizado pelo fato de que o um ou mais processadores são adicionalmente concebidos ou configurados para determinar uma correção de faseamento para um ciclo de chamada de bases imediatamente subsequente.

26. Sequenciador de ácidos nucleicos de acordo com a reivindicação 25, caracterizado pelo fato de que o um ou mais processadores são concebidos ou configurados para determinar a correção de faseamento para o ciclo de chamada de bases imediatamente subsequente pela análise

dos valores de cor parcialmente corrigidos na fase armazenados na memória, e

dos valores de cor armazenados no buffer do processador.

27. Sequenciador de ácidos nucleicos de acordo com a reivindicação 25, caracterizado pelo fato de que o um ou mais processadores são adicionalmente concebidos ou configurados para:

produzir valores de cor parcialmente corrigidos na fase para o

ciclo de chamada de bases imediatamente subsequente aplicando a correção de faseamento aos valores de cor da pluralidade de sítios armazenados na memória; e

armazenar os valores de cor parcialmente corrigidos na fase para o ciclo de chamada de bases imediatamente subsequente na memória.

28. Sequenciador de ácidos nucleicos de acordo com a reivindicação 27, caracterizado pelo fato de que o um ou mais processadores são concebidos ou configurados para produzir os valores de cor parcialmente corrigidos na fase para o ciclo de chamada de bases imediatamente subsequente somando

os valores de cor faseados corrigidos da pluralidade de sítios, e

os valores de cor da pluralidade de sítios a partir da imagem do substrato medido em (b).

29. Sequenciador de ácidos nucleicos de acordo com a reivindicação 27, caracterizado pelo fato de que o um ou mais processadores são concebidos ou configurados para armazenar os valores de cor parcialmente corrigidos na fase para o ciclo de chamada de bases imediatamente subsequente pelo armazenamento dos valores de cor parcialmente corrigidos nos buffers do bloco da memória.

30. Sequenciador de ácidos nucleicos de acordo com qualquer uma das reivindicações 20 a 29, caracterizado pelo fato de que o um ou mais processadores são concebidos ou configurados para executar (a) a (f) em tempo real durante a chamada de base.

31. Sequenciador de ácidos nucleicos de acordo com qualquer uma das reivindicações 20 a 30, caracterizado pelo fato de que compreende adicionalmente um sistema para sintetizar ácidos nucleicos na pluralidade de sítios.

32. Sequenciador de ácidos nucleicos de acordo com qualquer uma das reivindicações 20 a 31, caracterizado pelo fato de que o um ou mais

processadores são concebidos ou configurados para obter os valores de cor a partir de apenas dois canais.

33. Sequenciador de ácidos nucleicos de acordo com qualquer uma das reivindicações 20 a 31, caracterizado pelo fato de que o um ou mais processadores são concebidos ou configurados para obter os valores de cor a partir de quatro canais.

34. Sequenciador de ácidos nucleicos de acordo com qualquer uma das reivindicações 20 a 33, caracterizado pelo fato de que o substrato compreende uma célula de fluxo, sendo que a célula de fluxo está logicamente dividida em blocos, e em que cada bloco representa uma região da célula de fluxo que compreende um subconjunto de sítios, cujo subconjunto é capturado em uma única imagem do sistema de aquisição de imagem.

35. Sequenciador de ácidos nucleicos de acordo com a reivindicação 34, caracterizado pelo fato de que, na operação (d), os valores de cor parcialmente corrigidos na fase foram armazenados nos buffers do bloco da memória do sequenciador, e em que os buffers do bloco são designados para armazenar dados representando imagens de blocos individuais no substrato.

36. Sequenciador de ácidos nucleicos de acordo com a reivindicação 34, caracterizado pelo fato de que o um ou mais processadores são adicionalmente concebidos ou configurados para, antes da operação (a), prover reagentes à célula de fluxo e permitir que os reagentes interajam com os sítios para exibir as cores que representam os tipos de base de ácido nucleico durante o ciclo de chamada de bases.

37. Sequenciador de ácidos nucleicos de acordo com a reivindicação 36, caracterizado pelo fato de que o um ou mais processadores são adicionalmente concebidos ou configurados para, após a operação (f):

prover reagentes frescos para a célula de fluxo e permitir que

os reagentes frescos interajam com os sítios para exibir as cores que representam os tipos de base de ácido nucleico para um ciclo de chamada de bases seguinte; e

repetir as operações (a) a (e) para o próximo ciclo de chamada de bases.

38. Sequenciador de ácidos nucleicos de acordo com a reivindicação 37, caracterizado pelo fato de que o um ou mais processadores são adicionalmente concebidos ou configurados para criar um primeiro encadeamento de processador para executar as operações (a) a (f) para o ciclo de chamada de bases e criar um segundo encadeamento de processador para executar as operações (a) a (f) para o próximo ciclo de chamada de bases.

39. Sequenciador de ácidos nucleicos de acordo com qualquer uma das reivindicações 20 a 38, caracterizado pelo fato de que o um ou mais processadores são adicionalmente concebidos ou configurados para atribuir o buffer do processador e um segundo buffer do processador para determinar os valores de cor corrigidos em (f).

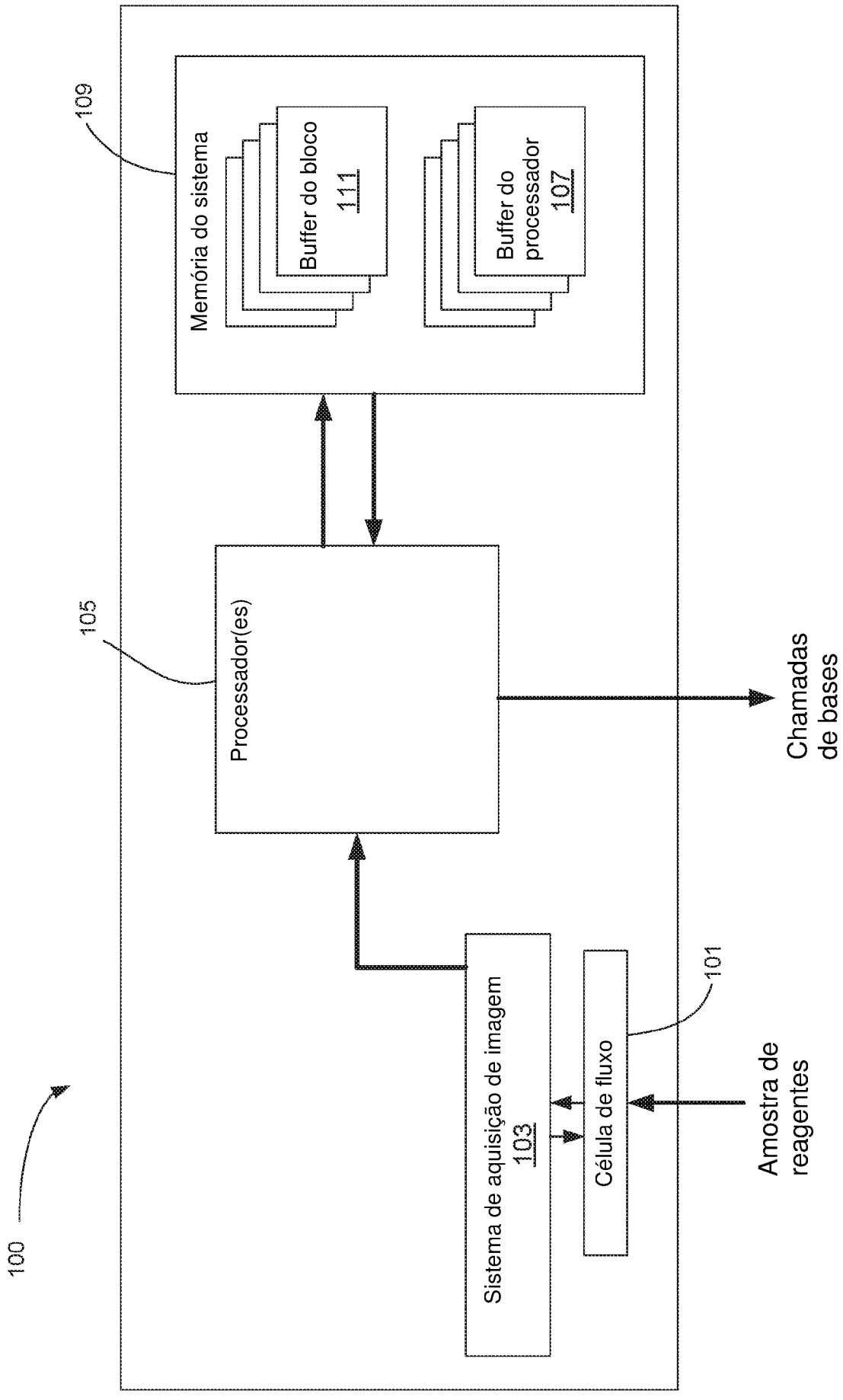

100
109
Memória do sistema
Buffer do bloco
111
Buffer do
processador
107
105
Processador(es)
Chamadas
de bases
Sistema de aquisição de imagem
103
Célula de fluxo
101
Amostra de
reagentes

Figura 1

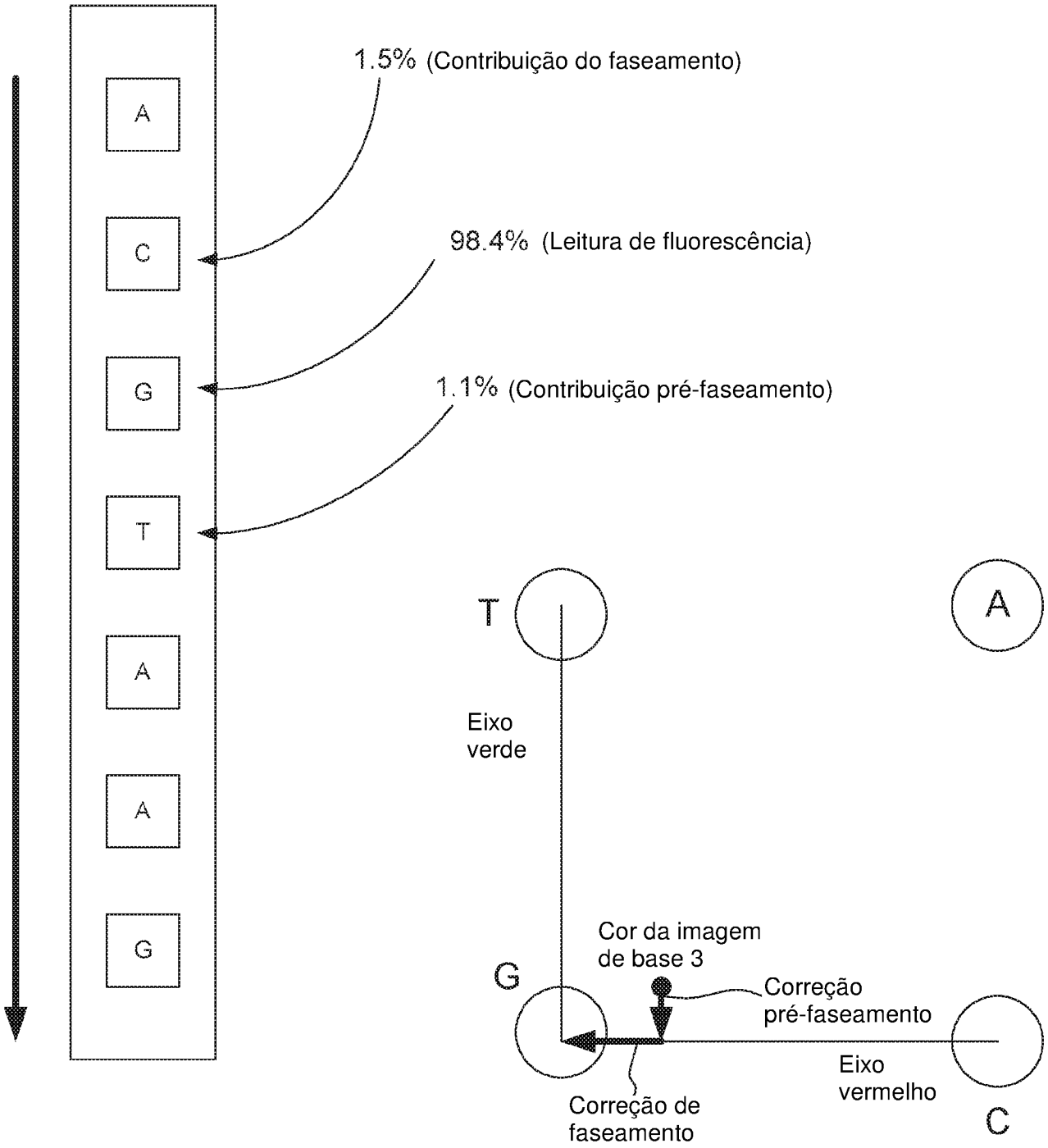

1.5% (Contribuição do faseamento)
98.4% (Leitura de fluorescência)
1.1% (Contribuição pré-faseamento)
A
C
G
T
A
A
G
T
A
Eixo
verde
Cor da imagem
de base 3
G
Correção
pré-faseamento
Eixo
vermelho
Correção de
faseamento
C

Figura 2

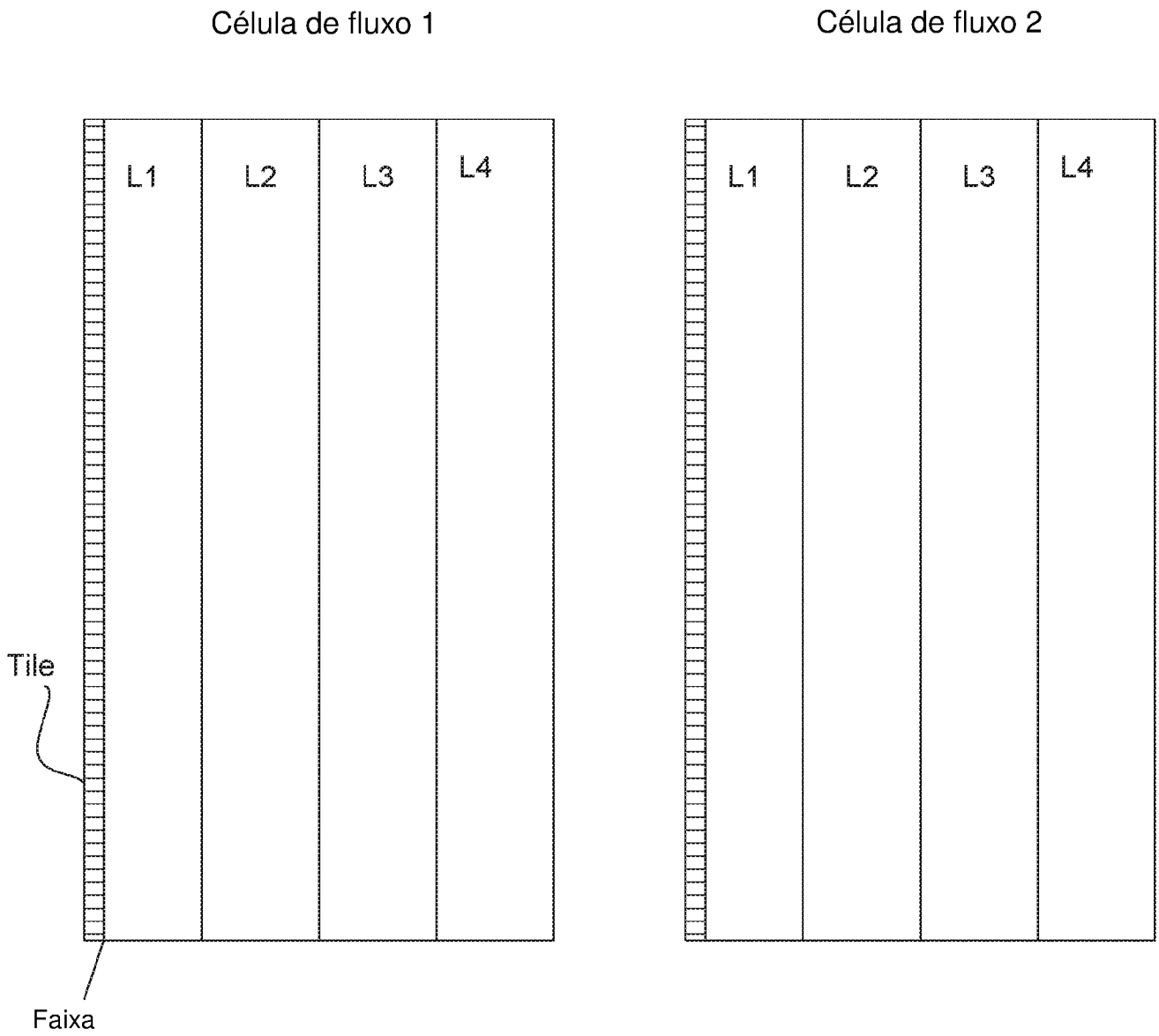

Célula de fluxo 1
Célula de fluxo 2
L1
L2
L3
L4
L1
L2
L3
L4
Tile
Faixa
Dados de imagem por ciclo
Processado em tempo real
2 Células de fluxo
2 Superfícies/célula de fluxo
4 Raias/superfície
6 Faixas/raia
120 blocos/faixa
Mais de 8000 blocos de dados por ciclo

Figura 3

Número do agrupamento	1	2	3	4	5	6	7	8	9	10	11	12
Valor da cor (canal 1)	468	695	1255	521	753	1452	1355	1255	598	720	1193	422
Valor da cor (canal 2)	1321	1466	1326	498	463	1384	632	558	452	1467	1369	1278

Figura 4

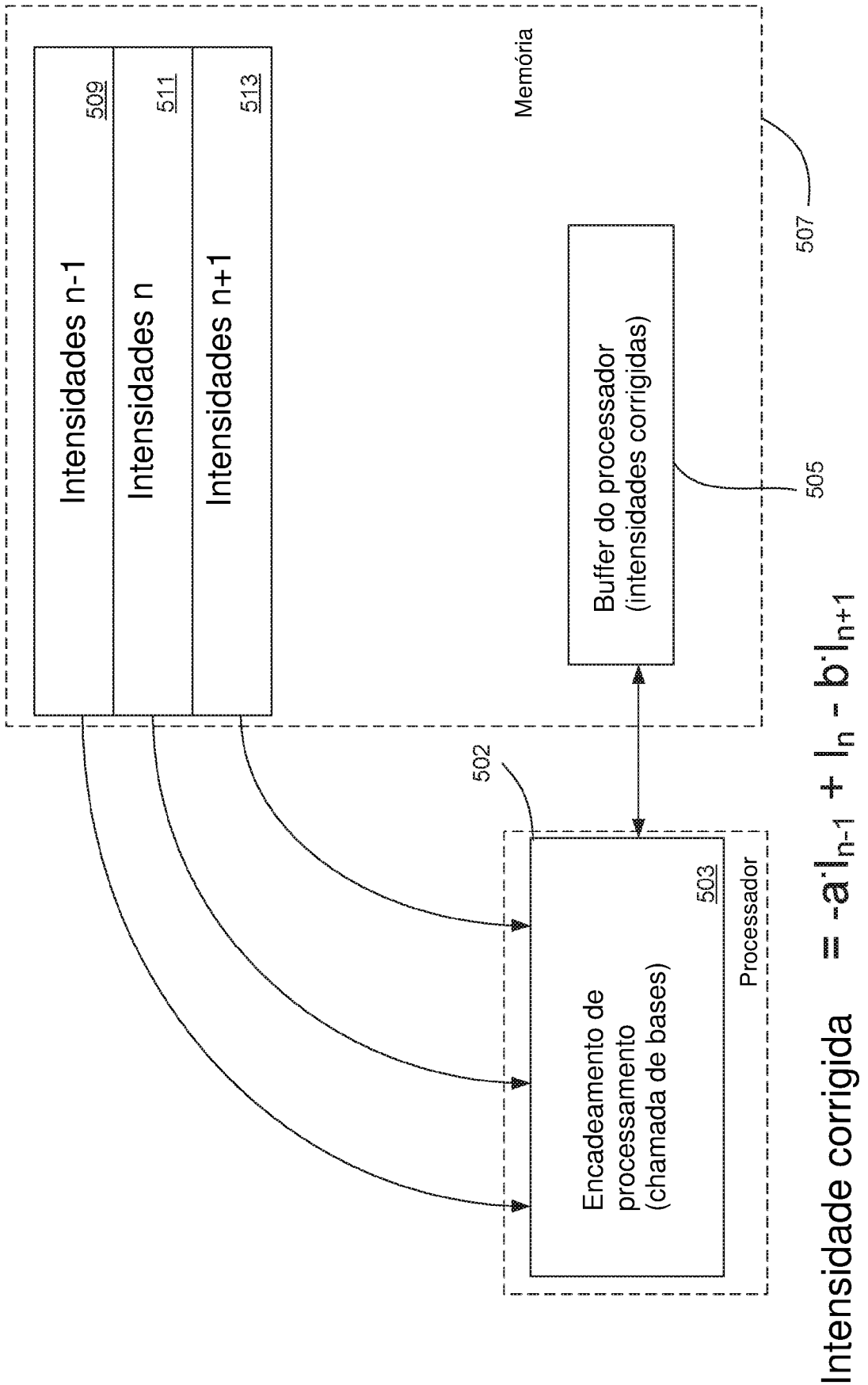

Memória
Intensidades n-1
509
Intensidades n
511
Intensidades n+1
513
507
Buffer do processador
(intensidades corrigidas)
505
502
Encadeamento de
processamento
(chamada de bases)
503
Processador
Intensidade corrigida = -a'I n-1 + I n - b'I n+1

Figura 5

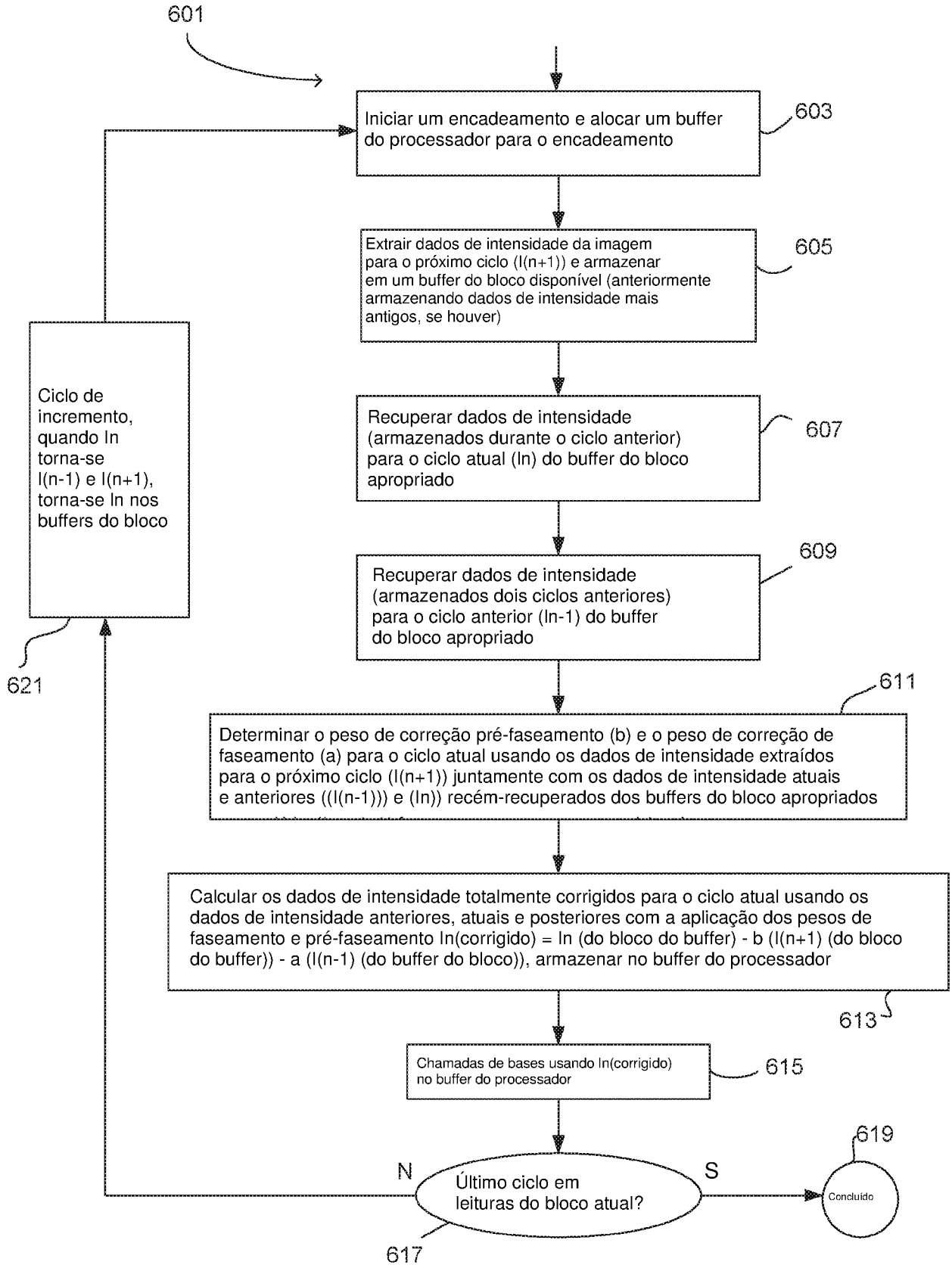

601
Iniciar um encadeamento e alocar um buffer do processador para o encadeamento
603
Extrair dados de intensidade da imagem para o próximo ciclo (I(n+1)) e armazenar em um buffer do bloco disponível (anteriormente armazenando dados de intensidade mais antigos, se houver)
605
Recuperar dados de intensidade (armazenados durante o ciclo anterior) para o ciclo atual (In) do buffer do bloco apropriado
607
Recuperar dados de intensidade (armazenados dois ciclos anteriores) para o ciclo anterior (In-1) do buffer do bloco apropriado
609
Determinar o peso de correção pré-faseamento (b) e o peso de correção de faseamento (a) para o ciclo atual usando os dados de intensidade extraídos para o próximo ciclo (I(n+1)) juntamente com os dados de intensidade atuais e anteriores ((I(n-1))) e (In)) recém-recuperados dos buffers do bloco apropriados
611
Calcular os dados de intensidade totalmente corrigidos para o ciclo atual usando os dados de intensidade anteriores, atuais e posteriores com a aplicação dos pesos de faseamento e pré-faseamento In(corrigido) = In (do bloco do buffer) - b (I(n+1) (do bloco do buffer)) - a (I(n-1) (do buffer do bloco)), armazenar no buffer do processador
613
Chamadas de bases usando In(corrigido) no buffer do processador
615
Último ciclo em leituras do bloco atual?
N
S
617
Concluído
619
Ciclo de incremento, quando In torna-se I(n-1) e I(n+1), torna-se In nos buffers do bloco
621

Figura 6

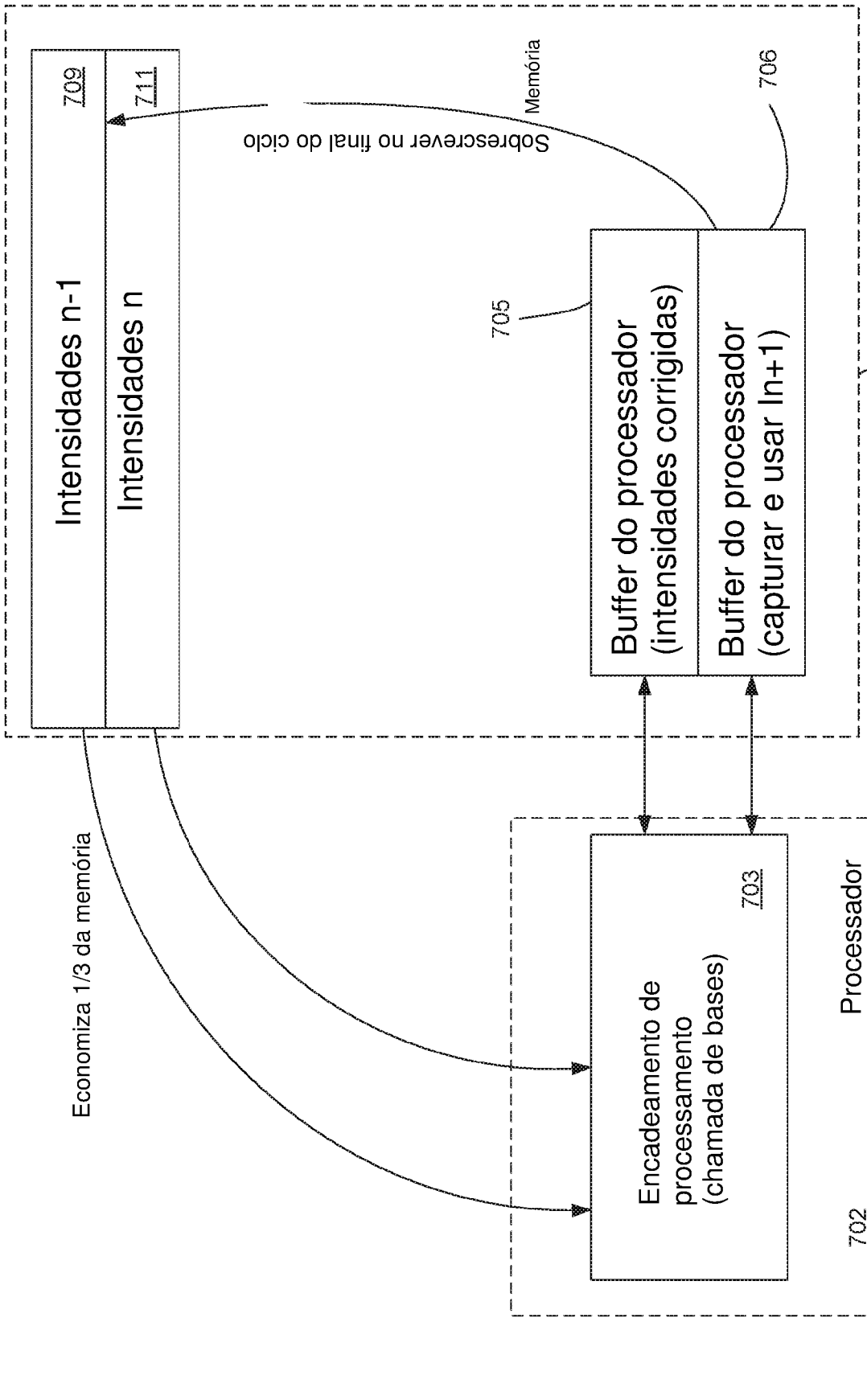
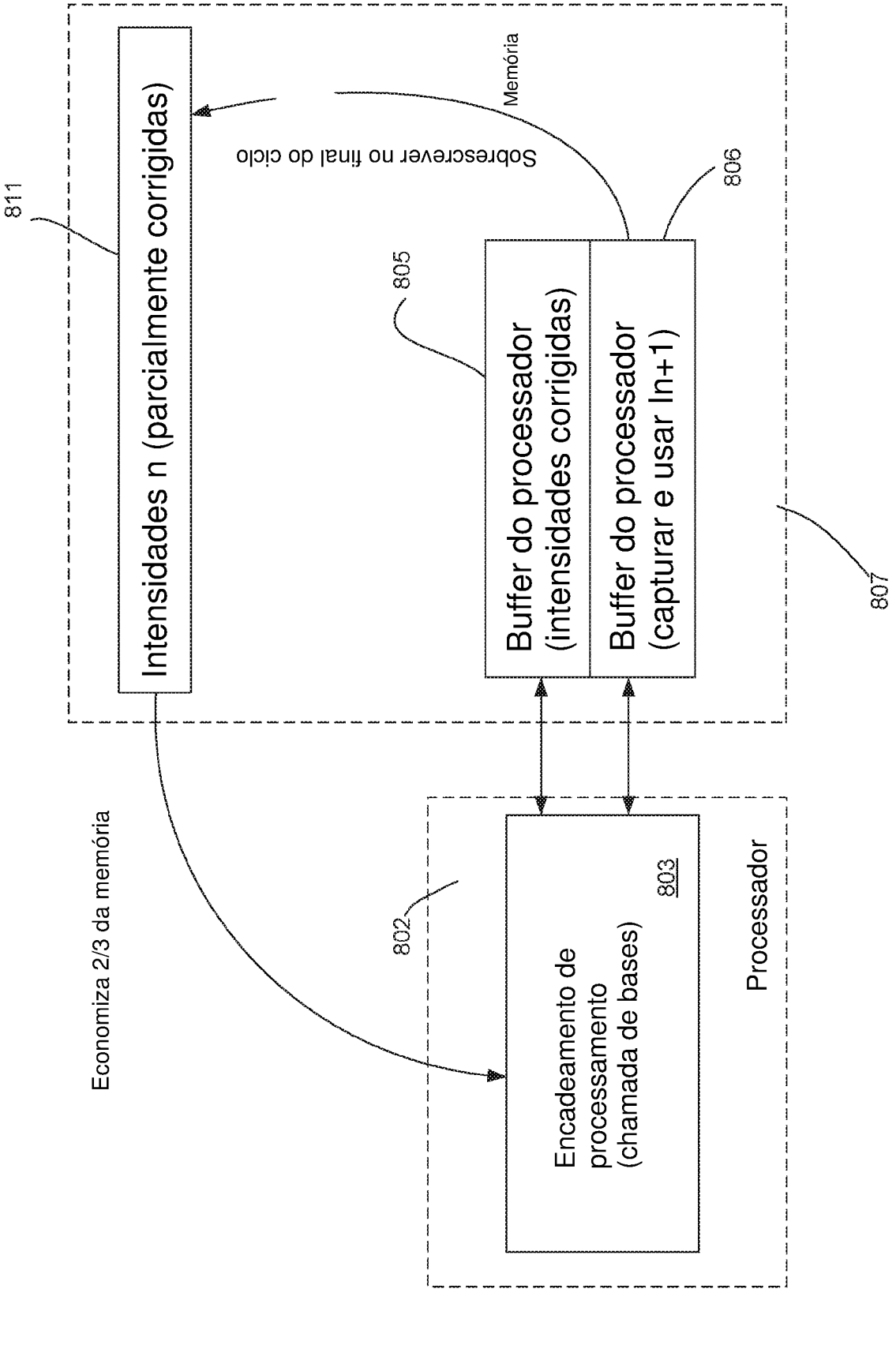

Intensidades n-1
709
Intensidades n
711
Sobrescrever no final do ciclo
Memória
706
705
Buffer do processador (intensidades corrigidas)
Buffer do processador (capturar e usar In+1)
707
Economiza 1/3 da memória
Encadeamento de processamento (chamada de bases)
703
Processador
702
Intensidade corrigida = -a'In-1 + In - b'In+1

Figura 7



Intensidade corrigida $= -a' I_{n-1} + I_n - b' I_{n+1}$

Figura 8

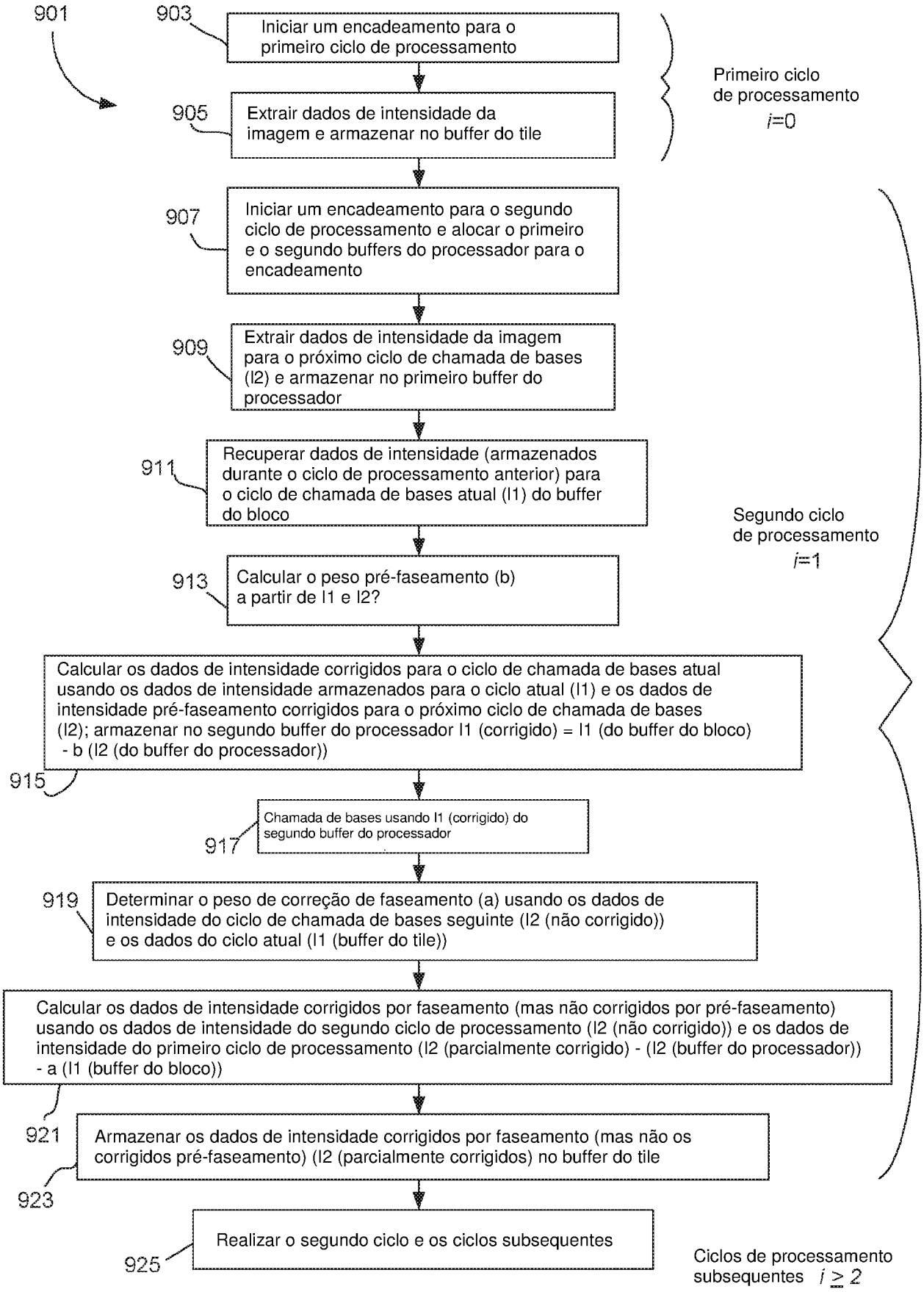

901
903
Iniciar um encadeamento para o primeiro ciclo de processamento
905
Extrair dados de intensidade da imagem e armazenar no buffer do tile
Primeiro ciclo de processamento i=0
907
Iniciar um encadeamento para o segundo ciclo de processamento e alocar o primeiro e o segundo buffers do processador para o encadeamento
909
Extrair dados de intensidade da imagem para o próximo ciclo de chamada de bases (I2) e armazenar no primeiro buffer do processador
911
Recuperar dados de intensidade (armazenados durante o ciclo de processamento anterior) para o ciclo de chamada de bases atual (I1) do buffer do bloco
913
Calcular o peso pré-faseamento (b) a partir de I1 e I2?
Segundo ciclo de processamento i=1
Calcular os dados de intensidade corrigidos para o ciclo de chamada de bases atual usando os dados de intensidade armazenados para o ciclo atual (I1) e os dados de intensidade pré-faseamento corrigidos para o próximo ciclo de chamada de bases (I2); armazenar no segundo buffer do processador I1 (corrigido) = I1 (do buffer do bloco) - b (I2 (do buffer do processador))
915
917
Chamada de bases usando I1 (corrigido) do segundo buffer do processador
919
Determinar o peso de correção de faseamento (a) usando os dados de intensidade do ciclo de chamada de bases seguinte (I2 (não corrigido)) e os dados do ciclo atual (I1 (buffer do tile))
Calcular os dados de intensidade corrigidos por faseamento (mas não corrigidos por pré-faseamento) usando os dados de intensidade do segundo ciclo de processamento (I2 (não corrigido)) e os dados de intensidade do primeiro ciclo de processamento (I2 (parcialmente corrigido) - (I2 (buffer do processador)) - a (I1 (buffer do bloco))
921
Armazenar os dados de intensidade corrigidos por faseamento (mas não os corrigidos pré-faseamento) (I2 (parcialmente corrigidos) no buffer do tile
923
Realizar o segundo ciclo e os ciclos subsequentes
925
Ciclos de processamento subsequentes i ≥ 2

Figura 9

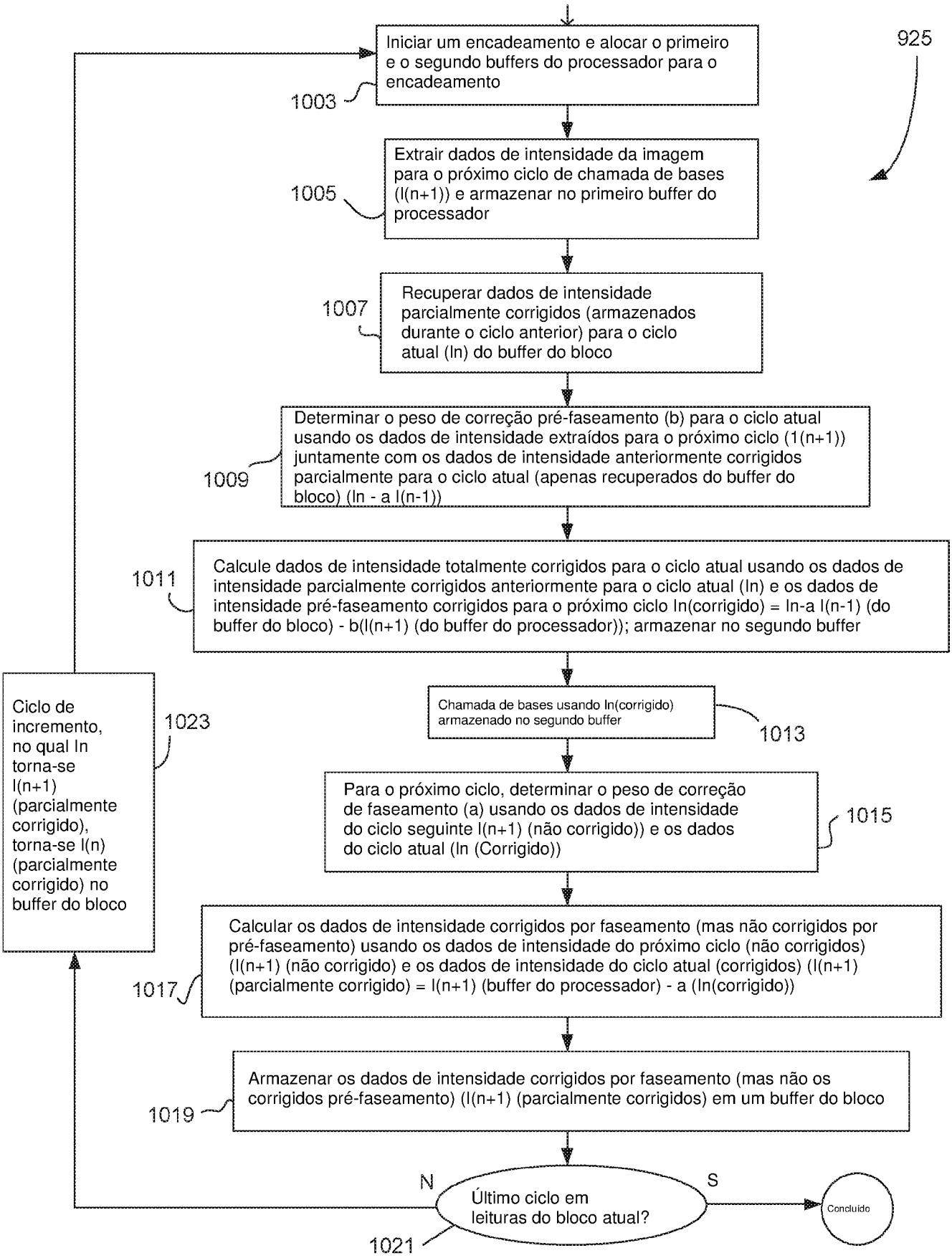

925
Iniciar um encadeamento e alocar o primeiro e o segundo buffers do processador para o encadeamento
1003
Extrair dados de intensidade da imagem para o próximo ciclo de chamada de bases (I(n+1)) e armazenar no primeiro buffer do processador
1005
Recuperar dados de intensidade parcialmente corrigidos (armazenados durante o ciclo anterior) para o ciclo atual (In) do buffer do bloco
1007
Determinar o peso de correção pré-faseamento (b) para o ciclo atual usando os dados de intensidade extraídos para o próximo ciclo (1(n+1)) juntamente com os dados de intensidade anteriormente corrigidos parcialmente para o ciclo atual (apenas recuperados do buffer do bloco) (In - a I(n-1))
1009
Calcule dados de intensidade totalmente corrigidos para o ciclo atual usando os dados de intensidade parcialmente corrigidos anteriormente para o ciclo atual (In) e os dados de intensidade pré-faseamento corrigidos para o próximo ciclo In(corrigido) = In-a I(n-1) (do buffer do bloco) - b(I(n+1) (do buffer do processador)); armazenar no segundo buffer
1011
Chamada de bases usando In(corrigido) armazenado no segundo buffer
1013
Para o próximo ciclo, determinar o peso de correção de faseamento (a) usando os dados de intensidade do ciclo seguinte I(n+1) (não corrigido)) e os dados do ciclo atual (In (Corrigido))
1015
Calcular os dados de intensidade corrigidos por faseamento (mas não corrigidos por pré-faseamento) usando os dados de intensidade do próximo ciclo (não corrigidos) (I(n+1) (não corrigido) e os dados de intensidade do ciclo atual (corrigidos) (I(n+1) (parcialmente corrigido) = I(n+1) (buffer do processador) - a (In(corrigido))
1017
Armazenar os dados de intensidade corrigidos por faseamento (mas não os corrigidos pré-faseamento) (I(n+1) (parcialmente corrigidos) em um buffer do bloco
1019
Último ciclo em leituras do bloco atual?
N
S
1021
Concluído
Ciclo de incremento, no qual In torna-se I(n+1) (parcialmente corrigido), torna-se I(n) (parcialmente corrigido) no buffer do bloco
1023

Figura 10

	Densidade (K/mm2)	Agrupamentos PF (%)	Erro de ciclo avaliado	Alinhado (%)	Taxa de erro (%)
Linha de base	206 +/- 0	81.32 +/- 0.69	150	1.43 +/- 0.01	1.38 +/- 0.00
Novo	206 +/- 0	81.80 +/- 0.00	150	1.42 +/- 0.00	1.41 +/- 0.00

	Blocos	Densidade (K/mm2)	Agrupamentos PF (%)	Alinhado (%)	Taxa de erro (%)
Linha de base	1	3815 +/- 0	65.39 +/- 0.00	1.43 +/- 0.00	1.89 +/- 0.00
Novo	1	3815 +/- 0	65.39 +/- 0.00	1.43 +/- 0.00	1.95 +/- 0.00

Figura 11

RESUMO

MÉTODO PARA DETERMINAR OS VALORES DE COR CORRIGIDOS, E, SEQUENCIADOR DE ÁCIDOS NUCLEICOS

Métodos eficientes de memória determinam valores de cor corrigidos de dados de imagem adquiridos por um sequenciador de ácidos nucleicos durante um ciclo de chamada de bases. Tais métodos podem: (a) obter uma imagem de um substrato (por exemplo, de uma porção de uma célula de fluxo) incluindo uma pluralidade de sítios onde as bases de ácido nucleico são lidas; (b) medir valores de cor da pluralidade de sítios a partir da imagem do substrato; (c) armazenar os valores de cor em um buffer do processador do um ou mais processadores do sequenciador; (d) recuperar valores de cor parcialmente corrigidos na fase da pluralidade de sítios, onde os valores de cor parcialmente corrigidos na fase foram armazenados na memória do sequenciador durante um ciclo de chamada de bases imediatamente anterior; (e) determinar uma correção pré-faseamento; e (f) determinar os valores de cor corrigidos. Em várias implementações, essas operações são todas realizadas durante um único ciclo de chamada de bases. Em certas modalidades, os métodos adicionalmente incluem o uso dos valores de cor corrigidos para fazer chamadas de base para a pluralidade de sítios. Sequenciadores podem ser concebidos ou configurados para implementar tais métodos.