



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

(21) BR 112020016516-6 A2



(22) Data do Depósito: 13/02/2019

(43) Data da Publicação Nacional: 15/12/2020

(54) Título: MÉTODOS E REAGENTES PARA DETECTAR E AVALIAR A GENOTOXICIDADE

(51) Int. Cl.: C12Q 1/6869; C12Q 1/6855; C12Q 1/6886.

(30) Prioridade Unionista: 13/02/2018 US 62/630,228; 26/09/2018 US 62/737,097.

(71) Depositante(es): TWINSTRAND BIOSCIENCES, INC..

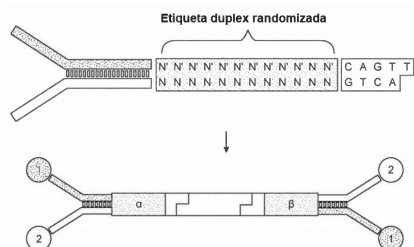
(72) Inventor(es): JESSE J. SALK; CHARLES CLINTON VALENTINE III.

(86) Pedido PCT: PCT US2019017908 de 13/02/2019

(87) Publicação PCT: WO 2019/160998 de 22/08/2019

(85) Data da Fase Nacional: 13/08/2020

(57) Resumo: Métodos, sistemas e kits com reagentes para avaliar a genotoxicidade, são divulgados aqui. A genotoxicidade e seus mecanismos de ação podem ser determinados alguns dias após a exposição do sujeito. Algumas modalidades da tecnologia são direcionadas à utilização do Sequenciamento Duplex para avaliar um potencial genotóxico de um composto (por exemplo, um composto químico) em um sujeito exposto. Outras modalidades da tecnologia são direcionadas à utilização do Sequenciamento Duplex para determinar uma assinatura de mutação associada a um agente genotóxico; e/ou um nível limiar seguro de exposição à genotoxina. Modalidades adicionais da tecnologia são direcionadas para identificar um ou mais agentes genotóxicos aos quais um sujeito pode ter sido exposto comparando o espectro de mutação do DNA do sujeito com os espectros de mutação de compostos mutagênicos conhecidos. Depois que uma exposição à genotoxina em um sujeito é identificada ou confirmada, é fornecido um curso terapêutico profilático e/ou inibitório.



MÉTODOS E REAGENTES PARA DETECTAR E AVALIAR A GENOTOXICIDADE REFERÊNCIA CRUZADA A PEDIDOS RELACIONADOS

[0001] Este pedido reivindica prioridade e o benefício do Pedido de Patente Provisório U.S. 62/630.228, depositado em 13 de fevereiro de 2018, e do Pedido de Patente Provisório U.S. 62/737.097, depositado em 26 de setembro de 2018, cujas divulgações são aqui incorporadas por referência na sua totalidade.

FUNDAMENTOS

[0002] Genotoxicidade se refere à propriedade destrutiva de agentes ou processos (isto é, genotoxinas) que causam danos ao material genético (por exemplo, DNA, RNA). Nas linhagens de células germinativas, o dano ao material de ácido nucleico tem o potencial de resultar em uma mutação hereditária da linhagem germinativa, enquanto o dano ao material de ácido nucleico em células somáticas pode resultar em uma mutação somática. Em alguns casos, essas mutações somáticas podem levar à malignidade ou outras doenças. Foi estabelecido que a exposição à genotoxina pode causar direta ou indiretamente esse dano ao ácido nucleico ou, em alguns casos, pode ser responsável por causar direta e indiretamente o dano ao ácido nucleico. Por exemplo, uma substância genotóxica pode interagir diretamente com o material genético para causar alterações na própria sequência nucleotídica ou na sua estrutura ou criar modificações químicas (por exemplo, adutos ou quebras) que quando sofrem tentativa de serem copiadas, reparadas ou processadas por máquinas celulares, induzem (ou aumentam a probabilidade de induzir) alterações na sequência nucleotídica. A genotoxina pode ser um produto químico ou processo de ocorrência natural (por exemplo, carvão, rádio ou luz UV) ou um produto químico ou processo ou terapia criado artificialmente (por exemplo, uretano industrial, aparelhos de raios-X, muitas drogas quimioterápicas e algumas formas de terapia genética).

[0003] Outras genotoxinas podem indiretamente desencadear danos aos ácidos nucleicos, ativando vias celulares que reduzem a fidelidade da replicação do DNA. Por exemplo, isso pode ser a ativação direta ou indireta de máquinas de ciclo celular que ignora os pontos de verificação normais ou reduz o reparo normal de ácidos nucleicos (como desregulação direta ou indireta de qualquer uma das muitas vias de reparo de ácido nucleico, incluindo reparo de incompatibilidade (MMR), reparo por excisão de nucleotídeo (NER), reparo por excisão de base (BER), reparo por quebra de fita dupla (DSBR), reparo acoplado por transcrição (TCR), junção

final não homóloga (NHEJ), entre outros). Outras genotoxinas podem atuar indiretamente promovendo o ambiente celular que é, ele próprio, genotóxico. Um exemplo desse ambiente é o "estresse oxidativo", que pode ser criado aumentando a produção de espécies reativas de oxigênio em um organismo (por exemplo, através da estimulação da inflamação mediada pelo sistema imunológico) ou célula que pode causar danos ao material genético, modificando a própria composição química da sequência ou alterando estruturalmente as fitas de ácido nucleico. Ainda outra forma indireta de genotoxinas são agentes ou processos que suprimem certos aspectos do sistema imunológico de um organismo. Tais reduções na vigilância imune podem levar à genotoxicidade em um organismo, permitindo a proliferação de micro-organismos que podem ser genotóxicos através de qualquer um de vários mecanismos (por exemplo, causando inflamação ou promovendo a progressão do ciclo celular em certos tecidos). Além disso, esses agentes ou processos podem contribuir para a carga genotóxica de um organismo através da redução da capacidade normal de purgar células portadoras de anormalidades genéticas que, de outra forma, seriam eliminadas e seriam cancerígenas por esse mecanismo. Os mecanismos de muitas genotoxinas ainda precisam ser descobertos.

[0004] As genotoxinas podem se originar de uma variedade de fontes externas e internas. Por exemplo, fontes externas (exógenas) podem incluir produtos químicos ou uma mistura de produtos químicos (por exemplo, produtos farmacêuticos, subprodutos industriais/de fabricação, resíduos químicos, cosméticos, produtos de limpeza, plastificantes, fumaça de tabaco, solventes etc.); metais pesados, partículas transportadas pelo ar, contaminantes, produtos alimentícios, radiação (por exemplo, fótons, como radiação gama, radiação X, radiação de partículas ou uma mistura dos mesmos), forças físicas (por exemplo, um campo magnético, campo gravitacional, forças de aceleração, etc.) do ambiente natural ou de um dispositivo; outro organismo (por exemplo, vírus, parasitas, bactérias, protozoários, fungos) ou produzido por outro organismo de ocorrência natural (por exemplo, fungos, plantas, animais, bactérias, bactérias, protozoários, etc.). Certas culturas em si (por exemplo, tabaco) contêm genotoxinas conhecidas em sua forma natural. As culturas alimentares básicas podem ficar contaminadas com genotoxinas durante o crescimento (por exemplo, contaminação da água de irrigação com resíduos industriais), colheita (por exemplo, co-colheita inadvertida de culturas com aristocholia, que produzem o ácido aristolóquico mutagênico), armazenamento (por exemplo, leguminosa úmida e silos de grãos que

levam ao crescimento de espécies de *aspergillus* que produzem a aflatoxina mutagênica) ou durante a preparação (por exemplo, fumo e alguns outros métodos de preservação de carnes, que criam muitas formas de genotoxinas ou cozimento de amidos a alta temperatura que podem produzir a acrilamida mutagênica). Alguns exemplos de fontes internas (isto é, endógenas) podem incluir processos bioquímicos ou os resultados de processos bioquímicos. Por exemplo, um agente químico pode ser determinado como sendo uma genotoxina se o agente for um precursor de um mutagênico que resulta da ativação metabólica. Outros exemplos podem incluir estimuladores de vias inflamatórias (por exemplo, estresse, doenças autoimunes) ou inibidores de apoptose ou vigilância imunológica. Independentemente da fonte, vários fatores desempenham um papel na determinação de se um agente ou processo é potencialmente genotóxico, mutagênico ou carcinogênico (isto é, causador de câncer).

[0005] Em certas aplicações, a capacidade de detectar e quantificar processos mutagênicos é importante para avaliar o risco de câncer e prever o impacto da exposição cancerígena em humanos. Da mesma forma, avaliar o potencial de compostos químicos ou outros agentes causarem mutações de ácidos nucleicos é um elemento essencial dos testes de segurança do produto antes da comercialização (por exemplo, produtos farmacêuticos, cosméticos, produtos alimentícios, subprodutos de fabricação e semelhantes). Os métodos atuais de identificação de genotoxinas são trabalhosos, dispendiosos e demorados (por exemplo, anos entre a exposição e os sintomas), podem não ser representativos do verdadeiro efeito humano (versos apenas certos organismos modelo) e, em alguns casos, apresentam dificuldade em identificar o agente causador exato. Por exemplo, ocasionalmente, é necessária a detecção de um aumento da incidência de uma população de sujeitos doentes (por exemplo, aglomerados de câncer) antes de iniciar uma busca por uma genotoxina (por exemplo, análises farmacêuticas e de segurança alimentar, contaminantes ambientais ou investigação de descarte ambiental, etc.).

[0006] As medidas convencionais de mutação somática *in vivo* são indiretamente inferidas a partir de ensaios baseados em seleção em bactérias, cultura de células ou animais transgênicos, onde o efeito em todo o genoma é extrapolado de um pequeno repórter artificial. Por conseguinte, os ensaios atualmente utilizados são substitutos imperfeitos para o verdadeiro potencial genotóxico de um composto *in vivo*, e são trabalhosos, enquanto fornecem apenas um subconjunto limitado de informações sobre o potencial mutagênico de um composto. É provável

que muitos compostos que apresentam potencial mutagênico em sistemas bacterianos artificiais (isto é, o ensaio de Ames), não reflitam com precisão um risco genuíno em seres humanos e façam com que outros compostos terapeuticamente promissores sejam desnecessariamente retirados do desenvolvimento ou uso comercial. Da mesma forma, alguns compostos com potencial carcinogênico o fazem através de mecanismos mutagênicos não diretos que são indetectáveis nas bactérias. Tais compostos podem causar danos aos sujeitos, pois o risco não pode ser adequadamente reconhecido precocemente.

[0007] Os sistemas repórteres de mamíferos *in vivo*, como ensaios de roedores transgênicos (por exemplo, o rato e o camundongo BigBlue[®] e o camundongo Muta[™]), oferecem uma melhor aproximação do efeito de drogas humanas do que as bactérias. Embora sejam limitados na medida em que os animais não são representações perfeitas dos seres humanos, os ensaios transgênicos em mamíferos permanecem valiosos para testes de segurança pré-clínicos precoces; no entanto, esses ensaios são complexos e ainda são algo artificiais. O ensaio BigBlue[®], por exemplo, baseia-se em um sistema baseado em repórter, no qual um subconjunto de mutações que ocorrem em um transgene lambda-fago de múltiplas cópias pode ser identificado fenotipicamente após a recuperação do repórter por um vetor de vaivém que é então transfectado para bactérias. Nem todas as mutações que ocorrem no gene repórter da 294 BP podem ser detectadas, pois muitas não conferem um fenótipo. O próprio transgene é altamente condensado, metilado e não representa o estado de transcrição e condensação altamente variável do genoma mais amplo. A passagem de moléculas mutantes através de máquinas virais e bacterianas tem o potencial de introduzir mutações artefatuais e o estrangulamento inerente que ocorre em cada etapa significa que a fração alelo das mutações é não quantitativa. Além disso, o teste requer o uso de cepas específicas de um subconjunto limitado de espécies. E os próprios roedores não são representações perfeitas dos seres humanos. Por exemplo, a aflatoxina é altamente mutagênica em humanos, mas não é significativamente cancerígena em camundongos após a maturidade sexual quando certas enzimas metabólicas se expressam, o que facilita sua desintoxicação. Embora os roedores transgênicos continuem sendo um padrão ouro atual aceito pela Food and Drug Administration (FDA) dos EUA e outras agências reguladoras como uma métrica válida de genotoxicidade que pode ser usada como substituto da carcinogenicidade em algumas situações

de teste, está longe de ser o ideal como amplamente utilizado ferramenta para avaliar o potencial de um composto causar câncer em humanos.

[0008] É necessário um método rápido, flexível e confiável que permita a medição direta do potencial genotóxico de fatores/agentes/ambientes aos quais um sujeito possa ser exposto, causando mutações e danos aos ácidos nucleicos, contribuindo para certos riscos à saúde (câncer/malignidade/neoplasia, neurotoxicidade, neurodegeneração, infertilidade, defeitos congênitos, etc.) O método deve ser utilizável em qualquer locus genômico de qualquer tipo de tecido e/ou tipo de célula em qualquer tipo de organismo e sem a necessidade de nenhuma seleção clonal (conforme exigido nos testes padrão-ouro da técnica anterior) e ao fornecer informações (inferido ou diretamente) no mecanismo de ação de como o fator cancerígeno causa mutações ou outros danos genotóxicos in vivo, levando ao desenvolvimento de câncer ou outras doenças ou distúrbios no sujeito/organismo, ou outro organismo modelado pelo sujeito/organismo.

[0009] Se uma ferramenta suficientemente precisa e conveniente com esses recursos estivesse disponível, ela teria muitas aplicações, por exemplo: nos testes de segurança de medicamentos pré-clínicos e clínicos; na prevenção, diagnóstico e tratamento de doenças e distúrbios associados à genotoxina; na detecção e identificação de fatores/agentes causadores de mutações e seus mecanismos de ação; e outras implicações em todo o setor (por exemplo, testes de poluição ambiental e determinação de níveis limiars de início de toxicidade, testes de segurança de produtos de consumo de alto rendimento, diagnóstico e tratamento de pacientes se houver suspeita de exposição tóxica, avaliação de risco de segurança nacional de liberação intencional ou não intencional de genotoxinas etc.).

SUMÁRIO

[0010] A presente tecnologia é direcionada a métodos, sistemas e kits de reagentes para avaliar a genotoxicidade. Em particular, algumas modalidades da tecnologia são direcionadas à utilização do Sequenciamento Duplex para avaliar um potencial genotóxico de um composto (por exemplo, um composto químico) e/ou um agente ambiental (por exemplo, radiação) em um sujeito exposto. Por exemplo, várias modalidades da presente tecnologia incluem a execução de métodos de Sequenciamento Duplex que permitem a medição direta de mutações induzidas por compostos em qualquer contexto genômico de qualquer organismo, e sem a necessidade de qualquer seleção clonal. Outros exemplos da presente tecnologia são direcionados a métodos para

detectar e avaliar a mutagênese genômica *in vivo* usando o Sequenciamento Duplex e reagentes associados. Vários aspectos da presente tecnologia têm muitas aplicações em testes de segurança de drogas pré-clínicas e clínicas, além de outras implicações em todo o setor.

[0011] Em uma modalidade, a presente tecnologia compreende um método para detectar e quantificar mutações genômicas desenvolvidas *in vivo* em um sujeito após a exposição do sujeito a um mutagênico, compreendendo: (1) Sequenciamento Duplex de uma ou mais moléculas de DNA de fita dupla alvo extraídas de um sujeito exposto a um mutagênico; (2) gerar uma sequência de consenso corrigida por erro para as moléculas de DNA de fita dupla alvo; e (3) identificar um espectro de mutação para as moléculas de DNA de fita dupla direcionadas; (4) calcular uma frequência mutante para as moléculas de DNA de fita dupla alvo, calculando o número de mutações únicas por par de bases duplex, de um ou mais tipos, sequenciadas.

[0012] Em outra modalidade, a presente tecnologia compreende um método para gerar uma assinatura mutagênica de um composto de teste, compreendendo: (1) fragmentos de DNA de sequenciamento duplex extraídos de um organismo vivo, por exemplo, um animal de teste, exposto ao composto de teste; e (2) gerar uma assinatura mutagênica do composto de teste. E o método pode ainda compreender calcular uma frequência mutante para uma pluralidade de fragmentos de DNA, calculando o número de mutações únicas por par de bases duplex sequenciado.

[0013] Em outra modalidade, a presente tecnologia compreende um método para avaliar um potencial genotóxico de um composto, compreendendo: (1) fragmentos de DNA direcionados de sequenciamento duplex extraídos de um animal de teste exposto ao composto para gerar sequências de consenso corrigidas por erro dos fragmentos de DNA direcionados; (2) gerar uma assinatura mutagênica do composto a partir das sequências de consenso corrigidas por erros; e (3) determinar se a exposição ao composto resultou em uma assinatura mutagênica representativa de um composto suficientemente genotóxico.

[0014] Em outra modalidade, a presente tecnologia compreende kits compreendendo reagentes com instruções para a condução dos métodos aqui divulgados para detecção e quantificação de genotoxinas. Os kits podem ainda compreender um produto de programa de computador instalado em um dispositivo de computação eletrônica (por exemplo, laptop/desktop, tablet, etc.) ou acessível via rede (por exemplo, servidor remoto com um banco de dados de

registros de sujeitos e genotoxinas detectadas). O produto do programa de computador é incorporado em um meio legível por computador não transitório que, quando executado em um computador, executa etapas dos métodos usando os kits aqui divulgados para detectar e identificar genotoxinas.

[0015] Em outra modalidade, a presente tecnologia compreende um sistema de computador em rede para identificar ou confirmar a exposição de um sujeito a pelo menos uma genotoxina, compreendendo: (1) um servidor remoto; (2) uma pluralidade de dispositivos de computação eletrônica do usuário capazes de utilizar os kits aqui divulgados para extrair, amplificar, sequenciar uma amostra de um sujeito; (3) um banco de dados de terceiros com perfis conhecidos de genotoxina (opcional); e (4) uma rede com ou sem fio para transmitir comunicações eletrônicas entre os dispositivos de computação eletrônica, o banco de dados e o servidor remoto. O servidor remoto compreende ainda: (a) um banco de dados que armazena resultados de registros de genotoxinas do usuário e registros de perfis de genotoxinas (por exemplo, espectro, frequências, mecanismo de ações etc.); (b) um ou mais processadores acoplados comunicativamente a uma memória; e um ou mais dispositivos ou meios de armazenamento legíveis por computador não transitórios compreendendo instruções para o(s) processador(es), em que os referidos processadores estão configurados para executar as referidas instruções para executar operações que compreendem as etapas de: correção de erros nos fragmentos de Sequenciamento Duplex; e computar o espectro de mutação, a frequência mutante e o espectro de mutação tripleto dos agentes detectados, a partir dos quais a identidade de pelo menos uma genotoxina pode ser determinada.

[0016] A presente tecnologia compreende ainda, um meio de armazenamento legível por computador não transitório, compreendendo instruções que, quando executadas por um ou mais processadores, executam um método para determinar se um sujeito está exposto e/ou a identidade de pelo menos uma genotoxina, o método compreendendo as etapas de correção de erros nos fragmentos de Sequenciamento Duplex; e computando o espectro de mutação, frequência mutante e espectro tripleto de agentes detectados, a partir dos quais é determinada a identidade de pelo menos uma genotoxina.

[0017] A presente tecnologia compreende ainda um método computadorizado para determinar se um sujeito é exposto e/ou a identidade de pelo menos uma genotoxina, o método compreendendo as etapas de corrigir erros nos fragmentos de Sequenciamento Duplex; e computar

o espectro de mutação, frequência mutante e espectro triplete de agentes detectados, a partir dos quais é determinada a identidade de pelo menos uma genotoxina.

[0018] Em outra modalidade, a presente tecnologia compreende um método, sistema e kit para diagnosticar e tratar um sujeito exposto a uma genotoxina. O diagnóstico compreende detectar pelo menos uma genotoxina à qual o sujeito foi exposto e/ou consumido; e o tratamento compreende remover a exposição futura e/ou o consumo da(s) genotoxina(s) e/ou administrar protocolos de tratamento (por exemplo, produtos farmacêuticos) para bloquear e/ou contrariar o efeito biológico da(s) genotoxina(s).

[0019] Em outra modalidade, a presente tecnologia compreende um método, sistema computadorizado e kit para testes de segurança de drogas pré-clínicas e clínicas; para detectar e identificar agentes cancerígenos e seus mecanismos de ação; e por outras implicações em todo o setor (por exemplo, poluentes ambientais tóxicos, produtos de consumo de alto rendimento e testes de segurança de drogas, etc.).

[0020] Em outra modalidade, a presente tecnologia compreende um método, sistema e kit que identifica novas genotoxinas usando o Sequenciamento Duplex corrigido por erro e/ou então determina uma quantidade de limiar de segurança (peso, volume, concentração, etc.) e/ou uma frequência de mutante de limiar de segurança de uma genotoxina a qual um sujeito pode ser exposto antes que o sujeito esteja em risco de desenvolver uma doença ou distúrbio associado à genotoxina (por exemplo, usado no estabelecimento de padrões da Agência de Proteção Ambiental; usado no diagnóstico e tratamento de um sujeito exposto à genotoxina, etc.).

[0021] Em outra modalidade, a presente tecnologia compreende um método, sistema e kit para impedir que um sujeito desenvolva uma doença ou distúrbio associado à mutação, determinando se o sujeito foi exposto a uma genotoxina em mais de um nível limiar de segurança (por exemplo, quantidade de genotoxina e/ou frequência mutante de genotoxina e assinatura triplete); e, nesse caso, fornecer tratamento profilático para prevenir, inibir ou impedir o início da doença.

[0022] Um aspecto da presente tecnologia compreende a capacidade de detectar mutações que causam uma doença, mas dentro de alguns dias ou algumas semanas ou alguns meses ou alguns anos após a exposição a uma mutação que causa genotoxina. Normalmente, o início completo da doença não é diagnosticado por muitos anos (por exemplo, 10 a 20 anos para o

desenvolvimento de câncer de pulmão após a exposição ao amianto). Os métodos e kits divulgados neste documento permitem a detecção de mutações genômicas que causam o aparecimento da doença imediatamente após a exposição, contra anos de espera pelo aparecimento dos sintomas.

[0023] Outro aspecto da presente tecnologia compreende a capacidade de prever se um sujeito tem um risco aumentado de desenvolver uma doença ou distúrbio devido a mutações causadas pela genotoxina dentro de cerca de 2-5 dias, em um mínimo de anos depois de uma exposição potencial à genotoxina; e, se for o caso, fornecer tratamento profilático e triagem periódica para detectar o início da doença nos estágios iniciais.

[0024] Outro aspecto compreende uma biblioteca de DNA e um método de fabricação, compreendendo uma pluralidade de fragmentos de DNA genômico isolados de fita dupla, em que cada fragmento é ligado a uma ou mais moléculas adaptadoras desejadas.

[0025] Outro aspecto compreende um método de alto rendimento para rastrear rapidamente uma pluralidade de compostos para identificar quais compostos são genotóxicos.

[0026] Outro aspecto compreende um método de alto rendimento para rastrear rapidamente uma pluralidade de tipos diferentes de tecidos/células do mesmo sujeito para determinar se o sujeito foi exposto a qualquer genotoxina.

[0027] Outro aspecto compreende um método de alto rendimento para rastrear rapidamente uma pluralidade de tecidos e células derivados de diferentes sujeitos para determinar a porcentagem da população exposta a qualquer genotoxina.

[0028] Outro aspecto compreende determinar direta ou inferencialmente o "mecanismo de ação" da genotoxina que causa a exposição a ela, resultando em uma mutação associada a uma doença ou distúrbio específico.

[0029] Outras modalidades, aspectos e vantagens da presente tecnologia são descritas mais adiante na descrição detalhada a seguir.

BREVE DESCRIÇÃO DOS DESENHOS

[0030] Muitos aspectos da presente divulgação podem ser mais bem compreendidos com referência aos desenhos a seguir. Os componentes nos desenhos não estão necessariamente em escala. Em vez disso, é enfatizada a ilustração clara dos princípios da presente divulgação.

[0031] A FIG. 1A ilustra uma molécula adaptadora de ácido nucleico para uso com algumas modalidades da presente tecnologia e um complexo adaptador-ácido nucleico de fita

dupla resultante da ligação da molécula adaptadora a um fragmento de ácido nucleico de fita dupla de acordo com uma modalidade da presente tecnologia.

[0032] As FIGS. 1B e 1C são ilustrações conceituais de várias etapas do método de Sequenciamento Duplex de acordo com uma modalidade da presente tecnologia.

[0033] A FIG. 2A é uma ilustração conceitual de vários esquemas de métodos para o uso de estudos em animais *in vivo* para prever o risco de câncer em humanos de um composto de teste, incluindo estudos convencionais de carcinogenicidade em roedores a longo prazo (esquema à esquerda), um estudo convencional de mutagenicidade em roedores transgênicos com seleção *ex vivo* (esquema intermediário) e avaliação da mutagênese por meio de um esquema de sequenciamento direto de DNA, de acordo com aspectos da presente tecnologia (esquema à direita).

[0034] As FIGS. 2B e 2C são ilustrações conceituais de esquemas de métodos para o uso do Sequenciamento Duplex para avaliar a mutagênese *in vitro* de um composto de teste em células humanas cultivadas em cultura (2B) e para avaliar a mutagênese *in vivo* de um composto de teste em um camundongo de tipo selvagem (2C) de acordo com os aspectos da presente tecnologia.

[0035] As FIGS. 3A-3D são gráficos de caixas que mostram frequências mutantes calculadas para o sequenciamento em duplex (FIGS. 3A e 3B) e o ensaio de placas BigBlue® *cII* (FIGS. 3C e 3D) no fígado e medula óssea após tratamento com mutagênico e de acordo com uma modalidade da tecnologia atual.

[0036] A FIG. 3E é um gráfico que ilustra o aumento relativo da dobra mutante de *cII* no ensaio de placas BigBlue® *cII* versus o ensaio de Sequenciamento Duplex das FIGS. 3A-3D, e de acordo com uma modalidade da presente tecnologia.

[0037] A FIG. 3F mostra a proporção de variantes de nucleotídeo único (SNV) dentro do gene *cII* para placas mutantes escolhidas individualmente produzidas a partir de tecido de camundongo BigBlue® e Sequenciamento Duplex do gDNA de *cII* a partir dos tecidos de camundongo BigBlue® de acordo com uma modalidade da presente tecnologia.

[0038] As FIGS. 3G e 3H mostram a distribuição de mutações identificadas por Sequenciamento Duplex direto (FIG. 3G) e entre placas mutantes coletadas individualmente (FIG. 3H) de *cII* em todos os tipos de tecido BigBlue® e grupos de tratamento por posição do códon e consequência funcional, de acordo com uma modalidade da presente tecnologia.

[0039] A FIG. 4 é um gráfico de barras que mostra a frequência mutante medida por Sequenciamento Duplex em várias amostras de cada grupo de tratamento e de acordo com uma modalidade da presente tecnologia.

[0040] As FIGS. 5A e 5B são gráficos de barras que mostram a frequência mutante de genes endógenos em comparação com o transgene *cII* no fígado (FIG. 5A) e medula óssea (FIG. 5B) e medidos por Sequenciamento Duplex e de acordo com uma modalidade da presente tecnologia.

[0041] A FIG. 5C são gráficos de caixa que mostram a frequência mutante de SNV (MF) calculada para o Sequenciamento Duplex por regiões gênicas para Fígado e Medula Óssea para as categorias de tratamentos indicadas e de acordo com uma modalidade da presente tecnologia.

[0042] A FIG. 5D é um gráfico de dispersão que mostra medições individuais de dados agregados mostrados na FIG. 5C de acordo com uma modalidade da presente tecnologia.

[0043] A FIG. 6 é um gráfico de barras que mostra um espectro de mutação conforme medido por Sequenciamento Duplex e de acordo com uma modalidade da presente tecnologia.

[0044] As FIGS. 7A-7C são gráficos que mostram espectros de mutação trinucleotídica para controle de veículo (7A), Benzo[a]pireno (7B) e N-etil-N-nitrosoureia (7C) de acordo com uma modalidade da presente tecnologia.

[0045] A FIG. 8 é um gráfico de barras que mostra a frequência mutante de amostras de pulmão, baço e sangue para animais de controle e experimentais submetidos a uretano, de acordo com uma modalidade da presente tecnologia.

[0046] A FIG. 9 é um gráfico de barras que mostra uma frequência mutante de ponto mínimo médio entre grupos de amostras de tecido de acordo com uma modalidade da presente tecnologia.

[0047] A FIG. 10A são gráficos de caixa que mostram SNV MF calculado para Sequenciamento Duplex por regiões gênicas para Pulmão, Baço e Sangue para as categorias de tratamentos indicadas e de acordo com uma modalidade da presente tecnologia.

[0048] A FIG. 10B é um gráfico de dispersão que mostra medições individuais de dados agregados mostrados na FIG. 10A, e de acordo com uma modalidade da presente tecnologia.

[0049] A FIG. 11 é um gráfico de barras que mostra o espectro de mutação do uretano e um controle de veículo dentro dos tecidos testados, medido por Sequenciamento Duplex e de acordo com uma modalidade da presente tecnologia.

[0050] As FIGS. 12A e 12B são gráficos que mostram espectros de mutação no contexto de nucleotídeos adjacentes (isto é, espectros de trinucleotídeos) para controle de veículo (12A) e uretano (12B) de acordo com uma modalidade da presente tecnologia.

[0051] A FIG. 13 mostra a tendência de fita espectral de variante de nucleotídeo único (SNV) em amostras tratadas com uretano de acordo com uma modalidade da presente tecnologia.

[0052] A FIG. 14 é um gráfico que ilustra a seleção clonal neoplásica em estágio inicial de frações de alelos variantes, conforme detectado por Sequenciamento Duplex, de acordo com uma modalidade da presente tecnologia.

[0053] A FIG. 15A é um gráfico que ilustra SNVs representados sobre os intervalos genômicos para os éxons capturados da família *Ras* de genes, incluindo os loci transgênicos humanos, no modelo de camundongo Tg-rasH2 e de acordo com uma modalidade da presente tecnologia.

[0054] A FIG. 15B é um gráfico que ilustra variantes de nucleotídeo único alinhadas ao éxon 3 do transgene *HRAS* humano de acordo com uma modalidade da presente tecnologia.

[0055] As FIGS. 16A-16B são representações gráficas de dados de sequenciamento de uma seção representativa de 400 pares de bases de *HRAS* humano no pulmão de camundongo após tratamento com uretano usando sequenciamento de DNA convencional (FIG. 16A) e Sequenciamento Duplex (FIG. 16B) de acordo com a modalidade da presente tecnologia.

[0056] As FIGS. 17A-17C são gráficos que mostram espectros de mutação no contexto de nucleotídeos adjacentes (isto é, espectros de trinucleotídeos) para a Assinatura 1 (FIG. 17A), Assinatura 4 (FIG. 17B) e Assinatura 29 (FIG. 17C) da COSMIC.

[0057] A FIG. 18 mostra agrupamento hierárquico não supervisionado de todas as 30 assinaturas COSMIC publicadas e os 4 espectros de corte dos Exemplos 1 e 2 de acordo com uma modalidade da presente tecnologia.

[0058] A FIG. 19 é um diagrama esquemático de um sistema de computador em rede para uso com os métodos e/ou kits aqui divulgados para identificar eventos mutagênicos e/ou

eventos de danos aos ácidos nucleicos resultantes da exposição genotóxica de acordo com uma modalidade da presente tecnologia.

[0059] A FIG. 20 é um diagrama de fluxo que ilustra uma rotina para fornecer dados de sequência de consenso de Sequenciamento Duplex de acordo com uma modalidade da presente tecnologia, de acordo com uma modalidade da presente tecnologia.

[0060] A FIG. 21 é um diagrama de fluxo que ilustra uma rotina para detectar e identificar eventos mutagênicos resultantes da exposição genotóxica de uma amostra de acordo com uma modalidade da presente tecnologia.

[0061] A FIG. 22 é um diagrama de fluxo que ilustra uma rotina para detectar e identificar eventos de dano ao DNA resultantes da exposição genotóxica de uma amostra de acordo com uma modalidade da presente tecnologia.

[0062] A FIG. 23 é um diagrama de fluxo que ilustra uma rotina para detectar e identificar uma exposição a cancerígenos ou cancerígenos em um sujeito de acordo com uma modalidade da presente tecnologia.

DESCRIÇÃO DETALHADA

[0063] Detalhes específicos de várias modalidades da tecnologia são descritos abaixo com referência às FIGS. 1A-23. As modalidades podem incluir, por exemplo, métodos, sistemas, kits, etc. para avaliar a genotoxicidade. Algumas modalidades da tecnologia são direcionadas à utilização do Sequenciamento Duplex para avaliar um potencial genotóxico de um agente (por exemplo, um composto químico) ou qualquer outro tipo de exposição (por exemplo, uma fonte de radiação) em um sujeito exposto, organismo modelo ou sistema de cultura celular modelo. Outras modalidades da tecnologia são direcionadas à utilização do Sequenciamento Duplex para determinar uma assinatura de mutação associada a um agente genotóxico. Modalidades adicionais da tecnologia são direcionadas para identificar um ou mais agentes genotóxicos aos quais um sujeito pode ter sido exposto comparando o espectro de mutação do DNA do sujeito com os espectros de mutação de compostos mutagênicos conhecidos. Modalidades adicionais da tecnologia são direcionadas para identificar um ou mais locais ou ambientes aos quais um sujeito pode ter sido exposto, comparando o espectro de mutação do DNA do sujeito a partir de um ou mais tipos de células em um ou mais tecidos com espectros de mutação de ambientes ou compostos conhecidos por estarem presentes nesses locais ou ambientes. Modalidades adicionais da

tecnologia são direcionadas para identificar um sujeito, comparando o espectro de mutação do DNA do sujeito de um ou mais tipos de células em um ou mais tecidos com espectros de mutação de indivíduos conhecidos ou de locais ou ambientes aos quais o indivíduo sabe ter sido exposto ou compostos conhecidos por estarem presentes em tais locais ou ambientes. Em certas modalidades, uma genotoxina pode ser avaliada quanto ao potencial carcinogênico. Modalidades adicionais incluem identificar e avaliar o risco de carcinogênese resultante de agentes cancerígenos mutagênicos ou não mutagênicos através da identificação de clones portadores de mutação que estão surgindo com mutações ativadoras de câncer. Modalidades adicionais incluem identificar e avaliar o risco de carcinogênese resultante de agentes cancerígenos mutagênicos ou não mutagênicos, identificando a emergência de clones portadores de mutação onde as mutações não são consideradas ativadoras de câncer (geralmente conhecidas como mutações "passageiro" ou "carona"), mas substancialmente marcar clones de forma exclusiva (Salk and Horwitz Sem Cancer Bio 2010 PMID: 20951806) Outras modalidades da tecnologia são direcionadas à utilização do Sequenciamento Duplex para detectar e avaliar danos de ácidos nucleicos (particularmente danos ao DNA, como adutos) resultantes da exposição à genotoxina ou outros processos genotóxicos endógenos (por exemplo, envelhecimento).

[0064] Embora muitas das modalidades sejam aqui descritas em relação ao Sequenciamento Duplex, outras modalidades de sequenciamento capazes de gerar leituras de sequenciamento corrigidas por erros, além daquelas aqui descritas, estão dentro do escopo da presente tecnologia. Além disso, outras modalidades da presente tecnologia podem ter configurações, componentes ou procedimentos diferentes daqueles descritos aqui. Portanto, uma pessoa versada na técnica entenderá que a tecnologia pode ter outras modalidades com elementos adicionais e que a tecnologia pode ter outras modalidades sem várias das características mostradas e descritas abaixo com referência às FIGS. 1A-23.

Definições

[0065] Para que a presente divulgação seja mais facilmente compreendida, certos termos são definidos primeiro a seguir. Definições adicionais para os seguintes termos e outros termos são estabelecidas em todo o relatório descritivo.

[0066] Neste pedido, a menos que seja claro do contexto, o termo "um" pode ser entendido como "pelo menos um". Conforme usado neste pedido, o termo "ou" pode ser entendido

como "e/ou". Neste pedido, os termos "compreendendo" e "incluindo" podem ser entendidos como abrangendo componentes ou etapas discriminadas, sejam elas apresentadas por elas mesmas ou em conjunto com um ou mais componentes ou etapas adicionais. Onde as faixas são fornecidas aqui, os pontos de extremidade estão incluídos. Conforme usado neste pedido, o termo "compreender" e variações do termo, como "compreendendo" e "compreende", não se destinam a excluir outros aditivos, componentes, números inteiros ou etapas.

[0067] *Cerca de:* O termo "cerca de", quando usado aqui em referência a um valor, se refere a um valor semelhante, no contexto do valor referenciado. Em geral, aqueles versados na técnica, familiarizados com o contexto, apreciarão o grau de variação relevante englobado por "cerca de" nesse contexto. Por exemplo, em algumas modalidades, o termo "cerca de" pode abranger uma faixa de valores que dentro de 25%, 20%, 19%, 18%, 17%, 16%, 15%, 14%, 13%, 12%, 11%, 10%, 9%, 8%, 7%, 6%, 5%, 4%, 3%, 2%, 1% ou menos do valor referido. Para variações de valores inteiros de um dígito em que um único valor numérico na direção positiva ou negativa excederia 25% do valor, "cerca de" é geralmente aceito pelos versados na técnica para incluir, pelo menos 1, 2, 3, 4 ou 5 valores inteiros na direção positiva ou negativa, que podem ou não cruzar zero, dependendo das circunstâncias. Um exemplo não limitativo disso é a suposição de que 3 centavos podem ser considerados cerca de 5 centavos em algumas situações que seriam evidentes para um versado na técnica.

[0068] *Análogo:* Como utilizado neste documento, o termo "análogo" se refere a uma substância que compartilha uma ou mais características estruturais, elementos, componentes ou frações estruturais particulares com uma substância de referência. Normalmente, um "análogo" mostra uma similaridade estrutural significativa com a substância de referência, por exemplo, compartilhando uma estrutura principal ou de consenso, mas também difere de certas maneiras distintas. Em algumas modalidades, um análogo é uma substância que pode ser gerada a partir da substância de referência, por exemplo, por manipulação química da substância de referência. Em algumas modalidades, um análogo é uma substância que pode ser gerada através da realização de um processo sintético substancialmente semelhante a (por exemplo, compartilhar uma pluralidade de etapas com) um que gera a substância de referência. Em algumas modalidades, um análogo é ou pode ser gerado através do desempenho de um processo sintético diferente daquele usado para gerar a substância de referência.

[0069] **Amostra biológica:** Como utilizado neste documento, o termo "amostra biológica" ou "amostra" geralmente se refere a uma amostra obtida ou derivada de uma fonte biológica (por exemplo, um tecido ou organismo ou cultura de células) de interesse, conforme descrito aqui. Em algumas modalidades, uma fonte de interesse compreende um organismo, como um animal ou humano. Em outras modalidades, uma fonte de interesse compreende um micro-organismo, como uma bactéria, vírus, protozoário ou fungo. Em outras modalidades, uma fonte de interesse pode ser um tecido sintético, organismo, cultura de células, ácido nucleico ou outro material. Em ainda outras modalidades, uma fonte de interesse pode ser um organismo à base de plantas. Em ainda outra modalidade, uma amostra pode ser uma amostra ambiental, como, por exemplo, uma amostra de água, amostra de solo, amostra arqueológica ou outra amostra coletada de uma fonte não viva. Em outras modalidades, uma amostra pode ser uma amostra de múltiplos organismos (por exemplo, uma amostra de organismo misto). Em algumas modalidades, uma amostra biológica é ou compreende tecido ou fluido biológico. Em algumas modalidades, uma amostra biológica pode ser ou compreender medula óssea; sangue; células sanguíneas; ascites; amostras de tecido, amostras de biópsia ou amostras de aspiração por agulha fina; fluidos corporais contendo células; ácidos nucleicos flutuantes livres; ácidos nucleicos ligados a proteínas, ácidos nucleicos ligados a riboproteínas; escarro; saliva; urina; líquido cefalorraquidiano, líquido peritoneal; líquido pleural; fezes; linfa; fluidos ginecológicos; cotonetes de pele; cotonetes vaginais; exame de Papanicolaou, cotonetes orais; cotonetes nasais; lavagens, tais como lavagens ductais ou lavagens bronco-alveolares; fluido vaginal, aspirados; raspados; amostras de medula óssea; amostras de biópsia de tecido; tecido ou fluidos fetais; espécimes cirúrgicos; fezes, outros fluidos corporais, secreções e/ou excreções; e/ou células das mesmas, *etc.* Em algumas modalidades, uma amostra biológica é ou compreende células obtidas de um indivíduo. Em algumas modalidades, as células obtidas são ou incluem células de um indivíduo do qual a amostra é obtida. Em algumas modalidades, derivados de células, como organelas ou vesículas ou exossomos. Numa modalidade particular, uma amostra biológica é uma biópsia líquida obtida de um sujeito. Em algumas modalidades, uma amostra é uma "amostra primária" obtida diretamente de uma fonte de interesse por qualquer meio apropriado. Por exemplo, em algumas modalidades, uma amostra biológica primária é obtida por métodos selecionados do grupo que consiste em biópsia (*por exemplo*, aspiração por agulha fina ou biópsia de tecido), cirurgia, coleta de fluido

corporal (*por exemplo*, sangue, linfa, fezes *etc.*), *etc.* Em algumas modalidades, como ficará claro no contexto, o termo "amostra" se refere a uma preparação que é obtida pelo processamento (por exemplo, removendo um ou mais componentes de e/ou adicionando um ou mais agentes a) uma amostra primária. Por exemplo, filtragem usando uma membrana semipermeável. Tal "amostra processada" pode compreender, por exemplo, ácidos nucleicos ou proteínas extraídas de uma amostra ou obtidas submetendo uma amostra primária a técnicas como amplificação ou transcrição reversa de mRNA, isolamento e/ou purificação de certos componentes, *etc.*

[0070] ***Doença de câncer:*** Em uma modalidade, a doença ou distúrbio genotóxico associado é uma "doença de câncer" que é familiar às pessoas experientes na técnica como sendo geralmente caracterizada pelo crescimento desregulado de células anormais, que podem sofrer metástases. As doenças cancerígenas detectáveis usando um ou mais aspectos da presente tecnologia compreendem, por meio de exemplos não limitativos, câncer de próstata (isto é, adenocarcinoma, célula pequena), câncer de ovário (por exemplo, adenocarcinoma de ovário, carcinoma seroso ou carcinoma embrionário, tumor no saco vitelino, teratoma), câncer de fígado (por exemplo, HCC ou hepatoma, angiossarcoma), tumores de células plasmáticas (por exemplo, mieloma múltiplo, leucemia plasmocítica, plasmocitoma, amiloidose, macroglobulinemia de Waldenstrom), câncer colorretal (por exemplo, colonic adenocarcinoma, colonic mucinous adenocarcinoma, carcinoid, lymphoma and rectal adenocarcinoma, rectal squamous carcinoma), leukemia (por exemplo, leucemia mieloide aguda, leucemia linfocítica aguda, leucemia mieloide crônica, leucemia linfocítica crônica, leucemia mieloblástica aguda, leucemia promielocítica aguda, leucemia mielomonocítica aguda, leucemia monocítica aguda, eritroleucemia aguda e leucemia crônica, leucemia de células T, síndrome de Sezary, mastocitose sistêmica, leucemia de células pilosas, crise crônica da explosão da leucemia mieloide crônica), síndrome mielodisplásica, linfoma (por exemplo, linfoma difuso de células B grandes, linfoma cutâneo de células T, linfoma periférico de células T, linfoma de Hodgkin, linfoma não Hodgkin, linfoma folicular, linfoma de células do manto, linfoma MALT, linfoma de células marginais, transformação de Richter, linfoma de duplo golpe, linfoma associado ao transplante, linfoma do CNS, linfoma extranodal, linfoma associado ao HIV, linfoma endêmico, linfoma de Burkitt, neoplasias linfoproliferativas associadas ao transplante, e linfoma linfocítico *etc.*), câncer do colo do útero (carcinoma escamoso do colo do útero, carcinoma de células claras, carcinoma associado ao HPV, sarcoma do colo do útero *etc.*)

câncer de esôfago (carcinoma de células escamosas do esôfago, adenocarcinoma, certos tipos de esôfago de Barretts, adenocarcinoma de esôfago), melanoma (melanoma dérmico, melanoma uveal, melanoma acral, melanoma amelanótico etc.), tumores do CNS (por exemplo, oligodendroglioma, astrocitoma, glioblastoma multiforme, meningioma, schwannoma, craniofaringioma etc.), câncer de pâncreas (por exemplo, adenocarcinoma, carcinoma adenoescamoso, carcinoma de células em anel de sinete, carcinoma hepatoide, carcinoma colóide, carcinoma de células de ilhotas, carcinoma neuroendócrino pancreático etc.), tumor estromal gastrointestinal, sarcoma (por exemplo, fibrossarcoma, mixossarcoma, lipossarcoma, condrosarcoma, sarcoma osteogênico, angiosarcoma, sarcoma de endotelioma, linfangiossarcoma, sarcoma de linfangioendotelioma, leiomiossarcoma, sarcoma de células de Ewing, e rabiomiossarcoma, tumor de células fusiformes etc.), câncer de mama (por exemplo, carcinoma inflamatório, carcinoma lobar, carcinoma ductal etc.), câncer ER-positivo, câncer HER-2 positivo, câncer de bexiga (câncer de bexiga escamosa, câncer de bexiga de pequenas células, câncer urotelial etc.), câncer de cabeça e pescoço (por exemplo, carcinoma espinocelular da cabeça e pescoço, carcinoma espinocelular associado ao HPV, carcinoma nasofaríngeo etc.), câncer de pulmão (por exemplo, carcinoma pulmonar de células não pequenas, carcinoma de células grandes, carcinoma broncogênico, câncer de células escamosas, câncer de células pequenas, etc.), câncer metastático, câncer de cavidade oral, câncer uterino (leiomiossarcoma, leiomioma etc.), câncer testicular (por exemplo, tumor de saco vitelino seminoma, não seminoma e carcinoma embrionário etc), câncer de pele (por exemplo, carcinoma espinocelular e carcinoma basocelular, carcinoma de células Merkel, melanoma, linfoma cutâneo de células T etc.), câncer de tireoide (por exemplo, carcinoma papilar, carcinoma medular, câncer de tireoide anaplásico etc.), câncer de estômago, intra- câncer epitelial, câncer ósseo, câncer do trato biliar, câncer ocular, câncer da laringe, câncer renal (por exemplo, carcinoma de células renais, tumor de Wilms etc.), câncer gástrico, blastoma (por exemplo, nefroblastoma, meduloblastoma, hemangioblastoma, neuroblastoma, retinoblastoma etc.), neoplasias mieloproliferativas (policitemia vera, trombocitose essencial, mielofibrose), etc.), cordoma, sinovioma, mesotelioma, adenocarcinoma, carcinoma da glândula sudorípara, carcinoma da glândula sebácea, cistadenocarcinoma, carcinoma do ducto biliar, coriocarcinoma, carcinoma epitelial, ependimoma, pinealoma, neuroma acústico, schwannoma, meningioma, adenoma hipofisário, tumor da bainha nervosa, câncer do intestino delgado,

feocromocitoma, câncer de pulmão de pequenas células, mesotelioma peritoneal, adenoma hiperparatireoidiano, câncer adrenal, câncer de origem primária desconhecida, câncer do sistema endócrino, câncer do pênis, câncer da uretra, melanoma cutâneo ou intra-ocular, tumor ginecológico, tumores sólidos da infância ou neoplasias do sistema nervoso central, tumor primário de células germinativas mediastinais, hematopoiese clonal de potencial indeterminado, mieloma latente, gamaglobulinopatia monoclonal de linfocitose monoclonal de células B significativa desconhecida, câncer de baixo grau, defeitos de campo clonal, neoplasias pré-neoplásicas, câncer ureteral, cânceres autoimunes (ou seja, colite ulcerativa, colangite esclerosante primária, doença celíaca), cânceres associados a uma predisposição herdada (ou seja, aqueles portadores de defeitos genéticos como *BRCA1*, *BRCA2*, *TP53*, *PTEN*, *ATM*, etc.) e várias síndromes genéticas como MEN1, MEN2 trissomia 21 etc.) e aquelas que ocorrem quando expostas a substâncias químicas no útero (isto é, câncer de células claras em filhos de mulheres expostas ao Dietilestilbestrol [DES]), entre muitas outras.

[0071] *Ativador de câncer ou Gene ativador de câncer:* Como utilizado neste documento, "ativador de câncer" ou "gene ativador de câncer" se refere a uma lesão genética que tem o potencial de permitir que uma célula, no contexto certo, sofra transformação maligna. Esses genes incluem supressores de tumores (por exemplo, *TP53*, *BRCA1*) que normalmente suprimem a transformação de malignidade e, quando mutados de certas maneiras, não o fazem mais. Outros genes ativadores podem ser oncogenes (por exemplo, *KRAS*, *EGFR*) que, quando mutados de certas maneiras, tornam-se constitutivamente ativos ou ganham novas propriedades que facilitam a célula a se tornar maligna. Outras mutações encontradas em regiões não codificantes do genoma podem ser ativadoras de câncer. Por exemplo, uma mutação da região promotora do gene da telomerase (*TERT*) pode resultar na superexpressão do gene e, assim, tornar-se um ativador de câncer. Certos rearranjos (por exemplo, fusão *BCR-ABL*) podem justapor uma região genética à de outra para acionar a tumorigênese através de mecanismos relacionados à superexpressão, perda de repressão ou genes de fusão quiméricos. Em termos gerais, mutações genéticas (ou epimutações) que conferem um fenótipo a uma célula que facilita sua proliferação, sobrevivência ou vantagem competitiva sobre outras células ou que tornam sua capacidade de evoluir mais robusta, podem ser consideradas uma mutação ativadora. Isso deve ser contrastado com mutações que carecem de tais características, mesmo que possam estar no mesmo gene (isto é, uma mutação sinônima). Quando

essas mutações são identificadas nos tumores, elas são comumente referidas como mutações de passageiros, porque “pegam carona” junto com a expansão clonal sem contribuir significativamente para a expansão. Como reconhecido por um versado comum na técnica, a distinção entre motorista e passageiro não é absoluta e não deve ser interpretada como tal. Alguns ativadores funcionam apenas em determinadas situações (por exemplo, certos tecidos) e outros podem não funcionar na ausência de outras mutações ou epimutações ou outros fatores.

[0072] **Amostra de controle:** Como utilizado neste documento, uma "amostra de controle" se refere a uma amostra isolada da mesma maneira que a amostra com a qual é comparada, exceto que a amostra de controle não está exposta a um agente, ambiente ou processo sendo avaliado quanto ao potencial genotóxico.

[0073] **Determinar:** Muitas metodologias descritas aqui incluem uma etapa de "determinação". Os versados na técnica, lendo o presente relatório descritivo, apreciarão que essa "determinação" pode utilizar ou ser realizada através do uso de qualquer uma das várias técnicas disponíveis para os versados na técnica, incluindo, por exemplo, técnicas específicas explicitamente referidas a aqui. Em algumas modalidades, a determinação envolve a manipulação de uma amostra física. Em algumas modalidades, a determinação envolve consideração e/ou manipulação de dados ou informações, por exemplo, utilizando um computador ou outra unidade de processamento adaptada para executar uma análise relevante. Em algumas modalidades, a determinação envolve o recebimento de informações e/ou materiais relevantes de uma fonte. Em algumas modalidades, a determinação envolve comparar um ou mais recursos de uma amostra ou entidade com uma referência comparável.

[0074] **Sequenciamento Duplex (DS):** Como utilizado neste documento, "Sequenciamento Duplex (DS)" é, em seu sentido mais amplo, se refere a um método de correção de erros baseado em etiqueta que atinge precisão excepcional comparando a sequência de ambas as fitas de moléculas de DNA individuais.

[0075] **Genotoxicidade:** Como utilizado neste documento, o termo "genotoxicidade" se refere à propriedade destrutiva de agentes ou processos (ou seja, genotoxinas) que causam danos ao material genético (por exemplo, DNA, RNA). O dano polinucleotídico, a formação de uma mutação genética e/ou a interrupção da estrutura normal dos ácidos nucleicos resultante direta ou indiretamente da exposição a uma genotoxina são aspectos da genotoxicidade.

Um sujeito exposto a uma genotoxina pode potencialmente desenvolver uma doença ou distúrbio (por exemplo, câncer) imediatamente ou anos depois. Em uma modalidade, a presente tecnologia é direcionada em parte à identificação de eventos e/ou fatores contribuintes (por exemplo, agentes, processos) que causam genotoxicidade em um sujeito, a fim de prevenir ou reduzir o risco do aparecimento da doença ou distúrbio e/ou combater seus efeitos adversos. Em outras modalidades, iniciar a genotoxicidade é por design, como para criar diversidade em uma biblioteca genética.

[0076] ***Genotoxina ou agente ou fator genotóxico:*** Como utilizado neste documento, o termo "genotoxina" ou "agente ou fator genotóxico" se refere a, por exemplo, qualquer produto químico ao qual uma fonte de ácido nucleico (por exemplo, fonte biológica, sujeito) esteja exposta e/ou consome, exposições ambientais e/ou qualquer evento desencadeante (mutação precursora endógena) que causa dano polinucleotídico, mutação genômica ou ruptura da estrutura normal dos ácidos nucleicos. Em algumas modalidades, uma genotoxina tem a capacidade de direta ou indiretamente (por exemplo, acionar um precursor mutagênico), ou ambos, causar um desenvolvimento de doença ou distúrbio em um sujeito. Os fatores ou agentes genotóxicos que podem ser detectados pela presente tecnologia compreendem, a título de exemplo não limitativo, um produto químico ou uma mistura de produtos químicos (por exemplo, produtos farmacêuticos, aditivos e subprodutos industriais - resíduos, destilados de petróleo, metais pesados, cosméticos, produtos de limpeza para uso doméstico, partículas no ar, produtos alimentícios, subprodutos da fabricação, contaminantes, plastificantes, detergentes, etc.); e radiação (radiação de partículas, fótons, ou ambos) e/ou forças físicas (por exemplo, um campo magnético, campo gravitacional, forças de aceleração, etc.) geradas pelo ambiente natural ou pelo homem (por exemplo, de um dispositivo). A genotoxina pode ainda compreender uma formulação líquida, sólida e/ou aerossol e a exposição da mesma pode ser por qualquer via de administração. Um agente ou fator genotóxico pode ser exógeno (por exemplo, a exposição se origina de fora da fonte biológica ou, em outros casos, o agente ou fator genotóxico pode ser endógeno à fonte biológica ou uma combinação dos mesmos. Um agente ou fator de origem exógena pode se tornar genotóxico uma vez que essa exposição é processada endogenamente. Ainda em outros exemplos, um agente ou fator pode se tornar genotóxico quando combinado com um ou mais agentes ou fatores adicionais e, em alguns casos, ter um efeito sinérgico. Exemplos adicionais de fatores ou agentes genotóxicos podem incluir ainda um organismo capaz de, direta ou indiretamente, causar

danos aos ácidos nucleicos em um sujeito em exposição (por exemplo, através de infecção do sujeito), como por meio de exemplos não limitativos, a esquistossomose contribuindo para câncer de bexiga, HPV contribuindo para câncer de colo uterino ou de cabeça e pescoço, poliomavírus contribuindo para carcinoma de células de Merkel, *Helicobacter pylori* contribuindo para câncer gástrico, infecção bacteriana crônica de uma ferida na pele que contribui para carcinoma espinocelular, etc. Agentes ou fatores genotóxicos adicionais podem incluir ainda um organismo capaz de produzir (por exemplo, dentro de si mesmo ou secretar) um agente genotóxico, como por meio de exemplos não limitativos, aflatoxina de *aspergillus flavus* ou ácido aristolóquico da família de plantas aristocholia, etc. Os fatores ou agentes genotóxicos que podem ser detectados usando vários aspectos da presente tecnologia podem compreender ainda genotoxinas endógenas, que podem não ser capazes de ser quantificadas com precisão ou controladas experimentalmente, como por meio de exemplos não limitativos, estresse, inflamação, efeitos de tratamentos terapêuticos (por exemplo, terapia genética, terapia de edição de genes, terapia com células-tronco, outras terapias celulares, produtos farmacêuticos, radiografia, etc.). Fatores endógenos também podem representar o acúmulo agregado de mutações e outros eventos genotóxicos nos tecidos de um sujeito que refletem os efeitos integrais das exposições do sujeito.

[0077] Doença ou distúrbio associado genotóxico: Conforme utilizado neste documento, o termo "doença ou distúrbio associado a genotóxico" se refere a qualquer condição médica resultante de uma mutação genômica ou outro dano ou rearranjo polinucleotídico em um sujeito causado direta ou indiretamente pela exposição a uma ou mais genotoxinas. Uma doença ou distúrbio associado a genotóxicos pode estar relacionado ou não ao câncer. Além disso, o dano/rearranjo ou mutação do polinucleotídeo pode estar em uma célula germinativa ou célula somática. Em exemplos, onde uma célula germinativa é afetada, é contemplado que a doença ou distúrbio associado a genotóxico possa se manifestar (ou de outra forma conferir um risco) a um sujeito que é descendente de um sujeito exposto.

[0078] Agente suficientemente genotóxico: Como utilizado neste documento, o termo "agente suficientemente genotóxico" se refere a um agente, fator, composto ou processo identificado pelo sistema, métodos e kits da presente tecnologia para ter cerca de 50%, cerca de 40%, cerca de 30%, cerca de 20%, cerca de 10%, cerca de 5%, cerca de 4%, cerca de 3%, cerca de 2%, cerca de 1%, cerca de 0,5%, cerca de 0,1%, cerca de 0,01%, cerca de 0,001%, cerca de

0,0001%, cerca de 0,00001%, cerca de 0,000001% etc. probabilidade de causar dano ou mutação de ácido nucleico em um ou mais resíduos de nucleotídeos em uma ou mais moléculas que podem derivar de um ou mais organismos biológicos que foram expostos. Em algumas modalidades, um agente suficientemente genotóxico pode ter mais de cerca de 50% de probabilidade de causar dano ou mutação por ácido nucleico que acima de um nível de controle de fundo. Em algumas modalidades, um agente suficientemente genotóxico se refere a um agente, fator, composto ou processo identificado pelo sistema, métodos e kits da presente tecnologia para ter cerca de 50%, cerca de 40%, cerca de 30%, cerca de 20%, cerca de 10%, cerca de 5%, cerca de 4%, cerca de 3%, cerca de 2%, cerca de 1%, cerca de 0,5%, cerca de 0,1%, cerca de 0,01%, cerca de 0,001%, cerca de 0,0001%, cerca de 0,00001% etc. de probabilidade de causar uma doença ou distúrbio em um sujeito exposto à genotoxina.

[0079] *Inibir o crescimento:* Como usado neste documento, o termo "inibir o crescimento" em uma doença cancerígena se refere a uma redução no crescimento celular (por exemplo, tamanho do tumor, taxa de divisão de células cancerígenas etc.) *in vivo* ou *in vitro* por, por exemplo, cerca de 5%, cerca de 10%, cerca de 20%, cerca de 30%, cerca de 40%, cerca de 50%, cerca de 60%, cerca de 70%, cerca de 80%, cerca de 90%, cerca de 95% ou cerca de 99% ou mais, como evidente por uma redução na proliferação de células e/ou no tamanho/massa de células expostas a um tratamento em relação à proliferação e/ou crescimento de tamanho de células nas células na ausência do tratamento. A inibição do crescimento pode ser o resultado de um tratamento que induz apoptose em uma célula, induz necrose em uma célula, retarda a progressão do ciclo celular, interrompe o metabolismo celular, induz a lise celular ou induz algum outro mecanismo que reduz a proliferação e/ou crescimento do tamanho de célula das células.

[0080] *Expressão:* Como utilizado neste documento, "expressão" de uma sequência de ácido nucleico se refere a um ou mais dos seguintes eventos: (1) produção de um modelo de RNA a partir de uma sequência de DNA (*por exemplo*, por transcrição); (2) processamento de um transcrito de RNA (por exemplo, junção, edição, formação de tampa 5' e/ou formação de extremidade 3'); (3) tradução de um RNA em um polipeptídeo ou proteína; e/ou (4) modificação pós-tradução de um polipeptídeo ou proteína.

[0081] *Mecanismo de ação:* Conforme utilizado neste documento, o termo "mecanismo de ação" se refere ao processo bioquímico que resulta em alteração no ácido nucleico

após a exposição a uma genotoxina. Em uma modalidade, o "mecanismo de ação" se refere à via bioquímica e/ou processos fisiopatológicos que seguem a mutação ou dano genômico até o início completo da doença ou distúrbio. Em outra modalidade, o "mecanismo de ação" inclui a via bioquímica e/ou processos fisiológicos que ocorrem em uma fonte biológica após a exposição à genotoxina e que resulta em dano genômico (por exemplo, lesões pré-mutagênicas) ou mutação. Em ainda outra modalidade, o mecanismo de ação de um agente ou processo genotóxico pode ser inferido a partir de um ou mais dos seguintes: a base nucleotídica afetada, a alteração nucleotídica introduzida, o tipo de dano ao DNA introduzido, a alteração estrutural introduzida, o contexto da sequência nucleotídica de flaqueamento do(s) nucleotídeo(s) afetado(s), o contexto genético ou a(s) sequência(s) afetada(s), o status transcricional ou a região afetada, o status de metilação da região afetada, o status ligado à proteína ou o status de condensação ou o local cromossômico da região afetada pela exposição à genotoxina.

[0082] *Mutação:* Como utilizado neste documento, o termo "mutação" se refere a alterações na sequência ou estrutura de ácidos nucleicos. Mutações em uma sequência polinucleotídica podem incluir mutações pontuais (por exemplo, mutações de base única), mutações multinucleotídicas, deleções nucleotídicas, rearranjos de sequências, inserções nucleotídicas e duplicações da sequência de DNA na amostra, entre alterações multinucleotídicas complexas. As mutações podem ocorrer em ambas as fitas de uma molécula de DNA duplex como alterações complementares de base (isto é, verdadeiras mutações) ou como uma mutação em uma fita, mas não na outra, ou seja, heteroduplex, que tem o potencial de ser reparada, destruída ou ser mal reparada/convertida em uma verdadeira mutação de fita dupla.

[0083] *Frequência mutante:* Como utilizado neste documento, o termo "frequência mutante", também chamado de "frequência mutante", se refere ao número de mutações únicas detectadas pelo número total de pares de bases duplex sequenciados. Em algumas modalidades, a frequência mutante é a frequência de mutações dentro de apenas um gene específico, ou um conjunto de genes ou um conjunto de alvos genômicos. Em algumas modalidades, a frequência mutante pode se referir apenas a certos tipos de mutações (por exemplo, a frequência de mutações A>T, que é calculada como o número de mutações A>T pelo número total de bases A). A frequência com que as mutações são introduzidas em uma população de células ou moléculas pode variar de acordo com a genotoxina, quantidade de tempo ou nível de exposição a uma genotoxina,

idade do sujeito, ao longo do tempo, tipo de tecido ou organização, região de um genoma, por tipo de mutação, pelo contexto dos trinucleotídeos, fundo genético herdado, entre outras coisas.

[0084] **Assinatura de mutação:** Conforme utilizado neste documento, o termo "assinatura de mutação" e "**espectro ou espectros de mutação**" se refere a combinações características de tipos de mutação decorrentes de processos de mutagênese, como infidelidade à replicação de DNA, exposições exógenas e endógenas a genotoxinas, vias de reparo de DNA com defeito e edição enzimática de DNA. Em uma modalidade, o espectro de mutação é gerado por correspondência de padrões computacionais (por exemplo, agrupamento hierárquico não supervisionado do espectro de mutação).

[0085] **Doença não cancerígena:** Em outra modalidade, a doença ou distúrbio genotóxico associado é uma doença não cancerígena; em vez disso, é outro tipo de doença ou distúrbio causado por, ou contribuído por, uma mutação ou dano genômico. A título de exemplos não limitativos, esses tipos de doenças ou distúrbios não cancerígenos que são detectáveis ou previstos usando um ou mais aspectos da presente tecnologia compreendem diabetes; doenças ou distúrbios autoimunes, infertilidade, neurodegeneração, progeria, doença cardiovascular, qualquer doença associada ao tratamento de outra doença mediada geneticamente (por exemplo, neuropatia mediada por quimioterapia e insuficiência renal associada a quimioterapia como cisplatina), Alzheimer/demência, obesidade, doença cardíaca, pressão alta, artrite, doença mental, outros distúrbios neurológicos (neurofibromatose) e um distúrbio de herança multifatorial (por exemplo, uma predisposição desencadeada por fatores ambientais).

[0086] **Ácido nucleico:** Como utilizado neste documento, no seu sentido mais amplo, se refere a qualquer composto e/ou substância que é ou pode ser incorporada a uma cadeia oligonucleotídica. Em algumas modalidades, um ácido nucleico é um composto e/ou substância que é ou pode ser incorporada a uma cadeia oligonucleotídica por meio de uma ligação fosfodiéster. Como ficará claro no contexto, em algumas modalidades, "*ácido nucleico*" se refere a um resíduo de ácido nucleico individual (por exemplo, um nucleotídeo e/ou nucleosídeo); em algumas modalidades, "*ácido nucleico*" se refere a uma cadeia oligonucleotídica compreendendo resíduos individuais de ácido nucleico. Em algumas modalidades, um "*ácido nucleico*"; é ou compreende RNA; em algumas modalidades, um "*ácido nucleico*"; é ou compreende DNA. Em algumas modalidades, um ácido nucleico é, compreende ou consiste em um ou mais resíduos naturais de

ácido nucleico. Em algumas modalidades, um ácido nucleico é, compreende ou consiste em um ou mais análogos de ácido nucleico. Em algumas modalidades, um análogo de ácido nucleico difere de um ácido nucleico pelo fato de não utilizar uma espinha dorsal de fosfodiéster. Por exemplo, em algumas modalidades, um ácido nucleico é, compreende ou consiste em um ou mais "ácidos nucleicos peptídicos", que são conhecidos na técnica e têm ligações peptídicas em vez de ligações fosfodiéster na espinha dorsal, são considerados dentro do escopo da tecnologia atual. Alternativamente, ou adicionalmente, em algumas modalidades, um ácido nucleico tem uma ou mais ligações fosforotioato e/ou ligações 5'-N-fosforamidita em vez de ligações fosfodiéster. Em algumas modalidades, um ácido nucleico é, compreende ou consiste em um ou mais nucleosídeos naturais (por exemplo, adenosina, timidina, guanosina, citidina, uridina, desoxiadenosina, desoxitimidina, desoxiacitana e desoxicidina). Em algumas modalidades, um ácido nucleico é, compreende ou consiste em um ou mais análogos de nucleosídeo (por exemplo, 2-aminoadenosina, 2-tiotimidina, inosina, pirrolo-pirimidina, 3-metil adenosina, 5-metilcitidina, C-5 propinil-citidina, C-5 propinil-uridina, 2-aminoadenosina, C5-bromouridina, C5-fluorouridina, C5-iodouridina, C5-propinil-uridina, C5-propinil-citidina, C5-metilcitidina, 2-aminoadenosina, 7-deaza-adenosina, 7-deazaguanosina, 8-oxoadenosina, 8-oxoguanosina, 0(6)-metilguanina, 2-tiocitidina, bases metiladas, bases intercaladas e combinações das mesmas). Em algumas modalidades, um ácido nucleico compreende um ou mais açúcares modificados (por exemplo, 2'-fluororibose, ribose, 2'-desoxirribose, arabinose e hexose) em comparação com os ácidos nucleicos naturais. Em algumas modalidades, um ácido nucleico tem uma sequência nucleotídica que codifica um produto genético funcional, como um RNA ou proteína. Em algumas modalidades, um ácido nucleico inclui um ou mais íntrons. Em algumas modalidades, os ácidos nucleicos são preparados por um ou mais isolamentos de uma fonte natural, síntese enzimática por polimerização com base em um modelo complementar (*in vivo* ou *in vitro*), reprodução em uma célula ou sistema recombinante e síntese química. Em algumas modalidades, um ácido nucleico tem pelo menos 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100, 110, 120, 130, 140, 150, 160, 170, 180, 190, 200, 225, 250, 275, 300, 325, 350, 375, 400, 425, 450, 475, 500, 600, 700, 800, 900, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000 ou mais resíduos de comprimento. Em algumas modalidades, um ácido nucleico é parcial ou totalmente de fita única; em algumas modalidades, um ácido nucleico é parcial ou totalmente de fita dupla. Em algumas modalidades,

um ácido nucleico pode ser ramificado de ter estruturas secundárias. Em algumas modalidades, um ácido nucleico tem uma sequência nucleotídica compreendendo pelo menos um elemento que codifica, ou é o complemento de uma sequência que codifica, um polipeptídeo. Em algumas modalidades, um ácido nucleico tem atividade enzimática. Em algumas modalidades, o ácido nucleico desempenha uma função mecânica, por exemplo, em um complexo de ribonucleoproteínas ou em um RNA de transferência.

[0087] ***Composição ou formulação farmacêutica:*** Como utilizado neste documento, o termo "composição farmacêutica" compreende uma quantidade farmacologicamente eficaz de uma droga ou agente ativo e um transportador farmacêuticamente aceitável. Em alguns exemplos, vários aspectos da presente tecnologia podem ser utilizados para avaliar a genotoxicidade da composição ou formulação farmacêutica, ou a droga ou agente ativo nela.

[0088] ***Dano polinucleotídico:*** Conforme utilizado neste documento, o termo "dano polinucleotídico" ou "dano ao ácido nucleico" se refere a dano na sequência de ácido desoxirribonucleico (DNA) de um sujeito ("dano ao DNA") ou na sequência de ácido ribonucleico (RNA) ("dano ao RNA") que é direta ou indiretamente (por exemplo, um metabólito ou indução de um processo que é prejudicial ou mutagênico) causado por uma genotoxina. O ácido nucleico danificado pode levar ao aparecimento de uma doença ou distúrbio associado à exposição à genotoxina em um sujeito. Em algumas modalidades, a detecção de ácido nucleico danificado em um sujeito pode ser uma indicação de uma exposição à genotoxina. O dano polinucleotídico pode ainda compreender modificação química e/ou física do DNA em uma célula. Em algumas modalidades, o dano é ou compreende, por meio de exemplos não limitativos, pelo menos um de oxidação, alquilação, desaminação, metilação, hidrólise, hidroxilação, corte, reticulações intra-fitas reticulações entre fitas, quebra de fita de extremidade cega, quebra de fita dupla de extremidade escalonada, fosforilação, desfosforilação, somalilação, glicosilação, desglicosilação, putrescinilação, carboxilação, halogenação, formilação, folgas de fita única, dano por calor, dano por dessecação, dano por exposição a UV, dano por radiação gama por radiação X, dano por radiação ionizante, dano por radiação não ionizante, dano por radiação de partículas pesadas, dano por decaimento nuclear, dano da radiação beta, dano da radiação alfa, dano da radiação de nêutrons, dano da radiação de prótons, dano da radiação cósmica, dano do pH alto, dano do pH baixo, dano das espécies oxidativas reativas, dano dos radicais livres, dano do peróxido, dano do hipoclorito,

dano da fixação do tecido como formalina ou formaldeído, danos causados pelo ferro reativo, danos causados por condições iônicas baixas, danos causados por condições iônicas elevadas, danos causados por condições sem tampão e danos por nucleases, danos por exposição ambiental, danos por incêndio, danos por estresse mecânico, danos por degradação enzimática, danos por micro-organismos, danos por cisalhamento mecânico preparativo, danos por fragmentação enzimática preparativa, danos ocorridos naturalmente *in vivo*, dano ocorrido durante a extração de ácido nucleico, dano ocorrido durante a preparação da biblioteca de sequenciamento, dano introduzido por uma polimerase, dano introduzido durante o reparo de ácido nucleico, dano ocorrido durante a finalização do ácido nucleico, dano ocorrido durante a ligação do ácido nucleico, danos ocorridos durante o sequenciamento, danos causados pelo manuseio mecânico do DNA, dano ocorrido durante a passagem através de um nanoporo, dano ocorrido como parte do envelhecimento em um organismo, dano causado como resultado caso ocorra a exposição química de um indivíduo, dano causado por um mutagênico, dano ocorrido por um cancerígeno, dano causado por um clastogênio, dano causado por dano à inflamação *in vivo* devido à exposição ao oxigênio, dano devido a uma ou mais quebras de fita e qualquer combinação dos mesmos.

[0089] **Referência:** Conforme utilizado neste documento, descreve um padrão ou controle em relação ao qual uma comparação é realizada. Por exemplo, em algumas modalidades, um agente, animal, indivíduo, população, amostra, sequência ou valor de interesse é comparado com um agente de referência ou controle, animal, indivíduo, população, amostra, sequência ou valor ou representação do mesmo em uma representação física ou banco de dados de computador que pode estar presente em um local ou acessado remotamente por meios eletrônicos. Em algumas modalidades, uma referência ou controle é testado e/ou determinado substancialmente simultaneamente com o teste ou determinação de interesse. Em algumas modalidades, uma referência ou controle é uma referência ou controle histórico, opcionalmente corporificado em um meio tangível. Tipicamente, como seria entendido pelos versados na técnica, uma referência ou controle é determinado ou caracterizado em condições ou circunstâncias comparáveis àquelas sob avaliação. Os versados na técnica apreciarão quando houver semelhanças suficientes para justificar a confiança e/ou comparação com uma referência ou controle possível particular. Uma "amostra de referência" se refere a uma amostra de um sujeito que é distinto do sujeito de teste e isolado da mesma maneira que a amostra com a qual é comparada e que foi exposta a uma quantidade

conhecida do mesmo agente genotóxico. O sujeito da amostra de referência pode ser geneticamente idêntico ao sujeito de teste ou pode ser diferente. Além disso, a amostra de referência pode ser derivada de vários sujeitos que foram expostos a uma quantidade conhecida do mesmo agente genotóxico.

[0090] **Nível de limiar seguro:** Conforme utilizado neste documento, o termo "nível de limiar seguro" se refere à quantidade (por exemplo, peso, volume, concentração, massa, abundância molar, unidade * tempo integral etc.) de uma genotoxina específica ou uma combinação de genotoxinas que um sujeito pode ser exposto antes que uma provável mutação genômica ocorra levando ao início da doença. Por exemplo, um nível de limiar seguro pode ser zero. Em outros exemplos, um nível de exposição à genotoxina pode ser tolerável. A tolerância ao risco aceitável de exposição pode variar dependendo do sujeito, idade, sexo, tipo de tecido, condição de saúde do paciente e outras considerações de risco-benefício familiares a alguém experiente na técnica etc.

[0091] **Frequência mutante do limiar seguro:** Como utilizado neste documento, o termo "frequência mutante do limiar seguro" se refere a uma taxa aceitável de mutação causada por um agente ou processo genotóxico, abaixo do qual um sujeito assume um risco aceitável de adquirir uma doença ou distúrbio associado a genotóxico. A tolerância ao risco aceitável de exposição e a taxa de mutação resultante podem diferir dependendo do sujeito, idade, sexo, tipo de tecido, condição de saúde do paciente, etc.

[0092] **Identificador de molécula única (SMI):** Como utilizado neste documento, o termo "identificador de molécula única" ou "SMI" (que pode ser chamado de "etiqueta", "código de barras", "código de barras molecular", um "Identificador Molecular Único", ou "UMI", entre outros nomes) se refere a qualquer material (por exemplo, uma sequência nucleotídica, um recurso de molécula de ácido nucleico) que é capaz de distinguir substancialmente uma molécula individual entre uma população heterogênea maior de moléculas. Em algumas modalidades, um SMI pode ser ou compreender um SMI aplicado exogenamente. Em algumas modalidades, um SMI aplicado exogenamente pode ser ou compreender uma sequência degenerada ou semidegenerada. Em algumas modalidades SMIs substancialmente degeneradas podem ser conhecidas como Identificadores Moleculares Aleatórios Exclusivos (R-UMIs). Em algumas modalidades, um SMI pode compreender um código (por exemplo, uma sequência de ácido

nucleico) de dentro de um conjunto de códigos conhecidos. Em algumas modalidades, os códigos SMI predefinidos são conhecidos como Identificadores Moleculares Exclusivos Definidos (D-UMIs). Em algumas modalidades, um SMI pode ser ou compreender um SMI endógeno. Em algumas modalidades, um SMI endógeno pode ser ou compreender informações relacionadas a pontos de cisalhamento específicos de uma sequência alvo, recursos relacionados às extremidades terminais de moléculas individuais que compreendem uma sequência alvo ou uma sequência específica a ou adjacente a ou dentro de uma distância conhecida do fim de moléculas individuais. Em algumas modalidades, um SMI pode se relacionar com uma variação de sequência em uma molécula de ácido nucleico causada por dano aleatório ou semialeatório, modificação química, modificação enzimática ou outra modificação na molécula de ácido nucleico. Em algumas modalidades, a modificação pode ser desaminação da metilcitosina. Em algumas modalidades, a modificação pode implicar sítios de cortes de ácidos nucleicos. Em algumas modalidades, um SMI pode compreender elementos exógenos e endógenos. Em algumas modalidades, um SMI pode compreender elementos SMI fisicamente adjacentes. Em algumas modalidades, os elementos SMI podem ser espacialmente distintos em uma molécula. Em algumas modalidades, um SMI pode ser um ácido não nucleico. Em algumas modalidades, um SMI pode compreender dois ou mais tipos diferentes de informações SMI. Várias modalidades de SMIs são divulgadas ainda na Publicação Internacional de Patente WO2017/100441, que é incorporada por referência aqui na sua totalidade.

[0093] *Elemento de definição de fita (SDE):* Conforme utilizado neste documento, o termo "Elemento de definição de fita" ou "SDE" se refere a qualquer material que permita a identificação de uma fita específica de um material de ácido nucleico de fita dupla e, assim, a diferenciação da outra/fita complementar (por exemplo, qualquer material que processe os produtos de amplificação de cada um dos dois ácidos nucleicos de fita única resultantes de um ácido nucleico de fita dupla alvo substancialmente distinguível um do outro após sequenciamento ou outra interrogação de ácido nucleico). Em algumas modalidades, um SDE pode ser ou compreender um ou mais segmentos de sequência substancialmente não complementar dentro de uma sequência adaptadora. Em modalidades particulares, um segmento de sequência substancialmente não complementar dentro de uma sequência adaptadora pode ser fornecido por uma molécula adaptadora compreendendo uma forma em Y ou uma forma de "alça". Em outras modalidades, um segmento de sequência substancialmente não complementar dentro de uma

sequência adaptadora pode formar uma "bolha" não emparelhada no meio de sequências complementares adjacentes dentro de uma sequência adaptadora. Em outras modalidades, um SDE pode abranger uma modificação de ácido nucleico. Em algumas modalidades, um SDE pode compreender separação física de fitas pareadas em compartimentos de reação fisicamente separados. Em algumas modalidades, um SDE pode compreender uma modificação química. Em algumas modalidades, um SDE pode compreender um ácido nucleico modificado. Em algumas modalidades, um SDE pode se relacionar com uma variação de sequência em uma molécula de ácido nucleico causada por dano aleatório ou semialeatório, modificação química, modificação enzimática ou outra modificação na molécula de ácido nucleico. Em algumas modalidades, a modificação pode ser desaminação da metilcitosina. Em algumas modalidades, a modificação pode implicar sítios de cortes de ácidos nucleicos. Várias modalidades de SDEs são divulgadas ainda na Publicação Internacional de Patente WO2017/100441, que é incorporada por referência aqui na sua totalidade.

[0094] *Sujeito:* Como utilizado neste documento, o termo "sujeito" se refere a um organismo, tipicamente um mamífero, como um humano (em algumas modalidades, incluindo formas humanas pré-natais), um animal não humano (por exemplo, mamíferos e não mamíferos, incluindo, mas não limitado a primatas não humanos, cavalos, ovelhas, cães, vacas, porcos, galinhas, anfíbios, répteis, vida marinha (geralmente excluindo macacos marinhos), outros organismos modelo, como vermes, moscas etc.) e animais transgênicos (por exemplo, roedores transgênicos) etc. Em algumas modalidades, um sujeito foi exposto à genotoxina ou fator ou agente genotóxico, ou em outra modalidade, o sujeito foi exposto a uma potencial genotoxina. Em algumas modalidades, um sujeito está sofrendo de uma doença, distúrbio ou condição relevante. Em algumas modalidades, um sujeito está sofrendo de uma doença ou distúrbio genotóxico associado. Em algumas modalidades, um sujeito é suscetível a uma doença, distúrbio ou condição. Em algumas modalidades, um sujeito exibe um ou mais sintomas ou características de uma doença, distúrbio ou condição. Em algumas modalidades, um sujeito não exibe nenhum sintoma ou característica de uma doença, distúrbio ou condição. Em algumas modalidades, um sujeito tem um ou mais aspectos característicos de suscetibilidade ou risco de uma doença, distúrbio ou condição. Em algumas modalidades, um sujeito está exibindo um sintoma ou característica de uma doença, distúrbio ou condição e, em algumas modalidades, esse sintoma ou característica está associado a

uma doença ou distúrbio genotóxico associado. Em algumas modalidades, um sujeito é um paciente. Em algumas modalidades, um sujeito é um indivíduo a quem o diagnóstico e/ou terapia é e/ou foi administrado. Em ainda outras modalidades, um sujeito se refere a quaisquer fontes biológicas vivas ou outro material de ácido nucleico, que pode ser exposto a genotoxinas e pode incluir, por exemplo, organismos, células e/ou tecidos, como para estudos *in vivo*, por exemplo: fungos, protozoários, bactérias, arqueobactérias, vírus, células isoladas em cultura, células que foram intencionalmente (por exemplo, transplante de células-tronco, transplante de órgãos) ou involuntariamente (por exemplo, microcimerismo fetal ou materno) ou ácidos nucleicos ou organelas isoladas (isto é, mitocôndrias, cloroplastos, genomas virais livres, plasmídeos livres, aptâmeros, ribozimas ou derivados ou precursores de ácidos nucleicos (isto é, oligonucleotídeos, trifosfatos de dinucleotídeos, etc.).

[0095] Substancialmente: Conforme utilizado neste documento, o termo "substancialmente" se refere à condição qualitativa de exibir extensão ou grau total ou quase total de uma característica ou propriedade de interesse. Alguém versado na técnica das ciências biológicas entenderá que os fenômenos biológicos e químicos raramente, se é que alguma vez, chegam à conclusão e/ou prosseguem à conclusão ou alcançam ou evitam um resultado absoluto. O termo "substancialmente" é, portanto, usado aqui para capturar a potencial falta de completude inerente a muitos fenômenos biológicos e químicos.

[0096] Quantidade terapeuticamente eficaz: Como utilizado neste documento, o termo "quantidade terapeuticamente eficaz" ou "quantidade farmacologicamente eficaz" ou simplesmente "quantidade eficaz" se refere à quantidade de uma droga ou agente ativo para produzir um resultado farmacológico, terapêutico ou preventivo pretendido. Em alguns exemplos, vários aspectos da presente tecnologia podem ser usados para avaliar ou determinar uma quantidade eficaz de uma droga ou agente ativo (por exemplo, uma droga ativa distribuída para induzir propositalmente eventos associados à genotoxicidade).

[0097] Contexto de trinucleotídeo ou trinucleotídeo: Como utilizado neste documento, os termos "trinucleotídeo" ou "contexto de trinucleotídeo" se referem a um nucleotídeo no contexto de bases nucleotídicas imediatamente precedentes e imediatamente após em sequência (por exemplo, um mononucleotídeo dentro de uma combinação de três mononucleotídeos).

[0098] ***Espectro ou assinatura de trinucleotídeo:*** Aqui, o termo "assinatura de trinucleotídeo" é usado de forma intercambiável com "espectro de trinucleotídeo", "assinatura de triplete" e "espectro de triplete" se refere a uma assinatura de mutação, como as associadas a uma exposição de genotoxina, em um contexto de trinucleotídeo. Numa modalidade, uma genotoxina pode ter um espectro/assinatura único, semiúnico e/ou identificável de outro modo triplete.

[0099] ***Tratamento:*** Como utilizado neste documento, o termo "tratamento" se refere à aplicação ou administração de um agente terapêutico a um sujeito, ou aplicação ou administração de um agente terapêutico a um tecido ou linhagem celular isolado de um sujeito que tenha um distúrbio, por exemplo, uma doença ou condição, um sintoma de doença ou uma predisposição para uma doença, com o objetivo de curar, sarar, aliviar, abrandar, alterar, remediar, melhorar, aprimorar ou afetar a doença, os sintomas da doença ou a predisposição para a doença. Em um exemplo, o distúrbio ou doença/condição é uma doença ou distúrbio genotóxico. Em outro exemplo, o distúrbio ou doença/condição não é uma doença ou distúrbio genotóxico. Em alguns exemplos, vários aspectos da presente tecnologia são usados para avaliar a genotoxicidade do tratamento ou de um potencial tratamento.

Modalidades Seleccionadas de Métodos de Sequenciamento Duplex e Adaptadores e Reagentes Associados

[00100] O Sequenciamento Duplex é um método para a produção de sequências de DNA corrigidas por erros a partir de moléculas de ácido nucleico de fita dupla e que foi originalmente descrito na Publicação Internacional de Patente WO 2013/142389 e na Patente U.S. 9.752.188, e WO 2017/100441, em Schmitt *et. al.*, PNAS, 2012 [1]; em Kennedy *et. al.*, PLOS Genetics, 2013 [2]; em Kennedy *et. al.*, Nature Protocols, 2014 [3]; e em Schmitt *et. al.*, Nature Methods, 2015 [4]. Cada uma das patentes, pedidos de patente e publicações acima mencionadas são incorporadas aqui por referência em sua totalidade. Como ilustrado nas FIGS. 1A-1C, e em certos aspectos da tecnologia, o Sequenciamento Duplex pode ser usado para sequenciar independentemente ambas as fitas de moléculas de DNA individuais, de modo que as leituras da sequência derivada possam ser reconhecidas como originárias da mesma molécula parental de ácido nucleico de fita dupla durante o sequenciamento massivamente paralelo (MPS), também conhecido como sequenciamento de próxima geração (NGS), mas também diferenciado entre si como entidades distinguíveis após o sequenciamento. As leituras de sequência resultantes de cada

fita são então comparadas com o objetivo de obter uma sequência corrigida por erro da molécula original de ácido nucleico de fita dupla conhecida como Sequência de Consenso Duplex (DCS). O processo de Sequenciamento Duplex permite confirmar explicitamente que ambas as fitas de uma molécula original de ácido nucleico de fita dupla estão representadas nos dados de sequenciamentos gerados usados para formar um DCS.

[00101] Em certas modalidades, os métodos que incorporam o DS podem incluir a ligação de um ou mais adaptadores de sequenciamento a uma molécula de ácido nucleico de fita dupla alvo, compreendendo uma sequência de ácido nucleico alvo de primeira fita e uma sequência nucleica alvo de segunda fita, para produzir um complexo de ácido nucleico alvo de fita dupla (por exemplo, FIG. 1A).

[00102] Em várias modalidades, um complexo de ácido nucleico alvo resultante pode incluir pelo menos uma sequência SMI, que pode implicar uma sequência degenerada ou semidegenerada aplicada exogenamente (por exemplo, etiqueta duplex aleatória mostrada na FIG. 1A, sequências identificadas como α e β na FIG. 1A), informações endógenas relacionadas aos pontos de cisalhamento específicos da molécula alvo de ácido nucleico de fita dupla ou uma combinação das mesmas. O SMI pode tornar a molécula de ácido nucleico alvo substancialmente distinguível da pluralidade de outras moléculas em uma população que é sequenciada sozinha ou em combinação com elementos distintivos dos fragmentos de ácido nucleico aos quais foram ligados. O recurso substancialmente distinguível do elemento SMI pode ser transportado independentemente por cada uma das fitas únicas que formam a molécula de ácido nucleico de fita dupla, de modo que os produtos de amplificação derivativos de cada fita possam ser reconhecidos como provenientes da mesma molécula original de ácido nucleico de fita dupla substancialmente única após o sequenciamento. Em outras modalidades, o SMI pode incluir informações adicionais e/ou pode ser usado em outros métodos para os quais essa funcionalidade de diferenciação de molécula é útil, como os descritos nas publicações acima mencionadas. Numa outra modalidade, o elemento SMI pode ser incorporado após a ligação do adaptador. Em algumas modalidades, o SMI é de natureza dupla. Em outras modalidades, é de natureza de fita única (por exemplo, o SMI pode estar na parte de fita única dos adaptadores). Em outras modalidades, é uma combinação de natureza de fita única e dupla.

[00103] Em algumas modalidades, cada complexo de sequência de ácido nucleico alvo de fita dupla pode incluir ainda um elemento (por exemplo, um SDE) que processa os produtos de amplificação dos dois ácidos nucleicos de fita única que formam a molécula de ácido nucleico de fita dupla alvo substancialmente distinguível uma da outra após a sequência. Em uma modalidade, um SDE pode compreender sítios primários assimétricos compreendidos nos adaptadores de sequenciamento ou, em outros arranjos, assimetrias de sequência podem ser introduzidas nas moléculas adaptadoras que não estão nas sequências iniciadoras, de modo que pelo menos uma posição nas sequências nucleotídicas do complexo de sequência de ácido nucleico alvo da primeira fita e da segunda fita do complexo de sequência de ácido nucleico alvo são diferentes um do outro após amplificação e sequenciamento. Em outras modalidades, o SMI pode compreender outra assimetria bioquímica entre as duas fitas que diferem das sequências nucleotídicas canônicas A, T, C, G ou U, mas é convertida em pelo menos uma diferença de sequência nucleotídica canônica nas duas moléculas amplificadas e sequenciadas. Em ainda outra modalidade, o SDE pode ser um meio de separar fisicamente as duas fitas antes da amplificação, de modo que os produtos de amplificação derivados da sequência de ácido nucleico alvo da primeira fita e a sequência de ácido nucleico alvo da segunda fita sejam mantidos em isolamento físico substancial um do outro com o objetivo de manter uma distinção entre os dois. Outros arranjos ou metodologias para fornecer uma função SDE que permita distinguir a primeira e a segunda fitas podem ser utilizados, como os descritos nas publicações mencionadas acima, ou outros métodos que atendem ao objetivo funcional descrito.

[00104] Depois de gerar o complexo de ácido nucleico alvo de fita dupla que compreende pelo menos um SMI e pelo menos um SDE, ou onde um ou ambos esses elementos serão subsequentemente introduzidos, o complexo pode ser submetido à amplificação de DNA, como com PCR, ou qualquer outro método bioquímico de amplificação de DNA (por exemplo, amplificação de círculo rolante, amplificação de deslocamento múltiplo, amplificação isotérmica, amplificação de ponte ou amplificação ligada à superfície, de modo que uma ou mais cópias da primeira fita direcionem a sequência de ácido nucleico e uma ou mais cópias da sequência de ácido nucleico alvo da segunda fita são produzidas (por exemplo, FIG. 1B). As uma ou mais cópias de amplificação da molécula de ácido nucleico alvo da primeira fita e as uma ou mais cópias de amplificação da segunda molécula de ácido nucleico alvo podem ser submetidas ao

sequenciamento de DNA, de preferência usando uma plataforma de sequenciamento de DNA massivamente paralela de "próxima geração" (por exemplo, FIG. 1B).

[00105] As leituras de sequência produzidas a partir da molécula de ácido nucleico alvo da primeira fita e da molécula de ácido nucleico alvo da segunda fita derivadas da molécula original de ácido nucleico alvo de fita dupla podem ser identificadas com base no compartilhamento de um SMI substancialmente único relacionado e distinto da molécula de ácido nucleico alvo de fita oposta em virtude de um SDE. Em algumas modalidades, o SMI pode ser uma sequência baseada em um código de correção de erros matematicamente baseado (por exemplo, um código de Hamming), pelo qual certos erros de amplificação, erros de sequenciamento ou erros de síntese SMI podem ser tolerados com a finalidade de relacionar as sequências SMI em fitas complementares de um Duplex original (por exemplo, uma molécula de ácido nucleico de fita dupla). Por exemplo, com um SMI exógeno de fita dupla, em que o SMI compreende 15 pares de bases de sequência totalmente degenerada de bases de DNA canônicas, existirão aproximadamente $4^{15} = 1.073.741.824$ variantes SMI estimadas em uma população de SMIs totalmente degeneradas. Se dois SMIs são recuperados de leituras de dados de sequenciamento que diferem em apenas um nucleotídeo dentro da sequência SMI de uma população de 10.000 SMIs amostrados, pode-se calcular matematicamente a probabilidade de isso ocorrer por acaso e tomar uma decisão se é mais provável que a diferença de um par de bases reflita um dos tipos de erros mencionados acima e que as sequências SMI possam ser determinadas como sendo de fato derivadas da mesma molécula duplex original. Em algumas modalidades em que o SMI é, pelo menos em parte, uma sequência aplicada exogenamente em que as variantes de sequência não são totalmente degeneradas entre si e são, pelo menos em parte, sequências conhecidas, a identidade das sequências conhecidas pode em algumas modalidades ser projetada de tal forma que um ou mais erros dos tipos mencionados acima não convertam a identidade de uma sequência SMI conhecida na de outra sequência SMI, de modo que a probabilidade de um SMI ser mal interpretado como a de outro SMI é reduzida. Em algumas modalidades, esta estratégia de projeto de SMI compreende uma abordagem do Código de Hamming ou derivado do mesmo. Uma vez identificada, uma ou mais leituras de sequência produzidas a partir da molécula de ácido nucleico alvo da primeira fita são comparadas com uma ou mais leituras de sequência produzidas a partir da molécula de ácido nucleico alvo da segunda fita para produzir uma sequência

da molécula de ácido nucleico alvo corrigida por erro (por exemplo, FIG. 1C). Por exemplo, as posições nucleotídicas nas quais as bases das sequências de ácido nucleico alvo da primeira e da segunda fita concordam são consideradas sequências verdadeiras, enquanto as posições nucleotídicas que discordam entre as duas fitas são reconhecidas como sítios potenciais de erros técnicos que podem ser descontados, eliminados, corrigidos ou identificados. Uma sequência corrigida por erro da molécula de ácido nucleico alvo de fita dupla original pode assim ser produzida (mostrada na FIG. 1C). Em algumas modalidades e após o agrupamento separado de cada uma das leituras de sequenciamento produzidas a partir da molécula de ácido nucleico alvo da primeira fita e da molécula de ácido nucleico alvo da segunda fita, uma sequência de consenso de fita única pode ser gerada para cada uma da primeira e da segunda fitas. As sequências de consenso de fita única da molécula de ácido nucleico alvo da primeira fita e a molécula de ácido nucleico alvo da segunda fita podem então ser comparadas para produzir uma sequência da molécula de ácido nucleico alvo corrigida por erro (por exemplo, FIG. 1C).

[00106] Alternativamente, em algumas modalidades, os sítios de desacordo da sequência entre as duas fitas podem ser reconhecidos como sítios potenciais de incompatibilidades derivadas biologicamente na molécula original de ácido nucleico alvo de fita dupla. Alternativamente, em algumas modalidades, os sítios de desacordo da sequência entre as duas fitas podem ser reconhecidos como sítios potenciais de incompatibilidades derivadas de síntese de DNA na molécula original de ácido nucleico alvo de fita dupla. Alternativamente, em algumas modalidades, os sítios de desacordo de sequência entre as duas fitas podem ser reconhecidos como sítios potenciais em que uma base nucleotídica danificada ou modificada estava presente em uma ou em ambas as fitas e foi convertida em uma incompatibilidade por um processo enzimático (por exemplo, uma DNA polimerase, uma DNA glicosilase ou outra enzima modificadora de ácido nucleico ou processo químico). Em algumas modalidades, esta última descoberta pode ser usada para inferir a presença de dano por ácido nucleico ou modificação de nucleotídeo antes do processo enzimático ou tratamento químico.

[00107] Em algumas modalidades, e de acordo com aspectos da presente tecnologia, as leituras de sequenciamento geradas a partir das etapas de Sequenciamento Duplex discutidas aqui podem ser filtradas ainda mais para eliminar leituras de sequenciamento de moléculas danificadas por DNA (por exemplo, danificadas durante o armazenamento, transporte, durante ou

após a extração de tecido ou sangue durante ou após a preparação da biblioteca, etc.). Por exemplo, enzimas de reparo de DNA, como Uracil-DNA Glicosilase (UDG), Formamidopirimidina DNA glicosilase (FPG) e 8-oxoguanina DNA glicosilase (OGG1), podem ser utilizadas para eliminar ou corrigir danos ao DNA (por exemplo, danos ao DNA *in vitro* ou danos *in vivo*). Essas enzimas de reparo do DNA, por exemplo, são glicosilases que removem bases danificadas do DNA. Por exemplo, o UDG remove o uracil resultante da desaminação da citosina (causada pela hidrólise espontânea da citosina) e o FPG remove a 8-oxo-guanina (por exemplo, uma lesão de DNA comum resultante de espécies reativas de oxigênio). O FPG também possui atividade de liase que pode gerar uma folga de 1 base em sítios abásicos. Tais sítios abásicos geralmente falham subsequentemente em amplificar por PCR, por exemplo, porque a polimerase falha em copiar o modelo. Consequentemente, o uso de tais enzimas de reparo/eliminação de danos ao DNA pode remover efetivamente o DNA danificado que não possui uma mutação verdadeira, mas que pode ser detectado como um erro após a sequência e a análise de sequência duplex. Embora um erro devido a uma base danificada possa frequentemente ser corrigido pelo Sequenciamento Duplex em casos raros, teoricamente, um erro complementar pode ocorrer na mesma posição em ambas as fitas, assim, reduzir o dano que aumenta o erro pode reduzir a probabilidade de artefatos. Além disso, durante a preparação da biblioteca, certos fragmentos de DNA a serem sequenciados podem ser de fita única a partir de sua fonte ou das etapas de processamento (por exemplo, cisalhamento mecânico de DNA). Essas regiões são tipicamente convertidas em DNA de fita dupla durante uma etapa de "reparo final" conhecida na técnica, na qual uma DNA polimerase e substratos de nucleosídeo são adicionados a uma amostra de DNA para estender as extremidades recuadas em 5'. Um sítio mutagênico de dano ao DNA na porção de fita única do DNA que está sendo copiado (isto é, saliência de fita única de 5' em uma ou ambas as extremidades do duplex de DNA ou cortes ou folgas internas de fita única) pode causar um erro durante a reação de preenchimento que poderia processar uma mutação de fita única, erro de síntese ou sítio de dano do ácido nucleico em uma forma de fita dupla que poderia ser mal interpretada na sequência de consenso duplex final como uma verdadeira mutação, pela qual a verdadeira mutação estava presente na molécula de ácido nucleico de fita dupla original, quando na verdade não estava. Esse cenário, denominado "pseudo-duplex", pode ser reduzido ou evitado pelo uso de tais enzimas destruidoras/reparadoras de danos. Em outras modalidades, essa ocorrência pode ser reduzida ou eliminada através do uso

de estratégias para destruir ou impedir a formação de porções de fita única da molécula duplex original (por exemplo, o uso de certas enzimas sendo usadas para fragmentar o material original de ácido nucleico de fita dupla, em vez de cisalhamento mecânico ou certas outras enzimas que podem deixar cortes ou folgas). Em outras modalidades, o uso de processos para eliminar porções de fita única de ácidos nucleicos originais de fita dupla (por exemplo, nucleases específicas de fita única, como nuclease S1 ou nuclease de feijão mungo), pode ser utilizado para uma finalidade semelhante.

[00108] Em outras modalidades, as leituras de sequenciamento geradas a partir das etapas de Sequenciamento Duplex discutidas aqui podem ser filtradas ainda mais para eliminar falsas mutações aparando as extremidades das leituras mais propensas a artefatos de pseudoduplex. Por exemplo, a fragmentação do DNA pode gerar porções de fita única nas extremidades terminais da molécula de fita dupla. Essas porções de fita única podem ser preenchidas (por exemplo, pela Klenow ou T4 polimerase) durante o reparo final. Em alguns casos, as polimerases cometem erros de cópia nessas regiões reparadas finais, levando à geração de "moléculas pseudoduplex". Esses artefatos de preparação da biblioteca podem incorretamente parecer verdadeiras mutações depois de sequenciados. Esses erros, como resultado de mecanismos de reparo final, podem ser eliminados ou reduzidos da análise pós-sequenciamento aparando as extremidades das leituras de sequenciamento para excluir quaisquer mutações que possam ter ocorrido em regiões de maior risco, reduzindo assim o número de falsas mutações. Numa modalidade, esse corte de leituras de sequenciamento pode ser realizado automaticamente (por exemplo, uma etapa normal do processo). Em outra modalidade, uma frequência mutante pode ser avaliada para regiões terminais do fragmento e se um nível limiar de mutações for observado nas regiões terminais do fragmento, o corte de leitura de sequenciamento pode ser realizado antes de gerar uma leitura de sequência de consenso de fita dupla dos fragmentos de DNA.

[00109] A título de exemplo específico, em algumas modalidades, são fornecidos aqui métodos para gerar uma sequência de leitura corrigida por erro de um material de ácido nucleico alvo de fita dupla, incluindo a etapa de ligar um material de ácido nucleico alvo de fita dupla a pelo menos uma sequência adaptadora, para formar um complexo de material de ácido nucleico alvo-adaptador, em que a pelo menos uma sequência adaptadora compreende (a) uma sequência de identificador de molécula única (SMI) degenerada ou semidegenerada que marca

exclusivamente cada molécula do material de ácido nucleico alvo de fita dupla, e (b) uma primeira sequência adaptadora de nucleotídeo que marca uma primeira fita do complexo de material de ácido nucleico alvo-adaptador e uma segunda sequência adaptadora de nucleotídeo que é pelo menos parcialmente não complementar à primeira sequência nucleotídica que marca uma segunda fita do complexo de material de ácido nucleico alvo-adaptador, de modo que cada fita do complexo de material de ácido nucleico alvo-adaptador tenha uma sequência nucleotídica distintamente identificável em relação à sua fita complementar. O método pode, em seguida, incluir as etapas de amplificação de cada fita do complexo de material de ácido nucleico alvo-adaptador para produzir uma pluralidade de amplicons do complexo de ácido nucleico alvo-adaptador da primeira fita e uma pluralidade de amplicons do complexo de ácido nucleico alvo-adaptador da segunda fita. O método pode ainda incluir as etapas de amplificação do primeiro e das fitas para fornecer um primeiro produto de ácido nucleico e um segundo produto de ácido nucleico. O método também pode incluir as etapas de sequenciar cada um do primeiro produto de ácido nucleico e do segundo produto de ácido nucleico para produzir uma pluralidade de leituras de sequência da primeira fita e pluralidade de leituras de sequência da segunda fita e confirmar a presença de pelo menos uma leitura da sequência da primeira fita e pelo menos uma leitura da sequência da segunda fita. O método pode ainda incluir comparar pelo menos uma leitura de sequência de primeira fita com a pelo menos uma leitura de sequência de segunda fita, e gerar uma leitura de sequência corrigida por erro do material de ácido nucleico alvo de fita dupla, descontando posições de nucleotídeos que não concordam, ou removendo, alternativamente, as leituras de sequência da primeira e da segunda fitas comparadas com uma ou mais posições nucleotídicas em que as leituras de sequência da primeira e da segunda fitas comparadas não são complementares.

[00110] Por meio de um exemplo específico adicional, em algumas modalidades, são aqui fornecidos métodos para identificar uma variante de DNA de uma amostra incluindo as etapas de ligação de ambas as fitas de um material de ácido nucleico (por exemplo, uma molécula de DNA alvo de fita dupla) a pelo menos uma molécula adaptadora assimétrica para formar um complexo de material de ácido nucleico alvo-adaptador tendo uma primeira sequência nucleotídica associada a uma primeira fita de uma molécula de DNA alvo de fita dupla (por exemplo, uma fita superior) e uma segunda sequência nucleotídica que é pelo menos parcialmente não complementar à primeira sequência nucleotídica associada a uma segunda fita da molécula de DNA alvo de fita

dupla (por exemplo, uma fita inferior) e amplificar cada fita do material de ácido nucleico alvo-adaptador, resultando em cada fita gerando um distinto, porém relacionado conjunto de produtos de ácido nucleico alvo-adaptador amplificado. O método pode ainda incluir as etapas de sequenciar cada uma de uma pluralidade de produtos de ácido nucleico alvo-adaptador da primeira fita e uma pluralidade de produtos de ácido nucleico alvo-adaptador da segunda fita, confirmando a presença de pelo menos uma leitura de sequência amplificada de cada fita do complexo de material de ácido nucleico alvo-adaptador e comparando pelo menos uma leitura de sequência amplificada obtida da primeira fita com a pelo menos uma leitura de sequência amplificada obtida da segunda fita para formar uma leitura de sequência de consenso do material de ácido nucleico (por exemplo, um molécula de DNA alvo de fita dupla) tendo apenas bases nucleotídicas nas quais a sequência de ambas as fitas do material de ácido nucleico (por exemplo, uma molécula de DNA alvo de fita dupla) está de acordo, de modo que uma variante que ocorre em uma posição específica na leitura de sequência de consenso (por exemplo, em comparação com uma sequência de referência) é identificada como uma verdadeira variante de DNA.

[00111] Em algumas modalidades, são aqui fornecidos métodos para gerar uma sequência de consenso de alta precisão a partir de um material de ácido nucleico de fita dupla, incluindo as etapas de marcação de moléculas de DNA duplex individuais com uma molécula adaptadora para formar material de DNA marcado, em que cada molécula adaptadora compreende (a) um identificador de molécula única degenerada ou semidegenerada (SMI) que marca exclusivamente a molécula de DNA duplex; e (b) a primeira e a segunda sequências adaptadoras de nucleotídeos não complementares que distinguem uma fita superior original de uma fita inferior original de cada molécula de DNA individual dentro do material de DNA marcado, para cada molécula de DNA marcada, e gerando um conjunto de duplicatas da fita superior original da molécula de DNA marcada e um conjunto de duplicatas da fita inferior original da molécula de DNA marcada para formar material de DNA amplificado. O método pode ainda incluir as etapas de criar uma primeira sequência de consenso de fita única (SSCS) a partir das duplicatas da fita superior original e uma segunda sequência de consenso de fita única (SSCS) a partir das duplicatas da fita inferior original, comparando o primeiro SSCS da fita superior original com o segundo SSCS da fita inferior original e gerar uma sequência de consenso de alta precisão com apenas

bases nucleotídicas nas quais a sequência do primeiro SSCS da fita superior original e do segundo SSCS da fita inferior original é complementar.

[00112] Em outras modalidades, são aqui fornecidos métodos para detectar e/ou quantificar danos ao DNA de uma amostra compreendendo moléculas de DNA alvo de fita dupla, incluindo as etapas de ligação de ambas as fitas de cada molécula de DNA alvo de fita dupla a pelo menos uma molécula adaptadora assimétrica para formar uma pluralidade de complexos de DNA alvo-adaptador, em que cada complexo de DNA alvo-adaptador tem uma primeira sequência nucleotídica associada a uma primeira fita de uma molécula de DNA de fita dupla e uma segunda sequência nucleotídica que é pelo menos parcialmente não complementar à primeira sequência nucleotídica associada a uma segunda fita da molécula de DNA alvo de fita dupla e para cada complexo de DNA alvo-adaptador: amplificar cada fita do complexo de DNA alvo-adaptador, resultando em cada fita gerando um conjunto distinto, mas relacionado, de amplicons de DNA alvo-adaptador amplificados. O método pode ainda incluir as etapas de sequenciar cada uma de uma pluralidade de amplicons de DNA alvo-adaptador da primeira fita e uma pluralidade de amplicons de DNA alvo-adaptador da segunda fita, confirmando a presença de pelo menos uma leitura de sequência de cada fita do complexo de DNA alvo-adaptador e comparar pelo menos uma leitura de sequência obtida da primeira fita com a pelo menos uma leitura de sequência obtida da segunda fita para detectar e/ou quantificar bases nucleotídicas nas quais a leitura de sequência de uma fita da molécula de DNA de fita dupla está em desacordo (por exemplo, não complementar) com a leitura de sequência da outra fita da molécula de DNA de fita dupla, de modo que os sítios de dano ao DNA possam ser detectados e/ou quantificados. Em algumas modalidades, o método pode incluir ainda as etapas de criar uma primeira sequência de consenso de fita única (SSCS) a partir dos amplicons de DNA alvo-adaptador da primeira fita e uma segunda sequência de consenso de fita única (SSCS) a partir dos amplicons de DNA alvo-adaptador da segunda fita, comparar o primeiro SSCS da primeira fita original com o segundo SSCS da segunda fita original e identificar as bases nucleotídicas nas quais a sequência do primeiro SSCS e do segundo SSCS não é complementar para detectar e/ou quantificar danos ao DNA associados às moléculas de DNA alvo de fita dupla na amostra.

Sequências de identificador de molécula única (SMIs)

[00113] De acordo com várias modalidades, os métodos e composições fornecidos incluem uma ou mais sequências SMI em cada fita de um material de ácido nucleico. O SMI pode ser transportado independentemente por cada uma das fitas únicas que resultam de uma molécula de ácido nucleico de fita dupla, de modo que os produtos de amplificação derivativos de cada fita possam ser reconhecidos como provenientes da mesma molécula original de ácido nucleico de fita dupla substancialmente única após o sequenciamento. Em algumas modalidades, o SMI pode incluir informações adicionais e/ou pode ser usado em outros métodos para os quais essa funcionalidade de diferenciação de molécula é útil, como será reconhecido por um versado na técnica. Em algumas modalidades, um elemento SMI pode ser incorporado antes, substancialmente simultaneamente, ou após a ligação da sequência adaptadora um material de ácido nucleico.

[00114] Em algumas modalidades, uma sequência SMI pode incluir pelo menos um ácido nucleico degenerado ou semidegenerado. Em outras modalidades, uma sequência SMI pode ser não degenerada. Em algumas modalidades, o SMI pode ser a sequência associada a ou próximo a uma extremidade do fragmento da molécula de ácido nucleico (por exemplo, extremidades cortadas aleatoriamente ou semialeatoriamente do material de ácido nucleico ligado). Em algumas modalidades, uma sequência exógena pode ser considerada em conjunto com a sequência correspondente às extremidades cortadas aleatoriamente ou semialeatoriamente do material de ácido nucleico ligado (por exemplo, DNA) para obter uma sequência SMI capaz de distinguir, por exemplo, moléculas de DNA únicas uma da outra. Em algumas modalidades, uma sequência SMI é uma porção de uma sequência adaptadora que está ligada a uma molécula de ácido nucleico de fita dupla. Em certas modalidades, a sequência adaptadora compreendendo uma sequência SMI é de fita dupla, de modo que cada fita da molécula de ácido nucleico de fita dupla inclui um SMI após a ligação à sequência adaptadora. Em outra modalidade, a sequência SMI é de fita única antes ou após a ligação a uma molécula de ácido nucleico de fita dupla e uma sequência SMI complementar pode ser gerada estendendo a fita oposta com uma DNA polimerase para produzir uma sequência SMI de fita dupla complementar. Em outras modalidades, uma sequência SMI está em uma porção de fita única do adaptador (por exemplo, um braço de um adaptador tendo uma forma em Y). Em tais modalidades, o SMI pode facilitar o agrupamento de famílias de leituras de sequência derivadas de uma fita original de uma molécula de ácido nucleico de fita dupla e, em

alguns casos, pode conferir relação entre a primeira e a segunda fita originais de uma molécula de ácido nucleico de fita dupla (por exemplo, a totalidade ou parte dos SMIs pode ser relacionada através da tabela de consulta). Nas modalidades, onde a primeira e a segunda fitas são marcadas com SMIs diferentes, a leitura das sequências das duas fitas originais pode ser relacionada usando um ou mais SMI endógenos (por exemplo, um recurso específico de fragmento, como a sequência associada a ou próximo a uma extremidade do fragmento da molécula de ácido nucleico) ou com o uso de uma etiqueta molecular adicional compartilhada pelas duas fitas originais (por exemplo, um código de barras em uma porção de fita dupla do adaptador, ou uma combinação dos mesmos. Em algumas modalidades, cada sequência SMI pode incluir entre cerca de 1 a cerca de 30 ácidos nucleicos (por exemplo, 1, 2, 3, 4, 5, 8, 10, 12, 14, 16, 18, 20 ou mais ácidos nucleicos degenerados ou semidegenerados).

[00115] Em algumas modalidades, um SMI é capaz de ser ligado a um ou ambos de um material de ácido nucleico e uma sequência adaptadora. Em algumas modalidades, um SMI pode ser ligado a pelo menos um dentre um excesso de T, um excesso de A, um excesso de CG, uma base desidroxilada e uma extremidade cega de um material de ácido nucleico.

[00116] Em algumas modalidades, uma sequência de um SMI pode ser considerada em conjunto com (ou projetada de acordo com) a sequência correspondente a, por exemplo, extremidades cortadas aleatoriamente ou semialeatoriamente de um material de ácido nucleico (por exemplo, um material de ácido nucleico ligado), para obter uma sequência SMI capaz de distinguir moléculas de ácido nucleico únicas uma da outra.

[00117] Em algumas modalidades, pelo menos um SMI pode ser um SMI endógeno (por exemplo, um SMI relacionado a um ponto de cisalhamento (por exemplo, uma extremidade de fragmento), por exemplo, usando o próprio ponto de cisalhamento ou usando um número definido de nucleotídeos no ácido nucleico material imediatamente adjacente ao ponto de cisalhamento [por exemplo, 2, 3, 4, 5, 6, 7, 8, 9, 10 nucleotídeos do ponto de cisalhamento]). Em algumas modalidades, pelo menos um SMI pode ser um SMI exógeno (por exemplo, um SMI compreendendo uma sequência que não é encontrada em um material de ácido nucleico alvo).

[00118] Em algumas modalidades, um SMI pode ser ou compreender uma fração de imageamento (por exemplo, uma fração fluorescente ou de outra forma detectável opticamente).

Em algumas modalidades, tais SMIs permitem a detecção e/ou quantificação sem a necessidade de uma etapa de amplificação.

[00119] Em algumas modalidades, um elemento SMI pode compreender dois ou mais elementos SMI distintos que estão localizados em sítios diferentes no complexo de ácido nucleico adaptador-alvo.

[00120] Várias modalidades de SMIs são divulgadas ainda na Publicação Internacional de Patente WO2017/100441, que é incorporada por referência aqui na sua totalidade.

Elemento de definição de fita (SDE)

[00121] Em algumas modalidades, cada fita de um material de ácido nucleico de fita dupla pode ainda incluir um elemento que processa os produtos de amplificação dos dois ácidos nucleicos de fita única que formam o material de ácido nucleico de fita dupla alvo substancialmente distinguível um do outro após o sequenciamento. Em algumas modalidades, um SDE pode ser ou compreender sítios primários assimétricos compreendidos em um adaptador de sequenciamento ou, em outros arranjos, assimetrias de sequência podem ser introduzidas nas sequências adaptadoras e não nas sequências iniciadoras, de modo que pelo menos uma posição nas sequências nucleotídicas de um complexo de sequência de ácido nucleico alvo da primeira fita e de uma segunda fita do complexo de sequência de ácido nucleico alvo são diferentes um do outro após amplificação e sequenciamento. Em outras modalidades, o SDE pode compreender outra assimetria bioquímica entre as duas fitas que diferem das sequências nucleotídicas canônicas A, T, C, G ou U, mas é convertida em pelo menos uma diferença de sequência nucleotídica canônica nas duas moléculas amplificadas e sequenciadas. Em ainda outra modalidade, o SDE pode ser ou compreender um meio de separar fisicamente as duas fitas antes da amplificação, de modo que produtos de amplificação derivados da sequência de ácido nucleico alvo da primeira fita e a sequência de ácido nucleico alvo da segunda fita sejam mantidos em isolamento físico substancial um do outro com o objetivo de manter uma distinção entre os dois produtos de amplificação derivados. Podem ser utilizados outros arranjos ou metodologias para fornecer uma função SDE que permita distinguir a primeira e a segunda fitas.

[00122] Em algumas modalidades, uma SDE pode ser capaz de formar um laço (por exemplo, uma alça em hairpin). Em algumas modalidades, uma alça pode compreender pelo menos um sítio de reconhecimento de endonucleases. Em algumas modalidades, o complexo de ácido

nucleico alvo pode conter um sítio de reconhecimento de endonucleases que facilita um evento de clivagem dentro da alça. Em algumas modalidades, uma alça pode compreender uma sequência nucleotídica não canônica. Em algumas modalidades, o nucleotídeo não canônico contido pode ser reconhecível por uma ou mais enzimas que facilitam a clivagem da fita. Em algumas modalidades, o nucleotídeo não canônico contido pode ser direcionado por um ou mais processos químicos que facilitam a clivagem da fita na alça. Em algumas modalidades, a alça pode conter um ligante de ácido nucleico modificado que pode ser direcionado por um ou mais processos enzimáticos, químicos ou físicos que facilitam a clivagem de fios na alça. Em algumas modalidades, este ligante modificado é um ligante fotoclivável.

[00123] Uma variedade de outras ferramentas moleculares poderia servir como SMIs e SDEs. Além de pontos de cisalhamento e marcadores baseados em DNA, os métodos de compartimentação de molécula única que mantêm as fitas pareadas na proximidade física ou outros métodos de marcação de ácido não nucleico podem servir à função de relacionamento das fitas. Da mesma forma, a marcação química assimétrica das fitas do adaptador de forma que eles possam ser separados fisicamente pode desempenhar um papel de SDE. Uma variação recentemente descrita do Sequenciamento Duplex utiliza a conversão de bissulfito para transformar as assimetrias de fita de ocorrência natural na forma de metilação de citosina em diferenças de sequência que distinguem as duas fitas. Embora essa implementação limite os tipos de mutações que podem ser detectadas, o conceito de capitalização na assimetria nativa é digno de nota no contexto de tecnologias emergentes de sequenciamento que podem detectar diretamente nucleotídeos modificados. Várias modalidades de SDEs são divulgadas ainda na Publicação Internacional de Patente WO2017/100441, que é incorporada por referência na sua totalidade.

Adaptadores e sequências adaptadoras

[00124] Em várias disposições, moléculas adaptadoras que compreendem SMIs (por exemplo, códigos de barras moleculares), SDEs, sítios de iniciação, sequências de células de fluxo e/ou outros recursos são contempladas para uso em muitas das modalidades aqui divulgadas. Em algumas modalidades, os adaptadores fornecidos podem ser ou compreender uma ou mais sequências complementares ou pelo menos parcialmente complementares aos iniciadores de PCR (por exemplo, sítios de iniciação) que possuem pelo menos uma das seguintes propriedades: 1)

alta especificidade de alvo; 2) capaz de ser multiplexado; e 3) exibem amplificação robusta e minimamente tendenciosa.

[00125] Em algumas modalidades, as moléculas adaptadoras podem ter a forma de “Y”, a forma de “U”, a forma de “hairpin”, ter uma bolha (por exemplo, uma porção da sequência que não é complementar) ou outros recursos. Em outras modalidades, as moléculas adaptadoras podem compreender uma forma de "Y", uma forma de "U", uma forma de "hairpin" ou uma bolha. Certos adaptadores podem compreender nucleotídeos modificados ou não padronizados, sítios de restrição ou outros recursos para manipulação da estrutura ou função *in vitro*. As moléculas adaptadoras podem se ligar a uma variedade de material de ácido nucleico com uma extremidade terminal. Por exemplo, as moléculas adaptadoras podem ser adaptadas para se ligarem a um excesso de T, um excesso de A, um excesso de CG, um excesso de múltiplos nucleotídeos, uma base desidroxilada, uma extremidade cega de um material de ácido nucleico e a extremidade de uma molécula onde o 5' do alvo é desfosforilado ou de outro modo bloqueado da ligação tradicional. Em outras modalidades, a molécula adaptadora pode conter uma modificação desfosforilada ou de outra forma impeditora de ligação na fita 5' no sítio da ligação. Nas duas últimas modalidades, tais estratégias podem ser úteis para prevenir a dimerização de fragmentos da biblioteca ou moléculas adaptadoras.

[00126] Uma sequência adaptadora pode significar uma sequência de fita única, uma sequência de fita dupla, uma sequência complementar, uma sequência não complementar, uma sequência parcial complementar, uma sequência assimétrica, uma sequência de ligação ao iniciador, uma sequência de células de fluxo, uma sequência de ligação ou outra sequência fornecida por uma molécula adaptadora. Em modalidades particulares, uma sequência adaptadora pode significar uma sequência usada para amplificação por meio de complemento a um oligonucleotídeo.

[00127] Em algumas modalidades, os métodos e composições fornecidos incluem pelo menos uma sequência adaptadora (por exemplo, duas sequências adaptadoras, uma em cada uma das extremidades 5' e 3' de um material de ácido nucleico). Em algumas modalidades, os métodos e composições fornecidos podem compreender 2 ou mais sequências adaptadoras (por exemplo, 3, 4, 5, 6, 7, 8, 9, 10 ou mais). Em algumas modalidades, pelo menos duas das sequências adaptadoras diferem uma da outra (por exemplo, por sequência). Em algumas modalidades, cada

sequência adaptadora difere uma da outra sequência adaptadora (por exemplo, por sequência). Em algumas modalidades, pelo menos uma sequência adaptadora é pelo menos parcialmente não complementar a pelo menos uma porção de pelo menos uma outra sequência adaptadora (por exemplo, não é complementar por pelo menos um nucleotídeo).

[00128] Em algumas modalidades, uma sequência adaptadora compreende pelo menos um nucleotídeo não padrão. Em algumas modalidades, um nucleotídeo não padrão é selecionado de um sítio abásico, um uracil, tetra-hidrofurano, 8-oxo-7,8-di-hidro-2'-deoxiadenosina (8-oxo-A), 8-oxo-7,8-di-hidro-2'-desoxiguananosina (8-oxo-G), desoxinossina, 5-nitroindol, 5-hidroximetil-2'-desoxicidina, iso-citosina, 5'-metil-isocitosina ou isoguanosina, um nucleotídeo metilado, um RNA nucleotídeo, um nucleotídeo ribose, um 8-oxo-guanina, um ligante fotoclivável, um nucleotídeo biotinilado, um nucleotídeo de destiobiotina, um nucleotídeo modificado por tiol, um nucleotídeo modificado por acridito um iso-dC, um iso dG, um nucleotídeo 2'-O-metil, um Ácido Nucleico Bloqueado de nucleotídeo de inosina, um ácido nucleico de peptídeo, um 5 metil dC, um 5-bromo desoxiuridina, um nucleotídeo de 2,6-Diaminopurina, 2-Aminopurina, um nucleotídeo abásico, um nucleotídeo de 5-Nitroindol, um nucleotídeo adenilado, um nucleotídeo de azida, um nucleotídeo de digoxigenina, um ligante I, um nucleotídeo modificado em 5' Hexinil, um 5-Octadi-inil dU, espaçador foto-clivável, um espaçador não fotoclivável, um nucleotídeo modificado compatível com química de cliques e qualquer combinação dos mesmos.

[00129] Em algumas modalidades, uma sequência adaptadora compreende uma fração com uma propriedade magnética (isto é, uma fração magnética). Em algumas modalidades, essa propriedade magnética é paramagnética. Em algumas modalidades em que uma sequência adaptadora compreende uma fração magnética (por exemplo, um material de ácido nucleico ligado a uma sequência adaptadora compreendendo uma fração magnética), quando um campo magnético é aplicado, uma sequência adaptadora compreendendo uma fração magnética é substancialmente separada das sequências adaptadoras que não compreendem uma fração magnética (por exemplo, um material de ácido nucleico ligado a uma sequência adaptadora que não compreende uma fração magnética).

[00130] Em algumas modalidades, pelo menos uma sequência adaptadora está localizada 5' para um SMI. Em algumas modalidades, pelo menos uma sequência adaptadora está localizada 3' para um SMI.

[00131] Em algumas modalidades, uma sequência adaptadora pode ser ligada a pelo menos um dentre um SMI e um material de ácido nucleico por meio de um ou mais domínios ligantes. Em algumas modalidades, um domínio ligante pode ser constituído por nucleotídeos. Em algumas modalidades, um domínio ligante pode incluir pelo menos uma molécula nucleotídica ou não nucleotídica modificada (por exemplo, como descrito em outra parte desta divulgação). Em algumas modalidades, um domínio ligante pode ser ou compreender uma alça.

[00132] Em algumas modalidades, uma sequência adaptadora em uma ou em ambas as extremidades de cada fita de um material de ácido nucleico de fita dupla pode incluir ainda um ou mais elementos que fornecem um SDE. Em algumas modalidades, uma SDE pode ser ou compreender sítios primários assimétricos compreendidos nas sequências adaptadoras.

[00133] Em algumas modalidades, uma sequência adaptadora pode ser ou compreender pelo menos um SDE e pelo menos um domínio de ligação (isto é, um domínio alterável à atividade de pelo menos uma ligase, por exemplo, um domínio adequado para a ligação a um material de ácido nucleico através da atividade de uma ligase). Em algumas modalidades, de 5' a 3', uma sequência adaptadora pode ser ou compreender um sítio de ligação ao iniciador, um SDE e um domínio de ligação.

[00134] Vários métodos para sintetizar adaptadores de Sequenciamento Duplex foram descritos anteriormente em, por exemplo, Patente U.S. 9.752.188, Publicação Internacional de Patente WO2017/100441 e Pedido de Patente Internacional PCT/US18/59908 (depositado em 8 de novembro de 2018), todos incorporados por referência aqui na íntegra.

Iniciadores

[00135] Em algumas modalidades, um ou mais iniciadores de PCR que possuem pelo menos uma das seguintes propriedades: 1) alta especificidade de alvo; 2) capaz de ser multiplexado; e 3) exibir amplificação robusta e minimamente tendenciosa são contemplados para uso em várias modalidades, de acordo com aspectos da presente tecnologia. Vários estudos anteriores e produtos comerciais projetaram misturas de iniciadores que atendem a alguns desses critérios para a PCR-CE convencional. No entanto, observou-se que essas misturas de iniciadores nem sempre são

ideais para uso com MPS. De fato, o desenvolvimento de misturas de iniciadores altamente multiplexados pode ser um processo desafiador e demorado. Convenientemente, a Illumina e a Promega desenvolveram recentemente misturas de iniciadores compatíveis com multiplex para a plataforma Illumina, que mostram amplificação robusta e eficiente de uma variedade de loci STR e SNP padrão e não padrão. Uma vez que estes kits utilizam PCR para amplificar as suas regiões alvo antes do sequenciamento, a extremidade 5' de cada leitura nos dados de sequenciamento na extremidade pareada corresponde à extremidade 5' dos iniciadores de PCR utilizados para amplificar o DNA. Em algumas modalidades, os métodos e composições fornecidos incluem iniciadores projetados para garantir amplificação uniforme, o que pode implicar concentrações variáveis de reação, temperaturas de fusão e minimização da estrutura secundária e interações intra/inter-iniciador. Muitas técnicas foram descritas para otimização de iniciadores altamente multiplexados para aplicações MPS. Em particular, essas técnicas são frequentemente conhecidas como métodos de amplificação, bem como descritas na técnica.

Amplificação

[00136] Os métodos e composições fornecidos, em várias modalidades, fazem uso de, ou são úteis em, pelo menos uma etapa de amplificação, em que um material de ácido nucleico (ou uma porção do mesmo, por exemplo, uma região ou locis alvo específico) é amplificado para formar um material de ácido nucleico amplificado (por exemplo, algum número de produtos de amplicons).

[00137] Em algumas modalidades, a amplificação de um material de ácido nucleico inclui uma etapa de amplificação do material de ácido nucleico derivado de cada um de uma primeira e uma segunda fita de ácido nucleico de um material original de ácido nucleico de fita dupla usando pelo menos um oligonucleotídeo de fita única pelo menos parcialmente complementar a uma sequência presente em uma primeira sequência adaptadora, de modo que uma sequência SMI seja pelo menos parcialmente mantida. Uma etapa de amplificação inclui ainda empregar um segundo oligonucleotídeo de fita única para amplificar cada fita de interesse, e esse segundo oligonucleotídeo de fita única pode ser (a) pelo menos parcialmente complementar a uma sequência de interesse alvo, ou (b) pelo menos parcialmente complementar a uma sequência presente em uma segunda sequência adaptadora de modo que o pelo menos um oligonucleotídeo

de fita única e um segundo oligonucleotídeo de fita única sejam orientados de maneira a amplificar efetivamente o material de ácido nucleico.

[00138] Em algumas modalidades, a amplificação do material de ácido nucleico em uma amostra pode incluir amplificar o material de ácido nucleico em "tubos" (por exemplo, tubos de PCR), em gotículas de emulsão, microcâmaras e outros exemplos descritos acima ou em outros recipientes conhecidos.

[00139] Em algumas modalidades, pelo menos uma etapa de amplificação inclui pelo menos um iniciador que é ou compreende pelo menos um nucleotídeo não padrão. Em algumas modalidades, um nucleotídeo não padrão é selecionado de um uracil, um nucleotídeo metilado, um nucleotídeo de RNA, um nucleotídeo ribose, uma 8-oxo-guanina, um nucleotídeo biotinilado, um ácido nucleico bloqueado, um ácido nucleico peptídico, uma variante de ácido nucleico de alta T_m, uma variante de ácido nucleico discriminador de alelo, qualquer outra variante nucleotídica ou de ligante descrita em outra parte deste documento e qualquer combinação dos mesmos.

[00140] Embora qualquer reação de amplificação apropriada para aplicação seja contemplada como compatível com algumas modalidades, a título de exemplo específico, em algumas modalidades, uma etapa de amplificação pode ser ou compreender uma reação em cadeia da polimerase (PCR), amplificação de círculo rolante (RCA), amplificação de deslocamento múltiplo (MDA), amplificação isotérmica, amplificação de polônio dentro de uma emulsão, amplificação de ponte em uma superfície, a superfície de uma esfera ou dentro de um hidrogel e qualquer combinação das mesmas.

[00141] Em algumas modalidades, a amplificação de um material de ácido nucleico inclui utilizar oligonucleotídeos de fita simples, pelo menos parcialmente complementares às regiões das sequências adaptadoras nas extremidades 5' e 3' de cada fita do material de ácido nucleico. Em algumas modalidades, a amplificação de um material de ácido nucleico inclui utilizar pelo menos um oligonucleotídeo de fita simples, pelo menos parcialmente complementar a uma região alvo ou uma sequência alvo de interesse (por exemplo, uma sequência genômica, uma sequência mitocondrial, uma sequência plasmídica, um ácido nucleico alvo produzido sinteticamente, etc.) e um oligonucleotídeo de fita única, pelo menos parcialmente complementar a uma região da sequência adaptadora (por exemplo, um sítio de iniciação).

[00142] Em geral, a amplificação robusta, por exemplo, a amplificação por PCR, pode ser altamente dependente das condições da reação. A PCR multiplex, por exemplo, pode ser sensível à composição do tampão, concentração de cátion monovalente ou divalente, concentração de detergente, concentração de agente de aglomeração (por exemplo, PEG, glicerol, etc.), concentração de iniciador, Tms do iniciador, design do iniciador, teor de GC do iniciador, propriedades nucleotídicas modificadas por iniciador e condições de ciclagem (*isto é*, tempos de temperatura e extensão e taxa de alterações de temperatura). A otimização das condições de tampão pode ser um processo difícil e demorado. Em algumas modalidades, uma reação de amplificação pode usar pelo menos um de um tampão, concentração do conjunto de iniciadores e condições de PCR de acordo com um protocolo de amplificação conhecido anteriormente. Em algumas modalidades, um novo protocolo de amplificação pode ser criado e/ou uma otimização da reação de amplificação pode ser usada. A título de exemplo específico, em algumas modalidades, um kit de otimização de PCR pode ser usado, como um Kit de Otimização de PCR da Promega[®], que contém vários tampões pré-formulados que são parcialmente otimizados para uma variedade de aplicações de PCR, como amplificações multiplex, em tempo real, ricas em GC e resistentes a inibidores. Esses tampões pré-formulados podem ser rapidamente suplementados com diferentes concentrações de Mg^{2+} e iniciadores, bem como razões de grupo de iniciadores. Além disso, em algumas modalidades, uma variedade de condições de ciclagem (por exemplo, ciclagem térmica) pode ser avaliada e/ou usada. Ao avaliar se uma modalidade específica é apropriada ou não para uma aplicação desejada específica, podem ser avaliadas uma ou mais especificidades, a razão de cobertura de alelos para loci heterozigotos, equilíbrio entre focos e profundidade, entre outros aspectos. As medições do sucesso da amplificação podem incluir sequenciamento de DNA dos produtos, avaliação de produtos por eletroforese em gel ou capilar ou HPLC ou outros métodos de separação por tamanho, seguidos de visualização de fragmentos, análise da curva de fusão usando corantes de ligação de ácido nucleico de fita dupla ou sondas fluorescentes, espectrometria de massa ou outros métodos conhecidos na técnica.

[00143] De acordo com várias modalidades, qualquer um de uma variedade de fatores pode influenciar o comprimento de uma etapa de amplificação específica (por exemplo, o número de ciclos em uma reação de PCR, etc.). Por exemplo, em algumas modalidades, um material de ácido nucleico fornecido pode ser comprometido ou de outro modo subideal (por exemplo,

degradado e/ou contaminado). Nesse caso, uma etapa de amplificação mais longa pode ser útil para garantir que um produto desejado seja amplificado em um grau aceitável. Em algumas modalidades, uma etapa de amplificação pode fornecer uma média de 3 a 10 cópias de PCR sequenciadas de cada molécula de DNA inicial, embora em outras modalidades, apenas uma cópia de cada uma das primeiras fitas e da segunda fita seja necessária. Sem desejar se apegar a uma teoria específica, é possível que muitas ou poucas cópias de PCR possam resultar em eficiência de ensaio reduzida e, finalmente, profundidade reduzida. Geralmente, o número de fragmentos de ácido nucleico (por exemplo, DNA) usados em uma reação de amplificação (por exemplo, PCR) é uma variável ajustável primária que pode ditar o número de leituras que compartilham a mesma sequência SMI/código de barras.

Material de ácido nucleico

Tipos

[00144] De acordo com várias modalidades, qualquer um de uma variedade de material de ácido nucleico pode ser usado. Em algumas modalidades, o material de ácido nucleico pode compreender pelo menos uma modificação em um polinucleotídeo dentro da espinha dorsal canônica de açúcar-fosfato. Em algumas modalidades, o material de ácido nucleico pode compreender pelo menos uma modificação dentro de qualquer base no material de ácido nucleico. Por exemplo, a título de exemplo não limitativo, em algumas modalidades, o material de ácido nucleico é ou compreende pelo menos um de DNA de fita dupla, DNA de fita única, RNA de fita dupla, RNA de fita simples, ácidos nucleicos de peptídeo (PNAs), ácidos nucleicos bloqueados (LNAs).

Modificações

[00145] De acordo com várias modalidades, o material de ácido nucleico pode receber uma ou mais modificações antes, substancialmente simultaneamente ou subsequentemente a qualquer etapa específica, dependendo da aplicação para a qual um método ou composição fornecida específica é usada.

[00146] Em algumas modalidades, uma modificação pode ser ou compreender reparo de pelo menos uma porção do material de ácido nucleico. Embora qualquer maneira apropriada de aplicação de reparo de ácido nucleico seja contemplada como compatível com algumas

modalidades, certos métodos e composições exemplificativas são, portanto, descritos abaixo e nos Exemplos.

[00147] A título de exemplo não limitativo, em algumas modalidades, enzimas de reparo de DNA, como Uracil-DNA Glicosilase (UDG), Formamidopirimidina DNA glicosilase (FPG) e 8-oxoguanina DNA glicosilase (OGG1), podem ser utilizadas para corrigir danos ao DNA (por exemplo, danos ao DNA *in vitro*). Como discutido acima, essas enzimas de reparo do DNA, por exemplo, são glicosilases que removem bases danificadas do DNA. Por exemplo, o UDG remove o uracil resultante da desaminação da citosina (causada pela hidrólise espontânea da citosina) e o FPG remove a 8-oxo-guanina (por exemplo, a lesão de DNA mais comum resultante de espécies reativas de oxigênio). O FPG também possui atividade de liase que pode gerar folga de 1 base em sítios abásicos. Tais sítios abásico falham subsequentemente em amplificar por PCR, por exemplo, porque a polimerase falha em copiar o modelo. Consequentemente, o uso de tais enzimas de reparo de danos ao DNA pode remover efetivamente o DNA danificado que não possui uma mutação verdadeira, mas que pode ser detectado como um erro após a sequência e a análise de sequência duplex.

[00148] Como discutido acima, em outras modalidades, as leituras de sequenciamento geradas a partir das etapas de processamento discutidas aqui podem ser filtradas ainda mais para eliminar falsas mutações aparando as extremidades das leituras mais propensas a artefatos. Por exemplo, a fragmentação do DNA pode gerar porções de fita única nas extremidades terminais da molécula de fita dupla. Essas porções de fita única podem ser preenchidas (por exemplo, por Klenow) durante o reparo final. Em alguns casos, as polimerases cometem erros de cópia nessas regiões reparadas finais, levando à geração de "moléculas pseudoduplex". Esses artefatos podem parecer verdadeiras mutações uma vez sequenciadas. Esses erros, como resultado de mecanismos de reparo final, podem ser eliminados da análise pós-sequenciamento aparando as extremidades das leituras de sequenciamento para excluir quaisquer mutações que possam ter ocorrido, reduzindo assim o número de falsas mutações. Em algumas modalidades, esse corte de leituras de sequenciamento pode ser realizado automaticamente (por exemplo, uma etapa normal do processo). Em algumas modalidades, uma frequência mutante pode ser avaliada para regiões terminais do fragmento e se um nível limiar de mutações for observado nas regiões terminais do fragmento, o

corte de leitura de sequenciamento pode ser realizado antes de gerar uma leitura de sequência de consenso de fita dupla dos fragmentos de DNA.

[00149] O alto grau de correção de erros fornecido pela tecnologia de comparação de fitas de Sequenciamento Duplex reduz os erros de sequenciamento de moléculas de ácido nucleico de fita dupla em várias ordens de magnitude em comparação com os métodos padrão de sequenciamento de próxima geração. Esta redução nos erros melhora a precisão do sequenciamento em quase todos os tipos de sequências, mas pode ser particularmente adequada para sequências bioquimicamente desafiadoras que são bem conhecidas na técnica por serem particularmente propensas a erros. Um exemplo não limitativo desse tipo de sequência são os homopolímeros ou outros microssatélites/repetições em tandem curto. Outro exemplo não limitativo de sequências propensas a erros que se beneficiam da correção de erros de Sequenciamento Duplex são moléculas danificadas, por exemplo, por aquecimento, radiação, estresse mecânico ou uma variedade de exposições químicas que criam adutos químicos propensos a erros durante a cópia por uma ou mais polimerases nucleotídicas e também aquelas que criam DNA de fita única nas extremidades das moléculas ou como cortes e folgas. Em outras modalidades, o Sequenciamento Duplex também pode ser usado para a detecção precisa de variantes de sequência minoritária entre uma população de moléculas de ácido nucleico de fita dupla. Um exemplo não limitativo desta aplicação é a detecção de um pequeno número de moléculas de DNA derivadas de um câncer, dentre um número maior de moléculas não mutadas de tecidos não cancerígenos dentro de um sujeito. Outra aplicação não limitativa para a detecção de variantes raras pelo Sequenciamento Duplex é a detecção precoce de danos ao DNA resultantes da exposição à genotoxina. Uma outra aplicação não limitativa do Sequenciamento Duplex é para a detecção de mutações geradas a partir de carcinógenos genotóxicos ou não genotóxicos, observando clones genéticos que estão surgindo com mutações ativadoras. Uma aplicação ainda não limitativa adicional para a detecção precisa de variantes de sequências minoritárias é gerar uma assinatura mutagênica associada a uma genotoxina.

Identificação e avaliação da genotoxicidade

[00150] A presente tecnologia é direcionada a métodos, sistemas, kits, etc. para avaliar a genotoxicidade. Em particular, algumas modalidades da tecnologia são direcionadas à utilização

do Sequenciamento Duplex para avaliar um potencial genotóxico de um composto (por exemplo, um composto químico) ou outro agente em uma fonte biológica. Por exemplo, várias modalidades da presente tecnologia incluem a execução de métodos de Sequenciamento Duplex que permitem a medição direta de mutações induzidas por agente em qualquer contexto genômico de qualquer organismo, e sem necessidade de seleção clonal. Outros exemplos da presente tecnologia são direcionados a métodos para detectar e avaliar a mutagênese genômica *in vivo* usando o Sequenciamento Duplex. Vários aspectos da presente tecnologia têm muitas aplicações em testes de segurança de drogas pré-clínicas e clínicas, além de outras implicações em todo o setor. Por exemplo, a presente tecnologia inclui métodos para detectar mutações de frequência ultra baixa que causam o aparecimento de doenças/distúrbios anos depois, em que as mutações ocorrem como resultado direto da exposição a pelo menos uma genotoxina (por exemplo, radiação, carcinógeno) e/ou como resultado de fontes endógenas, como erros na DNA polimerase, radicais livres e depurinação. A detecção pode ocorrer através do teste de um sujeito após uma exposição recente a uma genotoxina (por exemplo, dias após a exposição) e usando o Sequenciamento Duplex para identificar as mutações de frequência ultrabaixa. Em exemplos particulares, as mutações de frequência ultrabaixa detectadas podem ser comparadas com mutações conhecidas por causar uma doença ou distúrbio específico, incluindo aquelas doenças/distúrbios que geralmente se manifestam após muitos anos após a exposição (por exemplo, câncer de pulmão 20 anos após a exposição ao amianto). A presente tecnologia fornece, assim, um método conveniente para identificar a presença de genotoxinas e vítimas expostas a elas, a fim de evitar exposições futuras e fornecer tratamento médico precoce. A presente tecnologia também pode ser usada em uma variedade de métodos de triagem de alto rendimento para identificar produtos de consumo inseguros, produtos farmacêuticos e outros subprodutos industriais/comerciais/de fabricação que compreendem genotoxinas para removê-los do mercado ou do meio ambiente.

[00151] Numa modalidade particular, efeitos genotóxicos, como deleções, quebras e/ou rearranjos, podem levar ao câncer ou outra doença associada genotóxica a distúrbio se o dano não levar imediatamente à morte celular. Por exemplo, o dano do ácido nucleico pode ser suficiente para o sujeito desenvolver uma doença ou distúrbio associado genotóxico e/ou pode contribuir para a ativação ou progressão de outro tipo de doença ou distúrbio já existente em um sujeito exposto. Regiões sensíveis à quebra, chamadas sítios frágeis, podem resultar de agentes

genotóxicos (por exemplo, produtos químicos, como pesticidas ou certas drogas quimioterápicas). Alguns produtos químicos têm a capacidade de induzir sítios frágeis nas regiões do cromossomo onde estão presentes oncogenes, o que pode levar a efeitos carcinogênicos. Além disso, a exposição ocupacional a algumas misturas de pesticidas, compostos de fabricação ou outros materiais perigosos está positivamente correlacionada com o aumento de danos genotóxicos nos indivíduos expostos. A investigação do potencial de genotoxicidade, por exemplo, antes da exposição humana, é altamente desejável para qualquer genotoxina em potencial, como uma droga potencial, cosmético, produto de consumo, produto industrial/fabricação ou subproduto ou outro composto químico em desenvolvimento. Da mesma forma, nas modalidades em que a exposição a uma genotoxina é suspeita, se as genotoxinas puderem ser identificadas, o sujeito pode receber tratamentos terapêuticos direcionados e/ou a genotoxina pode ser removida para impedir a exposição futura ao sujeito e a outros.

[00152] A capacidade de detectar efeitos genotóxicos de um potencial agente ou fator genotóxico e quantificar um processo mutagênico potencialmente resultante de uma maneira que seja viável em tempo e custo e de importância comercial e médica. Em um exemplo particular, a capacidade de detectar e quantificar processos mutagênicos de uma genotoxina em potencial pode ser importante para avaliar o risco de câncer, identificar carcinógenos e prever o impacto da exposição em humanos. No entanto, as ferramentas atuais são lentas, pesadas e/ou limitadas nas informações fornecidas. Como descrito acima, os testes *in vivo* e os sistemas de repórteres de mamíferos, como o rato e o camundongo BigBlue[®], são atualmente utilizados sob os regulamentos da Food and Drug Administration (FDA) como uma métrica válida de genotoxicidade para determinar o potencial de compostos de causar danos ao DNA.

[00153] A FIG. 2A é uma ilustração conceitual que mostra várias metodologias para avaliar a mutagênese *in vivo* de uma genotoxina potencial (por exemplo, um potencial mutagênico). Em cada um dos esquemas ilustrados na FIG. 2A, um sujeito de teste (por exemplo, camundongo BigBlue[®], um organismo modelo de camundongo, um organismo modelo de rato, etc.) é exposto à genotoxina em potencial (por exemplo, o composto/agente/fator sob investigação) usando uma via de administração apropriada. Num esquema convencional mostrado no lado esquerdo da FIG. 2A, um bioensaio de carcinogenicidade de roedores a longo prazo observa o animal de teste por um longo período (por exemplo, 2 anos) para o desenvolvimento de lesões neoplásicas durante ou

após a exposição a várias doses da substância de teste. Os animais de teste podem ser doseados por exposições orais, dérmicas ou por inalação, com base no tipo esperado de exposição humana, por exemplo. No esquema convencional, a dosagem geralmente dura cerca de dois anos; no entanto, os parâmetros de dosagem (por exemplo, duração da dose, via de administração, níveis de dosagem ou outros parâmetros do regime de dosagem) podem ser definidos de acordo com o protocolo de teste desejado. Com referência à FIG. 2A, esquema à esquerda, certas características de saúde animal são monitoradas ao longo do estudo, mas a avaliação principal reside na análise patológica completa dos tecidos e órgãos dos animais em teste quando o estudo é finalizado.

[00154] Outro ensaio *in vivo* mostrado no esquema do meio da FIG. 2A, utiliza um roedor transgênico. Após um regime de dosagem adequado a curto prazo (por exemplo, na ordem dos dias ou semanas), o animal de teste é sacrificado, os tecidos desejados são colhidos e o DNA é extraído. A partir do DNA extraído, os fragmentos transgênicos são isolados e os plasmídeos purificados resultantes são embalados em fagos e infectados com *E. coli*. É realizado um ensaio de placa transgênica convencional e uma frequência mutante básica é calculada.

[00155] Ambos os esquemas descritos acima são lentos e fornecem informações muito limitadas com relação à genotoxicidade (por exemplo, mutagênese) da genotoxina potencial testada. A possibilidade de medir diretamente mutações somáticas de uma maneira que não seja restringida pelo locus genômico, tecido ou organismo é atraente, mas atualmente é impossível com o sequenciamento de DNA padrão devido a uma taxa de erro ($\sim 10^{-3}$) bem acima da frequência mutante de tecidos normais ($\sim 10^{-7}$ a 10^{-8}).

[00156] O sequenciamento massivamente paralelo oferece a possibilidade de pesquisar exaustivamente o genoma de qualquer organismo quanto ao efeito *in vivo* das exposições mutagênicas; no entanto, conforme discutido, os métodos convencionais são muito imprecisos para detectar tais mutações, que podem ocorrer em um nível abaixo de um em um milhão. Por exemplo, a taxa de erro do sequenciamento de próxima geração (NGS) em aproximadamente 0,1% cria um ruído de fundo que obscurece a detecção de variantes raras e perfis ou assinaturas moleculares exclusivas. Algumas fontes comuns de erros nas plataformas NGS incluem enzimas PCR (que surgem durante a amplificação), leituras do sequenciador e danos ao DNA durante o processamento (por exemplo, 8-oxo-guanina, citosina desaminada, sítios abásicos e outros).

[00157] De acordo com aspectos da presente tecnologia, as etapas do método de Sequenciamento Duplex podem gerar leituras de sequenciamento de DNA de alta precisão que podem fornecer ainda mais frequência mutante detalhada (por exemplo, resolução de mutações induzidas por genotoxina abaixo de um em um milhão e fornecer dados de espectro de mutação para caracterizar objetivamente diferentes processos mutagênicos e inferir o mecanismo de ação). Por exemplo, o esquema do lado direito mostrado na FIG. 2A inclui um método para detectar e avaliar rapidamente a genotoxicidade de uma potencial genotoxina (por exemplo, potencial mutagênico) no mesmo sujeito de teste que os esquemas da técnica anterior, além de fornecer informações detalhadas sobre a frequência mutante, espectro do(s) tipo(s) de mutação e dados de contexto genômico. Além disso, a análise de Sequenciamento Duplex pode fornecer detecção sensível de mutagênese em qualquer locus genético em qualquer tecido de qualquer organismo. Por exemplo, e como ilustrado nas FIGS. 2A e 2B, os esquemas do método de Sequenciamento Duplex podem ser usados para avaliar a mutagênese *in vitro* de um composto de teste em células (por exemplo, células humanas, células de roedores, células de mamíferos, células de não mamíferos, etc.) cultivadas em cultura (FIG. 2B) e para avaliar a mutagênese *in vivo* de um composto de teste em um roedor do tipo selvagem (por exemplo, camundongo) (FIG. 2C). Por exemplo, em uma modalidade, a presente tecnologia inclui etapas do método, incluindo a exposição de um organismo de teste (por exemplo, um roedor, células cultivadas em cultura) a um composto de teste (por exemplo, genotoxina/mutagênico potencial) por uma via de administração apropriada (por exemplo, por via oral), subcutânea, tópica, aerossol, intramuscular, etc.). Em uma modalidade, o organismo de teste pode ser exposto ao composto de teste por um período curto (por exemplo, uma dose única, alguns minutos, algumas horas, menos de 24 horas, alguns dias, 2-6 dias, etc.) ou uma duração moderada (por exemplo, vários dias, 3-12 dias, aproximadamente 1 semana, aproximadamente 2 semanas, aproximadamente 1 mês, aproximadamente 2 meses, aproximadamente 3-6 meses etc.) ou outra quantidade de tempo adequada. Se o organismo de teste for um animal (por exemplo, roedor), como ilustrado na FIG. 1A (esquema da direita) e FIG. 1C, o animal pode então ser sacrificado e/ou tecidos desejados colhidos para extração de DNA. Por exemplo, em certas modalidades, o animal de teste não é sacrificado e uma ou mais amostras de sangue (por exemplo, no mesmo ou em diferentes momentos após a administração ou exposição a uma substância de teste) podem ser coletadas do animal de teste para extração de DNA. Nas

modalidades em que o animal é sacrificado, um ou mais tecidos de interesse (por exemplo, fígado, medula óssea, pulmão, baço, sangue, etc.) podem ser colhidos para extração de DNA. Se o organismo de teste compreender células em cultura (FIG. 1B), todas ou uma porção das células poderá ser coletada para extração de DNA.

[00158] Após a extração de DNA da amostra biológica coletada ou colhida, uma biblioteca de DNA (por exemplo, uma biblioteca de sequenciamento) pode ser preparada. Em uma modalidade, uma abordagem para preparar uma biblioteca de DNA (ou outra biblioteca de sequenciamento de ácido nucleico) pode começar com a marcação (por exemplo, marcação) de material de ácido nucleico de fita dupla fragmentada (por exemplo, a partir da amostra de DNA) com códigos de barras moleculares de maneira semelhante como descrito acima e em relação a um protocolo de construção de biblioteca de Sequenciamento Duplex (por exemplo, como ilustrado na FIG. 1A). Em algumas modalidades, o material de ácido nucleico de fita dupla pode ser fragmentado (por exemplo, como com DNA livre de células, DNA danificado, etc.); no entanto, em outras modalidades, várias etapas podem incluir a fragmentação do material de ácido nucleico usando cisalhamento mecânico como sonicação ou outros métodos de corte de DNA (por exemplo, digestão enzimática, nebulização, etc.). Aspectos da marcação do material de ácido nucleico de fita dupla fragmentada podem incluir reparo final e cauda de 3'-dA, se necessário em uma aplicação específica, seguido pela ligação dos fragmentos de ácido nucleico de fita dupla com adaptadores adequados de Sequenciamento Duplex contendo um SMI (por exemplo, como ilustrado na FIG. 1A). Em outras modalidades, o SMI pode ser endógeno ou uma combinação de sequência exógena e endógena para relacionar exclusivamente informações de ambas as fitas de uma molécula de ácido nucleico original.

[00159] Após a ligação das moléculas adaptadoras ao material de ácido nucleico de fita dupla, o método pode continuar com a amplificação (por exemplo, amplificação por PCR, amplificação de círculo rolante, amplificação de deslocamento múltiplo, amplificação isotérmica, amplificação em ponte, amplificação ligada à superfície, etc.) (FIG. 1B). Em certas modalidades, os iniciadores específicos para, por exemplo, uma ou mais sequências adaptadoras, podem ser usados para amplificar cada fita do material de ácido nucleico, resultando em múltiplas cópias de amplicons de ácido nucleico derivadas de cada fita de uma molécula original de ácido nucleico de fita dupla, com cada amplicon mantendo o SMI originalmente associado (FIG. 1B). Após a

amplificação e as etapas associadas para remover os subprodutos da reação, a(s) região(ões) de ácido(s) nucleico(s) (por exemplo, regiões de interesse, loci etc.) pode ser opcionalmente enriquecida usando a captura direcionada baseada em hibridação ou em outra modalidade, com PCR multiplex usando iniciador(es) específico(s) para uma sequência adaptadora e iniciador(es) específico(s) da(s) região(ões) de ácido(s) nucleico(s) alvo de interesse (não mostrado).

[00160] Após as etapas de preparação e amplificação da biblioteca de DNA, os complexos de DNA-adaptador de fita dupla podem ser sequenciados com uma plataforma de sequenciamento de DNA massivamente paralela apropriada, usando métodos de sequenciamento padrão (FIG. 1B). Após o sequenciamento das múltiplas cópias da primeira fita, as múltiplas cópias da segunda fita, os dados de sequenciamento podem ser analisados usando uma abordagem de Sequenciamento Duplex e conforme descrito aqui, em que as leituras de sequências compartilhando o mesmo SMI exógeno (por exemplo, sequência adaptadora) e/ou endógeno derivadas da primeira ou da segunda fita da molécula original de ácido nucleico alvo de fita dupla são agrupadas separadamente. Em algumas modalidades, as leituras de sequenciamento agrupadas da primeira fita (por exemplo, "fita superior") são usadas para formar uma sequência de consenso da primeira fita (por exemplo, uma sequência de consenso de fita única (SSCS)) e as leituras de sequenciamento agrupadas da segunda fita (por exemplo, "fita inferior") é usada para formar uma sequência de consenso da segunda fita (por exemplo, SSCS). Voltando à FIG. 1C, o primeiro e o segundo SSCSs podem então ser comparados para gerar uma sequência de consenso duplex (DCS) com nucleotídeos que estão de acordo entre as duas fitas (por exemplo, variantes ou mutações são consideradas verdadeiras se aparecerem em leituras de sequenciamento derivadas de ambas as fitas) (ver, por exemplo, FIG. 1C). Da mesma forma, na etapa de comparação, as posições do DCS em que os nucleotídeos não estão de acordo entre as duas fitas também podem ser avaliadas como sítios potenciais de dano ao DNA, como dano causado pela exposição à genotoxina.

[00161] Voltando às FIGS. 2A-2C, e de acordo com aspectos da presente tecnologia, a análise de Sequenciamento Duplex pode ainda ser usada para quantificar com precisão a frequência de mutações induzidas no genoma. Por exemplo, aspectos da presente tecnologia são direcionados à geração de informações associadas à genotoxicidade capturadas nos dados da sequência derivada, incluindo, por exemplo, espectro de mutações, assinaturas mutacionais de trinucleotídeos, informações sobre as consequências funcionais de certas mutações na proliferação

e seleção neoplásica, comparação com informação associada à genotoxicidade derivada empiricamente relacionada a genotoxinas conhecidas (por exemplo, espectros de mutação, assinaturas mutacionais de trinucleotídeos) e semelhantes.

[00162] A presente tecnologia compreende ainda um método para detectar pelo menos uma mutação genômica em um sujeito como resultado da exposição a uma genotoxina, compreendendo as etapas de: 1) fornecer uma amostra de um sujeito após a exposição à genotoxina, em que a amostra compreende uma pluralidade de moléculas de DNA de fita dupla; 2) ligar moléculas adaptadoras assimétricas a moléculas de DNA de fita dupla individuais para gerar uma pluralidade de moléculas de DNA-adaptador; 3) para cada molécula de DNA-adaptador: (i) gerar um conjunto de cópias de uma primeira fita original da molécula de DNA-adaptador e um conjunto de cópias de uma segunda fita original da molécula de DNA-adaptador; (ii) sequenciar o conjunto de cópias da primeira e da segunda fitas originais para fornecer uma sequência da primeira fita e uma sequência da segunda fita; e (iii) comparar a sequência da primeira fita e a sequência da segunda fita para identificar uma ou mais correspondências entre as sequências da primeira e da segunda fitas; e 4) analisar as uma ou mais correspondências em cada uma das moléculas de DNA-adaptador para determinar pelo menos uma de uma frequência mutante e um espectro de mutação indicativo de uma genotoxina específica, uma classe de genotoxina e/ou um mecanismo de ação. Em algumas modalidades, o espectro de mutação é um espectro de mutação triplete. Em outras modalidades, a análise das uma ou mais correspondências em cada uma das moléculas de DNA-adaptador para determinar um espectro de mutação triplete compreende ainda gerar uma assinatura de mutação triplete para a genotoxina específica. Em certas modalidades, determinar uma frequência mutante compreende determinar uma frequência de um contexto de triplete/trinucleotídeo da base que é mutada.

[00163] Em algumas modalidades, a assinatura de mutação triplete e/ou o espectro de mutação são comparados com informações associadas à genotoxina derivada empiricamente para determinar (por exemplo, com base em semelhanças e/ou diferenças) um tipo de genotoxina a qual o sujeito foi exposto (se não for conhecido), o mecanismo de ação da genotoxina, uma probabilidade de o sujeito desenvolver uma doença ou distúrbio associado à genotoxina e/ou outras informações associadas à genotoxina. Por exemplo, um padrão de espectro de trinucleotídeo de Sequenciamento Duplex resultante de uma exposição conhecida ou suspeita de genotoxina (por

exemplo, a genotoxina de teste) em um sujeito pode ser comparado a padrões de espectro de trinucleotídeo derivados empiricamente associados à exposição a outras genotoxinas conhecidas (por exemplo, armazenadas em um banco de dados). Em certas modalidades, o padrão de espectro de trinucleotídeo de Sequenciamento Duplex pode ser substancialmente semelhante a um ou mais dos padrões de espectro de trinucleotídeo derivados empiricamente, de modo que um profissional possa ser informado sobre a identidade da genotoxina de teste, o nível de exposição ao teste de genotoxina, o mecanismo de ação da genotoxina de teste etc., com base na semelhança com um ou mais padrões espectrais de trinucleotídeos derivados empiricamente.

Frequência mutante

[00164] Em algumas modalidades, as etapas de análise de Sequenciamento Duplex podem identificar uma frequência mutante associada a uma genotoxina específica sob várias condições de exposição. Por exemplo, uma frequência mutante associada à exposição de uma amostra biológica a uma genotoxina pode variar dependendo da variedade de fatores, incluindo, mas não limitado a, organismo/sujeito, idade do sujeito, tipo de genotoxina, quantidade de tempo ou nível de exposição a uma genotoxina, tipo de tecido, grupo de tratamento, região do genoma (por exemplo, locus genômico), por tipo de mutação, por tipo de substituição e por contexto de trinucleotídeo, entre outros fatores. Em alguns exemplos, a frequência mutante é medida como o número de mutações únicas detectadas por par de bases duplex sequenciado. Em outras modalidades, a frequência mutante é a taxa de novas mutações em um único gene ou organismo ao longo do tempo.

Espectro de Mutação

[00165] Em várias modalidades, as leituras de sequência de alta precisão (por exemplo, corrigidas por erros) geradas usando o Sequenciamento Duplex podem ser analisadas ainda mais para gerar um espectro de mutação ou assinatura para uma genotoxina específica ou genotoxina em potencial. Numa modalidade, um espectro ou assinatura de mutação compreende as combinações características de tipos de mutação decorrentes de processos mutagênicos resultantes de uma exposição a uma genotoxina. Tais combinações de características podem incluir informações relacionadas ao tipo de mutações (por exemplo, alterações na sequência ou estrutura do ácido nucleico). Por exemplo, um espectro de mutação pode compreender uma informação padrão sobre o número, localização e contexto de mutações pontuais (por exemplo, mutações de

base única), deleções nucleotídicas, rearranjos de sequência, inserções nucleotídicas e duplicações da sequência de DNA na amostra. Em algumas modalidades, um espectro de mutação pode incluir informações relevantes para determinar um mecanismo de ação que resulta nos padrões de mutação determinados. Por exemplo, o espectro da mutação pode ser capaz de determinar se os processos mutagênicos foram causados diretamente por exposições exógenas ou endógenas à genotoxina ou indiretamente desencadeados pela exposição à genotoxina via perturbação da infidelidade à replicação do DNA, vias de reparo do DNA com defeito e edição enzimática do DNA, entre outros. Em algumas modalidades, o espectro de mutação pode ser gerado por correspondência de padrões computacionais (por exemplo, agrupamento hierárquico não supervisionado do espectro de mutação, fatoração de matriz não negativa etc.).

Espectro/Assinatura de mutação triplete/Assinatura

[00166] Em uma modalidade, as leituras de sequência de alta precisão (por exemplo, corrigidas por erros) geradas usando o Sequenciamento Duplex podem ser analisadas ainda mais para gerar um espectro de mutação triplete (também aqui referido como espectro ou assinatura trinucleotídica). Por exemplo, o espectro de mutação associado a uma genotoxina e/ou a um incidente de exposição à genotoxina pode ser analisado adicionalmente para detectar variações ou mutações de nucleotídeo único ou mutações em um contexto de trinucleotídeo ou trinucleotídeo. Sem estar limitado pela teoria, reconhece-se que a exposição à genotoxina ou outros processos (por exemplo, envelhecimento) podem causar danos variáveis e/ou específicos aos ácidos nucleicos, dependendo do contexto trinucleotídico (por exemplo, uma base nucleotídica e suas bases adjacentes imediatas). Em algumas modalidades, uma genotoxina pode ter um espectro/assinatura único, semiúnico e/ou identificável de outro modo triplete. Por exemplo, um espectro de trinucleotídeos de uma primeira genotoxina pode incluir predominantemente mutações C·G→A·T e pode ainda ter uma maior predileção por sítios CpG. Esse espectro de trinucleotídeos é semelhante às etiologias propostas, dirigidas principalmente pela exposição ao tabaco, onde Benzo[α]pireno e outros hidrocarbonetos aromáticos policíclicos são mutagênicos conhecidos. Em outro exemplo, o uretano é uma genotoxina que gera dano ao DNA em um padrão periódico de T·A→A·T em um contexto de trinucleotídeo 5'-NTG-3'. Por conseguinte, em algumas modalidades, determinar um espectro de mutação triplete pode ser vantajoso para identificar uma

exposição de genotoxina em um sujeito, determinar a genotoxicidade de uma potencial genotoxina e identificar um mecanismo de ação de um agente ou fator genotóxico entre outros benefícios.

Mecanismo de ação

[00167] Em algumas modalidades, as leituras de sequência de alta precisão (por exemplo, corrigidas por erros) geradas usando o Sequenciamento Duplex podem ser usadas para inferir os processos bioquímicos que resultam nas alterações detectadas no ácido nucleico após a exposição a uma genotoxina específica. Por exemplo, em uma modalidade, a frequência mutante e o espectro de mutação (incluindo o espectro de trinucleotídeo) gerado usando um método de Sequenciamento Duplex pode ser comparado a informações derivadas empiricamente ou *a priori* sobre os padrões e propriedades bioquímicas associadas aos tipos de mutações observados, como bem como localização genômica da mutação genética ou dano ao DNA causado pela exposição à genotoxina. Em modalidades em que a via bioquímica e/ou processos fisiopatológicos que seguem a pré-mutação genômica detectada, mutação ou dano são determinados, essas informações podem ser usadas, em algumas modalidades, para informar sobre as opções de tratamento (por exemplo, terapêuticas ou profiláticas) para sujeitos expostos à genotoxina, ou em outras modalidades, essas informações podem ser usadas para informar a viabilidade dos esforços de comercialização (por exemplo, nova droga), esforços de limpeza (por exemplo, de uma toxina ambiental ou subproduto da fabricação) ou em outras modalidades, essas informações podem ser usadas para informar sobre um composto testado, agente ou fator que pode ser alterado para eliminar e/ou reduzir a genotoxicidade associada ao composto, agente ou fator.

Fontes de material de ácido nucléico para avaliação da genotoxicidade

[00168] Como discutido acima, é contemplado que o material de ácido nucleico possa vir de qualquer uma de uma variedade de fontes. Por exemplo, em algumas modalidades, o material de ácido nucleico é fornecido a partir de uma amostra de pelo menos um sujeito (por exemplo, um sujeito humano ou animal) ou outra fonte biológica. Em algumas modalidades, um material de ácido nucleico é fornecido a partir de uma amostra depositada/armazenada. Em algumas modalidades, uma amostra é ou compreende pelo menos um de sangue, soro, suor, saliva, líquido cefalorraquidiano, muco, líquido de lavagem uterina, um cotonete vaginal, um cotonete nasal, um cotonete oral, uma raspagem de tecido, cabelo, uma impressão digital, urina, fezes, humor vítreo, lavagem peritoneal, escarro, lavagem brônquica, lavagem oral, lavagem pleural,

lavagem gástrica, suco gástrico, bile, lavagem do ducto pancreático, lavagem do ducto biliar, lavagem comum do ducto biliar, líquido da vesícula biliar, líquido sinovial e ferida infectada, ferida não infectada, amostra arqueológica, amostra forense, amostra de água, amostra de tecido, amostra de alimento, amostra de biorreator, uma amostra de planta, raspagem das unhas, sêmen, líquido prostático, lavagem das tubas uterinas, um ácido nucleico livre de célula, um ácido nucleico dentro de uma célula, uma amostra metagenômica, uma lavagem de um corpo estranho implantado, uma lavagem nasal, líquido intestinal, escovação epitelial, lavagem epitelial, biópsia de tecido, uma amostra de autópsia, uma amostra de necropsia, uma amostra de órgão, uma ampla identificação humana, uma amostra de ácido nucleico produzida artificialmente, uma amostra de gene sintético, uma amostra de armazenamento de dados de ácido nucleico, tecido tumoral e qualquer combinação dos mesmos. Em outras modalidades, uma amostra é ou compreende pelo menos um de um micro-organismo, um organismo baseado em plantas ou qualquer amostra ambiental coletada (por exemplo, água, solo, arqueológico, etc.). Em exemplos particulares discutidos aqui, o material de ácido nucleico pode vir de uma fonte biológica que foi exposta a uma genotoxina ou a uma potencial genotoxina. Em alguns exemplos, a genotoxina é um mutagênico e/ou um carcinógeno. Em um exemplo, o material de ácido nucleico é analisado para determinar se a fonte biológica da qual o material de ácido nucleico é derivado foi exposta à genotoxina.

[00169] Quando comparado a outros ensaios de toxicidade conhecidos ou convencionais, como o teste de Ames (por exemplo, teste de mutagênese em bactérias), teste in vitro em cultura de células de mamíferos, ensaio de roedores transgênicos, ensaio de Pig-a e o bioensaio de dois anos in vivo, o Sequenciamento Duplex oferece múltiplos avanços. Por exemplo, muitos dos métodos da técnica anterior são limitados ao interrogatório de genes repórteres como substitutos para obter informações informativas relacionadas à genotoxicidade de um agente/fator de teste (por exemplo, teste de Ames, cultura celular de mamíferos in vitro, ensaio de roedores transgênicos in vivo) ou testes em fontes não humanas (por exemplo, teste de Ames, ensaio de roedores transgênicos, ensaio de Pig-a, bioensaio de dois anos), podem exigir longos períodos de tempo para serem preenchidos com muito pouca informação fornecida (por exemplo, bioensaio de dois anos em roedores de tipo selvagem) ou podem ser muito caros (por exemplo, ensaio de roedores transgênicos, bioensaio de dois anos). Em contraste com muitas das desvantagens dos

ensaios e técnicas da técnica anterior para rastrear agentes/fatores de teste quanto à genotoxicidade, os ensaios de Sequenciamento Duplex podem ser amplamente implementáveis, econômicos, adequados para triagem precoce e tardia de agentes/fatores de teste, utilizados para fornecer dados de alta precisão em curtos períodos de tempo (por exemplo, menos de 2 semanas), podem ser usados para rastrear amostras testadas in vitro e in vivo de qualquer organismo/fonte biológica (isto é, incluindo amostras humanas in vivo entre outras) ou qualquer tecido/órgão, avalia vários loci genéticos e pode usar um genoma natural como repórter de genotoxicidade e pode informar sobre o mecanismo de ação de um determinado agente/fator de genotoxina.

Kits com reagentes

[00170] Os aspectos da presente tecnologia abrangem ainda kits para a condução de vários aspectos dos métodos de Sequenciamento Duplex (também aqui referidos como um "kit DS"). Em algumas modalidades, um kit pode compreender vários reagentes, juntamente com instruções para conduzir um ou mais dos métodos ou etapas do método divulgados aqui para extração de ácido nucleico, preparação da biblioteca de ácidos nucleicos, amplificação (por exemplo, via PCR) e sequenciamento. Em uma modalidade, um kit pode incluir ainda um produto de programa de computador (por exemplo, algoritmo codificado para execução em um computador, um código de acesso a um servidor baseado em nuvem para executar um ou mais algoritmos, etc.) para analisar dados de sequenciamento (por exemplo, dados brutos de sequenciamento, leituras de sequenciamento, etc.) para determinar, por exemplo, uma frequência mutante, espectro de mutação, espectro de mutação triplete, comparação com espectros de mutação de genotoxinas conhecidas, etc., associados a uma amostra e de acordo com aspectos da presente tecnologia.

[00171] Em algumas modalidades, um kit DS pode compreender reagentes ou combinações de reagentes adequados para executar vários aspectos da preparação de amostras (por exemplo, extração de DNA, fragmentação de DNA), preparação da biblioteca de ácidos nucleicos, amplificação e sequenciamento. Por exemplo, um kit DS pode opcionalmente compreender um ou mais reagentes de extração de DNA (por exemplo, tampões, colunas, etc.) e/ou reagentes de extração de tecido. Opcionalmente, um kit DS pode ainda compreender um ou mais reagentes ou ferramentas para fragmentar o DNA de fita dupla, como por meios físicos (por exemplo, tubos para facilitar o corte ou sonicação acústica, unidade de nebulização etc.) ou meios enzimáticos (por exemplo, enzimas para cisalhamento genômico aleatório ou semialeatório e enzimas de

reação apropriadas). Por exemplo, um kit pode incluir reagentes de fragmentação de DNA para fragmentar enzimaticamente o DNA de fita dupla que inclui uma ou mais enzimas para digestão direcionada (por exemplo, endonucleases de restrição, endonucleases CRISPR/Cas e guias de RNA e/ou outras endonucleases), coquetéis Fragmentase de fita dupla, enzimas DNase de fita simples (por exemplo, nuclease de feijão mungo, nuclease S1) para renderizar fragmentos de DNA predominantemente de fita dupla e/ou destruir DNA de fita única e tampões e soluções apropriadas para facilitar essas reações enzimáticas.

[00172] Em uma modalidade, um kit DS compreende iniciadores e adaptadores para preparar uma biblioteca de sequências de ácidos nucleicos a partir de uma amostra que é adequada para executar etapas do processo de Sequenciamento Duplex para gerar sequências corrigidas por erros (por exemplo, alta precisão) de moléculas de ácidos nucleicos de fita dupla na amostra. Por exemplo, o kit pode compreender pelo menos um conjunto de moléculas adaptadoras compreendendo sequências de identificador de molécula única (SMI) ou as ferramentas (por exemplo, oligonucleotídeos de fita única) para o usuário criá-lo. Em algumas modalidades, o grupo de moléculas adaptadoras compreenderá um número adequado de sequências SMI substancialmente únicas, de modo que uma pluralidade de moléculas de ácido nucleico em uma amostra possa ser substancialmente marcada exclusivamente após a ligação das moléculas adaptadoras, isoladamente ou em combinação com características únicas dos fragmentos aos quais estão ligados. Alguém experiente na técnica da marcação molecular reconhecerá que o que implica um número "adequado" de sequências SMI variará por várias ordens de magnitude, dependendo de vários fatores específicos (DNA de entrada, tipo de fragmentação de DNA, tamanho médio de fragmentos, complexidade versus repetitividade sequências sendo sequenciadas dentro de um genoma etc.) Opcionalmente, as moléculas adaptadoras incluem ainda um ou mais sítios de ligação ao iniciador de PCR, um ou mais sítios de ligação ao iniciador de sequenciamento, ou ambos. Em outra modalidade, um kit DS não inclui moléculas adaptadoras que compreendem sequências ou códigos de barras SMI, mas inclui moléculas adaptadoras convencionais (por exemplo, adaptadores de sequenciamento em forma de Y, etc.) e várias etapas do método podem utilizar SMIs endógenas para relacionar as leituras da sequência de moléculas. Em algumas modalidades, as moléculas adaptadoras são adaptadoras de indexação e/ou compreendem uma sequência de indexação.

[00173] Em uma modalidade, um kit DS compreende um conjunto de moléculas adaptadoras, cada uma tendo uma região não complementar e/ou algum outro elemento de definição de fita (SDE), ou as ferramentas para o usuário criá-lo (por exemplo, oligonucleotídeos de fita simples). Em outra modalidade, o kit compreende pelo menos um conjunto de moléculas adaptadoras em que pelo menos um subconjunto das moléculas adaptadoras cada uma compreende pelo menos um SMI e pelo menos um SDE, ou as ferramentas para criá-los. Recursos adicionais para iniciadores e adaptadores para a preparação de uma biblioteca de sequenciamento de ácido nucleico a partir de uma amostra que é adequada para executar as etapas do processo de Sequenciamento Duplex estão descritos acima, bem como divulgados na Patente U.S. 9.752.188, Publicação Internacional de Patente WO2017/100441 e Pedido de Patente Internacional PCT/US18/59908 (depositado em 8 de novembro de 2018), todos incorporados por referência aqui na íntegra.

[00174] Além disso, um kit pode incluir ainda materiais de quantificação de DNA, como, por exemplo, corante de ligação de DNA, como verde SYBR™ ou ouro SYBR™ (disponível na Thermo Fisher Scientific, Waltham, MA) ou similar para uso com um fluorômetro Qubit (por exemplo, disponível na Thermo Fisher Scientific, Waltham, MA) ou no corante PicoGreen™ (por exemplo, disponível na Thermo Fisher Scientific, Waltham, MA) para uso em um espectrômetro de fluorescência adequado. Outros reagentes adequados para quantificação de DNA em outras plataformas também são contemplados. Outras modalidades incluem kits que compreendem um ou mais reagentes de seleção de tamanho de ácido nucleico (por exemplo, esferas magnéticas, géis, colunas de Imobilização Reversível em Fase Sólida (SPRI), colunas para captura de DNA alvo usando hibridação isca/oração, reagentes qPCR (por exemplo, para determinação de número de cópias) e/ou reagentes de PCR para gotículas digitais. Em algumas modalidades, um kit pode opcionalmente incluir uma ou mais enzimas de preparação de bibliotecas (ligase, polimerase(s), endonuclease(s), transcriptase reversa para, por exemplo, interrogações de RNA), dNTPs, tampões, reagentes de captura (por exemplo, esferas, superfícies, tubos revestidos, colunas, etc.), iniciadores de indexação, iniciadores de amplificação (iniciadores de PCR) e iniciadores de sequenciamento. Em algumas modalidades, um kit pode incluir reagentes para avaliar tipos de danos ao DNA, como uma DNA polimerase propensa a erros e/ou uma DNA

polimerase de alta fidelidade. Aditivos e reagentes adicionais são contemplados para reações de PCR ou ligação em condições específicas (por exemplo, genoma/alvo alto em GC).

[00175] Em uma modalidade, os kits compreendem ainda reagentes, como enzimas de correção de erros de DNA que reparam erros na sequência de DNA que interferem nos processos de reação em cadeia da polimerase (PCR) (versus reparação de mutações que levam à doença). A título de exemplo não limitativo, as enzimas compreendem um ou mais dos seguintes: Uracil-DNA Glicosilase (UDG), Formamidopirimidina DNA glicosilase (FPG), 8-oxoguanina DNA glicosilase (OGG1), endonuclease apurínica / apirimidínica humana (APE 1), endonuclease III (Endo III), endonuclease IV (Endo IV), endonuclease V (Endo V), endonuclease VIII (Endo VIII), proteína de N-glicosilase/AP-liase NEIL 1 (hNEIL1), T7 endonuclease I (T7 Endo I), T4 pirimidina dímero glicosilase (T4 PDG), uracil-DNA glicosilase monofuncional seletiva de fita única humana (hSMUG1), DNA glicosilase de alquiladenina humana (hAAG), etc.; e pode ser utilizadas para corrigir danos ao DNA (por exemplo, dano aoDNA *in vitro*). Algumas dessas enzimas de reparo do DNA, por exemplo, são glicosilases que removem bases danificadas do DNA. Por exemplo, o UDG remove o uracil resultante da desaminação da citosina (causada pela hidrólise espontânea da citosina) e o FPG remove a 8-oxo-guanina (por exemplo, a lesão de DNA mais comum resultante de espécies reativas de oxigênio). O FPG também possui atividade de liase que pode gerar folga de 1 base em sítios abásicos. Tais sítios abásico falham subsequentemente em amplificar por PCR, por exemplo, porque a polimerase falha em copiar o modelo. Por conseguinte, o uso de tais enzimas de reparo de danos ao DNA, e/ou outros listados aqui e como conhecido na técnica, pode remover efetivamente o DNA danificado que não possui uma mutação verdadeira, mas que pode ser detectado como um erro após a sequência e a análise de sequência duplex.

[00176] Os kits podem ainda compreender controles apropriados, como controles de amplificação de DNA, controles de quantificação de ácido nucleico (modelo), controles de sequenciamento, moléculas de ácido nucleico derivadas de uma fonte biológica exposta a uma genotoxina/mutagênico conhecida (por exemplo, DNA extraído de um animal de teste ou células cultivadas em cultura que foram expostas à genotoxina) e/ou moléculas de ácido nucleico derivadas de uma fonte biológica que não foi exposta a uma genotoxina/mutagênico. Em outra modalidade, os reagentes de controle podem incluir ácido nucleico que foi intencionalmente danificado e/ou ácido nucleico que não foi danificado ou exposto a qualquer agente prejudicial.

Em modalidades adicionais, um kit também pode incluir um ou mais agentes genotóxicos e/ou não genotóxicos (por exemplo, compostos) a serem distribuídos em um experimento controlado de genotoxicidade e, opcionalmente, incluem protocolos para distribuir esses agentes a um sujeito, tecido, célula, etc. Por conseguinte, um kit pode incluir reagentes adequados (compostos de teste, ácido nucleico, biblioteca de sequenciamento de controle, etc.) para fornecer controles que produziram resultados de sequenciamento duplex (por exemplo, um espectro/assinatura de mutação esperado) que determinariam a autenticidade do protocolo para uma substância de teste (por exemplo, composto de teste, potencial agente ou fator genotóxico, etc.). Em uma modalidade, o kit compreende recipientes para o transporte de amostras do sujeito, como amostras de sangue, para análise para detectar mutações em uma amostra do sujeito, o padrão e o tipo indicando assim a quais genotoxinas o sujeito foi exposto. Em outra modalidade, um kit pode incluir padrões de controle de contaminação por ácidos nucleicos (por exemplo, sondas de captura de hibridação com afinidade com regiões genômicas em um organismo que é diferente do organismo de teste ou sujeito).

[00177] O kit pode ainda compreender um ou mais outros recipientes que compreendem materiais desejáveis do ponto de vista comercial e do usuário, incluindo tampões de PCR e sequenciamento, diluentes, ferramentas de extração de amostras de sujeitos (por exemplo, seringas, cotonetes, etc.) e bulas com instruções de uso. Além disso, um rótulo pode ser fornecido no recipiente com instruções de uso, como as descritas acima; e/ou as instruções e/ou outras informações também podem ser incluídas em uma bula que acompanha o kit; e/ou através de um endereço de site fornecido nele. O kit também pode compreender ferramentas de laboratório, como, por exemplo, tubos de amostra, vedadores de placas, abridores de tubos de microcentrífuga, rótulos, separador magnético de partículas, inserções de espuma, compressas de gelo, compressas de gelo seco, isolamento, etc.

[00178] Os kits podem ainda compreender um produto de programa de computador instalável em um dispositivo de computação eletrônica (por exemplo, laptop/desktop, tablet etc.) ou acessível via rede (por exemplo, servidor remoto), em que o dispositivo de computação ou servidor remoto compreende um ou mais processadores configurados para executar instruções para executar operações que compreendem as etapas de análise de Sequenciamento Duplex. Por exemplo, os processadores podem ser configurados para executar instruções para o processamento

de leituras de sequenciamento brutas ou não analisadas para gerar dados de Sequenciamento Duplex. Em modalidades adicionais, o produto do programa de computador pode incluir um banco de dados que compreende registros de sujeitos ou amostras (por exemplo, informações sobre um sujeito em particular ou amostra ou grupos de amostras) e informações derivadas empiricamente sobre genotoxinas conhecidas). O produto do programa de computador é incorporado em um meio legível por computador não transitório que, quando executado em um computador, executa etapas dos métodos divulgados neste documento (por exemplo, ver FIGS. 19 e 20).

Os kits podem ainda incluir instruções e/ou códigos/senhas de acesso e semelhantes para acessar servidores remotos (incluindo servidores baseados em nuvem) para carregar e baixar dados (por exemplo, sequenciamento de dados, relatórios, outros dados) ou software para ser instalado em um dispositivo local. Todo o trabalho computacional pode residir no servidor remoto e ser acessado por um usuário/kit via conexão à Internet, etc.

Triagem de Genotoxina de Alto Rendimento

[00179] A presente tecnologia compreende ainda esquemas de triagem de alto rendimento para avaliar a genotoxicidade de agentes ou fatores suspeitos (por exemplo, um composto, químico, agente farmacêutico, produto ou subproduto industrial, substância alimentar, fator ambiental, etc.). Numa modalidade, um agente/fator com um efeito de genotoxicidade desconhecido pode ser rastreado para determinar se o agente/fator de teste compreende um efeito genotóxico. Em algumas modalidades, agentes/fatores podem ser rastreados com o desejo de eliminar o uso de agentes/fatores que têm um efeito genotóxico ou excedem um efeito genotóxico limiar. Por exemplo, um agente/fator que é mutagenético de uma maneira que pode potencialmente causar uma doença ou distúrbio associado à genotoxicidade pode ser identificado de modo que o agente/fator possa ser adequadamente controlado, eliminado, descartado, armazenado etc. Em algumas modalidades, agentes/fatores que são cancerígenos podem ser identificados usando esquemas de triagem de alto rendimento, conforme descrito aqui. Em outra modalidade, um agente/fator com um efeito de genotoxicidade desconhecido pode ser rastreado com a intenção de descobrir um agente/fator que tenha um efeito genotóxico desejado e, em particular, um efeito genotóxico desejado em uma fonte biológica alvo. Por exemplo, amostras biológicas derivadas de um paciente com uma doença ou distúrbio (por exemplo, câncer) podem ser usadas em um esquema de triagem de alto rendimento para testar vários agentes/fatores quanto ao efeito

genotóxico desejado, que pode resultar em perturbação ou destruição da célula (por exemplo, células cancerígenas). Essa triagem pode ser realizada para a descoberta de novas drogas/terapias e/ou terapias direcionadas para uso em medicina personalizada.

[00180] Em algumas modalidades, a triagem de alto rendimento se refere à triagem de uma pluralidade de amostras simultaneamente e/ou com eficiência de tempo. Em um exemplo, testar um agente ou fator de genotoxicidade compreende expor (por exemplo, tratar, administrar, aplicar etc.) um sujeito (por exemplo, uma fonte biológica) a um agente ou fator de teste. Por conseguinte, para esquemas de triagem de alta produção, uma matriz de fontes/amostras biológicas pode ser tratada simultaneamente com o mesmo agente/fator de teste ou em outras modalidades, com vários agentes/fatores de teste. Em um exemplo particular, uma pluralidade de amostras biológicas (por exemplo, células humanas ou de outros organismos cultivadas em cultura, amostras de tecidos, sangue ou outras amostras de fluidos corporais, células de animais transgênicos, células humanas cultivadas em xenoenxertos, organoides de pacientes vivos, células alimentadoras, etc.) pode ser exposto a um agente/fator de teste substancialmente simultaneamente e sob condições consistentes. A triagem de alto rendimento também pode ser usada por órgãos em chips, como o uso de um chip de 10 órgãos com amostras de sangue ou tecido do mesmo sujeito extraído dos seguintes órgãos e tecidos: endócrino; pele; trato GI; pulmão; cérebro; coração; medula óssea; fígado; rim; e pâncreas. Os métodos de uso de órgãos em chips para triagem de alto rendimento são bem conhecidos na técnica (por exemplo, Chan et al. [5]). Em outras modalidades, linhagens de células geneticamente modificadas (por exemplo, possuindo vias de reparo de DNA deficientes ou prejudicadas para tornar essas células mais sensíveis aos efeitos de danos mutagênicos ou genotóxicos) podem ser incorporadas a um esquema de triagem de alto rendimento.

[00181] Em algumas modalidades, a pluralidade de amostras biológicas pode ser a mesma ou substancialmente semelhante (por exemplo, linhagens celulares idênticas cultivadas em cultura, amostras de tecido do mesmo sujeito e/ou mesmo tipo de tecido, etc.). Em outras modalidades, uma ou mais dentre a pluralidade de amostras biológicas podem ser diferentes. Por exemplo, um agente/fator de teste pode ser testado quanto a um efeito genotóxico em diferentes tipos de tecido/célula do mesmo organismo, um organismo diferente ou uma combinação dos mesmos. Em um exemplo particular, um agente ou fator genotóxico suspeito (por exemplo, um composto, uma droga farmacêutica, etc.) pode ser testado simultaneamente em amostras de tecido

de vários órgãos do mesmo sujeito (por exemplo, um chip de 10 órgãos). Em algumas modalidades, a triagem de alto rendimento pode abranger o teste de múltiplos agentes/fatores de teste simultaneamente. Por conseguinte, é contemplado que cada amostra testada pode ter propriedades diferentes que podem intencionalmente variar ou não (por exemplo, por tipo de célula, por tipo de tecido, por sujeito do qual uma célula ou tecido é extraído, por espécie, etc.) e/ou ser submetida a diferentes regimes de teste que podem variar de acordo com o projeto (por exemplo, por agente/fator de teste, nível de dose, tempo de exposição etc.), de modo que um esquema de triagem de alto rendimento possa ser usado para rastrear com eficiência várias amostras de uma maneira que fornece qualquer informação desejada.

[00182] Depois que as amostras biológicas são expostas e/ou um regime de exposição desejado é concluído, as células/tecido das amostras podem ser colhidas e o DNA pode ser extraído com a finalidade de usar o Sequenciamento Duplex para avaliar o impacto genotóxico/mutagênico do agente/fator de teste no DNA derivado de cada amostra. Em algumas modalidades, o DNA sem células (como liberado em meios de cultura) pode ser coletado das amostras biológicas para análise de Sequenciamento Duplex. Outras modalidades contempladas pela presente tecnologia incluem processamento de alta produtividade de amostras de DNA para gerar dados de Sequenciamento Duplex para avaliar danos ao DNA, mutagenicidade ou carcinogenicidade de uma genotoxina conhecida ou suspeita.

[00183] Os processos de triagem de alto rendimento aqui descritos podem compreender automação, como por meio do uso de robótica para realizar um ou mais tratamentos experimentais de amostras biológicas, extração de DNA, etapas de preparação de bibliotecas, etapas de amplificação (por exemplo, PCR) e/ou etapas de sequenciamento de DNA (por exemplo, usando várias técnicas e dispositivos para sequenciamento massivamente paralelo). O uso de triagem de alto rendimento permite que uma pluralidade de amostras (isto é, diferentes tipos de células do mesmo sujeito ou os mesmos tipos de células de diferentes sujeitos) sejam testadas em paralelo, para que um grande número de amostras seja rapidamente rastreado quanto a mutações associadas a genotóxicos e/ou dano ao DNA.

[00184] Em uma modalidade, microplacas, cada uma das quais consiste em uma matriz de poços, cada poço compreendendo uma amostra, são movidos através do sistema por manipulação robótica. Em um exemplo, os poços nas microplacas podem ser preenchidos por meio

de sistemas automatizados de manuseio de líquidos e sensores podem ser usados para avaliar as amostras na microplaca, por exemplo, frequentemente após um período de incubação. O software de automação de laboratório pode ser usado para controlar todo ou parte do processo de triagem, garantindo assim a precisão no processo e a repetibilidade entre os processos.

Genotoxinas ambientais/exógenas

[00185] Os aspectos da presente tecnologia compreendem a avaliação da genotoxicidade de agentes/fatores ambientais/exógenos, como o uso de qualquer um dos métodos de triagem de Sequenciamento Duplex *in vivo* ou *in vitro* descritos acima. Aspectos adicionais da presente tecnologia compreendem avaliar se sujeitos/organismos foram expostos a uma genotoxina em uma área ambiental. Por exemplo, amostras biológicas (por exemplo, tecido, sangue) podem ser coletadas de organismos vivos ou expostos a uma área suspeita de contaminação para, por exemplo, determinar se uma área está contaminada. Em outras modalidades, amostras biológicas podem ser coletadas de organismos presentes em uma área maior e avaliadas como um processo de triagem para identificar uma localização geográfica específica de uma fonte de contaminação por genotoxina (por exemplo, subproduto industrial vazado/liberado no sistema de água). Vários métodos, conforme descrito aqui, podem ser usados para analisar amostras biológicas (por exemplo, de sujeitos) expostos a uma área ambiental que está sob investigação quanto à presença de uma possível genotoxina. Em outra modalidade, vários métodos, conforme descrito aqui, podem ser usados para analisar amostras biológicas colhidas de sujeitos suspeitos de serem expostos a uma genotoxina conhecida em uma área ambiental (por exemplo, uma área geográfica, uma área de estar, um ambiente ocupacional etc.). De acordo com aspectos da presente tecnologia, as amostras biológicas podem ser obtidas de múltiplos organismos (por exemplo, vida marinha, mamífero, alimentador de filtro, organismo sentinela, etc.) ou de uma espécie específica (por exemplo, amostras humanas).

[00186] As genotoxinas ambientais detectáveis compreendem ainda a exposição a um ou mais agentes mutagênicos, tais como, mas não limitados a irradiação gama, raios-X; irradiação UV; micro-ondas; emissões eletrônicas; gás venenoso; partículas de ar venenosas (por exemplo, inalação de amianto); e compostos químicos e/ou lagos, rios, córregos, águas subterrâneas, etc. contaminados por patógenos. Fontes adicionais de genotoxinas exógenas podem incluir, por

exemplo, substâncias alimentares, cosméticos, utensílios domésticos, produtos relacionados a cuidados de saúde, produtos e ferramentas de cozinha e outros consumíveis fabricados.

[00187] Os resultados do Sequenciamento Duplex podem ainda ser utilizados em conjunto com outros métodos para identificar a presença de contaminantes causadores de doenças, como um estudo epidemiológico que identifique primeiro a localização de um aglomerado de câncer. Em algumas modalidades, os métodos aqui divulgados podem ser utilizados para identificar as genotoxinas específicas que afetaram os membros do agrupamento. A partir desses dados, a fonte da genotoxina pode ser determinada. Ao contrário dos meios convencionais de investigação que tradicionalmente usavam informações correlativas para vincular uma doença ou condição médica de um sujeito a um evento causal (por exemplo, exposição a um mutagênico ou carcinógeno ambiental ou outro exógeno), o Sequenciamento Duplex fornece dados reproduzíveis e de alta precisão, como espectro de mutação e mecanismo de ação, cujos resultados podem ser usados para determinar empiricamente o(s) evento(s) causador(es), por exemplo, exposição a um mutagênico ou carcinógeno específico.

Genotoxinas endógenas

[00188] Aspectos da presente tecnologia compreendem a avaliação da genotoxicidade de agentes/fatores endógenos (por exemplo, uma genotoxina endógena ou processo genotóxico), como o uso de qualquer um dos métodos de triagem de Sequenciamento Duplex *in vivo* ou *in vitro* descritos acima. Conseqüentemente, aspectos da presente tecnologia compreendem avaliar se sujeitos/organismos experimentaram uma genotoxina endógena ou um processo genotóxico que causou danos ao DNA. Por exemplo, amostras biológicas (por exemplo, tecido, sangue) podem ser coletadas de um sujeito (por exemplo, um paciente) para, por exemplo, determinar se o sujeito tem uma doença ou distúrbio associado à genotoxina ou corre o risco de desenvolver essa doença ou distúrbio.

[00189] Fatores endógenos podem compreender, por meio de exemplos não limitativos: incidentes biológicos que causam a incorporação incorreta de nucleotídeos, como erros de DNA polimerase, radicais livres e depurinação. Os fatores endógenos podem ainda compreender o aparecimento de condições biológicas, a curto ou longo prazo, que contribuem diretamente para a mutação polinucleotídica associada a doenças ou distúrbios, como, por exemplo, estresse, inflamação, ativação de um vírus endógeno, doença autoimune; exposições ambientais;

escolhas alimentares (por exemplo, alimentos e bebidas cancerígenas); fumo; composição genética natural; envelhecimento; neurodegeneração; e assim por diante. Por exemplo, se um sujeito é exposto a longo prazo a altos níveis de estresse, ele pode ser testado por Sequenciamento Duplex para qualquer mutação correlacionada com cânceres associados ao estresse (por exemplo, leucemia, câncer de mama, etc.).

[00190] Os fatores endógenos também podem representar o acúmulo agregado de mutações e outros eventos genotóxicos nos tecidos de um indivíduo humano que refletem os efeitos integrais das exposições do indivíduo e podem não ser capazes de ser quantificados com precisão ou controlados experimentalmente.

Métodos para determinar os níveis seguros de frequência mutante

[00191] Um nível ou quantidade de dano ao DNA resultante de uma exposição a uma genotoxina pode variar dependendo de vários fatores, incluindo, por exemplo, a eficácia de uma genotoxina em causar dano ao DNA (direta ou indiretamente), dose ou quantidade de exposição, rota ou maneira de exposição (por exemplo, ingestão, inalação, absorção transdérmica, intravenosa, etc.), duração (por exemplo, ao longo do tempo) da exposição, efeitos sinérgicos ou antagônicos de outros agentes ou fatores aos quais o sujeito está exposto, além de várias características do sujeito (por exemplo, nível de saúde, idade, sexo, composição genética, eventos anteriores de exposição à genotoxina, etc.). Como discutido acima, a exposição a uma genotoxina pode resultar em danos ao ácido polinuclear que podem ser avaliados, por exemplo, pelos métodos de Sequenciamento Duplex, conforme descrito aqui, para determinar um espectro ou assinatura mutagênica única, semiúnica e/ou de outra forma identificável associada a esse que pode compreender um padrão de mutação (por exemplo, tipo de mutação, frequência mutante, mutações identificáveis em um contexto trinucleotídico) suficientemente semelhante a um padrão de mutação associado à doença conhecida (por exemplo, uma mutação genômica distinta para câncer de mama). Vários aspectos da presente tecnologia são direcionados a métodos para determinar e/ou quantificar os níveis de frequência mutante que podem ser considerados seguros compreendem ainda um método para detectar uma frequência mutante de limiar seguro para uma genotoxina. Quando a frequência mutante dentro da amostra está acima de um nível seguro, isso indica que o sujeito corre um risco significativamente maior de desenvolver a doença ao longo do tempo.

[00192] A presente tecnologia compreende ainda um método para detectar e quantificar mutações genômicas desenvolvidas *in vivo* em um sujeito após a exposição do sujeito a um mutagênico, compreendendo: (1) sequenciamento duplex de uma ou mais moléculas de DNA de fita dupla alvo extraídas de um sujeito exposto a um mutagênico; (2) gerar uma sequência de consenso corrigida por erro para as moléculas de DNA de fita dupla alvo; e (3) identificar um espectro de mutação para as moléculas de DNA de fita dupla direcionadas; (4) calcular uma frequência mutante para as moléculas de DNA de fita dupla alvo, calculando o número de mutações únicas por par de bases duplex sequenciadas. Em uma modalidade da etapa (3), o espectro de mutação é o perfil único de uma amostra que compreende uma "assinatura de trinucleotídeo".

[00193] Em uma modalidade, as etapas (1) e (2) são realizadas por: a) ligar a molécula de ácido nucleico alvo de fita dupla a pelo menos uma molécula adaptadora, para formar um complexo de ácido nucleico alvo-adaptador, em que pelo menos uma molécula adaptadora compreende: i. uma sequência degenerada ou semidegenerada de identificador de molécula única (SMI) que sozinha ou em combinação com os pontos de cisalhamento de ácido nucleico alvo marca exclusivamente a molécula de ácido nucleico alvo de fita dupla; e ii. uma sequência nucleotídica que marca cada fita do complexo de ácido nucleico alvo-adaptador, de modo que cada fita do complexo de ácido nucleico alvo-adaptador tenha uma sequência nucleotídica distintamente identificável em relação à sua fita complementar, b) amplificar cada fita do complexo de ácido nucleico alvo-adaptador para produzir uma pluralidade de amplicons do complexo de ácido nucleico alvo da primeira fita e uma pluralidade de amplicons do complexo de ácido nucleico alvo-adaptador da segunda fita e; c) sequenciar os amplicons do complexo de ácido nucleico alvo-adaptador para produzir uma pluralidade de leituras de sequência da primeira fita e uma pluralidade de leituras de sequência da segunda fita; e d) comparar pelo menos uma leitura de sequência a partir da pluralidade de leituras de sequência da primeira fita com pelo menos uma leitura de sequência a partir da pluralidade de leituras de sequência da segunda fita e gerar uma leitura corrigida por erro da sequência da molécula de ácido nucleico alvo de fita dupla descontando as posições nucleotídicas que não concordam (ver Patente U.S. 9752188 B2 e WO 2017/100441).

Métodos para determinar níveis limiares seguros da quantidade de genotoxina

[00194] A presente tecnologia compreende ainda métodos experimentais *in vitro* e *in vivo* para determinar níveis seguros (quantidades de concentração em peso ou volume ou massa ou unidade*integrais de tempo etc.) de exposição por um sujeito a uma genotoxina específica; e/ou se um composto ou outro agente (por exemplo, ondas de rádio do dispositivo sem fio etc.) é genotóxico em qualquer nível de exposição. Essa determinação pode depender primeiro da determinação do nível de frequência mutante do limiar seguro. Em uma modalidade, a amostra de um sujeito de controle é testada quanto a genotoxinas (ou a falta dela) e comparada com o perfil de genotoxina das amostras de sujeitos expostos (por exemplo, uma pluralidade de camundongos; ou uma pluralidade de células do mesmo sujeito, um conjunto dos quais são as células de controle etc.). Os indivíduos expostos recebem quantidades de exposição designadas e predeterminadas de suspeita de genotoxina para determinar o nível limiar de exposição segura antes que ocorra uma mutação induzida por genotoxina detectada que contribua diretamente para o início da doença.

[00195] Em outra modalidade, os sujeitos do teste (por exemplo, animais de laboratório, células *in vitro*, etc.) são expostos a diferentes doses por diferentes períodos de tempo e a partir dos quais é determinado o nível de corte seguro da exposição à genotoxina: 1) a qual dose de exposição não são observadas mutações polinucleotídicas; e/ou 2) a qual dose de exposição são detectadas mutações polinucleotídicas, mas onde o nível equivalente à dose não causa câncer em sujeitos e usando o nível de mutações encontrado para inferir o mesmo de outros compostos; e/ou 3) determinar uma curva de resposta à dose de genotoxina e análise de regressão de mutações induzidas para extrapolar uma curva linear de resposta à baixa dose; e/ou 4) qual é a razão de risco para um dado resultado de saúde em uma população em questão que está associada a uma frequência/assinatura de genotoxina detectada.

[00196] Os níveis limiares de exposição segura podem ainda ser determinados por espécies - por exemplo, humanos, cães/gatos, cavalos etc. Os limiares seguros podem ainda ser determinados por vias de exposição à genotoxina. Por exemplo, experimentos usando várias quantidades de genotoxinas podem ser testados com os métodos de Sequenciamento Duplex aqui divulgados para determinar a quantidade (peso, volume, etc.) e/ou frequência pelo consumo oral, tópico ou em aerossol que resultaria em uma mutação e espectro triplete associado a um desenvolvimento específico da doença.

[00197] E/ou os métodos experimentais de Sequenciamento Duplex divulgados neste documento podem ser usados para determinar a quantidade limiar de exposição genotóxica com base no tempo e/ou temperatura. Por exemplo, a absorção pela pele de um chuveiro ou banho em água contendo uma genotoxina com base na duração da exposição, temperatura da água e concentração da genotoxina na água, pode ser usada para calcular a quantidade (dose) de genotoxina absorvida pela pele.

[00198] Os resultados do Sequenciamento Duplex corrigidos por erros, identificando os níveis de limiares seguros de genotoxina, podem ser combinados com outros dados de limiares de segurança (por exemplo, níveis existentes da FDA e EPA, níveis da Agência de Registro de Doenças Tóxicas de Substâncias, diretrizes do Programa Nacional de Toxicologia dos EUA, diretrizes da OECD, diretrizes Canadenses de Saúde, Diretrizes reguladoras Europeias, diretrizes ILSI/HESI etc.) para afirmar ou ajustar os padrões estabelecidos.

Métodos de detecção e tratamento

[00199] A doença ou o início da doença podem não ser capazes de serem diagnosticados por meio de técnicas tradicionais de teste e imageamento até muitos anos após a exposição à genotoxina (por exemplo, 20 anos); mas a presente tecnologia fornece métodos para detectar as mutações causadoras de doenças ou indicar processos genotóxicos com potencial para causar mutações ou precursores causadores de mutações, dentro de alguns dias ou algumas semanas ou meses após a exposição à genotoxina, a fim de tratar profilaticamente o sujeito ou rastrear ativamente o sujeito em busca de doenças (em virtude de estar em um nível de risco mais alto), bem como identificar a presença de uma genotoxina e eliminá-la para evitar exposições futuras.

[00200] Quando um sujeito é exposto a mais de um nível seguro de limiar de genotoxina e/ou quando se determina que um sujeito foi potencialmente exposto a níveis inseguros de uma genotoxina (por exemplo, departamento de saúde identificando níveis perigosos de exposição), então o sujeito está em um risco significativamente aumentado para o aparecimento da doença ou distúrbio genotóxico associado. O sujeito é então tratado profilaticamente com agentes que bloqueiam e/ou neutralizam a genotoxina; e/ou a exposição à genotoxina é reduzida ou eliminada (por exemplo, removendo a genotoxina do ambiente ou movendo o sujeito). Além disso, ou alternativamente, o sujeito é submetido a testes de diagnóstico com tempo sequencial

(por exemplo, exame de sangue para detecção de câncer) e/ou imageamento (por exemplo, CAT, MRI, PET, ultrassom, teste de biomarcadores séricos, etc.) para detectar se o sujeito desenvolveu um estágio inicial da doença ou distúrbio, durante o qual é tratado com mais eficácia. A título de exemplo não limitativo: para a exposição à aflatoxina ou ao ácido aristolocóico, o sujeito provavelmente seria submetido a um ultrassom hepático a cada 6 meses, o horário típico em que os pacientes com hepatite C crônica, outro hepatocarcinogênio, são rastreados quanto a carcinomas hepatocelulares. No momento em que os testes de diagnóstico tradicionais bem conhecidos na técnica detectam a doença (por exemplo, câncer), o tratamento é iniciado (por exemplo, cirurgia, quimioterapia, imunoterapia etc.).

[00201] Os métodos para fornecer tratamentos profiláticos (isto é, prevenir ou reduzir o risco de aparecimento) e/ou inibir o crescimento do câncer e/ou erradicar o câncer compreendem protocolos de tratamento bem conhecidos pelo clínico especializado e seriam adaptados ao tipo de genotoxina. Embora atualmente não existam tratamentos para reverter mutações que já foram induzidas, métodos terapêuticos para ajudar um sujeito a limpar certas genotoxinas residuais (por exemplo, metais pesados específicos por quelação), podem diminuir ainda mais a genotoxicidade.

[00202] Para tumores induzidos por mutagênico (por exemplo, câncer de pulmão em fumantes, melanoma em casos de câncer de boca exposto a UV, etc.), a carga de mutações nesses tumores tende a ser maior, o que se acredita levar a uma maior abundância de neoantígenos e explica sua tendência muito maior a responder favoravelmente às imunoterapias. É provável que a administração profilática de imunoterapias, como aquelas que compreendem inibidores de ponto de verificação (isto é, inibidores de PD1 e PDL1, como nivolumabe, pembrolizumabe e atezolizumabe, inibidores de CTLA4, como ipilizumabe), para permitir que o sistema imunológico do sujeito erradique tumores em formação precoce. Portanto, outro uso direcionado ao tratamento da identificação de uma assinatura de exposição é a previsão de resposta futura do tumor à imunoterapia e potencialmente até a prevenção de doenças com tratamento profilático, apesar de exigir testes cuidadosos no cenário de ensaios clínicos formais.

[00203] Os métodos de detecção e tratamento podem ainda compreender métodos de determinação direta ou inferencial do mecanismo de ação da genotoxina, que pode ser usado na determinação do curso apropriado do tratamento; e/ou monitoramento de variantes resistentes a droga (ver Schmitt et al [6]).

[00204] Uma vez que o sujeito é diagnosticado ou detectado como exposto a pelo menos uma genotoxina, o sujeito pode receber uma quantidade terapeuticamente eficaz de uma composição farmacêutica para evitar o início, retardar o início, reduzir os efeitos e/ou erradicar a doença associada à genotoxina ou distúrbio. Uma composição farmacêutica compreende uma quantidade terapeuticamente eficaz de uma composição compreendendo um inibidor ou erradicador de uma doença ou distúrbio associado à genotoxina e um transportador ou sal farmacêuticamente aceitável. E uma quantidade terapeuticamente eficaz compreende a faixa de doses terapêuticas, não tóxicas, da composição que compreende um inibidor ou erradicador de uma doença ou distúrbio associado à genotoxina, eficaz para produzir o resultado farmacológico, terapêutico ou profilático pretendido.

[00205] A composição farmacêutica é formulada e administrada por uma via de administração compreendendo: administração oral, intravenosa, intramuscular, subcutânea, intrauretral, retal, intraespinal, tópica, bucal ou parenteral. A composição farmacêutica pode ser misturada com transportadores e excipientes farmacêuticos convencionais e usada na forma de comprimidos, cápsulas, pílulas, líquidos, soluções intravenosas, bebidas e produtos alimentícios e semelhantes; e conterà de cerca de 0,1% a cerca de 99,9%, ou cerca de 1% a cerca de 98%, ou cerca de 5% a cerca de 95%, ou cerca de 10% a cerca de 80%, ou cerca de 15% a cerca de 60%, ou cerca de 20% a cerca de 55% em peso ou volume do ingrediente ativo.

[00206] Para administração oral, os comprimidos, pílulas e cápsulas podem adicionalmente transportadores convencionais, tais como agentes de ligação, por exemplo, goma arábica, gelatina, polivinilpirrolidona, sorbitol ou tragacanto; enchimentos, por exemplo, fosfato de cálcio, glicina, lactose, amido de milho, sorbitol ou sacarose; lubrificantes, por exemplo, estearato de magnésio, polietileno glicol, sílica ou talco; desintegrantes, por exemplo, amido de batata, agentes aromatizantes ou corantes ou agentes umectantes aceitáveis. As preparações líquidas orais podem ser formuladas em soluções aquosas ou oleosas, suspensões, emulsões, xaropes ou elixires e podem conter aditivos convencionais, tais como agentes de suspensão, agentes emulsificantes, agentes não aquosos, conservantes, corantes e aromatizantes.

[00207] Para vias de administração intravenosa, a composição farmacêutica pode ser dissolvida ou suspensa em qualquer um dos fluidos intravenosos comumente usados e

administrada por infusão. Os fluidos intravenosos incluem, sem limitação, solução salina fisiológica ou solução de Ringer.

[00208] As composições farmacêuticas para administração parental podem estar na forma de soluções ou suspensões estéreis isotônicas aquosas ou não aquosas. Estas soluções ou suspensões podem ser preparadas a partir de pós ou grânulos estéreis possuindo um ou mais dos carreadores mencionados para utilização nas formulações para administração oral. Os compostos podem ser dissolvidos em polietileno glicol, propileno glicol, etanol, óleo de milho, álcool benzílico, cloreto de sódio e/ou vários tampões.

[00209] A dose do efeito terapêutico pode ainda ser calculada com base em uma variedade de fatores, tais como: quantidade ou duração da exposição genotóxica; idade, peso, sexo ou raça do sujeito; estágio de desenvolvimento da doença ou distúrbio; e outros métodos bem conhecidos do clínico versado. Numa modalidade, o sujeito é testado após a descoberta de seu potencial ou suspeita de exposição a uma genotoxina, mesmo que a exposição tenha ocorrido muitos anos antes. Se diagnosticado como exposto acima de um nível limiar seguro, o sujeito é administrado ao composto farmacêutico imediatamente ou após a exibição dos sintomas. Em todas as modalidades, a genotoxina é removida do ambiente do sujeito, quando possível.

Exemplos Experimentais

[00210] A seção a seguir fornece exemplos de métodos para detectar e avaliar a mutagênese genômica *in vivo* usando o Sequenciamento Duplex e reagentes associados. Os exemplos a seguir são apresentados para ilustrar a presente tecnologia e para ajudar um versado na técnica a fazer e usar a mesma. Os exemplos não têm a intenção de limitar o escopo da tecnologia.

[00211] Geralmente, para comparar a eficácia do DS *na medição da mutagênese in vivo*, uma série de experimentos com camundongos que geraram 8,2 bilhões de bases corrigidas por erros em 62 amostras foi realizada para examinar o efeito de três mutagênicos em nove genes de cinco tecidos saudáveis em duas fitas de animais independentes. O Sequenciamento Duplex demonstrou quantitativamente uma frequência mutante aumentada entre os animais tratados, em uma extensão que variou por mutagênico específico, tipo de tecido e locus genômico, e espelhava de perto o de um ensaio de roedor transgênico padrão-ouro. Em vários exemplos, foi possível identificar amostras por seu grupo de tratamento com base apenas em padrões mutacionais

objetivos. Em alguns exemplos, a sensibilidade do mutagênico variou até quatro vezes entre os diferentes loci gênicos e, sem estar vinculado à teoria, os padrões espectrais sugeriram que isso fosse parcialmente o resultado de processos distintos regionalmente, que podem incluir transcrição e metilação. Em vários exemplos, a assinatura mutacional trinucleotídica entre os SNVs identificados por DS em frequência ultrabaixa em animais tratados com o carcinógeno benzo[a]pireno relacionado ao tabaco mostrou-se quase idêntica à observada entre os SNVs clonais nos genomas de câncer de pulmão associado ao tabagismo em bancos de dados publicamente disponíveis. Em alguns exemplos, o DS foi utilizado para identificar mutações oncogênicas de baixa frequência que se expandem clonicamente sob pressão seletiva, apenas 4 semanas após um tratamento com mutagênico. Por conseguinte, e como demonstrado em vários exemplos aqui descritos, o DS pode ser usado para quantificar diretamente processos genotóxicos e evolução neoplásica em tempo real, com diversas aplicações em biologia mutacional, toxicologia e avaliação de risco de câncer.

Exemplo 1

[00212] Aplicação do Sequenciamento Duplex para análise de mutação *in vivo* no transgene *cII* e genes endógenos em Camundongos BigBlue[®]. Esta seção descreve um exemplo em que o Sequenciamento de Próxima Geração (NGS) corrigido por erro foi usado para medir diretamente mutações induzidas quimicamente no transgene *cII* usado no ensaio de mutação de roedor transgênico (TGR) BigBlue[®] e nos genes nativos de camundongos. Atualmente, os ensaios de mutação TGR detectam mutantes raros de *cII* através da formação de placas. O NGS padrão é inutilizável para a detecção de mutações de baixa frequência devido à sua alta taxa de erros (~ 1 erro por 10³ bases sequenciadas). O NGS corrigido por erro, ou Sequenciamento Duplex, tem uma taxa de erro drasticamente mais baixa (~ 1/10⁸ bases), permitindo a detecção de mutações ultrarraras.

[00213] Neste exemplo, uma aplicação do Sequenciamento Duplex foi utilizada para avaliar a frequência mutante (MF) e o espectro no controle, N-etil-N-nitrosourea (ENU) e camundongos machos BigBlue[®] C57BL6 expostos a Benzo[a]pireno (B[a]P).

[00214] Os camundongos machos BigBlue[®] transgênicos C57BL/6 foram tratados por gavagem oral diária com veículo (azeite) ou B[a]P (50 mg/kg/dia) nos Dias 1-28 ou com ENU (40 mg/kg/dia em tampão pH 6) nos Dias 1-3 (n=6). Os tecidos foram coletados e congelados no dia

31 do estudo. O fígado e a medula óssea foram analisados quanto a mutantes. O DNA foi isolado e os mutantes analisados para placas mutantes de *cII* usando os métodos RecoverEase e Transpack descritos pela Agilent Technologies. O sequenciamento duplex foi usado para sequenciar *cII* e outros genes endógenos para mutações no fígado e medula óssea.

[00215] Os genes avaliados e os critérios utilizados para selecionar genes são os seguintes: (1) *Polr1c* (RNA polimerase), que é onipresentemente transcrito em todos os tipos de tecidos; (2) *Rho* (Rodopsina), que não é expresso em nenhum tecido além da retina; (3) *Hp* (Haptoglobina), que é altamente expresso no fígado, mas quase em nenhum outro lugar; (4) *Ctnnb1* (Beta-catenina), que é o gene mais comumente mutado no carcinoma hepatocelular humano; e (5) gene repórter transgênico *CII*: 360 pb presente em ~ 80 cópias em camundongos BigBlue[®].

[00216] As FIGS. 3A-3D são gráficos de caixas que mostram frequências mutantes calculadas para o sequenciamento em duplex (FIGS. 3A e 3B) e o ensaio de placas BigBlue[®] *cII* (FIGS. 3C e 3D) no fígado e medula óssea após tratamento com mutagênico, como descrito acima. A MF para Sequenciamento Duplex foi baseada no total de mutantes por par de bases duplex sequenciado (n = 5 camundongos/grupo). A MF para BigBlue[®] foi calculada como o número de placas mutantes em relação ao número de unidades formadoras de placas mutantes (n = 6 camundongos/grupo). Como mostrado, a MF medida por Sequenciamento Duplex e o tradicional ensaio de placa BigBlue[®] *cII* deu respostas semelhantes a ambos os mutagênicos. A medula óssea, que possui células que se dividem mais rapidamente, demonstrou maior MF do que o fígado usando os dois métodos.

[00217] A FIG. 3E ilustra o aumento relativo da dobra mutante de *cII* no ensaio de roedores transgênicos vs Sequenciamento Duplex. Como acima, a MF no ensaio de placa é calculado como o número de placas mutantes fenotipicamente ativas observadas em uma placa de seleção dividido pelo número total de placas formadas em uma placa permissiva. MF no ensaio de Sequenciamento Duplex é calculado como o número de observações de pares de bases mutantes dividido pelo número total de pares de bases sequenciados dentro do intervalo de transgene 297 BP *cII*. Apesar das diferenças nas medidas derivadas, a correlação entre o ensaio de sequenciamento duplex e o ensaio de placa BigBlue[®] *cII* é forte entre os tratamentos de tecidos e mutagênicos.

[00218] A FIG. 3F mostra a proporção de variantes de SNVs dentro do gene *cII* para placas mutantes escolhidas individualmente produzidas a partir de tecido de camundongo BigBlue® e Sequenciamento Duplex do gDNA de *cII* a partir dos tecidos de camundongo BigBlue®. Os SNVs são designados com pirimidina como referência. O Sequenciamento Duplex produz o mesmo espectro de mutação de cada grupo de tratamento obtido pela coleta manual de 3.510 placas (todos os três valores de $p > 0,999$ com teste do qui-quadrado). As proporções foram calculadas dividindo-se o total de observações de SNVs pelas contagens observadas de bases de referência dentro do intervalo *cII* e normalizando para uma.

[00219] A FIG. 3G mostra a distribuição de todas as mutações identificadas pelo Sequenciamento Duplex direto de *cII* em todos os tipos de tecidos e grupos de tratamento BigBlue® por posição do códon e consequência funcional. A FIG. 3H mostra dados de distribuição para mutações identificadas entre placas mutantes coletadas individualmente. Com referência às FIGS. 3G e 3H juntas, o Sequenciamento Duplex direto (FIG. 3G) identificam mutações ao longo de todo o gene que causam todas as classes de efeitos, enquanto as mutações das placas mutantes selecionadas (FIG. 3H) são desprovidas de variantes e mutações sinônimas nos terminais C e N não críticos da proteína. Sem estar limitado pela teoria, acredita-se que variantes e mutações sinônimas nos terminais C e N não críticos da proteína não causem interrupções na função do gene, o que é necessário para o crescimento seletivo e a pontuação no ensaio de placa.

[00220] A FIG. 4 é um gráfico de barras mostrando que a MF medida por Sequenciamento Duplex é consistente dentro de cada grupo de tratamento. A MF, agregada em todos os genes, foi medida no fígado e medula óssea pelo Sequenciamento Duplex. O número de mutantes únicos foi baixo em animais de controle de veículo (1-13 mutações/1,4 bilhões de pares de bases) em relação a camundongos expostos a mutagênicos (até 118 mutações/2,6 bilhões de pares de bases). A MF entre os animais de um grupo foi reproduzível em todas as condições de tratamento e o baixo número de mutações nos animais de controle (1 a 13) enfatiza a necessidade de sequenciamento profundo para gerar estimativas robustas da MF.

[00221] As FIGS. 5A e 5B são gráficos de barras que mostram MF de genes endógenos em comparação com o transgene *cII* no fígado (FIG. 5A) e medula óssea (FIG. 5B) e como medido por Sequenciamento Duplex. Cada gene (~ 3 a 6 kb) foi sequenciado a uma profundidade de aproximadamente 5000x, com o gene *cII* (~ 350 pb x 80 cópias por genoma)

sequenciado a uma profundidade de ~ 100K a 300K. A frequência mutante foi calculada como descrito acima e em relação às FIGS. **3A-3D**. Como mostrado, os genes endógenos exibem um aumento semelhante na MF como o transgene *cII*. O Sequenciamento Duplex demonstra que a MF é mais alta na medula óssea do que no fígado. Sem estar limitado pela teoria, a maior taxa de divisão celular na medula óssea pode explicar os níveis mais altos de MF detectados para ambos os mutagênicos testados. Além disso, as diferenças na resposta dos genes endógenos mostradas nas FIGS. **5A** e **5B** podem estar relacionados a diferenças no estado transcricional ou na estrutura cromática dos genes endógenos.

[00222] A FIG. **5C** são gráficos de caixa que mostram SNV MF calculado para o Sequenciamento Duplex por regiões gênicas para medula óssea e fígado e FIG. **5D** é um gráfico de dispersão mostrando medições individuais de dados agregados mostrados na FIG **5C**. Os pontos de dispersão mostram medições individuais com CI de 95% ao seu redor. O gráfico da caixa na FIG. **5C** mostra todos os quatro quartis de todos os pontos de dados para essa categoria de tecido e tratamento. As escalas do eixo Y são apresentadas linearmente e na magnitude 10^{-7} . Com referência à FIG. **5C**, o gráfico de caixa resume o agregado das frequências de mutação do SNV no fígado e tecidos da medula óssea nos quatro genes endógenos e o transgene *cII* do modelo de camundongo Big Blue[®] mostrado na FIG. **5D**. A extensão da indução de mutação é influenciada por mutagênico específico, tipo de tecido e locus genético.

[00223] A FIG. **6** é um gráfico de barras que mostra o espectro de mutação de cada mutagênico de teste (por exemplo, tratamento) dentro dos tecidos testados, conforme medido por Sequenciamento Duplex. Com referência à FIG. **6**, a porção de cada mutação, agregada em todos os genes e calculada para cada amostra e agrupada por análise de agrupamento hierárquico não supervisionada demonstra que o espectro da mutação é exclusivo para cada tratamento (por exemplo, mutagênico de teste). A análise não supervisionada de agrupamento de dados codificados permitiu o agrupamento de dados com base no espectro de mutações e demonstra que as amostras de ENU são facilmente identificadas em todos os tecidos por uma preponderância de mutações de $T \rightarrow C$, $T \rightarrow A$, e $C \rightarrow T$. Da mesma forma, as amostras de B[a]P são distinguidas pelas mutações $C \rightarrow A$ e $G \rightarrow T$.

[00224] As FIGS. **7A-7C** são gráficos que mostram espectros de mutação no contexto de nucleotídeos adjacentes (isto é, espectros de trinucleotídeos) para controle de veículo (**7A**),

B[a]P (7B) e ENU (7C). A assinatura mutacional no formato de espectro de trinucleotídeos fornece informações sobre diferentes mecanismos de mutagênese e/ou demonstra padrões mutacionais exclusivos para mutagênicos específicos. Por exemplo, os contextos CCG e CGC parecem ser mais vulneráveis ao agente cancerígeno associado ao tabaco, B[a]P, do que outros contextos (FIG. 7B). Esse padrão de assinatura pode ser semelhante aos padrões de assinatura demonstrados pela exposição à aflatoxina (por exemplo, pode ser um mecanismo semelhante de mutagênese). A FIG. 7C ilustra que o alquilador, ENU, tem dois contextos vulneráveis que correspondem ao código IUPAC GTS, onde S+[G][C], e é um forte indutor de mutações de transição.

[00225] Neste exemplo, foi demonstrado que a carga de mutação nas amostras de medula óssea e fígado tratados com ENU e B[a]P aumentou significativamente em relação aos controles, comparável à frequência tradicional de placa mutante BigBlue[®] *cII* (frequência mutante MF) e variou da mesma forma por tipo de tecido. A avaliação do espectro revelou padrões distintos de INDELS e substituições de base única em cada grupo de tratamento. a análise da base trinucleotídica demonstrou que o contexto nucleotídico adjacente modula fortemente o potencial mutagênico; os pontos críticos mais extremos foram CCG e CGC para B[a]P e GTG e GTC para ENU. O Sequenciamento Duplex foi estendido a 4 genes endógenos: *Polr1c*, rodopsina, haptoglobina e beta-catenina. Mais uma vez, a MF aumentou em animais expostos a ENU e B[a]P, mas variou significativamente de acordo com o locus genômico, provavelmente refletindo o status transcricional. Neste exemplo, o Sequenciamento Duplex demonstra ser um método bem-sucedido para detectar mutações no transgene *cII*, um biomarcador de segurança pré-clínico aceito em ensaios TGR, mas, além disso, este exemplo demonstra que o Sequenciamento Duplex pode ser a base de ferramentas de avaliação de risco baseadas em genes endógenos relacionados ao câncer.

Exemplo 2

[00226] Quantificação direta da mutagênese química *in vivo* em genomas de mamíferos usando sequenciamento duplex. Esta seção descreve um exemplo em que o Sequenciamento Duplex é usado para determinar se mutações precoces nos genes controladores de câncer refletem o potencial tumorigênico dos mutagênicos em teste.

[00227] Neste exemplo, o impacto de um uretano é examinado em diferentes tipos de tecido de camundongo (pulmão, baço, sangue) em um modelo de camundongo predisposto ao câncer aprovado pela FDA: Tg.rasH2 (Saitoh et al. Oncogene 1990. PMID 2202951). Este

camundongo contém ~ 3 cópias em tandem de *Hras* humano com uma mutação potenciadora ativadora para aumentar a expressão em um alelo hemizigótico. Estes camundongos estão predispostos a angiossarcomas esplênicos e adenocarcinomas de pulmão e são rotineiramente usados em estudos de carcinogenicidade de 6 meses para substituir estudos em animais nativos de 2 anos. Os tumores encontrados nos camundongos geralmente adquirem mutações ativadoras em uma cópia do proto-oncogene *Hras* humano. Nesta adição aos 4 genes nativos de camundongo (*Rho*, *Hp*, *Ctnnb1*, *Polr1c*), o transgene *Hras* de camundongo nativo e o *Hras* humano também são analisados neste exemplo.

[00228] Neste exemplo, os camundongos Tg.rasH2 (n = 5/grupo) foram dosados com veículo ou uma dose carcinogênica de uretano (dia 1,3,5) e sacrificados no dia 29 para detecção de mutação pelo Sequenciamento Duplex nos tecidos alvo (pulmão, baço) e sangue total. Os genes endógenos (*Rho*, *Hp*, *Ctnnb1*, *Polr1c*) e os genes nativos de camundongo e (trans)genes *Hras* humanos também foram sequenciados.

[00229] Os tumores (hemangiossarcomas esplênicos; adenocarcinoma de pulmão) foram coletados na semana 11 de animais (n = 5/grupo) dosados com uretano e submetidos a sequenciamento total do exoma (WES) para identificar mutações características do ativador de câncer (CDM) nesses tumores.

[00230] A FIG. 8 é um gráfico de barras que mostra a frequência mutante (MF) de amostras de pulmão, baço e sangue para animais de controle e experimentais submetidos a uretano. Nesta análise, cada variante única detectada foi contada como uma mutação, que foi somada por amostra. Isso foi dividido pelo número total de Bases Duplex sequenciadas e por toda a região de captura. O número de eventos é anotado acima de cada amostra. No total, em todas as 30 amostras, foram gerados 3.966.947.832 pares de bases Sequenciadas em Duplex. Como mostrado na FIG. 8, a indução da mutação é consistente entre animais no mesmo grupo de tratamento e a confiança aumenta com a profundidade do sequenciamento.

[00231] A FIG. 9 é um gráfico de barras que mostra a frequência média mutante de pontos mínimos em cada grupo de amostras de tecido (as barras de erro são +/- um desvio padrão).

Tabela 1

Tecido	Tratamento	Frequência de Mutação	Aumento em vezes	Valor p
Pulmão	Controle de veículo	0,67e-07		
Pulmão	Uretano	5,04e-07	7,5x	6,73e-05
Baço	Controle de veículo	0,83e-07		
Baço	Uretano	2,73e-07	3,3x	1,92e-04
Sangue	Controle de veículo	1,11e-07		
Sangue	Uretano	2,39e-07	2,2x	0,003025

[00232] Com referência à FIG. 9 e Tabela 1 juntas, as diferenças entre os grupos controle de veículo (VC) e tratamento foram altamente significativas. Foi utilizado um teste t de Welch (para variações desiguais) para determinar a significância da frequência mutante do tecido tratado com mutagênico sobre a do controle para esse tecido. Os intervalos de confiança um pouco mais amplos com o sangue refletem uma profundidade média mais baixa de sequenciamento nas amostras de VC de sangue neste exemplo em particular. Prevê-se que isso possa ser corrigido usando os métodos aqui descritos.

[00233] A FIG. 10A são gráficos de caixa que mostram SNV MF calculado para Sequenciamento Duplex por regiões gênicas para Pulmão, Baço e Sangue para as categorias de tratamentos indicadas, e a FIG. 10B é um gráfico de dispersão mostrando medições individuais de dados agregados mostrados na FIG. 10A. Os pontos de dispersão mostram medições individuais com CI de 95% ao seu redor. O gráfico da caixa na FIG. 10A mostra todos os quatro quartis de todos os pontos de dados para essa categoria de tecido e tratamento. As escalas do eixo Y são apresentadas linearmente e na magnitude 10^{-7} . Com referência à FIG. 10A, o gráfico de caixa resume o agregado das frequências de mutação SNV no pulmão, baço e sangue do modelo de camundongo Tg-rasH2 mostrado na FIG. 10B. Não há transgene *cII* no modelo de camundongo Tg-rasH2. A extensão da indução de mutação é influenciada por mutagênico específico, tipo de tecido e locus genético. A FIG. 11 é um gráfico de barras que mostra o espectro de mutação de uretano e VC dentro dos tecidos testados, conforme medido por Sequenciamento Duplex. Com referência à FIG. 11, a análise de agrupamento não supervisionada de dados codificados permitiu o agrupamento de dados com base no espectro de mutações. Esses dados demonstram que o

simples espectro de variação de nucleotídeos sozinho pode identificar a exposição. Por outras palavras, se o mutagênico fosse desconhecido, poderia ser identificado *de novo* por via de Sequenciamento Duplex de DNA de um organismo exposto, por natureza do espectro da mutação.

[00234] As FIGS. **12A** e **12B** são gráficos que mostram espectros de mutação no contexto de nucleotídeos adjacentes (isto é, espectros de trinucleotídeos) para controle de veículo (12A) e uretano (12B). A assinatura mutacional no formato de espectro de trinucleotídeos fornece informações sobre diferentes mecanismos de mutagênese e/ou demonstra padrões mutacionais exclusivos para mutagênicos específicos. Consequentemente, a discriminação detalhada de cada classe de mutação dentro de seu contexto trinucleotídico ("assinatura triplete") revela uma impressão digital altamente exclusiva para cada grupo de tratamento, consistente com assinaturas conhecidas de mutações clonais de tumores causadas por tais exposições. Nos animais não tratados, foram detectadas mutações C:G → A:T e C:G → G:C causadas, respectivamente, por oxidação da guanina e desaminação da citosina e 5-me-citosina, que é um padrão conhecido desde o envelhecimento. Após o tratamento com uretano, T:A → A:T dentro do motivo "NTG" é mostrado como a mutação mais comum.

[00235] A FIG. **13** mostra que a tendência de fita da variante de nucleotídeo único (SNV) foi observada em regiões genômicas *Ctnnb1* e *Polr1c* mas não em *Hp* ou *Rho*. A notação SNV é normalizada para o nucleotídeo de referência na direção para frente da fita transcrita. As réplicas individuais são mostradas com pontos e intervalos de confiança de 95%, com segmentos de linha. Todas as frequências de mutação foram corrigidas para as contagens de nucleotídeos de cada base de referência dentro da região de chamada variante. A hipótese nula para tendência de ausência de fita é a mesma frequência para mutações recíprocas. A tendência é evidente em *Ctnnb1* e *Polr1c* pois as variantes C>N e T>N estão em frequências uniformes e as variantes G>N e A>N estão em frequências elevadas. Comparado a *Hp* e *Rho*, e sem estar limitado pela teoria, acredita-se que essa diferença se deva ao reparo por excisão de nucleotídeos acoplados à transcrição e aos níveis de expressão relativa desses genes.

[00236] A FIG. **14** é um gráfico que ilustra a seleção clonal neoplásica em estágio inicial de frações de alelos variantes, conforme detectado por Sequenciamento Duplex. A grande maioria das mutações identificadas ocorreu em moléculas únicas e em frações alélicas variantes muito baixas (VAFs), por exemplo, da ordem de 1/10.000. Algumas variantes foram encontradas

em múltiplas moléculas em uma amostra e foram identificadas como tendo VAFs consideravelmente mais altas.

[00237] A FIG. 15A é um gráfico que ilustra SNVs representados sobre os intervalos genômicos para os éxons capturados da família *Ras* de genes, incluindo os loci transgênicos humanos, no modelo de camundongo Tg-rasH2. Singuletos são mutações encontradas em uma única molécula. Multipletos são uma mutação idêntica identificada dentro de múltiplas moléculas dentro do mesmo amostrador e pode representar um evento de expansão clonal. A altura de cada ponto corresponde à frequência do alelo variante (VAF) de cada SNV, com o tamanho do ponto corresponde ao somete para observações de multipletos. A localização e a frequência relativa dos pontos críticos de mutação do câncer humano da família *Ras* no COSMIC são indicados abaixo de cada gene. A FIG. 15B é um gráfico que ilustra variantes de nucleotídeo único (SNVs) alinhando-se ao éxon 3 do transgene *HRAS* humano. O resíduo central destacado no códon número 61 no éxon 3 do *HRAS* humano, o ponto de acesso ativador de câncer *HRAS* mais comum.

[00238] Com referência às FIGS. 15A e 15B juntos, um agrupamento de transversões T>A foi observado em 4/5 amostras de pulmão tratadas com uretano e 1/5 amostras esplênicas tratadas com uretano no ponto de acesso de 61 códon de *Hras* oncogênico humano. Em particular, quatro em cada cinco amostras de pulmão tratadas abrigavam essa mutação em frequências variantes de alelos de 0,1% -1,8%. Notavelmente estes clones são da transversão T>A no contexto NTG, que é característico da mutagênese do uretano (referindo-se a forte favorecimento dos sítios NTG na FIG. 12B). Além disso, duas amostras de baço tratadas apresentaram mutações nesse códon: uma nesta mesma posição e outra em um par de bases adjacentes. A observação de que 4/5 amostras de pulmão tratadas tinham mutações patogênicas expandidas clonalmente até o dia 29, ao passo que muito poucas mutações vistas em outras partes no painel eram vistas como clones de >1 membro ou eram repetidas em múltiplas amostras (como multipletos de VAF alta em um ativador de câncer bem estabelecido) é uma forte indicação de seleção positiva logo após a exposição. Além disso, os métodos de Sequenciamento Duplex, de acordo com modalidades da presente tecnologia, fornecem a sensibilidade necessária para detectar essa seleção clonal neoplásica em estágio inicial.

Tabela 2

Contagem de mutação	Número de famílias	
1	829	
2	8	
4	1	
17	1	} Mutações AA 61 T>A oncogênicas em gene HRAS Humano em tecido de pulmão tratado com uretano
58	1	
181	1	
300	1	

[00239] Referindo-se à Tabela 2, 97,5% das mutações foram identificadas apenas em uma única molécula, 1% foram observadas em duas moléculas e cerca de 0,5% foram observadas em > 2 moléculas. Todos os quatro clones de nível mais alto ocorreram com mutação oncogênica no AA 61, o ponto de acesso recorrente do tumor no *HRAS* humano. O fato de os clones de nível mais alto também aparecerem nos pontos de acesso do câncer enfatizou ainda mais a magnitude da forte pressão seletiva.

[00240] Uma quantidade muito maior de DNA foi extraída por amostra do que foi convertida em Moléculas Duplex sequenciadas. A porção de amostras de tecido extraída produziu aproximadamente 5µg de DNA genômico. Converter isso em equivalentes de genoma e multiplicar por três produz o número de cópias de tg.HRAS na extração. Apenas ~1/3% disso foi sequenciado, e aproximadamente 300 vezes mais mutantes estavam presentes na porção original do tecido amostrado do que o detectado.

Tabela 3

Amostra	DNA ng	Genomas	Cópias tg.HRAS	Profundidade a AA61	% cópias sequenciadas	Mutantes	Células mutantes em amostra original
9957-Pulmão 1	5.640	1.692.000	5.076.000	16.425	0,324%	300	92.712
9958-Pulmão 1	4.400	1.320.000	3.960.000	16.319	0,412%	181	43.922
9959-Pulmão 1	4.480	1.344.000	4.032.000	13.692	0,340%	58	17.080
9961-Pulmão 1	4.700	1.410.000	4.230.000	14.706	0,348%	17	4.890

[00241] Neste exemplo, os clones selecionados abrangem mais de 90.000 células no clone da fração de alelo mais alto. Como resultado, por cálculo, dentro dos 29 dias do estudo, por exemplo, a partir do momento da exposição à mutação e presumindo que não há morte celular, o

tempo de duplicação dessas células era aproximadamente a cada 1,8 dias $2^{(29/1,8)} \sim 90.000$. Sem estar limitado pela teoria, essa taxa calculada de duplicação de células sugere a capacidade provável de detectar essas mutações selecionadas em um curto espaço de tempo (por exemplo, apenas duas semanas).

[00242] As FIGS. **16A-16B** são representações gráficas de dados de sequenciamento de uma seção representativa de 400 pares de bases de *HRAS* humano no pulmão de camundongo após tratamento com uretano usando sequenciamento de DNA convencional (FIG. **16A**) e Sequenciamento Duplex (FIG. **16B**). O sequenciamento convencional de DNA tem uma taxa de erro entre 0,1% e 1%, o que obscurece a presença de mutações genuínas de baixa frequência. A FIG. **16A** mostra dados de sequenciamento convencionais de uma seção representativa de 400 PA de um gene (*HRAS* humano) de uma amostra (pulmão de camundongo) no presente estudo. Cada barra corresponde a uma posição nucleotídica. A altura de cada barra corresponde à fração alélica das bases que não são de referência nessa posição quando sequenciadas com profundidade $> 100.000x$. Toda posição parece estar mutada em alguma frequência; quase todos esses são erros. Com referência à FIG. **16B**, quando processado com o Sequenciamento Duplex, torna-se aparente que apenas uma mutação é autêntica.

[00243] Os resultados da análise experimental deste exemplo demonstram que o Sequenciamento Duplex quantifica a indução de mutações pelo uretano de maneira extremamente robusta e com intervalos de confiança replicados apertados. Além disso, a extensão da indução de mutação é específica do tecido, sendo o pulmão mais propenso que o baço e o sangue. O simples espectro mutacional da exposição ao uretano é limpo e o agrupamento sem tendência pode discriminar entre os grupos. O espectro de mutação tripleto do uretano mostra uma forte propensão para mutações de $T \rightarrow A$ e $T \rightarrow C$ no contexto de "NTG" e o espectro de mutação é distinguível do controle de veículo (e outros mutagênicos; ver o Exemplo 1).

[00244] Além disso, a indução de mutação no sangue periférico espelhava de perto a observada no baço e sugere que a amostragem in vivo de sangue periférico poderia, para alguns mutagênicos, substituir a necropsia (ou biópsia). Além disso, este exemplo demonstrou que mesmo no dia 29, evidências claras de seleção para mutações oncogênicas no transgene *HRAS* humano são demonstradas usando o Sequenciamento Duplex. O espectro de mutação neste ponto de acesso refletia com precisão os efeitos desse mutagênico conhecido. Portanto, o Sequenciamento Duplex

pode fornecer dados precoces e precisos no que diz respeito à avaliação de mutações precoces ativadoras de câncer como biomarcador de risco futuro de câncer. A contaminação entre espécies persistiu em níveis extremamente baixos, mas a remoção da contaminação por espécies estranhas foi realizada de forma automática e confiável.

Exemplo 3

[00245] Análise de assinaturas de mutagênicos em genomas de mamíferos usando o Sequenciamento Duplex. Esta seção descreve um exemplo em que os dados gerados a partir da análise de Sequenciamento Duplex podem ser usados para gerar e comparar assinaturas mutagênicas para a identificação de mutagênicos e/ou para identificar uma exposição a mutagênicos.

[00246] O banco de dados do Catálogo de Mutações Somáticas em Câncer (COSMIC) fornece referência a "assinaturas mutacionais", definidas como a combinação única de tipos de mutação encontrados presentes no genoma. Mutações somáticas que estão presentes em todas as células do corpo humano e ocorrem ao longo da vida. Tais mutações somáticas são a consequência de, por exemplo, múltiplos processos mutacionais, incluindo a pequena infidelidade intrínseca do mecanismo de replicação de DNA, exposições exógenas ou endógenas a mutagênicos, modificação enzimática do DNA e reparo defeituoso do DNA.

[00247] As FIGS. **17A-17C** são gráficos que mostram espectros de mutação no contexto de nucleotídeos adjacentes (isto é, espectros de trinucleotídeos) para a Assinatura 1 (FIG. **17A**), Assinatura 4 (FIG. **17B**) e Assinatura 29 (FIG. **17C**) da COSMIC. Com referência à FIG. **17A**, a assinatura 1 é vista em todos os tipos de câncer com uma etiologia proposta de ser causada pela desaminação espontânea de 5-metil-citosina, resultando em transições C>T nos sítios CpG. Com referência às FIGS. **17B-17C**, as assinaturas 4 e 29 estão correlacionadas com o tabagismo e são impulsionadas por um importante mutagênico no tabaco: benzo[a]pireno. Embora de padrão semelhante, a assinatura 4 é mais frequentemente observada em cânceres de pulmão em fumantes, enquanto a assinatura 29 é vista predominantemente no câncer de esôfago escamoso, que é mais frequente em fumantes e usuários de tabaco de mascar.

Tabela 4

Recurso	Exemplo 1	Exemplo 2	Total
---------	-----------	-----------	-------

Modelo de Camundongo	BigBlue	Tg-rasH2	2 cepas
Tecidos (amostras)	Fígado (15) Medula (17)	Pulmão (10) Baço (10) Sangue (10)	5 tipos
Mutagênico (animais <i>por</i> grupo)	B[a]P (10) ENU (11) VC (11)	Uretano (15) VC (15)	3 mutagênicos
Amostras	32	30	62
Loci Endógenos	<i>Polr1c</i> <i>Rho</i> <i>Ctnnb1</i> <i>Hp</i>	<i>Polr1c</i> <i>Rho</i> <i>Ctnnb1</i> <i>Hp</i> <i>Hras</i> <i>Kras</i> <i>Nras</i>	7 genes nativos
Loci transgênicos	<i>cII</i>	<i>Hras</i> humanos	2 transgenes
BP duplex	4.074.604.194	4.348.868.321	8.423.472.515

[00248] A Tabela 4 fornece parâmetros experimentais e dados derivados dos Exemplos 1 e 2 discutidos aqui. A **FIG. 18** mostra agrupamento hierárquico não supervisionado de todas as 30 assinaturas COSMIC publicadas e os 4 espectros de corte dos Exemplos 1 e 2. O agrupamento foi realizado com o método ponderado (WGMA) e métrica de similaridade de cosseno. Notavelmente, o benzo[a]pireno (BaP) é muito semelhante às Assinaturas 4 e 29, que foram correlacionadas com a exposição ao BaP através do consumo ou inalação de tabaco. O controle de veículo (VC) é como a Assinatura 1, um padrão ligado à desaminação espontânea da 5-metil-citosina e acredita-se que represente uma mistura tanto do efeito mutagênico das espécies oxidativas reativas quanto da desaminação espontânea da 5-metil-citosina.

[00249] Este exemplo demonstra que o Sequenciamento Duplex pode ser usado para gerar análises de espectros de mutações que podem ser comparadas ou referenciadas a assinaturas mutacionais conhecidas para fins de identificação e outras análises.

Ambientes de computação adequados

[00250] A discussão a seguir fornece uma descrição geral de um ambiente de computação adequado no qual aspectos da divulgação podem ser implementados. Embora não sejam necessários, aspectos e modalidades da divulgação serão descritos no contexto geral de instruções executáveis por computador, como rotinas executadas por um computador de uso geral, por exemplo, um servidor ou computador pessoal. Os versados na técnica relevante apreciarão que a divulgação pode ser praticada com outras configurações de sistema de computador, incluindo dispositivos de Internet, dispositivos portáteis, computadores portáteis, telefones celulares ou móveis, sistemas multiprocessadores, eletrônicos de consumo programáveis ou baseados em microprocessadores, decodificadores, PCs de rede, minicomputadores, computadores mainframe e semelhantes. A divulgação pode ser incorporada em um computador para fins especiais ou processador de dados que é especificamente programado, configurado ou construído para executar uma ou mais das instruções executáveis por computador explicadas em detalhes abaixo. De fato, o termo "computador", como geralmente utilizado neste documento, se refere a qualquer um dos dispositivos acima, bem como a qualquer processador de dados.

[00251] A divulgação também pode ser praticada em ambientes de computação distribuídos, onde tarefas ou módulos são executados por dispositivos de processamento remoto, que são ligados por meio de uma rede de comunicações, como Rede de Área Local ("LAN"), Rede de Área Ampla ("WAN") ou a Internet. Em um ambiente de computação distribuída, os módulos do programa ou sub-rotinas podem estar localizados em dispositivos de armazenamento de memória locais e remotos. Aspectos da divulgação descrita abaixo podem ser armazenados ou distribuídos em meio legível por computador, incluindo discos magnéticos e opticamente legíveis e removíveis, armazenados como firmware em chips (por exemplo, chips EEPROM), bem como distribuídos eletronicamente pela Internet ou por outras redes (incluindo redes sem fio). Os versados na técnica relevante reconhecerão que porções da divulgação podem residir em um computador servidor, enquanto as porções correspondentes residem em um computador cliente.

As estruturas de dados e a transmissão de dados específicos para aspectos da divulgação também estão incluídas no escopo da divulgação.

[00252] Modalidades de computadores, como um computador pessoal ou estação de trabalho, podem compreender um ou mais processadores acoplados a um ou mais dispositivos de entrada do usuário e dispositivos de armazenamento de dados. Um computador também pode ser acoplado a pelo menos um dispositivo de saída, como um dispositivo de exibição e um ou mais dispositivos de saída adicionais opcionais (por exemplo, impressora, plotadora, alto-falantes, dispositivos de saída táteis ou olfativos, etc.). O computador pode ser acoplado a computadores externos, como uma conexão de rede opcional, um transceptor sem fio ou ambos.

[00253] Vários dispositivos de entrada podem incluir um teclado e/ou um dispositivo apontador, como um mouse. Outros dispositivos de entrada são possíveis, como microfone, joystick, caneta, tela sensível ao toque, scanner, câmera digital, câmera de vídeo e semelhantes. Outros dispositivos de entrada podem incluir máquinas de sequenciamento (por exemplo, sequenciador massivamente paralelo), fluoroscópios e outros equipamentos de laboratório, etc. Os dispositivos de armazenamento de dados adequados podem incluir qualquer tipo de meio legível por computador que pode armazenar dados acessíveis pelo computador, como unidades magnéticas de disco rígido e de disquete, unidades de disco óptico, cassetes magnéticas, unidades de fita, cartões de memória flash, discos de vídeo digital (DVDs), Cartuchos Bernoulli, RAMs, ROMs, cartões inteligentes etc. De fato, qualquer meio para armazenar ou transmitir instruções e dados legíveis por computador pode ser empregado, incluindo uma porta de conexão ou nó em uma rede como uma rede de área local (LAN), rede de área ampla (WAN) ou a Internet.

[00254] Aspectos da divulgação podem ser praticados em uma variedade de outros ambientes de computação. Por exemplo, um ambiente de computação distribuído com uma interface de rede inclui pode incluir um ou mais computadores de usuário em um sistema em que eles podem incluir um módulo de programa de navegador que permite ao computador acessar e trocar dados com a Internet, incluindo sites na porção World Wide Web da Internet. Os computadores dos usuários podem incluir outros módulos de programas, como um sistema operacional, um ou mais programas aplicativos (por exemplo, aplicativos de processamento de texto ou planilha) e semelhantes. Os computadores podem ser dispositivos de uso geral que podem ser programados para executar vários tipos de aplicativos ou podem ser dispositivos de uso único

otimizados ou limitados a uma função ou classe de funções específica. Mais importante, enquanto mostrado nos navegadores de rede, qualquer programa aplicativo para fornecer uma interface gráfica do usuário aos usuários pode ser empregado, conforme descrito em detalhes abaixo; o uso de um navegador da web e uma interface da web são usados apenas como um exemplo familiar neste documento.

[00255] Pelo menos um computador servidor, acoplado à Internet ou à World Wide Web (“Web”), pode executar muitas ou todas as funções para receber, rotear e armazenar mensagens eletrônicas, como páginas da Web, fluxos de dados, sinais de áudio e imagens eletrônicas aqui descritas. Enquanto a Internet é mostrada, uma rede privada, como uma intranet, pode ser preferida em alguns aplicativos. A rede pode ter uma arquitetura cliente-servidor, na qual um computador é dedicado a servir outros computadores clientes, ou pode ter outras arquiteturas, como ponto a ponto, nas quais um ou mais computadores servem simultaneamente como servidores e clientes. Um banco de dados ou bancos de dados, acoplados ao(s) computador(es) servidor(es), podem armazenar grande parte das páginas da web e do conteúdo trocado entre os computadores dos usuários. O(s) computador(es) do servidor, incluindo o(s) banco(s) de dados, pode(m) empregar medidas de segurança para inibir ataques maliciosos no sistema e preservar a integridade das mensagens e dados armazenados nele (por exemplo, sistemas de firewall, SSL (Secure Socket Layer), esquemas de proteção por senha, criptografia e semelhantes).

[00256] Um computador servidor adequado pode incluir um mecanismo de servidor, um componente de gerenciamento de páginas da web, um componente de gerenciamento de conteúdo e um componente de gerenciamento de banco de dados, entre outros recursos. O mecanismo do servidor executa tarefas básicas de processamento e nível de sistema operacional. O componente de gerenciamento de páginas da web lida com a criação e exibição ou roteamento de páginas da web. Os usuários podem acessar o computador servidor por meio de um URL associado a ele. O componente de gerenciamento de conteúdo lida com a maioria das funções nas modalidades descritas neste documento. O componente de gerenciamento de banco de dados inclui tarefas de armazenamento e recuperação com relação ao banco de dados, consultas ao banco de dados, funções de leitura e gravação no banco de dados e armazenamento de dados como sinais de vídeo, gráficos e áudio.

[00257] Muitas das unidades funcionais descritas aqui foram rotuladas como módulos, a fim de enfatizar mais particularmente sua independência de implementação. Por exemplo, os módulos podem ser implementados em software para execução por vários tipos de processadores. Um módulo identificado de código executável pode, por exemplo, compreender um ou mais blocos físicos ou lógicos de instruções de computador que podem, por exemplo, ser organizadas como um objeto, procedimento ou função. Os blocos identificados de instruções do computador não precisam ser localizados fisicamente juntos, mas podem compreender instruções díspares armazenadas em diferentes locais que, quando unidos logicamente, compreendem o módulo e atingem o objetivo declarado para o módulo.

[00258] Um módulo também pode ser implementado como um circuito de hardware que compreende circuitos VLSI personalizados ou matrizes de portas, semicondutores prontos para uso, como chips lógicos, transistores ou outros componentes distintos. Um módulo também pode ser implementado em dispositivos de hardware programáveis, como matrizes de portas programáveis em campo, lógica de matriz programável, dispositivos de lógica programável ou semelhantes.

[00259] Um módulo de código executável pode ser uma única instrução, ou muitas instruções, e pode até ser distribuído por vários segmentos de código diferentes, entre diferentes programas e por vários dispositivos de memória. Da mesma forma, os dados operacionais podem ser identificados e ilustrados aqui dentro de módulos e podem ser incorporados em qualquer forma adequada e organizados dentro de qualquer tipo adequado de estrutura de dados. Os dados operacionais podem ser coletados como um único conjunto de dados, ou podem ser distribuídos em diferentes locais, incluindo diferentes dispositivos de armazenamento, e podem existir, pelo menos parcialmente, apenas como sinais eletrônicos em um sistema ou rede.

Sistema para Teste Genotóxico

[00260] A presente invenção compreende ainda um sistema (por exemplo, um sistema de computador em rede, um sistema automatizado de alto rendimento, etc.) para processar a amostra de um sujeito e transmitir os dados de sequenciamento por meio de uma rede com ou sem fio a um servidor remoto para determinar as leituras de sequência corrigidas por erro (por exemplo, leituras de sequência duplex, sequência de consenso duplex, etc.), espectro de mutação, frequência

de mutante, assinatura de mutação triplete e se houver uma semelhança entre os dados da amostra e os dados correspondentes associados a uma ou mais genotoxinas conhecidas.

[00261] Como descrito em detalhes adicionais abaixo, e com relação à modalidade ilustrada na FIG. 19, um sistema computadorizado de genotoxina compreende: (1) um servidor remoto; (2) uma pluralidade de dispositivos de computação eletrônica do usuário capazes de gerar e/ou transmitir dados de sequenciamento; (3) um banco de dados com perfis conhecidos de genotoxina e informações associadas (opcional); e (4) uma rede com ou sem fio para transmitir comunicações eletrônicas entre os dispositivos de computação eletrônica, o banco de dados e o servidor remoto. O servidor remoto compreende ainda: (a) um banco de dados que armazena resultados de registros de genotoxinas do usuário e registros de perfis de genotoxinas (por exemplo, espectro, frequências, mecanismo de ações etc.); (b) um ou mais processadores acoplados comunicativamente a uma memória; e um ou mais dispositivos ou meios de armazenamento legíveis por computador não transitórios compreendendo instruções para o(s) processador(es), em que os referidos processadores estão configurados para executar as referidas instruções para executar operações compreendendo uma ou mais das etapas descritas nas FIGS. 20-23.

[00262] Em uma modalidade, a presente tecnologia compreende ainda, um meio de armazenamento legível por computador não transitório compreendendo instruções que, quando executadas por um ou mais processadores, executam um método para determinar se um sujeito está exposto e/ou a identidade ou propriedades/características de pelo menos uma genotoxina. Em modalidades particulares, os métodos podem incluir uma ou mais das etapas descritas nas FIGS. 20-23.

[00263] Aspectos adicionais da presente tecnologia são direcionados a métodos computadorizados para determinar se um sujeito está exposto e/ou a identidade ou propriedades/características de pelo menos uma genotoxina. Em modalidades particulares, os métodos podem incluir uma ou mais das etapas descritas nas FIGS. 20-23.

[00264] A FIG. 19 é um diagrama de blocos de um sistema de computador 1900 com um produto de programa de computador 1950 instalado no mesmo e para uso com os métodos e/ou kits aqui divulgados para identificar eventos mutagênicos e/ou eventos de danos aos ácidos nucleicos resultantes da exposição genotóxica. Embora a FIG. 19 ilustre vários componentes do sistema de computação, é contemplado que outros ou diferentes componentes conhecidos dos

versados na técnica, como os discutidos acima, podem fornecer um ambiente de computação adequado no qual aspectos da divulgação podem ser implementados. A FIG. 20 é um diagrama de fluxo que ilustra uma rotina para fornecer dados de sequência de consenso de Sequenciamento Duplex de acordo com uma modalidade da presente tecnologia. As FIGS. 21-23 são diagramas de fluxo que ilustram várias rotinas para identificar eventos mutagênicos e/ou eventos de dano de ácido nucleico resultantes da exposição genotóxica de uma amostra. De acordo com aspectos da presente tecnologia, métodos descritos em relação às FIGS. 21-23 podem fornecer dados de amostra, incluindo, por exemplo, espectro de mutação de amostra, frequência mutante, espectro de mutação tripleto e informações derivadas da comparação de dados de amostra com conjuntos de dados de genotoxinas conhecidas.

[00265] Como ilustrado na FIG. 19, o sistema de computador 1900 pode compreender uma pluralidade de dispositivos de computação de usuário 1902, 1904; uma rede com ou sem fio 1910 e um servidor remoto (servidor “DupSeq™”) 1940 compreendendo processadores para analisar eventos mutagênicos e/ou eventos de danos aos ácidos nucleicos resultantes da exposição genotóxica de uma amostra. Nas modalidades, os dispositivos de computação do usuário 1902, 1904 podem ser usados para gerar e/ou transmitir dados de sequenciamento. Em uma modalidade, os usuários dos dispositivos de computação 1902, 1904 podem ser aqueles que executam outros aspectos da presente tecnologia, como etapas do método de Sequenciamento Duplex de amostras de sujeitos para avaliar a genotoxicidade. Em um exemplo, os usuários dos dispositivos de computação 1902, 1904 executam certas etapas do método de Sequenciamento Duplex com um kit (1,2) compreendendo reagentes e/ou adaptadores, de acordo com uma modalidade da presente tecnologia, para interrogar amostras em questão.

[00266] Como ilustrado, cada dispositivo de computação de usuário 1902, 1904 inclui pelo menos uma unidade de processamento central 1906, uma memória 1907 e uma interface de usuário e rede 1908. Em uma modalidade, os dispositivos de usuário 1902, 1904 compreendem um desktop, laptop ou tablet.

[00267] Embora dois dispositivos de computação do usuário 1902, 1904 estejam representados, é contemplado que qualquer número de dispositivos de computação do usuário possa ser incluído ou conectado a outros componentes do sistema 1900. Além disso, os dispositivos de computação 1902, 1904 também podem ser representativos de uma pluralidade de

dispositivos e software usados pelo Usuário (1) e Usuário (2) para amplificar e sequenciar as amostras. Por exemplo, um dispositivo de computação pode uma máquina de sequenciamento (por exemplo, Illumina HiSeq™, Ion Torrent PGM, sequenciador ABI SOLiD™, PacBio RS, Helicos Heliscope™, etc.), uma máquina de PCR em tempo real (por exemplo, ABI 7900, Fluidigm BioMark™, etc.), um instrumento de microarranjo, etc.

[00268] Além dos componentes descritos acima, o sistema **1900** pode ainda compreender um banco de dados **1930** para armazenar perfis de genotoxinas e informações associadas. Por exemplo, o banco de dados **1930**, que pode ser acessível pelo servidor **1940**, pode compreender registros ou coleções de espectro de mutações, espectro/assinaturas de mutação triplete, mecanismo de ação etc. para uma pluralidade de genotoxinas conhecidas e também pode incluir informações adicionais em relação aos perfis/padrões de mutação de cada genotoxina armazenada. Em um exemplo particular, o banco de dados **1930** pode ser um banco de dados de terceiros compreendendo perfis de genotoxina **1932**. Por exemplo, o site do Catálogo de Mutações Somáticas em Câncer (COSMIC) compreende uma coleção de "espectros mutacionais" que foram encontrados como mutações clonais em tumores que surgiram da exposição a agentes cancerígenos, por exemplo, câncer de pulmão em fumantes [8,9]. Em outra modalidade, o banco de dados pode ser um banco de dados independente **1930** (privado ou não privado) hospedado separadamente do servidor **1940**, ou um banco de dados pode ser hospedado no servidor **1940**, como o banco de dados **1970**, que compreende perfis de genotoxina derivados empiricamente de **1972**. Em algumas modalidades, como o sistema **1900** é usado para gerar novos perfis de agente/fator de teste, os dados gerados a partir do uso do sistema **1900** e métodos associados (por exemplo, métodos aqui descritos e, por exemplo, nas FIGS. **20-23**), pode ser carregado no banco de dados **1930** e/ou **1970**, para que perfis adicionais de genotoxina **1932**, **1972** possam ser criados para futuras atividades de comparação.

[00269] O servidor **1940** pode ser configurado para receber, computar e analisar dados de sequenciamento (por exemplo, arquivos de sequenciamento brutos) e informações relacionadas dos dispositivos de computação do usuário **1902**, **1904** através da rede **1910**. Os dados brutos de sequenciamento específicos da amostra podem ser calculados localmente usando um produto/módulo de programa de computador (Módulo de Sequência **1905**) instalado nos dispositivos **1902,1904**, ou acessível a partir do servidor remoto **1940** via rede **1910**, ou usando

outro software de sequenciamento conhecido na técnica. Os dados brutos da sequência podem então ser transmitidos via rede **1910** para o servidor remoto **1940** e os resultados do usuário **1974** podem ser armazenados no banco de dados **1970**. O servidor **1940** também compreende o produto/módulo de programa "Módulo DS"; **1912** configurado para receber os dados de sequenciamento brutos do banco de dados **1970** e configurado para gerar computacionalmente leituras de sequência de fita dupla corrigidas por erro usando, por exemplo, técnicas de Sequenciamento Duplex aqui divulgadas. Enquanto o DS Module **1912** é mostrado no servidor **1940**, um versado na técnica reconheceria que o DS Module **1912**, pode alternativamente, ser hospedado em operado nos dispositivos **1902**, **1904** ou em outro servidor remoto (não mostrado).

[00270] O servidor remoto **1940** pode compreender pelo menos uma unidade central de processamento (CPU) **1960**, um usuário e uma interface de rede **1962**(ou dispositivo de computação dedicado ao servidor com a interface conectada ao servidor), um banco de dados **1970**, como descrito acima, com uma pluralidade de arquivos/registros de computador para armazenar perfis de mutação de genotoxinas conhecidas e novas **1972**, e arquivos/registros para armazenar resultados (por exemplo, dados brutos de sequenciamento, dados de Sequenciamento Duplex, análise de genotoxicidade, etc.) para amostras testadas **1974**. O servidor **1940** compreende ainda uma memória de computador **1911** tendo armazenado nela o Produto de Programa de Computador de Genotoxina (Módulo de Genotoxina **1950**, de acordo com aspectos da presente tecnologia.

[00271] O produto/módulo de programa de computador **1950** é incorporado em um meio legível por computador não transitório que, quando executado em um computador (por exemplo, servidor **1940**), executa etapas dos métodos divulgados aqui para detectar e identificar genotoxinas. Outro aspecto da presente divulgação compreende o produto/módulo de programa de computador **1950** compreendendo um meio não transitório utilizável por computador, com códigos ou instruções de programa legíveis por computador incorporados para permitir que um processador realize análise de genotoxicidade (por exemplo, calcular frequência de mutante, espectro de mutação, espectro de mutação tripleto, relatórios de comparação de genotoxinas, relatórios de limiares etc.). Essas instruções de programa de computador podem ser carregadas em um computador ou outro aparelho programável para produzir uma máquina, de modo que as instruções executadas no computador ou em outro aparelho programável criem meios para implementar as funções ou etapas descritas aqui. Estas instruções do programa de computador também podem ser

armazenadas em uma memória ou meio legível por computador que pode direcionar um computador ou outro aparelho programável para funcionar de uma maneira particular, de modo que as instruções armazenadas na memória ou meio legível por computador produzem um artigo de fabricação incluindo meios de instruções que implementam a análise. As instruções do programa de computador também podem ser carregadas em um computador ou outro aparelho programável para fazer uma série de etapas operacionais serem realizadas no computador ou outro aparelho programável para produzir um processo implementado pelo computador de modo que as instruções que executam no computador ou outro aparelho programável fornecem etapas para implementar as funções ou etapas descritas acima.

[00272] Além disso, o produto/módulo de programa de computador **1950** pode ser implementado em qualquer idioma e/ou navegadores adequados. Por exemplo, ele pode ser implementado com Python, linguagem C e, de preferência, usando linguagens de programação de alto nível orientadas a objetos, como Visual Basic, SmallTalk, C++ e semelhantes. O aplicativo pode ser gravado para se adequar a ambientes como o ambiente Microsoft Windows™, incluindo Windows™ 98, Windows™ 2000, Windows™ NT e semelhantes. Além disso, o aplicativo também pode ser escrito para o ambiente MacIntosh™, SUN™, UNIX ou LINUX. Além disso, as etapas funcionais também podem ser implementadas usando uma linguagem de programação universal ou independente de plataforma. Exemplos dessas linguagens de programação de plataforma múltipla incluem, entre outras, linguagem de marcação de hipertexto (HTML), JAVA™, JavaScript™, linguagem de programação Flash, interface de gateway comum/linguagem de consulta estruturada (CGI/SQL), linguagem prática de relatório de extração (PERL), AppleScript™ e outras linguagens de script do sistema, linguagem de programação/linguagem de consulta estruturada (PL/SQL) e semelhantes. Podem ser usados navegadores habilitados para HotJava™, Microsoft™ Explorer™, ou Netscape™. Quando páginas da web de conteúdo ativo são usadas, elas podem incluir miniaplicativos Java™ ou controles ActiveX™ ou outras tecnologias de conteúdo ativo.

[00273] O sistema chama várias rotinas. Enquanto algumas das rotinas são aqui descritas, um versado na técnica é capaz de identificar outras rotinas que o sistema poderia executar. Além disso, as rotinas descritas neste documento podem ser alteradas de várias maneiras.

Como exemplos, a ordem da lógica ilustrada pode ser reorganizada, as subetapas podem ser executadas em paralelo, a lógica ilustrada pode ser omitida, outra lógica pode ser incluída, etc.

[00274] As FIGS. **20-23** são diagramas de fluxo ilustrando as rotinas **2000, 2100, 2200, 2300** para detectar e identificar eventos mutagênicos e/ou eventos de danos aos ácidos nucleicos resultantes da exposição genotóxica de uma amostra. A FIG. **20** é um diagrama de fluxo que ilustra a rotina **2000** para fornecer Dados de Sequenciamento Duplex para moléculas de ácido nucleico de fita dupla em uma amostra (por exemplo, uma amostra de um ensaio de genotoxicidade). A rotina **2000** pode ser invocada por um dispositivo de computação, como um computador cliente ou um servidor, acoplado a uma rede de computadores. Numa modalidade, o dispositivo de computação inclui gerador de dados de sequência e/ou um módulo de sequência. Como um exemplo, o dispositivo de computação pode invocar a rotina **2000** após um operador envolver uma interface de usuário em comunicação com o dispositivo de computação.

[00275] A rotina **2000** começa no bloco **2002** e o módulo de sequência recebe dados de sequência bruta de um dispositivo de computação do usuário (bloco **2004**) e cria um conjunto de dados específicos de amostra compreendendo uma pluralidade de leituras de sequência bruta derivadas de uma pluralidade de moléculas de ácido nucleico na amostra (bloco **2006**). Em algumas modalidades, o servidor pode armazenar o conjunto de dados específico da amostra em um banco de dados para processamento posterior. Em seguida, o módulo DS recebe uma solicitação para gerar dados do Sequenciamento de Consenso Duplex a partir dos dados brutos da sequência no conjunto de dados específico da amostra (bloco **2008**). O módulo DS agrupa a leitura de sequências de famílias que representam uma molécula original de ácido nucleico de fita dupla (por exemplo, com base em sequências SMI) e compara sequências representativas de fitas individuais entre si (bloco **2010**). Numa modalidade, as sequências representativas podem ser uma ou mais de uma leitura de sequência de cada molécula de ácido nucleico original. Em outra modalidade, as sequências representativas podem ser sequências de consenso de fita única (SSCSs) geradas a partir do alinhamento e correção de erros dentro de fitas representativas. Em tais modalidades, um SSCS de uma primeira fita pode ser comparado a um SSCS de uma segunda fita.

[00276] No bloco **2012**, o módulo DS identifica posições nucleotídicas de complementaridade entre as fitas representativas comparadas. Por exemplo, o módulo DS identifica as posições nucleotídicas ao longo da leitura comparada (por exemplo, alinhada) da

sequência em que as chamadas da base nucleotídica estão de acordo. Além disso, o módulo DS identifica posições de não complementaridade entre as fitas representativas comparadas (bloco **2014**). Da mesma forma, o módulo DS pode identificar posições de nucleotídeos ao longo das leituras de sequência comparadas (por exemplo, alinhadas) onde as chamadas da base de nucleotídeos estão em desacordo.

[00277] Em seguida, o módulo DS pode fornecer Dados de Sequenciamento Duplex para moléculas de ácido nucleico de fita dupla em uma amostra (bloco **2016**). Esses dados podem estar na forma de sequências de consenso duplex para cada uma das leituras de sequência processadas. As sequências de consenso duplex podem incluir, em uma modalidade, apenas posições de nucleotídeos em que as sequências representativas formam cada fita de uma molécula de ácido nucleico original estão de acordo. Por conseguinte, em uma modalidade, as posições de desacordo podem ser eliminadas ou de outro modo descontadas, de modo que a sequência de consenso duplex seja uma leitura de sequência de alta que foi corrigida por erro. Em outra modalidade, os Dados de Sequenciamento Duplex podem incluir informações de relatórios sobre posições de nucleotídeos de desacordo, a fim de que essas posições possam ser analisadas posteriormente (por exemplo, nos casos em que os danos ao DNA podem ser avaliados). A rotina **2000** pode continuar no bloco **2018**, onde termina.

[00278] A FIG. **21** é um diagrama de fluxo que ilustra uma rotina **2100** para detectar e identificar eventos mutagênicos resultantes da exposição genotóxica de uma amostra. A rotina pode ser invocada pelo dispositivo de computação da FIG. **19**. A rotina **2100** começa no bloco **2102** e o módulo de genotoxina compara os Dados de Sequenciamento Duplex da FIG. **20** (por exemplo, após o bloco **2016**) para fazer referência às informações da sequência (bloco **2104**) e identifica mutações (por exemplo, onde a sequência do sujeito varia da sequência de referência) (bloco **2106**). Em seguida, o módulo de genotoxina determina uma frequência mutante (bloco **2108**) e gera um espectro de mutação (bloco **2110**) para a amostra. Como tal, uma análise de padrão de mutação pode ser fornecida com informações sobre o tipo, localização e frequência de eventos de mutação nas moléculas de ácido nucleico analisadas a partir da amostra. Opcionalmente, o módulo de genotoxina pode gerar um espectro de mutação triplete (bloco **2112**) fornecendo informações sobre o contexto e o padrão de trinucleotídeos para analisar o resultado genotóxico da exposição.

[00279] O módulo de genotoxina também pode opcionalmente comparar um espectro de mutação e/ou espectro de mutação triplete (se determinado) com uma pluralidade de conjuntos de dados de genotoxina conhecidos, como aqueles armazenados nos registros de perfil de genotoxina em um banco de dados (bloco **2114**) para determinar, por exemplo, se a amostra foi exposta a uma genotoxina conhecida ou em outro exemplo, para determinar se um agente/fator de teste tem um perfil genotóxico semelhante ao de uma genotoxina conhecida anteriormente. Opcionalmente, o módulo de genotoxina pode determinar um provável mecanismo de ação de uma genotoxina com base, em parte, nas informações de comparação (bloco **2116**). Em seguida, o módulo de genotoxina pode fornecer dados de genotoxicidade (bloco **2118**) que podem ser armazenados no conjunto de dados específico da amostra no banco de dados. Em algumas modalidades, não mostradas, os dados de genotoxicidade podem ser usados para gerar um perfil de genotoxina a ser armazenado no banco de dados para futuras atividades de comparação. A rotina **2100** pode então continuar no bloco **2120**, onde termina.

[00280] A FIG. **22** é um diagrama de fluxo que ilustra uma rotina **2200** para detectar e identificar eventos de dano ao DNA resultantes da exposição genotóxica de uma amostra. A rotina pode ser invocada pelo dispositivo de computação da FIG. **19**. A rotina **2200** começa no bloco **2014** da FIG. **20** e no bloco de decisão **2202**, a rotina **2200** determina se as posições nucleotídicas da não complementaridade são erros de processo. Em várias modalidades, os parâmetros para determinar se uma posição de desacordo entre as leituras da sequência de ambas as fitas de uma molécula de DNA original pode ser especificada por um operador, pelas características conhecidas de danos ao DNA, pelas características conhecidas dos erros do processo, por um número mínimo de leituras de sequência a incompatibilidade é representada por e assim por diante.

[00281] Se a posição do nucleotídeo for determinada como um erro do processo (em oposição a um sítio de dano ao DNA *in vivo* antes da extração do DNA), o módulo DS pode eliminar ou descontar essas posições nucleotídicas de não complementaridade (bloco **2204**). A rotina **2200** pode continuar para o bloco **2016** da FIG. **20**.

[00282] Voltando ao bloco de decisão **2202**, e se a posição do nucleotídeo for determinada como um erro do processo, o módulo de genotoxina pode identificar posições de não complementaridade como sítios de possível dano *in vivo* ao DNA (bloco **2206**), como resultante

da exposição a uma genotoxina. Após a identificação, o módulo de genotoxina pode gerar um relatório de danos ao DNA a ser associado ao conjunto de dados específicos da amostra no banco de dados (bloco **2208**). Em algumas modalidades, o relatório de danos ao DNA pode ser usado para inferir o mecanismo de ação de uma potencial genotoxina (não mostrada). A rotina **2200** pode continuar para o bloco **2016** da FIG. **20**.

[**00283**] A FIG. **23** é um diagrama de fluxo que ilustra uma rotina **2300** para detectar e identificar uma exposição a carcinógenos ou carcinógenos em um sujeito. A rotina **2300** pode ser invocada pelo dispositivo de computação da FIG. **19**. A rotina **2300** começa no bloco **2302** e o módulo de genotoxina recebe dados de Sequenciamento Duplex da FIG. **20** (por exemplo, após o bloco **2016**) e, opcionalmente, dados de genotoxicidade da FIG. **21** (por exemplo, após o bloco **2116**) e confirma que a amostra foi exposta a uma genotoxina (bloco **2304**). Em seguida, o módulo de genotoxina identifica variantes na sequência de uma região genômica alvo (por exemplo, gene) (bloco **2306**). Por exemplo, o módulo de genotoxina pode analisar Dados de Sequenciamento Duplex e dados de genotoxicidade em loci genéticos específicos (por exemplo, genes ativadores de câncer, oncogenes, etc.). Em seguida, o módulo de genotoxina calcula uma frequência de alelo variante (VAF) (bloco **2308**).

[**00284**] No bloco de decisão **2310**, a rotina **2300** determina se a VAF é maior em um grupo de teste do que em um grupo de controle. Se a VAF do grupo de teste não for maior que um grupo de controle, o módulo de genotoxina marca o agente para diminuir a suspeita de ser cancerígeno (bloco **2312**). A rotina **2300** pode então continuar no bloco **2314**, onde termina. Se a VAF for maior no grupo de teste do que no grupo de controle, a rotina **2300** continuará no bloco de decisão **2316**, onde a rotina **2300** determina se uma mutação é não singleto.

[**00285**] Se a mutação for singleto, o módulo de genotoxina caracteriza o agente com um nível médio de suspeita de ser carcinógeno (bloco **2318**). Se a mutação for determinada como não singleto (isto é, um múltiplo), a rotina continuará no bloco de decisão **2320**, em que a rotina **2300** determina se uma variante é detectada no gene alvo e se a variante é consistente com uma mutação ativadora (por exemplo, uma mutação conhecida por ativar o crescimento/transformação do câncer).

[**00286**] Se a mutação não for uma mutação ativadora, o módulo de genotoxina caracteriza o agente com um nível médio de suspeita de ser cancerígeno (bloco **2318**). Se a(s)

variante(s) são consistentes com uma mutação ativadora, o módulo de genotoxina caracteriza o agente com um alto nível de suspeita de ser um agente carcinógeno (bloco **2322**).

[00287] Para agentes que foram caracterizados com um nível médio de suspeita (no bloco **2318**) ou um alto nível de suspeita (no bloco **2318**), o módulo de genotoxina pode avaliar um limiar de segurança para o agente carcinógeno e/ou determinar um risco associado ao desenvolvimento de uma doença ou distúrbio associado à genotoxina após a exposição no sujeito (bloco **2324**). A rotina **2300** pode então continuar no bloco **2314**, onde termina.

[00288] Outras etapas e rotinas também são contempladas pela presente tecnologia. Por exemplo, o sistema (por exemplo, o módulo de genotoxina ou outro módulo) pode ser configurado para analisar os dados de genotoxina para determinar se um sujeito foi exposto a uma genotoxina, se um agente/fator de teste é genotóxico, determinar sob quais características uma genotoxina é mutagênica ou carcinogênica e semelhantes. Outras etapas podem incluir determinar se um sujeito deve ser tratado profilaticamente ou terapeuticamente com base nos dados de genotoxina derivados da amostra biológica de um sujeito em particular. Por exemplo, depois que as genotoxinas são identificadas usando o sistema, o servidor pode determinar se o sujeito foi exposto a mais do que um nível limiar seguro de genotoxina. Nesse caso, podem ser iniciados tratamentos de doenças profiláticas ou inibidoras.

Exemplos adicionais

1. Um método para detectar e quantificar mutações genômicas desenvolvidas *in vivo* em um sujeito após a exposição do sujeito a um mutagênico, compreendendo:

fornecer uma amostra do sujeito, em que a amostra compreende moléculas de DNA de fita dupla;

gerar uma leitura de sequência corrigida por erro para cada uma de uma pluralidade de moléculas de DNA de fita dupla na amostra, compreendendo:

gerar um conjunto de cópias de uma primeira fita original da molécula de DNA adaptador e um conjunto de cópias de uma segunda fita original da molécula de DNA-adaptador;

sequenciar o conjunto de cópias da primeira e da segunda fitas originais para fornecer uma sequência da primeira fita e uma sequência da segunda fita; e

comparar a sequência da primeira fita e a sequência da segunda fita para identificar uma ou mais correspondências entre a sequência da primeira e da segunda fita; e

analisar as uma ou mais correspondências para determinar um espectro de mutação para as moléculas de DNA de fita dupla na amostra.

2. O método do exemplo 1, compreendendo ainda calcular uma frequência mutante para as moléculas de DNA de fita dupla alvo, calculando o número de mutações únicas por par de bases duplex sequenciado.

3. O método do exemplo 1, em que as moléculas de DNA de fita dupla alvo foram extraídas do fígado, baço, sangue, pulmão ou medula óssea do sujeito.

4. O método do exemplo 1, em que o sujeito foi exposto ao mutagênico 30 dias ou menos antes das moléculas de DNA de fita dupla alvo serem removidas do sujeito.

5. O método do exemplo 1, em que o espectro de mutação é gerado por agrupamento hierárquico não supervisionado do espectro de mutação.

6. O método do exemplo 1, em que o espectro de mutação é um espectro de mutação triplete.

7. O método do exemplo 1, em que a geração de uma leitura de sequência corrigida por erro para cada uma de uma pluralidade de moléculas de DNA de fita dupla inclui gerar leituras de sequência corrigidas por erro de uma ou mais regiões genômicas direcionadas.

8. O método do exemplo 7, em que a uma ou mais regiões genômicas direcionadas é um sítio propenso à mutação no genoma.

9. O método do exemplo 7, em que a uma ou mais regiões genômicas direcionadas é um gene ativador de câncer conhecido.

10. O método do exemplo 1, em que o sujeito é um animal transgênico e em que pelo menos algumas das moléculas de DNA de fita dupla alvo incluem uma ou mais porções de um transgene.

11. O método do exemplo 1, em que o sujeito é um animal não transgênico e em que as moléculas de DNA de fita dupla alvo compreendem regiões genômicas endógenas.

12. O método do exemplo 1, em que o sujeito é um humano e em que as moléculas de DNA de fita dupla alvo são extraídas de uma coleta de sangue retirada do humano.

13. Um método para gerar uma assinatura mutagênica de um agente de teste, compreendendo:

fragmentos de DNA de sequenciamento duplex extraídos de um sujeito de teste exposto ao agente de teste; e

gerar uma assinatura mutagênica do agente de teste, compreendendo:

calcular uma frequência mutante para uma pluralidade de fragmentos de DNA, calculando o número de mutações únicas por par de bases duplex sequenciado; e

determinar um padrão de mutação para a pluralidade dos fragmentos de DNA, em que o padrão de mutação inclui o tipo de mutação, o contexto de trinucleotídeo de mutação e a distribuição genômica de mutações.

14. O método do exemplo 13, compreendendo ainda comparar a assinatura de mutação do agente de teste com assinaturas de mutação de uma ou mais genotoxinas conhecidas.

15. O método do exemplo 13, em que a assinatura de mutação do agente de teste varia com base em um ou mais de um tipo de tecido, um nível de exposição ao agente de teste, uma região genômica e um tipo de sujeito.

16. O método do exemplo 15, em que o tipo de sujeito é células humanas cultivadas em cultura.

17. O método do exemplo 13, em que o animal de teste foi exposto ao composto de teste 30 dias ou menos antes do animal ser sacrificado.

18. O método do exemplo 13, em que a assinatura mutagênica é gerada pela correspondência de padrões computacionais.

19. O método do exemplo 13, em que a assinatura de mutação é uma assinatura de mutação triplete.

20. O método do exemplo 13, em que os fragmentos de DNA de sequenciamento duplex incluem o sequenciamento duplex de uma ou mais regiões genômicas direcionadas.

21. O método do exemplo 20, em que a uma ou mais regiões genômicas direcionadas é um sítio propenso à mutação no genoma.

22. O método do exemplo 20, em que a uma ou mais regiões genômicas direcionadas é um gene ativador de câncer conhecido.

23. O método do exemplo 13, em que o animal de teste é um animal transgênico e em que pelo menos alguns dos fragmentos de DNA incluem uma ou mais porções de um transgene.

24. O método do exemplo 13, em que o animal de teste é um animal não transgênico e em que os fragmentos de DNA compreendem regiões genômicas endógenas.

25. Um método para avaliar um potencial genotóxico de um agente de teste, compreendendo:

(a) preparar uma biblioteca de sequenciamento a partir de uma amostra compreendendo uma pluralidade de fragmentos de DNA de fita dupla a partir de uma fonte biológica exposta ao agente de teste, em que a preparação da biblioteca de sequências compreende a ligação de moléculas adaptadoras assimétricas à pluralidade de fragmentos de DNA de fita dupla para gerar uma pluralidade de moléculas de DNA-adaptador;

(b) sequenciar a primeira e a segunda fitas das moléculas adaptador-DNA para fornecer uma leitura de sequência da primeira e uma segunda leitura de sequência da segunda fita para cada molécula adaptador-DNA;

(c) para cada molécula adaptador-DNA, comparar a leitura de sequência da primeira fita com a leitura de sequência da segunda fita para identificar uma ou mais correspondências entre as leituras de sequências da primeira e da segunda fitas; e

(d) determinar uma assinatura de mutação do agente de teste analisando as uma ou mais correspondências entre as leituras de sequências da primeira e da segunda fitas para cada uma das moléculas de DNA-adaptador para determinar pelo menos um de um padrão de mutação, um tipo de mutação, uma frequência mutante, uma distribuição do tipo de mutação e uma distribuição genômica de mutações na amostra; e

(e) comparar a assinatura de mutação do agente de teste com um espectro de pluralidade de mutação derivado de genotoxinas conhecidas para determinar se a assinatura de mutação é suficientemente semelhante a um espectro de mutação de uma genotoxina conhecida; ou

(f) avaliar se pelo menos uma das frequências mutantes, o tipo de mutação ou a distribuição do tipo de mutação está acima de um nível limiar seguro; ou

(g) determinar se a frequência mutante excede um limiar seguro de frequência mutante.

26. O método do exemplo 25, em que uma assinatura de mutação do agente de teste compreende uma frequência mutante acima de uma frequência limiar segura.

27. O método do exemplo 25, em que a assinatura de mutação do agente de teste compreende um padrão de mutação suficientemente semelhante ao padrão de mutação conhecido associado ao câncer.

28. O método do exemplo 25, em que a fonte biológica é pelo menos uma das células cultivadas em cultura, um animal, um humano, uma linhagem celular humana, um animal transgênico, um animal não transgênico, uma amostra de tecido humano ou uma amostra de sangue humano.

29. O método do exemplo 25, em que a fonte biológica foi exposta ao agente de teste 30 dias ou menos antes da extração da amostra compreendendo uma pluralidade de fragmentos de DNA de fita dupla.

30. O método do exemplo 25, em que a assinatura de mutação é uma assinatura de mutação triplete.

31. O método do exemplo 25, em que antes de comparar a leitura de sequência da primeira fita com a leitura de sequência da segunda fita, o método compreende associar a leitura de sequência da primeira fita com a leitura de sequência da segunda fita usando um ou mais de uma sequência adaptadora, comprimento de leitura de sequência e informações originais da fita.

32. O método do exemplo 25, em que antes de preparar a biblioteca de sequenciamento, o método compreende ainda expor a fonte biológica ao agente de teste.

33. O método do exemplo 32, em que antes de expor a fonte biológica ao agente de teste, a fonte biológica é ou compreende um tecido de câncer.

34. O método do exemplo 32, em que antes de expor a fonte biológica ao agente de teste, a fonte biológica é ou compreende um tecido saudável.

35. O método do exemplo 25, em que a amostra é ou compreende uma amostra de sangue.

36. O método do exemplo 25, em que a amostra é ou compreende uma linhagem de células cancerígenas.

37. O método do exemplo 25, em que a fonte biológica compreende células cancerígenas e em que a substância é testada quanto à genotoxicidade seletiva em pelo menos uma porção das células cancerígenas.

38. O método do exemplo 37, em que a substância é um composto terapêutico.

39. O método do exemplo 38, em que para a porção das células cancerígenas que se mostra sensível à genotoxicidade seletiva do composto terapêutico, o método compreende ainda determinar uma ou mais de uma frequência mutante e um espectro de mutação para a porção das células cancerígenas antes da exposição ao composto terapêutico.

40. O método do exemplo 25, em que o agente de teste compreende um alimento, uma droga, uma vacina, uma substância cosmética, um aditivo industrial, um subproduto industrial, destilado de petróleo, metal pesado, produto de limpeza doméstico, particulados transportados pelo ar, subproduto de fabricação, contaminante, plastificante, detergente, produto emissor de radiação, produto de tabaco, material químico ou material biológico.

41. Um método para determinar a exposição de um sujeito a um agente genotóxico, compreendendo:

comparar o espectro de mutação do DNA de sujeitos com espectros de mutação de compostos mutagênicos conhecidos; e

identificar os espectros de mutação de compostos mutagênicos conhecidos mais semelhantes ao espectro de mutação de DNA do sujeito.

42. O método do exemplo 41, em que o espectro de mutação do DNA do sujeito é avaliado por Sequenciamento Duplex.

43. O método do exemplo 41, em que o espectro de mutação do DNA do sujeito é gerado a partir do DNA extraído do sangue do paciente.

44. O método do exemplo 41, em que o espectro de mutação do DNA do sujeito é um espectro de mutação triplete.

45. O método do exemplo 41, compreendendo ainda sequenciar o DNA do sujeito para gerar o espectro de mutação do DNA do sujeito.

46. O método do exemplo 45, em que o sequenciamento do DNA do sujeito inclui o sequenciamento de um ou mais genes ativadores de câncer conhecidos.

47. Um kit capaz de ser usado no sequenciamento duplex corrigido por erro de polinucleotídeos de fita dupla para identificar genotoxinas, o kit, compreendendo:

pelo menos um conjunto de iniciadores de reação em cadeia de polimerase (PCR) e pelo menos um conjunto de moléculas adaptadoras, em que os iniciadores e moléculas adaptadoras podem ser utilizados em experimentos de sequenciamento duplex corrigido por erro; e

instruções sobre métodos de uso do kit na condução do sequenciamento duplex corrigido por erro do DNA extraído da amostra de um sujeito para identificar se o sujeito foi exposto a pelo menos uma genotoxina.

48. O kit do exemplo 47, em que o reagente compreende uma enzima de reparo do DNA.

49. O kit do exemplo 47, em que cada uma das moléculas adaptadoras no conjunto de moléculas adaptadoras compreende pelo menos uma sequência de identificador de molécula única (SMI) e pelo menos um elemento de definição de fita.

50. O kit do exemplo 47 compreende ainda um produto de programa de computador incorporado em um meio legível por computador não transitório que, quando executado em um computador, executa etapas para determinar uma leitura de sequenciamento duplex corrigida por erro para uma ou mais moléculas de DNA de fita dupla em uma amostra e determinar a frequência mutante, espectro de mutação e/ou espectro tripleto de pelo menos uma genotoxina usando a leitura de sequenciamento duplex corrigida por erro.

51. O kit do exemplo 50, em que o produto do programa de computador determina ainda o mecanismo de ação da genotoxina na mutação do DNA de um sujeito; e tratamentos terapêuticos ou profiláticos adequados para administração ao sujeito com base no mecanismo de ação da genotoxina.

52. Um método para diagnosticar e tratar um sujeito exposto a uma genotoxina, compreendendo:

a) determinar se um sujeito foi exposto a uma genotoxina por:

i) obter uma amostra biológica do sujeito;

ii) fornecer leituras de sequenciamento corrigidas por erro duplex para uma pluralidade de sequências de DNA de fita dupla extraídas da amostra;

iii) determinar a frequência mutante, espectro de mutação e/ou espectro de mutação tripleto das sequências de DNA;

iv) determinar se a frequência mutante, espectro de mutação e/ou espectro de mutação tripleto é indicativo de que o sujeito foi exposto a uma genotoxina;

b) se o sujeito tiver sido exposto à genotoxina, fornecer um tratamento profilático e/ou terapêutico para prevenir ou inibir o início de uma doença ou distúrbio associado à genotoxina.

53. Um método para identificar um nível limiar de exposição segura a uma genotoxina e fornecer tratamento, compreendendo:

- a) determinar o nível limiar de exposição segura de uma genotoxina;
- b) determinar se um sujeito foi exposto à genotoxina em um nível superior ao nível limiar de exposição segura por:
 - i) obter uma amostra biológica do sujeito;
 - ii) fornecer leituras de sequenciamento corrigidas por erro duplex para uma pluralidade de sequências de DNA de fita dupla extraídas da amostra biológica;
 - iii) determinar a frequência mutante, espectro de mutação e/ou espectro de mutação triplete das sequências de DNA;
 - iv) determinar se a frequência mutante, espectro de mutação e/ou espectro de mutação triplete são indicativos de que o sujeito foi exposto a uma genotoxina específica;
 - v) calcular o nível de exposição do sujeito à genotoxina com base na frequência mutante, espectro de mutação e/ou espectro de mutação triplete; e
- c) se o sujeito tiver sido exposto a mais que um nível limiar de exposição segura à genotoxina, fornecer um tratamento profilático e/ou terapêutico para prevenir ou inibir o início de uma doença ou distúrbio associado à genotoxina.

54. Um sistema para detectar e identificar eventos mutagênicos e/ou eventos de danos aos ácidos nucleicos resultantes da exposição genotóxica de uma amostra, compreendendo:

uma rede de computadores para transmissão de informações relacionadas a dados de sequenciamento e dados de genotoxicidade, em que as informações incluem um ou mais dados brutos de sequenciamento, dados de sequenciamento duplex, informações de amostra e informações de genotoxina;

um computador cliente associado a um ou mais dispositivos de computação do usuário e em comunicação com a rede de computadores;

um banco de dados conectado à rede de computadores para armazenar uma pluralidade de perfis de genotoxinas e registros de resultados do usuário;

um módulo de sequenciamento duplex em comunicação com a rede de computadores e configurado para receber dados brutos de sequenciamento e solicitações do computador cliente para gerar dados de sequenciamento duplex, leituras de sequências de grupos de famílias que

representam uma molécula original de ácido nucleico de fita dupla e comparar sequências representativas de fitas individuais uma com a outra para gerar dados de sequenciamento duplex; e

um módulo de genotoxina em comunicação com a rede de computadores e configurado para comparar dados de sequenciamento duplex para referenciar informações de sequência para identificar mutações e gerar dados de genotoxina compreendendo pelo menos uma dentre uma frequência mutante, um espectro de mutação e um espectro de mutação triplete.

55. O sistema do exemplo 54, em que os perfis de genotoxina compreendem espectro de mutação de genotoxina a partir de uma pluralidade de genotoxinas conhecidas.

56. Um meio de armazenamento não transitório, legível por computador, compreendendo instruções que, quando executadas por um ou mais processadores, executam um método de qualquer um dos exemplos 1-46 e 52-53 para determinar se um sujeito está exposto a pelo menos uma genotoxina e/ou determinar uma identidade de pelo menos uma genotoxina.

57. O meio de armazenamento legível por computador não transitório do exemplo 56, compreendendo ainda calcular o espectro de mutação, a frequência mutante e/ou o espectro de mutação triplete de um agente detectado, a partir do qual é determinada a identidade de pelo menos uma genotoxina.

58. Um sistema de computador para executar um método de qualquer um dos exemplos 1-46 e 52-53 para determinar se um sujeito está exposto e/ou uma identidade de pelo menos uma genotoxina, o sistema compreendendo: pelo menos um computador com um processador, memória, banco de dados e um meio de armazenamento legível por computador não transitório compreendendo instruções para o(s) processador(es), em que o(s) referido(s) processador(es) está configurado para executar as referidas instruções para executar operações compreendendo os métodos de qualquer um dos exemplos 1-53.

59. O sistema do exemplo 58, compreendendo ainda um sistema de computador em rede que compreende:

a. uma rede com ou sem fio;

b. uma pluralidade de dispositivos de computação eletrônica do usuário capazes de receber dados derivados do uso de um kit compreendendo reagentes para extrair, amplificar e produzir

uma sequência polinucleotídica da amostra de um sujeito e transmitir a sequência polinucleotídica via rede para um servidor remoto; e

c. um servidor remoto compreendendo o processador, a memória, o banco de dados e o meio de armazenamento legível por computador não transitório compreendendo instruções para o(s) processador(es), em que o(s) referido(s) processador(es) está(ão) configurado(s) para executar as referidas instruções para executar operações compreendendo os métodos de qualquer um dos exemplos 1-46 e 52-53; e

d. em que o referido servidor remoto é capaz de detectar e identificar eventos mutagênicos e/ou eventos de danos aos ácidos nucleicos resultantes da exposição genotóxica de uma amostra.

60. O sistema do exemplo 59, em que o banco de dados e/ou um banco de dados de terceiros acessíveis via rede, compreende ainda uma pluralidade de registros compreendendo um ou mais de um perfil de genotoxina de genotoxinas conhecidas, um perfil de genotoxina de pelo menos uma amostra do sujeito, e em que o perfil de genotoxina compreende uma mutação ou um sítio de dano ao DNA.

61. Um meio legível por computador não transitório, cujo conteúdo faz com que pelo menos um computador execute um método para fornecer dados de sequenciamento duplex para moléculas de ácido nucleico de fita dupla em uma amostra de um ensaio de triagem de genotoxicidade, o método compreendendo:

receber dados de sequência bruta de um dispositivo de computação do usuário; e

criar um conjunto de dados específicos de amostra compreendendo uma pluralidade de leituras de sequência bruta derivadas de uma pluralidade de moléculas de ácido nucleico na amostra;

as leituras de sequência de agrupamento de famílias que representam uma molécula original de ácido nucleico de fita dupla, em que o agrupamento é baseado em uma sequência de identificador de molécula única compartilhada;

comparar uma sequência da primeira fita e a sequência da segunda fita a partir de uma molécula original de ácido nucleico de fita dupla para identificar uma ou mais correspondências entre as leituras de sequência da primeira e da segunda fitas; e

fornecer dados de sequenciamento duplex para as moléculas de ácido nucleico de fita dupla na amostra.

62. O meio legível por computador do exemplo 61, compreendendo ainda identificar posições nucleotídicas de não complementaridade entre a primeira e a segunda sequências comparadas, em que o método compreende ainda:

em posições de não complementaridade, identificar e eliminar ou descontar os erros do processo; e

em posições de não complementaridade que não são identificadas como erros de processo, identificar as posições restantes de não complementaridade como sítios de possíveis danos in vivo ao DNA resultantes da exposição a uma genotoxina.

63. Um meio legível por computador não transitório, cujo conteúdo faz com que pelo menos um computador execute um método para detectar e identificar eventos mutagênicos resultantes da exposição genotóxica de uma amostra, o método compreendendo:

comparar dados de sequência duplex para fazer referência a informações de sequência;

identificar mutações nos dados da sequência duplex, em que uma mutação é identificada como uma região de não concordância com a informação de referência;

determinar uma frequência mutante nos dados da sequência duplex;

gerar um espectro de mutação a partir dos dados da sequência duplex;

gerar um espectro de mutação tripleto a partir dos dados da sequência duplex; e

comparar o espectro de mutação e/ou o espectro de mutação tripleto a uma pluralidade de conjuntos de dados de genotoxina conhecidos.

64. Um meio legível por computador não transitório, cujo conteúdo faz com que pelo menos um computador execute um método para detectar e identificar uma exposição a carcinógenos ou carcinógenos em um sujeito, o método compreendendo:

identificar variantes de sequência em uma região genômica alvo usando dados de sequenciamento duplex gerados a partir de uma amostra do sujeito;

calcular uma frequência de alelo variante (VAF) de uma amostra de teste e uma amostra de controle;

determinar se uma VAF é maior em um grupo de teste do que em um grupo de controle;

em amostras com uma VAF mais alta, determinar se uma variante de sequência é um não singleto;

em amostras com uma VAF mais alta, determinar se a variante de sequência é uma mutação ativadora; e

caracterizar amostras com uma mutação não singleto e/ou ativadora como suspeita de ser um carcinógeno.

65. Um meio legível por computador não transitório do exemplo 64, compreendendo ainda avaliar um limiar de segurança para o carcinógeno e/ou determinar um risco associado ao desenvolvimento de uma doença ou distúrbio associado à genotoxina após a exposição no sujeito.

Referências

[00289] As referências listadas abaixo, assim como as patentes e os pedidos de patentes publicados citados no relatório descritivo acima, são incorporados por referência em sua totalidade, como se aqui totalmente estabelecido.

[1] Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, and Loeb LA. Detection of ultra-rare mutations by next-generation sequencing. *Proc Natl Acad Sci U S A*. 2012; 109(36): 14508–14513.

[2] Kennedy SR, Salk JJ, Schmitt MW, Loeb LA. Ultra-Sensitive Sequencing Reveals an Age-Related Increase in Somatic Mitochondrial Mutations that are inconsistent with oxidative damage. *PLOS Genetics*. 2013; 9(9): 1-10.

[3] Kennedy SR, Schmitt MW, Fox EJ, Kohn BF, Salk JJ, Ahn EH, et al. Detecting ultralow-frequency mutations by Duplex Sequencing. *Nat Protoc*. 2014; 9(11): 2586–2606.

[4] Schmitt MW, Fox EJ, Prindle MJ, Reid-Bayliss KS, True LD, et al. Sequencing small genomic targets with high efficiency and extreme accuracy. *Nature Methods*. 2015; 12(5): 423-5.

[5] Chan CY, Huang PH, Guo F, Ding X, Kapur V, Mai J D, et al. Accelerating drug discovery via organs-on-chips. *Lab Chip*. 2013; 12(24): 4697-4710.

[6] Schmitt MW, Loeb LA, and Salk JJ. The influence of subclonal resistance mutations on targeted cancer therapy. *Nat Rev Clin Oncol*. 2016; 13(6): 335–347.

[7] Salk JJ, Schmitt MW, Loeb L A. Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations. *Nature Reviews Genetics*. 2018. 19:269-283.

Conclusão

[00290] As descrições detalhadas acima de modalidades da tecnologia não pretendem ser exaustivas ou limitar a tecnologia à forma precisa divulgada acima. Embora modalidades específicas e exemplos para a tecnologia sejam descritos acima para fins ilustrativos, várias modificações equivalentes são possíveis dentro do escopo da tecnologia, como reconhecerão os versados na técnica. Por exemplo, enquanto as etapas são apresentadas em uma determinada ordem, modalidades alternativas podem executar etapas em uma ordem diferente. As várias modalidades descritas neste documento também podem ser combinadas para fornecer outras modalidades. Todas as referências aqui citadas são incorporadas por referência como se aqui fossem totalmente estabelecidas.

[00291] Pelo exposto, será apreciado que modalidades específicas da tecnologia foram descritas aqui para fins de ilustração, mas estruturas e funções conhecidas não foram mostradas ou descritas em detalhes para evitar obscurecer desnecessariamente a descrição das modalidades da tecnologia. Onde o contexto permitir, os termos no singular ou no plural também podem incluir o termo no plural ou no singular, respectivamente.

[00292] Além disso, a menos que a palavra “ou” seja expressamente limitada a significar apenas um único item exclusivo dos outros itens em referência a uma lista de dois ou mais itens, o uso de “ou” nessa lista deve ser interpretado como incluindo (a) qualquer item único da lista, (b) todos os itens da lista ou (c) qualquer combinação dos itens da lista. Além disso, o termo "compreendendo" é usado para significar a inclusão de pelo menos o(s) recurso(s) recitado(s), de modo que qualquer número maior do mesmo recurso e/ou tipos adicionais de outros recursos não sejam excluídos. Também será apreciado que modalidades específicas foram descritas aqui para fins de ilustração, mas que várias modificações podem ser feitas sem se desviar da tecnologia. Além disso, embora as vantagens associadas a certas modalidades da tecnologia tenham sido descritas no contexto dessas modalidades, outras modalidades também podem exibir essas vantagens, e nem todas as modalidades precisam necessariamente exibir essas vantagens para se enquadrarem no escopo da tecnologia. Por conseguinte, a divulgação e a tecnologia associada podem abranger outras modalidades não expressamente mostradas ou descritas aqui.

[00293] Os nomes dos produtos usados nesta divulgação são apenas para fins de identificação. Todas as marcas comerciais são de propriedade de seus respectivos donos.

REIVINDICAÇÕES

1. Método para detectar e quantificar mutações genômicas desenvolvidas *in vivo* em um sujeito após a exposição do sujeito a um mutagênico, caracterizado pelo fato de que compreende:

fornecer uma amostra do sujeito, em que a amostra compreende moléculas de DNA de fita dupla;

gerar uma leitura de sequência corrigida por erro para cada uma de uma pluralidade de moléculas de DNA de fita dupla na amostra, compreendendo:

gerar um conjunto de cópias de uma primeira fita original da molécula de DNA adaptador e um conjunto de cópias de uma segunda fita original da molécula de DNA-adaptador;

sequenciar o conjunto de cópias da primeira e da segunda fitas originais para fornecer uma sequência da primeira fita e uma sequência da segunda fita; e

comparar a sequência da primeira fita e a sequência da segunda fita para identificar uma ou mais correspondências entre a sequência da primeira e da segunda fita; e

analisar as uma ou mais correspondências para determinar um espectro de mutação para as moléculas de DNA de fita dupla na amostra.

2. Método, de acordo com a reivindicação 1, caracterizado pelo fato de que compreende ainda calcular uma frequência mutante para as moléculas de DNA de fita dupla alvo, calculando o número de mutações únicas por par de bases duplex sequenciado.

3. Método, de acordo com a reivindicação 1, caracterizado pelo fato de que as moléculas de DNA de fita dupla alvo foram extraídas do fígado, baço, sangue, pulmão ou medula óssea do sujeito.

4. Método, de acordo com a reivindicação 1, caracterizado pelo fato de que o sujeito foi exposto ao mutagênico 30 dias ou menos antes das moléculas de DNA de fita dupla alvo serem removidas do sujeito.

5. Método, de acordo com a reivindicação 1, caracterizado pelo fato de que o espectro de mutação é gerado por agrupamento hierárquico não supervisionado do espectro de mutação.

6. Método, de acordo com a reivindicação 1, caracterizado pelo fato de que o espectro de mutação é um espectro de mutação tripleto.

7. Método, de acordo com a reivindicação 1, caracterizado pelo fato de que a geração de uma leitura de sequência corrigida por erro para cada uma de uma pluralidade de moléculas de DNA de fita dupla inclui gerar leituras de sequência corrigidas por erro de uma ou mais regiões genômicas direcionadas.

8. Método, de acordo com a reivindicação 7, caracterizado pelo fato de que a uma ou mais regiões genômicas direcionadas é um sítio propenso à mutação no genoma.

9. Método, de acordo com a reivindicação 7, caracterizado pelo fato de que a uma ou mais regiões genômicas direcionadas é um gene ativador de câncer conhecido.

10. Método, de acordo com a reivindicação 1, caracterizado pelo fato de que o sujeito é um animal transgênico e em que pelo menos algumas das moléculas de DNA de fita dupla alvo incluem uma ou mais porções de um transgene.

11. Método, de acordo com a reivindicação 1, caracterizado pelo fato de que o sujeito é um animal não transgênico e em que as moléculas de DNA de fita dupla alvo compreendem regiões genômicas endógenas.

12. Método, de acordo com a reivindicação 1, caracterizado pelo fato de que o sujeito é um humano e em que as moléculas de DNA de fita dupla alvo são extraídas de uma coleta de sangue retirada do humano.

13. Método para gerar uma assinatura mutagênica de um agente de teste, caracterizado pelo fato de que compreende:

fragmentos de DNA de sequenciamento duplex extraídos de um sujeito de teste exposto ao agente de teste; e

gerar uma assinatura mutagênica do agente de teste, compreendendo:

calcular uma frequência mutante para uma pluralidade de fragmentos de DNA, calculando o número de mutações únicas por par de bases duplex sequenciado; e

determinar um padrão de mutação para a pluralidade dos fragmentos de DNA, em que o padrão de mutação inclui o tipo de mutação, o contexto de trinucleotídeo de mutação e a distribuição genômica de mutações.

14. Método, de acordo com a reivindicação 13, caracterizado pelo fato de que compreende ainda comparar a assinatura de mutação do agente de teste com assinaturas de mutação de uma ou mais genotoxinas conhecidas.

15. Método, de acordo com a reivindicação 13, caracterizado pelo fato de que a assinatura de mutação do agente de teste varia com base em um ou mais de um tipo de tecido, um nível de exposição ao agente de teste, uma região genômica e um tipo de sujeito.

16. Método, de acordo com a reivindicação 15, caracterizado pelo fato de que o tipo de sujeito é células humanas cultivadas em cultura.

17. Método, de acordo com a reivindicação 13, caracterizado pelo fato de que o animal de teste foi exposto ao composto de teste 30 dias ou menos antes do animal ser sacrificado.

18. Método, de acordo com a reivindicação 13, caracterizado pelo fato de que a assinatura mutagênica é gerada pela correspondência de padrões computacionais.

19. Método, de acordo com a reivindicação 13, caracterizado pelo fato de que a assinatura de mutação é uma assinatura de mutação triplete.

20. Método, de acordo com a reivindicação 13, caracterizado pelo fato de que os fragmentos de DNA de sequenciamento duplex incluem o sequenciamento duplex de uma ou mais regiões genômicas direcionadas.

21. Método, de acordo com a reivindicação 20, caracterizado pelo fato de que a uma ou mais regiões genômicas direcionadas é um sítio propenso à mutação no genoma.

22. Método, de acordo com a reivindicação 20, caracterizado pelo fato de que a uma ou mais regiões genômicas direcionadas é um gene ativador de câncer conhecido.

23. Método, de acordo com a reivindicação 13, caracterizado pelo fato de que o animal de teste é um animal transgênico e em que pelo menos alguns dos fragmentos de DNA incluem uma ou mais porções de um transgene.

24. Método, de acordo com a reivindicação 13, caracterizado pelo fato de que o animal de teste é um animal não transgênico e em que os fragmentos de DNA compreendem regiões genômicas endógenas.

25. Método para avaliar um potencial genotóxico de um agente de teste, caracterizado pelo fato de que compreende:

(a) preparar uma biblioteca de sequenciamento a partir de uma amostra compreendendo uma pluralidade de fragmentos de DNA de fita dupla a partir de uma fonte biológica exposta ao agente de teste, em que a preparação da biblioteca de sequências compreende a ligação de

moléculas adaptadoras assimétricas à pluralidade de fragmentos de DNA de fita dupla para gerar uma pluralidade de moléculas de DNA-adaptador;

(b) sequenciar a primeira e a segunda fitas das moléculas adaptador-DNA para fornecer uma leitura de sequência da primeira e uma segunda leitura de sequência da segunda fita para cada molécula adaptador-DNA;

(c) para cada molécula adaptador-DNA, comparar a leitura de sequência da primeira fita com a leitura de sequência da segunda fita para identificar uma ou mais correspondências entre as leituras de sequências da primeira e da segunda fitas; e

(d) determinar uma assinatura de mutação do agente de teste analisando as uma ou mais correspondências entre as leituras de sequências da primeira e da segunda fitas para cada uma das moléculas de DNA-adaptador para determinar pelo menos um de um padrão de mutação, um tipo de mutação, uma frequência mutante, uma distribuição do tipo de mutação e uma distribuição genômica de mutações na amostra; e

(e) comparar a assinatura de mutação do agente de teste com um espectro de pluralidade de mutação derivado de genotoxinas conhecidas para determinar se a assinatura de mutação é suficientemente semelhante a um espectro de mutação de uma genotoxina conhecida; ou

(f) avaliar se pelo menos uma das frequências mutantes, o tipo de mutação ou a distribuição do tipo de mutação está acima de um nível limiar seguro; ou

(g) determinar se a frequência mutante excede um limiar seguro de frequência mutante.

26. Método, de acordo com a reivindicação 25, caracterizado pelo fato de que uma assinatura de mutação do agente de teste compreende uma frequência mutante acima de uma frequência limiar segura.

27. Método, de acordo com a reivindicação 25, caracterizado pelo fato de que a assinatura de mutação do agente de teste compreende um padrão de mutação suficientemente semelhante ao padrão de mutação conhecido associado ao câncer.

28. Método, de acordo com a reivindicação 25, caracterizado pelo fato de que a fonte biológica é pelo menos uma das células cultivadas em cultura, um animal, um humano, uma linhagem celular humana, um animal transgênico, um animal não transgênico, uma amostra de tecido humano ou uma amostra de sangue humano.

29. Método, de acordo com a reivindicação 25, caracterizado pelo fato de que a fonte biológica foi exposta ao agente de teste 30 dias ou menos antes da extração da amostra compreendendo uma pluralidade de fragmentos de DNA de fita dupla.

30. Método, de acordo com a reivindicação 25, caracterizado pelo fato de que a assinatura de mutação é uma assinatura de mutação triplete.

31. Método, de acordo com a reivindicação 25, caracterizado pelo fato de que antes de comparar a leitura de sequência da primeira fita com a leitura de sequência da segunda fita, o método compreende associar a leitura de sequência da primeira fita com a leitura de sequência da segunda fita usando um ou mais de uma sequência adaptadora, comprimento de leitura de sequência e informações originais da fita.

32. Método, de acordo com a reivindicação 25, caracterizado pelo fato de que antes de preparar a biblioteca de sequenciamento, o método compreende ainda expor a fonte biológica ao agente de teste.

33. Método, de acordo com a reivindicação 32, caracterizado pelo fato de que antes de expor a fonte biológica ao agente de teste, a fonte biológica é ou compreende um tecido de câncer.

34. Método, de acordo com a reivindicação 32, caracterizado pelo fato de que antes de expor a fonte biológica ao agente de teste, a fonte biológica é ou compreende um tecido saudável.

35. Método, de acordo com a reivindicação 25, caracterizado pelo fato de que a amostra é ou compreende uma amostra de sangue.

36. Método, de acordo com a reivindicação 25, caracterizado pelo fato de que a amostra é ou compreende uma linhagem de células cancerígenas.

37. Método, de acordo com a reivindicação 25, caracterizado pelo fato de que a fonte biológica compreende células cancerígenas e em que a substância é testada quanto à genotoxicidade seletiva em pelo menos uma porção das células cancerígenas.

38. Método, de acordo com a reivindicação 37, caracterizado pelo fato de que a substância é um composto terapêutico.

39. Método, de acordo com a reivindicação 38, caracterizado pelo fato de que para a porção das células cancerígenas que se mostra sensível à genotoxicidade seletiva do composto terapêutico, o método compreende ainda determinar uma ou mais de uma frequência mutante e

um espectro de mutação para a porção da células cancerígenas antes da exposição ao composto terapêutico.

40. Método, de acordo com a reivindicação 25, caracterizado pelo fato de que o agente de teste compreende um alimento, uma droga, uma vacina, uma substância cosmética, um aditivo industrial, um subproduto industrial, destilado de petróleo, metal pesado, produto de limpeza doméstico, particulados transportados pelo ar, subproduto de fabricação, contaminante, plastificante, detergente, produto emissor de radiação, produto de tabaco, material químico ou material biológico.

41. Método para determinar a exposição de um sujeito a um agente genotóxico, caracterizado pelo fato de que compreende:

comparar o espectro de mutação do DNA de sujeitos com espectros de mutação de compostos mutagênicos conhecidos; e

identificar os espectros de mutação de compostos mutagênicos conhecidos mais semelhantes ao espectro de mutação de DNA do sujeito.

42. Método, de acordo com a reivindicação 41, caracterizado pelo fato de que o espectro de mutação do DNA do sujeito é avaliado por Sequenciamento Duplex.

43. Método, de acordo com a reivindicação 41, caracterizado pelo fato de que o espectro de mutação do DNA do sujeito é gerado a partir do DNA extraído do sangue do paciente.

44. Método, de acordo com a reivindicação 41, caracterizado pelo fato de que o espectro de mutação do DNA do sujeito é um espectro de mutação triplete.

45. Método, de acordo com a reivindicação 41, caracterizado pelo fato de que compreende ainda sequenciar o DNA do sujeito para gerar o espectro de mutação do DNA do sujeito.

46. Método, de acordo com a reivindicação 45, caracterizado pelo fato de que o sequenciamento do DNA do sujeito inclui o sequenciamento de um ou mais genes ativadores de câncer conhecidos.

47. Kit capaz de ser usado no sequenciamento duplex corrigido por erro de polinucleotídeos de fita dupla para identificar genotoxinas, o kit, caracterizado pelo fato de que compreende:

pelo menos um conjunto de iniciadores de reação em cadeia de polimerase (PCR) e pelo menos um conjunto de moléculas adaptadoras, em que os iniciadores e moléculas adaptadoras podem ser utilizados em experimentos de sequenciamento duplex corrigido por erro; e

instruções sobre métodos de uso do kit na condução do sequenciamento duplex corrigido por erro do DNA extraído da amostra de um sujeito para identificar se o sujeito foi exposto a pelo menos uma genotoxina.

48. Kit, de acordo com a reivindicação 47, caracterizado pelo fato de que o reagente compreende uma enzima de reparo do DNA.

49. Kit, de acordo com a reivindicação 47, caracterizado pelo fato de que cada uma das moléculas adaptadoras no conjunto de moléculas adaptadoras compreende pelo menos uma sequência de identificador de molécula única (SMI) e pelo menos um elemento de definição de fita.

50. Kit, de acordo com a reivindicação 47, caracterizado pelo fato de que compreende ainda um produto de programa de computador incorporado em um meio legível por computador não transitório que, quando executado em um computador, executa etapas para determinar uma leitura de sequenciamento duplex corrigida por erro para uma ou mais moléculas de DNA de fita dupla em uma amostra e determinar a frequência mutante, espectro de mutação e/ou espectro tripleto de pelo menos uma genotoxina usando a leitura de sequenciamento duplex corrigida por erro.

51. Kit, de acordo com a reivindicação 50, caracterizado pelo fato de que o produto do programa de computador determina ainda o mecanismo de ação da genotoxina na mutação do DNA de um sujeito; e tratamentos terapêuticos ou profiláticos adequados para administração ao sujeito com base no mecanismo de ação da genotoxina.

52. Método para diagnosticar e tratar um sujeito exposto a uma genotoxina, caracterizado pelo fato de que compreende:

a) determinar se um sujeito foi exposto a uma genotoxina por:

i) obter uma amostra biológica do sujeito;

ii) fornecer leituras de sequenciamento corrigidas por erro duplex para uma pluralidade de sequências de DNA de fita dupla extraídas da amostra;

iii) determinar a frequência mutante, espectro de mutação e/ou espectro de mutação triplete das sequências de DNA;

iv) determinar se a frequência mutante, espectro de mutação e/ou espectro de mutação triplete é indicativo de que o sujeito foi exposto a uma genotoxina;

b) se o sujeito tiver sido exposto à genotoxina, fornecer um tratamento profilático e/ou terapêutico para prevenir ou inibir o início de uma doença ou distúrbio associado à genotoxina.

53. Método para identificar um nível limiar de exposição segura a uma genotoxina e fornecer tratamento, caracterizado pelo fato de que compreende:

a) determinar o nível limiar de exposição segura de uma genotoxina;

b) determinar se um sujeito foi exposto à genotoxina em um nível superior ao nível limiar de exposição segura por:

i) obter uma amostra biológica do sujeito;

ii) fornecer leituras de sequenciamento corrigidas por erro duplex para uma pluralidade de sequências de DNA de fita dupla extraídas da amostra biológica;

iii) determinar a frequência mutante, espectro de mutação e/ou espectro de mutação triplete das sequências de DNA;

iv) determinar se a frequência mutante, espectro de mutação e/ou espectro de mutação triplete são indicativos de que o sujeito foi exposto a uma genotoxina específica;

v) calcular o nível de exposição do sujeito à genotoxina com base na frequência mutante, espectro de mutação e/ou espectro de mutação triplete; e

c) se o sujeito tiver sido exposto a mais que um nível limiar de exposição segura à genotoxina, fornecer um tratamento profilático e/ou terapêutico para prevenir ou inibir o início de uma doença ou distúrbio associado à genotoxina.

54. Sistema para detectar e identificar eventos mutagênicos e / ou eventos de danos aos ácidos nucleicos resultantes da exposição genotóxica de uma amostra, caracterizado pelo fato de que compreende:

uma rede de computadores para transmissão de informações relacionadas a dados de sequenciamento e dados de genotoxicidade, em que as informações incluem um ou mais dados brutos de sequenciamento, dados de sequenciamento duplex, informações de amostra e informações de genotoxina;

um computador cliente associado a um ou mais dispositivos de computação do usuário e em comunicação com a rede de computadores;

um banco de dados conectado à rede de computadores para armazenar uma pluralidade de perfis de genotoxinas e registros de resultados do usuário;

um módulo de sequenciamento duplex em comunicação com a rede de computadores e configurado para receber dados brutos de sequenciamento e solicitações do computador cliente para gerar dados de sequenciamento duplex, leituras de sequências de grupos de famílias que representam uma molécula original de ácido nucleico de fita dupla e comparar sequências representativas de fitas individuais uma com a outra para gerar dados de sequenciamento duplex;

e

um módulo de genotoxina em comunicação com a rede de computadores e configurado para comparar dados de sequenciamento duplex para referenciar informações de sequência para identificar mutações e gerar dados de genotoxina compreendendo pelo menos uma dentre uma frequência mutante, um espectro de mutação e um espectro de mutação tripleto.

55. Sistema, de acordo com a reivindicação 54, caracterizado pelo fato de que os perfis de genotoxina compreendem espectro de mutação de genotoxina a partir de uma pluralidade de genotoxinas conhecidas.

56. Meio de armazenamento não transitório, legível por computador, caracterizado pelo fato de que compreende instruções que, quando executadas por um ou mais processadores, executam um método como definido em qualquer um das reivindicações 1-46 e 52-53 para determinar se um sujeito está exposto a pelo menos uma genotoxina e/ou determinar uma identidade de pelo menos uma genotoxina.

57. Meio de armazenamento legível por computador não transitório, de acordo com a reivindicação 56, caracterizado pelo fato de que ainda calcular o espectro de mutação, a frequência mutante e/ou o espectro de mutação tripleto de um agente detectado, a partir do qual é determinada a identidade de pelo menos uma genotoxina.

58. Sistema de computador para executar um método como definido em qualquer uma das reivindicações 1-53 para determinar se um sujeito está exposto e/ou uma identidade de pelo menos uma genotoxina, o sistema, caracterizado pelo fato de que compreende: pelo menos um computador com um processador, memória, banco de dados e um meio de armazenamento legível

por computador não transitório compreendendo instruções para o(s) processador(es), em que o(s) referido(s) processador(es) está configurado para executar as referidas instruções para executar operações compreendendo os métodos como definidos em qualquer uma das reivindicações 1-46 e 52-53.

59. Sistema, de acordo com a reivindicação 58, caracterizado pelo fato de que compreende ainda um sistema de computador em rede que compreende:

a. uma rede com ou sem fio;

b. uma pluralidade de dispositivos de computação eletrônica do usuário capazes de receber dados derivados do uso de um kit compreendendo reagentes para extrair, amplificar e produzir uma sequência polinucleotídica da amostra de um sujeito e transmitir a sequência polinucleotídica via rede para um servidor remoto; e

c. um servidor remoto compreendendo o processador, a memória, o banco de dados e o meio de armazenamento legível por computador não transitório compreendendo instruções para o(s) processador(es), em que o(s) referido(s) processador(es) está(ão) configurado(s) para executar as referidas instruções para executar operações compreendendo os métodos como definidos em qualquer uma das reivindicações 1-46 e 52-53; e

d. em que o referido servidor remoto é capaz de detectar e identificar eventos mutagênicos e / ou eventos de danos aos ácidos nucleicos resultantes da exposição genotóxica de uma amostra.

60. Sistema, de acordo com a reivindicação 59, caracterizado pelo fato de que o banco de dados e/ou um banco de dados de terceiros acessíveis via rede, compreende ainda uma pluralidade de registros compreendendo um ou mais de um perfil de genotoxina de genotoxinas conhecidas, um perfil de genotoxina de pelo menos uma amostra do sujeito, e em que o perfil de genotoxina compreende uma mutação ou um sítio de dano ao DNA.

61. Meio legível por computador não transitório, cujo conteúdo faz com que pelo menos um computador execute um método para fornecer dados de sequenciamento duplex para moléculas de ácido nucleico de fita dupla em uma amostra de um ensaio de triagem de genotoxicidade, o método, caracterizado pelo fato de que compreende:

receber dados de sequência bruta de um dispositivo de computação do usuário; e

criar um conjunto de dados específicos de amostra compreendendo uma pluralidade de leituras de sequência bruta derivadas de uma pluralidade de moléculas de ácido nucleico na amostra;

as leituras de sequência de agrupamento de famílias que representam uma molécula original de ácido nucleico de fita dupla, em que o agrupamento é baseado em uma sequência de identificador de molécula única compartilhada;

comparar uma sequência da primeira fita e a sequência da segunda fita a partir de uma molécula original de ácido nucleico de fita dupla para identificar uma ou mais correspondências entre as leituras de sequência da primeira e da segunda fitas; e

fornecer dados de sequenciamento duplex para as moléculas de ácido nucleico de fita dupla na amostra.

62. Meio legível por computador, de acordo com a reivindicação 61, caracterizado pelo fato de que compreende ainda identificar posições nucleotídicas de não complementaridade entre a primeira e a segunda sequências comparadas, em que o método compreende ainda:

em posições de não complementaridade, identificar e eliminar ou descontar os erros do processo; e

em posições de não complementaridade que não são identificadas como erros de processo, identificar as posições restantes de não complementaridade como sítios de possíveis danos in vivo ao DNA resultantes da exposição a uma genotoxina.

63. Meio legível por computador não transitório, cujo conteúdo faz com que pelo menos um computador execute um método para detectar e identificar eventos mutagênicos resultantes da exposição genotóxica de uma amostra, o método, caracterizado pelo fato de que compreende:

comparar dados de sequência duplex para fazer referência a informações de sequência;

identificar mutações nos dados da sequência duplex, em que uma mutação é identificada como uma região de não concordância com a informação de referência;

determinar uma frequência mutante nos dados da sequência duplex;

gerar um espectro de mutação a partir dos dados da sequência duplex;

gerar um espectro de mutação triplete a partir dos dados da sequência duplex; e

comparar o espectro de mutação e/ou o espectro de mutação triplete a uma pluralidade de conjuntos de dados de genotoxina conhecidos.

64. Meio legível por computador não transitório, cujo conteúdo faz com que pelo menos um computador execute um método para detectar e identificar uma exposição a carcinógenos ou carcinógenos em um sujeito, o método, caracterizado pelo fato de que compreende:

identificar variantes de sequência em uma região genômica alvo usando dados de sequenciamento duplex gerados a partir de uma amostra do sujeito;

calcular uma frequência de alelo variante (VAF) de uma amostra de teste e uma amostra de controle;

determinar se uma VAF é maior em um grupo de teste do que em um grupo de controle;

em amostras com uma VAF mais alta, determinar se uma variante de sequência é um não singleto;

em amostras com uma VAF mais alta, determinar se a variante de sequência é uma mutação ativadora; e

caracterizar amostras com uma mutação não singleto e/ou ativadora como suspeita de ser um carcinógeno.

65. Meio legível por computador não transitório, de acordo com a reivindicação 64, caracterizado pelo fato de que compreende ainda avaliar um limiar de segurança para o carcinógeno e / ou determinar um risco associado ao desenvolvimento de uma doença ou distúrbio associado à genotoxina após a exposição no sujeito.

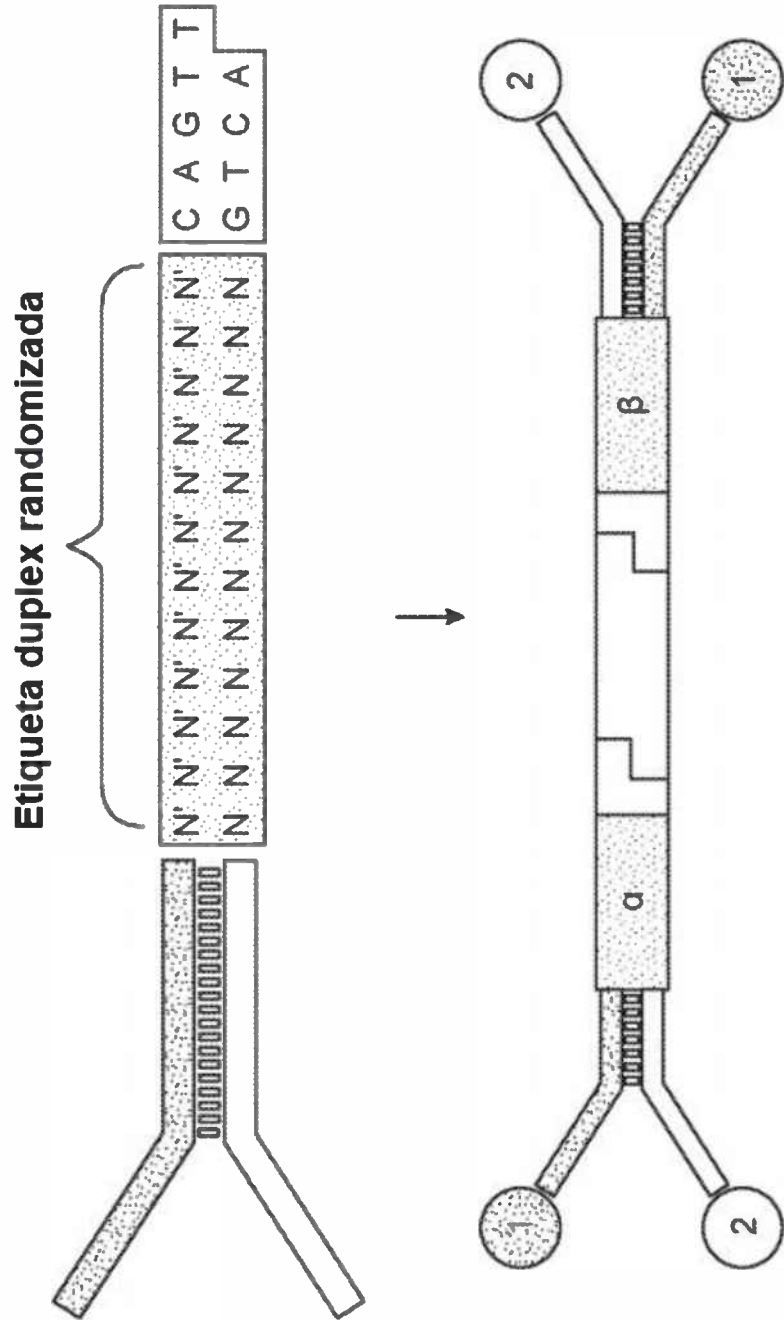


FIG. 1A

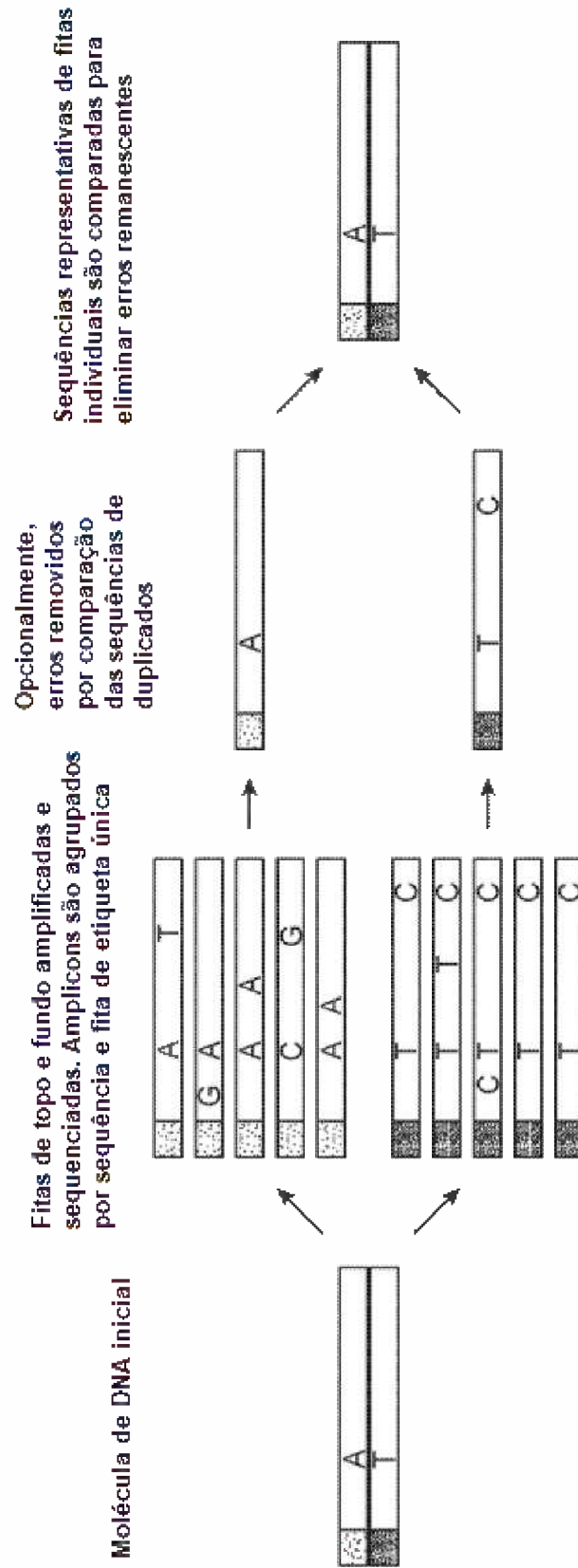


FIG. 1B

FIG. 1C

Quantificação de genotoxicidade

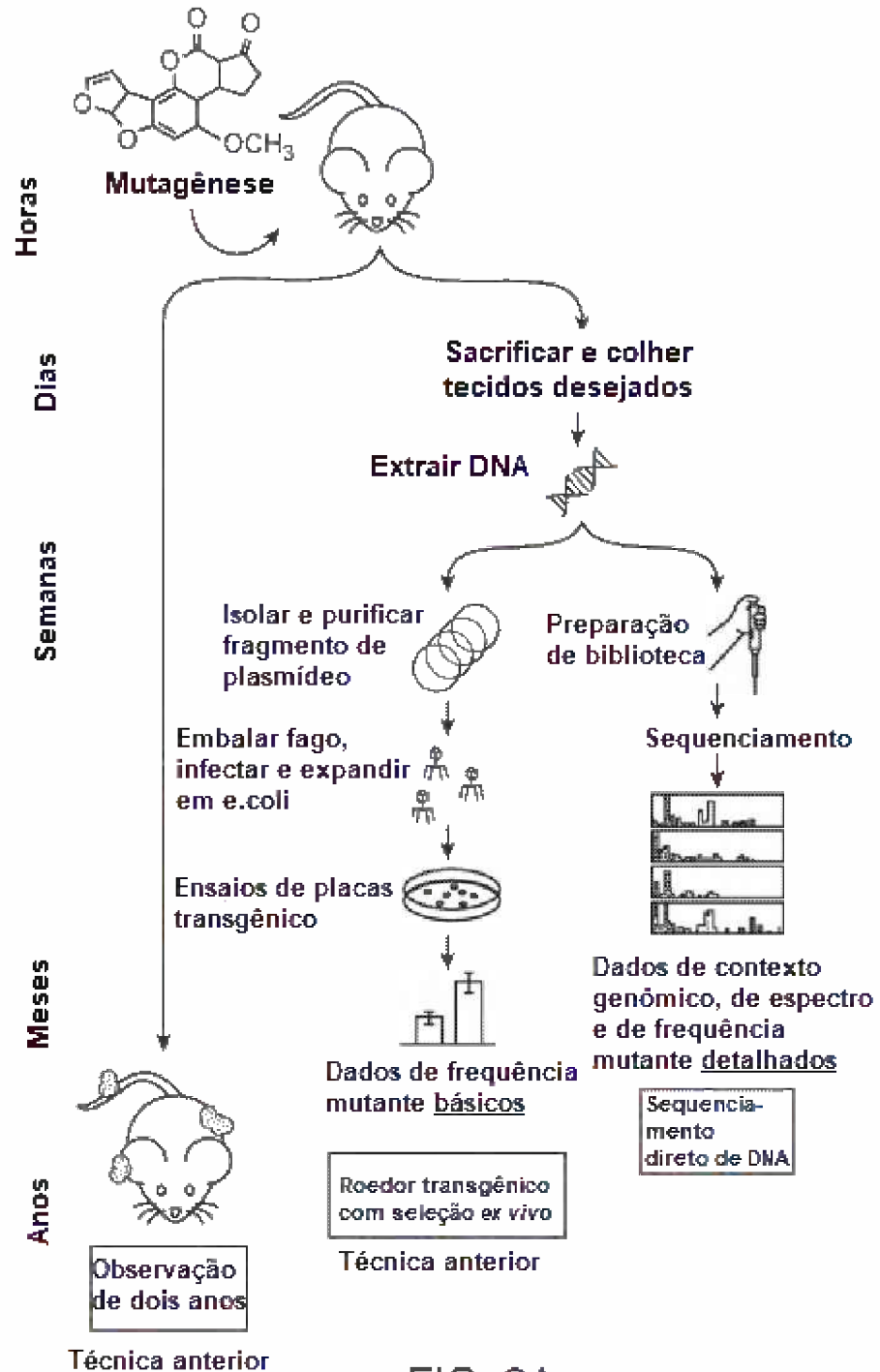


FIG. 2A

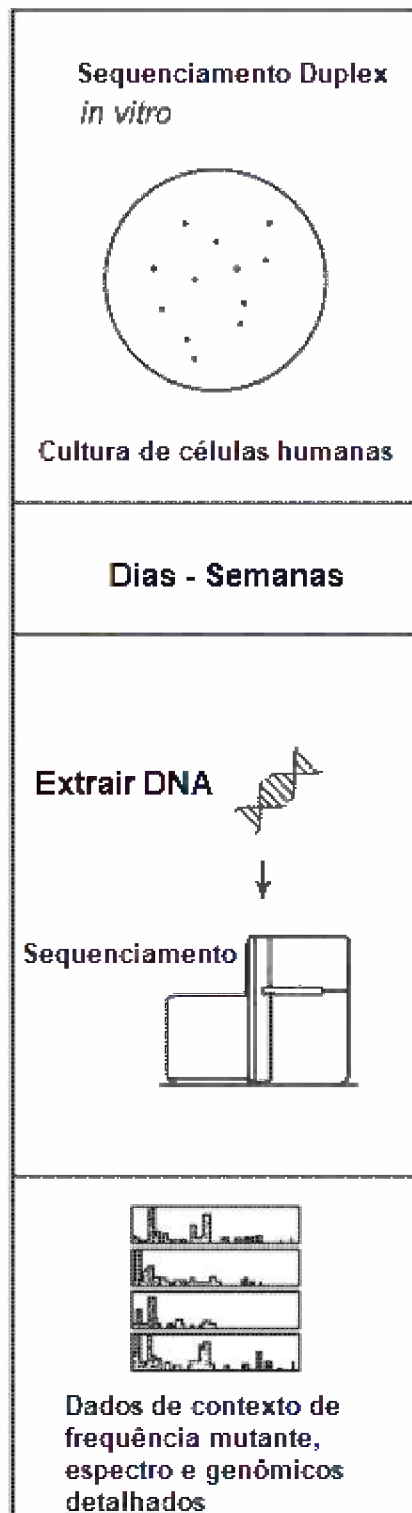


FIG. 2B

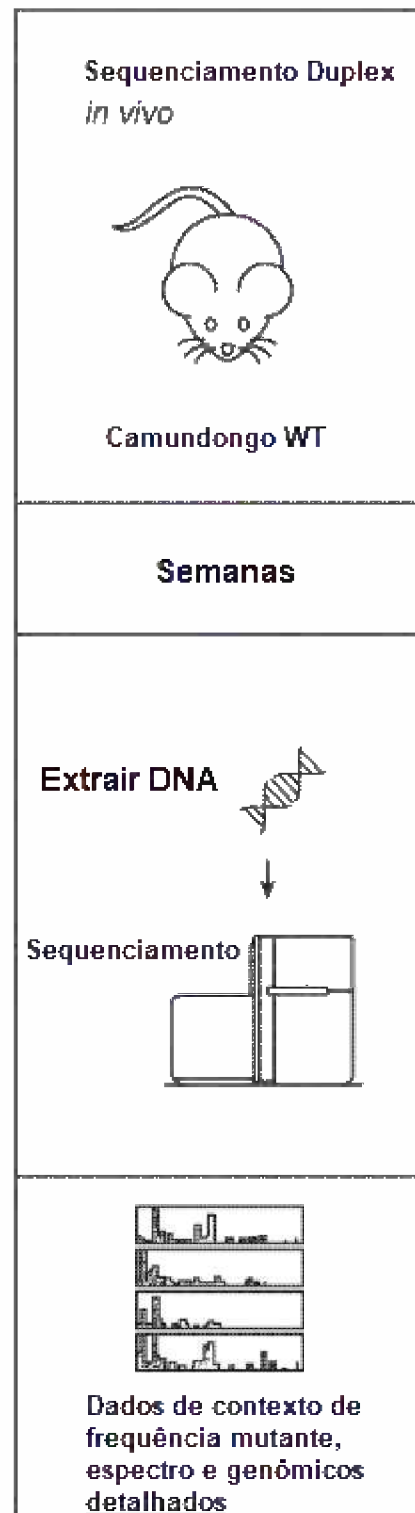


FIG. 2C

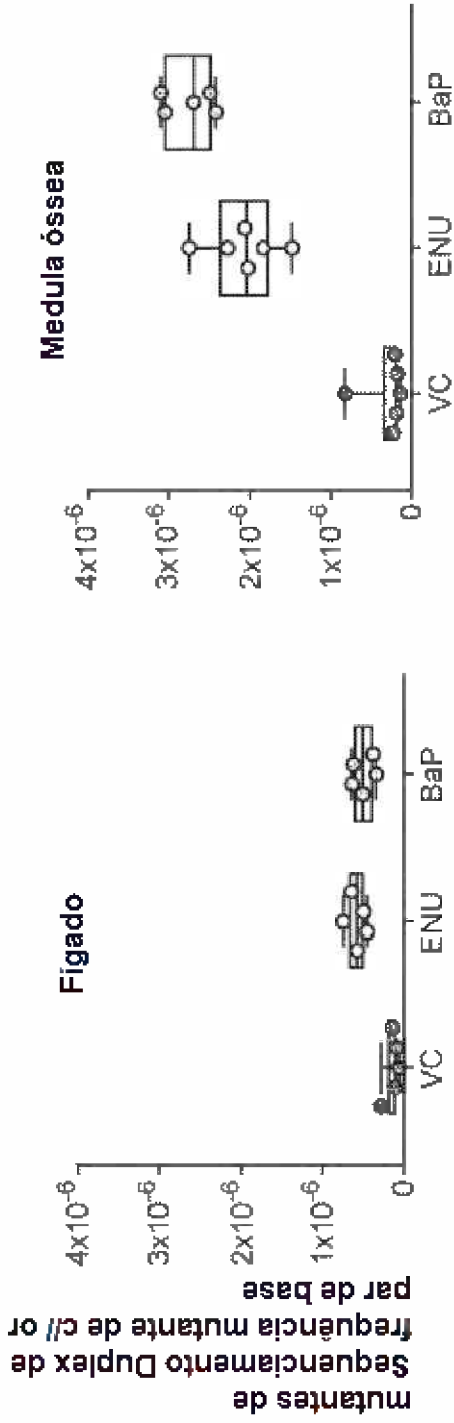


FIG. 3A

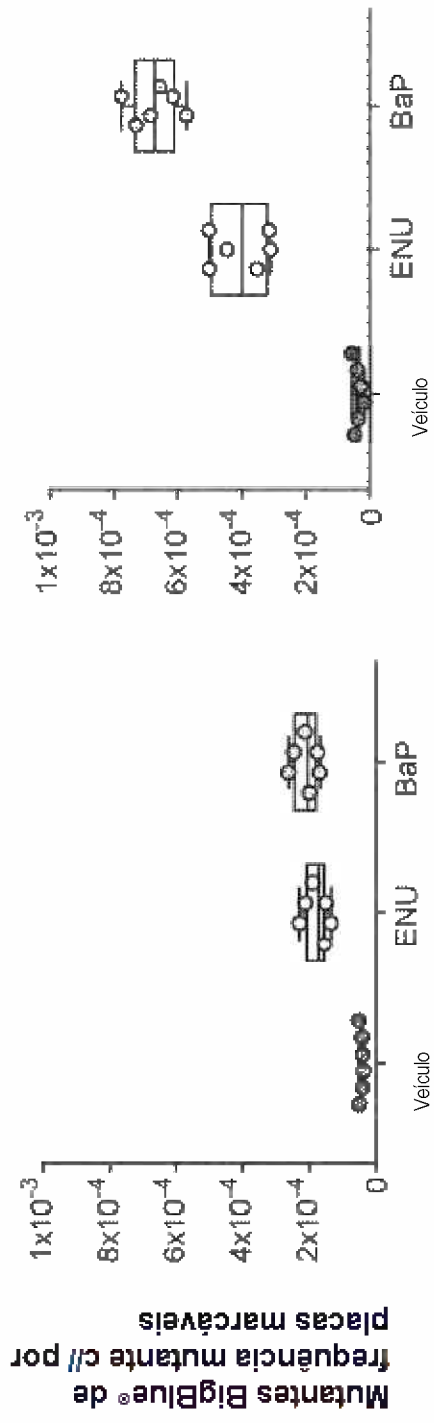


FIG. 3C

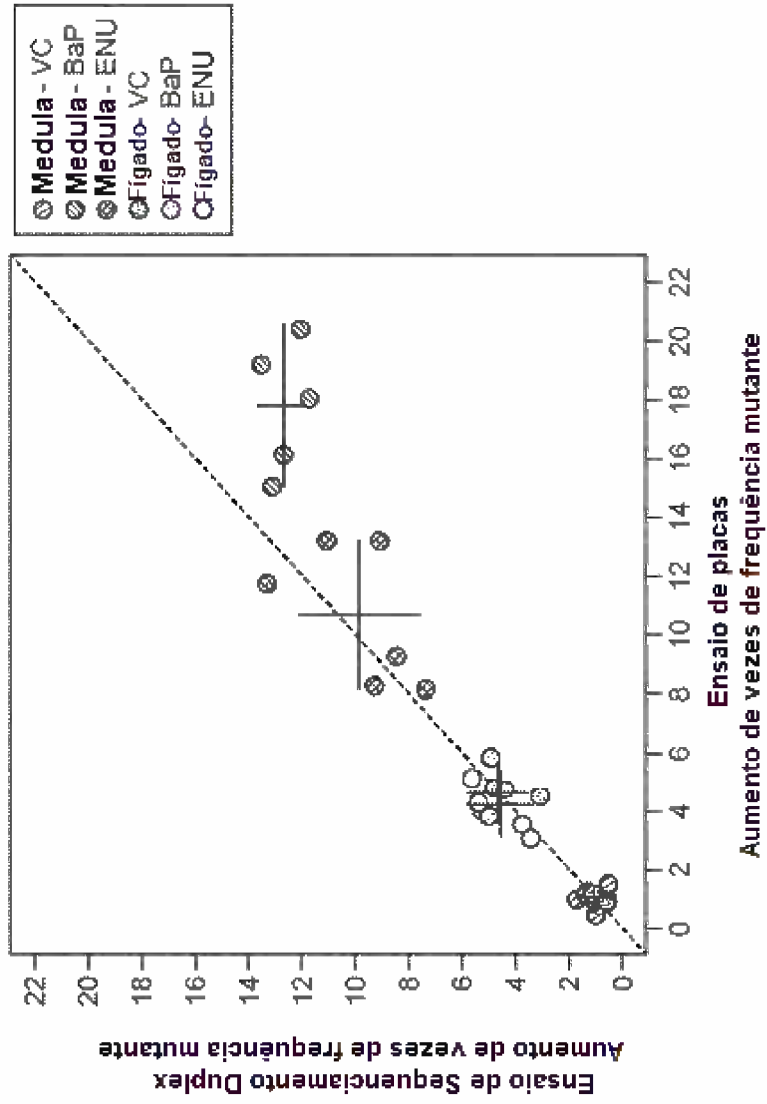


FIG. 3E

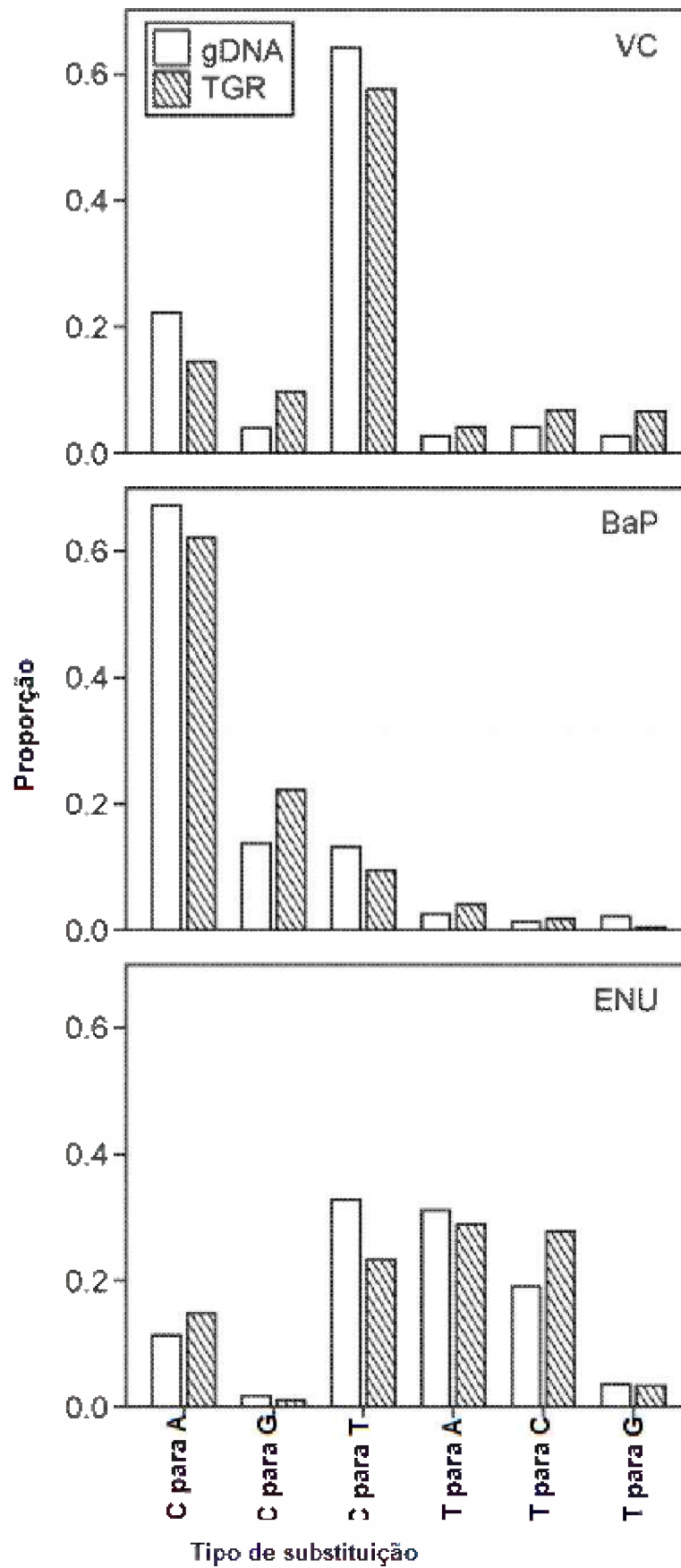


FIG. 3F

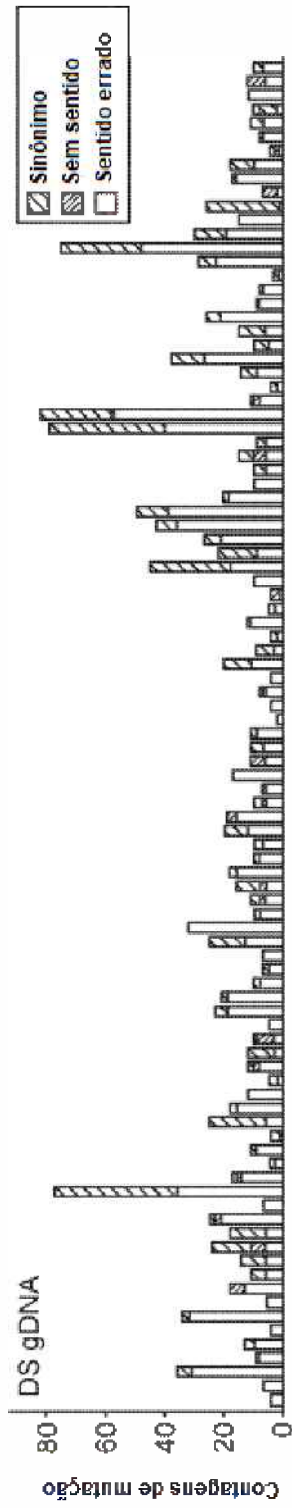


FIG. 3G

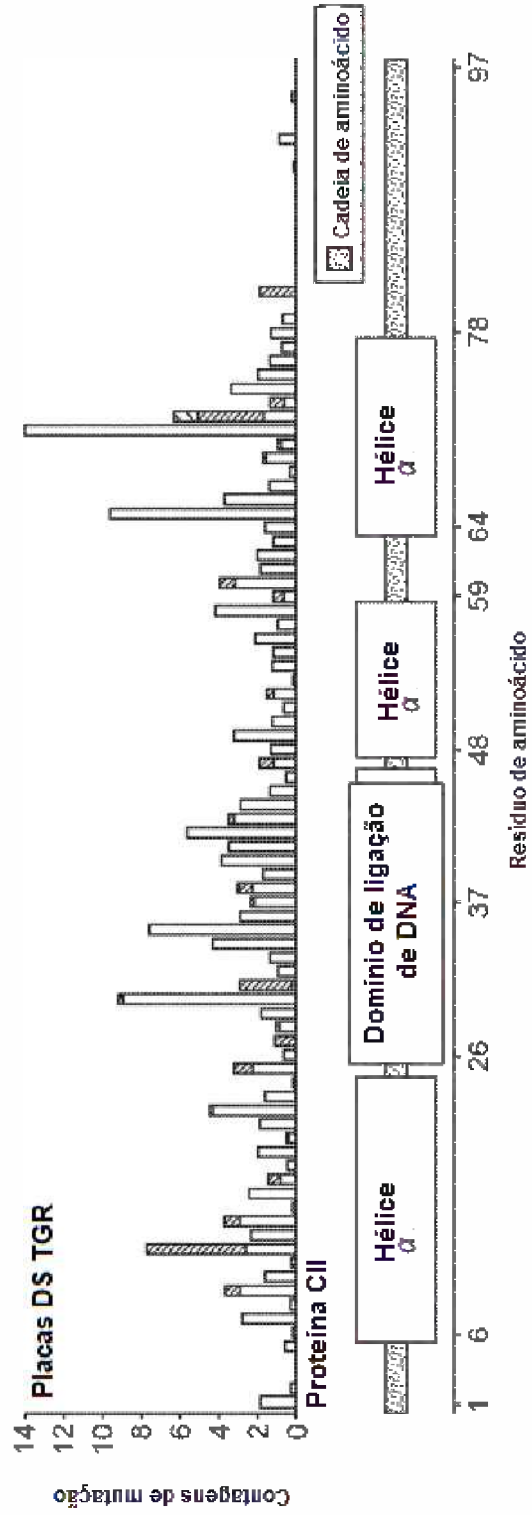
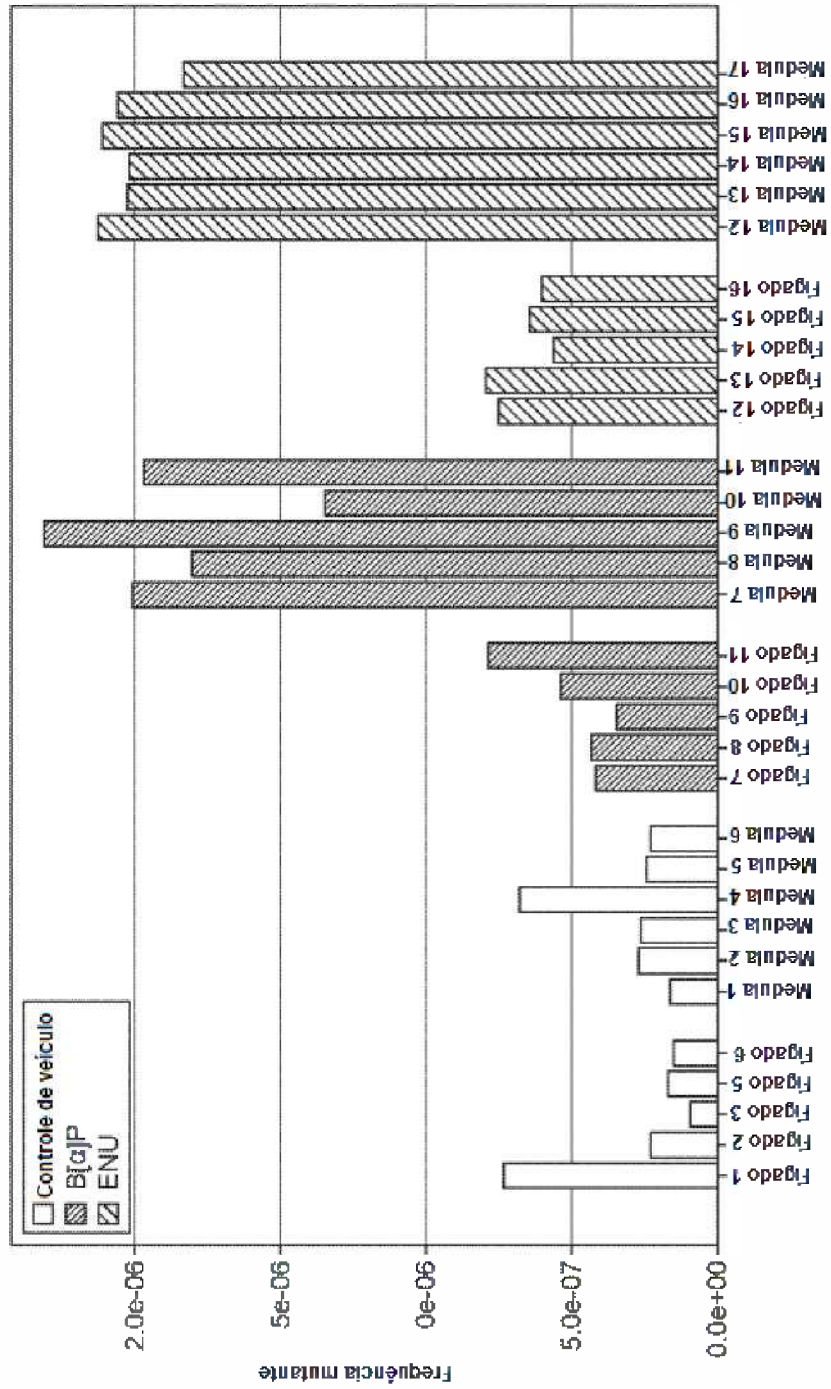


FIG. 3H



Amostra de tecido murino

FIG. 4

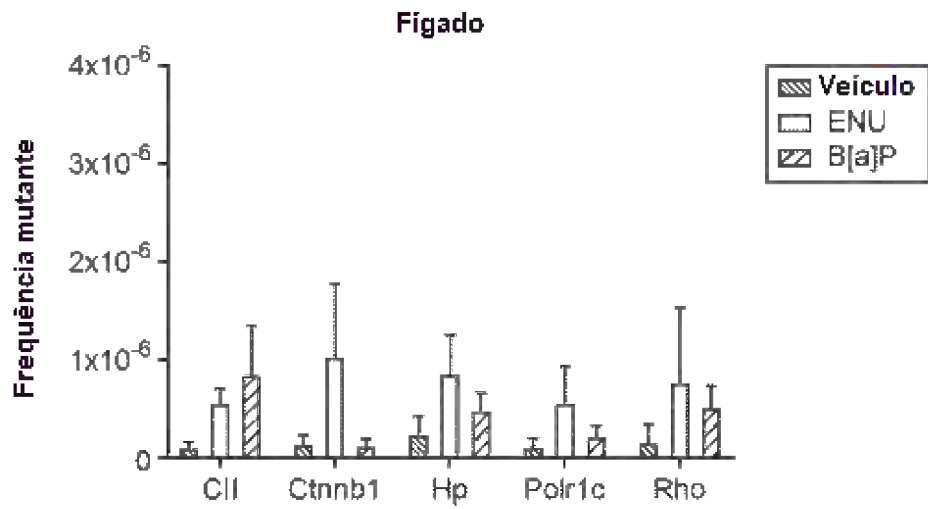


FIG. 5A

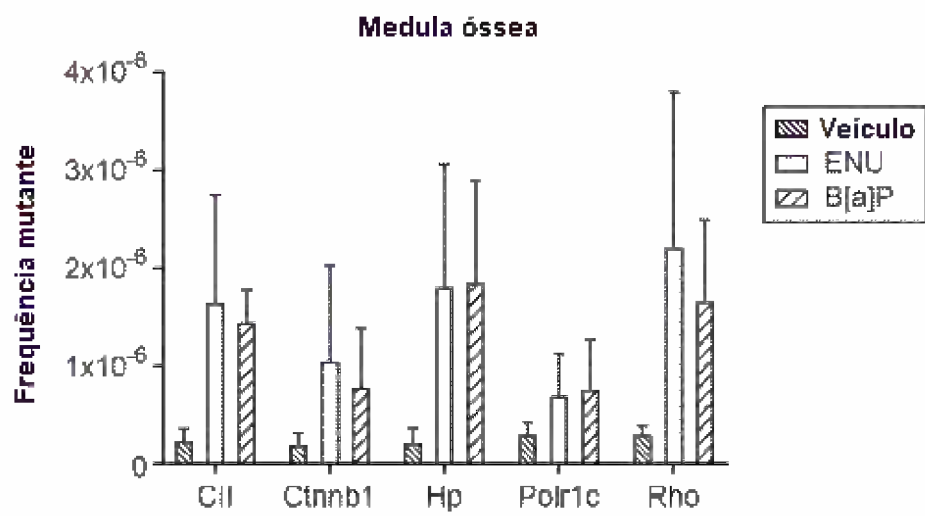


FIG. 5B

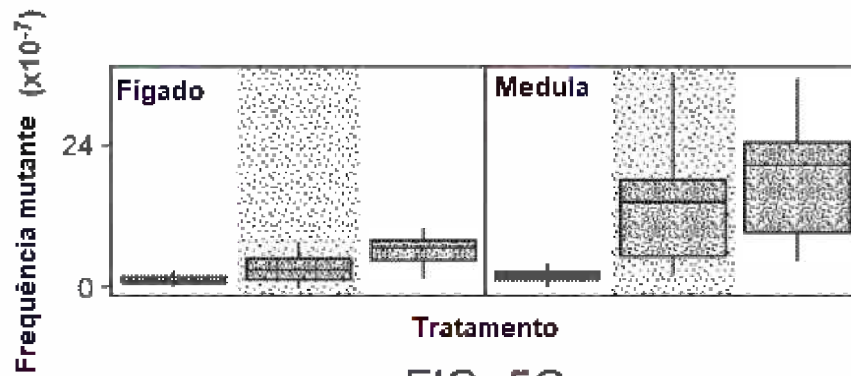


FIG. 5C

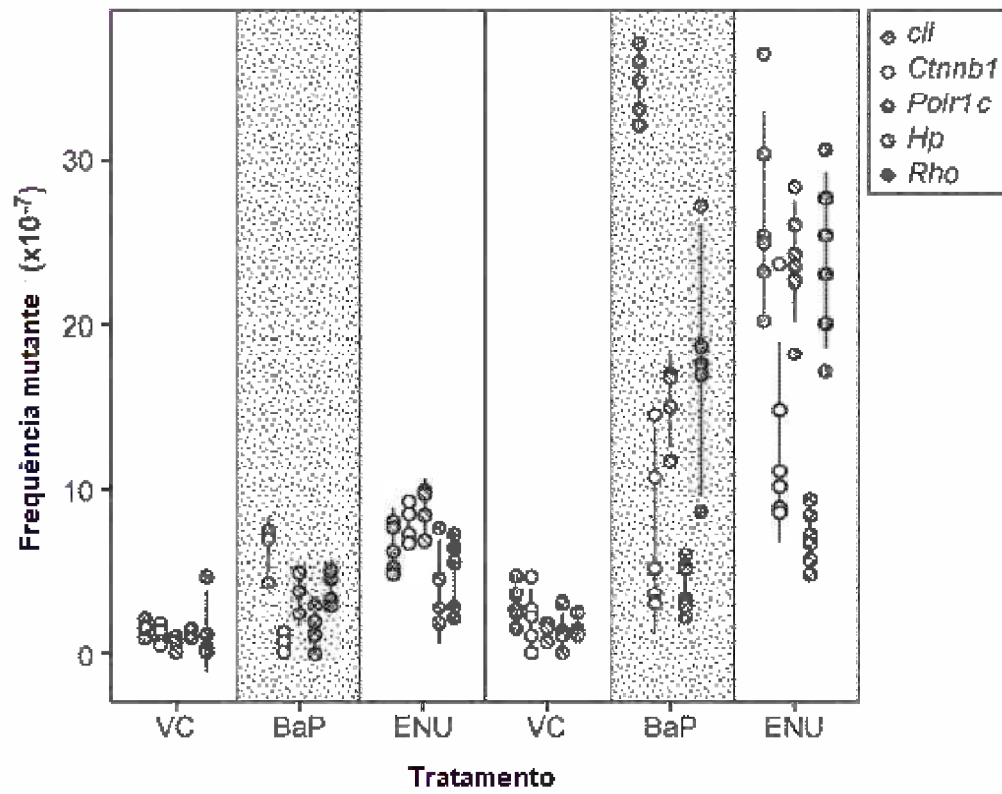


FIG. 5D

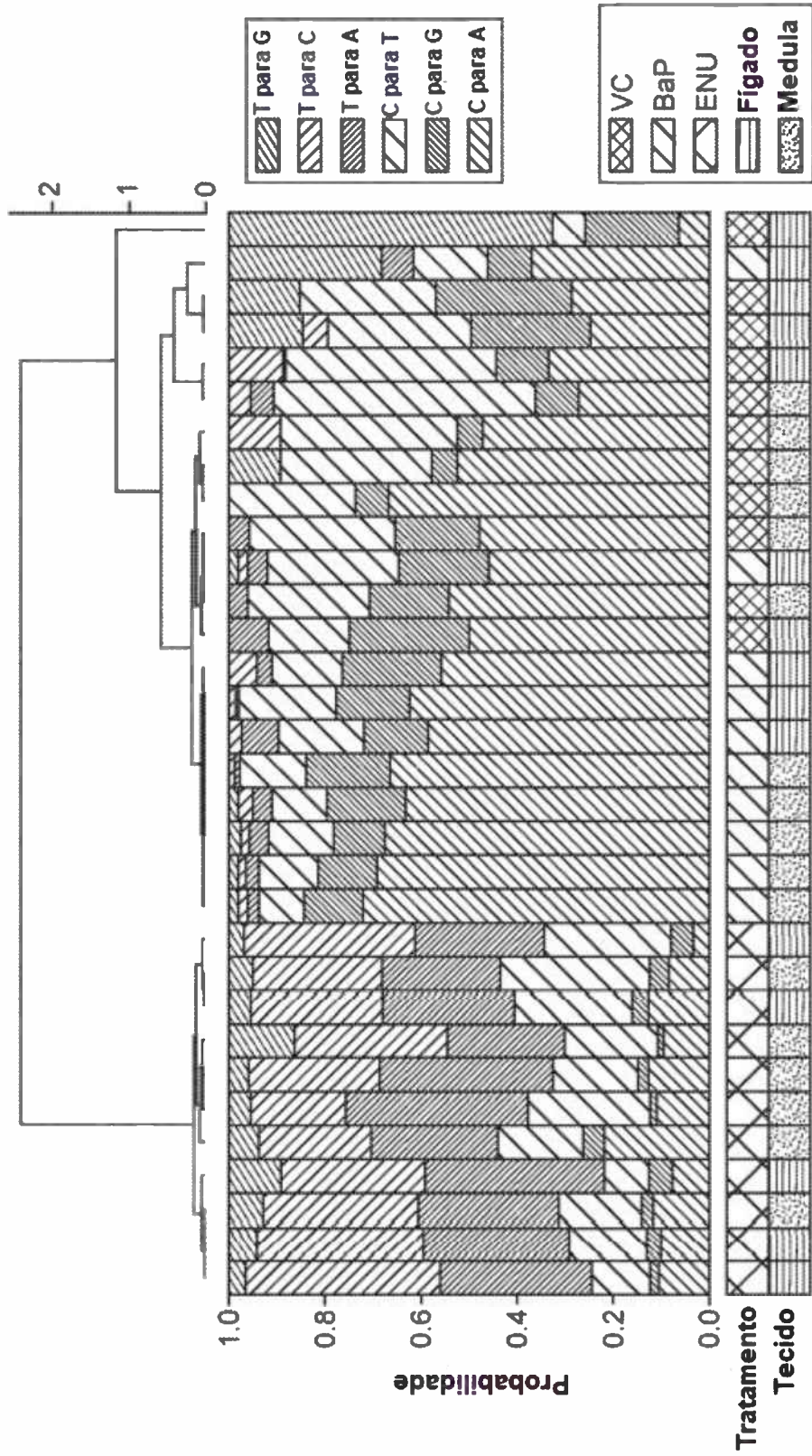
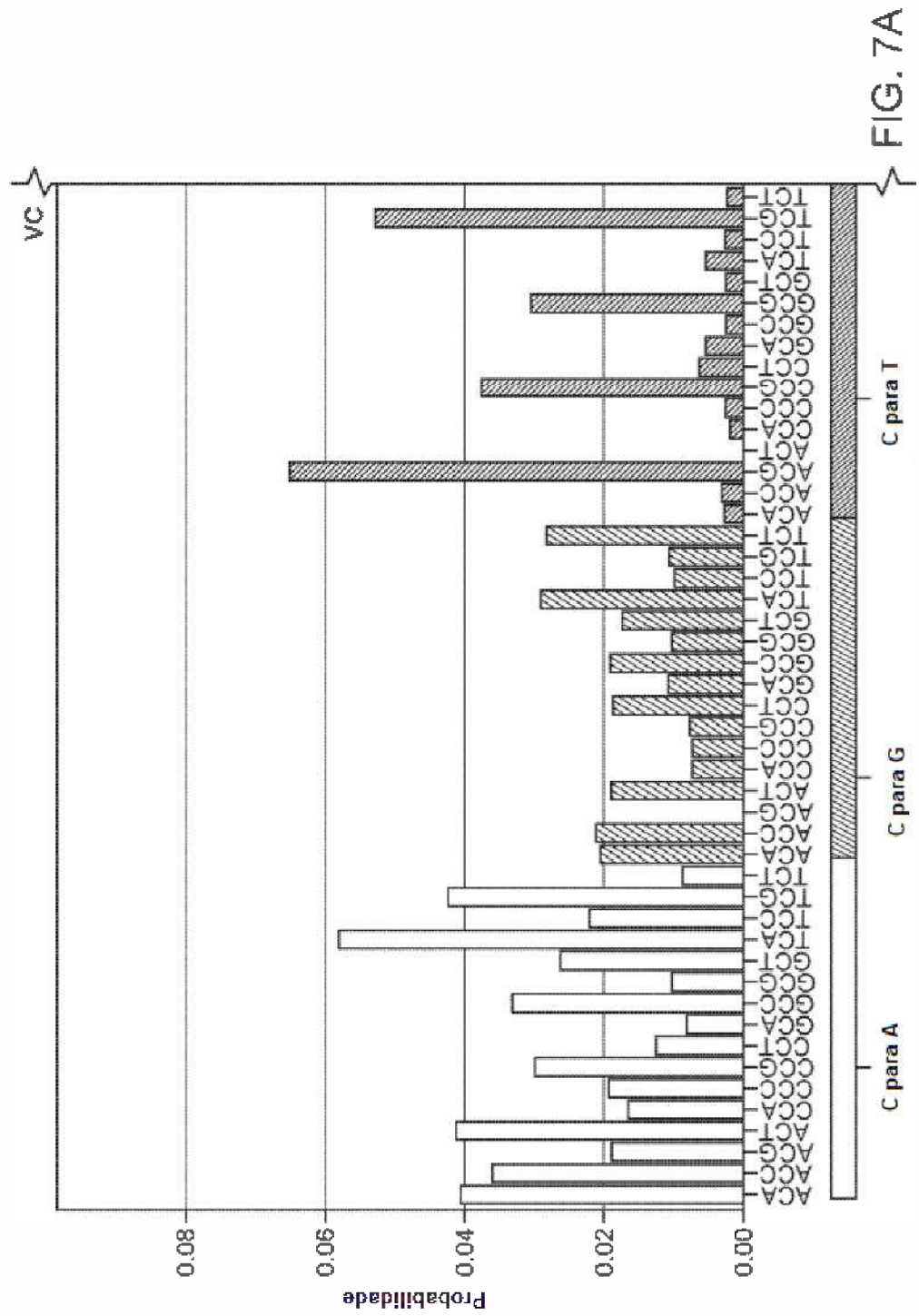


FIG. 6



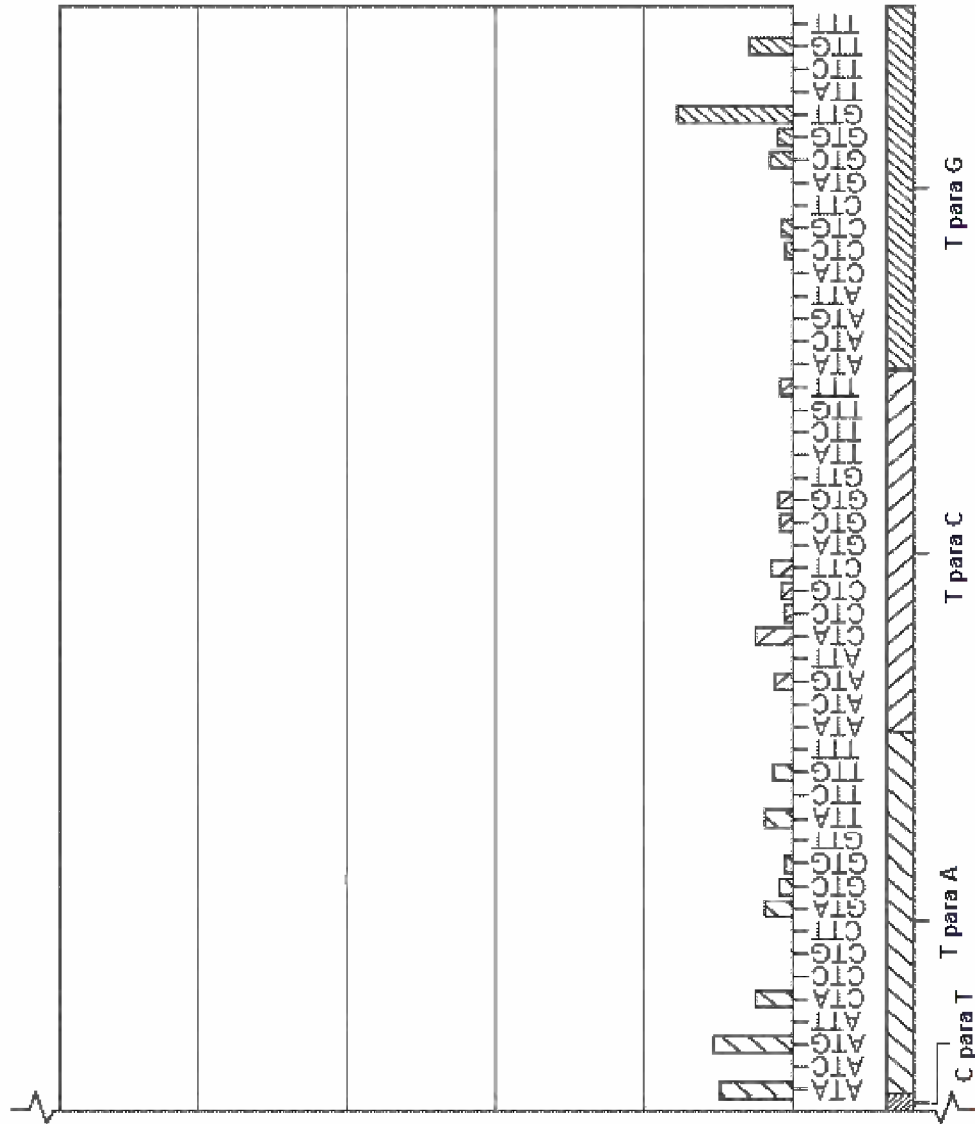


FIG. 7B (Cont.)

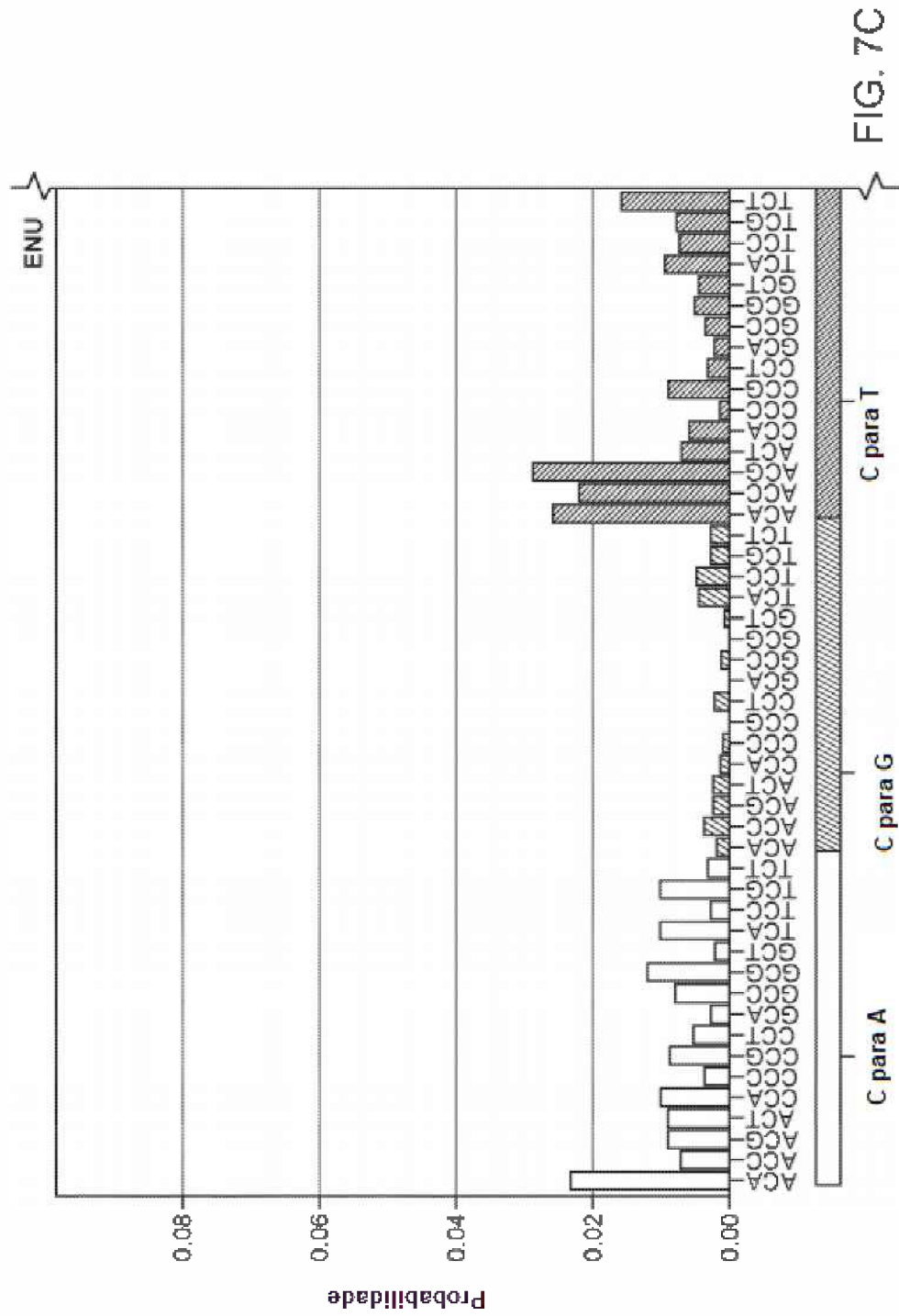


FIG. 7C

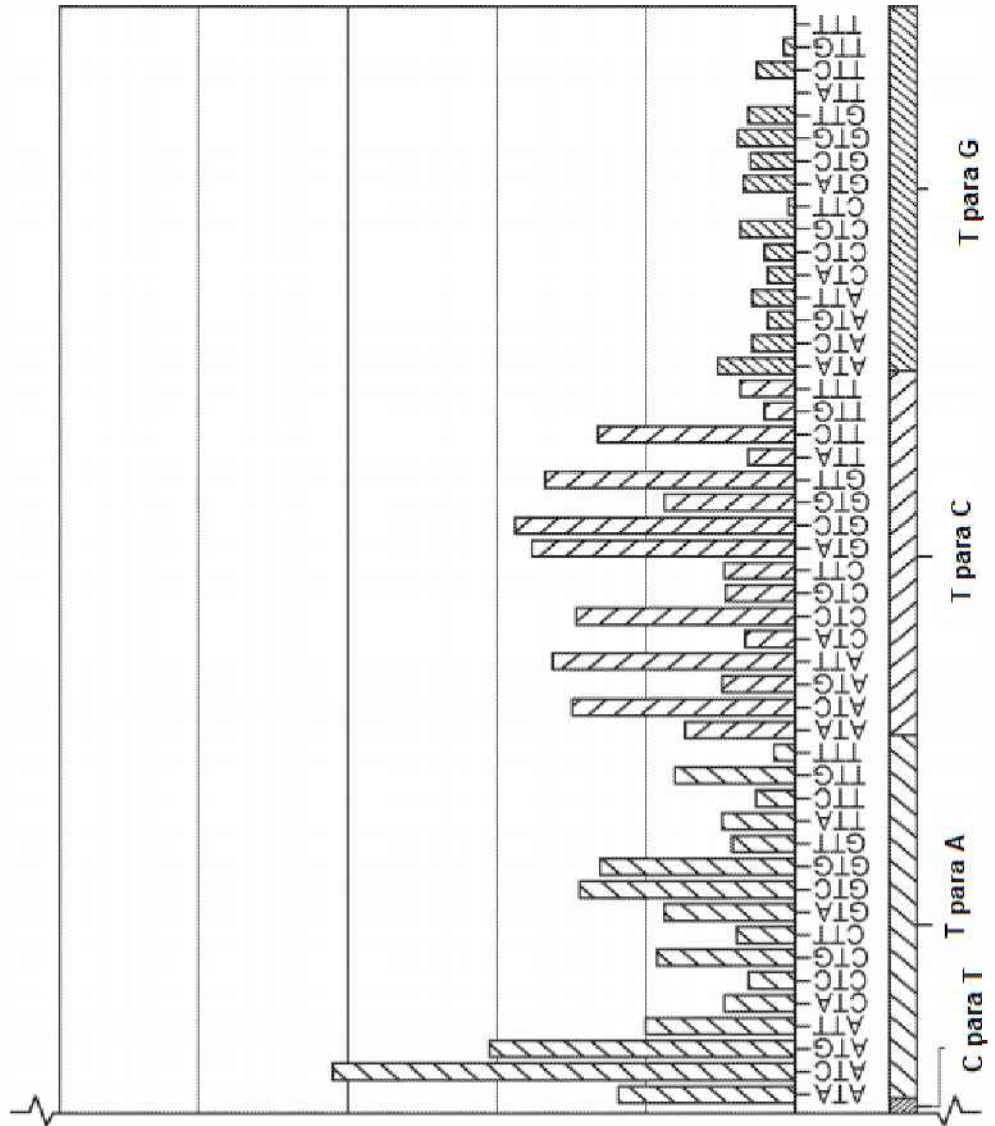


FIG. 7C (Cont.)

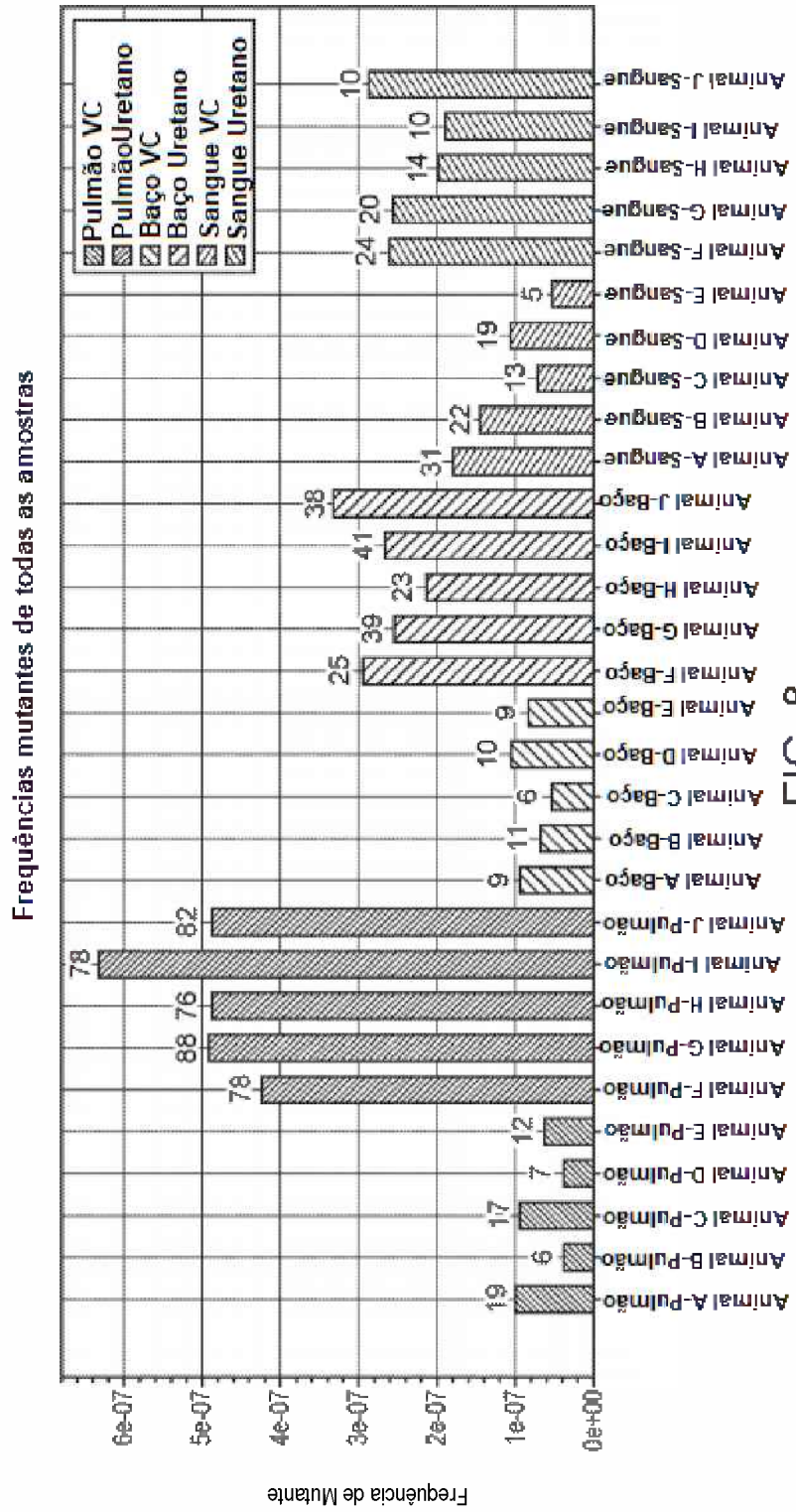


FIG. 8

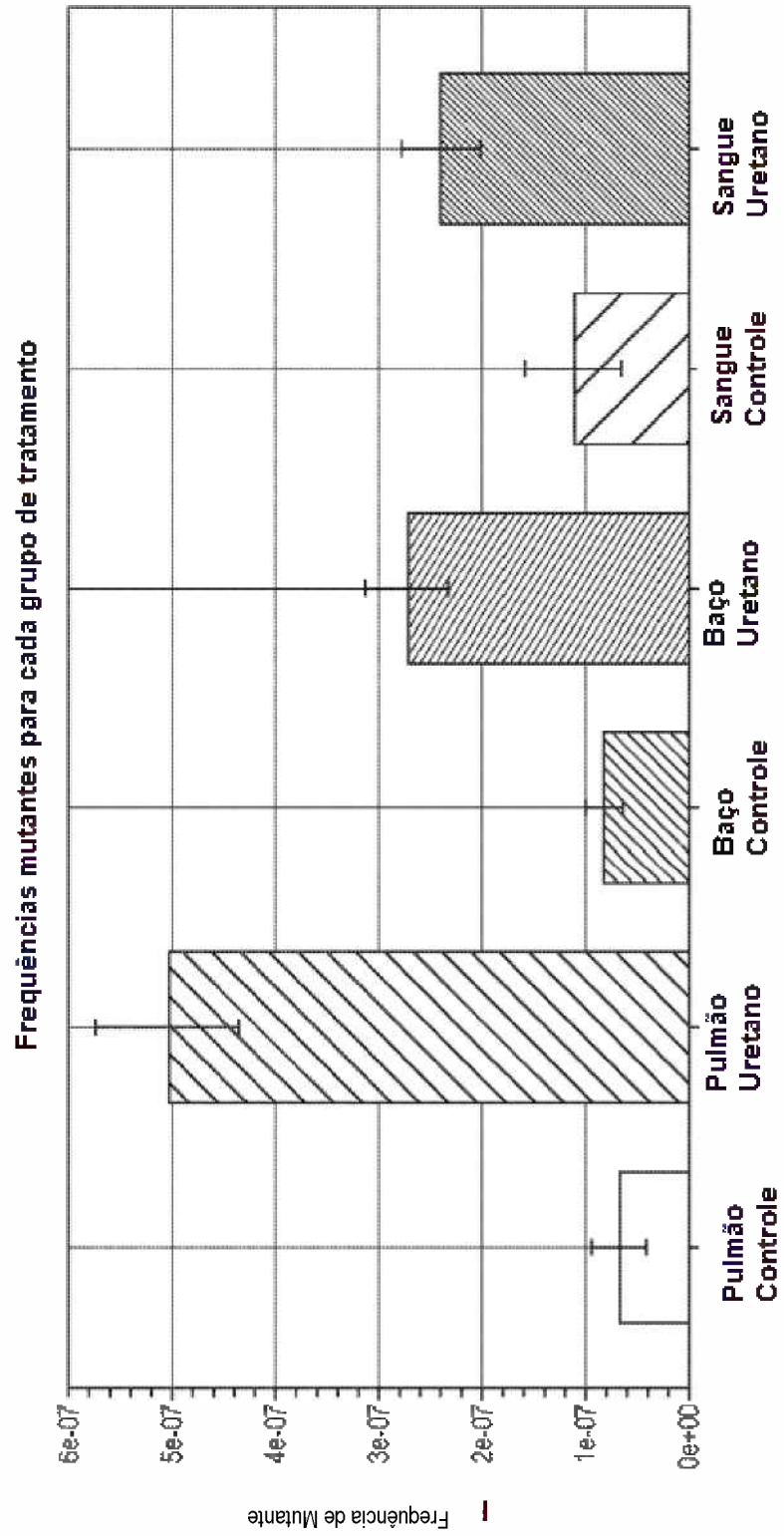


FIG. 9

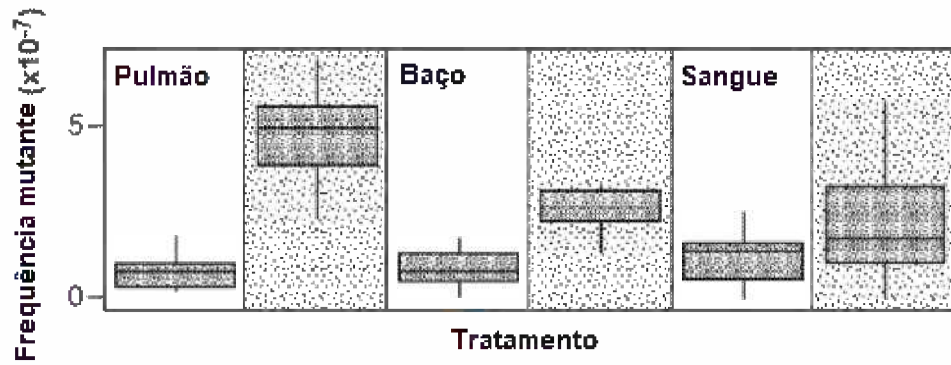


FIG. 10A

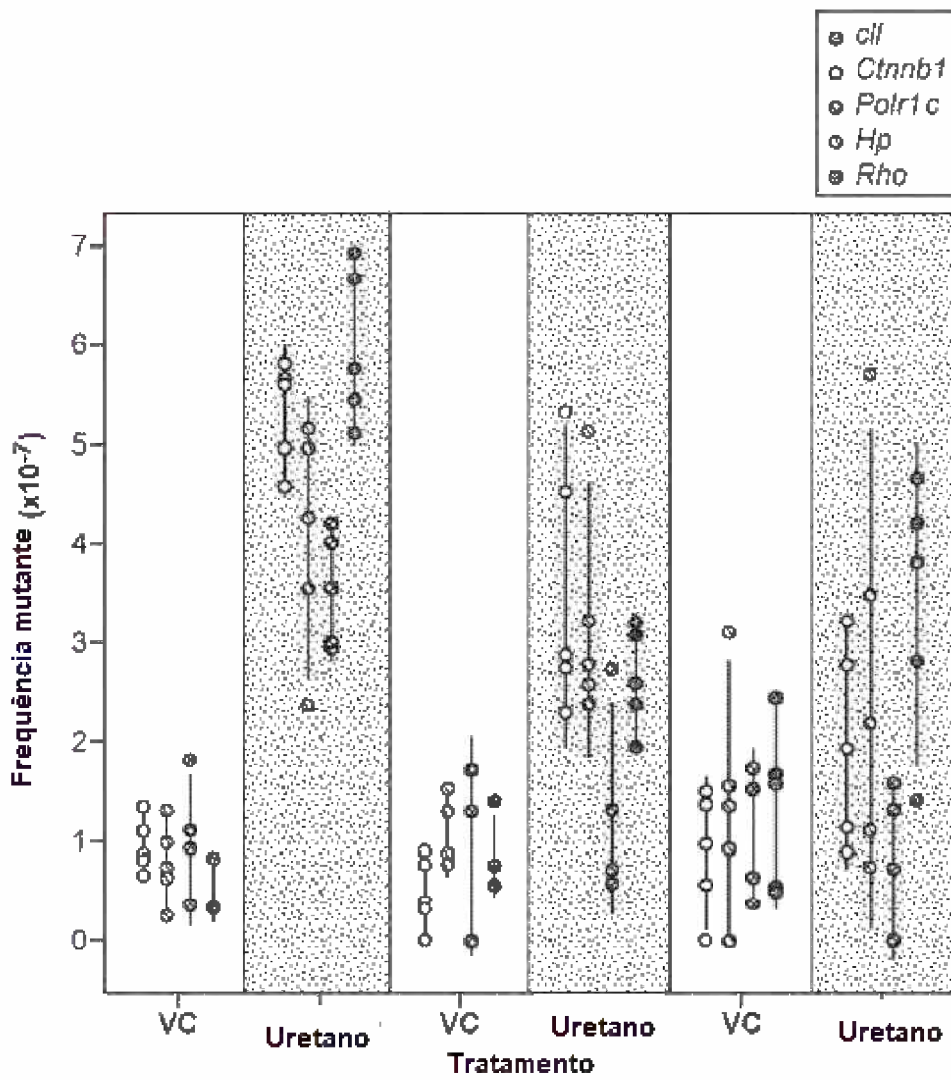


FIG. 10B



FIG. 11

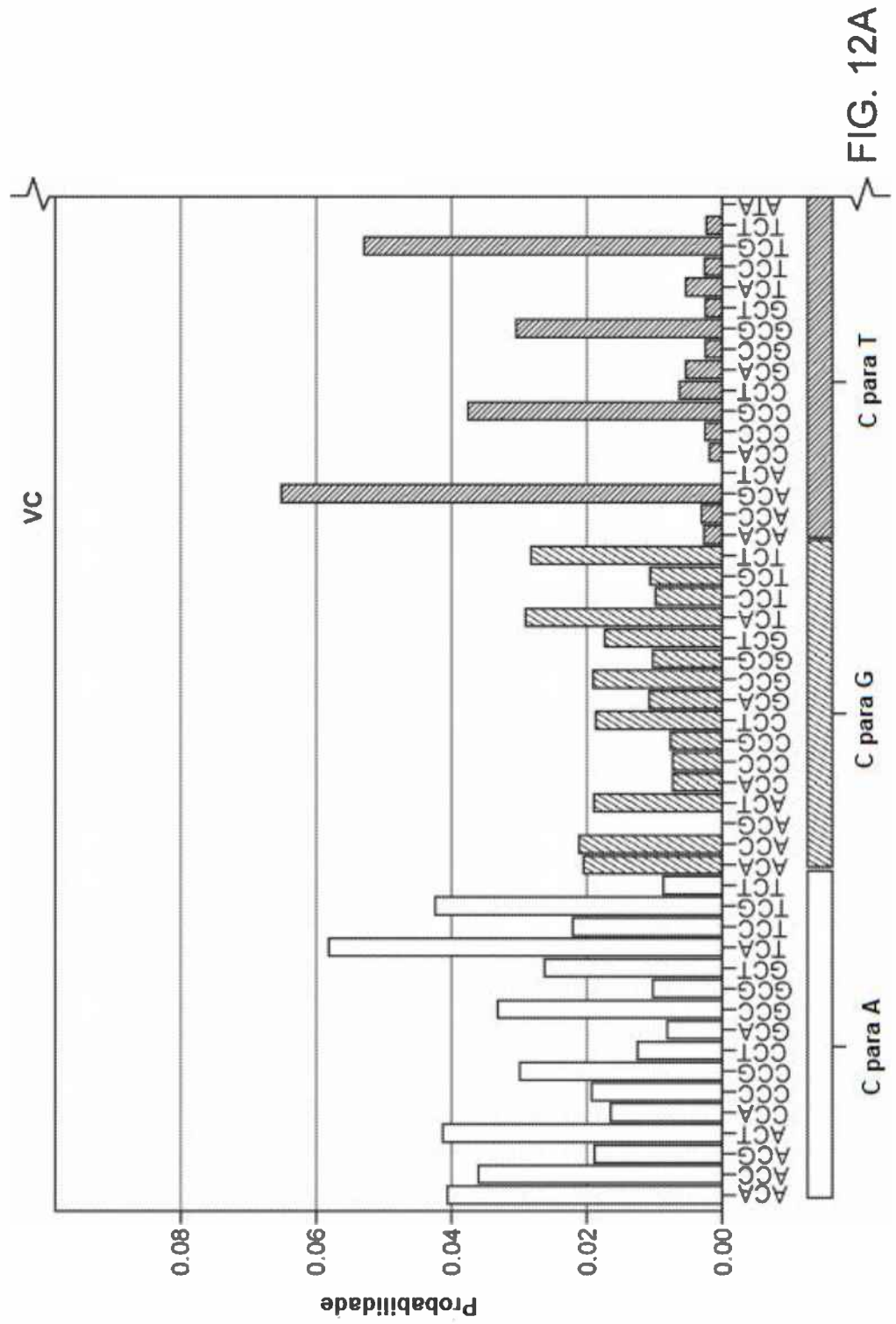


FIG. 12A

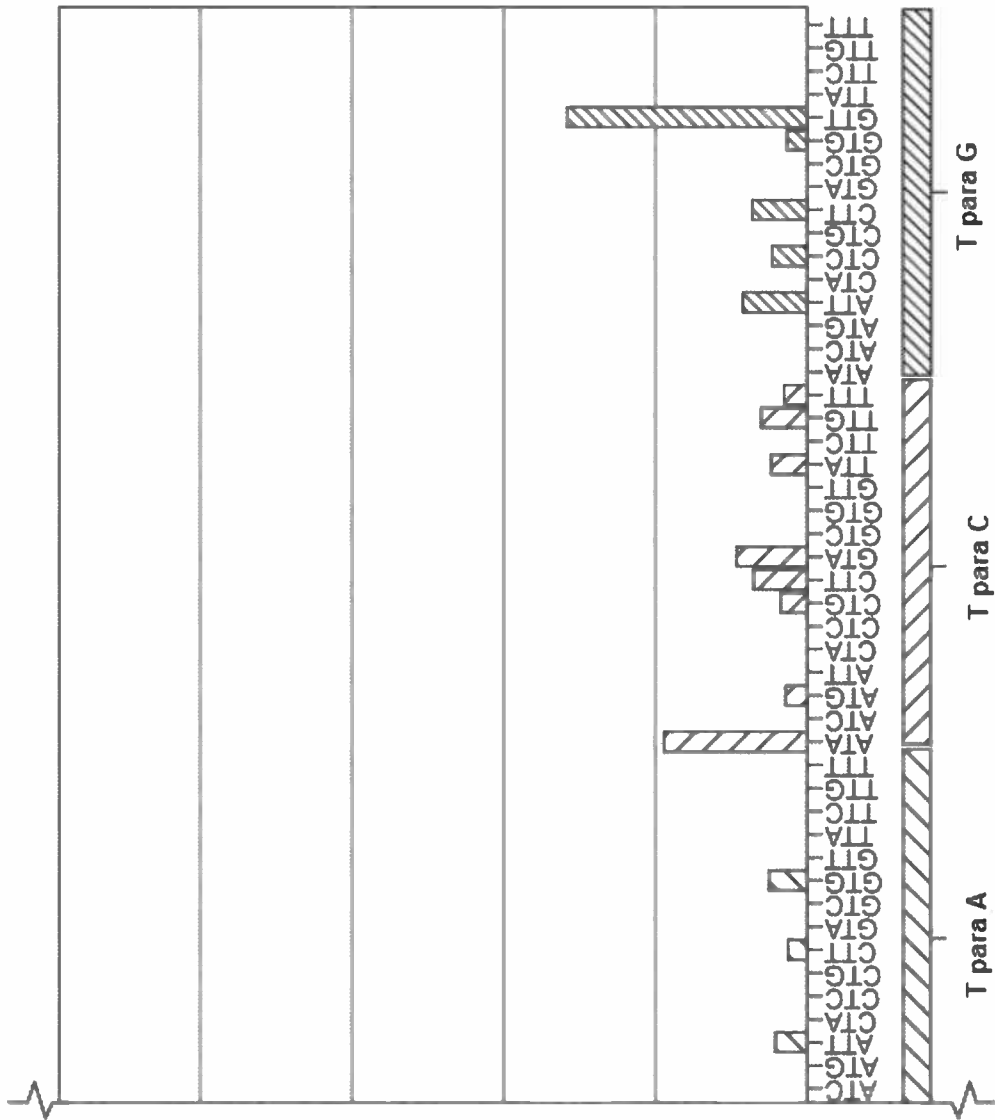


FIG. 12A (Cont.)

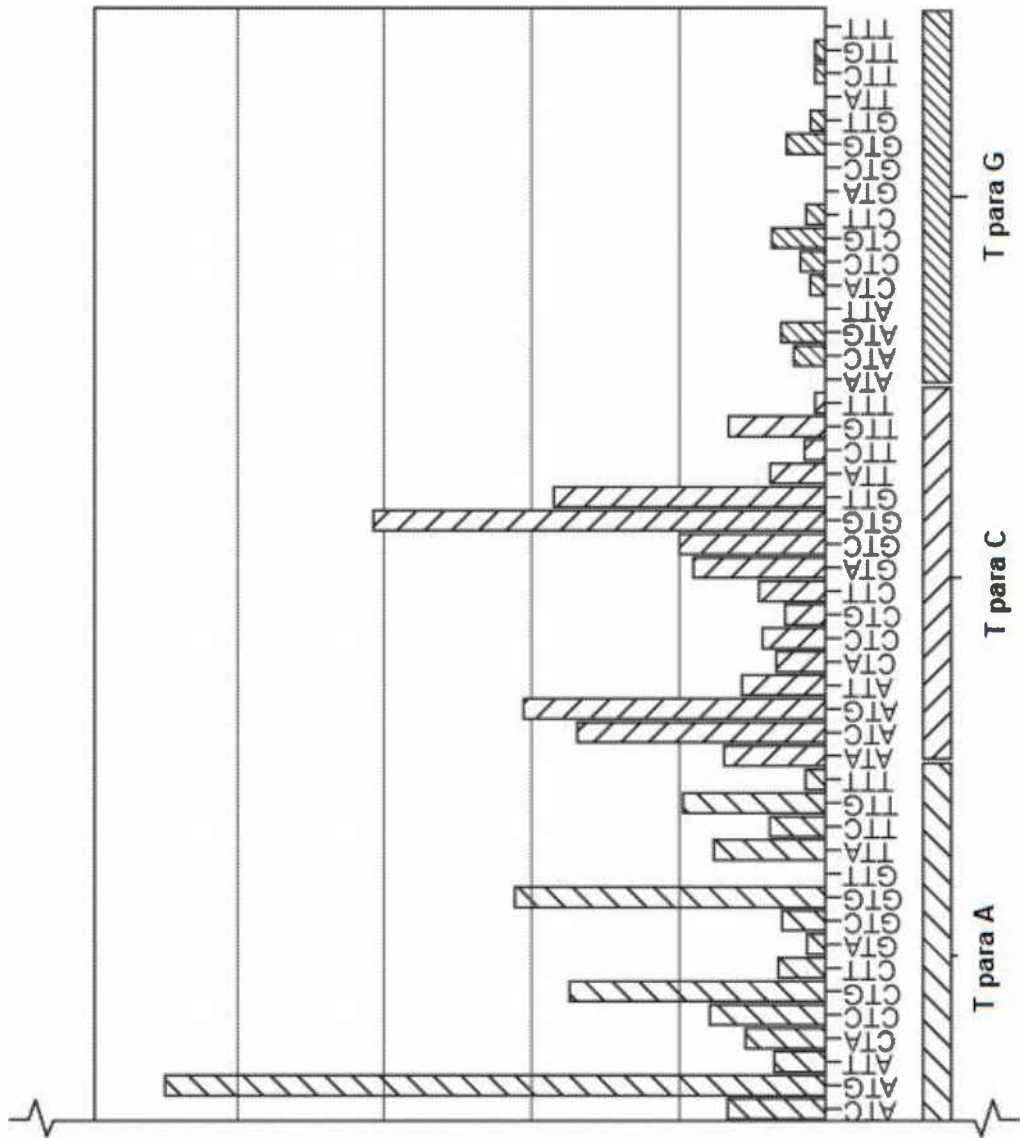
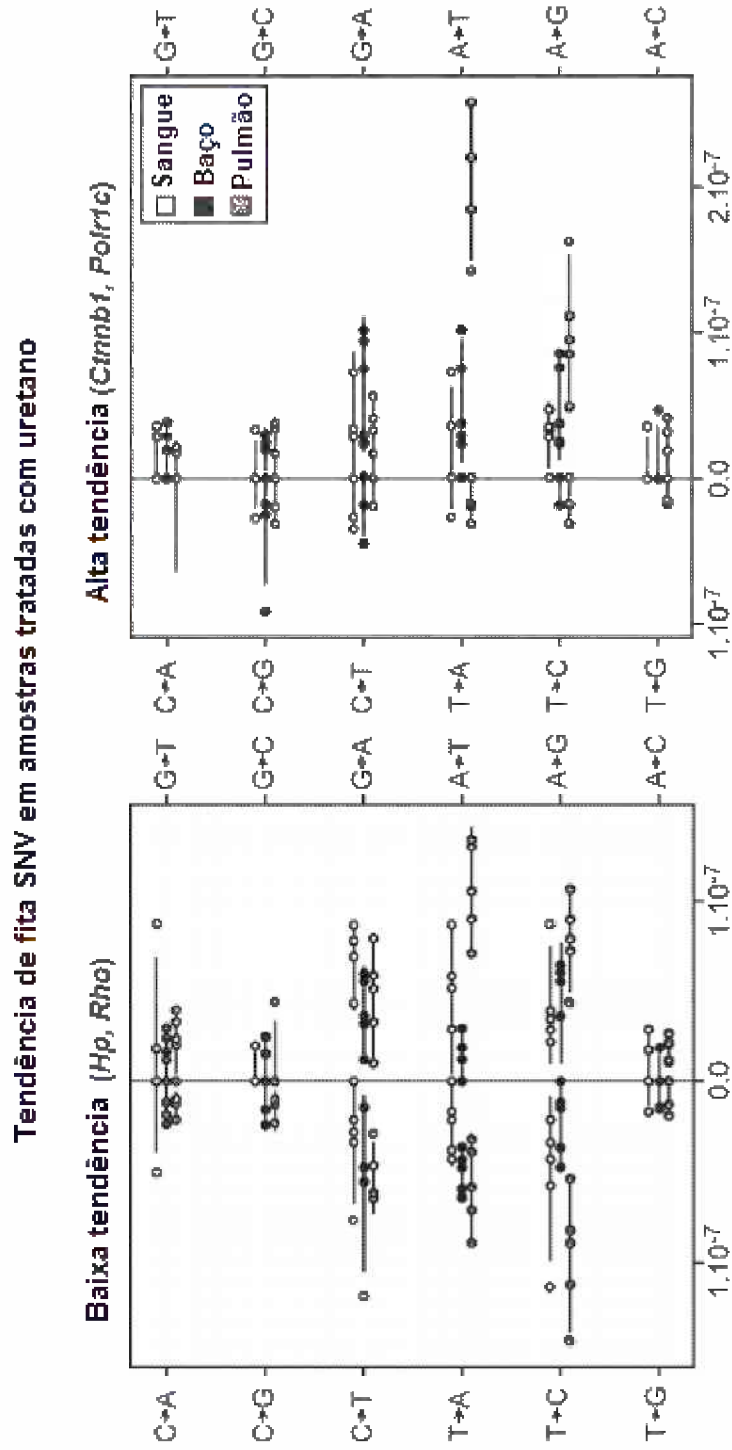


FIG. 12B (Cont.)



Frequência de mutações recíprocas em fita transcrita

FIG. 13

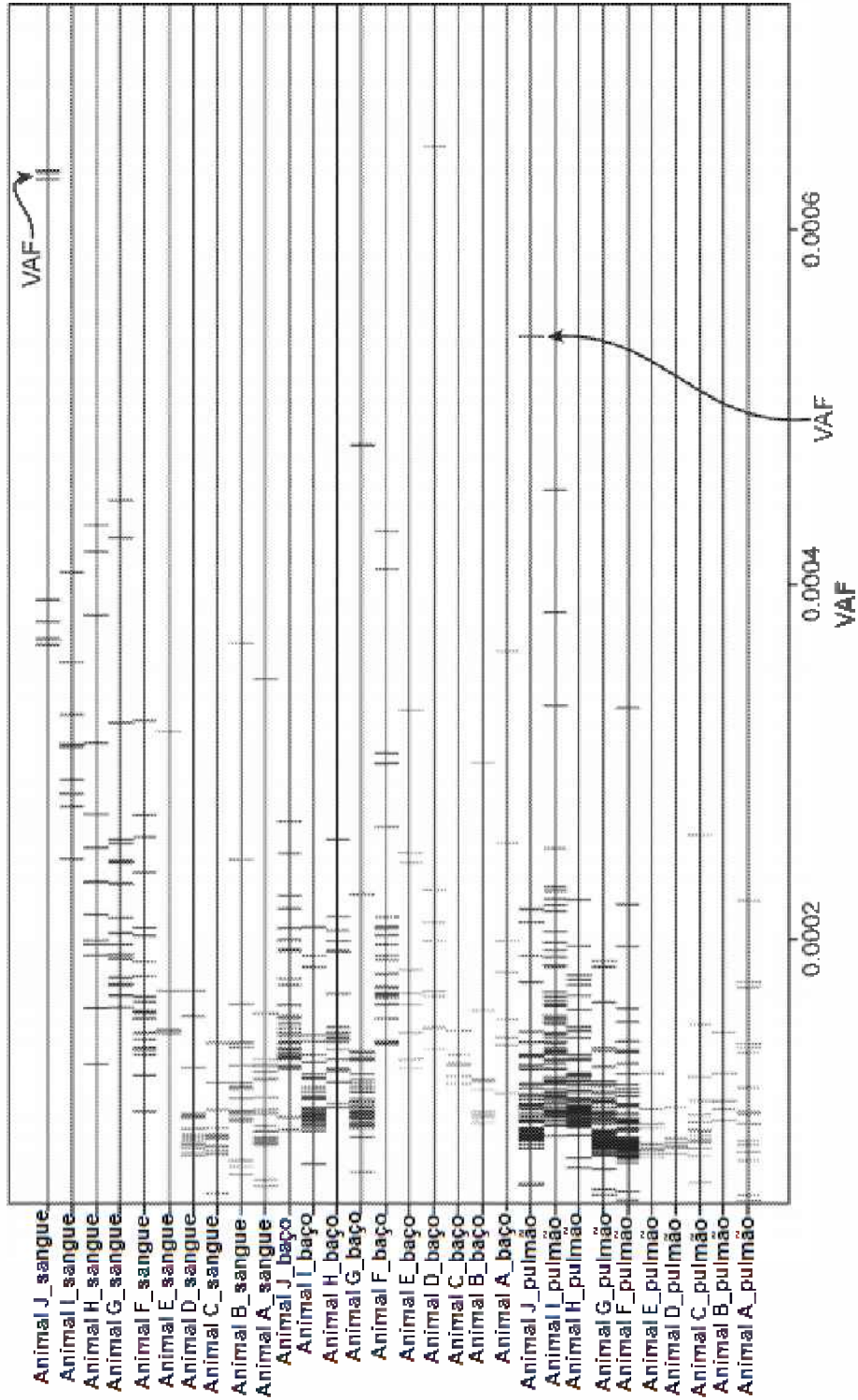


FIG. 14

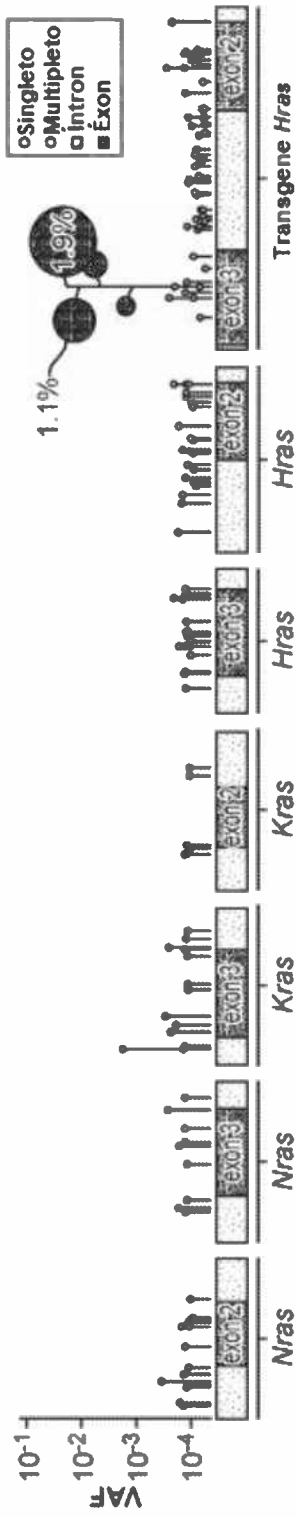


FIG. 15A

Éxon 3 HRAS humano

Amostra	Tecido	Tratamento	Subtipo	VD	Profundidade	Contexto	VAF
Animal A-sangue							
Animal A-pulmão							
Animal A-baço							
Animal B-sangue							
Animal B-pulmão	Pulmão	Uretano	T>A	300	16,425	CTG	1.82%
Animal B-baço							
Animal C-sangue							
Animal C-pulmão	Pulmão	Uretano	T>A	181	16,319	CTG	1.10%
Animal C-baço							
Animal D-sangue							
Animal D-pulmão	Pulmão	Uretano	T>A	58	13,692	CTG	0.42%
Animal D-baço							
Animal E-sangue	Pulmão	Uretano	T>A	17	14,706	CTG	0.11%
Animal E-pulmão							
Animal E-baço							
Animal F-sangue							
Animal F-pulmão							
Animal F-baço							
Animal G-sangue							
Animal G-pulmão							
Animal G-baço							
Animal H-sangue							
Animal H-pulmão							
Animal H-baço							
Animal I-sangue							
Animal I-pulmão							
Animal I-baço							
Animal J-sangue							
Animal J-pulmão							
Animal J-baço							
Sequência	G C T G T A C T C C T C C	G C T G T A C T C C T C C	G C T G T A C T C C T C C	G C T G T A C T C C T C C	G C T G T A C T C C T C C	G C T G T A C T C C T C C	G C T G T A C T C C T C C



FIG. 15B

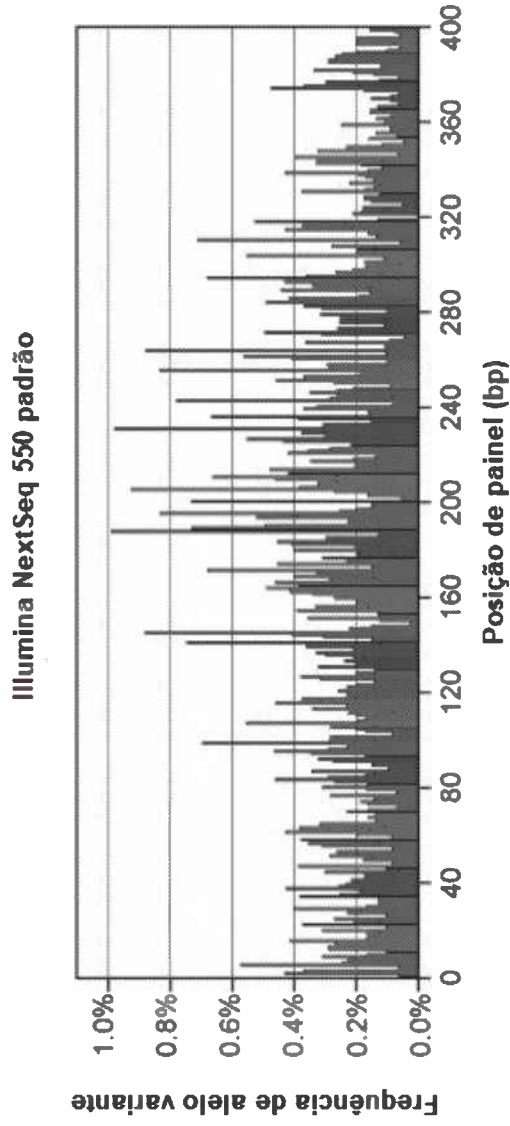


FIG. 16A

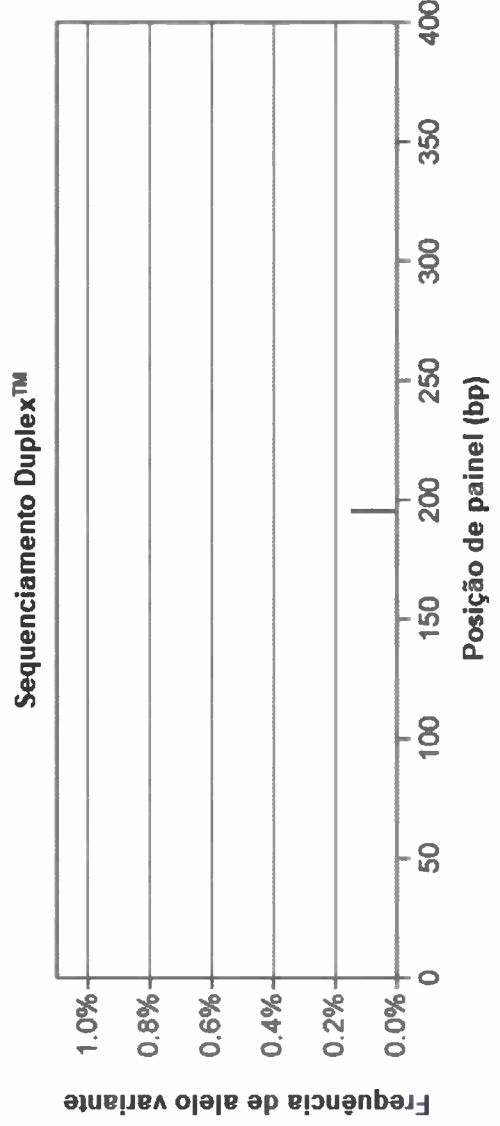


FIG. 16B

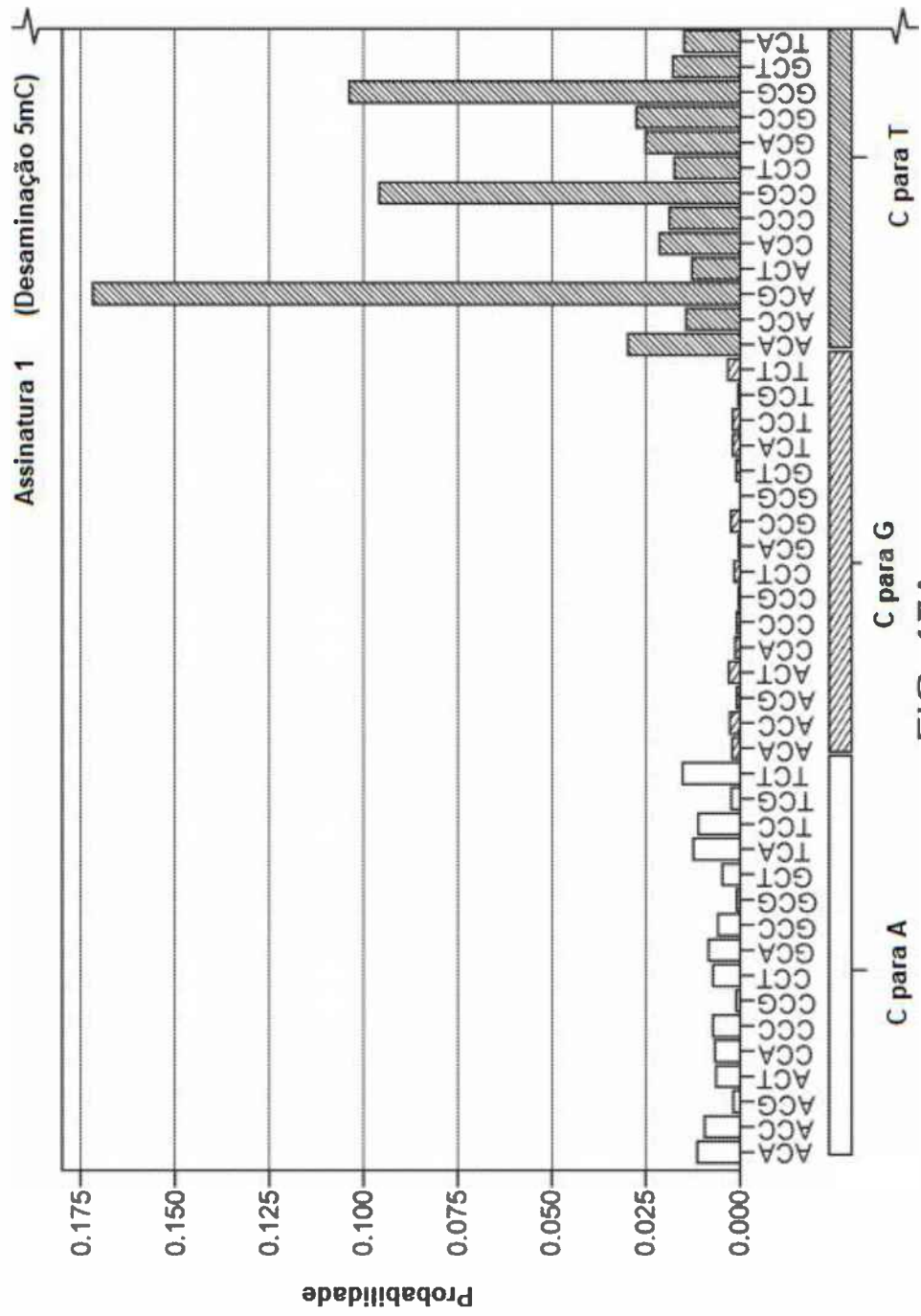


FIG. 17A

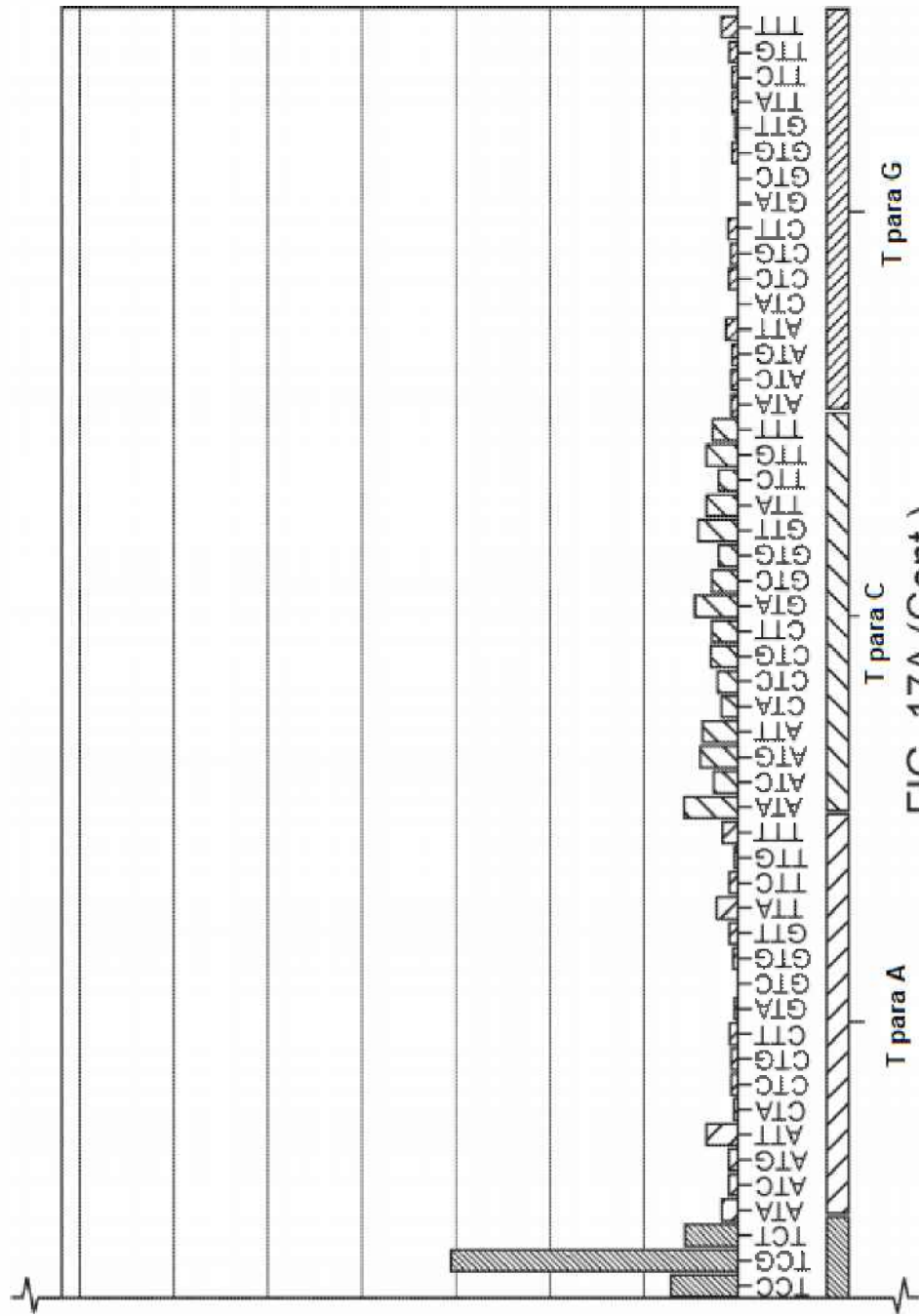


FIG. 17A (Cont.)

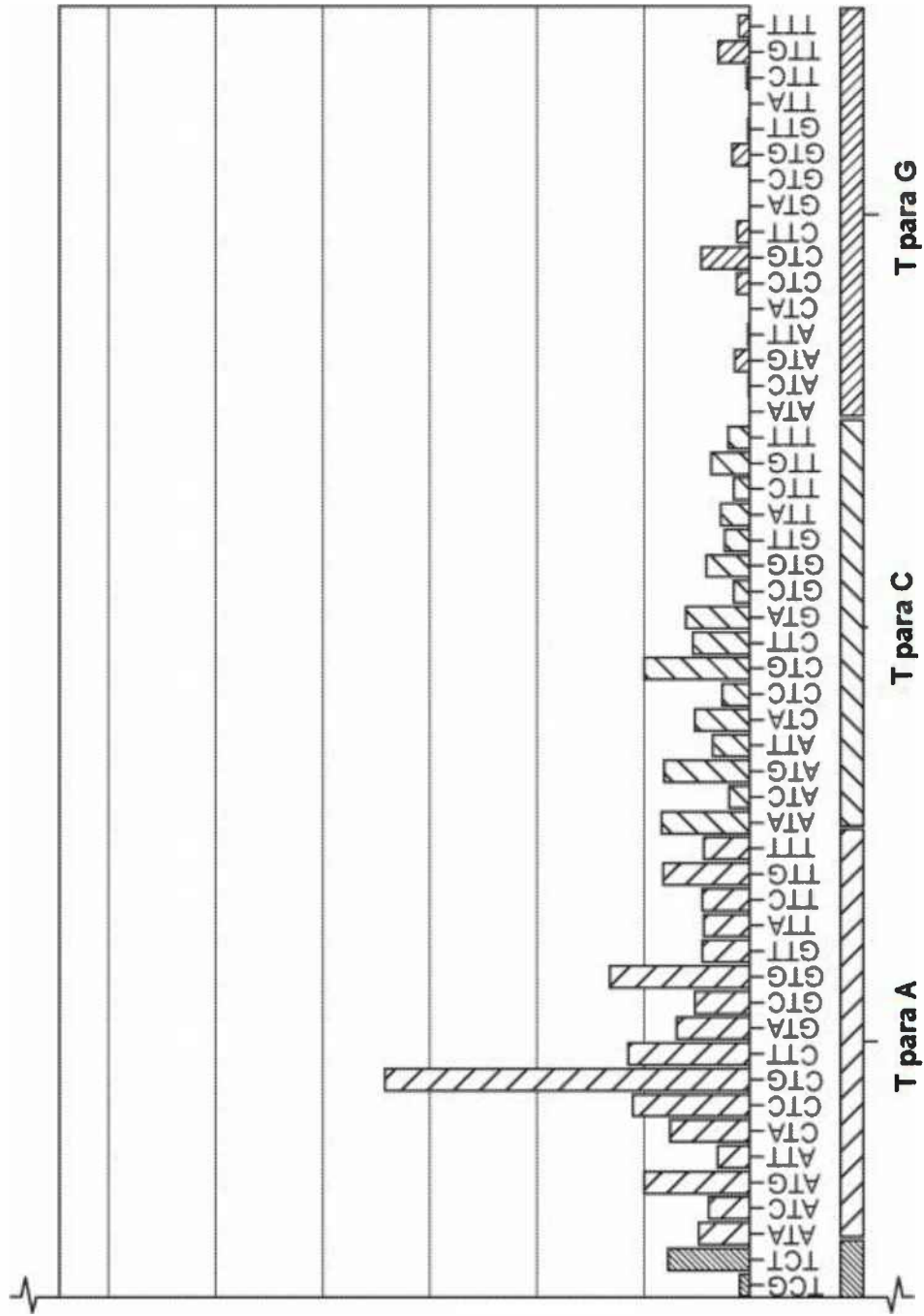


FIG. 17B (Cont.)

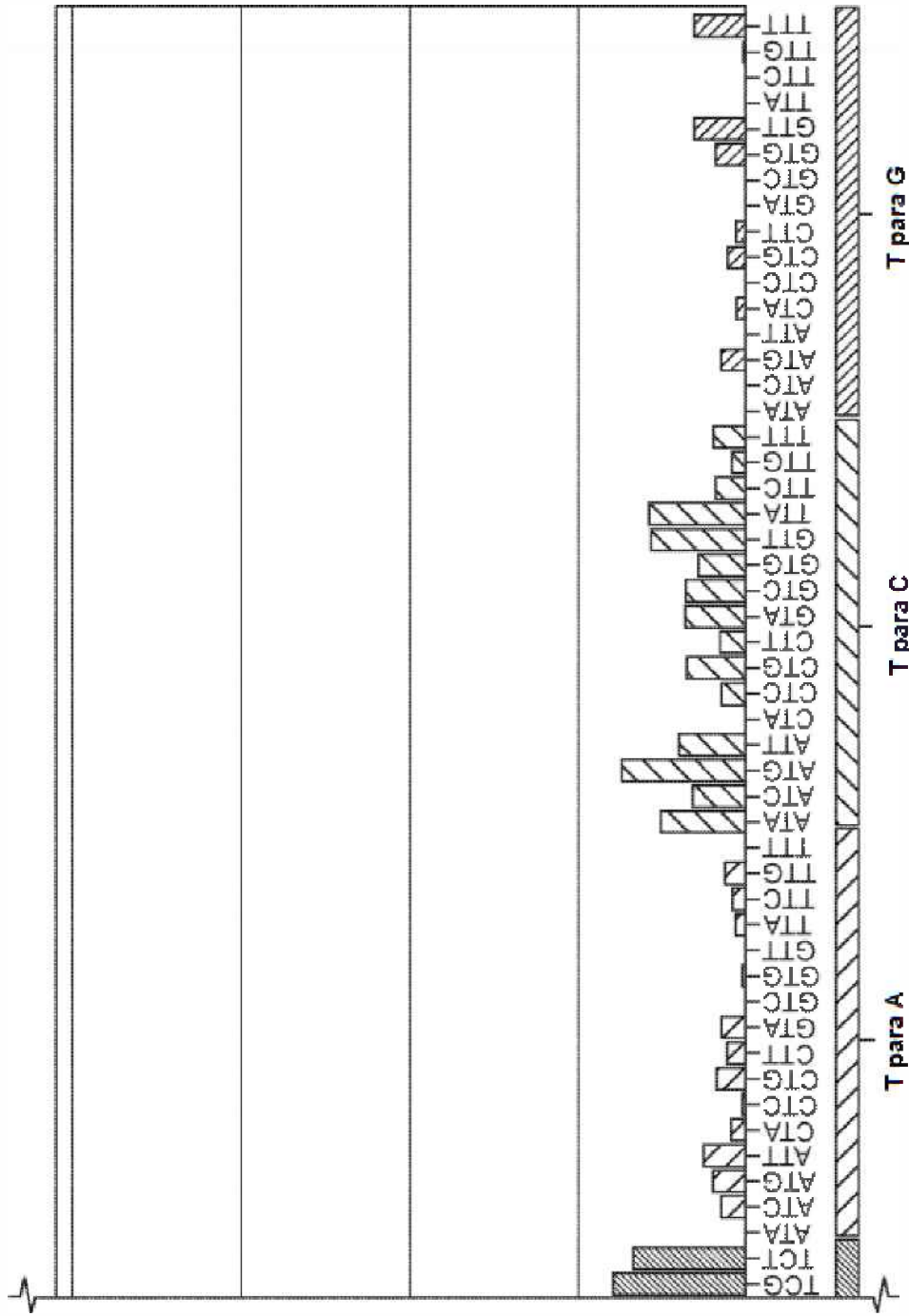


FIG. 17C (Cont.)

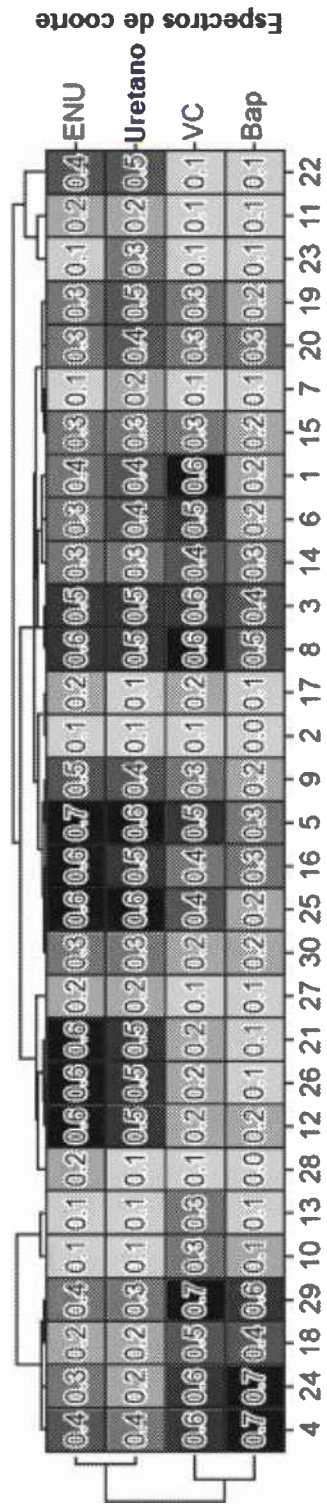


FIG. 18

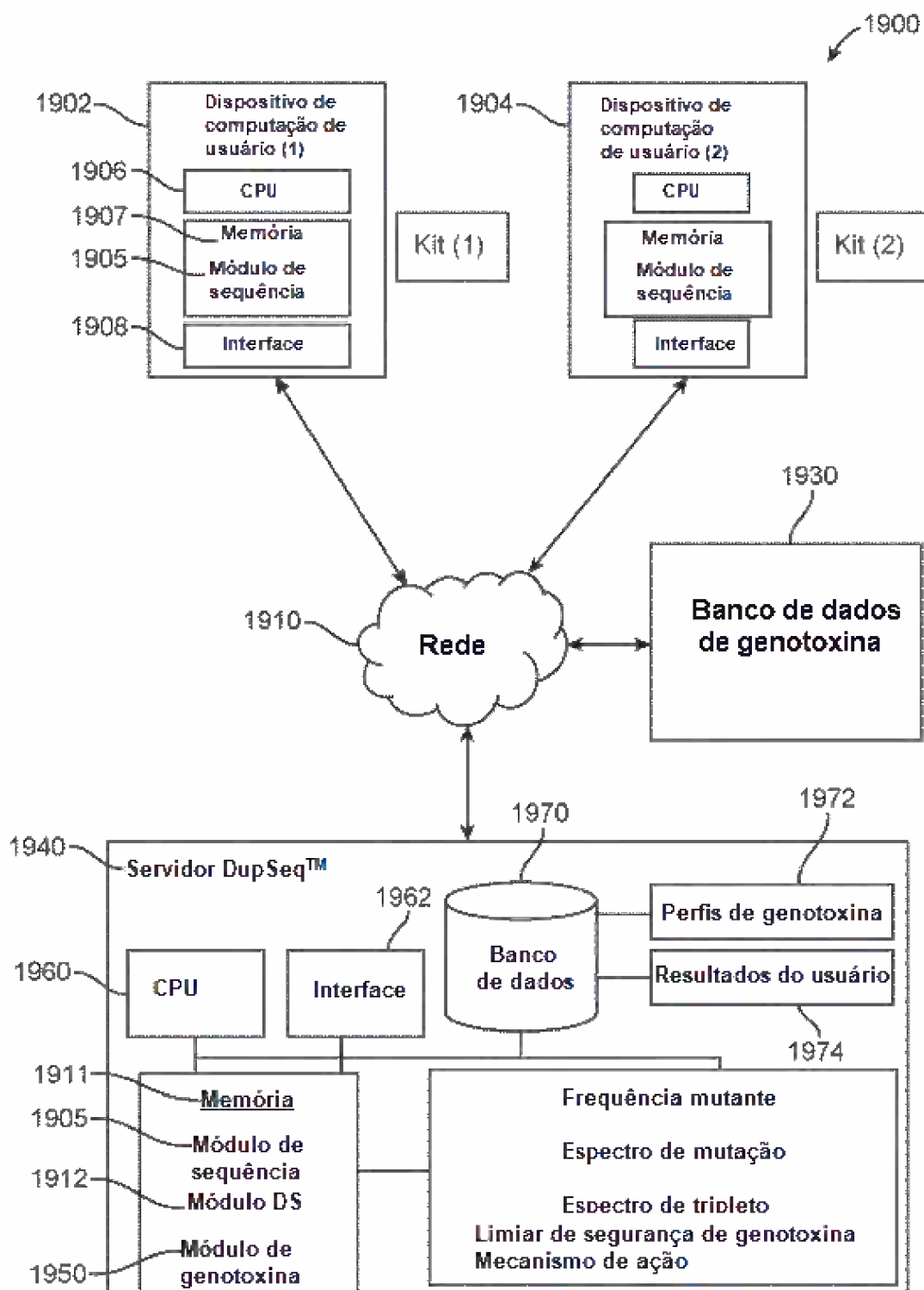


FIG. 19

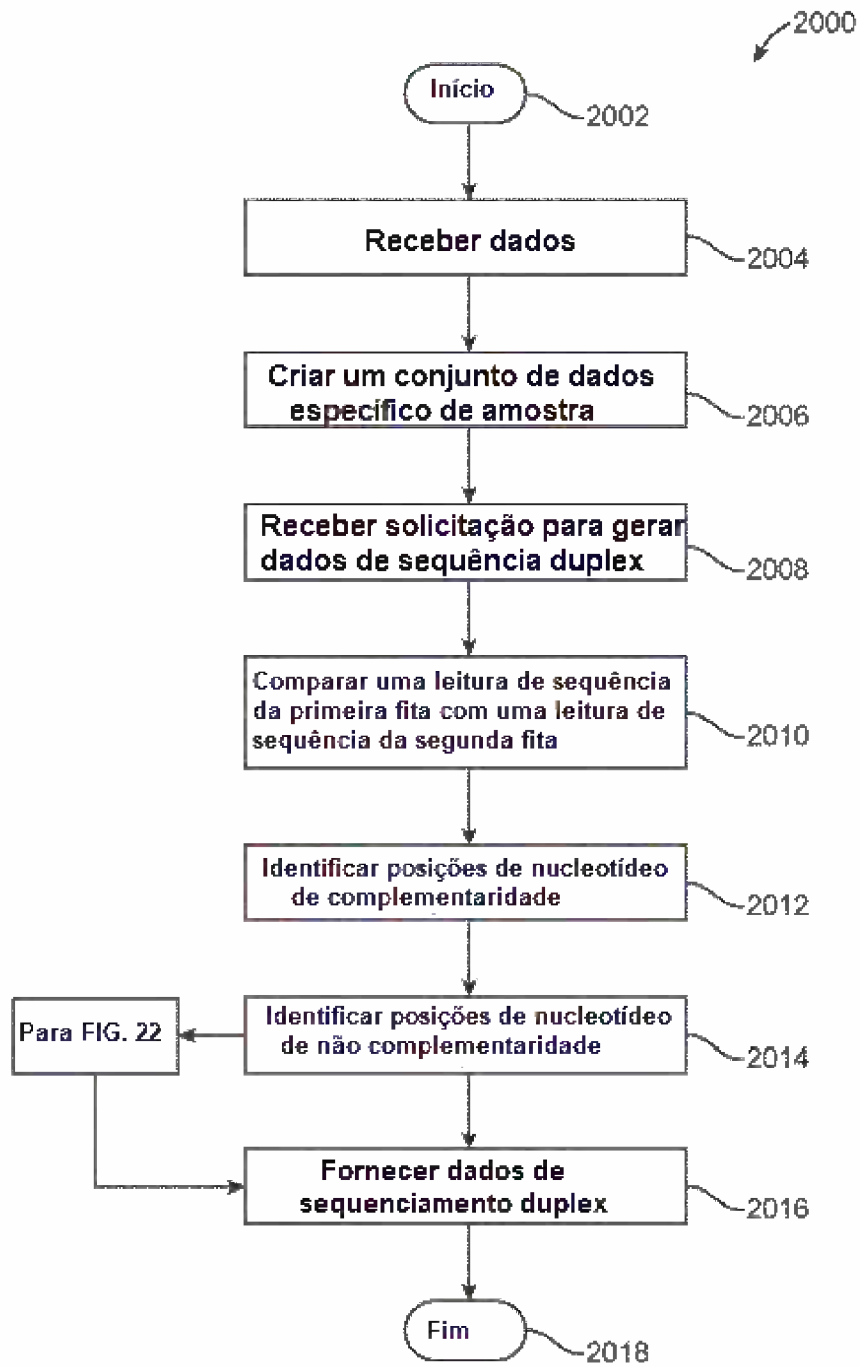


FIG. 20

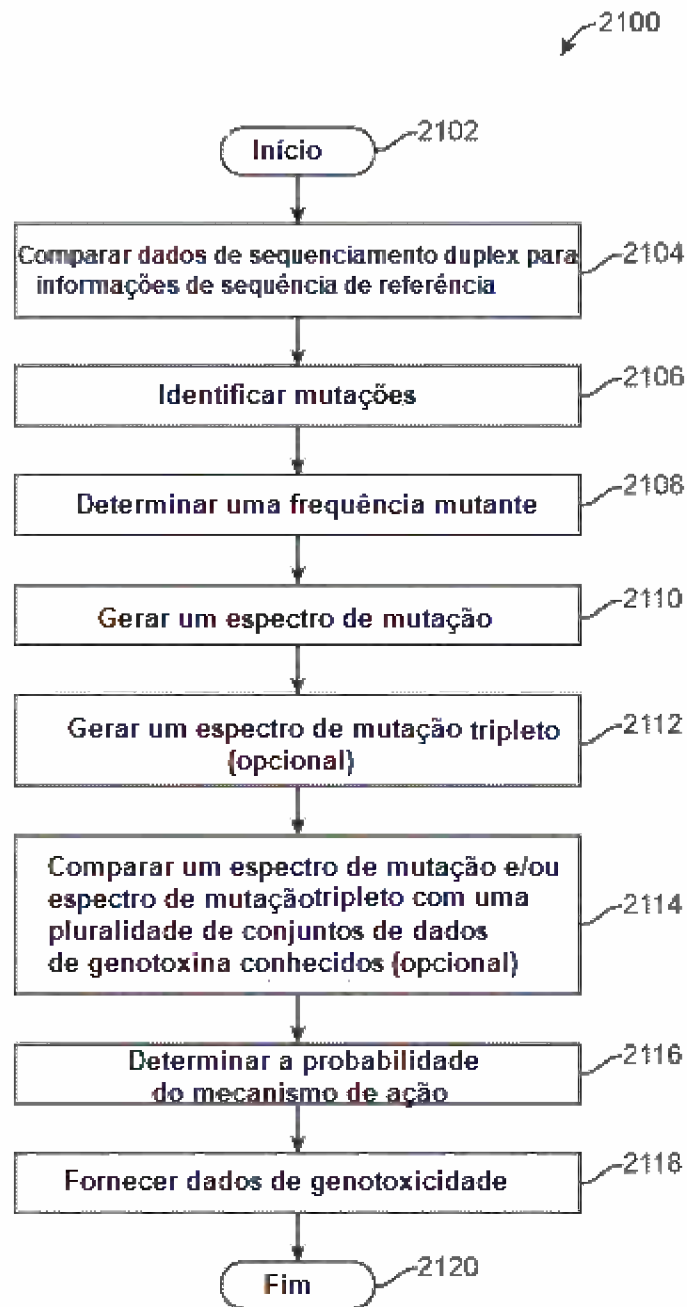


FIG. 21

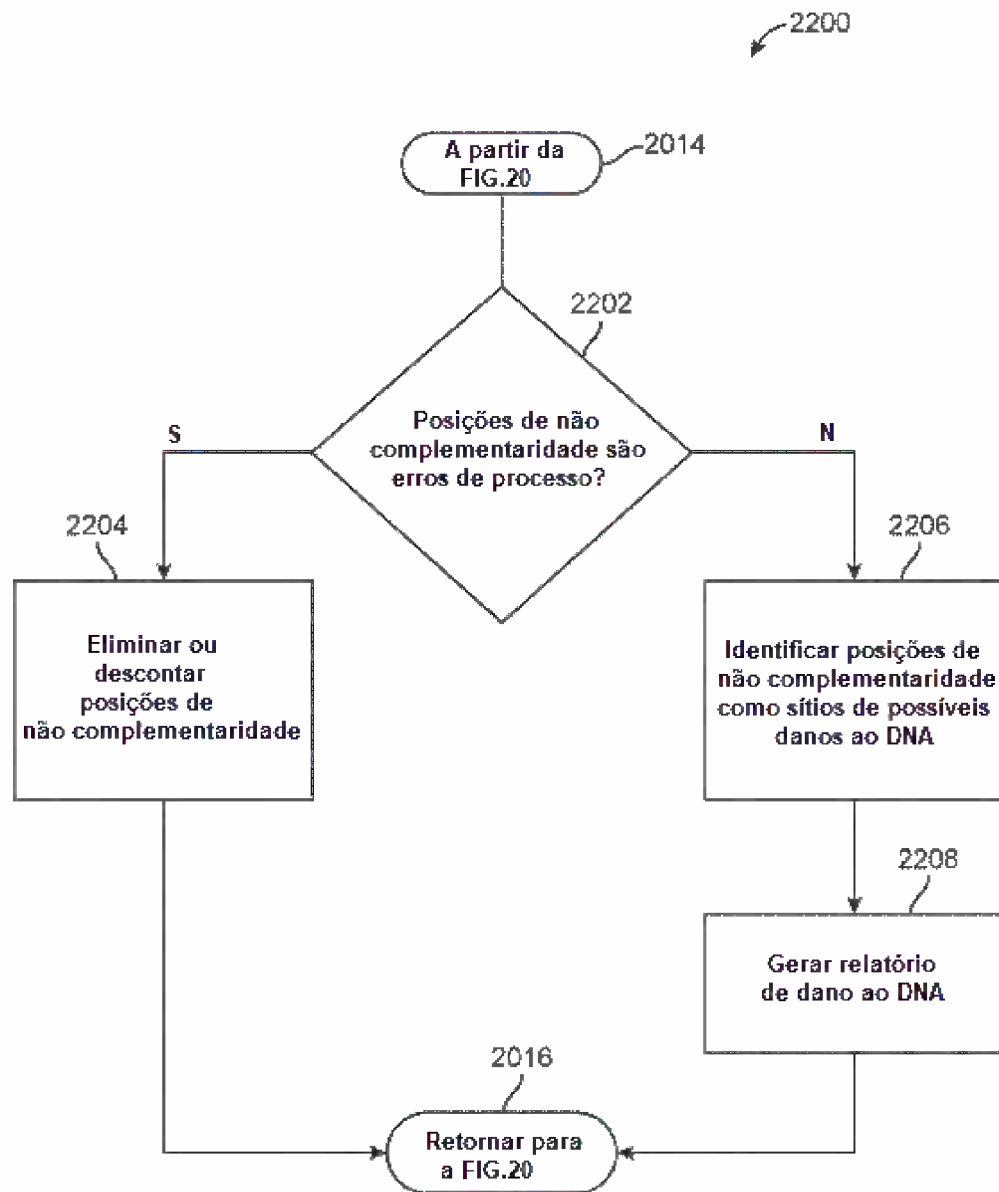


FIG. 22

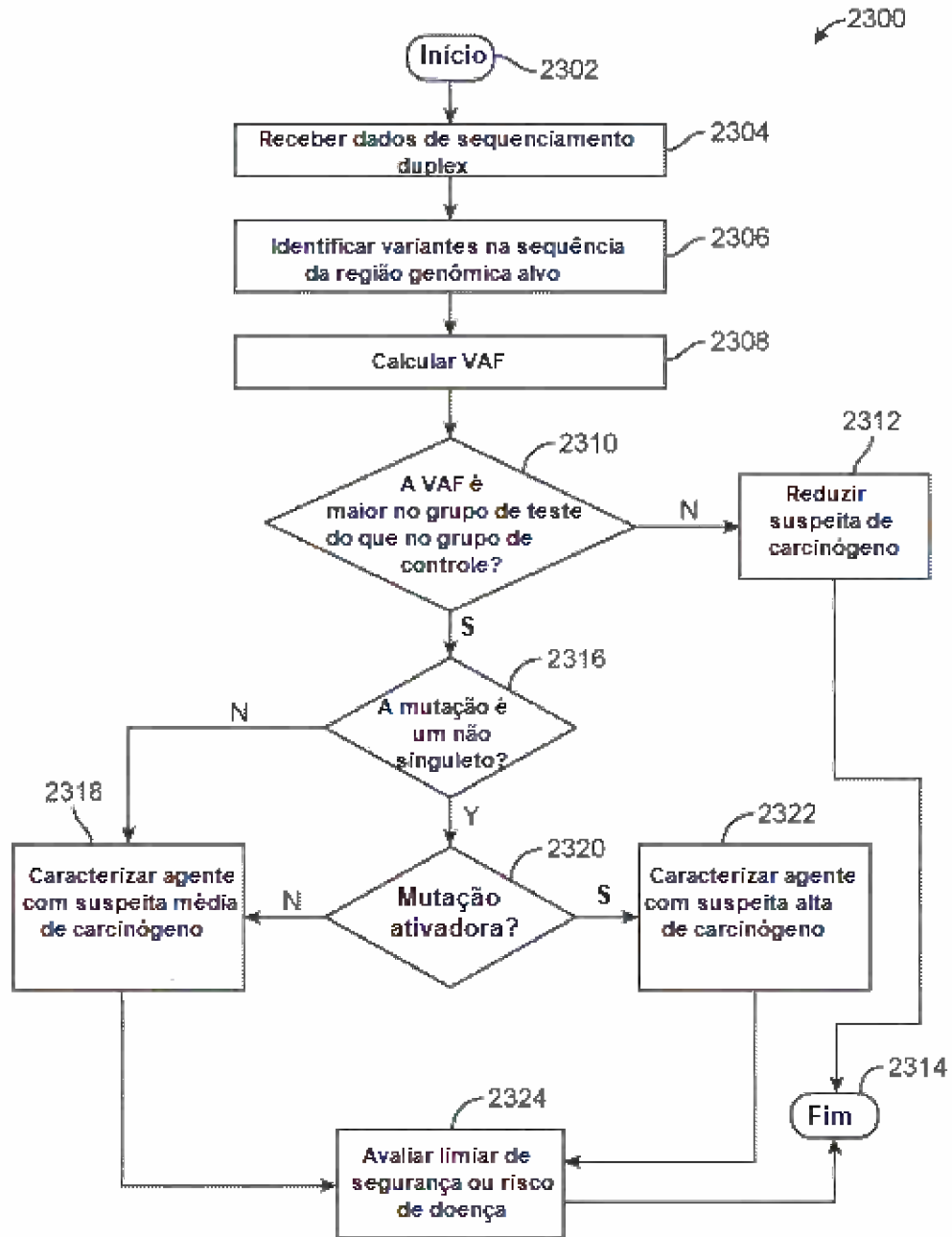


FIG. 23

RESUMO

MÉTODOS E REAGENTES PARA DETECTAR E AVALIAR A GENOTOXICIDADE

Métodos, sistemas e kits com reagentes para avaliar a genotoxicidade, são divulgados aqui. A genotoxicidade e seus mecanismos de ação podem ser determinados alguns dias após a exposição do sujeito. Algumas modalidades da tecnologia são direcionadas à utilização do Sequenciamento Duplex para avaliar um potencial genotóxico de um composto (por exemplo, um composto químico) em um sujeito exposto. Outras modalidades da tecnologia são direcionadas à utilização do Sequenciamento Duplex para determinar uma assinatura de mutação associada a um agente genotóxico; e/ou um nível limiar seguro de exposição à genotoxina. Modalidades adicionais da tecnologia são direcionadas para identificar um ou mais agentes genotóxicos aos quais um sujeito pode ter sido exposto comparando o espectro de mutação do DNA do sujeito com os espectros de mutação de compostos mutagênicos conhecidos. Depois que uma exposição à genotoxina em um sujeito é identificada ou confirmada, é fornecido um curso terapêutico profilático e/ou inibitório.