



República Federativa do Brasil  
Ministério do Desenvolvimento, Indústria  
e do Comércio Exterior  
Instituto Nacional da Propriedade Industrial

**(21) PI 0806565-9 A2**



\* B R P I 0 8 0 6 5 6 5 A 2 \*

(22) Data de Depósito: 10/01/2008  
**(43) Data da Publicação: 06/05/2014**  
**(RPI 2261)**

**(51) Int.Cl.:**  
**C12Q 1/68**

**(54) Título:** CAPTURA DE CONFORMAÇÃO DE  
CROMOSSOMO CIRCULAR

**(57) Resumo:**

**(30) Prioridade Unionista:** 05/10/2007 US 60/977.900,  
11/01/2007 US 60/999.750

**(66) Prioridade Interna:** 860446

**(73) Titular(es):** Erasmus University Medical Center

**(72) Inventor(es):** Frank Grosveld, Wouter De Laat

**(74) Procurador(es):** Alexandre Ferreira

**(86) Pedido Internacional:** PCT IB2008000625 de 10/01/2008

**(87) Publicação Internacional:** WO 2008/084405 de  
17/07/2008

## “CAPTURA DE CONFORMAÇÃO DE CROMOSSOMO CIRCULAR”

### Campo da invenção

A presente invenção diz respeito à análise da frequência de interação de duas ou mais sequências de nucleotídeos no espaço nuclear. Mudanças nas interações são usadas 5 como uma ferramenta para detectar rearranjos do genoma para diagnósticos e prognósticos.

### Antecedentes da invenção

Estudos na arquitetura nuclear de mamífero objetivam entender como 2 metros de DNA são dobrados em um núcleo de 10  $\mu\text{m}$  de diâmetro, permitindo ao mesmo tempo a expressão correta dos genes que especificam o tipo celular, e como esta é fielmente propagada durante cada ciclo celular. O progresso neste campo veio basicamente de estudos 10 microscópicos, que revelaram que os genomas são arranjados não aleatoriamente no espaço nuclear. Por exemplo, a heterocromatina densamente condensada é separada da eucromatina mais aberta e os cromossomos ocupam territórios distintos no espaço nuclear 2. Existe um relacionamento intrincado entre o posicionamento nuclear e a atividade transcricional. Embora a transcrição ocorra em todo o interior nuclear, os genes ativos que aglomeram-se nos cromossomos localizam-se preferivelmente na borda ou fora de seu território cromossômico. Genes individuais podem migrar mediante mudanças em seu estado de transcrição, medidos contra pontos de referência nuclear relativamente grandes, tais como territórios cromossômicos, centrômeros ou a periferia nuclear. Além disso, dezenas de genes ativamente transcritos de megabases separados no cromossomo podem ficar juntos no núcleo, demonstrado recentemente por hibridização *in situ* por fluorescência (FISH) para o locus de  $\beta$ -globina e alguns outros genes selecionados. Além da transcrição, a organização 15 genômica está associada com a coordenação de replicação, recombinação e a probabilidade de os loci translocarem (o que pode levar a malignidades) e o ajuste e reajuste de programas epigenéticos. Com base nestas observações, acredita-se que a organização arquitônica de DNA no núcleo celular seja uma contribuinte chave para a função genômica.

Diferentes ensaios foram desenvolvidos para permitir uma compreensão da organização espacial dos loci genômicos *in vivo*. Foi desenvolvido um ensaio, denominado RNA-TRAP (Carter et al. (2002) *Nat. Genet.* 32, 623), que envolve o alvejamento de peroxidase de raiz forte (HRP) para transcrições de RNA nascente, seguido por quantificação de deposição de biotina catalisada por HRP na cromatina ao redor.

Um outro ensaio que foi desenvolvido é denominado tecnologia de captura de conformação de cromossomo (3C), que fornece uma ferramenta para estudar a organização estrutural de uma região genômica. A tecnologia 3C envolve análise por PCR quantitativa de 35 frequências de reticulação entre dois dados fragmentos de restrição de DNA, que fornece uma medida de sua proximidade no espaço nuclear (ver figura 1). Originalmente desenvolvida para analisar a conformação de cromossomos em levedura (Dekker et al., 2002), esta

tecnologia foi adaptada para investigar o relacionamento entre expressão de gene e dobramento de cromatina em agrupamentos de gene de mamífero intricados (ver, por exemplo, Tolhuis et al., 2002; Palstra et al., 2003; e Drissen et al., 2004). Resumidamente, a tecnologia 3C envolve *in vivo* reticulação de formaldeído de células e digestão nuclear de cromatina 5 com uma enzima de restrição, seguida por ligação de fragmentos de DNA que foram reticulados em um complexo. Os produtos de ligação são a seguir quantificados por PCR. A etapa de amplificação por PCR exige o conhecimento da informação da sequência para cada um dos fragmentos de DNA que devem ser amplificados. Assim, a tecnologia 3C fornece uma medida de frequências de interação entre os fragmentos de DNA selecionados.

10 A tecnologia 3C foi desenvolvida para identificar elementos de interação entre partes selecionadas do genoma e ambas as técnicas exigem o desenho de iniciadores para todos os fragmentos de restrição analisados. Recentemente, foram desenvolvidas novas estratégias que permitem triar o genoma total de uma maneira não tendenciosa para segmentos de DNA que interagem fisicamente com um fragmento de DNA de escolha. Elas são 15 baseadas na tecnologia 3C e são coletivamente referidas como “tecnologia 4C”. A tecnologia 4C permite a triagem do genoma total de uma maneira não tendenciosa para segmentos de DNA que interagem fisicamente com um fragmento de DNA de escolha. A tecnologia 4C depende da ligação seletiva de fragmentos de DNA reticulado em um fragmento de restrição de escolha (a “isca”). Na tecnologia 4C, todos os fragmentos de DNA capturados pela isca 20 na população de células são simultaneamente amplificados por meio de PCR inversa, usando dois iniciadores específicos de isca que amplificam a partir de produtos de ligação circulares.

Essencialmente, duas estratégias podem ser adotadas para obter estes círculos de 25 DNA. Uma estratégia baseia-se na formação de círculos durante a etapa de ligação 3C padrão, isto é, enquanto o DNA ainda é reticulado (Zhao et al. (2006) *Nat Genet* 38, 1341-7). Aqui, a formação de círculo exige que ambas as extremidades do fragmento isca sejam ligadas em ambas as extremidades de um fragmento de restrição capturado. Se múltiplos 30 fragmentos de restrição forem reticulados um no outro, os círculos ainda podem ser formados, mas eles podem conter mais que um fragmento capturado e, portanto, serão maiores. Após a desreticulação, os fragmentos de DNA capturados são amplificados diretamente por 35 PCR inversa, usando iniciadores específicos de isca voltados para fora. Enzimas de restrição que reconhecem quatro ou seis pares de bases podem ser usadas nesta configuração. Cortadores quatro são preferidos neste método, embora, uma vez que eles produzem fragmentos de restrição menores (tamanho médio 256 bp, versus ~4 kb para cortadores seis) e a amplificação por PCR linear dos fragmentos de DNA capturados exige que o tamanho médio do produto seja pequeno. Essencialmente, este método compreende, portanto, as etapas de: (a) fornecer uma amostra de DNA reticulado; (b) digerir o DNA reticulado com

uma enzima de restrição primária, tal como um cortador de 4 bp ou um de 5 bp; (c) ligar as sequências de nucleotídeos reticulados; (d) inverter a reticulação e (e) amplificar uma ou mais sequências de nucleotídeos de interesse usando pelo menos dois oligonucleotídeos iniciadores, em que cada iniciador hibridiza nas sequências de DNA que flanqueiam as sequências de nucleotídeos de interesse. A(s) sequência(s) amplificada(s) pode(m) ser hibridizada(s) em um arranjo afim de auxiliar na determinação da frequência de interação entre as sequências de DNA.

A segunda estratégia baseia-se vantajosamente na formação de círculos de DNA após a cromatina ter sido desreticulada da maneira aqui descrita e em nosso pedido correspondente WO2007/004057. Da maneira aqui descrita, a tecnologia 4C permite uma pesquisa genômica não tendenciosa de fragmentos de DNA que interagem com um locus de escolha. Resumidamente, a análise 3C é realizada de maneira usual, mas omitindo a etapa de PCR. O molde de 3C contém uma sequência alvo ou "isca" (por exemplo, um fragmento de restrição de escolha que inclui um gene selecionado) ligada a muitas diferentes sequências de nucleotídeos de interesse (representando este ambiente genômico do gene). O molde é clivado por uma outra enzima de restrição secundária e, subsequentemente, religado para formar círculos de DNA pequenos. Vantajosamente, uma ou mais sequências de nucleotídeos de interesse que estão ligadas à sequência de nucleotídeos alvo são amplificadas usando pelo menos dois oligonucleotídeos iniciadores, em que pelo menos um iniciador hibridiza na sequência alvo. Preferivelmente, o segundo iniciador também hibridiza na sequência alvo, de maneira tal que ambos os iniciadores flanqueiam o nucleotídeo de interesse. Alternativamente, o segundo iniciador hibridiza em uma sequência adaptadora que está ligada ao sítio de restrição secundária, de maneira tal que os dois iniciadores flanqueiam o nucleotídeo de interesse. Tipicamente, isto produz um padrão de fragmentos de PCR que é altamente reproduzível entre reações de amplificação independentes e específico para um dado tecido. HindIII e DpnII podem ser usados como enzimas de restrição primária e secundária. A seguir, os fragmentos amplificados podem ser marcados e, opcionalmente, hibridizados em um arranjo, tipicamente contra uma amostra controle contendo DNA genômico digerido com a mesma combinação de enzimas de restrição. Portanto, a tecnologia 3C foi modificada de maneira tal que todas as sequências de nucleotídeos de interesse que interagem com uma sequência de nucleotídeos alvo sejam amplificadas. Praticamente, isto significa que em vez de realizar uma reação de amplificação com iniciadores que são específicos para os fragmentos que deseja-se analisar, uma amplificação é realizada usando oligonucleotídeo(s) iniciador(s) que hibridiza(m) em uma sequência de DNA que flanqueia as sequências de nucleotídeos de interesse. Vantajosamente, 4C não é tendenciosa para o desenho de iniciadores de PCR que estão incluídos na etapa de amplificação por PCR e, portanto, podem ser usados para pesquisar os elementos de DNA interagentes no genoma completo.

Existe uma importante necessidade da tecnologia de alto desempenho que possa triar sistematicamente todo o genoma de uma maneira não tendenciosa para os loci de DNA que fazem contato um com o outro no espaço nuclear.

Além disso, existe uma necessidade de melhorias em tais tecnologias que permitem a análise simultânea de múltiplas interações que ocorrem com múltiplas sequências no genoma, e de analisar o genoma para inserções, deleções, translocações, inversões e rearranjos que ocorrem em locais desconhecidos e que podem estar associados a uma doença.

A presente invenção busca fornecer melhorias na tecnologia 3C e 4C e técnicas relacionadas a isto.

#### 10 Sumário da invenção

A tecnologia 3C aplicada atualmente permite apenas análise de um número limitado de interações DNA-DNA selecionadas em virtude das limitações da etapa de amplificação por PCR, que exige conhecimento de informação da sequência específica para cada fragmento a ser analisado. Além disso, a seleção de fragmentos de restrição como candidatos para interações de DNA de longo alcance exige uma quantidade substancial de conhecimento anterior (por exemplo, a localização de sítios hipersensíveis) dos locus de interesse, que normalmente não é disponível. Dada a relevância funcional de muitas interações DNA-DNA de longo alcance descritas até agora, a capacidade de triar aleatoriamente elementos de DNA que laçam uma sequência de interesse, tal como um promotor, melhorador, isolante, silenciador, origem de replicação ou MAR/SAR de gene, ou uma região genômica de interesse, tal como uma região densa em gene ou pobre em gene ou elemento repetitivo, pode facilitar bastante o mapeamento de sequências envolvidas em uma rede regulatória.

A presente invenção diz respeito a tecnologia 4C (isto é, captura e caracteriza cromatina co-localizada) e melhorias nessa, que permitem a análise de alto desempenho da frequência de interação de duas ou mais sequências de nucleotídeos no espaço nuclear. A tecnologia 4C pode ser usada para identificar interações DNA-DNA de longo alcance (por exemplo, estudar dobramento de cromossomo), mas também para identificar rearranjos genômicos equilibrados e desequilibrados, tais como translocações, inversões, deleções, amplificações, etc., que podem fundamentar um traço ou doença em sujeitos humanos.

30 Rotineiramente, a tecnologia 4C envolve o uso de microarranjos para analisar os fragmentos de DNA capturados por uma única sequência alvo selecionada ("isca") (Simonis et al., *Nature Genetics* 2006). Os microarranjos têm a desvantagem que eles oferecem uma faixa dinâmica limitada, uma vez que sondas presentes nos arranjos podem ser saturadas, o que torna uma análise quantitativa de intensidades de sinal mais difícil. O sequenciamento de alto desempenho evita este problema, já que ele oferece uma faixa dinâmica ilimitada. É também quantitativo, uma vez que ele fornece números absolutos de sequências.

Além do mais, é preferível analisar interações de DNA com múltiplas sequências al-

vos simultaneamente. Isto é válido para todas as aplicações a base de 4C e, em particular, para a análise a base de 4C de rearranjos genômicos. A tecnologia 4C pode ser usada como uma ferramenta de diagnóstico para permitir a varredura do genoma total de uma maneira não tendenciosa a procura da presença de rearranjos genômicos. Uma série de sequências alvos ao longo de cada cromossomo que captura ao mesmo tempo todas as sequências (isto é, fragmentos de restrição) pode ser usada. Subsequentemente, afim de identificar o rearranjo genômico, é identificada a sequência que foi capturada por qual “isca”.

Os fragmentos capturados podem ser sequenciados para cada sequência alvo (“isca”) separadamente. Embora, preferivelmente, todos os produtos de ligação formados com todas as sequências alvos sejam analisados simultaneamente. Para isto, cada leitura precisa ser direcionada à junção de ligação e fornece informação da sequência suficiente para identificar inambiguamente tanto a sequência alvo quanto a sequência capturada.

#### Aspectos sumários da presente invenção

Os aspectos da presente invenção são apresentados nas reivindicações em anexo.

Em um primeiro aspecto, é fornecido um método para analisar a frequência de interação de uma sequência de nucleotídeos alvo com uma ou mais sequências de nucleotídeos de interesse (por exemplo, um ou mais loci genômicos) compreendendo as etapas de: (a) fornecer uma amostra de DNA reticulado; (b) digerir o DNA reticulado com uma enzima de restrição primária; (c) ligar as sequências de nucleotídeos reticulados; (d) inverter a reticulação; (e) digerir opcionalmente as sequências de nucleotídeos com uma enzima de restrição secundária; (f) ligar opcionalmente uma ou mais sequências de DNA de composição de nucleotídeo conhecida ao(s) sítio(s) de digestão de enzima de restrição secundária disponível(s) que flanqueia(m) uma ou mais sequências de nucleotídeos de interesse; (g) amplificar uma ou mais sequências de nucleotídeos de interesse usando pelo menos dois oligonucleotídeos iniciadores, em que cada iniciador hibridiza as sequências de DNA que flanqueiam as sequências de nucleotídeos de interesse; (h) hibridizar na(s) sequência(s) amplificada(s) em um arranjo; e (i) determinar a frequência de interação entre as sequências de DNA.

Em um segundo aspecto, é fornecido um método para analisar a frequência de interação de uma sequência de nucleotídeos alvo com uma ou mais sequências de nucleotídeos (por exemplo, um ou mais loci genômicos) compreendendo as etapas de: (a) fornecer uma amostra de DNA reticulado; (b) digerir o DNA reticulado com uma enzima de restrição primária; (c) ligar as sequências de nucleotídeos reticulados; (d) inverter a reticulação; (e) digerir opcionalmente as sequências de nucleotídeos com uma enzima de restrição secundária; (f) circular as sequências de nucleotídeos; (g) amplificar uma ou mais sequências de nucleotídeos que estão ligadas à sequência de nucleotídeos alvo; (h) hibridizar opcionalmente nas sequências amplificadas em um arranjo ou analisar as sequências amplificadas

pelo sequenciamento de alto desempenho; e (i) determinar a frequência de interação entre as sequências de DNA.

Em um terceiro aspecto, é fornecido um método para identificar uma ou mais interações DNA-DNA que são indicativas de uma condição de doença particular compreendendo a etapa de realizar as etapas (a)-(i) do método de acordo com o primeiro aspecto, em que na etapa (a) uma amostra de DNA reticulado é fornecida a partir de uma célula doente e uma não doente, e em que uma diferença entre a frequência de interação entre as sequências de DNA das células doentes e não doentes indica uma diferença na organização linear dos moldes do cromossomo (por exemplo, um rearranjo genômico), que é indicativa de um traço particular ou condição de doença.

Em um quarto aspecto, é fornecido um método de diagnose ou prognose de uma doença ou síndrome causada ou associada a uma mudança em uma interação DNA-DNA, compreendendo a etapa de realizar as etapas (a)-(i) do método de acordo com o primeiro aspecto, em que a etapa (a) compreende fornecer uma amostra de DNA reticulado de um sujeito; e em que a etapa (i) compreende comparar a frequência de interação entre as sequências de DNA com aquelas de um controle não afetado; em que uma diferença entre o valor obtido do controle e o valor obtido do sujeito é indicativa que o sujeito está sofrendo da doença ou síndrome ou é indicativa que o sujeito sofrerá da doença ou síndrome.

Em um quinto aspecto, é fornecido um método de diagnose ou prognose de uma doença ou síndrome causada ou associada a uma mudança em uma interação DNA-DNA compreendendo a etapa de: realizar as etapas (a)-(i) do método de acordo com o primeiro aspecto, em que a etapa (a) compreende fornecer uma amostra de DNA reticulado de um sujeito; e em que o dito método compreende a etapa adicional de: (j) identificar um ou mais loci que foram submetidos a um rearranjo genômico que está associado com uma doença.

Em um sexto aspecto, é fornecido um método de ensaio para identificar um ou mais agentes que modulam uma interação DNA-DNA compreendendo as etapas de: (a) colocar uma amostra em contato com um ou mais agentes; e (b) realizar as etapas (a) a (i) do método de acordo com o primeiro aspecto, em que a etapa (a) compreende fornecer DNA reticulado da amostra; em que uma diferença entre (i) a frequência de interação entre as sequências de DNA na presença do agente e (ii) a frequência de interação entre as sequências de DNA na ausência do agente é indicativa de um agente que modula a interação DNA-DNA.

Em um sétimo aspecto, é fornecido um método para detectar o local de um rearranjo equilibrado e/ou desequilibrado (por exemplo, uma translocação) compreendendo a etapa de: (a) realizar as etapas (a) a (i) do método de acordo com o primeiro aspecto; e (b) comparar a frequência de interação entre as sequências de DNA com aquelas de um controle; em que uma transição de baixa ou alta frequência de interação DNA-DNA na amostra compara-

da ao controle é indicativa do local de um ponto de quebra.

Em um oitavo aspecto, é fornecido um método para detectar o local de uma inversão equilibrada e/ou desequilibrada compreendendo as etapas de: (a) realizar as etapas (a) a (i) do método de acordo com o primeiro aspecto; e (b) comparar a frequência de interação entre as sequências de DNA com aquelas de um controle; em que um padrão inverso de frequências de interação DNA-DNA para a amostra comparada ao controle é indicativo de uma inversão.

Em um nono aspecto, é fornecido um método para detectar o local de uma deleção compreendendo as etapas de: (a) realizar as etapas (a) a (i) do método de acordo com o primeiro aspecto; e (b) comparar a frequência de interação entre as sequências de DNA com aquelas de um controle; em que uma redução na frequência de interação DNA-DNA para a amostra comparada ao controle é indicativa de deleção.

Em um décimo aspecto, é fornecido um método para detectar o local de uma duplicação compreendendo as etapas de: (a) realizar as etapas (a) a (i) do método de acordo com o primeiro aspecto; e (b) comparar a frequência de interação entre as sequências de DNA com aquelas de um controle; em que um aumento ou uma diminuição na frequência de interação DNA-DNA para a amostra em questão comparada ao controle é indicativo de uma duplicação ou inserção.

Em um décimo primeiro aspecto, é fornecido um agente obtido ou obtenível pelo método de ensaio aqui descrito.

Em um décimo segundo aspecto, é fornecido um método para analisar a frequência de interação de uma ou mais sequências alvos de nucleotídeos com uma ou mais sequências de nucleotídeos de interesse (por exemplo, um ou mais loci genômicos) compreendendo as etapas de: (a) fornecer uma amostra de DNA reticulado; (b) digerir o DNA reticulado com uma enzima de restrição primária; (c) ligar as sequências de nucleotídeos reticulados; (d) inverter a reticulação; e (e) sequenciar as sequências de nucleotídeos ligadas.

Em um décimo terceiro aspecto, é fornecido um método para determinar a presença de um rearranjo genômico em uma amostra compreendendo as etapas de: (a) fornecer uma amostra de ácido nucléico (por exemplo, DNA genômico), em que o dito ácido nucléico compreende uma sequência de nucleotídeos de sequência conhecida adjacente ao local do rearranjo genômico suspeito; (b) digerir o DNA com uma enzima de restrição primária para formar uma pluralidade de fragmentos de restrição; (c) purificar opcionalmente os fragmentos de restrição; (d) ligar os fragmentos de restrição para formar DNA circular; (e) purificar opcionalmente o DNA circular; (f) digerir o DNA circular com uma enzima de restrição secundária para formar uma pluralidade de fragmentos de restrição; (g) ligar os fragmentos de restrição para formar DNA circular; (h) amplificar o rearranjo genômico suspeito usando um ou mais iniciadores que hibridizam na sequência de nucleotídeos de sequência conhecida; e

(i) sequenciar o rearranjo genômico suspeito.

Em um décimo quarto aspecto, é fornecida uma base de dados de sequência de ácidos nucléicos de cerca de 6-50 pares de bases que flanqueiam diretamente e, opcionalmente, incluem o sítio de reconhecimento de enzima de restrição primária ou o sítio de reconhecimento de enzima de restrição secundária de cada sequência alvo.

Em um décimo quinto aspecto, é fornecida uma base de dados de sequência de ácidos nucléicos de cerca de 12- 50 pares de bases que flanqueiam diretamente todos os sítios de reconhecimento de enzima de restrição primária e secundária relevantes no genoma.

Em um décimo sexto aspecto, é fornecido o uso da base de dados de sequência de ácidos nucléicos para determinar a posição genômica de cada uma das sequências capturadas identificadas.

Em um décimo sétimo aspecto, é fornecido um método ou um agente ou uma base de dados ou um uso substancialmente da maneira aqui descrita e com relação a qualquer dos exemplos ou figuras.

#### Modalidades da invenção

Adequadamente, a reação de ligação na etapa (c) ou (f) resulta na formação de círculos de DNA.

Adequadamente, a etapa (h) compreende a análise de produtos de ligação entre sequências alvos e sequências reticuladas de interesse por meio de sequenciamento (por exemplo, o sequenciamento de alto desempenho).

Adequadamente, o método é para analisar a frequência de interação de duas ou mais sequências alvos de nucleotídeos com uma ou mais sequências de nucleotídeos de interesse, compreendendo o uso de PCR multiplex na etapa (g).

Adequadamente, o método é para analisar a frequência de interação de duas ou mais sequências alvos de nucleotídeos com uma ou mais sequências de nucleotídeos de interesse, compreendendo o agrupamento de alguns ou todos os produtos de PCR obtidos para cada uma das sequências alvos na etapa (g) e análise simultânea subsequente de suas interações de DNA.

Adequadamente, duas ou mais sequências amplificadas são marcadas diferencialmente antes do agrupamento e análise por hibridização em um arranjo.

Adequadamente, duas ou mais sequências amplificadas identicamente marcadas e analisadas por hibridização em um arranjo quando as sequências estão em cromossomos diferentes.

Adequadamente, duas ou mais sequências amplificadas são identicamente marcadas quando as sequências residem no mesmo cromossomo em uma distância que é grande o suficiente para sobreposição mínima entre sinais de interação DNA-DNA.

Adequadamente, sequenciamento de alto desempenho é usado para analisar as junções de ligação formadas entre sequências alvos e sequências capturadas de interesse.

Adequadamente, o sequenciamento é direcionado às junções de ligação formadas entre sequências alvos e sequências capturadas de interesse pela adição de sequências adaptadoras exigidas para o sequenciamento nas extremidades das sequências amplificadas.

Adequadamente, o sequenciamento é direcionado às junções de ligação formadas entre sequências alvos e sequências capturadas de interesse pela adição das sequências adaptadoras completas, ou parte delas, exigidas para sequenciamento como projeções 5' 10 nos oligonucleotídeos iniciadores usados para amplificar uma ou mais sequências de nucleotídeos de interesse.

Adequadamente, o sequenciamento é direcionado às junções de ligação formadas entre sequências alvos e sequências capturadas de interesse pela conjugação de uma substância de biotina ou outra fração nos oligonucleotídeos iniciadores usados para amplificar uma ou mais sequências de nucleotídeos de interesse, seguida por purificação mediada 15 por estreptavidina ou mediada de outra forma do material amplificado por PCR.

Adequadamente, o sequenciamento é direcionado às junções de ligação entre sequências alvos e sequências capturadas de interesse desenhando os oligonucleotídeos iniciadores usados para amplificar uma ou mais sequências de nucleotídeos de interesse em 20 400, 300, 200, 150, 100, 90, 80 70, 60, 50, 40, 30, 20, 10, 9, 8, 7, 6, 5, 4, 3, 2, 1 nucleotídeos do(s) sítio(s) de reconhecimento de enzima de restrição primária e/ou secundária analisados.

Adequadamente, o sequenciamento é direcionado às junções de ligação entre sequências alvos e sequências capturadas de interesse desenhando os oligonucleotídeos iniciadores usados para amplificar uma ou mais sequências de nucleotídeos de interesse, de maneira tal que elas se sobrepõem parcial ou completamente com o sítio de reconhecimento 25 de enzima de restrição primária e/ou secundária analisado.

Adequadamente, as sequências são lidas através da junção de ligação, de maneira tal que, quando as amostras de PCR multiplexadas ou agrupadas são analisadas, é obtida 30 informação da sequência suficiente (por exemplo, 12 nucleotídeos ou mais) em cada lado da junção de ligação para identificar inambiguamente cada sequência alvo e cada sequência capturada de interesse.

Adequadamente, a sequência de nucleotídeos alvo é selecionada do grupo que consiste em um rearranjo genômico, promotor, um melhorador, um silenciador, um isolante, 35 uma região de anexação de matriz, uma região controle de locus, uma unidade de transcrição, uma origem de replicação, um ponto quente de recombinação, um ponto de quebra de translocação, um centrômero, um telômero, uma região densa em gene, uma região pobre

em gene, um elemento repetitivo e um sítio de integração (viral).

Adequadamente, a sequência de nucleotídeos alvo é uma sequência de nucleotídeos que está associada ou causa uma doença, ou está localizada em até 15 Mb ou mais em um molde de DNA linear de um locus que está associado ou causa uma doença.

5 Adequadamente, a sequência de nucleotídeos alvo é selecionada do grupo que consiste em: *AMLI*, *MLL*, *MYC*, *BCL*, *BCR*, *ABLI*, *IGH*, *LYLI*, *TAL1*, *TAL2*, *LM02*, *TCR $\alpha$ / $\delta$* , *TCR $\beta$*  e *HOX* ou outros loci associados à doença descritos em "Catalogue of Balanced C-hromossome Aberrations in Man" 2nd edition. Albert Schinzel. Berlin: Walter de Gruyter, 2001. ISBN 3-11-011607-3.

10 Adequadamente, as sequências alvos são distribuídas ao longo do molde de genoma linear, de maneira tal que as sequências de interação cubram todo um cromossomo ou o genoma.

Adequadamente, a enzima de restrição primária é uma enzima de restrição que reconhece um sítio de reconhecimento de 6-8 bp.

15 Adequadamente, a enzima de restrição primária é selecionada do grupo que consiste em *BgRI*, *HindIII*, *BamHI*, *BamHI*, *Spel*, *PstI* e *Ndel*.

Adequadamente, a enzima de restrição primária é selecionada com base na sua ausência ou em representação nas sequências repetitivas.

20 Adequadamente, a enzima de restrição secundária é uma enzima de restrição que reconhece um sítio de reconhecimento de sequência de nucleotídeos de 4 ou 5 bp.

Adequadamente, o sítio de reconhecimento de enzima de restrição secundária está localizado a mais que cerca de 350 bp do sítio de restrição primária na sequência de nucleotídeos alvo.

25 Adequadamente, uma transição de baixas para altas frequências de interação é indicativa do local de um rearranjo genético equilibrado ou desequilibrado.

Adequadamente, um padrão inverso de frequências de interação DNA-DNA para a amostra em questão comparada ao controle é indicativa de uma inversão equilibrada e/ou desequilibrada.

30 Adequadamente, uma redução na frequência de interação DNA-DNA para a amostra em questão comparada ao controle, em combinação com um aumento na frequência de interação DNA-DNA para regiões mais distantes, é indicativa de uma deleção equilibrada e/ou desequilibrada.

Adequadamente, um aumento ou uma diminuição na frequência de interação DNA-DNA para a amostra em questão comparada ao controle é indicativa de uma duplicação ou 35 inserção equilibrada e/ou desequilibrada.

Adequadamente, cariotipagem espectral e/ou FISH são usadas antes de realizar o dito método.

Adequadamente, a doença é uma doença genética.

Adequadamente, a doença é câncer.

Adequadamente, sequências de nucleotídeos que interagem com duas ou mais sequências alvos são amplificadas.

5 Adequadamente, as sequências alvos estão posicionadas nos loci genômicos conhecidos por ser associados a uma condição doente, ou próximos a eles.

Adequadamente, as sequências alvos são selecionadas sem conhecimento prévio do local de um rearranjo e são espaçadas de maneira tal que as sequências de interação cubram todo um cromossomo ou o genoma, e em que as sequências de interação identificadas permitem a reconstrução de mapas cromossômicos lineares e rearranjos genômicos que ocorreram nos cromossomos e entre os mesmos.

Adequadamente, as sequências amplificadas são marcadas.

Adequadamente, as sequências amplificadas são marcadas diferencialmente de acordo com sua posição no genoma.

15 Adequadamente, o método é para a detecção de um rearranjo, translocação, inversão, deleção, duplicação ou inserção equilibrados e/ou desequilibrados.

Adequadamente, a etapa de hibridização do arranjo é substituída com uma etapa de sequenciamento.

20 Adequadamente, tanto a sequência de nucleotídeos alvo quanto a sequência de nucleotídeos de interesse são identificadas por sequenciamento.

Adequadamente, sequências adaptadoras estão ligadas aos produtos de PCR.

Adequadamente, sequências que interagem com duas ou mais sequências alvos são cada qual amplificadas em reações de PCR separadas.

25 Adequadamente, sequências que interagem com duas ou mais sequências alvos são cada qual amplificadas em reações de PCR separadas e subsequentemente agrupadas para análise simultânea.

Adequadamente, sequências que interagem com duas ou mais sequências alvos são amplificadas por PCR multiplex.

#### Vantagens

30 A presente invenção tem inúmeras vantagens. Estas vantagens ficarão aparentes na seguinte descrição.

A título de exemplo, a tecnologia 4C pode ser multiplexada, de maneira tal que interações com duas ou mais sequências alvos possam ser analisadas em um experimento único, por exemplo, em um arranjo único.

35 A título de exemplo adicional, a tecnologia 4C multiplexada pode ser usada para tratar rearranjos no DNA genômico em todo o genoma, em posições desconhecidas.

A título de exemplo adicional, o sequenciamento de alto desempenho pode ser u-

sado no lugar de microarranjos para analisar os fragmentos de DNA capturados. As melhorias na multiplexação e sequenciamento podem ainda ser combinadas.

A título de exemplo adicional, em vez de multiplexação, as sequências de interesse capturadas por sequências alvos diferentes podem ser amplificadas para cada sequência alvo separadamente e, a seguir, agrupadas subsequentemente para serem analisadas simultaneamente em um microarranjo.

A título de exemplo adicional, em vez de multiplexação, sequências de interesse capturadas por sequências alvos diferentes podem ser amplificadas separadamente para cada sequência alvo e agrupadas subsequentemente para serem analisadas simultaneamente por sequenciamento de alto desempenho.

A título de exemplo, a presente invenção é vantajosa, uma vez que ela fornece *inter alia* sequências de nucleotídeos, processos, sondas e arranjos comercialmente usados.

A título de exemplo adicional, a presente invenção é vantajosa, uma vez que ela permite a análise de alto desempenho da frequência de interação de duas ou mais sequências de nucleotídeos no espaço nuclear.

A título de exemplo adicional, a presente invenção é vantajosa, uma vez que, usando a tecnologia 3C convencional, cada interação DNA-DNA simples tem que ser analisada por uma reação de PCR exclusiva contendo um par exclusivo de iniciadores. Portanto, a análise de alto desempenho é possível apenas se a PCR for automatizada, mas os custos de muitos iniciadores serão muito altos. Dessa maneira, análise de alto desempenho (genômica) de interações DNA-DNA não é viável com a tecnologia 3C convencional. Ao contrário, a presente invenção permite atualmente a triagem simultânea de milhares de interações DNA-DNA. A análise de alto desempenho de interações DNA-DNA, de acordo com a presente invenção aumentarão bastante a escala e resolução de análise.

A título de exemplo adicional, a presente invenção é vantajosa, uma vez que, usando a tecnologia 3C convencional, a triagem é tendenciosa para aquelas sequências de DNA para as quais os oligonucleotídeos iniciadores foram desenhados, ordenados e incluídos na análise.

A escolha de tais oligonucleotídeos iniciadores baseia-se tipicamente no conhecimento acerca da posição, por exemplo, de melhoradores (distante) e/ou outros elementos regulatórios/sítios hipersensíveis que acredita-se reticularão com a sequência de nucleotídeos que está sendo investigada. Assim, 3C convencional é tendenciosa para o desenho de iniciadores de PCR que estão incluídos na etapa de amplificação por PCR, considerando que 4C é não tendenciosa e pode ser usada para pesquisar o genoma completo elementos de DNA interagentes. Isto se dá em virtude de a amplificação de sequências reticuladas em 4C não basearem-se no conhecimento predito de sequências que reticulam com a sequência de nucleotídeos que está sendo investigada. Preferivelmente, em uma modalidade de

4C, sequências que reticulam na primeira sequência de nucleotídeos (alvo) podem ser amplificadas usando iniciadores de PCR que hibridizam nessa sequência de nucleotídeos. Assim, a presente invenção permite uma triagem genômica não tendenciosa para interações DNA-DNA.

5 A título de exemplo adicional, a presente invenção é vantajosa em decorrência de que usar a tecnologia 3C convencional permite apenas a amplificação seletiva de uma interação DNA-DNA única. Isto não é informativo durante a hibridização em um arranjo. A tecnologia foi melhorada de maneira tal que todos os fragmentos que interagem com uma primeira sequência de nucleotídeos (alvo) são amplificados atualmente, por exemplo, amplificados seletivamente.

10 A título de exemplo adicional, a presente invenção é vantajosa em decorrência de a tecnologia 4C poder ser usada para detectar aberrações genéticas equilibradas ou desequilibradas, tais como todos os tipos de translocações, deleções, inversões, duplicações e outros rearranjos genômicos no ácido nucléico, por exemplo, cromossomos. A tecnologia 4C 15 (que mede a proximidade de fragmentos de DNA) pode ainda determinar uma predisposição do sujeito em adquirir certas translocações, deleções, inversões, duplicações e outros rearranjos genômicos (por exemplo, translocações, deleções, inversões, duplicações e outros rearranjos genômicos equilibrados ou desequilibrados). Uma vantagem com relação às estratégias atuais é que não exige-se conhecer a posição exata da mudança em decorrência 20 de a resolução da tecnologia 4C ser de maneira tal que ela possa ser usada para detectar rearranjos mesmo quando a "isca 4C" (definida pelos sítios de reconhecimento da enzima de restrição primária e secundária que são analisados) é localizada longe (por exemplo, até uma megabase ou ainda mais) da mudança. Uma outra vantagem com relação as estratégias atuais é que elas permitem uma pesquisa de genômica não tendenciosa simultânea 25 tanto para rearranjos genômicos equilibrados quanto desequilibrados. Uma outra vantagem é que a tecnologia 4C permite o mapeamento exato de mudanças, uma vez que ela pode ser usada para definir os dois sítios de restrição (primários) entre os quais as mudanças ocorreram. Uma outra vantagem é que células necessitam não ser cultivadas antes da fixação. Assim, por exemplo, tumores sólidos também podem ser analisados por rearranjos genômicos.

30 A título de exemplo adicional, a presente invenção é vantajosa em decorrência de a tecnologia 4C também poder detectar mudanças (por exemplo, rearranjos) em uma condição pré-maligna, isto é, antes de todas as células conterem estas mudanças. Assim, a tecnologia pode ser usada não apenas no diagnóstico da doença, mas também no prognóstico 35 da doença.

A título de exemplo adicional, o desenho do arranjo de acordo com a presente invenção é particularmente vantajoso comparado aos arranjos de ladrilhamento existentes,

tais como arranjos genômicos de ladrilhamento de Nimblegen, uma vez que o desenho permite a representação de uma parte muito maior do genoma por arranjo único.

A título de exemplo, para uma enzima de restrição que reconhece uma sequência de hexa-nucleotídeos, cerca de 3 arranjos com cerca de 385.000 sondas serão cada qual suficientes para cobrir, por exemplo, o genoma humano ou de camundongo completo. Para uma enzima de restrição que reconhece mais que 6 bp, um arranjo único de cerca de 385.000 sondas pode ser usado para cobrir, por exemplo, o genoma humano ou de camundongo completo. As vantagens do desenho do arranjo são que: (1) cada sonda é informativa, uma vez que cada uma analisa um evento de ligação independente, facilitando bastante a interpretação dos resultados; e (2) uma grande representação do genoma pode ser apontada em um arranjo único que é barato.

A tecnologia 4C pode ser usada vantajosamente para o mapeamento fino de rearranjos fracamente caracterizados detectados originalmente por abordagens citogenéticas (microscopia ótica, FISH, SKY, etc.).

A tecnologia 4C pode ser usada vantajosamente para a triagem simultânea, em um arranjo único, de combinações de rearranjos que ocorreram próximos aos múltiplos loci.

#### Descrição resumida das figuras

##### Figura 1

##### O princípio da tecnologia 3C

##### Figura 2

(a) O princípio de uma modalidade da tecnologia 4C. A análise 3C é realizada de maneira usual com, por exemplo, HindIII (H) como enzima de restrição. Após inversão das reticulações, a mistura de DNA conterá uma primeira sequência de nucleotídeos (alvos) ligada a muitos fragmentos diferentes. Estes fragmentos serão amplificados e marcados usando métodos de amplificação, tal como PCR inversa, por exemplo, em círculos de *DpnII*, usando primeiro iniciadores específicos da sequência de nucleotídeos (alvos). Produtos de amplificação marcados podem ser hibridizados nos arranjos da maneira aqui descrita. HindIII e *DpnII* são fornecidas como exemplos, mas outras combinações de enzimas de restrição, tais como cortadores 6 ou 8, e 4 ou 5, também podem ser usadas, (b) resultados de PCR separados por eletroforese em gel de duas amostras independentes de fígado (L1, L2) e cérebro fetais (B1, B2), (c) Representação esquemática do local das sondas de microarranjo. As sondas foram desenhadas em 100 bp de sítios HindIII. Assim, cada sonda analisa um possível parceiro de ligação.

##### Figura 3

A tecnologia 4C detecta o ambiente genômico de Rad23A (cromossomo 8). Razões não processadas são mostradas (sinais de 4C para Rad23A dividido pelo sinal obtido para amostra controle) por sondas localizadas em regiões genômicas ~15 Mb ou mais no cro-

5 cromossomo de camundongo 10, 11, 12, 14, 15, 7 e 8 (de cima para baixo; as regiões mostradas estão na mesma distância de cada centrômero correspondente). Note o grande agrupamento de fortes sinais ao redor da isca (Rad23A) no cromossomo 8 (linha 7), que demonstra que a tecnologia 4C detecta fragmentos genômicos próximos ao molde do cromossomo linear (de acordo com o fato de que frequências de interação são inversamente proporcionais à separação do sítio genômico). Note que a região ligada em *cis* ao redor da isca que mostra altas intensidades de sinal é grande (>5Mb), sugerindo, por exemplo, que translocações podem ser detectadas mesmo com iscas maiores que 1MB distante do ponto de quebra.

10 **Figura 4**

15 As interações 4C de  $\beta$ -globina no cromossomo 7 (~135 Mb) para um tecido de transcrição (fígado fetal) e um tecido de não transcrição (cérebro fetal) (analisados por uma abordagem de média corrida). Note que interações de longo alcance com  $\beta$ -globina diferem entre tecidos (dependendo provavelmente do estado de transcrição do gene). Independendo do tecido, sinais 4C fortes demarcam uma grande região (>5 Mb) ao redor da isca.

20 **Figura 5**

25 Uros e Eraf interagem com  $\beta$ -globina em células hepáticas fetais. A abordagem 4C revela que dois genes, Eraf e Uros, interagem mais que >30 Mb com o locus de  $\beta$ -globina localizado ~30 Mb distante. Estas duas interações foram previamente percebidas por uma tecnologia (Hibridização *in situ* por fluorescência) descrita em Osborne et al., *Nature Genetics* 36, 1065 (2004). Este exemplo mostra que interações de longo alcance detectadas por tecnologia 4C podem ser verificadas por FISH e refletem verdadeiramente proximidade nuclear.

30 **Figura 6**

35 A tecnologia 4C identifica exatamente transições entre regiões genômicas não relacionadas que estão ligadas em *cis*. Foram usados para estes experimentos camundongos transgênicos que contêm um cassete de região controle de locus (LCR) de  $\beta$ -globina humana (-20 kb) inserido (por meio de recombinação homóloga) no locus Rad23A no cromossomo de camundongo 8. A tecnologia 4C foi realizada em fígados fetais E 14.5 de camundongos transgênicos que foram homozigotos para esta inserção. Um fragmento *Hind*III no cassete de integração (HS2) foi usado como "isca 4C". Os dados mostram que a tecnologia 4C define exatamente ambas as extremidades do cassete transgênico (linha inferior: apenas sondas no LCR humano (~20kb) fornecem sinais de 4C e não as sondas no restante de ~380 kb da sequência de  $\beta$ -globina humana) e revelam claramente a posição de integração no cromossomo de camundongo 8 (paineel superior: sinais de comparação no cromossomo 8 (para posição de integração, ver seta) com sinais nos 6 outros cromossomo de camundongos) (cromossomos completos são representados). Este exemplo mostra que a tecnologia

4C pode ser usada para detectar a posição genômica de fragmentos de DNA ectopicamente integrados (vírus, transgene, etc.). Ela mostra que transições entre regiões genômicas não relacionadas que estão ligadas em *cis* podem ser identificadas exatamente, as quais podem ser usadas para identificar pontos de quebra genômicos e parceiros de translocação.

5 Figura 7

A tecnologia 4C produz dados reproduutíveis uma vez que os perfis para HS2 e  $\beta$ -globina são muito similares. Quatro experimentos de 4C biologicamente independentes foram realizados em fígados fetais E14.5, usando tanto o gene  $\beta$ -globina  $\beta$ -principal (2 linhas superiores) ou  $\beta$ -globina HS2 (duas linhas inferiores) como o isca. Estas iscas são separadas ~40 kb no molde do cromossomo linear, mas mostrou-se previamente que estavam próximas no espaço nuclear (Tolhuis et al, Molecular Cell 10, 1453 (2002)). É representada uma região de ~5 Mb no cromossomo de camundongo 7 que está 20-20 Mb longe do locus de  $\beta$ -globina. Os dados mostram alta reproduutibilidade entre experimentos independentes e demonstram que dois fragmentos próximos no espaço nuclear compartilham parceiros integrantes localizados em outra parte no genoma.

Figura 8

A tecnologia 4C é aplicada para medir frequências de interação DNA-DNA com sequência X (no cromossomo A) em células de uma pessoa saudável (topo) e um paciente com translocação (A;B) (fundo). Intensidades de sinal que representam frequências de interação DNA-DNA (eixo Y) são representadas graficamente por sondas ordenadas nos moldes do cromossomo linear (eixo X). Em células normais, interações DNA-DNA frequentes são detectadas no cromossomo A ao redor da sequência X. Em células de pacientes, uma redução em 50 % nas frequências de interação é observada por sondas no cromossomo A localizadas no outro lado do ponto de quebra (BP) (comparar curva cinza (paciente) com linha preta (pessoa saudável). Além disso, a translocação coloca parte do cromossomo B em proximidade física imediata com sequência X, e interações DNA-DNA frequentes são observadas atualmente para esta região no cromossomo B. A transição abrupta de baixas para altas frequências de interação neste cromossomo marca o local de seu ponto de quebra.

30 Figura 9

(Equilibrado) inversão(s) pode(m) ser detectada(s) por tecnologia 4C. Padrões inversos de frequências de interação DNA-DNA (medidos por tecnologia 4C como intensidades de sinal de hibridização) são observados em sujeito doente (curva sólida) comparados ao sujeito não doente (curva pontilhada), o que revela a presença e tamanho da inversão.

35 Figura 10

Detecção de deleção(s) heterozigota(s) por tecnologia 4C. Sondas com menores frequências de interação DNA-DNA (medidas por tecnologia 4C como intensidades de sinal

de hibridização) em sujeitos doentes (curva cinza) comparadas a sujeitos não doentes (curva negra), revelam a posição e tamanho da região deletada. Sinais de hibridização residuais na região deletada do sujeito doente vêm do alelo intacto (deleção heterozigota). A deleção é tipicamente acompanhada por um aumento nas intensidades de sinal para sondas localizadas diretamente além da região deletada (note que a curva cinza está acima da curva negra no lado direito da deleção), uma vez que estas regiões estão em proximidade física mais imediata às sequência 4C (isca).

Figura 11

Duplicação detectada por tecnologia 4C. Sondas com maiores sinais de hibridização em um paciente (curva sólida) comparados a um sujeito normal (curva pontilhada) indicam a posição e tamanho da duplicação. A duplicação detectada por tecnologia 4C é tipicamente acompanhada por menores sinais de hibridização em sujeitos doentes versus não doentes para sondas além da região duplicada (a duplicação aumenta sua separação do sítio genômico da sequência 4C).

Figura 12

Interações de longo alcance com  $\beta$ -globina reveladas por tecnologia 4C. a, Razões não processadas de 4C com relação a sinais de hibridização controle revelando interações de  $\beta$ -globina HS2 com cromossomo 7 e dois cromossomos não relacionados (8 e 14). b-c, Dados não processados para duas amostras independente de fígado fetal (topo, em vermelho) e cérebro fetal (fundo, em azul) representados graficamente ao longo de duas regiões diferentes 1-2 Mb no cromossomo 7. Agrupamentos altamente reproduzíveis de interações são observados tanto nas duas amostras de fígado fetal (b) quanto nas duas amostras de cérebro (c). d-e, Dados de média corrida para as mesmas regiões. A taxa de descoberta falsa foi ajustada em 5% (linha pontilhada), f, Representação esquemática de regiões de interação com  $\beta$ -globina ativa (fígado fetal, topo) e inativa (cérebro fetal, fundo) no cromossomo 7.

Figura 13

$\beta$ -globina ativa e inativa interagem com regiões cromossômicas ativas e inativas, respectivamente, a, Comparação entre interações de longo alcance de  $\beta$ -globina no fígado fetal (média corrida 4C, topo), análise de expressão de microarranjo no fígado fetal (escala log, meio) e o local dos genes (fundo) representado graficamente ao longo uma região 4 Mb que contém o gene *Uros* (~30 Mb longe da  $\beta$ -globina), mostrando que a  $\beta$ -globina ativa interage preferivelmente com outros genes ativamente transcritos, b, A mesma comparação no cérebro fetal ao redor um agrupamento de gene OR ~ 38 Mb longe da globina, mostrando que a  $\beta$ -globina inativa interage preferivelmente com regiões inativas, c, Caracterização de regiões que interagem com  $\beta$ -globina no fígado fetal (esquerda) e no cérebro (direita) em termos de conteúdo e atividade do gene.

## Figura 14

Rad23A ubiqüamente expresso interage com regiões ativas muito similares no fígado fetal e cérebro, a, Representação esquemática de regiões no cromossomo 8 que interagem com Rad23A ativo no fígado fetal (topo, vermelho) e cérebro (fundo, azul), b, Comparação entre interações de longo alcance de Rad23A (média corrida 4C) e análise de expressão de microarranjo (escala log) no fígado fetal (dois painéis superiores), interações de longo alcance de Rad23A (média corrida 4C) e análise de expressão de microarranjo (escala log) no cérebro fetal (paineis 3 e 4) e o local dos genes (paineil inferior) representado graficamente ao longo de uma região 3 Mb do cromossomo 8. c, Caracterização de regiões que interagem com Rad23A no fígado fetal (esquerda) e cérebro (direita) em termos de conteúdo e atividade do gene.

## Figura 15

Cryo-FISH confirma que a tecnologia 4C identifica verdadeiramente regiões interagentes, a, exemplo de parte de uma seção de cryo (200 nm) mostrando mais de 10 núcleos, alguns dos quais contendo o locus de  $\beta$ -globina (verde) e/ou Uros (vermelho). Em virtude do seccionamento, muitos núcleos não contêm sinais para estes dois loci, b-d, exemplos de sinais sobrepostos completamente (b) e parcialmente (c) e sinais de contato (d), que foram todos pontuados como positivos para interação, e-g, exemplos de núcleos contendo alelos não contactantes (e-f) e um núcleo contendo apenas  $\beta$ -globina (g), que foram todos pontuados como negativos para interação, h-i, Representação esquemática de resultados de cryo-FISH. Porcentagens de interação com  $\beta$ -globina (h) e Rad23A (i) são indicadas acima dos cromossomos para regiões positivamente identificadas (ponta da seta vermelha) e negativamente identificadas (ponta da seta azul) por tecnologia 4C. Os mesmos BACs foram usados para os dois tecidos. Frequências de interação medidas por cryo-FISH entre dois agrupamentos de genes OR distantes no fígado fetal e cérebro são indicadas abaixo dos cromossomos.

## Figura 16

A análise 4C de HS2 e  $\beta$ -principal fornece resultados altamente similares. (a) Dados de 4C não processados de quatro amostras de fígado E14.5 independentes mostram um padrão muito similar de interação com HS2 (topo) e  $\beta$ -principal (fundo), (b) Uma grande sobreposição existe entre sondas pontuadas positivas para interação no experimento HS-2 e sondas que foram pontuadas positivas para interação no experimento  $\beta$ -principal.

## Figura 17

Regiões que interagem com  $\beta$ -globina também colocam uma em contato com a outra frequentemente. Duas regiões (separadas praticamente 60 Mb), contendo genes ativamente transcritos e identificadas por tecnologia 4C interagem com  $\beta$ -globina no fígado fetal, mostraram frequências de co-localização por cryo-FISH de 5,5%, que foram significativa-

mente maiores que frequências de co-localização de fundo.

Figura 18

Exemplo de uma deleção heterozigota revelada por 4C multiplex usando um corante simples para a marcação de fragmentos de DNA que interagem com múltiplas sequências alvos. A razão de frequências de interação observada em um paciente (amostra) com relação a pessoa saudável (controle) é representada à direita.

Figura 19

A presença de uma deleção presente em um paciente com leucemia revelada por 4C usando uma sequência de nucleotídeos alvo que está tanto a 2 Mb (a) quanto a 1,3 Mb (B) à montante (“à esquerda”) do primeiro ponto de quebra. Note que deleções causam uma redução de sinais de interação de DNA na região deletada, mas também causam um aumento nas frequências de interação DNA:DNA para sequências diretamente à jusante (“à direita”) do último ponto de quebra. Isto é particularmente óbvio quando interações com sequência de nucleotídeos alvo B são examinadas cautelosamente (ver dois gráficos inferiores). Com base nos dados de 4C, iniciadores foram desenhados em cada lado da região deletada e o ponto de quebra foi identificado por sequenciamento: texto simples é sequência à montante de deleção, em negrito é indicado um nucleotídeo inserido, sublinhada é a sequência à jusante da deleção.

Figura 20

Uma inversão heterozigota revelada por 4C multiplex, usando um corante simples para a marcação de fragmentos de DNA que interagem com múltiplas sequências alvos. A razão de frequências de interação observada em paciente (amostra) com relação a pessoa saudável (controle) é descrita à direita. Note que as razões dos pontos de quebra próximos podem ser diferentes quando a posição dos pontos de quebra com relação às sequências alvos de nucleotídeos é diferente.

Figura 21

As cores alternam entre sequências de nucleotídeos alvos vizinhos, o que permite a detecção (em vermelho) de uma deleção próxima a uma sequência de nucleotídeos alvo (em azul) que não consegue detectar a deleção em virtude dos sinais de hibridização saturados. Caso a quantidade de sonda no arranjo não esteja saturando, o sinal azul também seria diminuído na deleção.

Figura 22

Cores alternam entre sequências de nucleotídeos alvos vizinhos, o que permite a detecção (em vermelho) de uma inversão. A razão de frequências de interação observada no paciente (amostra) com relação a pessoa saudável (controle) é descrita à direita. Note que comparado a um experimento de corante simples (ver figura 2), o uso de corantes alternantes facilitou a detecção de rearranjos, tais como inversões. Note também que razões

próximas aos pontos de quebra podem ser diferentes quando a posição dos pontos de quebra com relação às sequências de nucleotídeos alvos é diferente.

Figura 23

Exemplo de uma inversão heterozigota revelada por 4C multiplex usando corantes diferentes para a marcação de fragmentos de DNA que interagem com sequências alvos diferentes. Os pontos de quebra da inversão são indicados pela posição de sinais vermelho e verde no paciente que estão ausentes na amostra controle. Note que a introdução de mais cores facilita a detecção de rearranjos (compare, por exemplo, figura 4 e 5).

Figura 24

Detecção de translocações equilibradas. Cada cromossomo é marcado com dois corantes exclusivos que são usados de maneira tal que os corantes alternem entre as sequências alvos que são vizinhas no molde do cromossomo linear. Se as translocações forem equilibradas, cada um dos dois corantes específicos de cromossomo daria sinais fortes de hibridização em um conjunto mutuamente exclusivo de sondas diretamente vizinhas umas com as outras no molde linear do cromossomo não relacionado. O ponto de quebra neste cromossomo não relacionado está localizado entre os dois ajustes de sondas mostrando sinais positivos de hibridização. Os sinais de cromossomo pai no ponto de quebra e além dele teriam a metade da intensidade do controle (não mostrado na figura)

Figura 25

Demonstração do princípio para a detecção de translocações equilibradas. Detecção de translocação t(1;7) descrita em (R. Burnett et al., Blood, Vol 84, No 4 (August 15), 1994: pp 1232-1236). Sequências de nucleotídeos alvos flanqueiam o locus TCR $\beta$  no cromossomo 7, com os sinais vermelhos que representam interações DNA:DNA com a sequência alvo que está localizada à montante do locus TCR $\beta$ , e os sinais azuis que representam interações DNA:DNA com a sequência alvo que está localizada à jusante do locus TCR $\beta$ . São representados os sinais de DNA interagentes encontrados no cromossomo 1. O painel topo mostra a distribuição de sinal teórico. O painel médio e inferior mostra a distribuição de sinal atual. O painel inferior mostra sinais em uma resolução de sondas individuais justapostas no cromossomo molde. Note que, no caso de translocação equilibrada, sequências de nucleotídeos alvos que flanqueiam o ponto de quebra mostrariam um conjunto mutuamente exclusivo de sinais de interação de DNA intercromossômico que faz limite diretamente um com o outro no molde do cromossomo linear do cromossomo parceiro de translocação. A posição do ponto de quebra sequenciado (descrito em Burnett et al., 1994) é indicada por uma seta no painel inferior.

Figura 26

Exemplo teórico para a detecção de translocações desequilibradas. Cada cromossomo é marcado com dois corantes exclusivos que são usados de maneira tal que os coran-

tes alternem entre sequências alvos são vizinhas no molde do cromossomo linear. Se as translocações ocorreram com perda de DNA nos pontos de quebra (isto é, translocações desequilibradas), cada um dos dois corantes específicos de cromossomo fornecerá sinais fortes de hibridização em um conjunto mutuamente exclusivo de sondas no cromossomo 5 não relacionado que não são diretamente vizinhos um do outro no molde linear do cromossomo não relacionado. A região deletada é indicada.

Figura 27

Detecção de translocações desequilibradas. A detecção de translocação t(4;7) descrita em (RJ Galjaard et al., Am J Med Genet A. 2003 Aug 30; 121 (2): 168-73). Sequências de nucleotídeos alvos localizam-se no cromossomo 7; os sinais de DNA interagentes representados são localizados no cromossomo 4. Duas sequências alvos foram usadas, localizadas à montante (5') e à jusante (3') do ponto de quebra no cromossomo 7. Sinais de DNA interagentes localizados no cromossomo 4 são indicados (para ambas as sequências alvos em azul). A região entre os agrupamentos de fragmentos de DNA interagentes no cromossomo 4 foi deletada neste paciente. Topo: sinais para o cromossomo completo 4. Dados de 4C do painel inferior: sinais em uma região 11.5 MB ao redor dos pontos de quebra no cromossomo 4. Com base nos dados de 4C, o fragmento de restrição *Hind*III no cromossomo 4 contendo o ponto de quebra de translocação foi identificado e usado para mapear o ponto de quebra por sequenciamento. A sequência é fornecida na base da figura, onde a sequência sublinhada é do cromossomo 4, negrito é observado tanto no cromossomo 7 quanto 4 e a sequência comum é do cromossomo 7.

Figura 28

Marcação específica para cromossomo de interações de DNA. O sinal azul que aparece no cromossomo 3 e sinais laranja que aparecem no cromossomo 1 revelam os cromossomos parceiros de translocação e a posição aproximada dos pontos de quebra.

Figura 29

Sequenciamento de sequências de nucleotídeos amplificadas por PCR de interesse (azul: endereço de deslocamento) ligadas a sequência de nucleotídeos alvo (vermelho: endereço domiciliar). A amplificação foi realizada usando iniciadores (vermelho), pelo menos 30 um dos quais é complementar à sequência de nucleotídeos alvo. Opcionalmente, os adaptadores (verde) podem ser introduzidos de várias maneiras na extremidade dos produtos de PCR, da maneira indicada.

Figura 30

4C detecta exatamente uma translocação e inversão equilibrada (A-B). A tecnologia 35 4C detecta uma translocação t(1;7) equilibrada. (A) Em uma amostra controle saudável, fragmentos alvo a (vermelho) e b (azul) que estão localizados em lados opostos do locus TCRB no cromossomo 7 não capturam regiões no cromossomo 1. (B) Na linhagem celular

HSB- 2 contendo uma translocação t(l;7) (p35;q35) equilibrada, cada fragmento alvo de TCRB captura uma região no cromossomo 1. As regiões capturadas têm diversos megabases de tamanho (aumento 1), diretamente vizinhas uma da outra (aumento 2) e flanqueiam o ponto de quebra previamente clonado (seta). Ver figura S1 para resultados de outros cromossomos. (C-D) 4C detecta uma inversão equilibrada. (C) Em uma amostra controle saudável, fragmentos alvo a (vermelho) e b (azul) que estão localizados em lados opostos do locus TCRB no cromossomo 7 não capturam grandes regiões em outros locais no cromossomo 7. (D) Em uma amostra de paciente T-ALL, cada fragmento alvo de TCRB captura uma região adicional em outra extremidade do cromossomo 7. A maior parte do fragmento alvo 5' (a; vermelho) captura uma região 5' do agrupamento HOXA, a maior parte do fragmento alvo 3' (b; azul) captura uma região 3' do agrupamento HOXA, demonstrando uma inversão. As regiões capturadas têm diversos megabases de tamanho (aumento 1) e diretamente vizinhas uma da outra (aumento 2), mostrando que a inversão é equilibrada. Ambos os fragmentos alvo identificam o ponto de quebra (seta) próximo a HOXA9 em uma região de 6 kb. Dados de média corrida foram representados graficamente, usando um tamanho de janela de ~60kb. Aumentos mostram intensidades de sinal não processadas.

Figura 31

4C detecta exatamente rearranjos desequilibrados. (a) 4C detecta exatamente uma translocação t(4;7)(p15.2;q35) em combinação com uma microdeleção (isto é, rearranjo desequilibrado) em uma linhagem celular de uma criança por nascer com malformação congênita. Os fragmentos alvo a (vermelho) e b (azul), que estavam localizados em lados opostos dos pontos de quebra no cromossomo 7, ambos capturaram os fragmentos que cobrem diversos megabases no cromossomo 4 (para sinais em outros cromossomos, ver figura 35). As duas regiões capturadas com sinais altos não são diretamente vizinhas uma da outra, mostrando que a translocação é acompanhada por uma deleção no cromossomo 4. (B) A sequência de um dos pontos de quebra (seta), com sequências do cromossomo 7 e 4 em letras minúsculas e maiúsculas, respectivamente. (C) 4C identifica exatamente uma deleção homozigota em uma amostra de paciente T-ALL. Um fragmento alvo localizado no cromossomo 9 em 19,3 Mb identifica uma região (entre setas) que não tem sinais altos na amostra de paciente (inferior) comparados à amostra controle (superior), mostrando uma deleção de ~2 Mb em 9p21. Sinais 3' da deleção são maiores em paciente versus controle, uma vez que esta região está em proximidade mais imediata ao fragmento alvo em virtude da deleção. (D) Sequência através dos pontos de quebra indicada pelas setas em (C), confirmando a deleção. Intensidades de sinal não processadas são representadas graficamente nesta figura. Sinais altos raros em regiões deletadas indicam que estas sondas mostram hibridização não específica.

Figura 32

Triagem 4C identifica LM03 como um parceiro de translocação inédito de TCRB. Cinco amostras de paciente T-ALL não caracterizadas foram triadas por 4C, usando um fragmento alvo próximo do locus TCRB no cromossomo 7. (a) Em uma amostra de paciente, sinais altos apareceram especificamente no cromossomo 12, revelando uma translocação, 5 t(7;12)(q35;pl2.3). Para sinais em todos os outros cromossomos, ver figura S4. Uma deleção está presente em diversos megabases do sítio de translocação (seta) no cromossomo 12 (aumento 1). O sítio de translocação está presente em uma região de 6 kb próxima ao gene LMO3 (aumento 2). (B) Sequências de ambos os pontos de quebra de t(7;12)(q35;pl2.3); nucleotídeos em caixas de maiúsculas são a partir de 12, em caixas de minúsculas a partir 10 de 7 e em itálicos são a partir de origem desconhecida. (C) Representação esquemática do sítio de translocação de t(7;12)(q35;pl2.3). O melhorador de TCRB está posicionado 70 kb à jusante do gene LMO3. Dados de média corrida foram representados graficamente, usando um tamanho de janela de ~60kb. Aumentos mostram intensidades de sinal não processadas.

15 Figura 33

Sinais de 4C através de todos os cromossomos em um controle saudável e uma amostra que carrega t(1;7)(p35;q35). As pontas das flechas negras indicam a posição das sequências alvos. A ponta da seta vermelha indica a posição do sítio de translocação. Dados de média corrida foram representados graficamente, usando um tamanho de janela de 20 ~60kb. Escala no eixo Y (unidades arbitrárias) é idêntica para todos os cromossomos.

Figura 34

Sequenciamento de extremidade pareada de fragmento de restrição (a) Representação esquemática de sequenciamento de extremidade pareada de fragmento de restrição. (B) Sequências de ponto de quebra de uma inversão entre TCRB e HOXA no cromossomo 7 25 (ver figura 30). A sequência negra está localizada na posição em HOXA observada com tecnologia 4C, entre as sondas que marcam a transição dos fragmentos capturados para os não capturados. A sequência vermelha foi observada com sequenciamento de extremidade pareada de fragmento de restrição da sequência negra e está localizada no locus TCRB.

Figura 35

30 Sinais de 4C através de todos os cromossomos obtidos com dois fragmentos alvo de cromossomo 7 diferentes em uma amostra carregando t(4;7)(pl5.2;q35). As pontas das flechas negras indicam posição de sequências alvos. As pontas das flechas vermelhas indicam a posição dos sítios de translocação. Dados de média corrida foram representados graficamente, usando um tamanho de janela de ~60kb. Escala no eixo Y (unidades arbitrárias) é 35 idêntica para todos os cromossomos.

Figura 36

Sinais de 4C através de todos os cromossomos obtidos com uma sequência alvo

próxima ao locus TCRB no cromossomo 7 em duas amostras de paciente T-ALL, uma das quais carregando uma translocação t(7;12). As pontas das flechas negras indicam a posição de sequências alvos. A ponta da seta vermelha indica a posição do sítio de translocação. Dados de média corrida foram representados graficamente, usando um tamanho de janela de ~60kb. Escala no eixo Y (unidades arbitrárias) é idêntica para todos os cromossomos.

5 Figura 37

Expressão de LM03 em amostras de paciente T-ALL. A expressão de gene foi medida em arranjos de expressão de gene Affymetrix. LM03 é expresso no paciente que carrega t(7;12)(q35;pl2.3), mas não em outros pacientes.

10 Figura 38

15 Iniciadores e produtos de PCR 4C a ser analisados por sequenciamento Solexa. Sequências produzidas por Solexa (setas) primeiramente leram o "iniciador DpnII" (18 nucleotídeos, incluindo GATC (isto é, sítio de reconhecimento DpnII)), seguido pela sequência capturada.

15 Figura 39

20 Os resultados de PCR usando iniciadores com projeções 5' contendo sequências adaptadoras Solexa. Para comparação, os resultados obtidos com iniciadores padrões (raias 1,5,9 da esquerda para a direita; sem projeções) também são conhecidos.

Figura 40

25 Resultados do sequenciamento 4C.

Figura 41

Ajuste de iniciador 3 (139Mb) captura sequências de cromossomo 1 através do ponto de quebra na linhagem celular HSB-2 T-ALL (no cromossomo 7 em ~142Mb, isto é, 3Mb longe da isca). Para comparação, são mostrados os resultados de microarranjo. Note que a sequência alvo (isca) usada para o experimento de microarranjo foi mais próxima (<1 Mb) do ponto de quebra, explicando por que ela mapeia melhor o ponto de quebra no cromossomo 1.

#### Descrição detalhada da invenção

#### A tecnologia 3C

30 O método 3C foi descrito em detalhes em Dekker et al. (2002), Tolhuis et al. (2002), Palstra et al. (2003), Splinter et al. (2004) e Drissen et al. (2004). Resumidamente, 3C é realizada digerindo DNA reticulado com uma enzima de restrição primária seguida por ligação em concentrações de DNA muito baixas. Nestas condições, a ligação de fragmentos reticulados, que é intramolecular, é fortemente favorecida com relação a ligação de fragmentos aleatórios, que é intermolecular. A reticulação é então invertida e produtos de ligação individuais são detectados e quantificados pela reação em cadeia da polimerase (PCR) usando iniciadores específicos de locus. A frequência de reticulação (X) de dois loci específicos é

determinada por reações de PCR quantitativas usando moldes controle e reticulado, e X é expresso como a razão da quantidade do produto obtido com o molde reticulado e com o molde controle.

De acordo com a presente invenção, um molde de 3C é preparado usando os métodos descritos por Splinter et al, (2004) Methods Enzymol. 375, 493-507. (isto é, fixação de formaldeído, digestão de enzima de restrição (primária), religação de fragmentos de DNA reticulados e purificação de DNA). Resumidamente, uma amostra, tais como células, tecidos ou núcleos, é fixada usando um agente de reticulação, tal como formaldeído. A digestão de enzima de restrição primária é a seguir realizada de maneira tal que o DNA seja digerido no contexto do núcleo reticulado. A ligação intramolecular é a seguir realizada em baixas concentrações de DNA (por exemplo, cerca de 3,7 ng/µL), que favorece a ligação entre fragmentos de DNA reticulado (isto é, a ligação intramolecular) com relação a ligação entre fragmentos de DNA não reticulado (isto é, ligação intermolecular ou aleatória). A seguir, as reticulações são invertidas e o DNA pode ser purificado. O molde de 3C que é produzido contém fragmentos de restrição que são ligados em decorrência de que eles estavam originalmente próximos no espaço nuclear.

Uma vez que uma enzima de restrição primária é usada para digerir o DNA antes da etapa de ligação intramolecular, um sítio de reconhecimento de enzima para a enzima de restrição primária separará a primeira sequência de nucleotídeos (alvo) e a sequência de nucleotídeos que foi ligada. Dessa maneira, o sítio de reconhecimento primário está localizado entre a primeira sequência de nucleotídeos (alvo) e a sequência de nucleotídeos ligados (isto é, a segunda sequência ligada).

#### Sequência de nucleotídeos

A presente invenção envolve o uso de sequências de nucleotídeos (por exemplo, moldes 3C, moldes 4C, moldes de DNA, moldes de amplificação, fragmentos de DNA e DNA genômico), que podem estar disponíveis na base de dados.

A sequência de nucleotídeos pode ser DNA ou RNA de origem genômica, sintética ou recombinante, por exemplo, DNAc. Por exemplo, sequências de nucleotídeos recombinantes podem ser preparadas usando técnicas de clonagem de PCR. Isto envolverá preparar um par de iniciadores que flanqueiam uma região da sequência que deseja-se clonar, colocando os iniciadores em contato com RNAm ou DNAc obtidos, por exemplo, de uma célula de mamífero (por exemplo, célula animal ou humana) ou de não mamífero, realizando uma reação em cadeia da polimerase (PCR) em condições que realizam amplificação da região desejada, isolando o fragmento amplificado (por exemplo, purificando a mistura da reação em um gel de agarose) e recuperando o DNA amplificado. Os iniciadores podem ser desenhados para conter sítios de reconhecimento de enzima de restrição adequados de maneira tal que o DNA amplificado possa ser clonado em um vetor de clonagem adequado.

A sequência de nucleotídeos pode ser de fita dupla ou fita simples, quer representando a fita sentido ou antissenso ou combinações destas.

Para alguns aspectos, é preferível que a sequência de nucleotídeos seja DNA de fita simples, tais como iniciadores e sondas de fita simples.

5 Para alguns aspectos, é preferível que a sequência de nucleotídeos seja DNA de fita dupla, tais como moldes 3C e 4C de fita dupla.

Para alguns aspectos, é preferível que a sequência de nucleotídeos seja DNA genômico, tais como um ou mais loci genômicos.

10 Para alguns aspectos, é preferível que a sequência de nucleotídeos seja DNA cromossômico.

A sequência de nucleotídeos pode compreender uma primeira sequência de nucleotídeos (alvo) e/ou uma segunda sequência de nucleotídeos.

15 Os sítios de reconhecimento da enzima de restrição primária e secundária serão diferentes um do outro e ocorrerão tipicamente apenas uma vez na sequência de nucleotídeos.

20 Em um aspecto, é fornecida uma sequência de nucleotídeos circularizada compreendendo uma primeira sequência de nucleotídeos e (por exemplo, ligada) uma segunda sequência de nucleotídeos separada (por exemplo, dividida ou partida) por um sítio de reconhecimento de enzima de restrição primária e uma secundária, em que a dita primeira sequência de nucleotídeos é uma sequência de nucleotídeos alvo e a dita segunda sequência de nucleotídeos é obtêvel por DNA genômico de reticulação (por exemplo, *in vivo* ou *in vitro*). Os sítios de reconhecimento da enzima de restrição primária e secundária serão diferentes um do outro e ocorrerão tipicamente apenas uma vez na sequência de nucleotídeos.

25 Em um aspecto adicional, é fornecida uma sequência de nucleotídeos circularizada compreendendo uma primeira sequência de nucleotídeos e (por exemplo, ligada a) uma segunda sequência de nucleotídeos separada (por exemplo, dividida ou partida) por um sítio de reconhecimento de enzima de restrição primária e uma secundária, em que a dita primeira sequência de nucleotídeos é uma sequência de nucleotídeos alvo e em que as ditas primeira e segunda sequências de nucleotídeos são obtêveis por um processo compreendendo as etapas de: (a) DNA genômico de reticulação (por exemplo, *in vivo* ou *in vitro*) ; (b) digerir o DNA reticulado com uma enzima de restrição primária; (c) ligar as sequências de nucleotídeos reticulados; (d) inverter a reticulação; e (e) digerir as sequências de nucleotídeos com uma enzima de restrição secundária para circular as sequências de nucleotídeos.

35 Preferivelmente, a segunda sequência de nucleotídeos intercepta (por exemplo, bissecciona) a primeira sequência de nucleotídeos (alvo). Dessa maneira, a sequência de nucleotídeos compreende a segunda sequência de nucleotídeos, que separa a primeira sequência de nucleotídeos (alvo) em duas porções ou fragmentos, tais como aproximadamen-

te duas porções ou fragmentos igualmente dimensionados. Tipicamente, as porções ou fragmentos serão pelo menos cerca de 16 nucleotídeos de comprimento.

Em um aspecto adicional, é fornecida uma base de dados de sequências de 6-50 pares de bases que flanqueiam diretamente e, opcionalmente, incluem o sítio de reconhecimento de enzima de restrição primária de cada sequência alvo incluída, e que podem ser usadas nos métodos aqui descritos para identificar cada sequência alvo.

Em um outro aspecto, é fornecida uma base de dados de sequências de 12-50 pares de bases que flanqueiam diretamente todos os sítios de reconhecimento de enzima de restrição primária no genoma e que podem ser usadas nos métodos aqui descritos para determinar a posição genômica de cada uma das sequências capturadas identificadas.

Em um outro aspecto, é fornecida uma base de dados de sequências de 6-50 pares de bases que flanqueiam diretamente e, opcionalmente, incluem o sítio de reconhecimento de enzima de restrição secundária relevante de cada uma das sequências alvos incluídas, e que podem ser usadas nos métodos aqui descritos para identificar cada sequência alvo.

Em um outro aspecto, é fornecida uma base de dados de sequências de 12-50 pares de bases que flanqueiam diretamente todos os sítios de reconhecimento de enzima de restrição secundária relevantes no genoma e que podem ser usadas nos métodos aqui descritos para determinar a posição genômica de cada uma das sequências capturadas identificadas.

#### Primeira sequência de nucleotídeos

A primeira sequência de nucleotídeos é uma sequência de nucleotídeos alvo.

Da maneira aqui usada, o termo "sequência de nucleotídeos alvo" refere-se à sequência que é usada como uma sequência isca afim de identificar uma ou mais sequências nas quais ela reticula (por exemplo, uma ou mais sequências de nucleotídeos de interesse ou uma ou mais sequências de composição de sequência de nucleotídeos desconhecida).

A sequência de nucleotídeos alvo é de sequência conhecida.

A reticulação é indicativa de que a sequência de nucleotídeos alvo e a sequência reticulada nela foram originalmente próximas no espaço nuclear. Determinando a frequência cuja sequências estão próximas uma da outra, é possível entender, por exemplo, a conformação de cromossomos e regiões cromossômicas no contexto espacial do núcleo (por exemplo, *in vivo* ou *in vitro*). Além disso, é possível entender as organizações estruturais intrincadas no genoma, por exemplo, quando melhoradores ou outros elementos regulatórios transcripcionais comunicam com promotores distantes localizados em *cis* ou ainda em *trans*. Além disso, é ainda possível entender o posicionamento de uma dada região genômica com relação às sequências de nucleotídeos presentes no mesmo cromossomo (em *cis*), bem como às sequências de nucleotídeos em outros cromossomos (em *trans*). Assim, é possível mapear sequências de nucleotídeos nos cromossomos diferentes que compartilham fre-

quentemente sítios no espaço nuclear. Além disso, ainda é possível detectar aberrações genéticas equilibradas e/ou desequilibradas, tais como translocações, deleções, inversões, duplicações e outros rearranjos genômicos equilibrados e/ou desequilibrados (por exemplo, deleções ou translocações em um ou mais cromossomos). Com relação a isso, aberrações genéticas resultam em mudanças nas interações DNA-DNA na posição que a mudança ocorreu, que podem ser detectadas.

A primeira sequência de nucleotídeos (alvo) de acordo com a presente invenção pode ser qualquer sequência na qual deseja-se determinar a frequência de interação no espaço nuclear com uma ou mais outras sequências.

Em uma modalidade, a primeira sequência de nucleotídeos (alvo) será maior que cerca de 350 bp de comprimento, uma vez que é escolhida uma enzima de restrição secundária que corta a primeira sequência de nucleotídeos (alvo) em cerca de 350 bp ou mais do sítio de restrição primária. Isto pode minimizar uma tendência na formação de círculo em virtude das restrições topológicas (Rippe et al. (2001) Trends in Biochem. Sciences 26, 733-40).

Adequadamente, a primeira sequência de nucleotídeos (alvo) após a amplificação compreende pelo menos cerca de 32 bp, em virtude do fato de que o comprimento mínimo de pelo menos dois iniciadores de amplificação usados para amplificar a segunda sequência de nucleotídeos é cerca de 16 bases cada.

Em uma modalidade preferida, a primeira sequência de nucleotídeos (alvo) pode compreender completa ou parcialmente (por exemplo, um fragmento), ou estar perto (por exemplo, nas proximidades), de um promotor, um melhorador, um silenciador, um isolante, uma região de anexação de matriz, uma região controle de locus, uma unidade de transcrição, uma origem de replicação, um ponto quente de recombinação, um ponto de quebra de translocação, um centrômero, um telômero, uma região densa em gene, uma região pobre em gene, um elemento repetitivo, um sítio de integração (viral), uma sequência de nucleotídeos na qual deleções e/ou mutações estão relacionadas a um efeito (por exemplo, doença, efeito fisiológico, funcional ou estrutural, tal como um SNP (polimorfismo de nucleotídeo simples), ou sequência(s) de nucleotídeos(s) contendo tais deleções e/ou mutações, ou qualquer sequência na qual deseja-se determinar a frequência de interação no espaço nuclear com outras sequências.

Da maneira mencionada anteriormente, a primeira sequência de nucleotídeos (alvo) pode compreender completa ou parcialmente (por exemplo, um fragmento), ou estar perto (por exemplo, nas proximidades) de uma sequência de nucleotídeos na qual aberrações genéticas, tais como deleções e/ou mutações, estão relacionadas a um efeito (por exemplo, uma doença). De acordo com esta modalidade da invenção, a primeira (sequência de nucleotídeos alvo) pode ser, portanto, uma sequência de nucleotídeos (por exemplo, um gene

ou um locus), adjacente (no molde de DNA físico), ou na região genômica, na qual mudanças foram associadas ou correlacionadas a uma doença, tal como uma doença genética ou congênita. Em outras palavras, a primeira sequência de nucleotídeos (alvo) pode ser ou pode ser escolhida com base em sua associação com um fenótipo clínico. Em uma modalidade preferida, as mudanças são mudanças em um ou mais cromossomos, e a doença pode ser uma consequência, por exemplo, de uma ou mais deleções, uma ou mais translocações, uma ou mais duplicações, e/ou uma ou mais inversões, etc. nele. Exemplos não limitantes de tais genes/loci são AML1, MLL, MYC, BCL, BCR, ABL1, loci de imunoglobulina, LY11, TAL1, TAL2, LM02, TCR $\alpha$ / $\delta$ , TCR $\beta$ , HOX e outros loci em várias leucemias linfoblásticas.

Outros exemplos são descritos na base de dados eletrônica, tais como:

<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=cancerchromosomes>

<http://cgap.nci.nih.gov/Chromosomes/Mitelman>

<http://www.progenetix.net/progenetix/P14603437/ideogram.html>

<http://www.changbioscience.com/cytogenetics/cyto1.pl?query=47.xy>

<http://www.possible.net.au/>

<http://www.1mdatabases.com/>

<http://www.wiley.com/legacy/produtos/sujeito/life/borgaonkar/index.html>

<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM>

<http://www.sanger.ac.uk/PostGenomics/decipher/>

<http://agserver01.azn.nl:8080/ecaruca/ecaruca.jsp>

Outros exemplos são descritos em "Catalogue of Unbalanced Chromossome Aberrations in Man" 2nd edition. Albert Schinzel. Berlin: Walter de Gruyter, 2001. ISBN 3-11-011607-3.

Em uma modalidade, a palavra "adjacente" significa "diretamente adjacente", de maneira tal que não existam nucleotídeos dispostos entre duas sequências adjacentes.

Em uma outra modalidade, o termo "adjacente", no contexto da sequência de ácido nucléico e do sítio de reconhecimento de enzima de restrição primária, significa "diretamente adjacente", de maneira tal que não existam nucleotídeos dispostos entre a sequência de ácido nucléico e o sítio de reconhecimento de enzima de restrição primária.

Segunda sequência de nucleotídeos

A segunda sequência de nucleotídeos é obtenível, obtida, identificada ou identificável por DNA genômico de reticulação (por exemplo, *in vivo* ou *in vitro*).

A segunda sequência de nucleotídeos (por exemplo, sequência de nucleotídeos de interesse) torna-se ligada à primeira sequência de nucleotídeos (alvo) após tratar uma amostra com um agente de reticulação e digerir/ligar os fragmentos de DNA reticulado. Tais sequências são reticuladas na primeira sequência de nucleotídeos (alvo) em decorrência de que elas estavam originalmente próximas no espaço nuclear e ligadas à primeira sequência

de nucleotídeos (alvo), em decorrência das condições de ligação favorecerem a ligação entre fragmentos de DNA reticulado (intramolecular) com relação a eventos de ligação aleatórios.

Doenças com base em alterações, tais como translocações, deleções, inversões, 5 duplicações e outros rearranjos genômicos, são em geral causadas por interações DNA-DNA aberrantes. A tecnologia 4C mede frequências de interação DNA-DNA, que são basicamente uma função da separação do sítio genômico, isto é, frequências de interação DNA-DNA são inversamente proporcionais à distância linear (em quilobases) entre dois loci de DNA presentes no mesmo molde de DNA físico (Dekker et al., 2002). Assim, a(s) alteração(s) que cria(m) moldes de DNA fisicamente diferentes e/ou novos é acompanhada por interações DNA-DNA alteradas e isto pode ser medido por tecnologia 4C.

Adequadamente, a segunda sequência de nucleotídeos tem pelo menos 40 pares de bases.

Agentes de reticulação, tal como formaldeído, podem ser usados para reticular proteínas em outras proteínas e ácido nucléico vizinhos. Assim, duas ou mais sequências de nucleotídeos podem ser reticuladas apenas por meio de ligação de proteínas a (uma de) estas sequências de nucleotídeos. Agentes de reticulação sem ser formaldeído também podem ser usados de acordo com a presente invenção, incluindo aqueles agentes de reticulação que reticulam diretamente as sequências de nucleotídeos. Exemplos de agentes que 15 reticulam DNA incluem, mas sem limitações, luz UV, mitomicina C, nitrogênio mostarda, 20 melfalan, 1,3-butadieno diepoxido, cis diaminadichloroplatina (II) e ciclofosfamida.

Adequadamente, o agente de reticulação formará reticulações que ligam distâncias relativamente pequenas, tal como cerca de 2 Å, selecionando por meio disto interações particulares que podem ser revertidas.

25 A reticulação pode ser realizada, por exemplo, incubando as células em formaldeído a 2% em temperatura ambiente, tal como incubando 1 x 10 células em 10 mL de DMEM-10% FCS suplementado com formaldeído a 2% por 10 minutos a temperatura ambiente.

#### Enzima de restrição primária

30 Da maneira aqui usada, o termo "enzima de restrição primária" refere-se a uma primeira enzima de restrição que é usada para digerir o DNA reticulado.

A enzima de restrição primária será escolhida dependendo do tipo de sequência alvo (por exemplo, locus) a ser analisada. É desejável que experimentos preliminares sejam realizados para otimizar as condições de digestão.

35 A enzima de restrição primária pode ser selecionada de enzimas de restrição que reconhecem pelo menos sequências de 8 bp ou mais de DNA. A enzima de restrição primária pode ser selecionada de enzimas de restrição que reconhecem pelo menos sequências de 7 bp ou mais de DNA. A enzima de restrição primária pode ser selecionada de enzimas

de restrição que reconhecem pelo menos sequências de 6 bp ou mais de DNA. Para algumas modalidades, a enzima de restrição primária pode ser selecionada de enzimas de restrição que reconhecem uma sequência de 4 bp e/ou 5 bp de DNA.

O uso de cortadores menos frequentes aumentará a distância genômica capturada 5 (coberta) por cada isca.

Enzimas de restrição que reconhecem sequências de 6 bp de DNA incluem, mas sem limitações, *AclI*, *HindIII*, *SspI*, *BspLUI* II, *Agel*, *MluI*, *SpeI*, *BglIII*, *Eco47III*, *StuI*, *Seal*, *ClaI*, *Avall*, *VspI*, *MfeI*, *PmaCI*, *PvuII*, *NdeI*, *Ncol*, *Smal*, *SacII*, *AvrII*, *PvuI*, *XmaIII*, *SphI*, *Xhol*, *PstI*, *AfIII*, *BamHI*, *AatII*, *Sad*, *EcoRV*, *SphI*, *Nael*, *BsePI*, *NheI*, *BamHI*, *NarI*, *Apal*, *KpnI*, *SnaI*, 10 *SaiI*, *ApaLI*, *HpaI*, *SnaBI*, *BspHI*, *BspMII*, *NruI*, *XbaI*, *BclI*, *MstI*, *BalI*, *Bsp1407I*, *PsI*, *AsuI* e *AhaIII*.

Enzimas de restrição que reconhecem mais que uma sequência de 6 bp de DNA incluem, mas sem limitações *BbvC I*, *Ascl*, *AsiS I*, *Fse I*, *Not I*, *Pac I*, *Pme I*, *Sbf I*, *SgrA I*, *Swa I*, *Sap I*, *Cci NI*, *FspA I*, *Mss I*, *Sgf I*, *Smi I*, *Srf I* e *Sse8387 I*.

15 Para alguns aspectos da presente invenção, no caso de enzimas de restrição que reconhecem sequências de 6 bp, *BglI*, *HindI* ou *EcoBI* são preferidas.

Enzimas de restrição que reconhecem sequências de 4 ou 5 bp de DNA incluem, mas sem limitações, *TspEI*, *Maell*, *AluI*, *NlaIII*, *HpaII*, *FnuDII*, *Mael*, *DpnI*, *MboI*, *HhaI*, *HaeIII*, *RsaI*, *TaqI*, *CviRI*, *MseI*, *SthI32I*, *AcI*, *DpnII*, *Sau3AI* e *MnII*. Em uma modalidade, a enzima 20 de restrição secundária é *NlaIII* e/ou *DpnII*.

A expressão "sítio de reconhecimento de enzima de restrição primária" refere-se ao sítio em uma sequência de nucleotídeos que é reconhecida e clivada pela enzima de restrição primária.

Para algumas modalidades, a enzima de restrição não digere DNA repetitivo ou 25 DNA que é relativamente sub-representado no DNA repetitivo. Isto pode aumentar o número de leituras interpretáveis.

#### Enzima de restrição secundária

Da maneira aqui usada, o termo "enzima de restrição secundária" refere-se a uma 30 segunda enzima de restrição que é opcionalmente usada após a digestão da enzima de restrição primária, ligação de DNA reticulado, desreticulação e (opcional) purificação de DNA. Em uma modalidade, a enzima de restrição secundária é usada para fornecer extremidades de DNA definidas nas sequências de nucleotídeos de interesse, que permite a ligação de sequências de composição de nucleotídeo conhecidas nos sítios de reconhecimento de enzima de restrição secundária que flanqueiam as sequências de nucleotídeos de interesse.

35 Em uma modalidade, a ligação de sequências de composição de nucleotídeo conhecida nos sítios de reconhecimento de enzima de restrição secundária que flanqueiam (por exemplo, estão em cada lado ou extremidade de) as sequências de nucleotídeos de

interesse envolve ligação em condições diluídas para favorecer a ligação intramolecular entre os sítios de reconhecimento de enzima de restrição secundária que flanqueiam sequências de nucleotídeos alvos e as sequências de nucleotídeos ligados de interesse. Isto resulta efetivamente na formação de círculos de DNA nos quais a sequência de nucleotídeos alvo conhecida flanqueia a sequências desconhecidas de interesse.

Em uma outra modalidade, a ligação de sequências de composição de nucleotídeo conhecida nos sítios de reconhecimento de enzima de restrição secundária que flanqueiam (por exemplo, estão em cada lado ou extremidade de) as sequências de nucleotídeos de interesse envolve a adição de sequências de DNA exclusivas de composição de nucleotídeo conhecida, seguida pela ligação em condições que favorecem ligação intermolecular entre os sítios de reconhecimento de enzima de restrição secundária que flanqueiam as sequências de nucleotídeos de interesse e introduzem sequências de DNA exclusivas de composição de nucleotídeo conhecida.

Em uma modalidade, a enzima de restrição secundária é escolhida de maneira tal que nenhum dos sítios de enzima de restrição secundária esteja a cerca de 350 bp (por exemplo, 350-400 bp) do sítio de restrição primária.

Em uma outra modalidade, a enzima de restrição secundária é escolhida de maneira tal que o mesmo sítio de enzima de restrição secundária é provavelmente para ser localizado na sequência de nucleotídeos ligados (isto é, a sequência reticulada ligada). Uma vez que as extremidades da primeira sequência de nucleotídeos (alvo) e a sequência de nucleotídeos ligados podem ser extremidades coesivas compatíveis (ou rombas), as sequências podem ainda estar ligadas a fim de circular o DNA. Dessa maneira, a etapa de digestão é seguida por ligação em condições diluídas que favorecem interações intramoleculares e circulação opcional do DNA por meio de extremidades compatíveis.

Preferivelmente, o sítio de reconhecimento de enzima de restrição secundária é um sítio de reconhecimento de sequência de nucleotídeos de 4 ou 5 bp. Enzimas que reconhecem sequências de 4 ou 5 bp de DNA incluem, mas sem limitações, TspEI, Maell, Alul, NlaI-II, Hpall, FnuDII, Mael, Dpnl, Mbol, Hhal, HaeIII, Rsal, Taql, CviRI, Msel, SthI 321, Acil, DpnII, Sau3 Al e MnII.

Em uma modalidade preferida, a enzima de restrição secundária é NlaIII e/ou DpnII.

O termo "sítio de reconhecimento de enzima de restrição secundária" refere-se ao sítio na sequência de nucleotídeos que é reconhecida e clivada pela enzima de restrição secundária.

Após a digestão com a enzima de restrição secundária, uma reação de ligação adicional é realizada. Em uma modalidade, esta reação de ligação liga sequências de DNA de composição de sequência de nucleotídeos conhecida no sítio de digestão de enzima de restrição secundária de uma ou mais sequências que estão ligadas à sequência de nucleotí-

deos alvo.

Para algumas modalidades, o método exclui a etapa de digerir as sequências de nucleotídeos com uma enzima de restrição secundária.

5 Para algumas modalidades, o método exclui ligar uma ou mais sequências de DNA de composição de nucleotídeo conhecida no(s) sítio(s) de digestão de enzima de restrição secundária disponível(s) que flanqueia(m) uma ou mais sequências de nucleotídeos de interesse.

#### Enzima de restrição terciária

10 Da maneira aqui usada, o termo "enzima de restrição terciária" refere-se a uma terceira enzima de restrição que pode ser usada opcionalmente após a etapa da enzima de restrição secundária a fim de linearizar o DNA circular antes da amplificação.

Preferivelmente, a enzima de restrição terciária é uma enzima que reconhece um sítio de reconhecimento de nucleotídeo de 6 bp ou mais.

15 Preferivelmente, a enzima de restrição terciária digere a primeira sequência de nucleotídeos (alvo) entre os sítios de reconhecimento da enzima de restrição primária e secundária. Da maneira a ser entendida pelos versados na técnica, é desejável que a enzima de restrição terciária não digira a primeira sequência de nucleotídeos (alvo) tão perto dos sítios de reconhecimento da enzima de restrição primária e secundária, de maneira tal que os iniciadores de amplificação não possam mais hibridizar. Dessa maneira, é preferível que 20 o sítio de reconhecimento de enzima de restrição terciária esteja localizado pelo menos na mesma distância dos sítios de reconhecimento da enzima de restrição primária e secundária que o comprimento do iniciador a ser usado, de maneira tal que o(s) iniciador(s) de amplificação(s) ainda possa(m) hibridizar.

25 Em uma modalidade preferida, a enzima de restrição terciária é uma que reconhece uma sequência de 6 bp de DNA.

O termo "sítio de reconhecimento de enzima de restrição terciária" refere-se ao sítio na sequência de nucleotídeos que é reconhecida e clivada pela enzima de restrição terciária.

#### Sítio de reconhecimento

30 Endonucleases de restrição são enzimas que clivam o cerne de fosfato e açúcar do DNA. Na maioria dos cenários práticos, uma dada enzima de restrição corta ambas as fitas de DNA duplex em um trecho de apenas algumas bases. Os substratos para enzimas de restrição são sequências de DNA de fita dupla denominados sítios/sequências de reconhecimento.

35 O comprimento dos sítios de reconhecimento de restrição varia, dependendo da enzima de restrição que é usada. O comprimento da sequência de reconhecimento dita com que frequência a enzima cortará em uma sequência de DNA.

A título de exemplo, inúmeras enzimas de restrição reconhecem uma sequência de 4 bp de DNA. As sequências e a enzima que reconhecem a sequência de 4 bp de DNA incluem, mas sem limitações, AATT (TspEI), ACGT (MaeII), AGCT (AluI), CATG (NlaIII), CCGG (HpaII), CGCG (FnuDII), CTAG (MaeI), GATC (DpnI, DpnII, Sau3AI & MboI), GCGC 5 (HhaI), GGCC (HaeIII), GTAC (RsaI), TCGA (TaqI), TGCA (CviRI), TTAA (MseI), CCCG (SphI32I), CCGC (AciI) e CCTC (MnII)

A título de exemplo adicional, inúmeras enzimas de restrição reconhecem uma sequência de 6 bp de DNA. As sequências e a enzima que reconhecem a sequência de 6 pares de bases bp de DNA incluem, mas sem limitações, AACGTT (AclI), AAGCTT (HindIII),

10 AATATT (SspI), ACATGT (BspLUIII), ACCGGT (AgiI), ACGCGT (MluI), ACTAGT (SpeI), AGATCT (BglIII), AGCGCT (Eco47III), AGGCCT (StuI), AGTACT (Seal), ATCGAT (ClaI), ATGCAT (AvalI), ATTAAT (Vspl), CAATTG (MfeI), CACGTG (PmaCI), CAGCTG (PvuII), CATATG (NdeI), CCATGG (Ncol), CCCGGG (SmaI), CCGCGG (SacII), CCTAGG (AvrII), CGATCG (PvuI), CGGCCG (XmaII), CGTACG (SphII), CTCGAG (Xhol), CTGCAG (PstI), 15 CTTAAG (AflIII), GAATT (BamHI), GACGTC (AatII), GAGCTC (Sad), GATATC (EcoRV), GCATGC (SphI), GCCGGC (NaeI), GCGCGC (BsePI), GCTAGC (NheI), GGATCC (BamHI), GGCGCC (NarI), GGGCCC (Apal), GGTACC (KpnI), GTATAC (SnaI), GTCGAC (SaiI), GTGCAC (ApaLI), GTTAAC (HpaI), TACGTA (SnaBI), TCATGA (BspHI), TCCGGA (BspMII), TCGCGA (NruI), TCTAGA (XbaI), TGATCA (BclI), TGCGCA (MstI), TGGCCA (BalI), 20 TGTACA (Bsp1407I), TTATAA (PstI), TTTCGAA (AsuI) e TTTAAA (AhaIII).

A título de exemplo adicional, inúmeras enzimas de restrição reconhecem uma sequência de 7 bp de DNA. As sequências e a enzima que reconhecem a sequência de 7 bp de DNA incluem, mas sem limitações CCTNAGG (SauI), GCTNAGC (EspI), GGTNACC BstEII e TCCNGGA Pfol.

25 A título de exemplo adicional, inúmeras enzimas de restrição reconhecem uma sequência de 8 bp de DNA. As sequências e a enzima que reconhecem a sequência de 8 bp de DNA incluem, mas sem limitações ATTTAAAT (Swal), CCTGCAGG (Sse8387I), CGCCGGCG (Sse232I), CGTCGACG (SgrDI), GCCCAGGC (SrfI), GCGATCGC (SgfI), GCGGCCGC (NotI), GGCCGGCC (FseI), GGCGCGCC (Ascl), GTTTAAC (PmeI) e 30 TTAATTAA (Pad).

Inúmeras destas enzimas contêm sequência CG que pode ser metilada *in vivo*. Inúmeras enzimas de restrição são sensíveis a esta metilação e não clivarão a sequência metilada, por exemplo, HpaII não clivará a sequência CC<sup>m</sup>GG considerando que seu isosquizômero Mspl é insensível a esta modificação e clivará a sequência metilada. Dessa maneira, em alguns exemplos, não são usadas enzimas sensíveis a metilação eucariótica.

Em uma modalidade, um sítio de reconhecimento é um sítio de digestão.

Em uma modalidade, um sítio de reconhecimento de enzima de restrição é um sítio

de digestão de enzima de restrição.

Circulação

De acordo com uma modalidade da presente invenção, o material para 4C é preparado criando círculos de DNA digerindo o molde de 3C com uma enzima de restrição secundária, seguido por ligação.

Preferivelmente, é escolhida uma enzima de restrição secundária é escolhida que corta a primeira sequência de nucleotídeos (alvo) em mais que cerca de 350 bp (por exemplo, 350-400 bp) do sítio de restrição primária. Vantajosamente, isto minimiza uma tendência na formação de círculo em virtude das restrições topológicas (Rippe et al. (2001) Trends in Biochem. Sciences 26, 733-40).

Preferivelmente, a enzima de restrição secundária é um cortador frequente que reconhece um sítio de reconhecimento de enzima de restrição de 4 ou um 5 bp. Assim, é possível obter os menores fragmentos de restrição para eficiências de amplificação iguais de todos os fragmentos ligados durante a amplificação.

Antes da digestão e ligação da enzima de restrição secundária, o molde de DNA compreenderá um sítio de reconhecimento de enzima secundária na primeira sequência de nucleotídeos (alvo) localizada a mais que cerca de 350-400 bp do sítio de restrição primária e um outro sítio de reconhecimento de enzima secundária localizado na sequência de nucleotídeos que foi ligada (isto é, na segunda sequência de nucleotídeos).

Preferivelmente, a etapa de digestão de enzima de restrição secundária é realizada por mais de 1 hora até toda a noite e seguida por inativação pelo calor da enzima.

Preferivelmente, o DNA nesta mistura da reação é purificado usando métodos/estojos convencionais que são conhecidos na técnica.

Após a etapa de digestão da enzima de restrição secundária, um sítio de enzima de restrição secundária será localizado em mais que 350-400 bp do sítio de restrição primária na primeira sequência de nucleotídeos (alvo) e um outro sítio de enzima de restrição secundária será localizado na sequência de nucleotídeos ligados (isto é, a segunda sequência de nucleotídeos). Uma vez que as extremidades da primeira sequência de nucleotídeos (alvo) e a sequência de nucleotídeos ligados têm extremidades compatíveis, as sequências podem estar ligadas a fim de circular o DNA.

A etapa de digestão é então seguida pela ligação em condições diluídas que favorecem interações e circulação intramolecular do DNA por meio de extremidades compatíveis.

Preferivelmente, a reação de ligação é realizada em uma concentração de DNA de cerca de 1- 5 ng/ $\mu$ L.

Preferivelmente, a reação de ligação é realizada por mais de 1 hora (por exemplo, 2, 3, 4 ou mais horas) a cerca de 16-25 °C.

Dessa maneira, após a reação de ligação, DNA circular pode ser preparado. O DNA circular compreenderá os sítios de reconhecimento para pelo menos a enzima de restrição secundária ou as enzimas de restrição primária e secundária. No DNA circular contendo a primeira sequência de nucleotídeos (alvo), o sítio de reconhecimento de enzima de restrição primária e os sítios de reconhecimento de enzima de restrição secundária definirão as extremidades da primeira sequência de nucleotídeos (alvo) e a sequência de nucleotídeos ligados (isto é, a segunda sequência de nucleotídeos). Dessa maneira, a primeira sequência de nucleotídeos (alvo) e a sequência de nucleotídeos ligados dão separadas (por exemplo, divididas) pelo sítio de reconhecimento de enzima de restrição primária e o sítio de reconhecimento de enzima de restrição secundária.

#### Amplificação

Uma ou mais reações de amplificação podem ser realizadas a fim de amplificar os moldes de DNA 4C.

A amplificação de DNA pode ser realizada usando inúmeros métodos diferentes que são conhecidos na técnica. Por exemplo, o DNA podem ser amplificado usando a reação em cadeia da polimerase (Saiki et al., 1988); PCR mediada por ligação, amplificação de replicase de Qb (Cahill, Foster e Mahan, 1991; Chetverin e Spirin, 1995; Katanaev, Kurnasov e Spirin, 1995); a reação em cadeia da ligase (LCR) (Landegren et al., 1988; Barany, 1991); o sistema de replicação de sequência auto-sustentável (Fahy, Kwok e Gingeras, 1991) e amplificação por substituição de fita (Walker et al., 1992).

Adequadamente, o DNA é amplificado usando PCR. "PCR" refere-se ao método de K. B. Mullis em patentes U.S. 4.683.195, 4.683.202, e 4.965.188 que descrevem um método para aumentar a concentração de um segmento de uma sequência de nucleotídeos em uma mistura de DNA genômico sem clonagem ou purificação.

Em uma modalidade, PCR inversa é usada. PCR inversa (IPCR) (descrita por Ochman et al (1988) Genetics 120(3), 621-3) é um método para a amplificação rápida *in vitro* de sequências de DNA que flanqueiam uma região de sequência conhecida. O método usa a reação em cadeia da polimerase (PCR), mas ele tem os iniciadores orientados na direção reversa da orientação usual. O molde para os iniciadores reversos é um fragmento de restrição que foi ligado mediante ele mesmo formar um círculo. PCR inversa tem muitas aplicações na genética molecular, por exemplo, a amplificação e identificação de sequências que flanqueiam elementos transponíveis. Para aumentar a eficiência e reprodutibilidade da amplificação, é preferível que os círculos de DNA sejam linearizados antes da amplificação usando uma enzima de restrição terciária. Preferivelmente, uma enzima de restrição terciária corta a primeira sequência de nucleotídeos (alvo) entre os sítios de enzima de restrição primária e secundária.

A digestão do molde de 3C com a enzima de restrição secundária, circulação opcional, ligação (por exemplo, ligação em condições diluídas) e linearização opcional da primeira sequência de nucleotídeos (alvo) contendo círculos produzem um molde de DNA para amplificação ("molde de DNA 4C").

5 Para a etapa de amplificação, pelo menos dois oligonucleotídeos iniciadores são usados, nos quais cada iniciador hibridiza em uma sequência de DNA que flanqueia as sequências de nucleotídeos de interesse. Em uma modalidade preferida, pelo menos dois oligonucleotídeos iniciadores são usados, nos quais cada iniciador hibridiza na sequência alvo que flanqueia as sequências de nucleotídeos de interesse.

10 Em uma modalidade, a palavra "flanquear", no contexto da hibridização do iniciador, significa que pelo menos um iniciador hibridiza em uma sequência de DNA adjacente a uma extremidade (por exemplo, a extremidade 5') da sequência de nucleotídeos de interesse, e pelo menos um iniciador hibridiza em uma sequência de DNA na outra extremidade (por exemplo, a extremidade 3') da sequência de nucleotídeos de interesse. Preferivelmente, 15 pelo menos um iniciador direto hibridiza em uma sequência de DNA adjacente a uma extremidade (por exemplo, a extremidade 5') da sequência de nucleotídeos de interesse e pelo menos um iniciador reverso hibridiza em uma sequência de DNA na outra extremidade (por exemplo, a extremidade 3') da sequência de nucleotídeos de interesse.

20 Em uma modalidade preferida, a palavra "flanquear", no contexto da hibridização do iniciador, significa que pelo menos um iniciador hibridiza em uma sequência alvo adjacente a uma extremidade (por exemplo, a extremidade 5') da sequência de nucleotídeos de interesse e pelo menos um iniciador hibridiza em uma sequência alvo na outra extremidade (por exemplo, a extremidade 3') da sequência de nucleotídeos de interesse. Preferivelmente, 25 pelo menos um iniciador direto hibridiza em uma sequência alvo adjacente a uma extremidade (por exemplo, a extremidade 5') da sequência de nucleotídeos de interesse e pelo menos um iniciador reverso hibridiza em uma sequência alvo na outra extremidade (por exemplo, a extremidade 3') da sequência de nucleotídeos de interesse.

30 Da maneira aqui usada, a palavra "iniciador" refere-se a um oligonucleotídeo, quer de ocorrência natural como em uma digestão de restrição purificada quer produzido sinteticamente, que é capaz de agir como um ponto de iniciação de síntese quando colocado em condições em que a síntese de um produto de extensão de iniciador, que é complementar a uma fita de ácido nucléico, é induzida (isto é, na presença de nucleotídeos e um agente de indução tal como DNA polimerase e em uma temperatura e pH adequados). O iniciador é preferivelmente de fita simples para eficiência máxima na amplificação, mas pode ser de fita dupla. Se for de fita dupla, o iniciador é primeiro tratado para separar suas fitas antes de ser usado para preparar produtos de extensão. Preferivelmente, o iniciador é um oligodesoxirribonucleotídeo. O iniciador tem que ser suficientemente longo para iniciar a síntese de pro-

dutos de extensão na presença do agente de indução. Os comprimentos exatos dos iniciadores dependerão de muitos fatores, incluindo temperatura, origem do iniciador e o uso do método.

Adequadamente, os iniciadores terão pelo menos 15, preferivelmente pelo menos 5 16, 17, 18, 19 ou 20, por exemplo, pelo menos 25, 30 ou 40 nucleotídeos de comprimento. Preferivelmente, os iniciadores de amplificação têm de 16 a 30 nucleotídeos de comprimento.

Preferivelmente, os iniciadores são desenhados para estar o mais próximo possível dos sítios de reconhecimento da enzima de restrição primária e secundária que separam a 10 primeira sequência de nucleotídeos (alvo) e a segunda (capturada) sequência de nucleotídeos. Os iniciadores podem ser desenhados de maneira tal que eles estejam em cerca de 100 nucleotídeos, tais como cerca de 90, 80, 70, 60, 50, 40, 30, 20, 10, 9, 8, 7, 6, 5, 4, 3, 2 ou 1 nucleotídeo(s) longe dos sítios de reconhecimento de enzima de restrição primária e secundária.

15 Adequadamente, os iniciadores podem ser desenhados de maneira tal que eles se sobreponham parcial ou completamente nos sítios de reconhecimento da enzima de restrição primária e secundária.

20 Adequadamente, os iniciadores de amplificação são desenhados de maneira tal que suas extremidades 3' fiquem voltadas para fora no sentido da segunda sequência de nucleotídeos.

Em uma modalidade, a amplificação pode ser combinada com a adição de sequências adicionais às extremidades dos produtos amplificados. Preferivelmente, estas sequências adicionais são sequências adaptadoras exigidas para o sequenciamento de alto desempenho. Adequadamente, iniciadores contêm uma projeção (por exemplo, uma projeção 25 5'). Adequadamente, a projeção adiciona parte ou a sequência adaptadora completa exigida para o sequenciamento de alto desempenho. Adequadamente, os iniciadores contêm uma projeção (por exemplo, uma projeção 5') que adiciona parte, ou a sequência completa usada para iniciar a reação de sequenciamento in o sequenciamento de alto desempenho. Dessa maneira, em uma modalidade, os iniciadores sobrepõem-se parcial ou completamente com 30 os sítios de reconhecimento de enzima de restrição primária e/ou secundária com as sequências adaptadoras e de iniciação de sequência adicionadas como projeções 5' no iniciador. Além do mais, ou alternativamente, os iniciadores contêm uma fração conjugada (por exemplo, biotina) que permite a separação subsequente de produtos de PCR do molde 4C genômico.

35 Se o método de amplificação que é usado for PCR inversa, então é preferível que as reações de amplificação sejam realizadas em cerca de 100-400 ng de DNA do molde 4C (por cerca de 50  $\mu$ L de reação de mistura de PCR) ou outras quantidades de DNA para as

quais PCR reações replicadas fornecem resultados reproduutíveis (ver figura 1) e incluem um número máximo de eventos de ligação por reação de PCR.

Preferivelmente, a reação de amplificação de PCR inversa é realizada usando o Expand Long Molde PCR System (Roche), usando tampão 1, de acordo com as instruções do fabricante.

#### Amostras

O termo "amostra" da maneira aqui usada tem seu significado natural. Uma amostra pode ser qualquer entidade física compreendendo DNA que é ou é capaz de ser reticulado. A amostra pode ser ou pode ser derivada de material biológico.

10 A amostra pode ser ou pode ser derivada de uma ou mais entidades, tais como uma ou mais células, um ou mais núcleos, ou uma ou mais amostras de tecido. As entidades podem ser ou podem ser derivadas de quaisquer entidades nas quais o DNA, tal como cromatina, está presente. A amostra pode ser ou pode ser derivada de uma ou mais células isoladas ou uma ou mais amostras de tecido isoladas, ou um ou mais núcleos.

15 A amostra pode ser ou pode ser derivada de células vivas e/ou células mortas e/ou lisados nucleares e/ou cromatina isolada.

A amostra pode ser ou pode ser derivada de sujeitos doentes e/ou não doentes.

A amostra pode ser ou pode ser derivada de um sujeito que suspeita-se estar sofrendo de uma doença.

20 A amostra pode ser ou pode ser derivada de um sujeito que deve ser testado com relação a probabilidade que ele sofrerá de uma doença no futuro.

A amostra pode ser ou pode ser derivada de material de paciente viável ou não viável.

25 A fixação de células e tecidos para uso na preparação do molde de 3C é descrita em detalhes em Splinter et ali, (2004) Methods Enzymol. 375, 493-507.

#### Marcador

30 Preferivelmente, as sequências de nucleotídeos (por exemplo, moldes de DNA 4C amplificados, iniciadores ou sondas etc.) são marcadas a fim de assistir em suas aplicações à jusante, tal como hibridização de arranjo. A título de exemplo, os moldes de DNA 4C podem ser marcados usando iniciação aleatória ou tradução com corte.

Uma grande variedade de marcadores (por exemplo, reportadores) pode ser usada para marcar as sequências de nucleotídeos aqui descritas, particularmente durante a etapa de amplificação. Marcadores adequados incluem radioisótopos, enzimas, agentes fluorescentes, quimioluminescentes ou cromogênicos, bem como substratos, cofatores, inibidores, partículas magnéticas e similares. Patentes que preceituam o uso de tais marcadores incluem US-A-3.817.837, US-A-3.850.752, US-A-3.939.350, US-A-3.996.345, US-A-4.277.437, US-A-4.275.149 e US-A-4.366.241.

Marcadores adicionais incluem, mas sem limitações,  $\beta$ -galactosidase, invertase, proteína verde fluorescente, luciferase, cloranfenicol, acetiltransferase,  $\beta$ -glucuronidase, exo-glucanase e glicoamilase. Marcadores fluorescentes também podem ser usados, bem como reagentes fluorescentes especificamente sintetizados com propriedades químicas particulares. Uma grande variedade de maneiras de medir fluorescência está disponível. Por exemplo, alguns marcadores fluorescentes exibem uma mudança no espectro de excitação ou emissão, algumas exibem transferência de energia de ressonância onde um reportador fluorescente perde a fluorescência, enquanto um segundo ganha fluorescência, alguns exibem uma perda (finalização) ou aparência de fluorescência, enquanto alguns reportam movimentos rotacionais.

A fim de obter material suficiente para marcar, múltiplas amplificações podem ser agrupadas, em vez de aumentar o número de ciclos de amplificação por reação. Alternativamente, nucleotídeos marcados podem ser incorporados nos últimos ciclos da reação de amplificação (por exemplo, 30 ciclos de PCR (sem marcador) + 10 ciclos de PCR (mais marcador)).

#### Arranjo

Em uma modalidade particularmente vantajosa, os moldes de DNA 4C que são preparados de acordo com os métodos aqui descritos podem ser hibridizados em um arranjo. Dessa maneira, a tecnologia do arranjo (por exemplo, micro-arranjo) pode ser usada para identificar sequências de nucleotídeos, tais como fragmentos genômicos, que compartilham frequentemente um sítio nuclear com uma primeira sequência de nucleotídeos (alvo).

Arranjos existentes, tais como arranjos de expressão e genômicos, podem ser usados de acordo com a presente invenção. Entretanto, a presente invenção também busca fornecer arranjos inéditos (por exemplo, arranjos de DNA) da maneira aqui descrita.

Um "arranjo" é uma coleção intencionalmente criada de ácidos nucléicos que pode ser preparada tanto sinteticamente quanto biossinteticamente e triada com relação à atividade biológica em uma variedade de formatos diferentes (por exemplo, bibliotecas de moléculas solúveis; e bibliotecas de oligos presas à microesferas de resina, chips de sílica ou outros suportes sólidos). Adicionalmente, o termo "arranjo" inclui aquelas bibliotecas de ácidos nucléicos que podem ser preparadas apontando ácidos nucléicos de essencialmente qualquer comprimento (por exemplo, de 1 a cerca de 1000 monômeros de nucleotídeos de comprimento) em um substrato.

Tecnologia do arranjo e as várias técnicas e aplicações associadas a ela estão descritas em geral em inúmeros livros texto e documentos. Este incluem Lemieux et al, 1998, Molecular Breeding 4, 277-289, Schena e Davis. Parallel Analysis with Biological Chips, em PCR Methods Manual (eds. M. Innis, D. Gelfand, J. Sninsky), Schena e Davis, 1999, Genes, Genomas and Chips. In DNA Microarrays: A Practical Approach (ed. M. Sche-

na), Oxford University Press, Oxford, UK, 1999), The Chipping Forecast (Nature Genetics special issue; January 1999 Supplement), Mark Schena (Ed.), Microarray Biochip Technology, (Eaton Publishing Company), Cortes, 2000, The Scientist 14[17]:25, Gwynne e Page, Microarrays analysis: the next revolution in molecular biology, Science, 1999 August 6; e 5 Eakins e Chu, 1999, Trends in Biotechnology, 17, 217-218.

Tecnologia do arranjo supera as desvantagens com métodos tradicionais em biologia molecular, que trabalham em geral em uma base "um gene em um experimento", resultando em baixo desempenho e na incapacidade de apreciar a "figura total" da função do gene. Simultaneamente, as principais aplicações para tecnologia do arranjo incluem a identificação de sequência (gene/mutação de gene) e a determinação do nível de expressão (abundância) de genes. O perfilamento da expressão de gene pode promover o uso da tecnologia do arranjo, opcionalmente em combinação com técnicas proteômicas (Celis et al, 2000, FEBS Lett, 480(1):2-16; Lockhart e Winzeler, 2000, Nature 405(6788):827-836; Khan et al., 10 1999, 20(2):223-9). Outras aplicações de tecnologia do arranjo também são conhecidas na técnica; por exemplo, descoberta de gene, pesquisa de câncer (Marx, 2000, Science 289: 15 1670-1672; Scherf, et al, 2000, Nat Genet;24(3):236-44; Ross et al, 2000, Nat Genet. 2000 Mar;24(3):227-35), análise SNP (Wang et al, 1998, Science, 280(5366):1077-82), descoberta de medicamento, farmacogenômicos, diagnóstico de doença (por exemplo, utilizando dispositivos de microfluidos: Chemical & Engineering News, February 22, 1999, 77(8):27- 36), 20 toxicologia (Rockett e Dix (2000), Xenobiotica, 30(2): 155-77; Afshari et al., 1999, Cancer Resl;59(19):4759-60) e toxicogenômicos (um híbrido de genômicos funcionais e toxicologia molecular).

Em geral, qualquer biblioteca pode ser arranjada de uma maneira em ordem em um arranjo, separando espacialmente os elementos da biblioteca. Exemplos de bibliotecas adequadas para arranjar incluem bibliotecas de ácido nucléico (incluindo bibliotecas de DNA, 25 DNAc, oligonucleotídeo, etc.), bibliotecas de peptídeo, polipeptídeo e proteína, bem como bibliotecas compreendendo quaisquer moléculas, tais como bibliotecas ligantes, entre outras.

As amostras (por exemplo, elementos de uma biblioteca) são em geral fixas ou imobilizadas em uma fase sólida, preferivelmente um substrato sólido, para limitar difusão e mistura das amostras. Em uma modalidade preferida, bibliotecas de ligantes de ligação de DNA podem ser preparadas. Em particular, as bibliotecas podem ser imobilizadas em uma fase sólida substancialmente planar, incluindo membranas e substratos não porosos, tais como plástico e vidro. Além disso, as amostras são preferivelmente arranjadas em uma maneira tal que esta que a indexação (isto é, referência ou acesso a uma amostra particular) seja facilitada. Tipicamente, as amostras são aplicadas como pontos em uma formação de grade. Sistemas de ensaio comuns podem ser adaptados com este propósito. Por exemplo, 30 35

um arranjo pode ser imobilizado na superfície de uma microplaca, tanto com múltiplas amostras em um poço, quanto com uma única amostra em cada poço. Além disso, o substrato sólido pode ser uma membrana, tal como uma membrana de nitrocelulose ou de náilon (por exemplo, membranas usadas em experimentos de blotting). Substratos alternativos incluem 5 vidro ou substratos a base de sílica. Assim, as amostras são imobilizadas por qualquer método adequado conhecido na técnica, por exemplo, por interações de carga, ou por acoplamento químico nas paredes ou fundo dos poços, ou na superfície da membrana. Outros meios de arranjo e fixação podem ser usados, por exemplo, pipetagem, remoção de gota residual com um leve toque, meios piezelétricos, tecnologia jato de tinta e jato de bolha, a- 10 plicação eletrostática, etc. No caso de chips a base de silício, a fotolitografia pode ser utilizada para arranjar e fixar as amostras no chip.

As amostras podem ser arranjadas sendo "apontadas" no substrato sólido; isto pode ser realizado manualmente ou elaborando o uso de robótica para depositar a amostra. Em geral, os arranjos podem ser descritos como macroarranjos ou microarranjos, a diferen- 15 ça sendo o tamanho dos pontos da amostra. Os macroarranjos contêm tipicamente tama- nhos de ponto de amostra de cerca de 300 microns ou maiores e podem ser facilmente ima- geados por varredores de gel e blot existentes. Os tamanhos de ponto de amostra em mi- croarranjos são tipicamente menores que 200 microns em diâmetro e estes arranjos contêm 20 normalmente milhares de pontos. Assim, microarranjos podem exigir robótica e equipamento de imageamento especializados, que podem necessitar ser feitos sob medida. A instrumen- tação é descrita em geral em uma revisão de Cortese, 2000, O Scientist 14[11]:26.

Técnicas para produzir bibliotecas imobilizadas de moléculas de DNA são descritas na técnica. Em geral, a maioria dos métodos da técnica anterior descreveu como sintetizar bibliotecas de moléculas de ácido nucléico de fita simples usando, por exemplo, técnicas de 25 mascaramento para constituir várias permutas de sequências nas várias posições discretas no substrato sólido. A patente U.S. 5.837.832 descreve um método melhorado para produzir arranjos de DNA imobilizados em substratos de silício com base na tecnologia de integração de escala muito grande. Em particular, patente U.S. 5.837.832 descreve uma estratégia de- 30 nominada "ladrilhamento" para sintetizar conjuntos específicos de sondas em locais espacialmente definidos em um substrato que pode ser usado para produzir as bibliotecas de DNA imobilizadas da presente invenção. A patente U.S. 5.837.832 também fornece referências para técnicas anteriores que também podem ser usadas.

Arranjos também pode ser construídos usando química de fotodeposição.

Arranjos de peptídeos (ou peptideomiméticos) também podem ser sintetizados em 35 uma superfície de uma maneira que coloque cada elemento distinto da biblioteca (por exemplo, sequência de peptídeo exclusiva) em um local discreto predefinido no arranjo. A identidade de cada elemento da biblioteca é determinada por seu local espacial no arranjo.

Os locais no arranjo onde interações de ligação entre uma molécula predeterminada (por exemplo, um alvo ou sonda) e elementos de biblioteca reativos ocorrem são determinados, identificando por maio disto as sequências de os elementos de biblioteca reativos com base no local espacial. Estes métodos são descritos na patente U.S. 5.143.854; WO90/15070 e 5 WO92/10092; Fodor et al. (1991) *Science*, 251: 767; Dower e Fodor (1991) *Ann. Rep. Med. Chem.*, 26: 271.

Para auxiliar a detecção, marcadores foram tipicamente usados (da maneira discutida anteriormente), tal como qualquer reportador facilmente detectável, por exemplo, um reportador fluorescente, bioluminescente, fosforescente, radioativo, etc. Tais reportadores, 10 sua detecção, acoplamento a alvos/sondas, etc. são discutidos em outra parte deste documento. A marcação de sondas e alvos também é discutida em Shalon et al, 1996, *Genoma Res* 6(7):639-45.

Exemplos específicos de arranjos de DNA são como a seguir:

Formato I: DNAc de sonda (500-5.000 bases de comprimento) é imobilizado em 15 uma superfície sólida, tal como vidro usando apontamento por robô, e exposto em um conjunto de alvos tanto separadamente quanto em uma mistura. Este método, é amplamente considerado como desenvolvido em Stanford University (Ekins e Chu, 1999, *Trends in Biotechnology*, 1999, 17, 217-218).

Formato II: um arranjo de sondas de oligonucleotídeos (ologos de 20-25-mer, preferivelmente, oligos de 40-60 mer) ou ácido nucléico de peptídeo (PNA) é sintetizado tanto *in situ* (no chipe) quanto por síntese convencional seguido por imobilização no chipe. O arranjo é exposto ao DNA de amostra marcada, hibridizado, e a identidade/abundância de sequências complementares é determinada. Um chipe de DNA como este é vendido por Affymetrix, Inc., com o nome comercial GeneChip®. Agilent e Nimblegen também fornecem arranjos 25 adequados (por exemplo, arranjos de ladrilhamento genômico).

Exemplos de alguns formatos de microarranjo comercialmente disponíveis são demonstrados na tabela 1 a seguir (ver também Marshall e Hodgson, 1998, *Nature Biotechnology*, 16(1), 27-31).

Empresa	Nome do produto	Método de arranjo	Etapa de hibridização	Saída
Affymetrix, Inc., Santa Clara, Califórnia	GeneChip®	Síntese fotolito-gráfica (no chipe) <i>in situ</i> de oligos de ~ 20-25 mer em pastilhas de silício, que são	10.000-260.000 características oligo sondadas com 30-40 fragmentos de nucleotídeo marcados de DNAc de amostra	Fluorescência

		cortadas em chips de 1,25 cm <sup>2</sup> ou 5,25 cm <sup>2</sup>	ou RNA antissentido	
Brax, Cambridge, UK		oligo sintético pequeno, sintetizado de fora do chip	1.000 oligos em um "chipe universal" sondado com ácido nucléico alvejado	espectrometria de massa
<u>Gene Logic, Inc.</u> , Columbia, Maryland	READS™			
<u>Geometrix, Inc.</u> , The Woodlands, Texas	Universal Arrays™			
<u>Hyseq Inc.</u> , Sunnyvale, Califórnia	HyChip™	amostras de DNA de 500-2.000 nt impressas em membranas de 0,6 cm <sup>2</sup> (HyGnostics) ou ~ 18 cm <sup>2</sup> (Gene Discovery) 5 oligômeros fabricados impressos como arranjos de 1,15 cm <sup>2</sup> em vidro (HyChip)	pontos de DNAc de 64 amostras sondados com 8.000 oligos 7-mer (HyGnostics) ou <= pontos de DNAc de 55.000 amostras sondados com 300 oligo de 7-mer (Gene Discovery) Pontos de oligo 1024 universais sondaram DNAcs de amostra de 10 kb, marcado com oligo 5-mer e ligase	Radioisótopo f
<u>Incyte Pharmaceutica 1s, Inc.</u> , PAlo Alto, Califórnia	GEM	Impressão piezo-elétrica para pontuação de fragmentos de PCR e síntese no chip de oligos	<=1.000 (eventualmente 10.000) pontos de oligo/fragmento de PCR sondados com RNA marcado	fluorescência e radioisótopo
<u>Molecular Dy-</u>	Storm® Flu-	DNAcs de 500-	pontos de DNAc de	fluorescência

<u>namics Inc.</u> , Sunnyvale, Califórnia	orlImager®	5.000 nt impressos à caneta em lâmina de vidro de ~ 10 cm <sup>2</sup>	~ 10.000 sondados com DNAcs de amostra marcada de 200-400 nt	
Nanogen, San Diego, Califórnia	Semiconductor Microchip	Oligos ~20-mer, capturados em pontos eletroatrativos em pastilhas de silício, que são cortadas em chips ≤1 cm <sup>2</sup>	pontos de oligos 25, 64, 400 (e eventualmente 10.000) polarizados para melhorar a hibridização de DNAcs de amostra marcados de 200-400 nt	fluorescência
Protogene Laboratories, Palo Alto, Califórnia		Síntese no chipe de oligos de 40-50 mer em chipe de vidro de 9 cm <sup>2</sup> por meio da impressão em um arranjo de tensão de superfície	pontos de oligo ≤ 8.000 sondados com ácidos nucléicos de amostra marcada de 200-400 nt	fluorescência
Sequenom, Hamburg, Alemanha e San Diego, Califórnia	MassArray SpectroChip	Impressão de arranjo; em torno de oligos de 20-25-mer	250 localizações por SpectroChip interrogado por dessorção a laser e espectroscopia de massa	espectroscopia de massa
<u>Synteni Inc.</u> , Fremont, Califórnia	UniGEM™	DNAcs de 500-5.000 nt impressos pela ponta em chipe de vidro de ~ 4 cm <sup>2</sup>	<= pontos de DNAc de 10.000 sondados com DNAcs de amostra marcada de 200-400 nt	fluorescência
<u>Nimblegen Systems Inc.</u> , Madison	<i>Homo sapiens</i> Whole-Genome 60mer Microarray	38.000 transcrições com 5 sondas por gene de 17,4 mm x 13 mm		plataforma de varredura de 5 micron
The German Cancer Insti-		macrochipe PNA prototípico com	em torno de 1.000 pontos em um chipe	fluorescência/espectroscopia

tute, Heidel- berg, Ale- manha		síntese no chipe de sondas usan- do química f-moc ou t-moc	de 8 x 12 cm	a de massa
--------------------------------------	--	---	--------------	------------

Tabela 1: Exemplos de formatos de microarranjo de hibridização atualmente disponíveis

A fim de gerar dados de ensaios baseados em arranjo, é detectado um sinal que significa a presença ou ausência de hibridização entre uma sonda e uma sequência de nucleotídeos. A presente invenção contempla adicionalmente técnicas de marcação direta e indireta. Por exemplo, marcação direta incorpora corantes fluorescentes diretamente nas sequências de nucleotídeos que hibridizam no arranjo associado a sondas (por exemplo, corantes são incorporados na sequência de nucleotídeos por síntese enzimática na presença de nucleotídeos marcados ou iniciadores de PCR). Esquemas de marcação direta produzem sinais fortes de hibridização, usando tipicamente famílias de corantes fluorescentes com estruturas e características químicas similares e são simples de implementar. Em modalidades preferidas compreendendo marcação direta de ácidos nucléicos, análogos de cianina ou alexa são utilizados em análise de arranjo comparativo multifluor. Em outras modalidades, esquemas de marcação indireta podem ser utilizados para incorporar epítopos nos ácidos nucléicos tanto antes quanto após a hibridização nas sondas de microarranjo. Um ou mais procedimentos e reagentes de coloração são usados para marcar o complexo hibridizado (por exemplo, uma molécula fluorescente que se liga aos epítopos, fornecendo por meio disto um sinal fluorescente em virtude da conjugação de molécula corante no epítopo da espécie hibridizada).

A análise de dados também é uma importante parte de um experimento envolvendo arranjos. Os dados brutos de um experimento de arranjo são tipicamente imagens que precisam ser transformadas em matrizes, tabelas onde linhas representam, por exemplo, genes, colunas representam, por exemplo, várias amostras tais como tecidos ou condições experimentais, e números em cada célula, por exemplo, caracterizam a expressão de uma sequência particular (preferivelmente, uma segunda sequência que foi ligada na primeira sequência de nucleotídeos (alvo)) na amostra particular. Estas matrizes têm que ser analisadas adicionalmente, se qualquer conhecimento sobre os processos biológicos fundamentais tiver que ser extraído. Métodos de análise de dados (incluindo análise de dados supervisada e não supervisada, bem como abordagens de bioinformática) são revelados em Brazma e Vilo J (2000) FEBS Lett 480(1): 17-24.

Da maneira aqui descrita, uma ou mais sequências de nucleotídeos (por exemplo, o molde de DNA) que são marcadas e, subsequentemente hibridizadas em um arranjo, com-

5 preendem uma sequência de nucleotídeos que é enriquecida por pequenos trechos de sequências com uma assinatura distinta, isto é, cobrindo a sequência de nucleotídeos entre o sítio de reconhecimento de enzima de restrição primária que foi ligado durante o procedimento 3C na primeira sequência de nucleotídeos (alvo), e seus respectivos sítios de reconhecimento de enzima de restrição secundária vizinhos.

Um arranjo único pode compreender múltiplas (por exemplo, duas ou mais) sequências de iscas.

#### Mapa cromossômico

10 É descrito adicionalmente aqui um método envolvendo a caracterização de fragmentos de DNA reticulados como uma consequência de sua proximidade espacial ou física, com o propósito da (re)construção de mapas cromossômicos (por exemplo, mapas cromossômicos lineares) e a identificação diagnóstica de mudanças nestes mapas cromossômicos.

15 Vantajosamente, tais técnicas podem ser ampliadas com propósitos diagnósticos, tal como reconstruir mapas cromossômicos físicos e identificar mudanças nestes mapas em decorrência de rearranjos genômicos.

A metodologia também pode ser usada para identificar variação genômica (natural) que não está necessariamente associada com doença, mas pode predispor um sujeito a um certo traço (por exemplo, um traço mental ou comportamental).

20 Em um aspecto, é fornecido um método para construir pelo menos um mapa cromossômico linear de um indivíduo compreendendo as etapas de: (a) fornecer uma amostra de ácido nucléico do dito indivíduo; (b) reticular o ácido nucléico na dita amostra; (c) digerir o ácido nucléico com uma enzima de restrição primária; (d) ligar as sequências de nucleotídeos reticulados; (e) reverter a reticulação; (f) analisar os produtos de ligação; (g) construir pelo menos um mapa cromossômico linear; e (h) identificar um ou mais rearranjos genômicos no mapa cromossômico linear dos ditos indivíduos.

25 Em um aspecto adicional, é fornecido um método para diagnosticar uma doença ou identificar um traço causado por um ou mais rearranjos genômicos em um cromossomo compreendendo as etapas de: (a) digerir uma amostra de DNA reticulado com uma enzima de restrição primária; (b) ligar as sequências de nucleotídeos reticulados; (c) reverter a reticulação; (d) analisar os produtos de ligação; (e) construir pelo menos um mapa cromossômico linear; (f) identificar um ou mais rearranjos genômicos no mapa cromossômico linear; e (g) correlacionar um ou mais rearranjos genômicos com uma doença ou um traço.

#### Sondas

30 Da maneira aqui usada, a palavra "sonda" refere-se a uma molécula (por exemplo, um oligonucleotídeo, quer de ocorrência natural como em uma digestão de restrição purificada quer produzido sinteticamente, recombinantemente ou por amplificação de PCR), que é capaz de hibridizar em uma outra molécula de interesse (por exemplo, um outro oligonu-

cleotídeo). Quando as sondas são oligonucleotídeos, elas podem ser de fita simples ou fita dupla. As sondas são usadas na detecção, identificação e isolamento de alvos particulares (por exemplo, sequências de gene). Da maneira aqui descrita, contempla-se que as sondas usadas na presente invenção podem ser marcadas com um marcador, de maneira tal que 5 seja detectável em qualquer sistema de detecção, incluindo, mas sem limitação, sistema de enzima (por exemplo, ELISA, bem como ensaios histoquímicos a base de enzima), fluorescente, radioativo e luminescente.

10 Com relação a arranjos e microarranjos, a palavra "sonda" é usada para se referir a qualquer material hibridizável que está fixado ao arranjo com o propósito de detectar uma sequência de nucleotídeos que foi hibridizada na dita sonda. Preferivelmente, estas sondas têm 25- 60 mers ou mais.

15 Estratégias para desenho de sonda são descritas em WO95/11995, EP 717,113 e WO97/29212.

Uma vez que 4C permite uma pesquisa genômica não tendenciosa de interações, é 20 vantajoso preparar um arranjo com sondas que interroguem cada sítio de reconhecimento de enzima de restrição primária (por exemplo, exclusivo/não repetitivo) possível no genoma. Assim, o desenho do arranjo depende apenas da escolha de enzima de restrição primária e não das reais sequências de nucleotídeos primárias ou secundárias.

25 Embora arranjos existentes possam ser usados de acordo com a presente invenção, é preferível usar configurações alternativas.

Em uma configuração, uma ou mais sondas no arranjo são desenhadas de maneira tal que elas possam hibridizar próximo aos sítios que foram digeridos pela enzima de restrição primária. Mais preferivelmente, a(s) sonda(s) estão em cerca de 20 bp do sítio de reconhecimento de enzima de restrição primária. Mais preferivelmente, a(s) sonda(s) estão em cerca de 50 bp do sítio de reconhecimento de enzima de restrição primária.

30 Adequadamente, a(s) sonda(s) estão em cerca de 100 bp (por exemplo, cerca de 0-100 bp, cerca de 20-100 bp) do sítio de reconhecimento de enzima de restrição primária.

Em uma configuração preferida, uma sonda única exclusiva é desenhada em 100 bp em cada lado dos sítios que são digeridos pela enzima de restrição primária.

35 Em uma outra configuração preferida, as posições de sítios digeridos pela enzima de restrição secundária com relação às posições de sítios digeridos pelos sítios de restrição primária são levadas em consideração. Nesta configuração, uma sonda única exclusiva é desenhada apenas em cada lado dos sítios digeridos pela enzima de restrição primária que tem o sítio de reconhecimento de enzima de restrição secundária mais próximo em um distância longa o suficiente de uma sonda de um dado comprimento a ser desenhada entre o sítio de reconhecimento de enzima de restrição primária e secundária. Nesta configuração, por exemplo, nenhuma sonda é desenhada ao lado de um sítio de reconhecimento de enzi-

ma de restrição primária particular que tem um sítio de reconhecimento de enzima de restrição secundária em 10 bp naquele mesmo lado.

Em uma outra configuração, as sondas no arranjo são desenhadas de maneira tal que elas possam hibridizar em cada um dos lados dos sítios que são digeridos pela enzima de restrição primária. Adequadamente, uma sonda única em cada lado do sítio de reconhecimento de enzima de restrição primária pode ser usada.

Ainda em uma outra configuração, duas ou mais sondas (por exemplo, 3, 4, 5, 6, 7 ou 8 ou mais) podem ser desenhadas em cada lado do sítio de reconhecimento de enzima de restrição primária, que pode então ser usado para investigar o mesmo evento de ligação.

Para o número e posição de sondas com relação a cada sítio de reconhecimento de enzima de restrição primária, o local genômico exato de seu sítio de reconhecimento de enzima de restrição secundária vizinho pode ser levado em consideração.

Ainda em uma outra configuração, duas ou mais sondas (por exemplo, 3, 4, 5, 6, 7 ou 8 ou mais) podem ser desenhadas próximas a cada sítio de reconhecimento de enzima de restrição primária sem restrição do sítio de reconhecimento de enzima de restrição secundária mais próximo. Nesta configuração, todas as sondas estariam próximas aos sítios de reconhecimento de enzima de restrição primária (preferivelmente em 300 bp do sítio de restrição).

Vantajosamente, o último desenho e também o desenho que usa 1 sonda por sítio de reconhecimento de enzima de restrição primária (ao lado de um) permitem o uso de enzimas de restrição secundárias diferentes em combinação com uma dada enzima de restrição primária.

Vantajosamente, o uso de múltiplas (por exemplo, 2, 3, 4, 5, 6, 7 ou 8 ou mais) sondas por sítio de reconhecimento de enzima de restrição primária pode minimizar o problema de obter resultados falso negativos, em virtude do pouco desempenho das sondas individuais. Além disso, elas também podem aumentar a confiabilidade de dados obtidos com um experimento de chipe único e reduzir o número de arranjos exigidos para tirar conclusões estatisticamente concretas.

As sondas para uso no arranjo podem ter mais que 40 nucleotídeos de comprimento e podem ser isotérmicas.

Preferivelmente, as sondas contendo sequências repetitivas de DNA são excluídas.

Espera-se que o diagnóstico de sondas para os sítios de restrição que flanqueiam diretamente ou estão próximos da primeira sequência de nucleotídeos forneça sinais de hibridização muito fortes e também possa ser excluído do desenho da sonda.

O arranjo pode cobrir qualquer genoma, incluindo genomas de mamífero (por exemplo, humano, camundongo (por exemplo, cromossomo 7)), vertebrado (por exemplo, peixe zebra)), ou invertebrado (por exemplo, bacteriano, levedura, fúngico ou de inseto (por

exemplo, *Drosofila*)).

Em uma modalidade adicional preferida, o arranjo contém 2-6 sondas ao redor de cada único sítio de restrição primária e tão próximo quanto possível do sítio de digestão de enzima de restrição.

5 Preferivelmente, a distância máxima do sítio de digestão de enzima de restrição é cerca de 300 bp.

Em uma modalidade adicional preferida da presente invenção, são fornecidos arranjos para enzimas de restrição, tais como *HindIII*, *BamHI*, *BgIII* e *NotI*, que cobrem os genomas de mamífero ou não mamífero. Vantajosamente, o desenho dos arranjos aqui descritos evitam a necessidade de redesenhar arranjos para cada sequência alvo, a análise fornecida é realizada na mesma espécie.

#### Conjuntos de sondas

Da maneira aqui usada, o termo "conjunto de sondas" refere-se a uma suíte ou uma coleção de sondas que hibridizam em cada um dos sítios de reconhecimento de enzima de restrição primária para uma enzima de restrição primária em um genoma.

Dessa maneira, é fornecido em um aspecto adicional, um conjunto de sondas complementares em sequência à sequência de ácido nucléico adjacente em cada um dos sítios de reconhecimento de enzima de restrição primária para uma enzima de restrição primária no DNA genômico.

20 Adequadamente, o conjunto de sondas é complementar em sequência aos primeiros 25-60 (por exemplo, 35-60, 45-60, ou 50-60) ou mais nucleotídeos que são adjacentes a cada um dos sítios de reconhecimento de enzima de restrição primária no DNA genômico. O conjunto de sondas pode ser complementar em sequência a um (por exemplo, qualquer um) lado ou ambos ou lados do sítio de reconhecimento de enzima de restrição primária. Dessa maneira, as sondas podem ser complementares em sequência à sequência de ácido nucléico adjacente a cada lado de cada um dos sítios de reconhecimento de enzima de restrição primária no DNA genômico.

Também é possível definir uma janela (por exemplo, 300 bp ou menos, tais como 250 bp, 200 bp, 150 bp ou 100 bp, do sítio de reconhecimento de enzima de restrição primária) na qual uma ou mais sondas podem ser desenhadas para o conjunto. Tais fatores que são importantes na definição da janela na qual o desenho das sondas são, por exemplo, conteúdo de GC, ausência de sequências palindrômicas que podem formar estruturas tipo alça, tamanho máximo dos trechos de um único tipo de nucleotídeo. Dessa maneira, o conjunto de sondas pode ser complementar em sequência à sequência de ácido nucléico que é menos que 300 bp de cada um dos sítios de reconhecimento de enzima de restrição primária no DNA genômico.

Também é possível definir uma janela de cerca de 100 bp do sítio de reconheci-

mento de enzima de restrição primária afim de identificar sondas ideais próximas de cada sítio de restrição.

Em modalidades adicionais da presente invenção, o conjunto de sondas é complementar à sequência que é menos que 300 bp de cada um dos sítios de reconhecimento de enzima de restrição primária no DNA genômico, complementar à sequência que está entre 200 e 300 bp de cada um dos sítios de reconhecimento de enzima de restrição primária no DNA genômico e/ou complementar à sequência que está entre 100 e 200 bp de cada um dos sítios de reconhecimento de enzima de restrição primária no DNA genômico.

Em modalidades adicionais da presente invenção, o conjunto de sondas é complementar à sequência que está de 0 a 300 bp de cada um dos sítios de reconhecimento de enzima de restrição primária no DNA genômico, complementar à sequência que está entre 0 a 200 bp de cada um dos sítios de reconhecimento de enzima de restrição primária no DNA genômico e/ou complementar à sequência que está entre 0 to 100 bp de cada um dos sítios de reconhecimento de enzima de restrição primária no DNA genômico (por exemplo, cerca de 10, 20, 30, 40, 50, 60, 70, 80 ou 90 bp de cada um dos sítios de reconhecimento de enzima de restrição primária no DNA genômico).

Podem ainda ser desenhadas duas ou mais sondas que são capazes de hibridizar na sequência adjacente a cada sítio de reconhecimento de enzima de restrição primária no DNA genômico.

As sondas podem sobrepor-se, ou sobrepõem-se parcialmente. Se as sondas sobrepõem-se, então a sobreposição é preferivelmente menos que 10 nucleotídeos.

Fragments de PCR que representam os primeiros 1-300 nucleotídeos (por exemplo, 1-20, 1-40, 1-60, 1-80, 1- 100, 1-120, 1-140, 1-160, 1-180, 1-200, 1-220, 1-240, 1-260 ou 1-280 nucleotídeos) que flanqueiam cada sítio de reconhecimento de enzima de restrição primária também podem ser usados.

Fragments de PCR também podem ser usados como sondas que correspondem exatamente a cada sítio genômico que é flanqueado pelo sítio de reconhecimento de enzima de restrição primária e o primeiro sítio de reconhecimento de enzima de restrição vizinho do segundo. Dessa maneira, a sequência de sonda pode corresponder a toda a sequência ou parte dela entre cada um dos sítios de reconhecimento de enzima de restrição primária e cada um dos primeiros sítios de reconhecimento de enzima de restrição secundária vizinhos.

Tipicamente, as sondas, arranjo de sondas ou conjunto de sondas serão imobilizados em um suporte. Os suportes (por exemplo, suportes sólidos) podem ser feitos de uma variedade de materiais, tais como vidro, sílica, plástico, náilon ou nitrocelulose. Os suportes são preferivelmente rígidos e têm uma superfície plana. Os suportes têm tipicamente cerca de 1-10.000.000 regiões endereçáveis espacialmente discretas, ou células. Suportes com

cerca de 10-1.000.000 ou cerca de 100-100.000 ou cerca de 1.000-100.000 células são comuns. A densidade de células é tipicamente pelo menos cerca de 1.000, 10.000, 100.000 ou 1.000.000 células em um centímetro quadrado. Em alguns suportes, todas as células são ocupadas por misturas agrupadas de sondas ou um conjunto de sondas. Em outros suportes, 5 algumas células são ocupadas por misturas agrupadas de sondas ou um conjunto de sondas, e outras células são ocupadas, pelo menos no grau de pureza obtinível por métodos de síntese, por um único tipo de oligonucleotídeo.

Preferivelmente, o arranjo aqui descrito compreende mais que uma sonda por sítio de reconhecimento de enzima de restrição primária que, no caso de uma enzima de restrição que corta de 6 bp, ocorre, por exemplo, aproximadamente 750.000 vezes por genoma humano ou de camundongo.

10 Para uma enzima de restrição que reconhece uma sequência de reconhecimento > 6 bp, um arranjo único de cerca de  $2 \times 750.000$  sondas pode ser usado para cobrir, por exemplo, o genoma humano ou de camundongo completo, com 1 sonda em cada lado de 15 cada sítio de restrição.

Em um desenho do arranjo preferido, o número total de moléculas de sonda de uma dada sequência de nucleotídeos presente no arranjo está em grande excesso nos fragmentos homólogos presentes na amostra 4C a ser hibridizada em tal arranjo. Dada a 20 natureza da tecnologia 4C, os fragmentos que representam as regiões genômicas próximas à sequência de nucleotídeos analisada no molde de cromatina linear estará em grande excesso na amostra de hibridização 4C (descrita na figura 2). Para obter informação quantitativa sobre eficiências de hibridização de tais fragmentos abundantes, pode ser necessário reduzir a quantidade de amostra a ser hibridizada e/ou aumentar o número de moléculas de uma dada sonda de sequência de oligonucleotídeos no arranjo.

25 Assim, para a detecção de elementos de DNA regulatórios que colocam freqüentemente em contato, por exemplo, um elemento promotor de gene, pode ser necessário usar um arranjo com sondas que representam apenas a região genômica selecionada (por exemplo, cerca de 0,5-10 Mb), mas com cada sonda exclusiva presente em múltiplas (por exemplo, cerca de 100, 200, 1.000) posições no arranjo. Tais desenhos também podem ser 30 preferidos com propósitos diagnósticos de detectar rearranjos genômicos locais (por exemplo, em cerca de 10 Mb), tais como deleções, inversões, duplicações, etc., ao redor de um sítio (por exemplo, gene de interesse).

O arranjo pode compreender cerca de  $3 \times 750.000$  sondas,  $4 \times 750.000$  sondas,  $5 \times 750.000$  sondas, ou preferivelmente,  $6 \times 750.000$  sondas. Mais preferivelmente, o arranjo 35 compreende  $6 \times 750.000$  sondas com 2, 3, 4, 5, 6, 7 ou 8 ou mais sondas em cada lado de cada sítio de restrição. Acima de tudo preferivelmente, o arranjo compreende  $6 \times 750.000$  sondas com 3 sondas em cada lado de cada sítio de restrição.

Arranjos de sondas ou conjuntos de sondas podem ser sintetizado de uma maneira etapa por etapa em um suporte ou podem ser anexados na forma pré-sintetizada. Um método de síntese é VLSIPS.TM. (descrito em US 5.143.854 e EP 476,014), que implica o uso da luz para direcionar a síntese de sondas de oligonucleotídeo em arranjos de alta densidade miniaturizados. Algoritmos para desenho de máscaras para reduzir o número de ciclos de síntese são descritos em US 5.571.639 e US. 5.593.839. Os arranjos também podem ser sintetizados em uma maneira combinatória liberando monômeros para as células de um suporte por caminhos de fluxo limitados mecanicamente, descritos em EP 624,059. Os arranjos também pode ser sintetizados apontando reagentes em um suporte usando uma impressora jato de tinta (ver, por exemplo, EP 728.520).

No contexto da presente invenção, os termos "substancialmente um conjunto de sondas" e "substancialmente o arranjo de sondas" significam que o conjunto ou o arranjo de sondas compreende pelo menos cerca de 50, 60, 70, 80, 90, 95, 96, 97, 98 ou 99% do conjunto ou arranjo total ou completo de sondas. Preferivelmente, o conjunto ou o arranjo de sondas é um conjunto de sondas total ou completo (isto é, 100%).

Em uma modalidade preferida, o arranjo compreende uma sonda exclusiva única por lado de cada sítio de reconhecimento de enzima de restrição primária que está presente em um dado genoma. Se este número de sondas excede o número de sondas que pode estar contido em um arranjo único, o arranjo pode preferivelmente conter ainda uma representação do genoma completo de uma dada espécie, mas em resolução menor, por exemplo, com uma de cada 2, 3, 4, 5, 6, 7, 8, 9, 10,  $10^2$ ,  $10^3$ ,  $10^4$  etc. sondas da maneira ordenada no molde do cromossomo linear presente no arranjo. Tais arranjos que cobrem o genoma completo humano, ou outro, em resolução subideal podem ser preferidos com relação a arranjos de alta resolução que cobrem parte do mesmo genoma, por exemplo, em casos onde observa-se parceiros de translocação.

Preferivelmente, a representação do genoma completo de uma dada espécie menor resolução é obtida por sondas no arranjo em que cada qual representa um único fragmento de restrição obtido após a digestão com uma enzima de restrição primária. Preferivelmente, isto é obtido desconsiderando cada segunda, terceira, quarta, quinta, sexta, sétima, oitava, nona, décima, vigésima, trigésima, quadragésima, quinquagésima, sexagésima, septuagésima, octogésima, nonagésima, ou uma centésima sonda, etc. (por exemplo, segunda à centésima sonda) que hibridiza no mesmo fragmento de restrição.

Preferivelmente, a representação do genoma completo de uma dada espécie em menor resolução compreende sondas que são distribuídas igualmente ao longo dos moldes lineares do cromossomo. Preferivelmente, isto é obtido desconsiderando uma ou mais sondas nestas regiões genômicas que mostram maior densidade de sonda.

#### Hibridização

A palavra "hibridização" da maneira aqui usada, pode incluir "o processo pelo qual uma fita de ácido nucléico se liga a uma fita complementar através do pareamento de bases", bem como o processo de amplificação da maneira realizada, por exemplo, em tecnologias da reação em cadeia da polimerase (PCR).

5 Sequências de nucleotídeos capazes de hibridização seletiva serão em geral pelo menos 75%, preferivelmente pelo menos 85 ou 90% e mais preferivelmente pelo menos 95% ou 98% homólogas à sequência de nucleotídeos complementar correspondente com relação a uma região de pelo menos 20, preferivelmente pelo menos 25 ou 30, por exemplo, pelo menos 40, 60 ou 100 ou mais nucleotídeos contíguos.

10 "Hibridização específica" refere-se à ligação, duplexação ou hibridização de uma molécula apenas em um sequência particular de nucleotídeos em condições rigorosas (por exemplo, 65 °C e 0,1xSSC {1xSSC = NaCl 0,15 M, citrato de Na 0,015 M pH 7,0}). Condições rigorosas são condições em que uma sonda hibridizará em sua sequência alvo, mas em nenhuma das outras sequências. Condições rigorosas são dependentes de sequência e 15 são diferentes em circunstâncias diferentes. Sequências maiores hibridizam especificamente em temperaturas maiores. Em geral, condições rigorosas são selecionadas entre cerca de 5 °C abaixo do ponto de fusão (Tm) para a sequência específica em uma concentração iônica e pH definidos. O Tm é a temperatura (em concentração iônica, pH, e concentração de ácido nucléico definidos) na qual 50% das sondas complementares a uma sequência alvo hibridizam na sequência alvo em equilíbrio. (Uma vez que as sequências alvos estão em geral presentes em excesso, em Tm, 50% das sondas são ocupadas em equilíbrio). Tipicamente, 20 condições rigorosas incluem uma concentração de sal de pelo menos cerca de concentração iônica de Na 0,01 a 1,0 M (ou outros sais) em pH 7,0 a 8,3 e a temperatura é pelo menos cerca de 30 °C para sondas curtas. Condições rigorosas também podem ser atingidas 25 com a adição de agentes desestabilizantes, tais como formamida ou sais de amônio de tetra-alquila.

30 Da maneira a ser entendida pelos versados na técnica, uma hibridização de rigor máxima pode ser usada para identificar ou detectar sequências de nucleotídeos idênticas enquanto uma hibridização de rigor intermediário (ou baixa) pode ser usada para identificar ou detectar sequências de polinucleotídeos similares ou relacionados.

Também são descritos métodos para a hibridização de arranjos de sondas em sequências de nucleotídeos marcadas ou não marcadas. As condições de reação de hibridização particulares podem ser controladas para alterar hibridização (por exemplo, aumento ou diminuição do rigor de ligação de sonda/alvo). Por exemplo, temperatura da reação, concentrações de ânions e cátions, adição de detergentes e similares, podem todos alterar as características de hibridização de sondas de arranjo e moléculas alvos.

#### Frequência de interação

A quantificação de frequências de ligação de fragmentos de restrição fornece uma medida de suas frequências de reticulação. Adequadamente, isto pode ser atingido usando PCR da maneira usada na tecnologia 3C convencional descrita por Splinter et al. (2004) (supra). Resumidamente, a formação de produtos de PCR pode ser medida varrendo as intensidades de sinal após a separação em géis de agarose corados com brometo de etídio, usando um imageador Typhoon 9200 (Molecular Dynamics, Sunnyvale, CA). Adequadamente, diversos controles são usados para a interpretação correta de dados também descritos em Splinter et al. (2004) (supra).

Uma vez que, a tecnologia 4C aqui descrita proporciona a análise de alto desempenho da frequência de interação de duas ou mais sequências de nucleotídeos no espaço nuclear, é preferível que as frequências de ligação de fragmentos de restrição sejam quantificadas usando os arranjos aqui descritos.

Para quantificação, sinais obtidos de uma amostra 4C podem ser normalizados em sinais obtidos para uma amostra controle. A(s) amostra(s) 4C e amostra(s) controle(s) serão marcadas com marcadores diferentes e discerníveis (por exemplo, corantes) e serão simultaneamente hibridizadas no arranjo. A(s) amostra(s) controle(s) conterão tipicamente todos os fragmentos de DNA (isto é, todas as segundas sequências de nucleotídeos potenciais que foram ligadas na primeira sequência de nucleotídeos (alvo)) em quantidades equimolares e, para excluir uma tendência na eficiência da hibridização, elas devem ser similares em tamanho à(s) segunda(s) sequência(s) de nucleotídeos(s). Assim, o molde controle conterá tipicamente DNA genômico (do mesmo fundo genético, uma vez que é usado para obter o molde 4C), digerido tanto com a enzima de restrição primária quanto a secundária e marcado pelo mesmo método (por exemplo, iniciação aleatória) como o molde 4C. Tal molde controle torna possível corrigir as diferenças sonda a sonda na eficiência da hibridização. A normalização de sinais de arranjo 4C para controlar sinais de arranjo torna possível expressar resultados em termos de enriquecimento com relação a eventos aleatórios.

O molde marcado 4C pode ainda ser hibridizada em um arranjo com ou sem uma amostra controle diferencialmente marcada e com ou sem uma ou mais outros moldes 4C diferencialmente marcados. Outros moldes 4C podem não estar relacionados a este molde 4C, por exemplo, eles podem ser obtidos de tecido diferente e/ou obtidos com um conjunto diferente de iniciadores de PCR inversa. Por exemplo, o primeiro molde 4C pode ser material de paciente e o segundo molde 4C pode ser obtido de um sujeito saudável ou de uma amostra controle.

Dados os padrões de hibridização notáveis que são esperados para rearranjos genéticos, nem sempre será necessário comparar sujeitos doentes com sujeitos saudáveis. Dessa maneira, múltiplos moldes 4C (por exemplo, dois ou mais), cada qual interrogando um locus diferente do mesmo paciente ou sujeito, podem ser hibridizados em um arranjo

(por exemplo, um ou mais).

Os moldes 4C podem ser diferencialmente marcados (por exemplo, com hibridização de duas ou multicores) e/ou podem ser identicamente marcados no caso em que tais loci residem normalmente em cromossomos diferentes ou no mesmo cromossomo em uma 5 distância longe o suficiente para sobreposição mínima entre sinais de interação DNA-DNA. Como um exemplo, material de um sujeito com leucemia de célula T pode ser processado para obter moldes 4C para TCRA/  $\delta$  (marcado em uma cor, a fim de detectar translocações), e MLL, TALI, HOXII e LMO2 (cada qual marcado na mesma segunda cor, a fim de detectar 10 outros rearranjos genéticos). Estes cinco moldes 4C podem ser hibridizados em um arranjo, que permitirá a análise simultânea em múltiplos loci para um rearranjo genômico associado 15 à doença.

Para a quantificação de frequências de interação, intensidades de sinal absolutas ou razões com relação à amostra controle também podem ser considerados. Além do mais, 20 sinais de sondas adjacentes no molde do cromossomo linear podem ser usados para identificar 15 regiões cromossômicas interagentes. Tal informação posicional é analisada preferivelmente ordenando as sondas no molde do cromossomo linear e analisando as intensidades de sinal absolutas ou razões com relação a sinais de molde controle, por abordagens de janela corrediça, usando, por exemplo, média corrida ou abordagens de média corrida.

A frequência de interação de uma ou mais sequências de nucleotídeos alvo com 25 uma ou mais sequências de nucleotídeos de interesse (por exemplo, um ou mais loci genômicos) obtida de acordo com os métodos aqui descritos, pode ser usada para reconstruir partes, ou na íntegra, mapas cromossômicos lineares e identificar rearranjos genômicos equilibrados e desequilibrados que ocorreram nos cromossomos e entre os mesmos, em que tais rearranjos são indicativos de um traço ou doença.

#### 25 Método de ensaio

Em um aspecto adicional da presente invenção, é fornecido um método de ensaio para identificar um ou mais agentes que modulam uma interação DNA-DNA.

Da maneira aqui usada, a palavra "modular" refere-se a prevenir, diminuir, suprimir, restaurar, elevar, aumentar ou de outra forma afetar a interação DNA-DNA.

30 Em alguns casos, pode ser desejável avaliar dois ou mais agentes juntos para uso na modulação da interação DNA-DNA. Nestes casos, os ensaios podem ser facilmente modificados adicionando tal agente(s) adicional(s) tanto simultaneamente quanto subsequentemente ao primeiro agente.

O método da presente invenção também pode ser uma triagem, por meio da qual 35 inúmeros agentes são testados para modular a atividade da interação DNA-DNA.

Espera-se que os métodos de ensaio da presente invenção sejam adequados tanto para triagem em pequena escala quanto em grande escala de agentes, bem como em en-

saios quantitativos.

Usos médicos de tais agentes terapêuticos estão no escopo da presente invenção bem como os próprios programas de desenvolvimento de medicamento e composições farmacêuticas compreendendo tais agentes. Um programa de desenvolvimento de medicamento, por exemplo, pode envolver preparar um agente identificado ou identificável pelos métodos aqui descritos, modificando-o opcionalmente (por exemplo, modificando sua estrutura e/ou fornecendo uma composição inédita compreendendo a dita fração) e realizando estudos adicionais (por exemplo, estudos de toxicidade e /ou estudos na atividade, estrutura ou função). Experiências podem ser realizadas em animais não humanos e podem eventualmente ser realizadas em humanos. Tais experiências incluirão em geral determinar o(s) efeito(s) de níveis de dosagem diferentes. Os programas de desenvolvimento de medicamento podem utilizar computadores para analisar frações identificadas trinado (por exemplo, para prognosticar estrutura e/ou função, para identificar possíveis agonistas ou antagonistas, para pesquisar frações que podem ter estruturas ou funções similares, etc.).

#### TESTE DE DIAGNÓSTICO

Atualmente, vários rearranjos genômicos permanecem difíceis de detectar pelas técnicas molecular-citogênicas disponíveis. Embora a técnica de hibridização genômica comparativa de arranjo (arranjo-CGH) seja uma técnica recém desenvolvida para a detecção de amplificação cromossômica e/ou deleções com uma resolução de 35 a 300 Kb, esta técnica não é adequada para detectar translocações equilibradas e inversões cromossômicas. Por outro lado, cariotipagem espectral (SKY) ou cariotipagem convencional é frequentemente realizada no material do paciente para a detecção de translocações cromossômicas bem como mudanças numéricas, mas a resolução para definir translocação dos pontos de quebra é baixa, normalmente 10 a 50 Mb e 5 a 10 Mb, respectivamente. Consequentemente, resultados obtidos por ambos métodos e especialmente SKY levarão a experimentos de validações demorados e de muita mão de obra do tipo hibridização *in situ* de fluorescência (FISH) e estratégias de clonagem de ponto de quebra molecular.

Tecnologia 4C envolve um procedimento que pode detectar quaisquer rearranjos cromossômicos com base nas frequências de interação mudadas entre as sequências de DNA fisicamente ligadas. Tecnologia 4C é portanto usada para a identificação de (recorrente) rearranjos cromossômicos para a maioria de malignidade/má formações congênitas múltiplas humanas ou retardamento mental. Uma vantagem importante de tecnologia 4C é que ela permite o mapeamento muito preciso do ponto de quebra de uma região de apenas diversos milhares de pares de base. Uma outra vantagem de tecnologia 4C é que não é exigido nenhum conhecimento anterior na posição exata do ponto de quebra, uma vez que os pontos de quebra serão detectáveis mesmo quando a sequência 4C-isca for localizada 1 a 5 Mb distante do ponto de quebra. Isto tem também a vantagem de que a mesma sequência

isca pode ser usada para a detecção de rearranjos cromossômicos específicos cobrindo áreas de ponto de quebra grandes. O mapeamento preciso de rearranjos genômicos pela tecnologia 4C facilitarão enormemente a identificação de doenças de base de gene(s) expresso(s) aberrantemente ou distúrbios genéticos, que contribuirão de forma muito importante para um melhor entendimento das correlações genótipo-fenótipo, ajudarão na tomada de decisão do tratamento e adicionarão importante informação de prognóstico.

Em uma modalidade da presente invenção, a fim de fornecer uma base para o diagnóstico ou prognóstico de doença, são estabelecidos valores normais ou padrões de um sujeito. Isto pode ser realizado testando amostras tiradas de sujeitos normais - tais como animais ou humanos. A frequência da interação DNA-DNA pode ser quantificada comparando-a com uma série de diluição de controles positivos. Então, valores padrões obtidos de amostras normais podem ser comparados com valores obtidos de amostras de sujeitos afetados ou potencialmente afetados por uma doença ou um distúrbio. Desvio entre valores padrões e do sujeito estabelece a presença da condição da doença.

Tais ensaios de diagnóstico podem ser talhados para avaliar a eficácia de um regime de tratamento terapêutico particular e podem ser usados em estudos de animal, em experiências clínicas, ou no monitoramento do tratamento de um paciente individual. A fim de fornecer uma base para o diagnóstico de doença, um perfil normal ou padrão para a interação DNA-DNA pode ser estabelecido. Os valores padrões obtidos de amostras normais podem ser comparados com os valores obtidos de amostras de sujeitos potencialmente afetados por um distúrbio ou doença. O desvio entre valores padrões e o sujeito estabelece a presença da condição da doença. Se a doença for estabelecida, um agente terapêutico existente pode ser administrado, e perfil ou valores do tratamento podem ser gerados. Finalmente, o método pode ser repetido em uma base regular para avaliar se os valores progredem a favor ou contra para o modelo normal ou padrão. Perfis de tratamento sucessivo podem ser usados para mostrar a eficácia de tratamento durante um período de diversos dias ou diversos meses.

Tecnologia 4C detecta precisamente pelo menos 5 Mb de DNA genômico ligado em cis com a sequência de nucleotídeos que é analisada (ver Figura 2-3 e 5). Vantajosamente, a tecnologia 4C pode ser usada para detectar qualquer aberração genômica que é acompanhada por uma mudança na separação do sítio genômico entre sequências rearranjadas e uma sequência 4C (isca) de escolha. Tal mudança pode ser, por exemplo, um aumento ou diminuição na separação do sítio genômico ou pode ser uma subrepresentação (como em deleções) ou sobrerepresentação (como em duplicações) de sequências proximais (por exemplo, até ou maior que 15 Mb) na sequência 4C (isca). Tipicamente, tais aberrações ou rearranjos genômicos são uma causa ou são associadas com doenças - tais como câncer (por exemplo, leucemia) e outras doenças genéricas ou congênitas da maneira aqui descri-

ta.

Aberrações genéticas (por exemplo, aberrações genômicas ou cromossômicas - tais como aberrações genômicas ou cromossômicas equilibradas e/ou ou não equilibradas) incluem, mas sem limitações, rearranjos, translocações, inversões, inserções, deleções e 5 outras mutações de ácido nucléico (por exemplo, cromossomos) e também perdas ou ganhos de parte ou todos os cromossomos. Eles são uma causa que leva a distúrbios genéticos ou doenças, incluindo distúrbios congênitos e doenças adquiridas - tais como malignidades. Em quaisquer rearranjos, dois cromossomos diferentes são envolvidos. Desta maneira, genes (ou fragmentos de genes) são removidos do contexto fisiológico normal de um 10 cromossomo particular e são localizados em um cromossomo do recipiente, adjacente a genes não relacionados ou fragmentos de genes (frequentemente oncogenes ou proto-oncogenes).

Malignidades podem incluir leucemias agudas, linfomas malignos e tumores sólidos. Exemplos não limitantes de alterações são t(14;18) que ocorre frequentemente em 15 NHL; t(12;21) que é frequentemente encontrado em precursor-B-ALL infantil; e a presença de llq23 (MLL (leucemia mielóide-linfóide ou leucemia de linhagem misturada) gene) aberrações em leucemias agudas.

O gene MLL na região do cromossomo llq23 é envolvido em diversas translocações tanto em ALL quanto em leucemias mielóides agudas (AML). Até hoje, pelo menos dez 20 genes parceiros foram identificados. Algumas dessas translocações, - tais como t(4;11) (q21;q23), t(11;19) (q23;p13) e t(1;11) (p32;q23), ocorrem predominantemente em ALL, onde como outros, do tipo t(1;11) (q21;q23), t(2;11) (p21;q23), t(6;11) (q27;q23) e t(9;11) (p22;q23) são mais frequentemente observados em AML. Rearranjos que envolvem a região llq23 ocorrem muito frequentemente em leucemias agudas infantis (em torno de 60 a 70 %), 25 e com uma extensão muito menor em leucemias infantis e de adultos (cada qual em torno de 5 %).

Rearranjos em malignidades linfóides envolvem frequentemente genes Ig ou TCR. Exemplos incluem os três tipos de translocações (t(8;14), t(2;8), e t(8;22)) que são encontrados em linfomas de Burkitt, em que o gene MYC é acoplado em segmentos de gene Ig de 30 cadeia pesada Ig (IGH), kappa Ig (IGK), ou lambda Ig (IGL), respectivamente. Um outro tipo comum de translocação nesta categoria é t(14;18) (q32;q21) que são observados em cerca de 90 % de linfomas foliculares, um dos principais tipos de NHL. Nesta translocação, o gene BCL2 é rearranjado nas regiões no locus IGH dentro ou adjacente aos segmentos do gene JH. O resultado desta aberração do cromossomo é o sobre-expresão da proteína BCL2, 35 que exerce um papel como um fator de sobrevivência no controle de crescimento inibindo a morte celular programada.

O gene BCL2 consiste em três exons, mas estes são espalhados em uma grande

área. Desses, o último exon codifica uma grande região 3' não traduzida (3<sup>1</sup> UTR). Este UTR 3' é uma das duas regiões nas quais muitos pontos de quebra t(14;18) são agrupados e é chamado a "região de ponto de quebra principal"; a outra região de ponto de quebra envolvida em translocações t(14;18), é localizada 20 a 30 kb à jusante do locus BCL2 e é chamada a "região de agrupamento secundário". Uma terceira área de ponto de quebra de BCL2, o VCR (região de agrupamento variante), é localizada no lado 5' do locus BCL2 e está entre outros envolvidos em translocações variantes, isto é, t(2;18) e t(18;22), na qual os segmentos IGK e IGL do gene são os genes parceiros.

Assim, a título de exemplo, a tecnologia 4C pode ser aplicada para a triagem de material do paciente para aberrações genéticas próximas ou em loci que foram escolhidos com base na sua frequente associação com um dado fenótipo clínico. Exemplos não limitantes adicionais de tal loci são AML1, MLL, MYC, BCL, BCR, ABL1, imunoglobulina loci, LYL1, TAL1, TAL2, LMO2, TCR $\alpha/\delta$ , TCR $\beta$ , HOX e outro loci em várias leucemias linfoblásticas.

Vantajosamente, se suspeita-se de uma aberração genética, a tecnologia 4C pode ser aplicada como a primeira e única triagem para verificar e mapear a presença da aberração da maneira aqui explicada.

#### *Detecção de rearranjos genômicos*

Em uma modalidade particularmente preferida da presente invenção, os métodos aqui descritos podem ser usados para a detecção de rearranjos genômicos.

Atualmente, rearranjos genômicos - tais como pontos de quebra de translocação - são muito difíceis para detectar. Por exemplo, hibridização genômica comparativa (CGH) micro-arranjos pode detectar diversos tipos de rearranjos mas falta detectar translocações. Se a translocação é suspeita em um paciente mas cromossomos parceiros são desconhecidos, cariotipagem espectral (SKY) pode ser realizada para encontrar parceiros de translocação e obter uma estimativa aproximada de localização de ponto de quebra. Entretanto, a resolução é muito fraca (normalmente não superior a ~ 50 Mb) e mapeamento fino adicional (que tanto é demorado quanto caro) é normalmente exigido. Isto é normalmente feito usando Hibridização *in situ* de fluorescência (FISH), que novamente fornece resolução limitada. Usando FISH, pontos de quebra podem ser localizados na região +/-50 kb em resolução máxima.

Frequências de interação de DNA-DNA são basicamente uma função da separação do sítio genômico, isto é, frequências de interação de DNA-DNA são inversamente proporcionais à distância linear (em kilobases) entre dois loci de DNA presentes no mesmo molde de DNA físico (Dekker *et al.*, 2002). Assim, uma translocação, que cria um ou mais moldes de DNA físico inéditos, é acompanhada por interações DNA-DNA alteradas próximas aos pontos de quebra, e isto pode ser medido pela tecnologia 4C. Doenças com base em translocações são tipicamente causadas por interações DNA-DNA aberrantes, as translocação é

o resultado da ligação física (interação) de braços do cromossomo quebrados (DNA).

Consequentemente, para a detecção de translocações, a tecnologia 4C pode ser usada para identificar aquelas interações DNA-DNA que são diferentes entre os sujeitos doentes e não doentes.

5 A título de exemplo, a tecnologia 4C pode ser aplicada para a triagem de material do paciente para loci próximo a translocações que foram escolhidos com base em sua frequente associação com um dado fenótipo clínico da maneira aqui descrita.

Se houver suspeita de translocação em um paciente, mas os cromossomos parceiros forem desconhecidos, um mapeamento inicial pode ser realizado usando métodos atualmente disponíveis similares a cariotipagem espectral (SKY). Isto pode identificar os parceiros de translocação e fornecer uma estimativa muito grosseira de localização de ponto de quebra (normalmente resolução não superior a ~ 50 Mb). Tecnologia 4C pode então ser aplicada, usando sequências 'isca' nesta região localizada por exemplo, em cada 2 Mb, 5 Mb, 10 Mb, 20 Mb (ou outros intervalos da maneira aqui descrita) para mapear bem o ponto 10 de quebra e identificar, por exemplo, o(s) gene(s) que deixam de ser expressos em decorrência da translocação.

Tipicamente, uma translocação será identificada por meio de uma transição abrupta de frequências baixas para altas de interação em um cromossomo sem ser aquela contendo a sequência 4C-isca, ou em outro lugar naquele mesmo cromossomo.

20 Em uma modalidade preferida, a amostra do sujeito é em um estado pré-maligno.

Em uma modalidade preferida, a amostra do sujeito consiste em amniócitos cultivados ou não cultivados obtidos por aminocentese para diagnóstico pré-natal.

Em um desenho de arranjo preferido, sondas apresentadas em um arranjo simples representam o genoma completo de uma dada espécie em resolução máxima. Assim, arranjos para detectar translocações e similares por tecnologia 4C, contêm sondas da maneira aqui descrita complementares para cada lado de cada sítio de reconhecimento de enzima de restrição primário no genoma de uma dada espécie (por exemplo, humano).

Em um outro projeto preferido, sondas apresentadas em um arranjo simples representam o genoma completo de uma dada espécie, mas não em resolução máxima. Assim, arranjos para detectar translocações e similares pela tecnologia 4C contêm sondas da maneira aqui descrita que são complementares apenas para um lado de cada sítio de reconhecimento de enzima de restrição primário no genoma de uma dada espécie (por exemplo, humano).

35 Em um outro projeto preferido, sondas apresentadas em um arranjo simples representam o genoma completo de uma dada espécie, mas não em resolução máxima. Assim, arranjos para detectar translocações, deleções, inversões, duplicações e outros rearranjos genômicos pela tecnologia 4C contêm sondas da maneira aqui descrita que são comple-

mentares para um lado de cada outro sítio de reconhecimento de enzima de restrição primário da maneira ordenada ao longo do molde linear do genoma de uma dada espécie (por exemplo, humano).

Assim, arranjos para detectar translocações, deleções, inversões, duplicações e outros rearranjos genômicos pela tecnologia 4C contêm sondas da maneira aqui descrita, que cada qual representa um fragmento de restrição simples da maneira obtida após digestão com uma enzima de restrição primária. Preferivelmente, isto é obtido ignorando cada segunda, terceira, quarta, quinta, sexta, sétima, oitava, nona, décima, vigésima, trigésima, quadragésima, quinquagésima, sexagésima, septuagésima, octagésima, nonagésima, ou centésima, etc. sonda que hibridiza no mesmo fragmento de restrição. Arranjos para detectar translocações, deleções, inversões, duplicações e outros rearranjos genômicos pela tecnologia 4C podem conter sondas da maneira aqui descrita que são distribuídas igualmente ao longo dos moldes do cromossomo linear. Preferivelmente, isto é obtido ignorando uma ou mais sondas nestas regiões genômicas que apresenta densidade de sonda mais alta.

Em um outro projeto preferido, sondas apresentadas em um arranjo simples representam o genoma completo de uma dada espécie, mas não em resolução máxima. Assim, arranjos para detectar translocações, deleções, inversões, duplicações e outros rearranjos genômicos pela tecnologia 4C, contêm sondas da maneira aqui descrita complementares para um lado de cada terceiro, quarto, quinto, sexto, sétimo, oitavo, nono, décimo, vigésimo, trigésimo, quadragésimo, quinquagésimo, sexagésimo, septuagésimo, octagésimo, nonagésimo, ou centésimo, etc. sítio de reconhecimento de enzima de restrição primário da maneira ordenada ao longo do molde linear do genoma de uma dada espécie (por exemplo, humano). Arranjos para detectar translocações, deleções, inversões, duplicações e outros rearranjos genômicos pela tecnologia 4C podem conter sondas da maneira aqui descrita, que representam o genoma completo, mas com uma sonda simples a cada 100 kilobases. Arranjos para detectar translocações, deleções, inversões, duplicações e outros rearranjos genômicos pela tecnologia 4C podem conter sondas da maneira aqui descrita que representam cada sítio de reconhecimento de enzima de restrição primário simples no genoma que pode ser representado por uma sequência de sonda exclusiva.

Em um outro desenho de arranjo preferido, sabe-se que sondas da maneira aqui descrita em um arranjo simples representam regiões genômicas de um dado tamanho - tais como cerca de 50 kb, 100 kb, 200 kb, 300 kb, 400 kb, 500 kb, 1 Mb, 2 Mb, 3 Mb, 4 Mb, 5 Mb, 6 Mb, 7 Mb, 8 Mb, 9 Mb ou 10 Mb - (por exemplo, cerca de 50 kb- 10 Mb) em torno de todo loci estão envolvidas em translocações, deleções, inversões, duplicações e outros rearranjos genômicos.

Em um outro desenho de arranjo preferido, sabe-se que sondas da maneira aqui descrita em um arranjo simples representam regiões genômicas de um dado tamanho - tais

como cerca de 50 kb, 100 kb, 200 kb, 300 kb, 400 kb, 500 kb, 1 Mb, 2 Mb, 3 Mb, 4 Mb, 5 Mb, 6 Mb, 7 Mb, 8 Mb, 9 Mb ou 10 Mb - (por exemplo, cerca de 50 kb- 10 Mb) em torno de uma triagem de loci estão envolvidas em translocações, deleções, inversões, duplicações e outros rearranjos genômicos. Seleções podem ser feitas em critério educado, por exemplo, 5 elas podem representar apenas os loci que são implicados em um dado tipo de doença.

Em um outro desenho de arranjo preferido, sondas da maneira aqui descrita em um arranjo simples representam uma região genômica de interesse de, por exemplo, 100 kb, 200 kb, 300 kb, 400 kb, 500 kb, 600 kb, 700 kb, 800 kb, 900 kb, 1 Mb, 2 Mb, 3 Mb, 4 Mb, 5 Mb, 6 Mb, 7 Mb, 8 Mb, 9 Mb, 10 Mb, 20 Mb, 30 Mb, 40 Mb, 50 Mb, 60 Mb, 70 Mb, 80 Mb, 90 Mb, ou 100 Mb (por exemplo, 100 kb- 10 Mb) (parte de) um cromossomo ou múltiplos cromossomos, com cada sonda senda representada múltiplas vezes (por exemplo, 10, 100, 1000) para permitir medições quantitativas de intensidades de sinal de hibridização em cada sequência de sonda.

Em um projeto experimental preferido, ao sequência 4C (isca) é cerca de 0 kb, 10 kb, 20 kb, 30 kb, 40 kb, 50 kb, 100 kb, 200 kb, 300 kb, 400 kb, 500 kb, 1 Mb, 2 Mb, 3 Mb, 4 Mb, 5 Mb, 6 Mb, 7 Mb, 8 Mb, 9 Mb 10 Mb, HMb, 12Mb, 13Mb, 14Mb ou 15 Mb (por exemplo, cerca de 0 a 15 Mb) ou mais da sequência rearranjada real (isto é, ponto de quebra no caso de uma translocação).

Em uma hibridização preferida, dois moldes 4C marcados diferencialmente, obtidos com uma sequência (4C isca) de um sujeito doente e não doente são hibridizados simultaneamente no mesmo arranjo. Diferenças nas interações DNA-DNA permitem a detecção do ponto de quebra em cis (no mesmo cromossomo como o 4C-isca) e em trans (no parceiro de translocação).

Em uma hibridização preferida, múltiplos moldes 4C marcados diferencialmente, obtidos com uma sequência (4C isca) de sujeitos doentes e não doentes são hibridizados simultaneamente no mesmo arranjo. Diferenças nas interações DNA-DNA permite a detecção do ponto de quebra em cis (no mesmo cromossomo como o 4C-isca) e em trans (no parceiro de translocação).

Vantajosamente, análise de múltiplas cores, em vez de cores duplas em micro-arranjos pode ser utilizada permitindo a hibridização simultânea de mais que duas amostras em um arranjo simples. Consequentemente, hibridização de múltiplas cores pode ser usada na tecnologia 4C.

Em uma hibridização preferida, múltiplos moldes 4C marcados diferencialmente, obtidos com uma sequência (4C isca) de sujeitos doentes e um molde 4C marcado diferencialmente e um sujeito não doente são hibridizados simultaneamente no mesmo arranjo. Diferenças nas interações DNA-DNA permitem a detecção do ponto de quebra em cis (no mesmo cromossomo como o 4C-isca) e em trans (no parceiro de translocação).

Em uma outra hibridização preferida, dois moldes 4C marcados diferencialmente do mesmo sujeito não doente, obtidos com duas diferentes sequências (4C-iscas) que cada qual representa um outro parceiro de translocação possível, são hibridizados simultaneamente no mesmo arranjo. Agrupamentos de sinais de hibridização fortes observados no 5 molde linear de cromossomos não relacionados com o cromossomo carregam a sequência de interesse (4C-isca) identificarão o cromossomo parceiro de translocação e o ponto de quebra no parceiro de translocação.

Em uma outra hibridização preferida, múltiplos moldes 4C marcados diferencialmente do mesmo sujeito não doente, obtidos com múltiplas sequências diferentes (4C-iscas) 10 que cada qual representa um outro parceiro de translocação possível, são hibridizados simultaneamente no mesmo arranjo. Agrupamentos de sinais de hibridização fortes observados no molde linear de cromossomos não relacionados com o cromossomo carregam a sequência de interesse (4C-isca) identificarão o cromossomo parceiro de translocação e seu ponto de quebra para a sequência de interesse.

15 Material usado para a detecção de translocações, deleções, inversões, duplicações e outros rearranjos genômicos pela tecnologia 4C pode ser obtido por reticulação (e processamento adicional, da maneira descrita) de células vivas e/ou células mortas e/ou lisatos nucleares e/ou cromatina isolada etc. (da maneira aqui descrita) de sujeitos doentes e/ou não doentes.

20 *Detecção de inversões*

Inversões (por exemplo, inversões equilibradas) não podem ser detectadas por métodos - tais como Técnicas de hibridização genômica comparativas - mas pode ser detectadas pela tecnologia 4C particularmente quando a inversão (equilibrada) é próxima (por exemplo, até cerca de 1 a 15 Mb ou mais) da sequência 4C (isca).

25 Detecção de inversões (equilibradas) é baseada na identificação daquelas interações DNA-DNA que foram diferentes entre o sujeito doente e o não doentes. Inversões mudarão a posição relativa (em kilobases) no molde de DNA físico de todas as sequências (mas a maioria centralmente localizada) da região rearranjada medido contra uma sequência próxima do mesmo cromossomo que é considerada a sequência 4C (isca). Uma vez que 30 frequências de interação de DNA-DNA são inversamente relacionadas à separação do sítio genômico, sujeitos doentes darão padrões invertidos de intensidades de hibridização para todas as sondas localizadas na região genômica rearranjada, comparado com um sujeito não doente. Assim, a tecnologia 4C permite a identificação de posição e tamanho de inversões (equilibradas).

35 De acordo com este aspecto da presente invenção, um desenho de arranjo consagrado preferido compreende sondas em um arranjo simples representando regiões genômicas de um dado tamanho - tais como cerca de 50 kb, 100 kb, 200 kb, 300 kb, 400 kb, 500

kb, 1 Mb, 2 Mb, 3 Mb, 4 Mb, 5 Mb, 6 Mb, 7 Mb, 8 Mb, 9 Mb ou 10 Mb) (por exemplo, 50 kb- 10 Mb) em torno do locus no qual suspeita-se da inversão ou outro rearranjo.

Em um outro desenho de arranjo consagrado preferido, sondas em um arranjo simples representam regiões genômicas de um dado tamanho (50 kb, 100 kb, 200 kb, 300 kb, 5 400 kb, 500 kb, 1 Mb, 2 Mb, etc.) em torno do locus no qual suspeita-se da inversão ou outro rearranjo. Para análise quantitativa de confiança de intensidades de sinal, a quantidade de sonda presente no arranjo é tipicamente em grande excesso para a quantidade de cognatos de fragmentos que são hibridizados no arranjo. Portanto, pode ser necessário ter cada sonda presente múltiplas vezes (por exemplo, 10, 20, 50, 100, 1000 vezes, etc.) no arranjo. A 10 Além do mais, pode ser necessário titular a quantidade de molde que deve ser hibridizada no arranjo.

#### *Detecção de deleções*

Detecção de deleções é baseada em identificar aquelas interações DNA-DNA que foram diferentes entre os sujeitos doentes e não doentes. Deleções resultarão na ausência 15 de interações de DNA com uma sequência 4C (isca) localizada próximo a região deletada (por exemplo, cerca de 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14 ou 15 Mb ou mais). Isto pode resultar na ausência completa de sinais de hibridização para todas as sondas localizada 20 na região rearranjada se a deleção for presente em ambos alelos (homozigoto), ou uma redução para sujeitos doentes versus não doentes de intensidades de sinal se a deleção for presente em apenas um alelo (heterozigoto). Deleção resulta em mais sequências distais na proximidade imediata no molde de DNA físico para a sequência 4C (isca) analisada, que resultará em sinais de hibridização mais fortes para sondas localizadas diretamente além da 25 região deletada.

#### *Detecção de duplicação(s)*

Detecção de duplicação é tipicamente baseada em identificar aquelas interações 25 DNA-DNA que são diferentes entre sujeitos doentes e não doentes. As sondas na região duplicada mostrarão maiores sinais de hibridização com uma sequência 4C (isca) localizada próximo à região rearranjada (por exemplo, cerca de 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14 ou 15 Mb ou mais), comparado com os sinais de um sujeito não doente de controle. Sondas além da região duplicada são adicionalmente distantes da sequência 4C e consequentemente, apresentarão maiores sinais de hibridização comparados com os sinais de um sujeito não doente de controle.

Preferivelmente, um aumento ou uma diminuição na frequência de interação DNA-DNA para a amostra do sujeito comparado com o controle é indicativo de uma duplicação ou 35 inserção.

Preferivelmente, um aumento na frequência de interação DNA-DNA para a amostra do sujeito comparado com o controle e/ou uma redução na frequência de interação DNA-

DNA para regiões mais distantes é indicativo de uma duplicação ou inserção.

*Diagnóstico Pré-natal*

Vantajosamente, a tecnologia 4C pode também ser usada em diagnóstico pré-natal.

5 Ácido nucléico pode ser obtido de um feto usando vários métodos que são conhecidos na tecnologia. A título de exemplo, amniocentese pode ser usado para obter fluido amniótico do qual células fetais em suspensão são extraídas e cultivadas por diversos dias (Mercier & Bresson (1995) Ann. Gnt., 38, 151-157). Ácido nucléico das células pode ser então extraído. A coleta de chorial villi pode tornar possível dispensar a etapa de cultura e evitar a coleta de fluido amniótico. Estas técnicas podem ser aplicadas anteriormente (até 7 10 semanas de gestação para a coleta de chorial villi e 13 a 14 semanas para amniocentese), mas com um risco ligeiramente maior de aborto.

Uma coleta direta de sangue fetal no nível do cordão umbilical pode também ser usada para obter ácido nucléico, mas exige tipicamente uma equipe de clínicos especializados nesta técnica (Dormer *et al.* (1996) *Fetal Diagn. Ther.*, 10, 192-199).

Vantajosamente, as aberrações genéticas (por exemplo, aberrações genômicas ou cromossômicas) - tais como rearranjos, translocações, inversões, inserções, deleções e outras mutações em cromossomos e ácido nucléico - podem ser detectadas neste estágio.

20 Preferivelmente, as aberrações genéticas (por exemplo, aberrações genômicas ou cromossômicas) - tais como rearranjos, translocações, inversões, inserções, deleções e outras mutações em cromossomos 21, 18, 13, X ou Y e também perdas ou ganhos de parte ou todos cromossomos 21, 18, 13, X ou Y podem ser detectadas uma vez que estes são os cromossomos nos quais a maioria de aberrações ocorrem no feto.

*Determinação de sítios de integração genômica*

25 A tecnologia 4C também permite a determinação de sítios de integração genômica de vírus e transgenes, etc., também quando múltiplas cópias são inseridas em diferentes posições no genoma (da maneira descrita na Figura 3).

*Predisposição determinante de adquirir certas translocações*

30 Vantajosamente, a tecnologia 4C pode também ser aplicada a sujeitos não doentes para medir o ambiente genômico de loci frequentemente envolvido em aberrações genéticas. Desta maneira, é possível determinar a predisposição de o sujeito adquirir certas aberrações genéticas.

Assim, além dos usos médicos aqui descritos, a presente invenção pode ser usada em diagnóstico.

35 4C multiplex

A presente invenção permite a análise simultânea da frequência de interações de múltiplas sequências de nucleotídeos alvos com uma ou mais sequências de nucleotídeos

de interesse. Amplificação pode ser obtida usando multiplex PCR. Tais métodos permitem uma triagem não tendenciosa para rearranjos genômicos equilibrados e não equilibrados isto é, translocações, inversões, deleções e duplicações que podem ter ocorrido em qualquer lugar no genoma. Os métodos podem ser usados para identificar pontos de quebra de rearranjos em resolução muito alta, tipicamente em vinte kilobases (em média 5 kb). O método pode ser usado em aplicações de diagnóstico da maneira apresentada a seguir, por exemplo, diagnóstico pré-natal, diagnóstico pós-natal e para a análise de tumor e outras amostras doentes para identificar doenças de base de rearranjos genômicos ou sujeitos com predisposição a doença. Sequências de nucleotídeos amplificadas de interesse podem ser analisadas em microarranjos 4C feitos sob medida (da maneira descrita anteriormente), ou em micro-arranjos de ladrilhamentos genômicos, ou por sequenciamento da maneira aqui adicionalmente descrita.

*A triagem simultânea de rearranjos em múltiplos loci conhecidos por ser associados com uma doença.*

Para muitas doenças, síndromes ou fenótipos, múltiplas alterações de DNA causativas possíveis são conhecidas e os sujeitos precisam ser classificados para identificar o exato rearranjo básico da doença/síndrome/fenótipo. Por exemplo, no caso de rearranjos de leucemias linfoblásticas que envolvem AML1, MLL, MYC, BCL, BCR, ABL, imunoglobulina loci, LYL1, TAL1, TAL2, TCR $\alpha/\delta$ , TCR $\beta$ , HOX e possivelmente outro loci frequentemente fundamentam a doença; a tecnologia 4C pode ser aplicada para identificar qual locus e rearranjo são envolvidos em um paciente com a dada doença.

Nesta modalidade, a tecnologia 4C é direcionada a cada um dos loci de interesse. Cada locus pode ser analisado separadamente, mas múltiplos loci podem também ser analisados simultaneamente em um arranjo simples. Assim, uma triagem 4C pode envolver a amplificação de PCR (inversa) de elementos de DNA interagindo com uma ou mais sequências de nucleotídeos alvos próximas, cada qual dos loci que necessita ser analisado para rearranjos. Sequências alvos em, ou próximas, a esses loci são escolhidas com base no critério aqui mencionado.

A amplificação de sequências de nucleotídeos de interesse pode ser realizada separadamente para cada sequência de nucleotídeos alvo, ou pode ser realizada simultaneamente em um volume de reação por PCR multiplex (inversa). Prefere-se o último nos casos quando sequências de nucleotídeos de interesse interagindo com diferentes sequências de nucleotídeos alvos podem ser identicamente marcadas sem comprometer a análise. Isto pode ser o caso, por exemplo, quando loci são localizados em cromossomos diferentes ou quando loci são localizados no mesmo cromossomo a uma distância grande o bastante para a sobreposição mínima entre sinais de interação DNA-DNA, ou nos casos quando a sobreposição entre as interações DNA-DNA amplificada de diferentes sequências de nucleotídeos

alvos não interfere com a detecção de rearranjos genômicos.

A amplificação de sequências de nucleotídeos de interesse interagindo com diferentes sequências de nucleotídeos alvos é preferivelmente realizada separadamente quando cada conjunto de interações DNA-DNA precisa ser marcado diferentemente ou quando a eficiência de amplificação de conjuntos de iniciador de PCR inversa interfere um no outro.

Sequências diferencialmente ou identicamente marcadas de interesse interagindo com as várias sequências de nucleotídeos alvos são hibridizadas em micro-arranjos simples ou múltiplos contendo sondas representando o genoma completo (por exemplo, arranjos ou arranjos de ladrilhamento de alta densidade) ou uma parte selecionada do genoma, da maneira descrita anteriormente. Sinais de hibridização serão comparados aos obtidos com uma amostra controle, onde um aumento ou diminuição nas frequências de interação de DNA-DNA medido no teste versus amostra controle é indicativo para um rearranjo de DNA na amostra teste.

Uma triagem 4C genômica não tendencioso para rearranjos em posições desconhecidas no genoma.

Em uma segunda modalidade, a tecnologia 4C é aplicada para uma triagem 4C genômica não tendencioso para identificar rearranjos em uma amostra de um sujeito doente (ou não doente), onde os rearranjos são previamente desconhecidos e/ou ocorrem em localizações desconhecidas. Nesta modalidade, as sequências alvos não podem ser escolhidas para ficar próximas ao rearranjo. As sequências alvos são assim não conhecidas ou suspeitas de estarem associadas com a doença. Ao contrário, elas são escolhidas para ser distribuídas por todo o genoma ou seção de genoma escolhido, para fornecer assim cobertura suficiente da cromatina a ser investigada. Preferivelmente, o genoma total é coberto.

Por exemplo, as sequências alvos são escolhidas de maneira que suas sequências de interação de interesse, sendo que maioria as sequências diretamente em volta da sequência alvo (isto é, em 5, 10, 15, 20, 25, 30, 35, 40, ou 45 a 50 Mb da sequência alvo) cobre completamente, ou uma parte substancial, do genoma ou um cromossomo ou uma parte de um cromossomo. Isto permitirá reconstruir mapas cromossômicos físicos presentes em qualquer tipo de sujeito ou célula.

Para que duas sequências alvos adjacentes no molde do cromossomo linear tenham sequências de DNA de interação sobrepostas (isto é, ambientes genômicos sobrepostos) elas precisam estar, por exemplo, 5, 10, 15, 20, 25, 30, 35, 40, 45, ou 50 a 100 Mb longe uma da outra. Assim, para cobrir o genoma completo ( $2-3 \times 10^9$  bp) com as sequências alvos separadas - são exigidas 10 Mb, 200-300 sequências alvo. Para cobrir o genoma completo com as sequências alvos separadas ~ 50 Mb, 40 a 60 sequências alvos são exigidas. Em uma última modalidade, apenas uma sequência alvo por cromossomo é exigida para cobrir o genoma completo.

Todas as sequências de nucleotídeos de interesse que interagem com as sequências alvos são amplificadas como em 4C, tanto juntas em um mistura de reação simples (ou um número limitado de misturas da reação) por PCR multiplex (inverso), quanto em reações de PCR separadas (inverso) que podem ser agrupadas posteriormente. Prefere-se PCR

5 Multiplex nos casos quando as sequências de nucleotídeos de interesse interagindo com diferentes sequências de nucleotídeos alvos podem ser identicamente marcadas sem comprometer a análise. A amplificação será feita separadamente para cada sequência de nucleotídeos alvo quando cada conjunto de interações DNA-DNA pertencentes a uma dada sequência alvo precisar ser marcada diferentemente e/ou quando eficiência de amplificação

10 dos conjuntos de iniciador de PCR inversa interferir um no outro.

*Identificação de rearranjos intracromossômicos tais como deleções, inserções e duplicações e inversões (equilibradas e não equilibradas).*

Em uma modalidade preferida, todas as sequências de DNA amplificadas interagindo com o conjunto de sequências alvos de uma amostra teste (por exemplo, uma amostra do paciente) são identicamente marcadas e frequências de interação amplas do genoma são comparadas com estas de uma amostra controle (por exemplo, de sujeito saudável). Amostras controle e teste podem ser hibridizadas no mesmo arranjo em diferentes cores, ou elas podem ser hibridizadas nos diferentes arranjos e comparadas. Um aumento ou uma diminuição na frequência de interação DNA-DNA para a amostra teste comparado com a amostra controle é indicativo para uma duplicação/inserção ou uma deleção na amostra teste. Pode-se também ser indicativo para uma inversão.

As Figuras 18 e 19 mostram uma deleção identificada por 4C desta maneira.

Em uma outra modalidade preferida, sequências de DNA amplificadas interagindo com o conjunto de sequências alvos de uma amostra teste (por exemplo, amostra do paciente) são marcadas em duas cores, com cores alternando para sequências alvos que são vizinhas no molde do cromossomo linear. Sequências alvos vizinhas são bastante próximas no molde do cromossomo linear para que suas sequências de interação sejam sobrepostas. Assim, sequências alvos vizinhas podem estar, por exemplo, 5, 10, 15, 20, 25, 30, 35, 40, 45 a 50 Mb longe uma da outra. Um projeto experimental como esse impede que rearranjos faltarem em virtude de sinais de hibridização saturados próximos a uma dada sequência alvo. As frequências de interação amplas do genoma são comparadas com as de uma amostra controle (por exemplo, de sujeito saudável). Amostras controle e teste podem ser diferencialmente marcadas e hibridizadas no mesmo arranjo, ou elas podem ser hibridizadas em diferentes arranjos e comparadas. Um aumento ou uma diminuição na frequência de interação DNA-DNA para a amostra teste comparada com a amostra controle é indicativo para um duplicação/inserção ou uma deleção na amostra teste. Pode-se também ser indicativo de uma inversão.

Uma análise 4C subsequente direcionada a sequências alvos que flanqueiam ou ficam dentro da parte rearranjada do genoma pode ser realizada para identificar inversões. Um padrão inverso de sinais de hibridização comparado a amostra controle identifica a inversão na amostra teste. Isto está descrito na Figura 20.

5 Em uma modalidade preferida adicional, sequências de nucleotídeos de interesse interagindo com diferentes sequências alvos justapostas no molde do cromossomo são marcadas com diferentes corantes. Rearranjos são detectados pelo aparecimento aparência ou desaparecimento de sinais de DNA interagentes na amostra do paciente comparado com a amostra controle. Isto está descrito nas Figuras 21 a 23.

10 *Identificação de translocações equilibradas e não equilibradas.*

Em uma modalidade preferida adicional, múltiplos corantes são disponíveis (por exemplo, 48 corantes) e cada cromossomo é marcado com dois corantes exclusivos que são usados de maneira que os corantes alternados entre sequências alvos que são vizinhos no molde do cromossomo linear. Todos os fragmentos de DNA podem ser hibridizados juntos com um arranjo contendo sondas representando o genoma completo. A identificação de interações DNA-DNA entre cromossomos que ocorrem na amostras teste mas não nas amostras controle são indicativos de uma translocação e identifica os dois cromossomos rearranjados. A transição de sinais baixos para altos em um cromossomo não relacionado identifica o sítio de reconhecimento de enzima de restrição primário próximo ao ponto de quebra do DNA.

20 Se translocações forem equilibradas, cada um dos dois corantes específicos para cromossomo deve dar sinais de hibridização fortes em um conjunto mutuamente exclusivo de sondas diretamente vizinhas umas das outras no molde linear do cromossomo não relacionado. O ponto de quebra neste cromossomo não relacionado é localizado entre os dois conjuntos de sondas mostrando sinais positivos de hibridização. Ver Figura 24.

25 Quando testados em uma amostra do paciente que envolve uma translocação cromossomo 1:cromossomo 7, os resultados mostrados na Figura 25 são obtidos. Nesta amostra particular as sequências alvos no arranjo representado uma triagem de sequências localizada próximas aos os sítios *Hind*III (ver anteriormente para a derivação de sequências alvo).

30 Se as translocações ocorrerem com perda de DNA nos pontos de quebra (isto é, translocações não equilibradas), cada um dos dois corantes específicos para cromossomo darão sinais de hibridização fortes em um conjunto mutuamente exclusivo de sondas no cromossomo não relacionado que não é diretamente vizinho um do outro no molde linear do cromossomo não relacionado. As sondas localizadas no cromossomo não relacionado entre os dois conjuntos de sondas apresentando interações DNA-DNA intercromossômicas representam a região genômica que foi perdida. A maioria sondas externas desta região marca

os sítios de restrição próximos aos pontos de quebra no cromossomo não relacionado (ver Figura 26).

Quando testados em uma amostra do paciente que envolve uma translocação cromossomo 4:cromossomo 7 com uma quantidade desconhecida de DNA deletado no ponto de quebra, os resultados mostrados na Figura 27 são obtidos. Nesta amostra particular as sequências alvos no arranjo novamente representaram uma triagem de sequências localizadas próximas aos sítios *HindIII*. Os resultados mostram que tanto a translocação quanto a deleção são detectadas (aproximadamente 2 Mb).

Ainda em uma modalidade preferida adicional, múltiplos corantes (por exemplo, 24 corantes) são disponíveis e usados para marcar diferencialmente cada cromossomo (Figura 28). Assim, todas as sequências de DNA que interagem com as sequências alvos presentes no mesmo cromossomo são marcados identicamente e diferentes daqueles localizados em outros cromossomos. A identificação de interações DNA-DNA entre cromossomos que ocorrem em amostras teste mas não em amostras controle são indicativos de uma translocação e identifica os dois cromossomos rearranjados.

Uma análise 4C subsequente direcionada especificamente para os cromossomos envolvidos na translocação permite a identificação de pontos de quebra. Aqui, um ou cada um cromossomo é marcado com dois corantes exclusivos que são usados de maneira que eles alternem entre fragmentos de DNA interagindo com as sequências alvos que vizinhas no molde do cromossomo linear.

Se as translocações forem equilibradas, cada dos dois corantes específicos para cromossomo poderão dar sinais de hibridização fortes em um conjunto mutuamente exclusivo de sondas diretamente vizinhas umas das outras no molde linear do cromossomo não relacionado. O ponto de quebra deste cromossomo não relacionado é localizado em entre os dois conjuntos de sondas que apresentam sinais positivos de hibridização.

Se as translocações ocorrerem com perda de DNA nos pontos de quebra (isto é, translocações não equilibradas), cada um dos dois corantes específicos para cromossomo darão sinais de hibridização fortes em um conjunto mutuamente exclusivo de sondas no cromossomo não relacionado que não é diretamente vizinho um do outro no molde linear do cromossomo não relacionado. As sondas localizadas no cromossomo não relacionado entre os dois conjuntos de sondas mostrando interações DNA-DNA intercromossônicas representam a região genômica que foi perdida. A maioria das sondas externas desta região marca os sítios de restrição próximos aos pontos de quebra no cromossomo não relacionado.

Em uma outra modalidade preferida, menos que 24 corantes são disponíveis e cada corante é usado para marcar exclusivamente todos os fragmentos de DNA que interagem com as sequências alvos presentes no mesmo cromossomo. Assim, com 1, 2, 3, 4, 6, 8, 10, 12 etc. corantes, 1, 2, 3, 4, 6, 8, 10, 12 etc. cromossomos podem ser exclusivamente mar-

cados. Fragmentos de DNA marcados pertencentes a cada um desses cromossomos podem ser hibridizados juntos com um arranjo contendo as sondas representando o genoma completo. A aparência de sinais de hibridização representando interações DNA-DNA em um cromossomo não relacionado identifica este cromossomo como um parceiro de translocação. Uma análise 4C subsequente direcionada especificamente para os cromossomos envolvidos na translocação permite a identificação de pontos de quebra (da maneira descrita anteriormente).

Em uma outra modalidade preferida, menos que 48 corantes são disponíveis e cada cromossomo é marcado com dois corantes exclusivos que são usados de maneira que 5 eles alternem entre os fragmentos de DNA interagindo com as sequências alvos que vizinhas no molde do cromossomo linear. Assim, com 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24 etc. corantes 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 etc. cromossomos podem ser exclusivamente marcados. Todos os fragmentos de DNA podem ser hibridizados juntos com um arranjo contendo as sondas representando o genoma completo. Sinais de hibridização em um cromossomo não relacionado representam interações DNA-DNA intercromossômicas e identificam 10 este cromossomo como um parceiro de translocação. A transição dos sinais baixos para altos em um cromossomo não relacionado identifica o sítio de reconhecimento de enzima de 15 restrição primário próximo ao ponto de quebra do DNA.

Ainda em uma modalidade preferida adicional, menos que 48 corantes são disponíveis 20 e cada cromossomo é marcado com mais que dois corantes exclusivos que são usados de maneira que eles alternem entre os fragmentos de DNA interagindo com as sequências alvos que vizinhas no molde do cromossomo linear. Todos os fragmentos de DNA podem ser hibridizados juntos com um arranjo contendo as sondas representando o genoma completo. Sinais de hibridização em um cromossomo não relacionado representam interações 25 DNA-DNA intercromossômicas e identificam este cromossomo como um parceiro de translocação. A transição dos sinais baixos para altos em um cromossomo não relacionado identifica o sítio de reconhecimento de enzima de restrição primário próximo ao ponto de quebra do DNA.

#### *Como identificar outros pontos de quebra de translocação*

30 Os pontos de quebra identificados em um cromossomo de acordo com quaisquer das estratégias supramencionadas podem ser usados subsequentemente para identificar os pontos de quebra no cromossomo parceiro de translocação. Isto pode ser feito, por exemplo, por meio de sequenciamento de produtos obtidos através de PCR de longo alcance nas Junções de DNA ou PCR mediado por ligação (LM) ou PCR inversa em círculos de DNA 35 criados por digestão e religação de enzima de restrição de DNA genômico (não reticulado), usando iniciadores específicos para o cromossomo que lê nas sequências do outro cromossomo.

Em uma modalidade preferida, cada uma das classificações supramencionadas para rearranjos genômicos pode ser seguida por um experimento 4C dedicados direcionados contra as sequências alvos próximas candidatas aos pontos de quebra para identificá-las de forma inequívoca como tal.

- 5 Aspectos dos métodos descritos anteriormente para a detecção de diferentes tipos de rearranjos genômicos podem ser combinados para classificar simultaneamente o genoma para sua ocorrência.

Se arranjos de ladrilhamento genômicos forem usados em vez de arranjos 4C, transições em intensidades de sinal para sondas justapostas no molde do cromossomo linear 10 que são observados em paciente, mas não no controle, amostras, indicam a posição de pontos de quebra associados com um rearranjo genômico (em vez do sítio de reconhecimento de enzima de restrição primário próximo ao ponto de quebra do DNA).

#### Sequenciamento

Sequenciamento de DNA de alta produção prometem tornar uma alternativa disponível 15 e mais quantitativa para micro-arranjos para analisar grandes coleções de sequências de DNA. Exemplos de abordagens de sequenciamento de alta produção são listados em E. Y. Chan, *Mutation Research* 573 (2005) 13-40 e incluem, mas sem limitações, abordagens de sequenciamento de termo próximas tais como abordagens de extensão do ciclo, abordagens de leitura polimerase e sequenciamento exonuclease, abordagens de sequenciamento 20 revolucionárias tais como varredura de DNA e sequenciamento nanoporo e análise linear direta. Exemplos de métodos de sequenciamento atual de alta produção são (pиро)sequenciamento 454, método de sequenciamento de Solexa Genoma Analysis System, Agencourt SOLiD (Biosistemas aplicados), sequenciamento MS-PET (Ng *et al.*, 2006, <http://nar.oxfordjournals.org/cgi/content/full/34/12/e84>).

25 Sequenciamento pode substituir hibridização do arranjo em análise de alta produção dos resultados de 4C e outro abordado com base em detecção de interações genômicas. A frequência de ocorrência de uma sequência é indicativa da frequência de associação no genoma, e pode ser analisada da mesma maneira que os resultados de hibridização são analisados anteriormente.

30 Sequenciamento é realizado no molde da maneira fornecida pelas etapas a-g, descritas na reivindicação 1. Alternativamente, o sequenciamento pode ser realizado nos produtos de PCR obtidos por métodos da maneira descrita por Lomvardas *et al.*, *Cell* 126, 403-413, July 28, 2006 ou por Ling *et al.*, *Science* 312, 14 April 2006, 269-272.

35 Sequenciamento é iniciado de uma ou ambas extremidades dos produtos de PCR. Ambas extremidades do produto de PCR consistem em sequências de nucleotídeos de composições de nucleotídeo conhecidas, com pelo menos uma extremidade sendo sequência de nucleotídeos alvo, e flanqueia uma sequência de nucleotídeos de interesse que foi

interagida e ligada às sequência de nucleotídeos alvo. Dependendo do método de sequenciamento usado, adaptadores podem precisar ser adicionados a uma ou ambas extremidades dos produtos de PCR. Adaptadores podem ser sequências de oligonucleotídeo exigidos para o método de sequenciamento de interesse, que pode ou não conter frações, por exemplo, que permitam que elas sejam capturadas. Adaptadores podem ser ligados aos produtos de PCR diretamente ou após tornar as extremidades dos produtos de PCR rombas. Alternativamente, conjuntos iniciador de PCR, da maneira usada em etapa g (reivindicação 1), podem conter projeções que representam sequências de adaptadores ou podem conter projeções que introduzem sítios de clivagem de enzima de restrição que pode ser usada para ligação subsequente de adaptadores específicos ou não específicos para cada extremidade do produto de PCR.

Em experimentos 4C que ajuda identificar as sequências de DNA interagindo com uma sequência alvo de nucleotídeo simples, sequenciamento precisa ler através do evento de ligação primária (etapa c) e/ou o evento de ligação secundária (etapa f) de maneira que informação de sequência suficiente seja obtida para identificar as sequências de nucleotídeos de interesse. Tipicamente, isto exige o sequenciamento e identificação de trechos de minimamente 8-30 nucleotídeos além da junção de ligação na sequência de nucleotídeos de interesse (ver figura 28).

No sequenciamento 4C multiplex precisa-se ler no evento de ligação primária (etapa c) e/ou no evento de ligação secundária (etapa f) de maneira que informação de sequência suficiente seja obtida para identificar tanto a sequência de nucleotídeos alvo quanto a sequência de nucleotídeos de interesse que juntas formam o produto de ligação. Tipicamente, isto exige o sequenciamento e identificação de trechos de 8-30 nucleotídeos em cada lado da junção de ligação. A identificação da sequência de nucleotídeos alvo fornecerá cada produto de ligação com um “endereço domiciliar”. Iniciadores usados na etapa g que hibridizam na sequência de nucleotídeos alvo precisam ser localizados a uma distância da junção de ligação primária e secundária que seja grande o bastante para o sequenciamento identificar de forma inequívoca este “endereço domiciliar”. Dependendo do método de sequenciamento, esta distância pode ser minimamente 0, 10, 20 30, 40, 50, 60, 70, 80, 90 ou 100 nucleotídeos longe da junção de ligação.

Em 4C uniplex e multiplex, quando os produtos de PCR consistem em uma sequência de nucleotídeos de interesse flanqueados em cada lado pelas sequências de nucleotídeos alvo, a leitura de uma extremidade do produto de PCR através da junção de ligação primária, e leitura da outra extremidade do produto de PCR através da junção de ligação secundária, fornece a mesma informação. Assim, as sequências obtidas de um lado do produto de PCR são suficientes para a análise de interações DNA-DNA. Sequências obtidas do outro lado do produto de PCR podem ser usadas para complementar ou para verificar os

dados. Ver Figura 29.

A iniciação das sequência de reações ideais pode ser feita usando os iniciadores padrões rotineiramente usados pelas plataformas de sequenciamento de alta produção respectivas. Ela pode também ser feita com iniciadores de sequenciamento customizados que 5 sobrepõem parcialmente ou completamente os iniciadores de PCR inversa usados para amplificar as sequências capturadas por uma dada sequência alvo. Isto impediria a releitura da sequência de iniciador de PCR inversa total e permitiria a leitura de mais nucleotídeos do fragmento capturado, que portanto podem ser identificados mais facilmente e mapeados no genoma. Estes iniciadores de sequenciamento customizados podem então também anelar 10 parcialmente nas sequências de adaptadores normalmente usadas como a fita de DNA complementar ao iniciador de sequenciamento.

Em uma configuração multiplex, múltiplos iniciadores de sequenciamento que sobrepõem tanto com a sequência de adaptador quanto com as extremidades externas dos produtos de DNA amplificados por PCR inversa podem ser usados simultaneamente, de 15 maneira que cada fragmento alvo incluído na análise tenha seu próprio iniciador de sequência exclusivo. A sobreposição com as extremidades externas dos produtos de DNA amplificados por PCR inversa (isto é, a sequência de iniciador de PCR inversa) deve ser de maneira que nucleotídeos suficientes da sequência alvo sejam ainda disponíveis para sequenciamento e de maneira que cada sequência alvo (endereço domiciliar) possa ser identificada de 20 forma inequívoca. Dependendo da natureza e número de sequências alvos analisadas simultaneamente, isto significa que em uma modalidade, os iniciadores de sequenciamento podem hibridizar 1 a 20 nucleotídeos longe do sítio de restrição que forma a junção entre a sequência alvo e a sequência capturada. A identificação de cada combinação de sequência 25 alvo e sequência capturada pode também ser feita usando sequenciamento di-tag, que fornece informação de sequência de ambas extremidades de cada fragmento de DNA analisado.

Em uma outra modalidade, os iniciadores de sequenciamento são desenhados para cada sequência alvo de maneira que eles sejam próximos aos sítios de reconhecimento de enzima de restrição primários e/ou secundários analisados e podem sobrepor parcialmente 30 ou completamente nos sítios de reconhecimento de enzima de restrição primários e/ou secundários analisado.

Em uma modalidade, a tecnologia de sequenciamento de alta produção usada é o sequenciamento Solexa (Illumina).

Em uma modalidade, o sequenciamento pode ser direcionado para o lado do sítio 35 de reconhecimento de enzima de restrição primário. Isto pode impedir a leitura de eventos de ligação aleatórios que ocorrem durante a segunda etapa de ligação.

Em uma modalidade, sequenciamento pode ser direcionada a o lado do sítio de re-

conhecimento de enzima de restrição secundário.

*Análise dos dados*

Produtos de ligação analisados pela tecnologia 4C são compostos de um “endereço domiciliar” (sequência de nucleotídeos alvo) e uma sequência de nucleotídeos interagente de interesse (“endereço de deslocamento”). No caso de uniplex 4C este “endereço domiciliar” é conhecido (sendo a uma sequência de nucleotídeos alvo analisada).

Nos casos de 4C multiplex, o endereço “domiciliar” é identificado comparando eletronicamente a sequência de nucleotídeos alvo obtida com um biblioteca/base de dados contendo todas as sequências de nucleotídeos alvos incluídas na análise. O ‘endereço de deslocamento’ obtido (isto é, sequência de nucleotídeos de interesse) é identificado comparando eletronicamente sua sequência com um biblioteca/base de dados contendo todos os fragmentos de DNA genômico que são localizados entre um sítio de reconhecimento de enzima de restrição primário e secundário de escolha.

Para visualizar os dados e facilitar a análise, em uma modalidade preferida cada produto de ligação sequenciado é traçados graficamente ao longo do molde linear dos cromossomos, na localização genômica do ‘endereço de deslocamento’ (sequência de nucleotídeos de interesse). Códigos de cores exclusivas revelam o “endereço domiciliar” de cada produto de ligação e a frequência de detecção de cada produto de ligação é indicada graficamente. Rearranjos genômicos podem ser detectados dividindo cada frequência de interação medida em uma amostra (por exemplo, obtida de um paciente) daquelas medidas na outra amostra (por exemplo, de sujeito saudável). Estes valores podem ser colocados igualmente em gráfico ao longo dos moldes do cromossomo linear.

Uma diminuição na frequência de interação DNA-DNA na amostra teste é indicativa de uma deleção. Tipicamente, isto coincide com um aumento nas frequências de interação de DNA-DNA para sequências além do ponto de quebra mais distal medido a partir da sequência alvo. Um aumento na frequência de interação DNA-DNA na amostra teste é indicativo de uma duplicação. Tipicamente, isto coincide com uma diminuição nas frequências de interação de DNA-DNA para sequências além do ponto de quebra mais distal medido a partir da sequência alvo. Uma inversão na frequência de interação DNA-DNA na amostra teste é indicativa de uma inversão genômica. A detecção de interações DNA-DNA através dos cromossomos é indicativa de uma translocação. Os pontos de quebra são detectados da maneira descrita para a análise de micro-arranjo.

Biomarcadores

A identificação de rearranjos - tais como translocações, inversões e deleções, que são associados com uma doença permite a identificação de biomarcadores que podem ser usados para diagnosticar a doença. Por exemplo, sondas de hibridização ou iniciadores de PCR que detectam um dado rearranjo podem ser desenhados e usados para diagnosticar a

doença em um paciente. Sondas PCR podem ser desenhadas de acordo com as técnicas conhecidas na tecnologia, de maneira que uma região suscetível para rearranjo em uma condição da doença é amplificada usando os iniciadores; a natureza do produto de amplificação será indicativa da presença ou ausência da doença. Alternativamente, podem ser 5 desenhadas sondas de hibridização ou iniciadores que hibridizarão exclusivamente na presença ou ausência do rearranjo. Proteínas de fusão resultantes de rearranjos podem ser detectadas por técnicas tais como detecção de anticorpo com anticorpos desenhados de acordo com as técnicas conhecidas na tecnologia ou espectrometria de massa.

Sujeiro

10 O termo "sujeito" inclui mamíferos - tais como animais e humanos.

Agente

O agente pode ser um composto orgânico ou outra substância química. O agente pode ser um composto, que é obtêivel de ou produzido por qualquer fonte adequada, seja 15 natural ou artificial. O agente pode ser uma molécula de aminoácido, um polipeptídeo, ou uma substância química derivada destes, ou uma combinação destes. O agente pode ainda ser uma molécula de polinucleotídeo - que pode ser uma molécula sentido ou um anti-sentido, ou um anticorpo, por exemplo, um anticorpo policlonal, um anticorpo monoclonal ou um anticorpo humanizado monoclonal.

Várias estratégias foram desenvolvidas para produzir anticorpos monoclonais com 20 caráter humano, que desviam o necessário para uma linhagem celular humana de produção de anticorpo. Por exemplo, anticorpos monoclonais de camundongos usados foram "humanizados" ligando regiões variáveis de roedor e regiões constantes de humano (Winter, G. e Milstein, C. (1991) *Nature* 349, 293-299). Isto reduz a imunogenicidade de anticamundongos humano do anticorpo, mas a imunogenicidade residual é retida em virtude da estrutura da 25 região V anterior. Além disso, a especificidade de ligação de antígeno é essencialmente aquela do doador murino. Enxerto CDR e manipulação da estrutura (EP 0239400) melhoraram e refinaram a manipulação do anticorpo até o ponto onde é possível produzir anticorpos humanizados murinos que são aceitáveis para o uso terapêutico em humanos. Anticorpos humanizados podem ser obtidos usando outros métodos bem conhecidos na tecnologia (por 30 exemplo, da maneira descrita em US-A-239400).

Os agentes podem ser anexados a uma entidade (por exemplo, uma molécula orgânica) por um ligante que pode ser um ligante bifuncional hidrolisável.

A entidade pode ser desenhada ou obtida de uma biblioteca de compostos, que podem 35 compreender peptídeos, bem como outros compostos, tais como pequenas moléculas orgânicas.

A título de exemplo, a entidade pode ser uma substância natural, uma macromolécula biológica, ou um extrato feito de materiais biológicos tais como bactéria, fungo, ou célu-

las ou tecidos de animal (particularmente mamífero), uma molécula orgânica ou inorgânica, um agente sintético, um agente semi-sintético, um mimético estrutural ou funcional, um peptídeo, um peptidomiméticos, um peptídeo clivado de uma proteína total, ou um peptídeo sintetizado sinteticamente (tais como, a título de exemplo, tanto usando um sintetizador de peptídeo quanto por técnicas recombinantes ou combinações destes, um agente recombinante, 5 um anticorpo, um agente natural ou um não natural, uma proteína de fusão ou equivalente destes e mutantes, derivados ou combinações destes.

Tipicamente, a entidade será um composto orgânico. Para alguns exemplos, os compostos orgânicos compreenderão dois ou mais grupos hidrocarbila. Aqui, o termo "agrupamento hidrocarbila" significa um agrupamento compreendendo pelo menos C e H e pode compreender opcionalmente um ou mais outros substituintes adequados. Exemplos de tais substituintes podem incluir um agrupamento halo-, alcóxi-, nitro-, alquila, um agrupamento cílico etc. Além das possibilidades de os substituintes serem um agrupamento cílico, uma combinação de substituintes pode formar um agrupamento cílico. Se o agrupamento hidrocarbila compreender mais que um C, então aqueles carbonos não precisam necessariamente ser ligados um no outro. Por exemplo, pelo menos dois dos carbonos podem ser ligados por meio de um elemento ou agrupamento adequado. Assim, o agrupamento hidrocarbila pode conter heteroátomos. Heteroátomos adequados ficarão aparente aos versados na tecnologia que e incluirão, por exemplo, enxofre, nitrogênio e oxigênio. Para algumas aplicações, a entidade preferivelmente compreende pelo menos um agrupamento cílico. O agrupamento cílico pode ser um agrupamento policílico, tal como um agrupamento policílico não fundido. Para algumas aplicações, a entidade compreende pelo menos um dos ditos grupos cílicos ligados a um outro agrupamento hidrocarbila.

A entidade pode conter grupos halo - tais como grupos flúor, cloro, bromo ou iodo.

25 A entidade pode conter um ou mais dos grupos alquila, alcóxi, alquenila, alquileno e alquenileno - que pode ser de cadeia não ramificada ou ramificada.

#### Doença

Aspectos da presente invenção podem ser usados para o tratamento e/ou prevenção e/ou diagnóstico e/ou prognóstico de uma doença - tais como aquelas listadas em WO-30 A- 98/09985.

Para facilidade de referência, parte desta que lista é agora fornecida: atividade inibitória de macrófago e/ou inibitória de célula T e assim, atividade anti-inflamatória; atividade anti-imune, isto é, efeitos inibitórios contra uma resposta imune celular e/ou humoral, incluindo uma resposta não associada com inflamação; doenças associadas com vírus e/ou outro patógenos intracelulares; inibir a capacidade de macrófagos e células T de aderir nos componentes da matriz extracelular e fibronectina, bem como expressão do receptor Fas suprarregulado em células T; inibir reação imune e inflamação indesejável incluindo artrite,

incluindo artrite reumatóide, inflamação associada com hipersensibilidade, reações alérgicas, asma, lupus eritematoso sistêmico, doenças de colágeno e outras doenças autoimunes, inflamação associada com aterosclerose, arteriosclerose, doença do coração aterosclerótico, lesão de reperfusão, parada cardíaca, infarto do miocárdio, distúrbios inflamatórios vasculares, síndrome de dificuldade respiratória ou outras doenças cardiopulmonares, inflamação associada com úlcera péptica, colite ulcerativa e outras doenças do trato gastrintestinal, fibrose hepática, cirrose do fígado ou outras doenças hepáticas, tireoidite ou outras doenças glandulares, glomerulonefrites ou outras doenças renais e urológicas, otite ou outras doenças otorrino-laringológica, dermatite ou outras doenças dérmicas, doenças periodontais ou outras doenças dentais, infertilidade orquite ou epidídimoorquite, trauma orquidal ou outras doenças testiculares relacionadas imunes, disfunção da placenta, insuficiência da placenta, aborto habitual, eclampsia, pré-eclâmpsia e outras doenças imune e/ou genicológica relacionada inflamatória, uveite posterior, uveite intermediária, uveite anterior, conjuntivite, coriorretinite, uveoretinite, neurite ótica, inflamação intraocular, por exemplo, retinite ou oedema macular cistóide, oftalmia simpatética, esclerite, retinite pigmentosa, componentes imunes e inflamatórios de doença de fundo degenerativo, componentes inflamatórios de trauma ocular, inflamação ocular causada por infecção, vitreoretinopatia proliferativa, neuropatia ótica de isquemia aguda, cicatriz excessiva, por exemplo, após a operação de filtração de glaucoma, reação imune e/ou inflamação contra implantes oculares e outras doenças oftálmicas inflamatória relacionadas imunes e inflamação associada com doenças ou condições ou distúrbios autoimunes onde, tanto no sistema nervoso central (CNS) quanto em qualquer outro órgão, supressão imune e/ou inflamação seria benéfica, doença de Parkinson, complicação e/ou efeitos colaterais do tratamento de doença de Parkinson, demência complexa relacionada a AIDS, encefalopatia relacionada a HFIV, doença de Devic, corea de Sydenham, doença de Alzheimer e outras doenças degenerativas, condições ou distúrbios do CNS, componentes inflamatórios de acidente vascular cerebral, síndrome pós-polio, componentes imunes e inflamatórios de distúrbios psiquiátricos, mielite, encefalite, pan-encefalite de esclerose subaguda, encefalomielite, neuropatia aguda, neuropatia subaguda, neuropatia crônica, síndrome de Guillain-Barre, corea de Sydenham, miastenia grave, pseudotumor cerebral, Síndrome de Down, doença de Huntington, esclerose lateral amiotrófica, componentes inflamatórios de compressão de CNS ou trauma do CNS ou infecções do CNS, componentes inflamatórios de atrofias e distrofias musculares, e doenças relacionadas imunes e inflamatórias, condições ou distúrbios dos sistemas nervosos central e periférico, inflamação pós-traumática, choque séptico, doenças infecciosos, complicações inflamatórias ou efeitos colaterais de cirurgia, transplante de medula óssea ou outras complicações e/ou efeitos colaterais de transplante, complicações inflamatórias e/ou imunes e efeitos colaterais de terapia de gene, por exemplo, em virtude de infecção com um veículo viral, ou inflamação asso-

ciada com AIDS, para suprimir ou inibir uma resposta humoral e/ou imune celular, para tratar ou melhorar doenças proliferativas de monócito ou leucócito, por exemplo, leucemia, reduzindo a quantidade de monócitos ou linfócitos, para a prevenção e/ou tratamento de rejeição a enxerto no casos de transplante de células naturais ou artificiais, tecido e órgãos tais como 5 córnea, medula óssea, órgãos, lentes, marcapassos, tecido de pele natural ou artificial. Distúrbios relacionados a câncer específicos incluem, mas sem limitações, tumores sólidos; tumores carregados no sangue tais como leucemia; metástase tumoral; tumores benignos, por exemplo, hemangiomas, neuromas acústicos, neurofibromas, tracoma, e granulomas 10 piogênicas; artrite reumatóide; psoríase; doenças angiogênicas oculares, por exemplo, retinopatia diabética, retinopatia de pré-maturidade, degeneração macular, rejeição a enxerto de córnea, glaucoma neovascular, fibroplasia retroletal, rubeose; síndrome de Osier- Webber; angiogênese miocardial; neovascularização de placa; telangiectasia; juntas hemofílicas; 15 angiofibroma; granulação de ferida; colaterais coronárias; colaterais cerebrais; más formações arteriovenosa; angiogênese de elemento isquêmico; glaucoma neovascular; fibroplasia retroletal; neovascularização diabética; doenças relacionadas a Helicobacter, fraturas, vasculogênese, hematopoiese, ovulação, menstruação e placentação.

Preferivelmente, a doença é câncer - tais como leucemia linfócita aguda (ALL), leucemia mielóide aguda (AML), câncer adrenocortical, câncer anal, câncer da bexiga, câncer sanguíneo, câncer ósseo, tumor cerebral, câncer de mama, câncer do sistema genital feminino, câncer do genital sistema masculino, linfoma do sistema nervoso central, câncer cervical, rabdomiosarcoma infantil, sarcoma infantil, leucemia linfócita crônica (CLL), leucemia mielóide crônica (CML), câncer do cólon e retal, câncer cólon, câncer endométrico, sarcoma endométrico, câncer esofagial, câncer do olho, câncer da vesícula biliar, câncer gástrico, câncer do trato gastrintestinal, leucemia da célula capilar, câncer da cabeça e pescoço, câncer 20 hepatocelular, doença de Hodgkin, câncer hipofaringeal, sarcoma de Kaposi, câncer do rim, câncer da laringe, leucemia, câncer do fígado, câncer do pulmão, histiocitoma fibroso maligno, timoma maligno, melanoma, mesotelioma, mieloma múltiplo, mieloma, câncer da cavidade nasal e seio paranasal, nasofaringeal câncer, câncer do sistema nervoso, neuroblastoma, linfoma de não Hodgkin, câncer da cavidade oral, câncer orofaringeal, osteosarcoma, câncer do ovário, câncer pancreático, câncer da paratireóide, câncer do pênis, câncer da faringe, tumor pituitário, neoplasma de célula plasmática, linfoma CNS primário, câncer 25 próstata, câncer retal, sistema respiratório, retinoblastoma, câncer da glândula salivar, câncer de pele, câncer do intestino delgado, sarcoma do tecido macio, câncer de estômago, câncer testicular, câncer da tireóide, câncer do sistema urinário, sarcoma uterino, câncer 30 vaginal, sistema vascular, macroglobulinemia de Waldenstrom e tumor de Wilms.

#### Estojos

Os materiais para uso nos métodos da presente invenção são idealmente adequa-

dos para a preparação de estojos.

Um estojo como este pode compreender recipientes, cada qual com um ou mais dos vários reagentes (tipicamente em forma concentrada) utilizados nos métodos aqui descritos, incluindo, por exemplo, uma enzima de restrição primária, uma enzima de restrição secundária, um agente de reticulação, uma enzima de ligação (por exemplo, uma ligase) e um agente para reverter a reticulação (por exemplo, proteinase K).

Os oligonucleotídeos podem também ser fornecidos em recipientes que podem ser em qualquer forma, por exemplo, liofilizada, ou em solução (por exemplo, uma água destilada ou solução tamponada), etc.

Em um aspecto preferido da presente invenção, é fornecido um estojo compreendendo um conjunto de sondas da maneira aqui descrita, um arranjo e opcionalmente uma ou mais marcadores.

Um conjunto de instruções será também tipicamente incluído.

#### Usos

Vantajosamente, a presente invenção pode ser usada a fim de obter informação a cerca da organização espacial de sequências de nucleotídeos - tais como loci genômicos in vitro ou in vivo.

A título de exemplo, a tecnologia 4C pode ser usada para estudar a organização tridimensional de um ou mais loci do gene. Em particular, esta tecnologia pode ser usada para estudar o papel de um ou mais fatores de transcrição na organização tridimensional de um ou mais loci do gene.

A título de exemplo adicional, a tecnologia 4C pode ser usada para estudar o papel de fatores de transação e elementos cis-regulatórios de DNA.

A título de exemplo adicional, a tecnologia 4C pode ser usada para estudar regulação do gene de longo alcance in vitro ou in vivo.

A título de exemplo adicional, a tecnologia 4C pode ser usada para estudar proximidade e interação intracromossômica.

A título de exemplo adicional, a tecnologia 4C pode ser usada para estudar proximidade e interação inter-cromossômica.

A título de exemplo adicional, a tecnologia 4C pode ser usada para identificar sequências de nucleotídeos que funcionam com um promotor, melhorador, silenciador, isolante, região de controle dos locus, origem de replicação, MAR, SAR, centrômero, telômero ou qualquer outra sequência de interesse em uma rede regulatória.

A título de exemplo adicional, a tecnologia 4C pode ser usada para identificar genes responsáveis por um fenótipo (doença) nos casos onde acontece de uma mutação e/ou deleção afetar um elemento regulatório distante e seu mapeamento falta portanto fornecer tal informação.

A título de exemplo adicional, a tecnologia 4C pode ser usada para reconstruir eventualmente a conformação espacial de loci do gene, grandes regiões genômicas ou mesmo cromossomos completos.

5 A título de exemplo adicional, a tecnologia 4C pode ser usada para definir sequências âncoras potenciais que mantêm certos cromossomos juntos no espaço nuclear.

A título de exemplo adicional, a tecnologia 4C pode ser usada para reconstruir eventualmente em alta resolução o posicionamento de cromossomos uns com relação aos outros.

10 A título de exemplo adicional, a tecnologia 4C pode ser usada em diagnóstico (por exemplo, diagnóstico pré-natal) para detectar ou identificar rearranjos genômicos e/ou aberações - tais como translocações, deleções, inversões, duplicações.

#### Técnicas de Metodologia de DNA Recombinante Gerais

A presente invenção emprega, a menos que de outra forma indicada, técnicas convencionais de substância química, biologia molecular, microbiologia, DNA recombinante e imunologia, que estão na capacidade de versados na tecnologia. Tais técnicas são explicadas na literatura. Ver, por exemplo, *J. Sambrook, E. F. Fritsch, and T. Maniatis, 1989, Molecular Cloning: A Laboratory Manual, Second Edition, Books 1-3, Cold Spring Harbor Laboratory Press; Ausubel, F. M. et al. (1995 and periodic supplements; Current Protocols in Molecular Biology, ch. 9, 13, and 16, John Wiley & Sons, New York, N. Y.); B. Roe, J. Crabtree, and A. Kahn, 1996, DNA Isolation and Sequencing: Essential Techniques, John Wiley & Sons; M. J. Gait (Editor), 1984, Oligonucleotide Synthesis: A Practical Approach, Irl Press; and, D. M. J. Lilley and J. E. Dahlberg, 1992, Methods of Enzymology: DNA Structure Part A: Synthesis and Physical Analysis of DNA Methods in Enzymology, Academic Press*. Cada um destes textos gerais está aqui incorporado pela referência.

25 A invenção será agora descrita com mais detalhes a título de exemplo, que visa servir aos versados na tecnologia na realização da invenção e não devem de qualquer maneira se limitar ao escopo da invenção.

#### EXEMPLO 1

30 Seção Materiais & Métodos que está relacionada com as figuras 2, 13, 14, 15, 16, 17, 19.

#### Tecnologia 4C

As etapas iniciais do procedimento da tecnologia 3C foram realizadas da maneira descrita previamente (Splinter *et al.* (2004). *Methods Enzymol* 375, 493-507 (2004), produzindo produtos de ligação entre fragmentos de *Hind*III. Este molde 3C ligado a *Hind*III (~ 50 µg) foi digerido por toda a noite a 100 ng/µL com 50U de uma enzima de restrição secundária, frequente de corte, sendo tanto *Dpn*II (HS2, Rad23A) quanto *Nla*III ( $\beta$ -principal). Para evitar restrição na formação de círculo de DNA (Rippe *et al.* (1995) *Trends Biochem Sci* 20,

500- 6), tomou-se cuidado para escolher uma enzima de restrição secundária que não cortou em cerca de 350 a 400 bp do sítio de restrição *Hind*III que marca o fragmento de restrição de interesse (isto é, a 'isca'). Após digestão da enzima de restrição secundária, o DNA foi extraído com fenol, precipitado com etanol e subsequentemente ligado em baixa concentração (amostra de 50  $\mu$ g em 14 mL usando 200 U ligase (Roche), 4 horas a 16 °C) para promover formação de círculo *Dpn*II ou *Dpn*II. Produtos de ligação foram extraídos com fenol e precipitado com etanol, usando glicogênio (Roche) como um veículo (20  $\mu$ g/mL). Os círculos de interesse foram linearizados digerindo por toda a noite com um 50U de uma enzima de restrição terciária que corta a isca entre os sítios de reconhecimento de enzima de restrição primária e secundária, usando as seguintes enzimas de restrição: Spel (HS2), PstI (Rad23A) e PflMI ( $\beta$ -principal). Esta etapa de linearização foi realizada para facilitar hibridização de iniciador subsequente durante as primeira rodadas de amplificação de PCR. Produtos digeridos foram purificados usando uma coluna de remoção de nucleotídeo QIAquick (250) (Qiagen).

15 As reações de PCR foram realizadas usando o sistema de PCR Expand Long Molde (Roche), usando condições cuidadosamente otimizadas para assegurar a amplificação linear de fragmentos dimensionado até 1,2 kb (80 % de fragmentos de PCR 4C são menores que 600 bp). As condições de PCR foram como a seguir: 94 °C por 2 minutos, 30 ciclos de 94 °C por 15 segundos, 55 °C por 1 minuto e 68 °C por 3 minutos, seguido por uma etapa final de 68 °C por 7 minutos. Foi determinada a quantidade máxima de molde que ainda mostra faixa linear de amplificação. Para isto, diluições em série de molde foram adicionadas às reações de PCR, material de DNA amplificado correu em um gel de agarose e produtos de PCR foram quantificados usando software ImageQuant. Tipicamente, 100 a 200 ng de molde por 50  $\mu$ L de reação de PCR fornece produtos na faixa linear de amplificação. 16 a 25 32 reações de PCR foram agrupadas e purificaram este molde 4C usando o sistema de remoção de nucleotídeo QIAquick (250) (Qiagen). Molde 4C purificado foi marcado e hibridizado em arranjos, de acordo com protocolos ChIP-chip padrões (Nimblegen Systems of Iceland, LLC). DNA genômico marcados diferencialmente, que foi digerido com a enzima primária e secundária usada no procedimento 4C, serviu como um molde controle para corrigir as 30 diferenças nas eficiências de hibridização. Para cada experimento duas amostras independentemente processadas foram marcadas com orientações de corante alternadas.

Sequências de iniciador 4C usadas:

HS2: 5'-ACTTCCTACACATTAACGAGCC-3',

5'- GCTGTTATCCCTTCTCTTCTAC-3'

35 Rad23A: 5'- TCACACGCGAAGTAGGCC-3',

5'- CCTTCCTCCACC ATGATGA-3'

$\beta$ -principal: 5'-AACGCATTTGCTCAATCAACTACTG-3',

5'-GTTGCTCCTCACATTTGCTTCTGAC-3'

Arranjos 4C

Arranjos e análise foram baseados na construção m34 de NCBI. Sondas (60-mers) foram selecionadas das sequências 100 bp à montante e à jusante de sítios *Hind*III. O conteúdo de CG foi otimizado no sentido de 50 %, para sinais de hibridização uniformes. Para impedir hibridização cruzada, sondas que tiveram qualquer similaridade com repetições altamente abundantes (RepBase 10,09)<sup>3</sup> foram removidas do conjunto de sonda. Além do mais, sondas que dão mais que duas colisões BLAST no genoma foram também removidas do conjunto de sonda. Alinhamentos da sequência foram realizados usando MegaBLAST (Zhang *et al.* (2000) *J Comput Biol* 7, 203-14) usando os ajustes padrões. Uma colisão foi definida como um alinhamento de 30 nt ou mais.

Análise dos dados 4C

A razão de sinal amostra 4C/DNA genômico foi calculada para cada sonda e os dados foram visualizados com software SignalMap fornecidos por Nimblegen Systems. Os dados foram analisados usando o pacote R (<http://www.r-project.org>) Spotfire and Excel. Razões de hibridização não processada apresentaram agrupamentos de 20 a 50 sinais 4C positivos ao longo do molde do cromossomo. Para definir estes agrupamentos, uma média corrida foi aplicada. Vários tamanhos de janela foram usados, variando de 9 a 39 sondas, todos os quais identificaram os mesmos agrupamentos. Os resultados mostrados foram baseados em um tamanho de janela de 29 sondas (em média 60 kb) e foram comparados com a média corrida realizada através de dados aleatórios. Isto foi feito para cada arranjo separadamente. Consequentemente, todas as medições foram apreciadas com relação à amplitude e ruído de tal arranjo específico. A Taxa de falsa descoberta (FDR), definida como (número de falsos positivos) / (número de falsos positivos + número de positivos verdadeiros) foi determinada da maneira a seguir: (número de positivos no conjunto aleatório) / (número de positivos nos dados). O nível do patamar foi determinado usando uma abordagem de cima para baixo para estabelecer o valor mínimo para cada: FDR < 0,05.

Em seguida, experimentos de duplicatas biológicas foram comparados. Janelas que atendem o patamar em ambas duplicatas foram consideradas positivas. Durante a comparação dos dados aleatórios, nenhuma janela esteve acima do patamar em ambas duplicatas. Janelas positivas diretamente adjacentes no molde do cromossomo foram unidas (não foi permita nenhuma folga), criando áreas positivas.

Análise de expressão

Para cada tecido, três microarranjos independentes foram realizados de acordo com o protocolo Affymetrix (2 arranjos - 430 camundongos). Dados foram normalizados usando RMA cat-tools; [www.bioconductor.org](http://www.bioconductor.org)) e para cada sonda-conjunto a média das medições foi calculada dos três microarranjos. Além do mais, quando múltiplos sonda-conjuntos

representaram o mesmo gene, a média deles foi também calculada. Mas5calls (Affy library: [www.bioconductor.org](http://www.bioconductor.org)) foi usado para estabelecer chamadas "presentes", "ausentes" e "marginais". Genes com uma chamada "presente" em todos os três arranjos e um valor de expressão maior que 50 foram chamados expressos. "Genes específicos de fígado fetal" foram classificados como genes que atendem seu critério de ser expresso no fígado fetal e teve valores de expressão mais que cinco vezes maior comparados ao cérebro fetal. Para fornecer uma medida da atividade transcrional total em torno de cada gene, uma soma da corrida foi aplicada. Para isso, foram usados valores de expressão log-transformados. Para cada gene foi calculada a soma de expressão de todos os genes encontrados em uma janela 100 kb à montante do início e 100 kb à jusante do final do gene, incluindo o próprio gene. Valores resultantes para genes ativos de regiões 4C positivas encontradas dentro (n = 124, 123 e 208 respectivamente para HS2 em fígado, Rad23A em cérebro e Rad23A em fígado) foram comparados com os valores obtidos para genes ativos fora das áreas 4C positivas (n= 153, 301 e 186, respectivamente, onde n= 153 corresponde ao número de genes ativos não interagentes, presentes entre a maioria de região interagente centromérica e o telômero do cromossomo 7); os dois grupos foram comparados usando um teste de soma de classificação 4C Wilcoxon de uma cauda.

#### Sondas FISH

Os clones BAC seguintes (BACPAC Resources Centre) foram usados; RP23-370E12 para Hbb-1, RP23-317H16 por chr.7a 80,1 Mb (agrupamento de gene OR), RP23-334E9 para Uros, RP23-32C19 para chr.7 a 118,3 Mb, RP23-143F10 para chr.7 a 130,1 Mb, RP23-470N5 para chr.7 a 73,1 Mb, RP23-247L11 para chr.7 a 135,0 Mb (agrupamento de gene OR), RP23-136A15 para Rad23A, RP23-307P24 para chr.8 a 21,8 Mb e RP23-460F21 para chr.8 a 122,4 Mb. Para uma sonda específica do centrômero do cromossomo 7 foi usado clone P1 5279 (Genoma Sistemas Inc.) que anela ao segmento de DNA D7Mit21. Sondas marcadas com iniciador aleatório foram preparadas usando BioPrime Array CGH Genomic Labeling Sistema (Invitrogen). Antes da marcação, o DNA foi digerido com *Dpn*II e purificado com um estojo de DNA clean & concentrator-5 (Zymo research). O DNA digerido (300 ng) foi marcado com SpectrumGreen dUTP (Vysis) ou Alexa fluor 594 dUTP (Molecular Probes) e purificados através de um estojo de purificação GFX PCR DNA e Gel Band (Amersham Biosciences) para remover nucleotídeos não incorporados. A especificidade das sondas marcadas foi testada em espalhamento de metáfase preparada a partir de células ES murino.

#### Cryo-FISH

Cryo-FISH foi realizado da maneira descrita antes <sup>5</sup>. Resumidamente, E14.5 de fígado e cérebro foram fixados por 20 minutos em paraformaldeído 4 %/HEPES 250 mM, pH 7,5 e cortado em pequenos blocos de tecido, seguido por uma outra etapa de fixação de 2

horas em paraformaldeído 8 % a 4 °C. Blocos de tecido fixados foram imersos em 2,3 M de sacarose por 20 minutos a temperatura ambiente, montados em um suporte de corpo de prova e congelados rápido em nitrogênio líquido. Blocos de tecido foram armazenados em nitrogênio líquido até o seccionamento. Crioseções de Ultrathin de aproximadamente 200 5 mn foram cortados usando um Reichert Ultramicrotome E equipado com crioacessório (Leica). Usando um laço cheio com sacarose, as seções foram transferidas para lamínulas e armazenadas a -20 °C. Para hibridização, as seções foram lavadas com PBS para remover sacarose, tratadas com RNase 250 ng/mL em 2 x SSC por 1 hora a 37 °C, incubadas por 10 minutos em M HCL 0,1, desidratadas em uma série de etanol e desnaturadas por 8 minutos 10 a 80 °C em formamida 70 %/2 x SSC, pH 7,5. Seções foram novamente desidratadas diretamente antes da hibridização da sonda. 500 ng de sonda marcada foram co-precipitados 15 com 5 µg de DNA camundongos COT1 (Invitrogen) e dissolvidos em hybmix (formamida 50 %, sulfato de dextrano 10 %, 2 x SSC, tampão de fosfato 50 mM, pH 7,5). As sondas foram desnaturadas por 5 minutos a 95 °C, ré-aneladas por 30 minutos a 37 °C e hibridizadas por 20 pelo menos 40 horas a 37 °C. Após as lavagens de pós-hibridização, os núcleos foram contracorados com DAPI 20 ng/mL (Sigma) em PBS/Tween-20 0,05 % e montados em reagente 25 de antidescoloração Prolong Gold (Molecular Probes).

As imagens foram coletadas com um microscópio de epifluorescência Zeiss Axio Imager Z1 (x 100 plan apochromat, 1,4 oil objective), equipado com uma câmera CCD e 20 software Isis FISH Imaging System (Metasystems). Um mínimo de alelos 250 β-globina ou Rad23A foi analisado e pontuado de maneira sobreposta ou não sobreposta com BACs localizados em outro lugar no genoma, por uma pessoa que desconhece a combinação de sonda aplicada às seções. Testes de ajuste replicados (estatística G-) <sup>6</sup> foram realizados 25 para avaliar a significância de diferenças entre valores medidos para regiões 4C positivas versus 4C negativas. O resumo dos resultados é fornecido na Tabela 2.

Embora tenhamos observado diferenças estatisticamente significativas entre frequências de interação de fundo (0,4-3,9 %) e verdadeiras (5 a 20,4 %) pode-se ficar claro 30 que frequências medidas por cryo-FISH são menores que aquelas medidas por outros usando protocolos FISH diferentes. Seccionamento pode separar alguns loci interagentes e crio- medições FISH portanto subestimarão ligeiramente frequências de interação verdadeiras. Por outro lado, procedimentos FISH 2D e 3D atuais sobreestimarão estas porcentagens 35 por causa da resolução limitada na direção z. No futuro, melhores técnicas de microscopia em combinação com sondas FISH mais específicas revelarão melhor as frequências de interação verdadeiras.

### 35 EXEMPLO 2

A digestão de enzima de restrição do procedimento 3 C (isto é, fixação de formaldeído, (primário), ré-ligação de fragmentos reticulados de DNA e purificação de DNA) é reali-

zada essencialmente da maneira descrita (Splinter *et al.*, (2004) *Methods Enzymol.* 375: 493-507), produzindo uma mistura de DNA (molde '3C') contendo fragmentos de restrição que são ligados em virtude de eles estarem originalmente próximos no espaço nuclear.

PCR inversa é realizado para amplificar todos os fragmentos ligados a um dado 5 fragmento de restrição ('isca'; escolhido em virtude de ele conter um promotor, melhorador, isolante, região de anexação da matriz, origem da replicação ou qualquer outro primeira sequência de nucleotídeo(alvo)).

Para isto, círculos de DNA são criados digerindo o molde 3 C com uma enzima de restrição secundária (preferivelmente um cortador frequente que reconhece a sequência de 10 tetra ou pentanucleotídeos), seguido por ligação em condições diluídas de maneira que interações intramolecular sejam favorecidas. Para minimizar uma tendência na formação de círculo em virtude da restrição topológica (Rippe *et al.*, (2001) *Trends in Biochem. Sciences* 26, 733-40), uma enzima de restrição secundária deve ser escolhida que preferivelmente corte a isca a >350-400 bp a partir do sítio de restrição primária. Para aumentar a eficiência 15 de amplificação de PCR inversa e reproduzibilidade, círculos são mais bem linearizados antes da amplificação de PCR por uma enzima de restrição (por exemplo, um cortador de 6 ou mais bp) que corta a isca entre o sítio de restrição primária e secundária do diagnóstico.

Digestão do molde 3 C com a enzima de restrição secundária, circularização através 20 da ligação em condições diluídas e linearização de isca contendo círculos são realizados em condições padrões para tais manipulações de DNA para produzir um molde de DNA para amplificação de PCR inversa ('molde 4C').

Consequentemente, 10 µg de molde 3 C são digeridos em 100 µL com 20U da enzima de restrição secundária (por toda a noite), seguido por inativação térmica da enzima e purificação do DNA. A ligação é realizada em 10 mL (1 ng/µL DNA) com 50U T4 ligase (4 25 horas a 16 °C, 30 minutos a RT), seguido por purificação de DNA. Finalmente, linearização dos círculos de interesse é feita em 100 µL com 20U de enzima de restrição (por toda a noite), seguido novamente por purificação do DNA.

Para PCR inversa, dois iniciadores específicos de isca são desenhados, cada qual 30 tão próximo quanto possível do sítio de reconhecimento de enzima de restrição primária e diretamente vizinho à secundária, respectivamente, e cada qual com sua extremidade 3' voltada para fora de maneira que a extensão se dê imediatamente através dos sítios de restrição em um fragmento ligado à isca. PCR inversa com estes iniciadores é preferivelmente realizada em 100 a 400 ng de DNA de molde 4C (por 50 µL de mistura de reação PCR), para incluir um número máximo de eventos de ligação por reação de PCR. Nós realizamos 35 PCR inversa aplicando o Expand Long Molde PCR Sistema (Roche), usando tampão 1 de acordo com os procedimentos do fabricante.

Os ciclos PCR seguinte são realizados:

1. 2 minutos 94 °C
2. 15 segundos 94 °C
3. 1 minuto 55 °C
4. 3 minutos 68 °C
5. repetir a etapa 2- 4 29 x (ou algo em torno de 25-40 x)
6. 7 minutos 68 °C
7. extremidade

Eletroforese de gel é realizada para analisar reprodutibilidade entre reações de PCR individuais. Tipicamente, padrões do produto idênticos devem ser obtidos.

A fim de obter material suficiente para marcação por iniciação aleatória e hibridização do arranjo, múltiplas reações de PCR (cada qual obtida após 30 ciclos de PCR) podem ser agrupadas, (em vez de aumentar o número de ciclos de PCR por reação). Como uma alternativa para marcação iniciada aleatoriamente, nucleotídeos marcados podem ser incorporados nos últimos ciclos de PCR (por exemplo, 30 ciclos (sem marcação) + 10 ciclos (marcação)).

### EXEMPLO 3

#### *Detecção de translocação usando tecnologia 4C*

A tecnologia 4C é usada para medir as frequências de interação para uma dada sequência X presente em um dado cromossomo A em células de um sujeito saudável e em células de um paciente carregando uma translocação simples recíproca entre os cromossomos A e B, com o ponto de quebra sendo próximo a sequência X (da maneira mostrada na Figura 8).

Em células normais esta análise revela sinais elevados de hibridização (isto é, interações frequentes com X) para (praticamente) cada sonda localizada dentro de 0,2- 10 Mb de sequência X no cromossomo A (o tamanho real da região cromossômica mostrando fortes sinais de reticulação depende principalmente da complexidade da amostra que foi hibridizada no arranjo). Em outro lugar no mesmo cromossomo A, bem como em outros cromossomos, não foi observada nenhuma região grande como essa (no molde de DNA linear) de sondas com sinais elevados de hibridização.

Em células de paciente entretanto, sinais de hibridização com todas as sondas do cromossomo A localizadas no outro lado do ponto de quebra são reduzidas por -50 % (uma cópia do cromossomo A está ainda intacta e produzirá sinais normais), enquanto uma concentração exclusiva (isto é, não presente em células normais) de sinais elevados de hibridização é observada para sondas fazem limite com o ponto de quebra no cromossomo B. De fato, a transição abrupta entre sondas que não mostram sinais fortes de hibridização versus sondas que mostram sinais fortes de hibridização no cromossomo B revela a localização do ponto de quebra no cromossomo B.

## EXEMPLO 4

## Análise dos resultados de tecnologia 4C

A tecnologia 4C foi usada para caracterizar o ambiente genômico da região de controle do locus  $\beta$ -globina de camundongos (LCR), focando em um fragmento de restrição 5 contendo seu sítio 2 hipersensível (HS2). A LCR é um forte elemento regulatório de transcrição específico de eritróide exigido para altos níveis de expressão do gene de  $\beta$ -globina. O locus  $\beta$ -globina é presente no cromossomo 7 na posição 97 Mb, onde ele reside em um grande agrupamento de 2,9 Mb dos genes do receptor olfatório que são transcritos apenas 10 em neurônios olfatórios. Interações foram analisadas em dois tecidos: E 14.5 de fígado fetal, onde a LCR é ativa e os genes de  $\beta$ -globinas são altamente transcritos, e E 14.5 de cérebro fetal, onde a LCR é inativa e os genes de globina são silenciosos. Em ambos tecidos, a grande maioria das interações foi encontrada com as sequências no cromossomo 7 e muito poucas interações LCR foram detectadas com seis cromossomo não relacionados (8, 10, 11, 12, 13, 14) (Figura 12a). Os sinais mais fortes no cromossomo 7 foram encontrados 15 dentro de uma região de 5 a 10 Mb centrada em torno da posição cromossômica de  $\beta$ -globina, de acordo com a idéia de que as frequências de interação são inversamente proporcionais à distância (em pares de base) entre sequências de DNA ligadas fisicamente. Não foi possível interpretar quantitativamente as interações nesta região. Inferimos que estas sequências próximas estejam juntas com  $\beta$ -globina de maneira que frequentemente sua grande sobre- 20 representação em nossas amostras de hibridização saturou as sondas correspondentes. Isto foi confirmado quando nós realizamos hibridizações com amostras diluídas 1:10 e 1:100 e observamos que a intensidade de sinal foi reduzida nas sondas fora e na borda, mas não dentro desta região (dados não apresentados).

O procedimento 4C produziu altos dados reproduzíveis. A Figura 2b-c razões não 25 processadas de sinais 4C sob sinais de hibridização de controle para duas regiões 1.5 Mb no cromossomo 7, aproximadamente 25 Mb e 80 Mb longe do gene de  $\beta$ -globina. Neste nível de resolução, os resultados das amostras independentemente processadas foram praticamente idênticos. Tanto no fígado fetal quanto no cérebro, agrupamentos de sinais positivos foram identificados no cromossomo 7, frequentemente em dezenas de localização cromossômica de megabases longe da  $\beta$ -globina. Estes agrupamentos tipicamente consistem 30 em minimamente 20 a 50 sondas com maiores razões de sinal justapostos no molde do cromossomo (Figura 12b-c). Cada sonda no arranjo analisa um evento de ligação independente. Além disso, apenas duas cópias do fragmento de restrição HS2 estão presentes por célula, cada qual que pode apenas ligar a um outro fragmento de restrição. Portanto, a detecção de eventos de ligação independente com 20 ou mais fragmentos de restrição vizinhos indica fortemente que o locus correspondente fazem contato com a LCR  $\beta$ -globina em 35 múltiplas células.

Para determinar significância estatística destes agrupamentos, dados de experimentos individuais foram ordenados em mapas cromossômicos e analisados usando um algoritmo de média corrida com um tamanho de janela de aproximadamente 60 kb. A distribuição da média corrida de dados embaralhados aleatoriamente foi usada para ajustar um valor do patamar, permitindo uma taxa de falsa descoberta de 5 %. Esta análise identificou 66 grupos no fígado fetal e 45 em cérebro cuja reproduzibilidade foi encontrada em experimentos duplicados (Figura 12d-f). Realmente, FISH de alta resolução confirmou que tais agrupamentos representam verdadeiramente os loci que interagem frequentemente (ver a seguir).

Assim, a tecnologia 4C identifica loci interagentes de longo alcance pela detecção de eventos de ligação independentes com múltiplos fragmentos de restrição agrupados em uma posição cromossômica.

Uma série completamente independente de experimentos 4C foi realizada com um conjunto de iniciador de PCR inversa diferente que investigou o ambiente genômico do gene  $\beta$  principal, localizado  $\sim$  50 kb à jusante de HS2. No fígado fetal, o gene  $\beta$  principal é altamente transcrito e frequentemente colocado em contato pela LCR. Agrupamentos praticamente idênticos de interações de longo alcance com  $\beta$  principal como com HS2 foram encontrados, tanto no fígado fetal quanto em cérebro, substanciando ainda mais que estes loci frequentemente fazem contato com o locus  $\beta$ -globina (Figura 16).

#### 20 EXEMPLO 5

*O locus  $\beta$ -globina ativo e inativo ocupa ambientes genômicos distintos.*

Uma comparação entre os dois tecidos revelou que o locus  $\beta$ -globinaativamente transcrito no fígado fetal interage com um conjunto de loci completamente diferente da sua contraparte transcrecionalmente silenciosa no cérebro ( $r=-0,03$ ; correlação de triagem de Spearman) (Figura 12f). Isto excluiu que os resultados tenham sido influenciados pelas composições da sequência das sondas. No fígado fetal, os segmentos de DNA interagentes foram localizados dentro de uma região 70 Mb centrada em torno do locus  $\beta$ -globina, com a maioria (40/66) localizada em direção ao telômero do cromossomo 7. Em cérebro fetal, os loci interagentes foram encontrados em similar ou mesmo grandes distâncias da  $\beta$ -globina comparada ao fígado fetal e com a grande maioria de interações (43/45) localizadas em direção ao centrômero do cromossomo 7. Estes dados demonstraram que o locus  $\beta$ -globina ativo e inativo faz contato com diferentes partes do cromossomo 7.

Seis outros cromossomos (8, 10, 11, 12, 13 e 14) foram representados nos microarranjos. Sinais de hibridização fortes nestes cromossomos foram raros, apareceram tipicamente isolado sem o molde de DNA linear e frequentemente foram ausentes dos experimentos duplicados. Também, níveis de média corrida através desses cromossomos nunca a reproduzibilidade chegou próximo aos níveis pontuados para cromossomo 7 (Figura 17).

Assim, nossos dados mostraram que o locus  $\beta$ -globina principalmente faz contato com os loci em outro lugar no mesmo cromossomo, de acordo com a localização preferida deste locus dentro do seu próprio território do cromossomo. Notamos que o locus  $\alpha$ -globina esteve também presente no arranjo (cromossomo 11) e não pontuou positivo para a interação com 5  $\beta$ -globina, de acordo com a recente demonstração pelo FISH que globina  $\alpha$  e  $\beta$  de camundongos não se encontram frequentemente no espaço nuclear (Brown, J. M. *et al.* (2006) *J Cell Biol* 172, 177-87).

A fim de entender melhor a relevância das interações de longo alcance observadas no cromossomo 7, comparamos os loci interagentes com as posições cromossômicas dos 10 genes. Além do mais, análise de arranjo de expressão Affymetrix foi realizada para determinar a atividade de transcrição nestas posições nos dois tecidos. Embora o tamanho médio de áreas interagentes no fígado fetal e cérebro tenha sido comparável (183 kb e 159 kb, respectivamente), diferenças drásticas foram observadas em seu conteúdo e atividade do 15 gene. No fígado fetal, 80 % dos loci interagentes de  $\beta$ -globina continham um ou mais genesativamente transcritos, embora no cérebro fetal a grande maioria (87 %) não tenha mostrado nenhuma atividade de gene detectável (Figura 14). Assim, o locus  $\beta$ -globina é embutido em 20 um ambiente genômico muito diferente nos dois tecidos. No cérebro, onde o locus não é ativo, ele basicamente faz contato com o loci silencioso transcripcional localizado em direção ao centrômero do cromossomo 7. No fígado fetal, onde o locus é altamente ativo, ele interage preferencialmente com regiõesativamente transcritas localizadas mais proeminentemente em direção ao lado telomérico do cromossomo 7. De forma muito importante, a tecnologia 25 4C identificou tanto Uros quanto Eraf, (~ 30 Mb longe da  $\beta$ -globina) como genes interagente com o locus  $\beta$ -globina ativo no fígado fetal, de acordo com observações anteriores feitas por FISH (Osborne, C. S. *et al.* (2004) *Nat Genet* 36, 1065-71 (2004)). Interessantemente, em 30 cérebro foram observados dois outros genes do receptor olfatório em contato com os agrupamentos presentes no cromossomo 7 que foram localizados em cada lado de, e 17 e 37 Mb longe da  $\beta$ -globina.

Nem todas as regiões transcritas no cromossomo 7 interagem com o locus  $\beta$ -globina ativo no fígado fetal. Portanto, pesquisamos um denominador compartilhado exclusivamente pelos loci interagentes mas não por outras regiões ativas no fígado fetal. Os genes de  $\beta$ -globina, Uros e Eraf são todos genes específicos de eritróide que podem ser regulados pelo mesmo conjunto de fatores de transcrição, e é uma idéia atrativa que esses fatores coordenem a expressão de seus genes alvos no espaço nuclear. Comparados os dados do arranjo de expressão Affymetrix de E 14.5 fígado fetal com o do cérebro fetal para identificar 35 genes expressos preferencialmente (>5-vezes mais) no fígado fetal. Como tal, 28 % dos genes ativos no cromossomo 7 foram classificados como "específico de fígado fetal", dos quais 25 % foram encontrados em uma área colocalizada. Assim, não observamos nenhum

enriquecimento de genes "específico de fígado fetal" nas áreas colocalizadas. De forma mais muito importante, 49 de 66 (74 %) regiões interagentes não contiveram um "específico de fígado fetal" e conclui-se portanto que nossos dados não mostraram nenhuma evidência para expressão coordenada de genes específicos do tecido no espaço nuclear. Os genes de 5  $\beta$ -globina são transcritos em taxas excepcionalmente altas e pergunta-se em seguida se o locus preferencialmente interagiu com outras regiões de alta atividade transcrecional, sendo tanto genes altamente expressos quanto áreas com uma alta densidade de genes ativos. Usando contagem Affymetrix como uma medida para a atividade do gene, realizamos um 10 algoritmo de soma corrida para medir a atividade transcrecional geral dentro de 200 kb regiões em torno de genes ativamente transcritos. Esta análise revelou que a atividade transcrecional em torno de genes interagentes não foi maior que em torno dos genes não interagentes ativos no cromossomo 7 ( $p = 0,9867$ ; soma da triagem 4C Wilcoxon).

#### EXEMPLO 6

15 *O ambiente genômico de um gene de manutenção é amplamente conservado entre os tecidos*

Investigou-se em seguida se um gene que é expresso similarmente em ambos tecidos também muda seu ambiente genômico. Rad23A é um gene onipresentemente expresso que reside em um agrupamento denso de gene de principalmente de genes manutenção no 20 cromossomo 8. Tanto no E14.5 do fígado fetal quanto no cérebro, este gene e muitos de seus vizinhos diretos são ativo. A análise 4C foi realizada e identificou muitas interações de longo alcance com os loci até 70 Mb longe de Rad23A. De forma muito importante, interações com Rad23A foram altamente correlacionadas entre o fígado fetal e cérebro ( $r=0,73$ ; correlação de triagem de Spearman) (Figura 14a). Uma marca registrada compartilhada desses loci foi novamente que eles continham genes ativamente transcritos. Assim, em ambos tecidos aproximadamente 70 % continham pelo menos um gene ativo (Figura 14b-c). Regiões em torno de genes interagentes apresentaram níveis estatisticamente mais significativos de atividade do gene comparados com genes ativos em outro lugar no cromossomo, da maneira determinada por um algoritmo de soma corrida ( $p < 0,001$  para ambos tecidos). Assim, diferente do locus  $\beta$ -globina, o gene Rad23A que é localizado em uma região rica em 30 gene interage preferencialmente a distância com outras regiões cromossômica de maior atividade transcrecional. Observou-se que FISH que a área cromossômica contendo Rad23A reside principalmente na borda (90 %) ou fora (10 %) de seu território de cromossomo (unpublished, D. Noordermeer, M. Branco, A. Pombo and W. de Laat). Entretanto, a análise 4C apenas revelou interações intracromossômicas e nenhuma área na reprodutibilidade do 35 cromossomo 7, 10, 11, 12, 13 ou 14 atende nosso critério rigoroso para interação. Assim, Rad23A é principalmente envolvido em interações intra-cromossômica que são similares em dois tecidos muito diferentes. Se Rad23A tem loci vizinhos preferidos neste cromossomo

não relacionado, ele não interage frequentemente o bastante para ser detectado nas condições usadas aqui para a tecnologia 4C.

#### EXEMPLO 7

##### *Validação de tecnologia 4C por microscopia de alta resolução*

5 Para validar os resultados obtidos pela tecnologia 4C, experimentos cryo-FISH foram realizados. Cryo-FISH é uma técnica de microscopia recentemente desenvolvida, que tem a vantagem sobre protocolos 3D-FISH atuais que preservam melhor a estrutura ultra nuclear embora ofereça melhor resolução no eixo z pela preparação de crio-seções ultrafinas (Branco, M. R. & Pombo, A (2006). PLoS Biol 4, e138). Dados 4C foram verificados medindo com que frequência alelos  $\beta$ -globina ou Rad23A (sempre  $n > 250$ ) são colocalizados com mais que 15 regiões cromossômica selecionadas em 200 nm de seções ultrafinas preparadas de E14.5 de fígado e cérebro. De forma muito importante, todas as frequências de interação medidas por cryo-FISH foram em perfeito acordo com os resultados 4C (Figura 16). Por exemplo, regiões distantes que foram identificadas para interagir com  $\beta$ -globina 10 pela tecnologia 4C co-localizaram mais frequentemente aquelas áreas intervenientes não detectadas por 4C (7,4 % e 9,7 %, versus 3,6 % e 3,5 %, respectivamente). Também, os 15 dois agrupamentos de gene do receptor olfatório distantes identificados pela tecnologia 4C para interagir com  $\beta$ -globina no cérebro fetal mas não no fígado pontuou frequências de co-localização respectivamente de 12,9 % e 7 % nas seções no cérebro, versus 3,6 % e 1,9 % 20 no fígado. Em resumo, frequências de colocalização medidas por loci positivamente identificados pela tecnologia 4C foram todas significativamente maiores que as frequências medidas para loci de fundo ( $p < 0,05$ ; teste G). Concluímos que a tecnologia 4C identificou fielmente loci interagentes de DNA. Finalmente, usamos cryo-FISH para demonstrar que loci 25 identificados para interagir com  $\beta$ -globina também frequentemente fazem contato um com o outro. Isto foi verdadeiro para duas regiões ativas separadas por grande distância cromossômica no fígado fetal (Figura 17) bem como para dois agrupamentos de genes OR inativos bem separados no cromossomo do cérebro (Figura 16). Interessantemente, contatos frequentes entre estes dois agrupamentos de genes OR distantes foram também encontrados 30 no fígado fetal, onde eles não interagiram com o agrupamento de gene OR que continha o locus  $\beta$ -globina ativamente transcrito. Estes dados indicaram que interações nucleares entre agrupamento de genes OR distintos não foram uma peculiaridade do tecido do cérebro fetal analisado. Tenta-se conjecturar que tais contatos espaciais facilitam a comunicação entre os muitos genes OR exigidos para assegurar que apenas um alelo simples seja transcritos por neurônio olfatório (Shykind, B. (2005) *Hum Mol Genet* 14 Spec No 1, R33-9.

#### EXEMPLO 8

##### *Organização nuclear de domínios cromatina ativos e inativos*

As observações aqui descritas demonstram que não apenas regiões genômicas ati-

vas, mas também inativas formam regiões distintas no espaço nuclear que envolvem muitos contatos de longo alcance, sugerindo fortemente que cada segmento de DNA tenha seu próprio conjunto de interações preferido. Nossos dados sugerem que quando o locus  $\beta$ -globina é mudado, ele abandona um ambiente genômico silencioso transcrecional e entra em 5 uma área nuclear onde interações com domínios ativos são favorecidos. Antecipou-se que um reposicionamento drástico como esse mediante ativação transcrecional pode bem ser uma marca registrada apenas de genes específicos de tecido que atingem um certo nível de expressão e, de forma muito mais importante, ficam isolados dos outros genes ativos no molde do cromossomo linear, como é o caso de  $\beta$ -globina. Propõe-se que a rede extensiva 10 de interações de longo alcance que são identificadas tanto entre loci genômico inativo quanto entre ativo, reflete diferenças célula-a-célula nas conformações do cromossomo mais que sendo uma consequência de movimentos dinâmicos durante a interfase (Chakalova *et al.* (2005) *Nat Rev Genet* 6, 669-77 (2005). Presumivelmente, diferentes graus de descondensação 15 após a divisão celular expulsou as regiões genômicas ativas da cromatina inativa (Gilbert, N. *et al.* (2004) *Célula* 118, 555-66 (2004)) e contatos entre loci distantes de composições de cromatina similares são estabelecidas principalmente através das afinidades 20 entre proteínas ligadas a cromatina. Justaposição espacial entre os loci distantes pode ser funcional, mas pode também ser simplesmente a consequência do desdobramento padrão de um cromossomo. Embora loci individuais possam mover-se em um volume nuclear restrito, a conformação geral de um cromossomo seria mantida amplamente por todo o ciclo celular 25 e exigiria divisão celular para restabelecimento. Esta idéia está de acordo com estudos de imageamento de célula viva mostrando movimento restrito de loci de DNA marcado no interior nuclear (Chubb *et al.* (2002) *Curr Biol* 12, 439-45 (2002)) e ajusta bem com estudos mostrando que informação de posição de cromatina nuclear é frequentemente propagada 30 durante a divisão celular sem ser conservada na população de células (Essers, J. *et al.* *Mol Biol Cell* 16, 769-75 (2005); Gerlich, D. *et al.* *Cell* 112, 751-64 (2003)).

#### EXEMPLO 9

Demonstração do princípio: *tecnologia 4C detecta precisamente deleções em uma amostra do paciente (Figura 19)*

30 A presença de uma deleção presente em um paciente com leucemia da maneira revelada por 4C usando uma sequência de nucleotídeos alvo que tanto é em 2 Mb (A) quanto em 1,3 Mb (B) à montante ('a esquerda') do primeiro ponto de quebra. Note que as deleções causam uma redução de sinais de interação de DNA na região deletada, mas também causam um aumento em frequências de interação DNA:DNA para sequências diretamente à 35 jusante ('a direita') do último ponto de quebra. Isto é particularmente óbvio quando interações com sequência de nucleotídeos alvo B são proximamente examinadas (ver gráficos de duas bases). Com base em dados 4C iniciadores foram desenhados em cada lado da região

deletada e o ponto de quebra foi identificado por sequenciamento: texto comum é sequência à montante de deleção, em negrito está indicado um nucleotídeo inserido, grifado é a sequência à jusante da deleção.

#### EXEMPLO 10

5 *Demonstração do princípio: a tecnologia 4C detecta precisamente uma translocação equilibrada em uma amostra do paciente (Figura 25).*

Demonstração de princípios para a detecção de translocações equilibradas. Detecção de translocação t(1;7) da maneira descrita em (R. Burnett *et al.*, Blood, Vol 84, No 4 (August 15), 1994: pp 1232-1236). Sequências de nucleotídeos alvos flanqueiam o locus *TCRB* no cromossomo 7, com os sinais vermelhos representando interações DNA:DNA com a sequência alvo que é localizada à montante do locus *TCRB*, e os sinais azuis representando interações DNA:DNA com a sequência alvo que é localizada à jusante do locus *TCRB*. São descritos os sinais de DNA interagentes encontrados no cromossomo 1. Painel superior mostra a distribuição de sinal teórico. O painel médio e inferior mostra a distribuição de sinal real. O painel inferior mostra os sinais em uma resolução de sondas justapostas individuais no molde do cromossomo. Note que, no caso de uma translocação equilibrada, sequências de nucleotídeos alvos que flanqueiam o ponto de quebra apresentarão um conjunto mutuamente exclusivo de sinais de DNA de interação intercromossômicas que limitam diretamente um do outro no molde do cromossomo linear do cromossomo parceiro de translocação. A posição de ponto de quebra sequenciado (descrita em Burnett *et al.*, 1994) é indicada por uma seta no painel inferior.

#### EXEMPLO 11

5 *Demonstração do princípio: tecnologia 4C detecta precisamente uma translocação não equilibrada em uma amostra do paciente (Figura 27).*

25 Detecção de translocações não equilibradas. Detecção de t(4;7) translocação da maneira descrita em (RJ Galjaard *et al.*, Am J Med Genet Um. 2003 Aug 30;121(2):168-73). Sequências de nucleotídeos alvos localizada no cromossomo 7; os sinais de DNA interagentes descritos são localizados no cromossomo 4. Duas sequências alvos foram usadas localizadas à montante (5') e à jusante (3') do ponto de quebra no cromossomo 7. Sinais de DNA interagentes localizados no cromossomo 4 são indicados (para ambas sequências alvos em azul). A região entre os agrupamentos de fragmentos de DNA interagentes no cromossomo 4 foi deletada neste paciente. Superior: sinais para o cromossomo 4 completo. Dados 4C do painel inferior: sinais em uma região 11.5 Mb em torno dos pontos de quebra no cromossomo 4. Com base nesses dados 4C, o fragmento de restrição *Hind*III no cromossomo 4 contendo o ponto de quebra da translocação foi identificado e usado para mapear o ponto de quebra por sequenciamento. A sequência é provida no fundo da figura, onde a sequência sublinhada é do cromossomo 4, em negrito é encontrada tanto no 7 quanto no 4 e sequên-

cia comum é do cromossomo 7.

### EXEMPLO 12

*Identificação de alta resolução rápida de rearranjos genômicos equilibrados pela tecnologia 4C*

5

#### Resumo

Técnicas atuais para estudar na variação genética, não conseguem identificar precisamente rearranjos cromossômicos equilibrados (inversões, translocações), ou não conseguem identificá-los que precisamente, que frequentemente ocorrem na população humana e pode causar doença. Aqui nós demonstramos que tecnologia 4C detecta inversões e 10 translocações equilibradas, bem como translocações e deleções não equilibradas, em uma resolução (~7 kilobases) permitindo sequenciamento imediato dos pontos de quebra. A tecnologia 4C é usada para caracterizar rearranjos que fundamentam anormalidades congênitas e leucemia. O gene *LMO3* é identificado como um parceiro de translocação inédito do gene  $\beta$  receptor de célula T (*TCRB*) em leucemia linfoblástica aguda de célula T (T-ALL). 15 Estes resultados estabelecem a tecnologia 4C como uma nova e poderosa ferramenta de pesquisa clínica para a análise precisa de rearranjos genômicos, importante para diagnóstico de doença, prognóstico e, finalmente, cuidado ideal com o paciente.

#### Introdução

Rearranjos cromossômicos (deleções, amplificações, inversões, translocações) podem ser a causa de doença, particularmente quando eles afetam o gene de expressão em virtude de ganho ou perda de genes, criação de genes de fusão, ou reposicionamento de elementos de DNA regulatórios de transcrição. Rearranjos que surgem na linhagem germinal pode dar origem a defeitos congênitos, aqueles no tecido somático podem resultar em neoplasia. Após o término dos projetos de sequenciamento do genoma humano, tem-se agora tornado uma tarefa principal caracterizar variantes estruturais no genoma humano, uma vez que tornou-se cada vez mais claro que a diversidade genômica ocorre naturalmente na população humano e pode estar ligada com suscetibilidade a doença (1-6).

Hibridização genômica comparativa com base em microarranjo (arranjo-CGH) é uma abordagem genômica de alta produção amplamente usada que pode detectar amplificações ou deleções cromossômica em uma resolução de poucos kilobases ou ainda menos. CGH baseia-se na medição de mudanças no número de cópia do DNA tais como visto nas deleções ou amplificações e portanto não consegue identificar translocações e inversões que ocorrem sem perda ou ganho de conteúdo de DNA. Desconhece-se com que frequência tais eventos equilibrados ocorrem, mas estima-se que eles constituem até 20 % de todas variações estruturais (7). Atualmente sua detecção depende de grandes abordagens citogenéticas tais como cariotipagem cromossômica, que tem a desvantagem de que eles omitem eventos (aproximadamente 20 %) e fornecem resolução limitada (máximo 5 a 10 megabá-

ses). Isto necessita adicionalmente, de análise de grande mão de obra para identificar a aberração genética real que fundamenta a doença. Aqui, nós demonstramos que tecnologia de Captura Conformação de Cromatina no Chipe (4C) (8) identifica tanto rearranjos genômicos equilibrados quanto não equilibrados em uma resolução (~ 7 kb, ver a seguir) que permite clonagem e sequenciamento imediatos dos pontos de quebra cromossômicos. De forma muito importante, esta estratégia de alta resolução exige o uso de um microarranjo simples para triar o genoma total e portanto é barato.

Nossos resultados iniciais com 4C mostraram que qualquer que seja a dobra da cromatina das interações de longo alcance de um dado locus, fragmentos de DNA próximos ao molde do cromossomo linear são sempre capturados mais eficientemente, resultando em sinais de hibridização fortes e frequentemente mesmo saturados para sondas dentro de uma região de pelo menos 5 a 10 megabases em volta da sequência alvo. Isto é de acordo com a idéia de que os segmentos de DNA localizados muito próximos em uma fibra de cromatina flexível interagirão mais frequentemente (9). Fragmentos locais são também capturados muito mais eficientemente do que segmentos megabases longe que frequentemente laçam em direção à sequência alvo (8). Captura aleatória de fragmentos de restrição é rara, as demonstrado pela frequência muito baixa de sinais das sondas localizadas nos cromossomos não relacionados. Assim, a tecnologia 4C permite reconstruir os mapas físicos de moldes do cromossomo em torno de sequências alvos e consequentemente deve também ser capaz de identificar mudanças nesses mapas em decorrência de rearranjos genômicos.

#### *Materiais & Métodos*

##### *Preparação da amostra*

Amostra T-ALL dos pacientes e amostras de célula T controle saudável foram manuseados da maneira previamente descrita (Vlierberghe et al., *Leukemia* 20, 1245 (Jul, 2006); Simonis et al., *Nat Genet* 38, 1348 (Nov, 2006)). A linhagem celular transformada por EBV derivada do paciente PAP foi cultivada e manuseada da maneira antes descrita (Simonis et al., *Nat Genet* 38, 1348 (Nov, 2006); Galjaard et al., *Am J Med Genet A* 121, 168 (Aug 30, 2003)).

##### *Desenho do arranjo 4C*

As sondas de 60 bp foram desenhadas dentro de 100 bp de um sítio *HindIII*, usando critério descrito previamente que, por exemplo, seleciona apenas sequências de DNA exclusivas (Simonis et al., *Nat Genet* 38, 1348 (Nov, 2006)). Para poder cobrir o genoma total com as 400.000 sondas que ajustam no microarranjo NimbleGen, foi feita uma seleção. Os números de sondas foram primeiramente reduzidos mantendo apenas uma sonda por fragmento *HindIII*, em vez de uma em cada lado. Em segundo lugar, as sondas foram selecionadas de maneira que o espaçamento das sondas seja tão possível quanto no genoma.

##### *análise 4C*

análise 4C foi realizada da maneira previamente descrita (2), usando as seguintes sequências iniciadoras:

extremidade 5' de TCRBCATGAAGAACGAGCACCC

CCTTGATGTTCTCCCTTACC

5 extremidade 3' de TCRBTGTCAGGCTCTCCTACAC

GTCGTCCAGAACACTCACC

Centromérico t(4;7)AATCCAGGGCTACTTCCAG CCGTGATGCTATCTGCCA

Telomérico t(4;7) TGTTGGAAGACCAGGTGAAG TGCGTGGAAAGCGAGTG

Deleção 9 CAATCCCAGATACTTCCTCATACAAATACTTCCAAGACTGG AC

10 3' de TCRAGAATATGTTATGCTTGATCC TTCCATGAGAGAAGTCTAG

Os dados 4C foram visualizados usando software SignalMap. Para criar uma vista do cromossomo total dos dados 4C, uma média corrida com um tamanho de janela de 29 sondas foi calculada usando o pacote R (<http://www.r-project.org>).xxx

Extremidade pareada de fragmento de restrição-sequenciamento

15 10 µg de DNA genômico foi primeiramente digerido em 500 µL com 10 U de uma enzima que reconhece 6 bases (*Hind*III, *Bgl*III ou *Eco*RI) (37 °C por 2 horas). As amostras foram purificadas pela extração de fenol-clorofórmio e precipitação de etanol. Subsequentemente, as amostras foram ligadas em 2 mL com 40U de ligase (Roche) por 4 horas a 16 °C e 30 minutos a 20 °C.

20 As amostras ligadas foram purificadas por extração de fenol-clorofórmio e precipitação de etanol.

Uma segunda digestão foi realizada com uma enzima de restrição que reconhece 4 bases (por exemplo, *Nla*III ou *Dpn*II) nas mesmas condições da maneira descrita para a enzima de reconhecimento 6-base. Ligação subsequente foi também da maneira descrita anteriormente. As amostras foram purificadas por extração de fenol-clorofórmio e precipitação de etanol. Fragmentos selecionados foram PCR amplificada de 50 a 100 ng de DNA, usando as seguintes condições: 94 °C por 3 minutos, seguido por 30 ciclos de 15 segundos a 94 °C, 1 minuto a 55 °C e 2 minutos a 72 °C e uma etapa final de 7 minutos a 72 °C.

### Resultados

30 *Identificação dos pontos de quebra de translocação*

Para este teste, a tecnologia 4C foi primeiramente aplicada à linhagem celular HSB-2 T-ALL, contendo um translocação recíproca t(1;7)(p35;q35) entre o locus do receptor β da célula T (*TCRB*) em 7q35 e o locus *LCK* em 1p35 (10). Dois experimentos 4C independentes foram realizados, cada qual analisando interações com um fragmento de restrição diferente no cromossomo 7 localizado tanto no lado do ponto de quebra no locus *TCRB* quanto a uma distância de 462 kb e 239 kb, respectivamente. Em ambos casos, foram observados sinais de hibridização fortes em torno do locus *TCRB* no cromossomo 7 em uma amostra controle

saudável e a amostra HSB-2. A amostra controle não apresentou sinal nem sinal de fundo em todos os outros cromossomos (figura 30A; fig. 33). Ao contrário, a amostra HSB-2 apresentou sinais especificamente muito altos adicionais em uma região megabase em 1p35 (figura 30B). Estes sinais representaram os fragmentos de restrição no cromossomo 1 capturado pelos dois fragmentos no cromossomo 7 em HSB-2, indicando que partes desses cromossomos foram colocadas em proximidade física imediata. A maioria das sequências alvos de *TCRB* teloméricos no fragmento de restrições capturado do cromossomo 7 em direção ao lado centromérico do gene *LCK* no cromossomo 1. Ao contrário, a maioria dos fragmentos capturados da sequência alvo de *TCRB* centromérico em direção ao lado telomérico de *LCK*. Isto é de acordo com a orientação da translocação t(1;7). Além disso, os primeiros fragmentos de restrição capturados no cromossomo 1 em ambos experimentos flanqueiam diretamente o ponto de quebra cromossômico previamente identificado. Assim, 4C localiza pontos de quebra de translocação em uma posição entre o par de sondas que representa a transição de fragmentos de restrição não capturados para capturado. Neste caso, os dois experimentos 4C analisando as interações com sequências alvos de cromossomo 7 identifica cada qual o ponto de quebra no cromossomo 1 envolvido na translocação equilibrada t(1;7) dentro de 27 kb.

#### *Identificação de inversões*

Em seguida, testamos se 4C pode também identificar inversões. Aplicamos a tecnologia 4C a um amostra T-ALL pediátrica do paciente que, com base em estudos de expressão FISH e microarranjo, foi suspeito de carregar uma inversão no cromossomo 7, inv(7) (p15q35). Esta anormalidade leva a rearranjo do locus *TCRB* no agrupamento do gene *HOXA*, como foi previamente descrito para outros pacientes (11, 12). Novamente dois experimentos foram realizados, usando o mesmo conjunto de fragmento alvos *TCRB* que identificou a translocação t(1;7) descrita anteriormente. As duas sequências alvos capturaram eficientemente muitos fragmentos cobrindo megabases de DNA no outro lado do cromossomo 7 em torno do locus *HOXA* na amostra do paciente 4 apenas (figura. 30C-D). Além disso, cada sequência alvo capturou uma região cromossômica distinta em torno do agrupamento do gene *HOXA*, indicando que os dois fragmentos são ligados em diferentes partes do cromossomo 7. A maioria dos fragmentos alvos *TCRB* 3' capturou a maioria fragmentos *HOXA* 3', ao passo que a maioria dos fragmentos alvos *TCRB* 5' capturou a maioria dos fragmentos *HOXA* 5', revelando assim uma inversão entre os loci. As duas regiões capturadas em torno de *HOXA* flanqueiam diretamente umas nas outras, mostrando que a inversão foi equilibrada e não acompanhada por perda (extensiva) de sequências *HOXA*. As duas sondas que marcam a transição entre fragmento não capturado e capturado revelaram a posição do ponto de quebra, que localizou uma região de 6 kb próxima ao gene *HOXA9* do agrupamento *HOXA* (figura 30D). A confirmação que esta região certamente carrega o ponto de quebra

foi obtida por sequenciamento, usando uma abordagem de sequenciamento da extremidade pareada de fragmento de restrição (fig. 34). Concluímos, portanto, que a tecnologia 4C é a primeira abordagem genômica de alta produção que pode detectar translocações e inversões equilibradas. A resolução fornecida pela tecnologia 4C permite clonagem imediata dos 5 pontos de quebra. A tecnologia 4C é portanto a primeira técnica que pode detectar eventos genéticos equilibrados em uma alta resolução como essa.

*Identificação de translocações não equilibrada*

O potencial de tecnologia 4C foi adicionalmente explorado aplicando-o a uma linhagem celular transformado por EBV derivada de um paciente com Polidactilia pós-axial (PAP). PAP é um distúrbio hereditário dominante autossomal **CARACTERIZADO** por ulnar extra de dedo fibular. As células do paciente foram previamente **CARACTERIZADAS** por cariotipagem e FISH para conter uma translocação não equilibrada entre os cromossomos 4 e 7 com uma microdeleção,  $t(4;7)(p15.2;q35)$ . Entretanto, a resolução limitada de FISH impedia definir a extensão da deleção e as posições exatas dos pontos de quebra (13). Dois 10 experimentos 4C foram realizados, cada qual analisando interações de DNA com um fragmento alvo localizado em um outro lado da parte rearranjada do cromossomo 7, sendo um 15 dos quais 4 megabases longe do ponto de quebra mais próximo (ver a seguir). Em ambos experimentos, os fragmentos genômicos foram capturados não apenas no cromossomo 7, mas também no cromossomo 4, a 4p15.2 (figura 31A; fig.35). Ao contrário do que foi encontrado para a translocação equilibrada, os fragmentos do cromossomo 4 capturados pelas 20 duas sequências alvos não serão diretamente vizinhos um do outro. Um ponto de quebra foi localizado na posição 17,28 Mb (NCBI 36) e o outro foi encontrado na posição 20,08 Mb do cromossomo 4, mostrando que a translocação  $t(4;7)$  foi acompanhada por uma deleção de 2,8 Mb no cromossomo 4. Para verificar que as transições dos fragmentos de restrição de 25 capturados para não capturados de fato marcam as posições dos pontos de quebra, o ponto de quebra que foi menos óbvio dos dados 4C, localizados a 20,08 Mb, foi clonado e sequenciado (figura 31B). Isto confirmou a posição do ponto de quebra no cromossomo 4 a 20,08 Mb dentro do o gene SLIT2 e revelou que ele foi rearranjado com uma sequência intergênica no cromossomo 7 que estava 4 Mb longe da sequência alvo usada para identificar 30 o ponto de quebra. Isto mostra que sequências alvos 4C podem capturar fragmentos de DNA e identificar rearranjos mesmo quando os pontos de quebra estão diversos megabases distantes. Quando a análise 4C é direcionada a ambos os lados de um ponto de quebra genômico, ela pode imediatamente identificar se uma translocação ou inversão é equilibrada 35 ou é acompanhada por rearranjos adicionais tais como uma deleção (isto é, não equilibrada).

*Identificação de deleções*

Em seguida investigamos se a tecnologia 4C pode identificar uma deleção que não

é associada com uma translocação. Para isto, analisamos uma outra amostra de T-ALL pediátrica do paciente que, baseado nos dados CGH do arranjo, conteve uma deleção homozigota do loci p15/p16 em 9p21 de cromossomo 5. O tamanho exato da deleção dos pontos de quebra atuais não foi conhecido. Definimos um fragmento alvo localizado ~2 Mb longe de 5 um dos pontos de quebra estimados. Conforme esperado, foi observada uma região que laqueia os sinais da sonda, demarcando a área deletada (figura 31C). Um aumento nos sinais de hibridização para a região imediatamente à jusante da deleção é observado na amostra do paciente versus a amostra controle saudável. Isto se dá em virtude de a região 10 estar em proximidade muito imediata no molde linear do fragmento alvo em virtude da deleção. Com base nos dados 4C, foram desenhados os iniciadores de PCR que flanqueiam a região ~2Mb deletada que permitiram amplificação nos pontos de quebra; o sequenciamento 15 dos produtos de PCR confirmou as posições dos dois pontos de quebra que flanqueiam a deleção (figura 31D). Concluímos que a tecnologia 4C pode identificar deleções homozigotas. As deleções revelam por si mesmas as regiões contendo menores sinais de hibridização em combinação com mais sequências à jusante que mostraram menores sinais de hibridização.

*Amostra do pacientes não caracterizadas*

A tecnologia 4C foi aplicada para a triagem de amostra T-ALL do paciente não caracterizada pra rearranjos genéticos associados com o locus *TCRB* ou o locus  $\alpha$  do receptor 20 de célula T (TCRA). Em T-ALL, frequentemente ocorrem translocações cromossômicas durante uma tentativa de recombinação VDJ dos loci TCR. As amostras dos cinco pacientes T-ALL, que foram previamente mostradas não carregam qualquer das anormalidades genética 25 recorrentes associadas com T-ALL (dados não apresentados), foram analisada pela tecnologia 4C, usando um *TCRB* próximo a sequência alvo e um *TCRA* próxima. Nenhuma das amostras apresentaram rearranjos com o locus *TCRA* (dados não apresentados), e quatro das cinco amostras do paciente também não apresentaram nenhum rearranjo com *TCRB* 30 (fig. 36), que foi subsequentemente confirmado por FISH. Uma amostra do paciente entretanto, apresentou uma translocação entre *TCRB* e o braço p do cromossomo 12 (figura 32 A; fig. 36). Além do mais, observou-se que o paciente carrega uma grande deleção no cromossomo 12, que foi confirmado por um experimento oligoarray CGH (dados não apresentados). Esta deleção é localizada ~3 Mb distante do ponto de quebra de translocação, novamente mostrando que sequências alvos 4C podem identificar rearranjos em grandes distâncias. A translocação t(7;12)( q35;pl2.3) não foi descrita antes em T-ALL. As duas sondas 35 no cromossomo 12 que marcam a transição entre fragmentos de restrição capturados e não capturados foram 6 kb aparte e localizados logo à jusante apenas do domínio Lim do gene *LMO3*. O sequenciamento da extremidade pareada do fragmento de restrição foi usado para confirmar que estas sondas demarcam a região contendo o ponto de quebra. O sequencia-

mento dos pontos de quebra presentes em ambos cromossomos derivados demonstrou que o cromossomo 12 foi rearranjado sem a perda de um nucleotídeo simples, embora a quebra no cromossomo 7 tenha sido acompanhada por um deleção de quase 400 kb de sequências TCRB (figura 32B). Esta deleção possivelmente se deu em virtude de eventos de deleção 5 associados com a tentativa de recombinação de VDJ do locus TCRB. Ambos os pontos de quebra também contiveram pares intervenientes de base de origem desconhecida (4 e 18 bp, respectivamente), que podem representar os nucleotídeos aleatórios que também são normalmente incorporados durante recombinação de VDJ. Interessantemente, a translocação posiciona o melhorador de TCRB 70 kb à jusante do gene *LMO-3* (figura 32C), que é 10 comparável a sua posição normal com relação ao *TCRB*. Dados de expressão do microarranjo mostraram que embora *LMO3* seja normalmente fora da amostra T-ALL do paciente, o gene é altamente expresso nesta amostra T-ALL (figura 37). Os elementos da família de proteína *LMO-1* e *LMO-2*, mas não *LMO3*, foram previamente considerados como parceiros 15 oncogênicos de translocação do TCR loci em T-ALL. Interessantemente, observou-se recentemente que *LMO3* age como um oncogene em neuroblastoma (14). Assim, a tecnologia 4C aplicada à triagem de amostra do paciente não caracterizada para aberrações genéticas leva a descoberta de uma translocação previamente não detectada e estabeleceu *LMO-3* 20 como um oncogene célula T inédito putativo.

#### Conclusões

Estes dados estabelecem a tecnologia 4C como a primeira abordagem genômica 25 que pode identificar anormalidades genéticas equilibradas tais como translocações e inversões recíprocas. Além do mais, fica claro que a tecnologia 4C pode identificar deleções homozigotas e deleções associadas com translocações. A tecnologia 4C pode também identificar eventos heterozigotos não equilibrados com base em mudanças no número de cópias 30 de DNA capturado. Uma vantagem principal da tecnologia 4C sobre abordagens de sequenciamento de extremidade pareada (15) ou mesmo sequenciamento do genoma total (alta produção) é que a identificação de rearranjos equilibrados não se baseia em capturar a sequência fragmento simples que carrega o ponto de quebra; em vez disto, a tecnologia 4C identifica rearranjos equilibrados com base na captura de muitos fragmentos cobrindo diversas 35 megabases no ponto de quebra. Por exemplo, translocações recíprocas são identificadas com base na captura de muitos fragmentos que se localizam em uma parte de um cromossomo por uma sequência alvo localizada em um outro cromossomo. Isto torna a tecnologia mais robusta do que outras abordagens. A resolução de tecnologia 4C é alta e permite a imediata clonagem e sequenciamento dos pontos de quebra genéticos, como foi demonstrado para quatro diferentes rearranjos. Mesmo os pontos de quebra localizados diversos 40 megabases distantes da sequência alvo podem facilmente ser identificados. A resolução é essencialmente idêntica ao tamanho médio dos fragmentos criados pela enzima de restri-

ção. Usando uma enzima que cria menores fragmentos deve ser aumentado ainda mais a resolução desta técnica. A tecnologia 4C pode ser aplicada a todos os tipos celulares tais como células cancerígenas do sangue, tumores sólidos, amniócitos coletados para diagnóstico pré-natal, etc., desde que células suficientes possam ser isoladas (ou cultivadas) contendo DNA intacto. Iniciamos atualmente com cerca de 10 milhões de células, mas esta quantidade pode ser reduzida, uma vez que hibridizamos junções de ligação amplificadas de PCR originando de ~ 0,5 milhões de genomas equivalentes. A tecnologia 4C atual exige triagem de uma sequência alvo e é portanto particularmente adequada para a triagem de rearranjos próximos dos loci frequentemente envolvidos em uma doença, tais como os genes do receptor da célula T em T-ALL ou o receptor de célula B (BCR) loci de cadeia pesada e leve em linfomas humanos. A técnica é também muito usada para o mapeamento fino de rearranjos fracamente caracterizados, por exemplo, de translocações ou inversões que foram encontradas com base em cariotipagem cromossômica.

#### EXEMPLO 13

15 *Sequenciamento Multiplex 4C*

##### Desenho experimental

Este exemplo é com base em uma das diversas tecnologias de sequenciamento de alta produção (Solexa), mas pode ser modificado para ajustar outras plataformas. A fim de usar sequenciamento Solexa (Illumina) par a análise dos resultados 4C, sequenciamento é direcionado às junções de ligação. Portanto, os iniciadores de PCR inversa foram desenhados para cada sequência alvo de maneira que eles estejam próximos aos sítios de reconhecimento de enzima de restrição primária e secundária analisada. Desenhamos aqui, pequenos iniciadores de PCR inversa (18-mers) que cada sobreposição parcialmente ou completamente com os sítios de reconhecimento de enzima de restrição primária (*Hind*III) e secundária (*Dpn*II) analisada. O adaptador Solexa e a sequência de iniciação de sequências foram adicionados como projeções 5' para os iniciadores de PCR inversa (figura 38).

Três sequências alvos (iscas) foram selecionadas em cromossomo 7 humano, respectivamente na posição 85, 105 e 139 Mb. Os conjuntos iniciadores usados para amplificar os fragmentos capturados por cada uma dessas iscas são da maneira a seguir:

30 85 Mb:

iniciador *Dpn*II: atgtgactcctcttagatc

iniciador *Dpn*II com adaptador: aatgatacggcgaccaccgaacactcttccctacacgacgcttccgatct- atgtgactcctcttagatc

iniciador *Hind*III: ccctgaacctttaagct

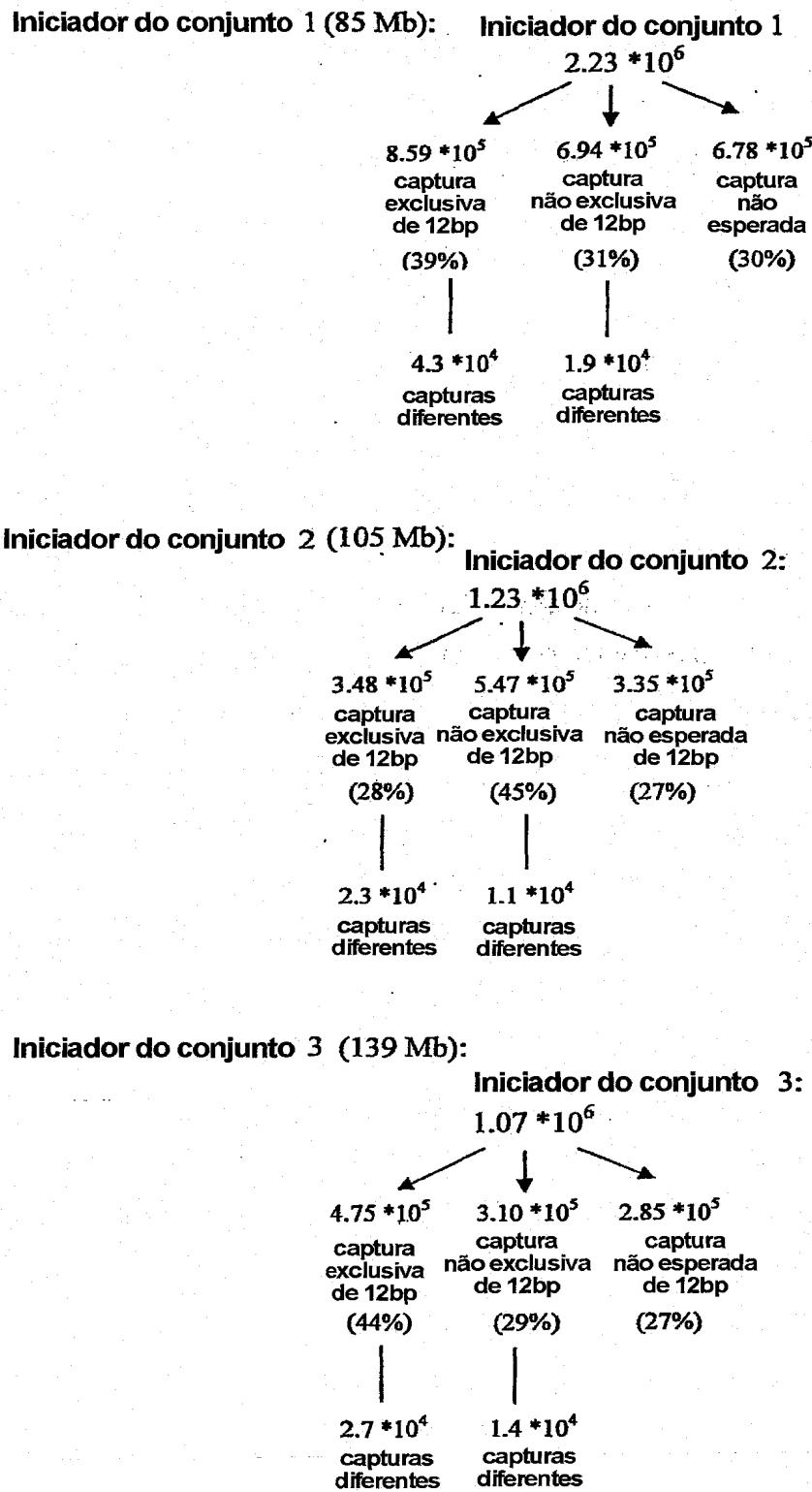
35

iniciador *Hind*III com adaptador: caagcagaagacggcatacga-ccctgaacctttaagct

105 Mb:

Iniciador *Dpn*II: cggcctccaattgtgatc

iniciador *DpnII* com adaptador:  
 aatgatacggcgaccaccgaacactcttcctacacgacgcttcgatct- cggcctccaattgtgatc iniciador *HindIII*-  
 I: gaattgctttggtaagctt  
 iniciador *HindIII* com adaptador: caagcagaagacggcatacga-gaattgctttggtaagctt  
 5 139 Mb:  
 Iniciador *DpnII*: ttttagccctgacagatc  
 iniciador *DpnII* com adaptador:  
 aatgatacggcgaccaccgaacactcttcctacacgacgcttcgatct-ttttagccctgacagatc iniciador *HindIII*-  
 agtcaaacataaggctaagg  
 10 iniciador *HindIII*- com adaptadores: caagcagaagacggcatacga-agtcaaacataaggctaagg  
 Cada conjunto de iniciador (com adaptadores) foi usado em uma reação separada  
 de PCR; 3 reações de PCR (usando 200 ng de molde por reação) foram realizadas em condições padrões (descritas em Simonis *et al.*, *Nature Methods* 2007, vol.4, 895-901). O molde  
 15 4C foi preparado a partir da linhagem celular HSB-2 T-ALL, contendo uma translocação re-  
 cíproca t(1;7)(p35;q35) entre o locus beta receptor de célula T (*TCRB*) em 7q35 e o locus  
*LCK* em 1p35 (Burnett, R. C, *et al.*, *Blood* 84, 1232-6 (1994)). Os produtos de PCR são mos-  
 trados na Figura 39.  
 Subsequentemente, as reações de PCR foram agrupadas por conjunto de iniciador  
 e purificadas em colunas em branco Amersham. A concentração de DNA foi medida e quan-  
 20 tidades iguais de material amplificado por cada conjunto de iniciador foram misturadas. Esta  
 mistura foi analisada por sequenciamento Solexa.  
Resultados de sequenciamento Solexa  
 Número total de leituras (uma raia):  $4.9 \times 10^6$  sequências.  
 25 93 % das sequências começam com uma das sequências iniciadoras esperada.  
 Durante análise dos primeiros 12 pares de base (bp) da junção de ligação *DpnII* e  
 comparando estes 12 pares de base com uma base de dados local contendo todos os frag-  
 mentos genômicos de 12 bp que flanqueiam os sítios *DpnII* relevantes no genoma (isto é,  
 aqueles diretamente adjacentes a um sítio *HindIII*) descobrimos que: 37 % contêm captura  
 30 exclusivo de 12bp, 34 % contêm captura não exclusivo de 12bp, e 29 % contêm captura não  
 esperado de 12 bp.  
 De todas as sequências  $4.9 \times 10^6$ : 34 % contiveram iniciador + captura exclusivo,  
 32% contiveram iniciador + captura não exclusivo. No total portanto 66 % contiveram uma  
 primeira sequência esperada de 30 bp.  
 Por conjunto de iniciador o número total de sequências e a natureza dessas se-  
 35 quências são da maneira a seguir:



Quando sequências capturadas são colocadas em gráfico nas suas posições cromossômicas, os dados são mostrados nas Figuras 40 e 41.

Em resumo, os dados fornecem uma demonstração do princípio para sequencia-

mento multiplex-4C. Para cada sequência, a isca e a sequência capturada podem ser identificadas. A técnica pode ser adicionalmente melhorada:

1. O sequenciamento pode ser mais bem direcionado para o lado *Hind*III (em vez de o lado *Dpn*II); isto impede que eventos de ligação aleatórios a leitura ocorram durante a

5 segunda etapa de ligação.

2. O uso de menos cortadores de frequente (7 ou 8 cortadores em vez de 6 cortadores) aumentaria a distância genômica capturada (coberta) por cada isca.

3. O uso de uma enzima de restrição que não digere o DNA repetitivo (ou é relativamente sub-representado em DNA repetitivo) aumentaria o número de leituras interpretáveis.

#### Aspectos Adicionais 1

1. Um método para analisar a frequência de interação de uma sequência de nucleotídeos alvo com uma ou mais sequências de nucleotídeos (por exemplo, um ou mais loci genômicos) compreendendo o uso de uma sequência de nucleotídeos ou um arranjo de sondas ou um conjunto de sondas ou um arranjo da maneira aqui descrita.

2. Um método para identificar uma ou mais interações DNA-DNA que são indicativos de uma condição da doença particular ou estado do portador compreendendo o uso de uma sequência de nucleotídeos ou um arranjo de sondas ou um conjunto de sondas ou um arranjo da maneira aqui descrita.

20 3. Um método de diagnóstico ou prognóstico de uma doença ou síndrome ou estado do portador causados ou associados a uma mudança em um DNA-DNA compreendendo o uso de uma sequência de nucleotídeos ou um arranjo de sondas ou um conjunto de sondas ou um arranjo da maneira aqui descrita.

25 4. Um método do ensaio para identificar um ou mais agentes que modulam uma interação DNA-DNA compreendendo o uso de uma sequência de nucleotídeos ou um arranjo de sondas ou um conjunto de sondas ou um arranjo da maneira aqui descrita.

5. Um método para detectar a localização de um ponto de quebra (por exemplo, uma translocação) compreendendo o uso de uma sequência de nucleotídeos ou um arranjo de sondas ou um conjunto de sondas ou um arranjo da maneira aqui descrita.

30 6. Um método para detectar a localização de uma inversão compreendendo o uso de uma sequência de nucleotídeos ou um arranjo de sondas ou um conjunto de sondas ou um arranjo da maneira aqui descrita.

35 7. Um método para detectar a localização de uma deleção compreendendo o uso de uma sequência de nucleotídeos ou um arranjo de sondas ou um conjunto de sondas ou um arranjo da maneira aqui descrita.

8. Um método para detectar a localização de uma duplicação compreendendo o uso de uma sequência de nucleotídeos ou um arranjo de sondas ou um conjunto de sondas

ou um arranjo da maneira aqui descrita.

9. Um método para analisar a frequência de interação de uma sequência de nucleotídeos alvo com uma ou mais sequências de nucleotídeos de interesse (por exemplo, um ou mais loci genômicos) compreendendo a etapa de:

- 5 (a) fornecer uma amostra de DNA reticulado;  
(b) digerir o DNA reticulado com uma enzima de restrição primária;  
(c) ligar as sequência de nucleotídeos reticuladas  
(d) reverter a reticulação;  
(e) digerir opcionalmente as sequências de nucleotídeos com uma enzima de restrição secundária;

10 (f) ligar opcionalmente a sequência de nucleotídeos;  
(g) amplificar uma ou mais sequências de nucleotídeos de interesse que são ligadas à sequência de nucleotídeos alvo usando pelo menos dois oligonucleotídeo iniciadores, em que cada iniciador hibridiza em uma sequência de DNA conhecida que flanqueia aa sequências de nucleotídeos de interesse;

15 (h) hibridizar a(s) sequência(s) amplificada(s) em um arranjo ou sequenciar as sequências amplificadas; e  
(i) determinar a frequência de interação entre as sequências de DNA.

## Aspectos Adicionais 2

20 Ainda adicionalmente os aspectos da presente invenção são apresentados a seguir nos parágrafos numerados.

1. Um sequência de nucleotídeos circularizada compreendendo uma primeira e uma segunda sequência de nucleotídeos separadas pelos sítios de reconhecimento enzima de restrição primária, em que a dita primeira sequência de nucleotídeos é uma sequência de nucleotídeos alvo e a dita segunda sequência de nucleotídeos é obtêivel pela reticulação do DNA genômico.

2. A sequência de nucleotídeos circularizada de acordo com o parágrafo 1, em que a sequência de nucleotídeos alvo é selecionada do agrupamento que consiste em um promotor, um melhorador, um silenciador, um isolante, uma região de anexação de matriz, uma região de controle dos locus, uma unidade de transcrição, uma origem de replicação, um ponto quente de recombinação, um ponto de quebra de translocação, um centrômero, um telômero, um região densa em gene, uma região fraca em gene, um elemento repetitivo e um sítio integração (viral).

35 3. A sequência de nucleotídeos circularizada de acordo com o parágrafo 1, em que  
a sequência de nucleotídeos alvo é uma sequência de nucleotídeos que é associada com,  
ou causa uma doença, ou é localizada menos que 15 Mb em um molde de DNA linear de  
um locus que é associado com, ou causa um doença.

4. A sequência de nucleotídeos circularizada de acordo com os parágrafos 1-3, em que a sequência de nucleotídeos alvo é selecionada do agrupamento que consiste em: *AML1, MLL, MYC, BCL, BCR, ABL1, IGH, LYL1, TAL1, TAL2, LMO2, TCRα/δ, TCRβ* e *HOX* ou outros loci associados com doença da maneira descrita em "Catalogue of Unbalanced C-hromosome Aberrations in Man" 2nd edition. Albert Schinzel. Berlin: Walter de Gruyter, 2001. ISBN 3-11-011607-3.

5. A sequência de nucleotídeos circularizada de acordo com qualquer dos parágrafos 1 a 4 em que o sítio de reconhecimento de enzima de restrição primária é um sítio de reconhecimento de 6 a 8 bp, preferivelmente selecionado do agrupamento que consiste em *BglIII, HindIII, EcoRI, BamIII, SpeI, PstI* e *NdeI*.xxx

10. 6. A sequência de nucleotídeos circularizada de acordo com qualquer dos parágrafos anteriores, em que o sítio de reconhecimento de enzima de restrição secundária é um sítio de reconhecimento de sequência de nucleotídeos de 4 ou 5 bp.

15. 7. A sequência de nucleotídeos circularizada de acordo com qualquer dos parágrafos anteriores, em que o sítio de reconhecimento de enzima de restrição secundária é localizado em mais que cerca de 350bp do sítio de restrição primária.

8. A sequência de nucleotídeos circularizada de acordo com qualquer dos parágrafos anteriores, em que a sequência de nucleotídeos é marcada.

9. Um método para preparar uma sequência de nucleotídeos circularizada compreendendo a etapa de:

- (a) fornecer uma amostra de DNA reticulado;
- (b) digerir o DNA reticulado com uma enzima de restrição primária;
- (c) ligar o sequências de nucleotídeos reticuladas para circularização;

10. Um método de acordo com o parágrafo 9, em que a sequência de nucleotídeos reticulada é amplificada usando PCR.

25. 11. Um método de acordo com o parágrafo 10, em que a sequência de nucleotídeos reticulada é amplificada usando PCR inversa.

12. Um de acordo com o parágrafo 10 ou parágrafo 11, em que o Sistema de PCR Expand Long Molde (Roche) é usado.

30

Tabela 2

Interação	em 4C	N	% de sobreposição	em Cyro-FISH	valor de P
B-globina – C-hr.7 73,1 Mb	+	258	7,4	+	P < 0,001
B-globina – Chr.7 80,1 Mb (OR)	-	254	3,6	-	

B-globina – C- hr.7 118,3 Mb	-	255	3,5	-	
B-globina – C- hr.7 129,9 Mb (Uros)	+	259	6,6	+	P < 0,001
B-globina – C- hr.7 130,1 Mb	+	413	9,	+	P < 0,001
B-globina – C- hr.7 135,0 Mb	-	261	1,9	-	
B-globina – D7Mit21	x	258	0,4	-	
Chr.7 80,1 Mb – Chr.7 135,0 Mb	x	253	5,9	+	P < 0,05
Chr.7 73,1 Mb – Chr.7 130,1 Mb	x	254	5,5	+	P < 0,05
Rad23A – Chr. 8 21,8 Mb	+	255	5,9	+	P < 0,05
Rad23A – Chr. 8 122,4 Mb	+	261	8	+	P < 0,001

Interação	em 4C	N	% de sobreposi- ção	em Cyro- FISH	valor de P
B-globina – C- hr.7 73,1 Mb	-	256	3,9	-	
B-globina – Chr.7 80,1 Mb (OR)	+	256	12,9	+	P < 0,001
B-globina – C- hr.7 118,3 Mb	-	242	4,1	-	
B-globina – C- hr.7 130,1 Mb	-	263	3	-	
B-globina – C- hr.7 135,0 Mb	+	256	7	+	P < 0,05
B-globina – D7Mit21		258	6,2	+	P < 0,05
Chr.7 80,1 Mb – Chr.7 135,0 Mb		261	5	+	P < 0,1

Rad23A – Chr. 8 21,8 Mb	-	260	3,8	-	
Rad23A – Chr. 8 122,4 Mb	+	258	8,1	+	P < 0,001

### Referências

Blanton J, Gaszner M, Schedl P. 2003. Protein:protein interactions and the pairing of boundary elements in vivo. *Genes Dev* 17:664-75.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. 2002. Capturing chromosome conformation. *Science* 295: 1306-11.

Driissen R5 Palstra RJ, Gillemans N, Splinter E, Grosveld F, Philipsen S, de Laat W. 2004. The active spatial organization of the beta-globin locus requires the transcription factor EKLF. *Genes Dev* 18:2485-90.

Horike S, Cai S, Miyano M, Cheng JF, Kohwi-Shigematsu T. 2005. Loss of silent-chromatin looping and impaired imprinting of DLX5 in Rett syndrome. *Nat Genet* 37:31-40.

Murrell A, Heeson S, Reik W. 2004. Interaction between differentially methylated regions partitions the imprinted genes Igf2 and H19 into parent-specific chromatin loops. *Nat Genet* 36:889-93.

Palstra, R.J., Tolhuis, B., Splinter, E., Nijmeijer, R., Grosveld, F., and de Laat, W. 2003. The beta-globin nuclear compartment in development and erythroid differentiation. *Nat Genet* 35: 190-4.

Patrinos, G.P., de Krom, M., de Boer, E., Langeveld, A., Imam, A.M.A., Strouboulis, J., de Laat, W., and Grosveld, F.G. (2004). Multiple interactions between regulatory regions are required to stabilize an active chromatin hub. *Genes & Dev.* 18: 1495- 1509.

Spilianakis CG, Flavell RA. 2004. Long-range intrachromosomal interactions in the T helper type 2 cytokine locus. *Nat Immunol* 5:1017-27.

Tolhuis, B., Palstra, R.J., Splinter, E., Grosveld, F., and de Laat, W. 2002. Looping and interaction between hypersensitive sites in the active beta-globin locus. *Molecular Cell* 10: 1453-65.

Vakoc CR5 Letting DL, Gheldof N, Sawado T, Bender MA, Groudine M, Weiss MJ, Dekker J, Blobel GA. 2005. Proximity among distant regulatory elements at the beta-globin locus requires GATA-I and FOG-I. *Mol Cell*. 17:453-62

### **REFERÊNCIAS PARA O EXEMPLO 12**

1. Nature 447, 661 (Jun 7, 2007).

2. D. F. Easton et al, Nature 447, 1087 (Jun 28, 2007).

3. L. Feuk, A. R. Carson, S. W. Scherer, Nat Rev Genet 7, 85 (Feb, 2006).

4. A. J. Sharp, Z. Cheng, E. E. Eichler, Annu Rev Genomics Hum Genet 7, 407

(2006).

5. A. J. Iafrate et al, *Nat Genet* 36, 949 (Sep, 2004).
6. M. R. Mehan, N. B. Freimer, R. A. Ophoff, *Hum Genomics* 1, 335 (Aug, 2004).
7. E. E. Eichler et al, *Nature* 447, 161 (May 10, 2007).
8. M. Simonis et al, *Nat Genet* 3%, 1348 (Nov, 2006).
9. K. Rippe, *Trends Biochem Sci* 26, 733 (Dec, 2001).
10. R. C. Burnett, M. J. Thirman, J. D. Rowley, M. O. Diaz, *Blood* 84, 1232 (Aug 15, 1994).
11. F. Speleman et al, *Leukemia* 19, 358 (Mar, 2005).
12. J. Soulier et al, *Blood* 106, 274 (Jul 1, 2005).
13. R. J. Galjaard et al, *Am J Med Genet A* 121", 168 (Aug 30, 2003).
14. M. Aoyama et al, *Cancer Res* 65, 4587 (Jun 1, 2005).
15. E. Tuzun et al, *Nat Genet* 37, 727 (Jul, 2005).

Todas as publicações mencionadas na especificação anterior estão aqui incorporadas pela referência. Várias modificações e variações dos métodos e sistemas descritos da invenção serão evidentes aos versados na tecnologia sem fugir do escopo e espírito da invenção. Embora a invenção tenha sido descrita com relação às modalidades preferidas específicas, deve-se entender que a invenção reivindicada não deve ser indevidamente limitada a tais modalidades específicas. De fato, várias modificações dos modos descritos para realizar a invenção que são óbvias aos versados em biologia molecular ou campos relacionados estão no escopo das seguintes reivindicações.

## LISTAGEM DE SEQUÊNCIA

<110> Erasmus University Medical Center

<120> 4C

<130> P031700BR

<140> PCT/IB2008/000625

<141> 10-01-2008

<150> US 60/999,750

<151> 11-01-2007

<150> US 60/977,900

<151> 05-10-2007

<160> 32

<170> PatentIn versão 3.4

<210> 1

<211> 22

<212> DNA

<213> Artificial

<220>

<223> Iniciador de oligonucleotídeo

<400> 1

acttcctaca cattaacgag cc

22

<210> 2

<211> 23

<212> DNA

<213> Artificial

<220>

<223> Iniciador de oligonucleotídeo

<400> 2

gctgttatcc ctttcttcc tac

23

<210> 3

<211> 18

<212> DNA

<213> Artificial

<220>

<223> Iniciador de oligonucleotídeo

<400> 3

tcacacgcga agtaggcc

18

<210> 4  
<211> 19  
<212> DNA  
<213> Artificial

<220>  
<223> Iniciador de oligonucleotídeo

<400> 4  
ccttcctcca ccatgatga

19

<210> 5  
<211> 25  
<212> DNA  
<213> Artificial

<220>  
<223> Iniciador de oligonucleotídeo

<400> 5  
aacgcatttg ctcaatcaac tactg

25

<210> 6  
<211> 25  
<212> DNA  
<213> Artificial

<220>  
<223> Iniciador de oligonucleotídeo

<400> 6  
gttgctcctc acatttgctt ctagc

25

<210> 7  
<211> 41  
<212> DNA  
<213> Artificial

<220>  
<223> Iniciador de oligonucleotídeo

<400> 7  
catgaagaaa cgagcacccccc ctgtatgttt ctcccttac c

41

<210> 8  
<211> 40  
<212> DNA  
<213> Artificial

<220>  
<223> Iniciador de oligonucleotídeo

<400> 8  
tgtcaggctc ttctccata cgtcgccag aacactcacc

40

<210> 9  
<211> 37  
<212> DNA  
<213> Artificial

<220>  
<223> Iniciador de oligonucleotídeo

<400> 9  
aatccagggc tacttccagc cgtgatgcta tctgccca

37

<210> 10  
<211> 38  
<212> DNA  
<213> Artificial

<220>  
<223> Iniciador de oligonucleotídeo

<400> 10  
tgggaaaga ccaggtgaag tgcgtggaa agcgagtg

38

<210> 11  
<211> 45  
<212> DNA  
<213> Artificial

<220>  
<223> Iniciador de oligonucleotídeo

<400> 11  
caatccaga tacattccct atacaaatac ttccaagac tggac

45

<210> 12  
<211> 39  
<212> DNA  
<213> Artificial

<220>  
<223> Iniciador de oligonucleotídeo

<400> 12  
gaatatgtta tgcttgatcc ttccatgaga gaagtctag

39

<210> 13  
<211> 18  
<212> DNA  
<213> Artificial

&lt;220&gt;

&lt;223&gt; Iniciador de oligonucleotídeo

&lt;400&gt; 13

atgtgactcc tcttagatc

18

&lt;210&gt; 14

&lt;211&gt; 71

&lt;212&gt; DNA

&lt;213&gt; Artificial

&lt;220&gt;

&lt;223&gt; Iniciador de oligonucleotídeo

&lt;400&gt; 14

aatgatacgg cgaccaccga acactcttc cctacacgac gctctccga tctatgtac 60

tcctcttagat c

71

&lt;210&gt; 15

&lt;211&gt; 19

&lt;212&gt; DNA

&lt;213&gt; Artificial

&lt;220&gt;

&lt;223&gt; Iniciador de oligonucleotídeo

&lt;400&gt; 15

ccctgaacct cttgaagct

19

&lt;210&gt; 16

&lt;211&gt; 40

&lt;212&gt; DNA

&lt;213&gt; Artificial

&lt;220&gt;

&lt;223&gt; Iniciador de oligonucleotídeo

&lt;400&gt; 16

caagcagaag acggcatacg accctgaacc tcttgaagct

40

&lt;210&gt; 17

&lt;211&gt; 18

&lt;212&gt; DNA

&lt;213&gt; Artificial

&lt;220&gt;

&lt;223&gt; Iniciador de oligonucleotídeo

&lt;400&gt; 17

cggcctccaa ttgtgatc

18

<210> 18  
<211> 71  
<212> DNA  
<213> Artificial

<220>  
<223> Iniciador de oligonucleotideo

<400> 18

aatgatacgg cgaccaccga acactcttc cctacacgac gctctccga tctcggcctc 60

caattgtgat c

71

<210> 19  
<211> 20  
<212> DNA  
<213> Artificial

<220>  
<223> Iniciador de oligonucleotideo

<400> 19

gaattgctt tggtaagctt

20

<210> 20  
<211> 41  
<212> DNA  
<213> Artificial

<220>  
<223> Iniciador de oligonucleotideo

<400> 20

caagcagaag acggcatacg agaattgctt ttggtaagct t

41

<210> 21  
<211> 18  
<212> DNA  
<213> Artificial

<220>  
<223> Iniciador de oligonucleotideo

<400> 21

tttagccct gacagatc

18

<210> 22  
<211> 71  
<212> DNA  
<213> Artificial

<220>

<223> Iniciador de oligonucleotídeo

<400> 22

aatgatacgg cgaccaccga acactcttc cctacacgac gctttccga tcttttagc 60

cctgacagat c

71

<210> 23

<211> 20

<212> DNA

<213> Artificial

<220>

<223> Iniciador de oligonucleotídeo

<400> 23

agtcaaacat aaggctaa

20

<210> 24

<211> 41

<212> DNA

<213> Artificial

<220>

<223> Iniciador de oligonucleotídeo

<400> 24

caaggcagaag acggcatacg aagtcaaaca taaggctaa

c 41

<210> 25

<211> 157

<212> DNA

<213> Homo sapiens

<400> 25

aaggaaaccc catgcccata agacgtcact aattctgaa ctcttgttt tttttttt 60

tttcaagta gttctcatct aagttagtgtt ttttgtcat gagaaaaatca gatatgtgc 120

taaaaaattca caactattgc aaaaaaaaaat aaaagac

157

<210> 26

<211> 100

<212> DNA

<213> Homo sapiens

<400> 26

ctccaatgta actgtggatt acacctaaaaa gagccagaaaa acacagactc tctgtggaa 60

catgacacaa cagtgcgtgg tattatttt tccttagttag

100

<210> 27  
 <211> 60  
 <212> DNA  
 <213> Homo sapiens

<400> 27  
 aagagccaga aaacacagac tctctgtgga accatgacac aacagtgc tt ggtattttt 60

<210> 28  
 <211> 50  
 <212> DNA  
 <213> Homo sapiens

<400> 28  
 aaacccatg cccataagac gtcactaatt tctgaactct tgttttttt 50

<210> 29  
 <211> 45  
 <212> DNA  
 <213> Homo sapiens

<400> 29  
 cgaagaactg ctcattgtag gtcaagttc aaggcttga accac 45

<210> 30  
 <211> 52  
 <212> DNA  
 <213> Homo sapiens

<400> 30  
 tctgtgccag cagcttgc cgcgagccgg ccgatactct gcctaggact cc 52

<210> 31  
 <211> 30  
 <212> DNA  
 <213> Homo sapiens

<400> 31  
 gaccaggc a ttggattttt ttcagagatc 30

<210> 32  
 <211> 26  
 <212> DNA  
 <213> Homo sapiens

<400> 32  
 gatcctgaca ccttagagct aagctt 26

## REIVINDICAÇÕES

1. Método para analisar a frequência de interação de duas ou mais sequências de nucleotídeos alvos com uma ou mais sequências de nucleotídeos de interesse (por exemplo, um ou mais loci genômicos), **CARACTERIZADO** pelo fato de que compreendendo as

5 etapas de:

- (a) fornecer uma amostra de DNA reticulado;
- (b) digerir o DNA reticulado com uma enzima de restrição primária;
- (c) ligar as sequências de nucleotídeos reticulados;
- (d) reverter a reticulação;
- 10 (e) digerir opcionalmente as sequências de nucleotídeos com uma enzima de restrição secundária;

(f) ligar opcionalmente uma ou mais sequências de DNA de composição de nucleotídeo conhecida no(s) sítio(s) de digestão de enzima de restrição secundária disponível(s) que flanqueiam uma ou mais sequências de nucleotídeos de interesse;

15 (g) amplificar uma ou mais sequências de nucleotídeos de interesse usando pelo menos dois oligonucleotídeos iniciadores, em que cada iniciador hibridiza nas sequências de DNA que flanqueiam as sequências de nucleotídeos de interesse usando PCR multiplex;

- (h) hibridizar a(s) sequência(s) amplificada(s) em um arranjo; e
- (i) determinar a frequência de interação entre as sequências de DNA.

20 2. Método, de acordo com a reivindicação 1, **CARACTERIZADO** pelo fato de que a reação de ligação na etapa (c) ou (f) resulta na formação de círculos de DNA.

25 3. Método, de acordo com a reivindicação 1 ou reivindicação 2, **CARACTERIZADO** pelo fato de que a etapa (h) compreende a análise de produtos de ligação entre sequências alvos e sequências reticuladas de interesse por meio de sequenciamento (por exemplo, o sequenciamento de alto desempenho).

30 4. Método, de acordo com qualquer das reivindicações anteriores para analisar a frequência de interação de duas ou mais sequências de nucleotídeos alvos com uma ou mais sequências de nucleotídeos de interesse, **CARACTERIZADO** pelo fato de que compreende o agrupamento de alguns ou todos os produtos de PCR obtidos por cada uma das sequências alvos na etapa (g) e análise simultânea subsequente de suas interações de DNA.

35 5. Método, de acordo com a reivindicação 4, **CARACTERIZADO** pelo fato de que as duas ou mais sequências amplificadas são marcadas diferencialmente antes do agrupamento e análise por hibridização em um arranjo.

6. Método, de acordo com a reivindicação 4 ou reivindicação 5, **CARACTERIZADO** pelo fato de que as duas ou mais sequências amplificadas são identicamente marcadas e analisadas por hibridização em um arranjo quando as sequências residem em cromossomos

diferentes.

7. Método, de acordo com a reivindicação 4, **CARACTERIZADO** pelo fato de que as duas ou mais sequências amplificadas são identicamente marcadas quando as sequências residem no mesmo cromossomo em uma distância que é longe o suficiente para sobreposição mínima entre sinais de interação DNA-DNA.

8. Método, de acordo com qualquer das reivindicações anteriores, **CARACTERIZADO** pelo fato de que o sequenciamento de alto desempenho é usado para analisar as junções de ligação formadas entre as sequências alvos e sequências capturadas de interesse.

9. Método, de acordo com a reivindicação 8, **CARACTERIZADO** pelo fato de que o sequenciamento é direcionado às junções de ligação formadas entre as sequências alvos e sequências capturadas de interesse pela adição de sequências adaptadoras exigidas para sequenciamento nas extremidades das sequências amplificadas.

10. Método de acordo com a reivindicação 9, **CARACTERIZADO** pelo fato de que o sequenciamento é direcionado às junções de ligação formadas entre sequências alvos e sequências capturadas de interesse pela adição das sequências adaptadoras completas ou parte delas exigidas para sequenciamento como projeções 5' nos oligonucleotídeos iniciadores usados para amplificar uma ou mais sequências de nucleotídeos de interesse.

11. Método, de acordo com a reivindicação 9, **CARACTERIZADO** pelo fato de que o sequenciamento é direcionado às junções de ligação formadas entre sequências alvos e sequências capturadas de interesse pela conjugação de uma substância de biotina ou outra fração nos oligonucleotídeos iniciadores usados para amplificar uma ou mais sequências de nucleotídeos de interesse, seguido por purificação mediada por estreptavidina ou mediada de outra forma do material amplificado por PCR.

12. Método, de acordo com qualquer uma das reivindicações 8 a 11, **CARACTERIZADO** pelo fato de que o sequenciamento é direcionado às junções de ligação entre sequências alvos e sequências capturadas de interesse desenhando os oligonucleotídeos iniciadores usados para amplificar uma ou mais sequências de nucleotídeos de interesse em 400, 300, 200, 150, 100, 90, 80 70, 60, 50, 40, 30, 20, 10, 9, 8, 7, 6, 5, 4, 3, 2, 1 nucleotídeos do(s) sítio(s) de reconhecimento de enzima de restrição primária e/ou secundária analisado(s).

13. Método, de acordo com qualquer uma das reivindicações 8 to 11, **CARACTERIZADO** pelo fato de que o sequenciamento é direcionado às junções de ligação entre sequências alvos e sequências capturadas de interesse desenhando os oligonucleotídeos iniciadores usados para amplificar uma ou mais sequências de nucleotídeos de interesse, de maneira tal que elas se sobreponham parcial ou completamente com o sítio de reconhecimento de enzima de restrição primária e/ou secundária analisado.

14. Método, de acordo com qualquer uma das reivindicações 8 a 13, **CARACTERIZADO** pelo fato de que as sequências são lidas através da junção de ligação de maneira tal que, quando amostras de PCR multiplexadas ou agrupadas são analisadas, seja obtida informação de sequência suficiente (por exemplo, 12 nucleotídeos ou mais) em qualquer lado da junção de ligação para identificar inambiguamente cada sequência alvo e cada sequência capturada de interesse.

5 15. Método, de acordo com qualquer das reivindicações anteriores, **CARACTERIZADO** pelo fato de que a sequência de nucleotídeos alvo é selecionada do grupo que consiste em um rearranjo genômico, promotor, um melhorador, um silenciador, 10 um isolante, uma região de anexação de matriz, uma região de controle de locus, uma unidade de transcrição, uma origem de replicação, um ponto quente de recombinação, um ponto de quebra de translocação, um centrômero, um telômero, uma região densa em gene, uma região pobre em gene, um elemento repetitivo e um sítio de integração (viral).

16. Método, de acordo com qualquer das reivindicações anteriores, 15 **CARACTERIZADO** pelo fato de que a sequência de nucleotídeos alvo é uma sequência de nucleotídeos que está associada ou causa uma doença, ou está localizada em até ou mais que 15 Mb em um molde de DNA linear de um locus que está associado com ou causam uma doença.

20 17. Método, de acordo com qualquer das reivindicações anteriores, **CARACTERIZADO** pelo fato de que a sequência de nucleotídeos alvo é selecionada do grupo que consiste em: *AML1*, *MLL*, *MYC*, *BCL*, *BCR*, *ABLI*, *IGH*, *LYL1*, *TAL1*, *TAL2*, *LMO2*, *TCRcr/S*, *TCR/3* e *HOX* ou outros loci associados com doença descritos em "Catalogue of Unbalanced Chromossome Aberrations in Man" 2nd edition. Albert Schinzel. Berlin: Walter de Gruyter, 2001. ISBN 3-11-011607-3.

25 18. Método, de acordo com qualquer das reivindicações anteriores, **CARACTERIZADO** pelo fato de que as sequências alvos são distribuídas ao longo do molde do genoma linear, de maneira tal que as sequências interagentes cubram um cromossomo total ou o genoma.

30 19. Método, de acordo com qualquer das reivindicações anteriores, **CARACTERIZADO** pelo fato de que a enzima de restrição primária é uma enzima de restrição que reconhece um sítio de reconhecimento de 6-8 bp.

20. Método, de acordo com a reivindicação 19, **CARACTERIZADO** pelo fato de que a enzima de restrição primária é selecionada do grupo que consiste em *BglIII*, *HindIII*, *BamHI*, *BamHI*, *Spa*, *PstI* e *NdeI*.

35 21. Método, de acordo com a reivindicação 19 ou reivindicação 20, **CARACTERIZADO** pelo fato de que a enzima de restrição primária é selecionada com base em sua ausência ou em representação em sequências repetitivas.

22. Método, de acordo com qualquer das reivindicações anteriores, **CARACTERIZADO** pelo fato de que a enzima de restrição secundária é uma enzima de restrição que reconhece um sítio de reconhecimento de sequência de nucleotídeos de 4 ou 5 bp.

5 23. Método, de acordo com qualquer das reivindicações anteriores, **CARACTERIZADO** pelo fato de que o sítio de reconhecimento de enzima de restrição secundária está localizado em mais que cerca de 350 bp do sítio de restrição primária na sequência de nucleotídeos alvo.

10 24. Método para analisar a frequência de interação de uma sequência de nucleotídeos alvo com uma ou mais sequências de nucleotídeos (por exemplo, um ou mais loci genômicos), **CARACTERIZADO** pelo fato de que compreende as etapas de:

- (a) fornecer uma amostra de DNA reticulado;
- (b) digerir o DNA reticulado com uma enzima de restrição primária;
- (c) ligar as sequências de nucleotídeos reticulados;
- 15 (d) inverter a reticulação;
- (e) digerir opcionalmente as sequências de nucleotídeos com uma enzima de restrição secundária;
- (f) circular as sequências de nucleotídeos;
- (g) amplificar uma ou mais sequências de nucleotídeos que estão ligadas na sequência de nucleotídeos alvo;
- 20 (h) hibridizar opcionalmente as sequências amplificadas em um arranjo ou analisar as sequências amplificadas por sequenciamento (por exemplo, o sequenciamento de alto desempenho); e
- (i) determinar a frequência de interação entre as sequências de DNA.

25 25. Método para identificar uma ou mais Interações DNA-DNA que são indicativas de uma condição de doença particular, **CARACTERIZADO** pelo fato de que compreende as etapa de realizar as etapas (a)-(i) das reivindicações 1-23, em que na etapa (a) uma amostra de DNA reticulado é fornecido de uma célula doente e uma não doente, e em que uma diferença entre a frequência de interação entre as sequências de DNA da células doentes e não doentes indica uma diferença na organização linear dos moldes do cromossomo (por exemplo, um rearranjo genômico), que é indicativo de um traço particular ou condição de doença.

30 26. Método de diagnose ou prognose de uma doença ou síndrome causada ou associada com uma mudança em uma interação DNA-DNA, **CARACTERIZADO** pelo fato de que compreende a etapa de realizar as etapas (a)-(i) de acordo com qualquer uma das reivindicações 1-23, em que etapa (a) compreende fornecer uma amostra de DNA reticulado de um sujeito; e em que etapa (i) compreende comparar a frequência de interação entre as

sequências de DNA com aquela de um controle não afetado; em que uma diferença entre o valor obtido do controle e o valor obtido do sujeito é indicativa que o sujeito está sofrendo da doença ou síndrome ou é indicativa que o sujeito sofrerá da doença ou síndrome.

27. Método, de acordo com a reivindicação 26, **CARACTERIZADO** pelo fato de que uma transição de baixas para altas frequências de interação é indicativa do local de um rearranjo genético equilibrado e/ou desequilibrado.

28. Método, de acordo com a reivindicação 26, **CARACTERIZADO** pelo fato de que um padrão inverso de frequências de interação DNA-DNA na amostra em questão comparado ao controle é indicativo de uma inversão equilibrada e/ou desequilibrada.

29. Método, de acordo com a reivindicação 26, **CARACTERIZADO** pelo fato de que uma redução na frequência de interação DNA-DNA na amostra em questão comparada ao controle, em combinação com um aumento na frequências de interação DNA-DNA em regiões mais distantes, é indicativa de uma deleção equilibrada e/ou desequilibrada.

30. Método, de acordo com a reivindicação 26, **CARACTERIZADO** pelo fato de que um aumento ou uma diminuição nas frequências de interação DNA-DNA das amostras em questão comparadas ao controle é indicativo de uma duplicação ou inserção equilibrada e/ou desequilibrada.

31. Método, de acordo com qualquer uma das reivindicações 26-30, **CARACTERIZADO** pelo fato de que a cariotipagem espectral e/ou FISH é usada antes de realizar o dito método.

32. Método, de acordo com qualquer uma das reivindicações 26-30, **CARACTERIZADO** pelo fato de que a doença é uma doença genética.

33. Método, de acordo com qualquer uma das reivindicações 26-30, **CARACTERIZADO** pelo fato de que a doença é câncer.

34. Método de diagnose ou prognose de um doença ou síndrome causada ou associada com uma mudança em uma interação DNA-DNA, **CARACTERIZADO** pelo fato de que comprehende as etapa de: realizar as etapas (a)-(i) de acordo com qualquer uma das reivindicações 1-23, em que etapa (a) comprehende fornecer uma amostra de DNA reticulado de um sujeito; e em que o dito método comprehende a etapa adicional de: (j) identificar um ou mais loci que foram submetidos a um rearranjo genômico que está associado com uma doença.

35. Método de ensaio para identificar um ou mais agentes que modulam uma interação DNA-DNA, **CARACTERIZADO** pelo fato de que comprehende as etapas de:

(a) colocar uma amostra em contato com um ou mais agentes; e  
(b) realizar as etapas (a) a (i) de acordo com qualquer uma das reivindicações 1-23, em que a etapa (a) comprehende fornecer DNA reticulado da amostra; em que uma diferença entre (i) a frequência de interação entre as sequências de

DNA na presença do agente e (ii) a frequência de interação entre as sequências de DNA na ausência do agente é indicativa de um agente que modula a interação DNA-DNA.

36. Método para detectar o local de um rearranjo equilibrado e/ou desequilibrado (por exemplo, uma translocação), **CARACTERIZADO** pelo fato de que compreende as etapas de:

- (a) realizar as etapas (a) a (i) de acordo com qualquer uma das reivindicações 1-23; e  
(b) comparar a frequência de interação entre as sequências de DNA com aquela de um controle;

em que uma transição de altas para baixas frequências de interação DNA-DNA na amostra comparada ao controle é indicativa do local de um ponto de quebra.

37. Método para detectar o local de uma inversão equilibrado e/ou desequilibrado, **CARACTERIZADO** pelo fato de que compreende as etapas de:

- (a) realizar as etapas (a) a (i) de acordo com qualquer uma das reivindicações 1-23; e  
(b) comparar a frequência de interação entre as sequências de DNA com aquela de um controle;

em que um padrão inverso de frequências de interação DNA-DNA da amostra comparado ao controle é indicativo de uma inversão.

38. Método para detectar o local de uma deleção, **CARACTERIZADO** pelo fato de que compreende as etapas de:

- (a) realizar as etapas (a) a (i) de acordo com qualquer uma das reivindicações 1-23; e  
(b) comparar a frequência de interação entre as sequências de DNA com aquela de um controle;

em que uma redução nas frequências de interação DNA-DNA da amostra comparada ao controle é indicativa de deleção.

39. Método para detectar o local de uma duplicação, **CARACTERIZADO** pelo fato de que compreende as etapas de:

- (a) realizar as etapas (a) a (i) de acordo com qualquer uma das reivindicações 1-23; e  
(b) comparar a frequência de interação entre as sequências de DNA com aquela de um controle;

em que um aumento ou uma diminuição nas frequências de interação DNA-DNA da amostra em questão comparado ao controle é indicativo de uma duplicação ou inserção.

40. Método, de acordo com qualquer das reivindicações anteriores, **CARACTERIZADO** pelo fato de que as sequências de nucleotídeos interagentes com duas

ou mais sequências alvos são amplificadas.

41. Método, de acordo com a reivindicação 40, **CARACTERIZADO** pelo fato de que as sequências alvos estão posicionadas nos loci genômicos ou próximas a eles, conhecidos para estar associados a uma condição de doença.

5 42. Método, de acordo com a reivindicação 41, **CARACTERIZADO** pelo fato de que as sequências alvos são selecionadas sem conhecimento prévio no local de um rearranjo e são espaçadas de maneira tal que as sequências interagentes cubram um cromossomo completo ou o genoma, e em que as sequências interagentes identificadas permitem reconstruir mapas cromossômicos lineares e rearranjos genômicos que ocorreram nos cromossomos e entre eles.

10 43. Método, de acordo com qualquer uma das reivindicações 40 a 42, **CARACTERIZADO** pelo fato de que as sequências amplificadas são marcadas.

15 44. Método, de acordo com qualquer uma das reivindicações 43, **CARACTERIZADO** pelo fato de que as sequências amplificadas são marcadas diferencialmente de acordo com sua posição no genoma.

45. Método, de acordo com qualquer uma das reivindicações 40 a 44, **CARACTERIZADO** pelo fato de que é para a detecção de um rearranjo, translocação, inversão, deleção, duplicação ou inserção equilibrados e/ou desequilibrados.

20 46. Agente, **CARACTERIZADO** pelo fato de que é obtido ou obtenível pelo método de ensaio, de acordo com a reivindicação 34.

47. Método para analisar a frequência de interação de uma ou mais sequências de nucleotídeos alvos com uma ou mais sequências de nucleotídeos de interesse (por exemplo, um ou mais loci genômicos), **CARACTERIZADO** pelo fato de que compreende as etapas de:

25 (a) fornecer uma amostra de DNA reticulado;  
(b) digerir o DNA reticulado com uma enzima de restrição primária;  
(c) ligar as sequências reticuladas de nucleotídeos;  
(d) inverter a reticulação; e  
(e) sequenciar as sequências de nucleotídeos ligados.

30 48. Método para determinar a presença de um rearranjo genômico em uma amostra, **CARACTERIZADO** pelo fato de que compreende as etapas de:

(a) fornecer uma amostra de ácido nucléico (por exemplo, DNA genômico), em que o dito ácido nucléico compreende uma sequência de nucleotídeos de sequência adjacente conhecida no local do rearranjo genômico suspeito;  
35 (b) digerir o DNA com uma enzima de restrição primária para formar uma pluralidade de fragmentos de restrição;  
(c) purificar opcionalmente os fragmentos de restrição;

(d) ligar os fragmentos de restrição para formar DNA circular;

(e) purificar opcionalmente o DNA circular;

(f) digerir o DNA circular com uma enzima de restrição secundária para formar uma pluralidade de fragmentos de restrição;

5 (g) ligar os fragmentos de restrição para formar DNA circular;

(h) amplificar o rearranjo genômico suspeito usando um ou mais iniciadores que hibridizam na sequência de nucleotídeos de sequência conhecida; e

(g) sequenciar o rearranjo genômico suspeito.

49. Método, de acordo com qualquer uma das reivindicações 1 a 45,

10 **CARACTERIZADO** pelo fato de que a etapa de hibridização etapa pode ser substituída com uma etapa de sequenciamento.

50. Método, de acordo com a reivindicação 47 ou 48, **CARACTERIZADO** pelo fato de que tanto a sequência de nucleotídeos alvo quanto a sequência de nucleotídeos de interesse são identificadas por sequenciamento.

15 51. Método, de acordo com qualquer uma das reivindicações 47 a 50, **CARACTERIZADO** pelo fato de que as sequências adaptadoras estão ligadas nos produtos de PCR.

52. Base de dados, **CARACTERIZADO** pelo fato de que é de sequências de ácidos nucléicos de cerca de 6-50 pares de bases que flanqueiam diretamente e, opcionalmente, incluem o sítio de reconhecimento de enzima de restrição primária ou o sítio de reconhecimento de enzima de restrição secundária de cada sequência alvo.

53. Base de dados, **CARACTERIZADO** pelo fato de que é de sequências de ácido nucléico de cerca de 12-50 pares de bases, **CARACTERIZADA** pelo fato de que flanqueiam diretamente todos os sítios de reconhecimento de enzima de restrição primária relevantes no genoma.

25 54. Uso da base de dados das sequências de ácido nucléico da reivindicação 52 ou 53, **CARACTERIZADO** pelo fato de que é para determinar a posição genômica de cada uma das sequências capturadas identificadas.

55. Método ou um agente ou uma base de dados ou um uso, **CARACTERIZADO** pelo fato de que é substancialmente da maneira aqui descrita e com relação a qualquer dos exemplos ou figuras.

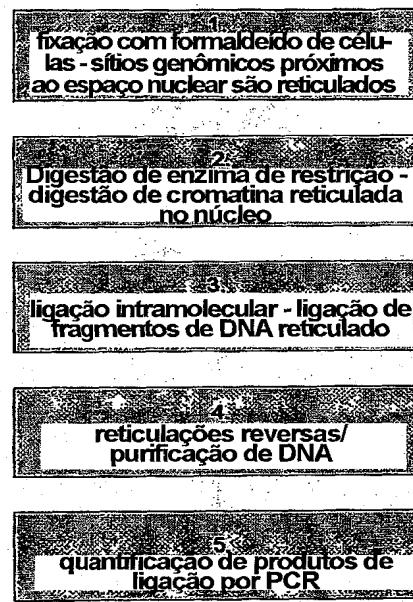
**FIGURA 1**

FIGURA 2

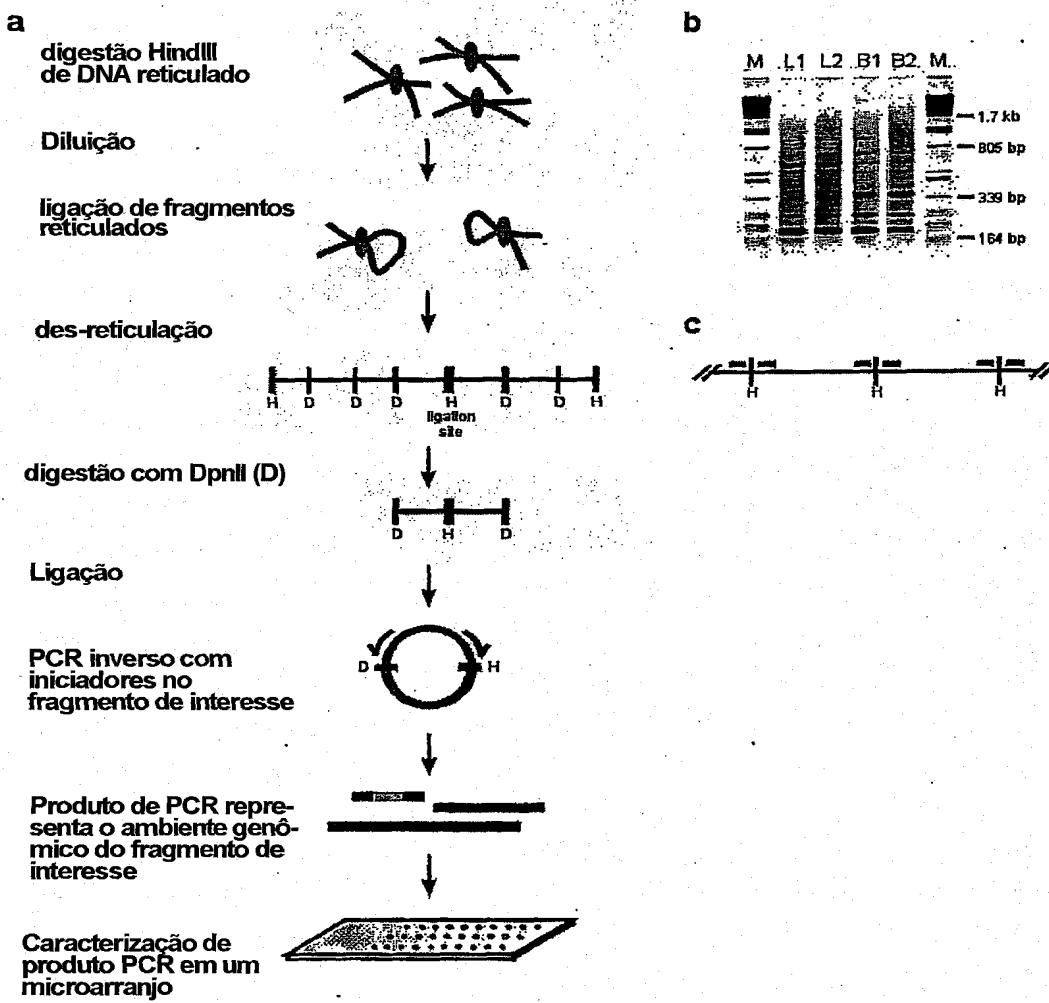


FIGURA 3

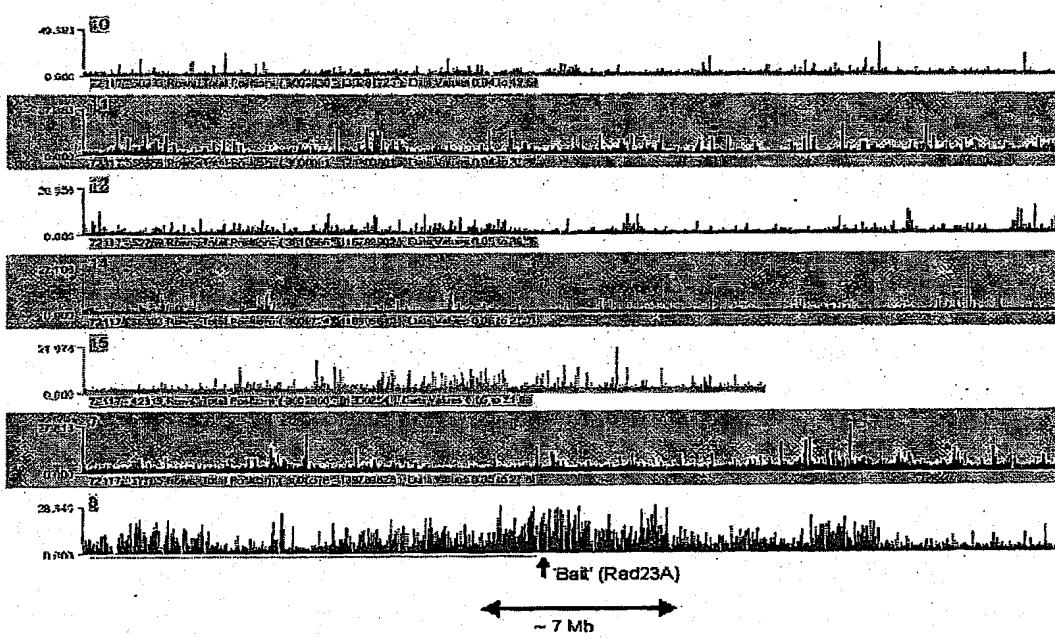


FIGURA 4

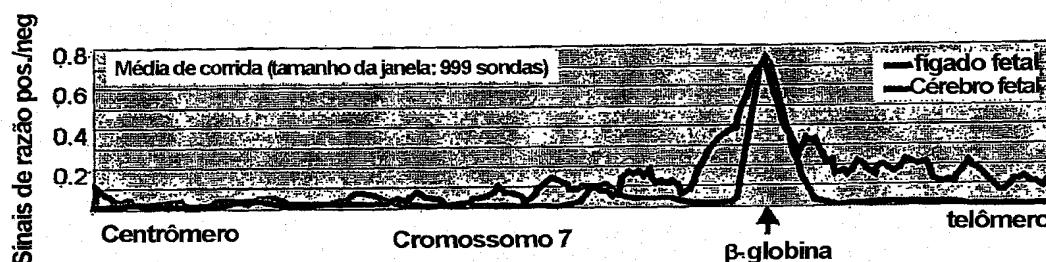


FIGURA 5

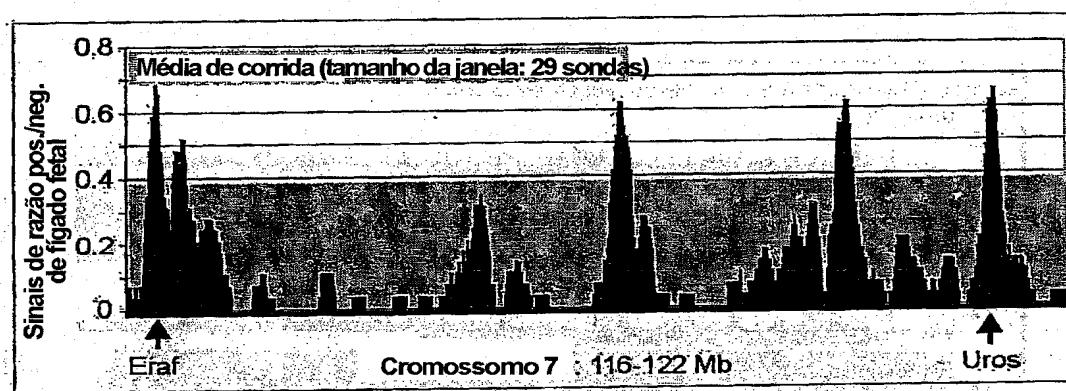
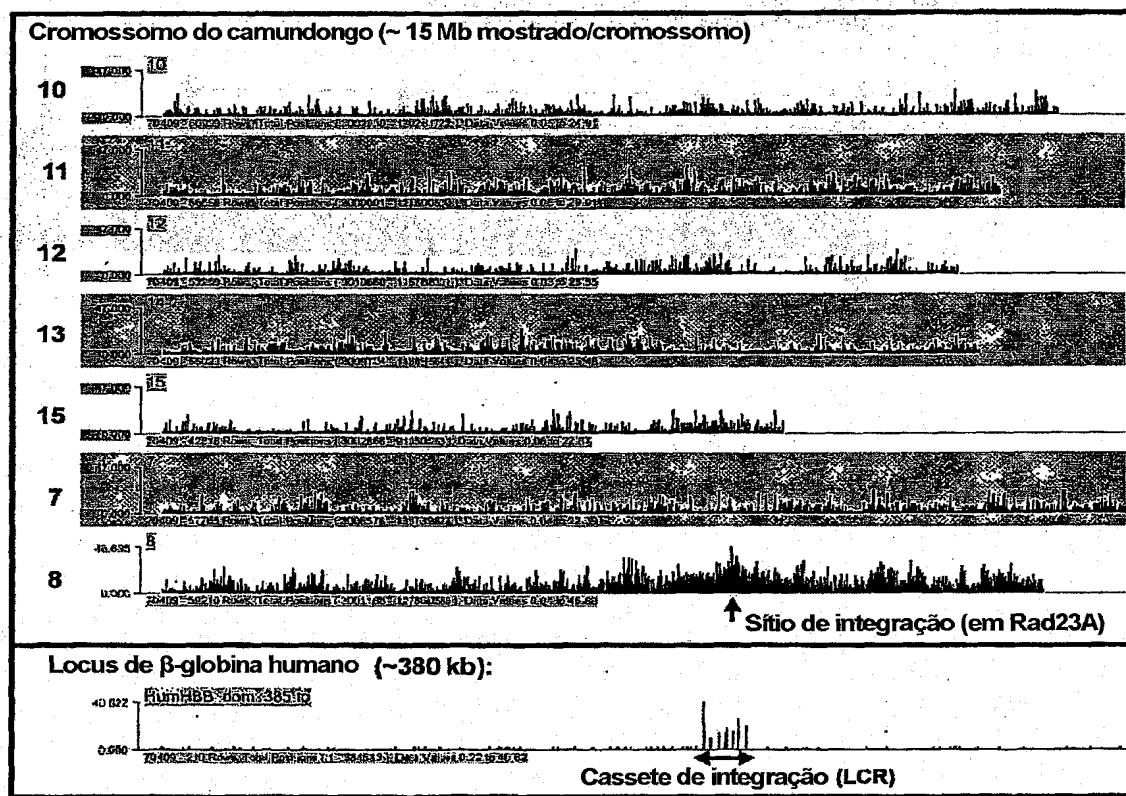


FIGURA 6



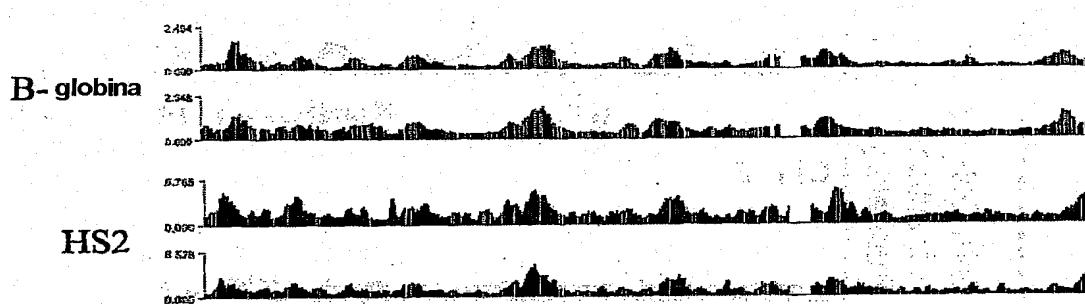
**FIGURA 7**

FIGURA 8

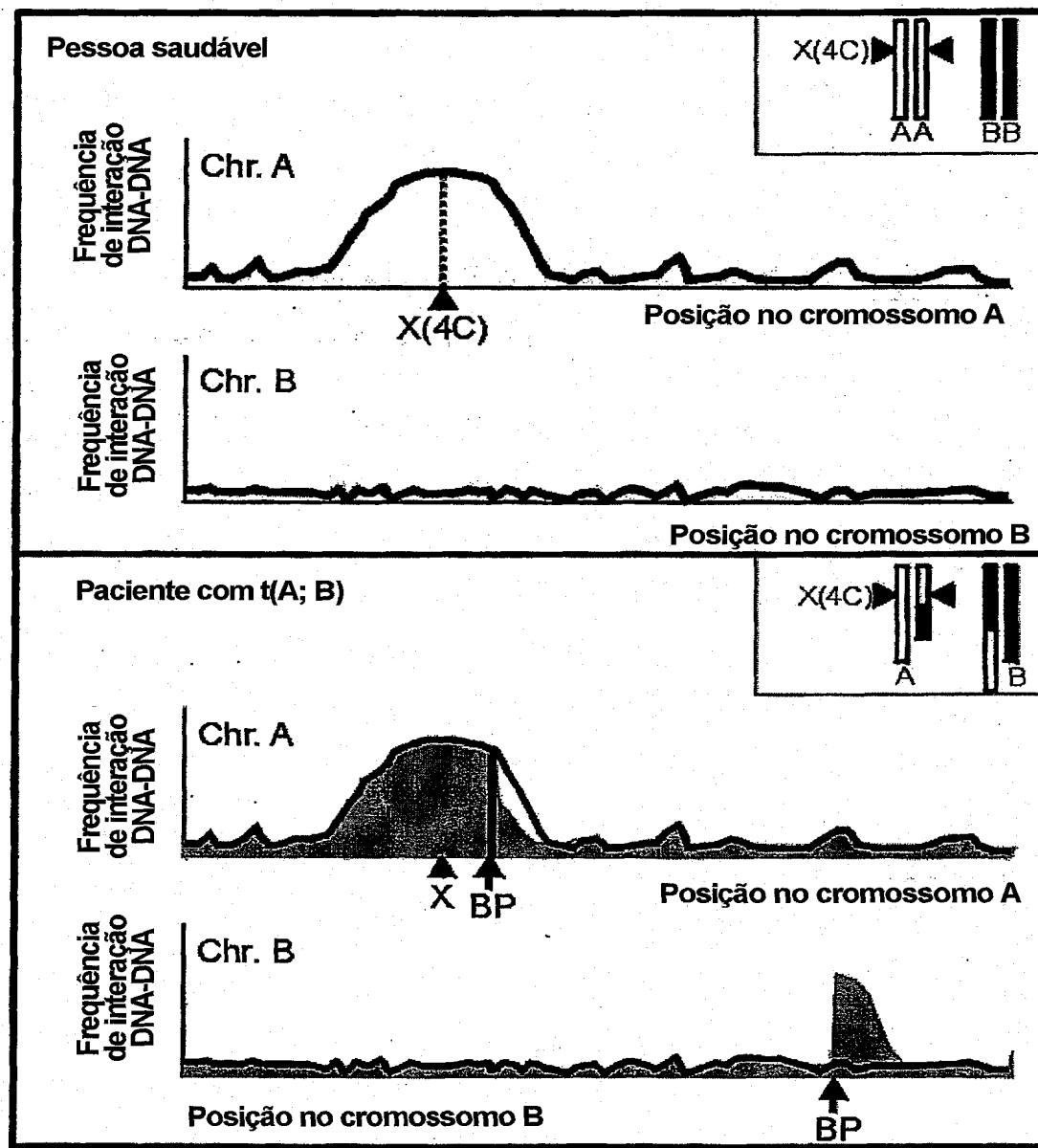


FIGURA 9

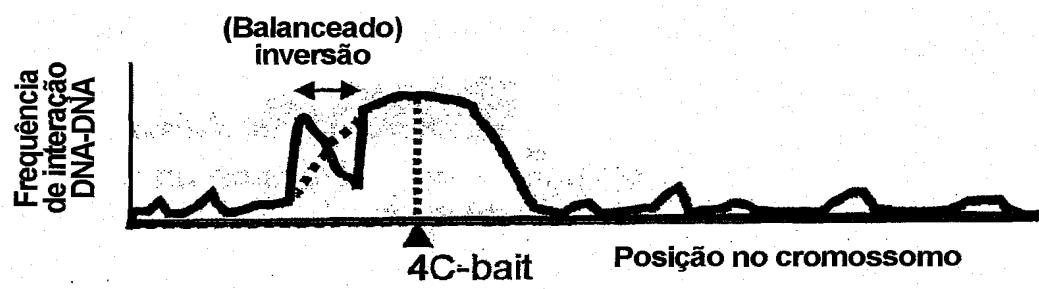
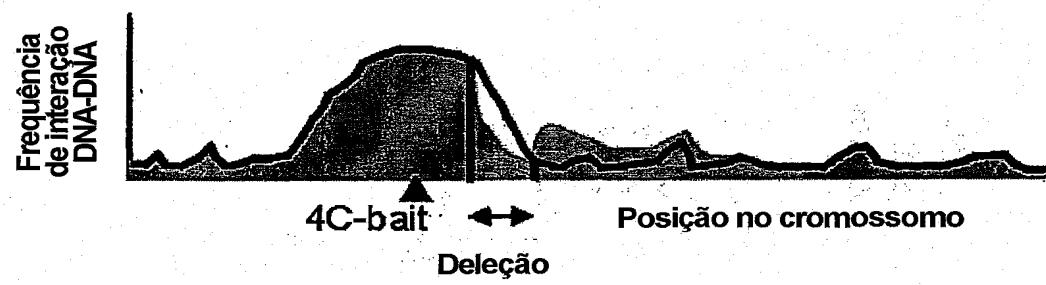


FIGURA 10



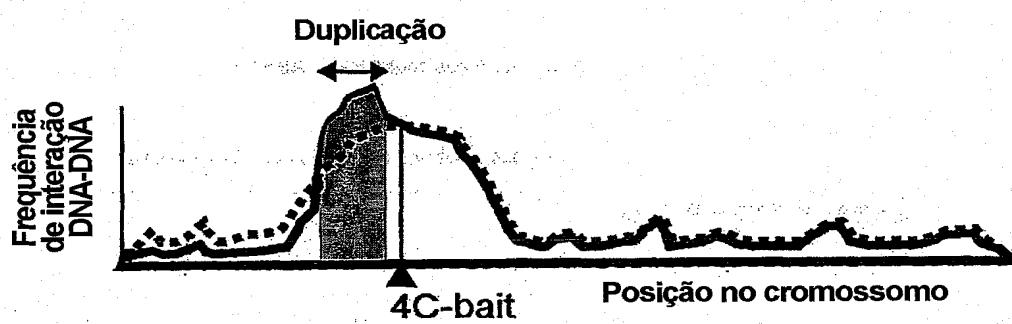
**FIGURA 11**

FIGURA 12

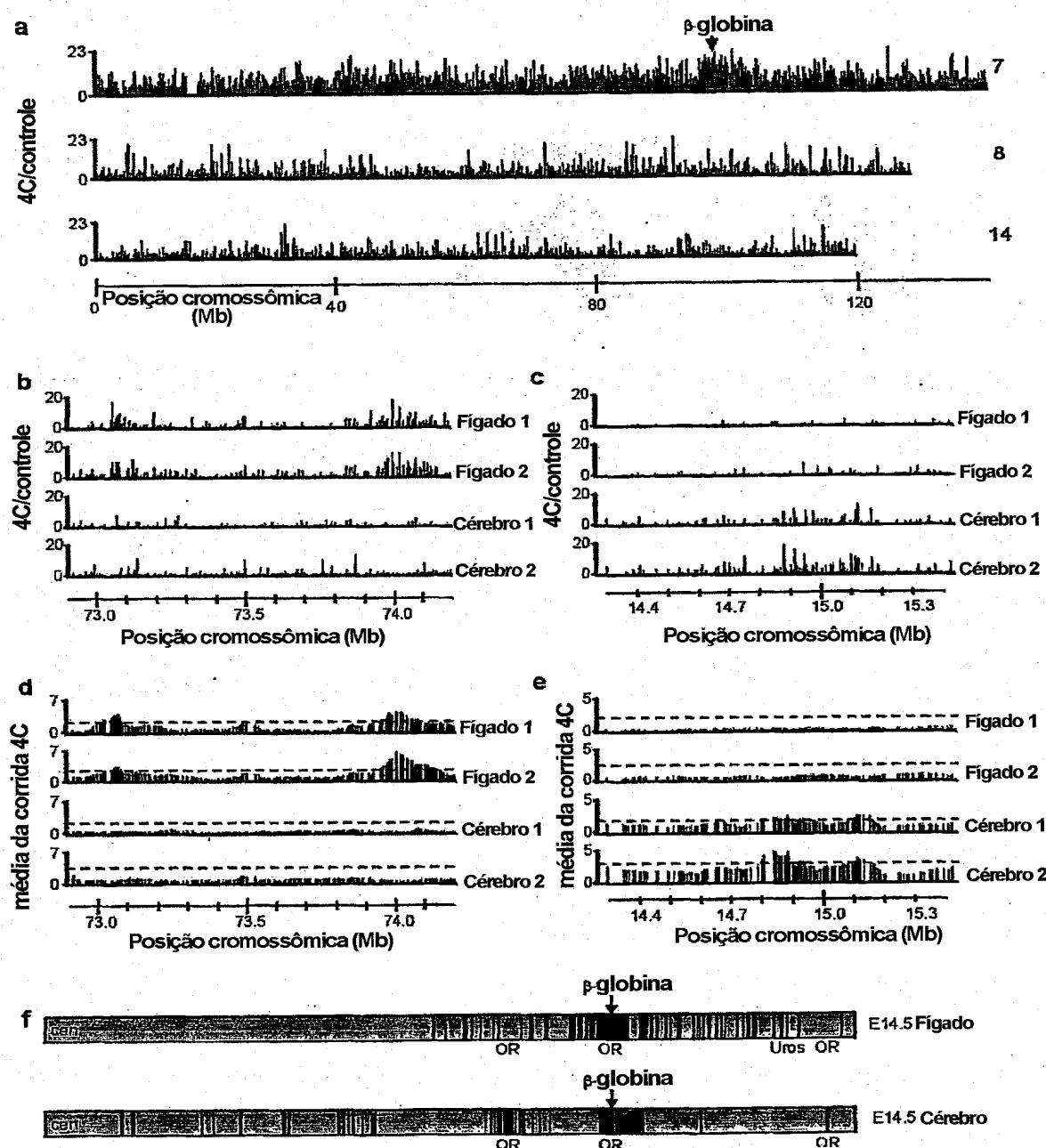


FIGURA 13

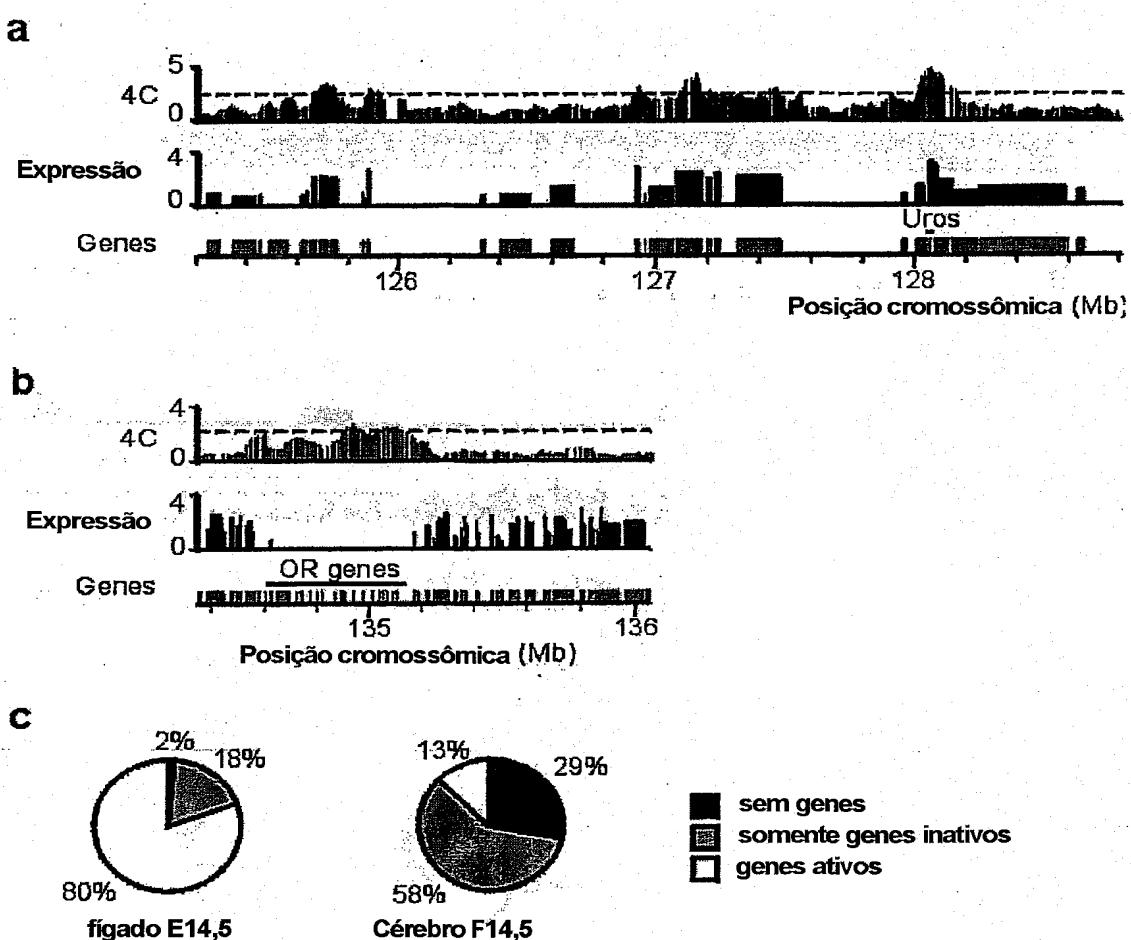


FIGURA 14

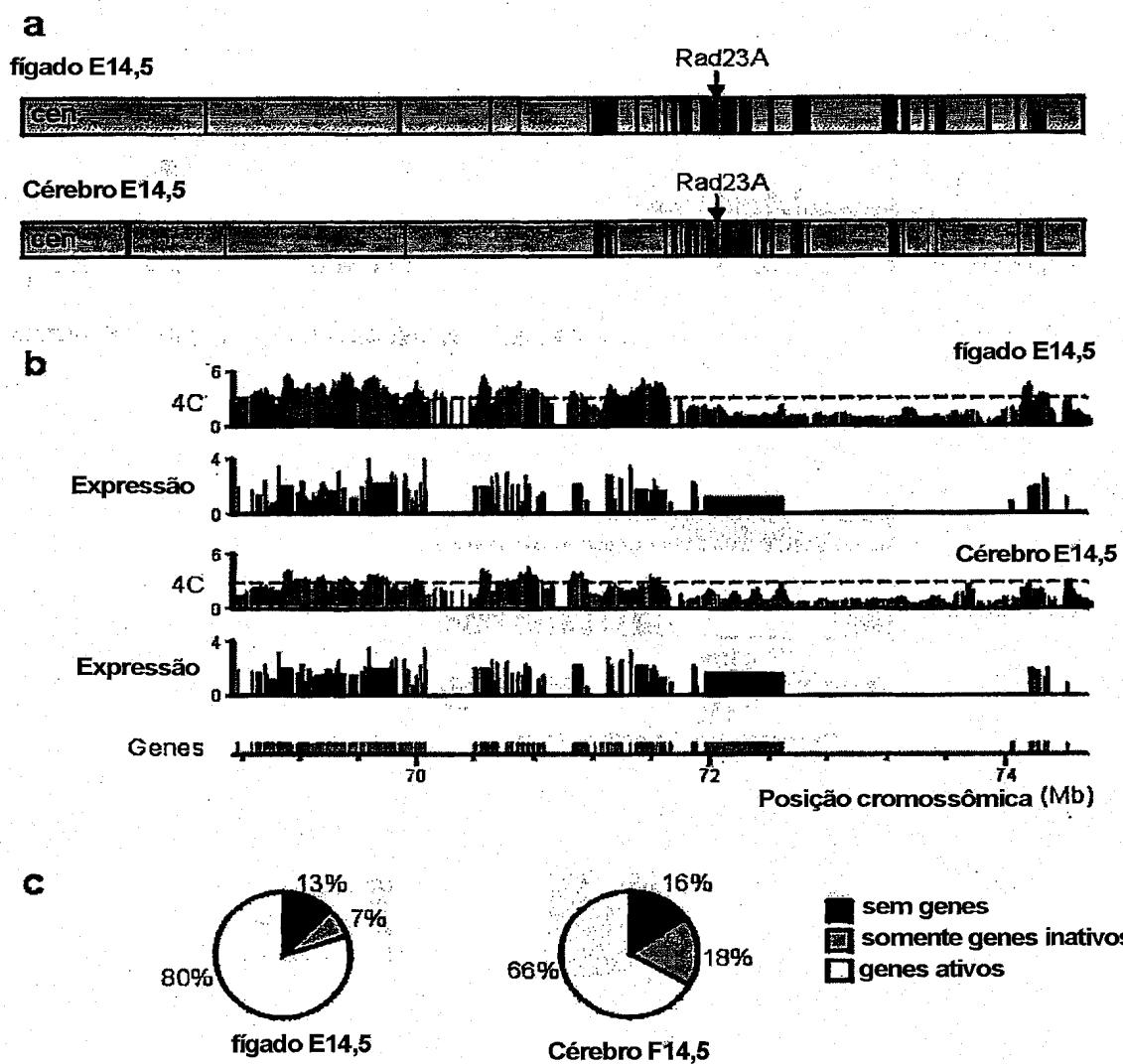


FIGURA 15

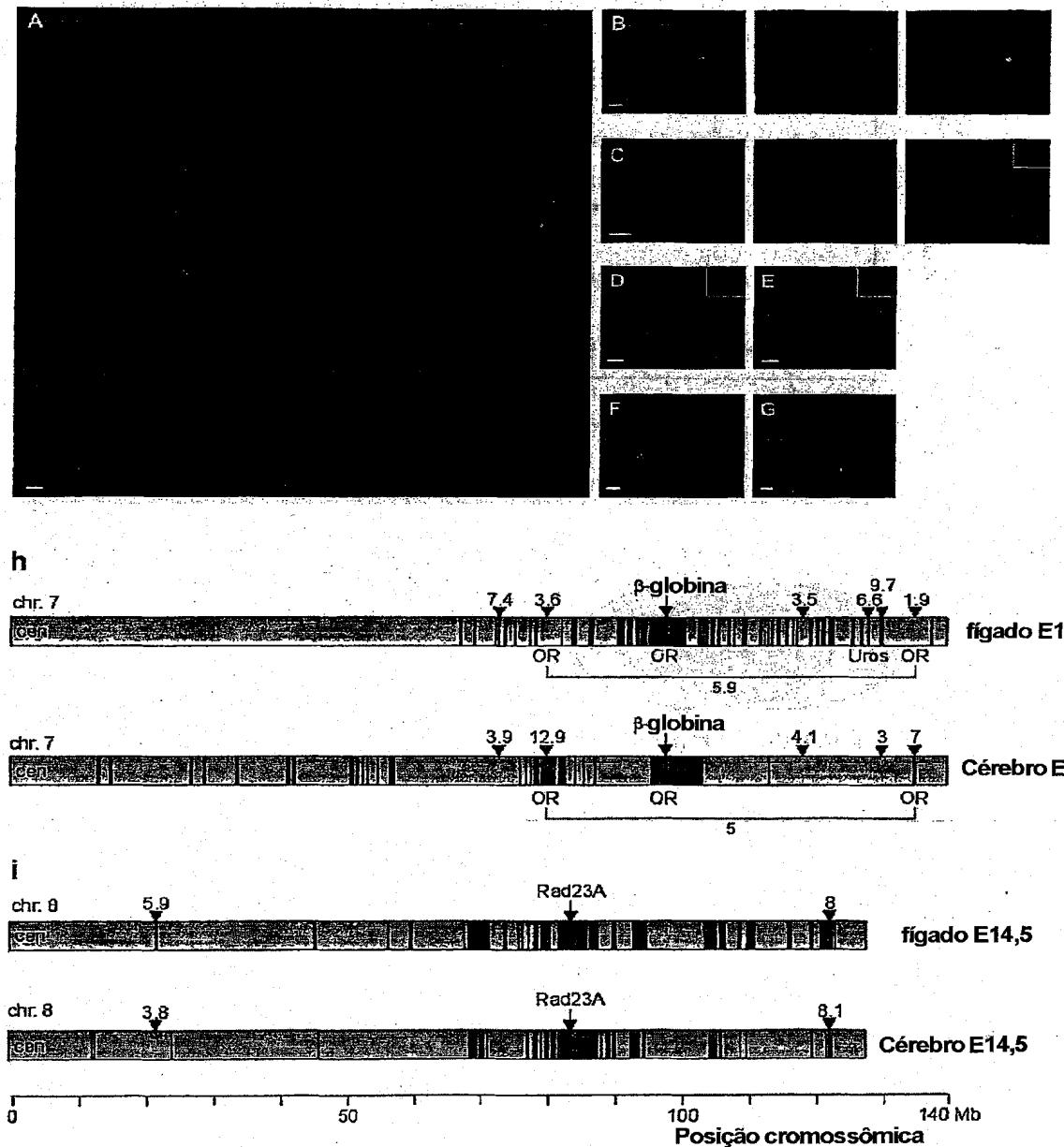
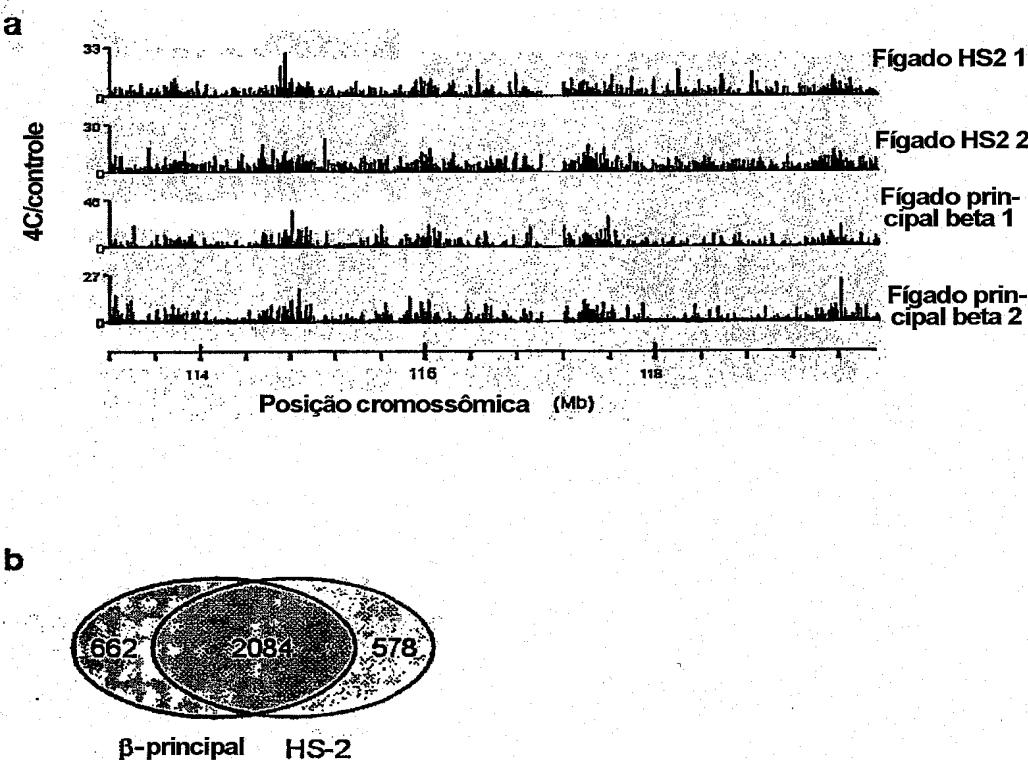


FIGURA 16



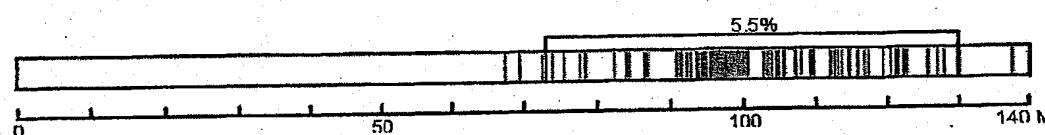
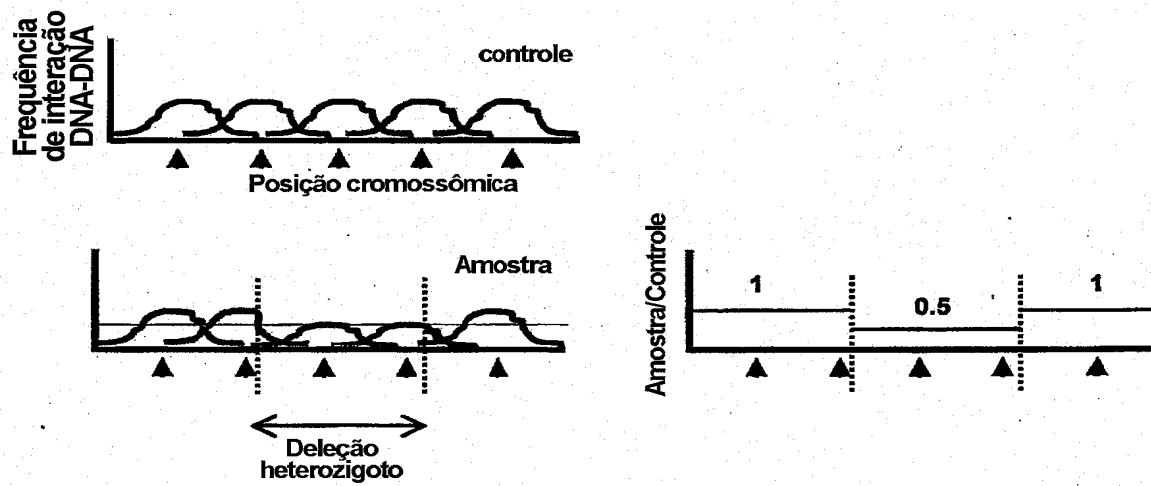
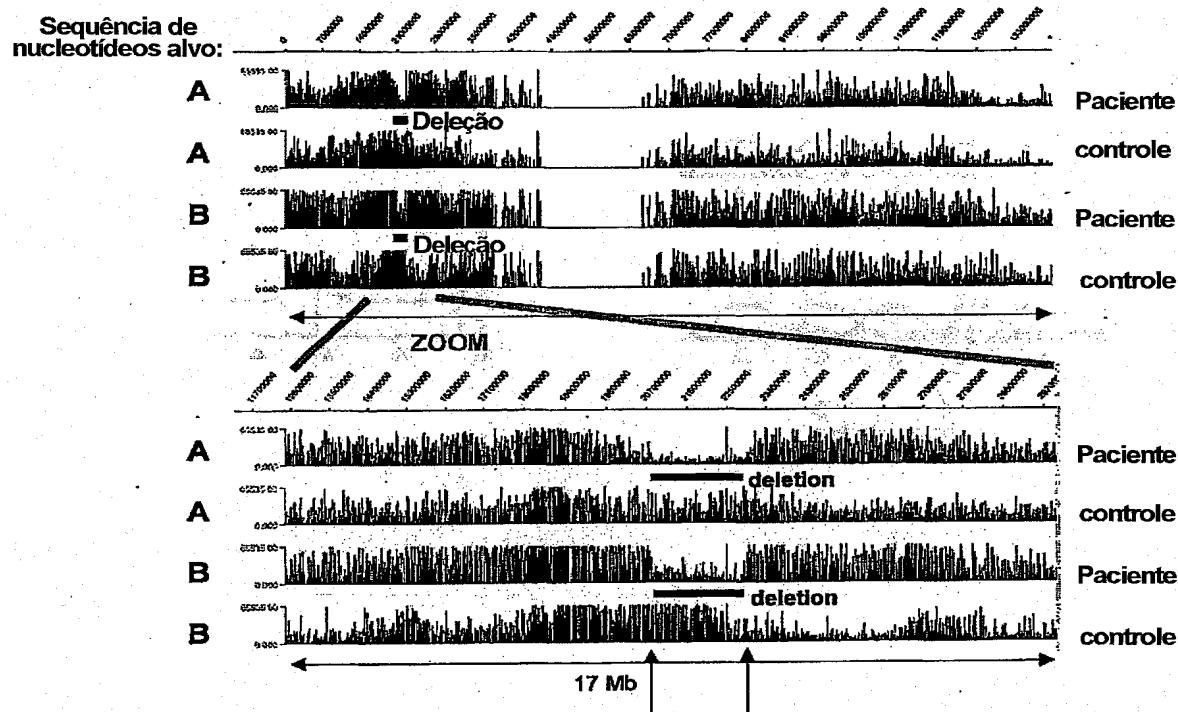
**FIGURA 17****Figado E14.5**

FIGURA 18



## FIGURA 19



### Sequência sobre o ponto de quebra

AAGGAAACCCCATGCCCATAAGACGTCACTAATTCTGAACCTTGTCCCCCCCCCCCCCCCC  
CAAGTGTCTCATCTAAGTAGTTGTTTGTCACTGAGAAAATCAGATATGTTGCTAAAAATTCAACAA  
CTATTGCAAGAAAAATAAAAGAC

FIGURA 20

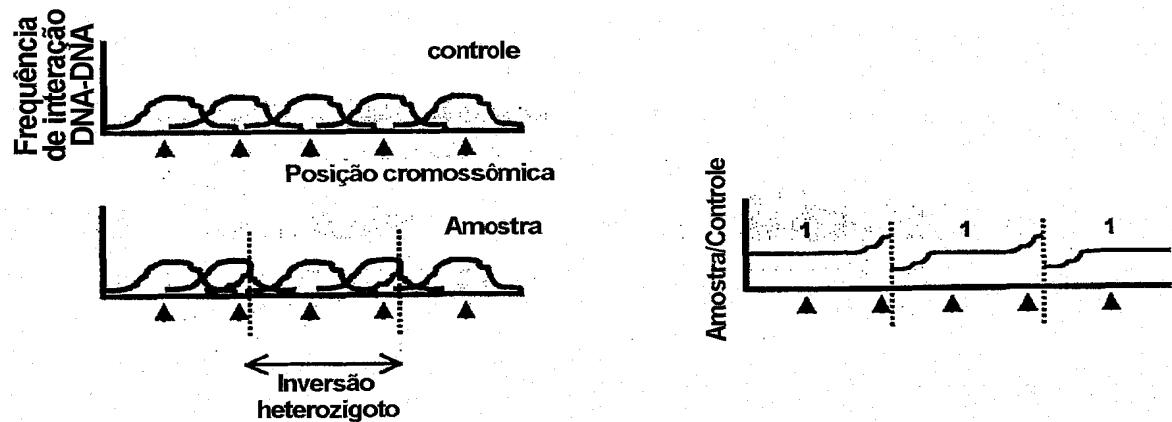


FIGURA 21

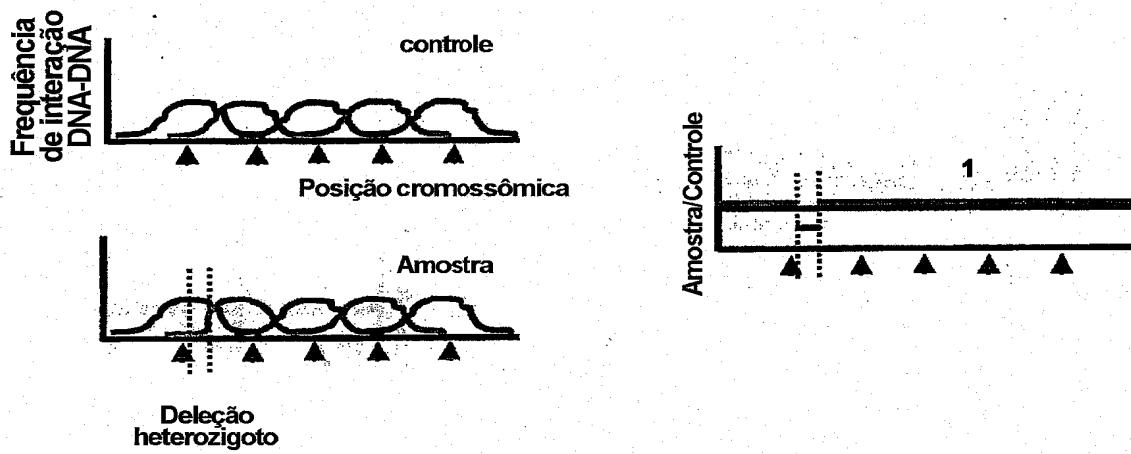


FIGURA 22

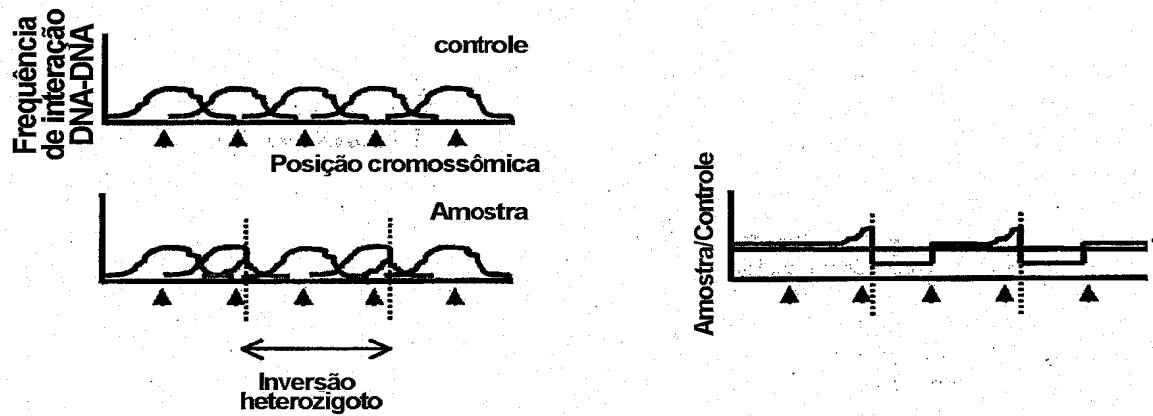


FIGURA 23

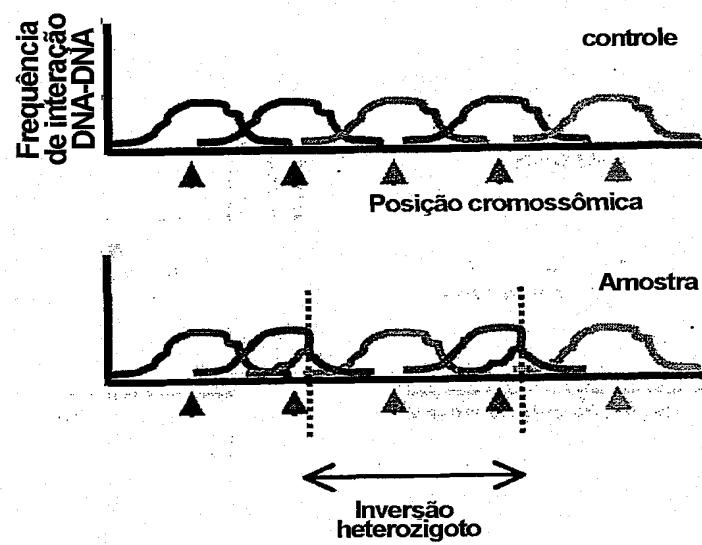
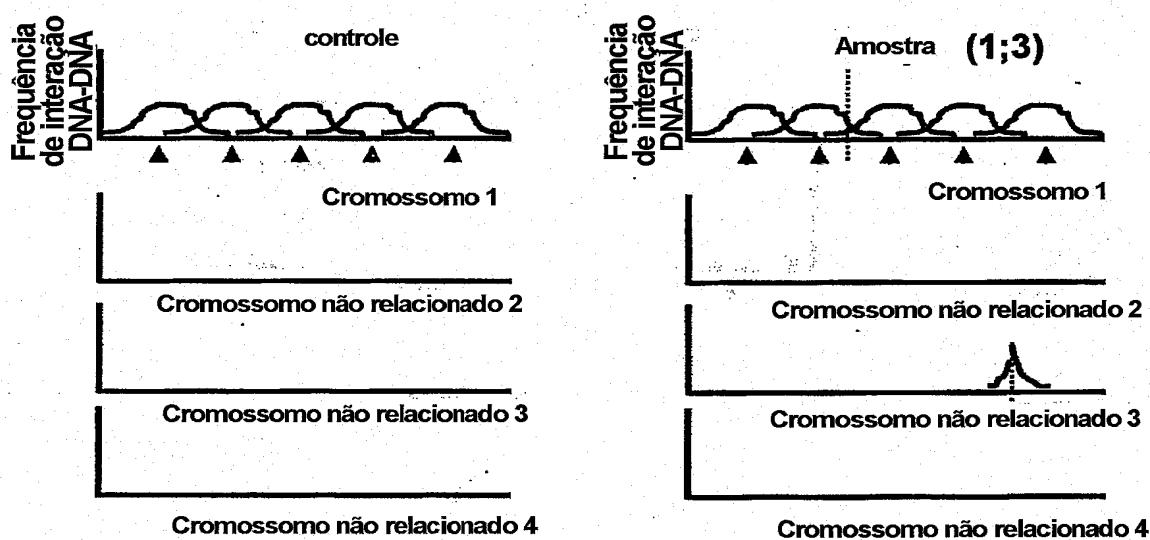


FIGURA 24



**FIGURA 25**

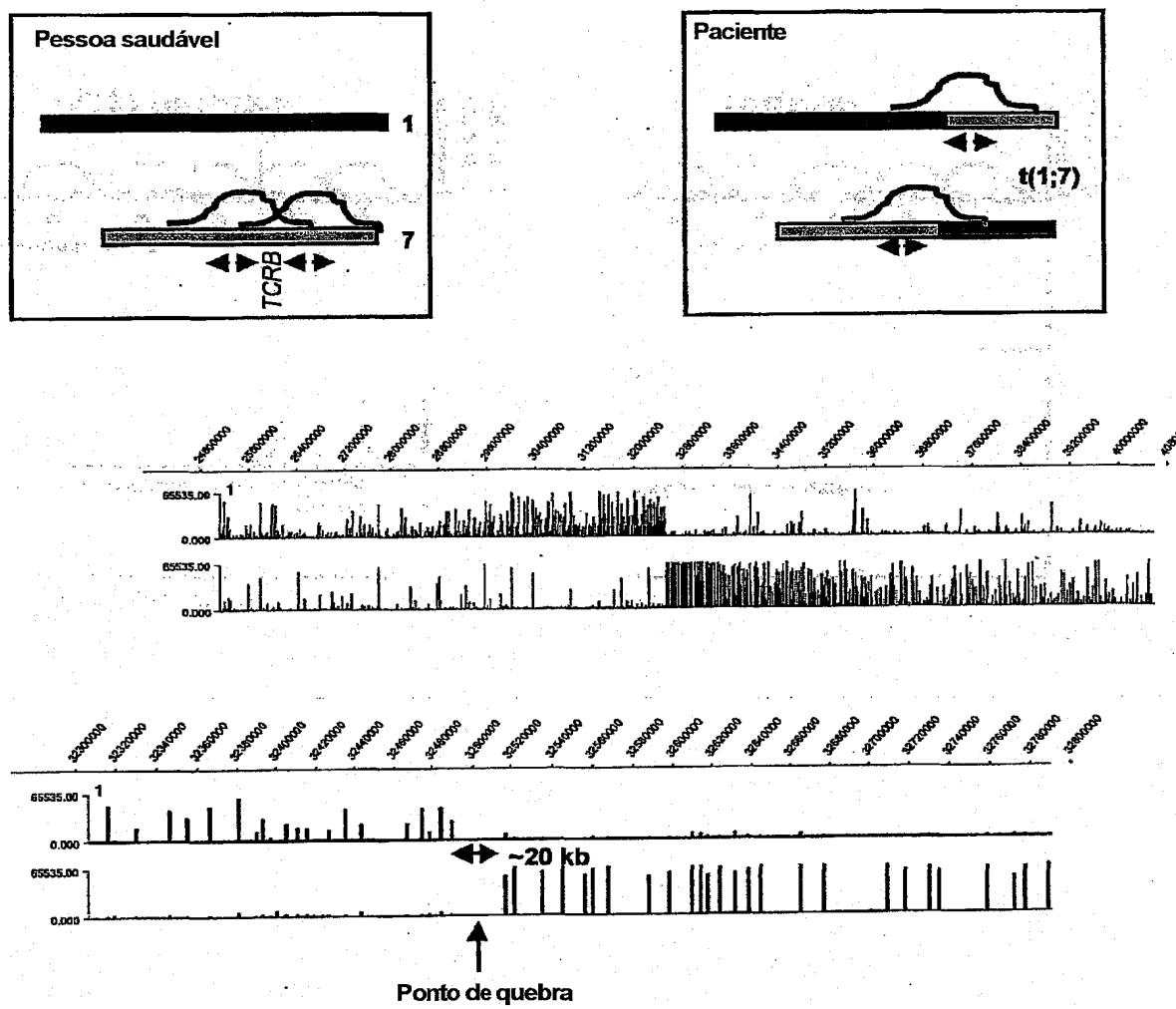


FIGURA 26

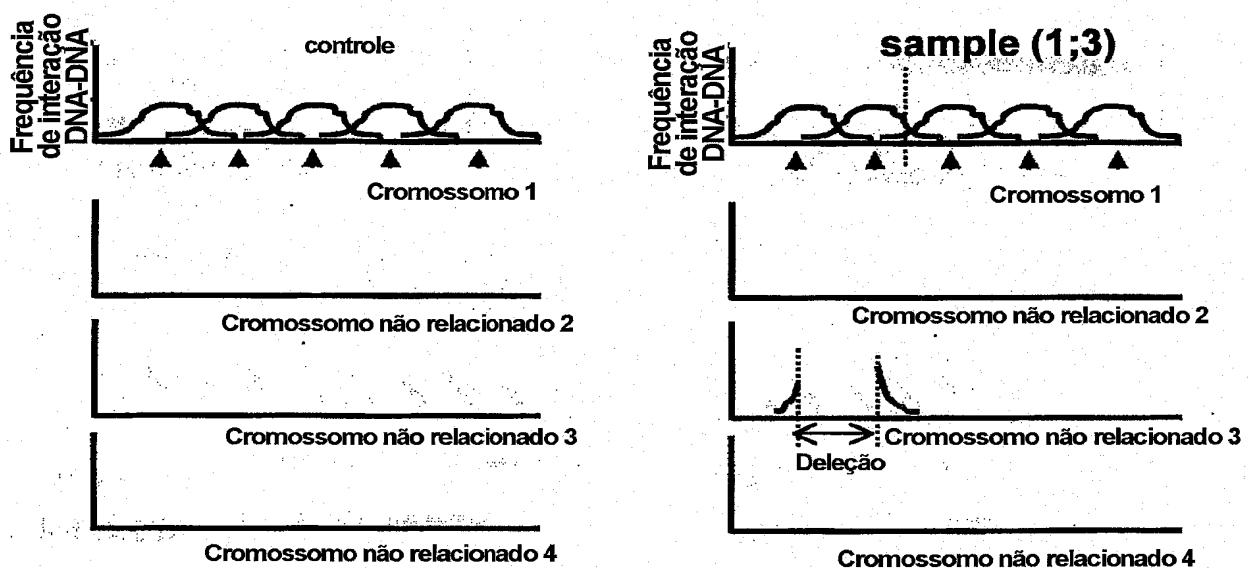


FIGURA 27

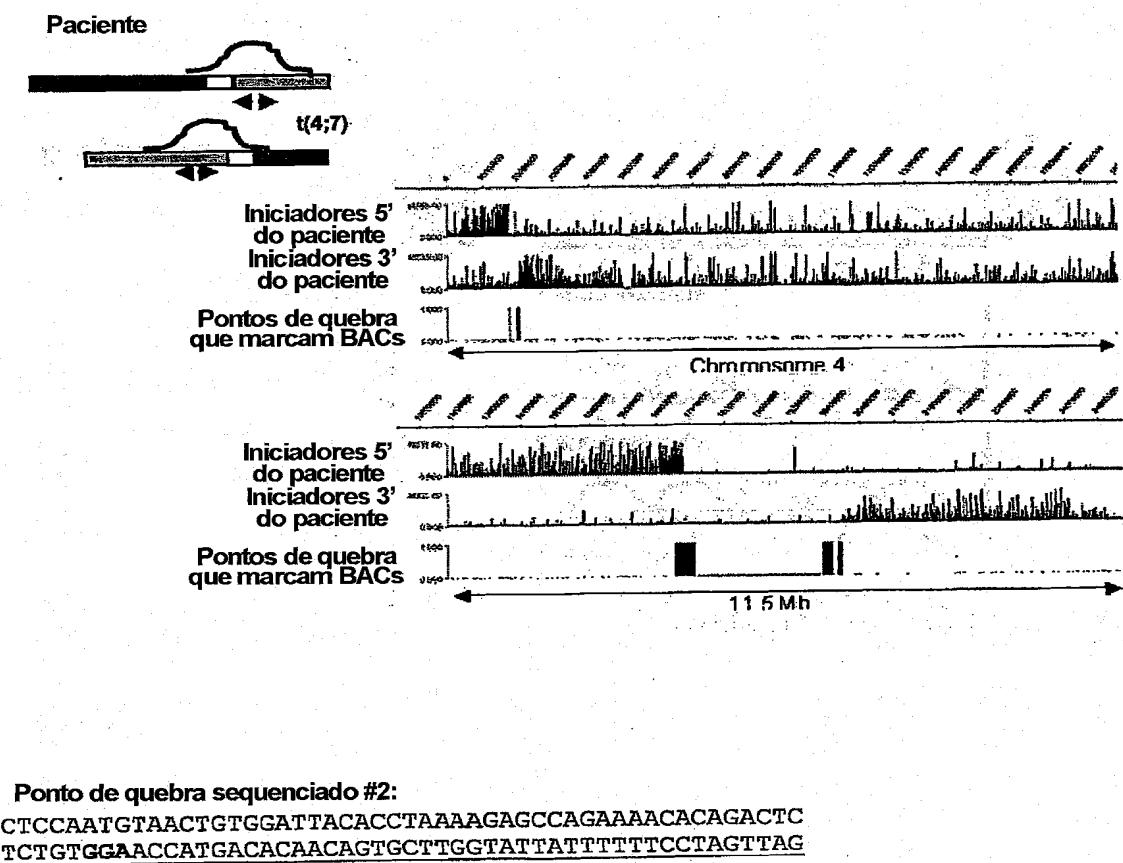


FIGURA 28

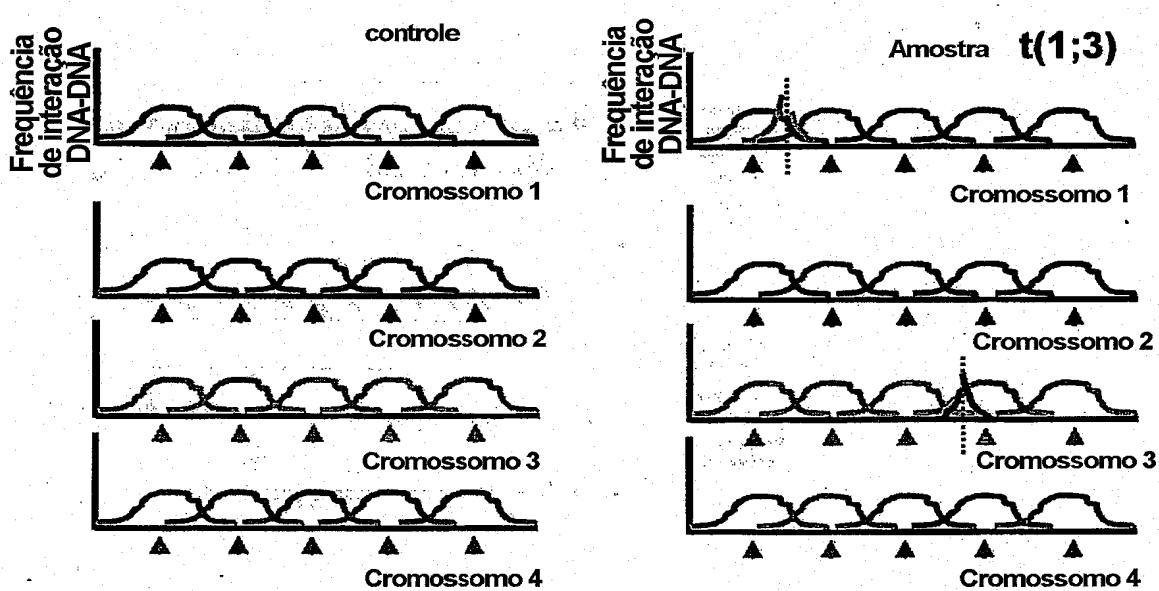
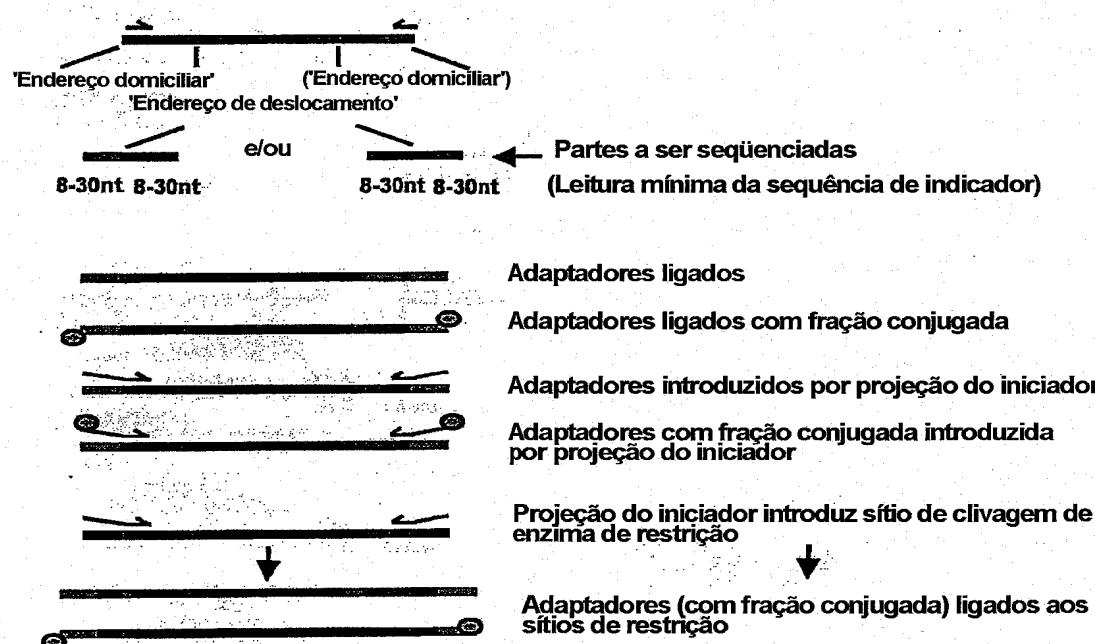


FIGURA 29



**NB.** Adaptadores em cada extremidade do produto de PCR podem ser os mesmos ou podem ser diferentes

### FIGURA 30

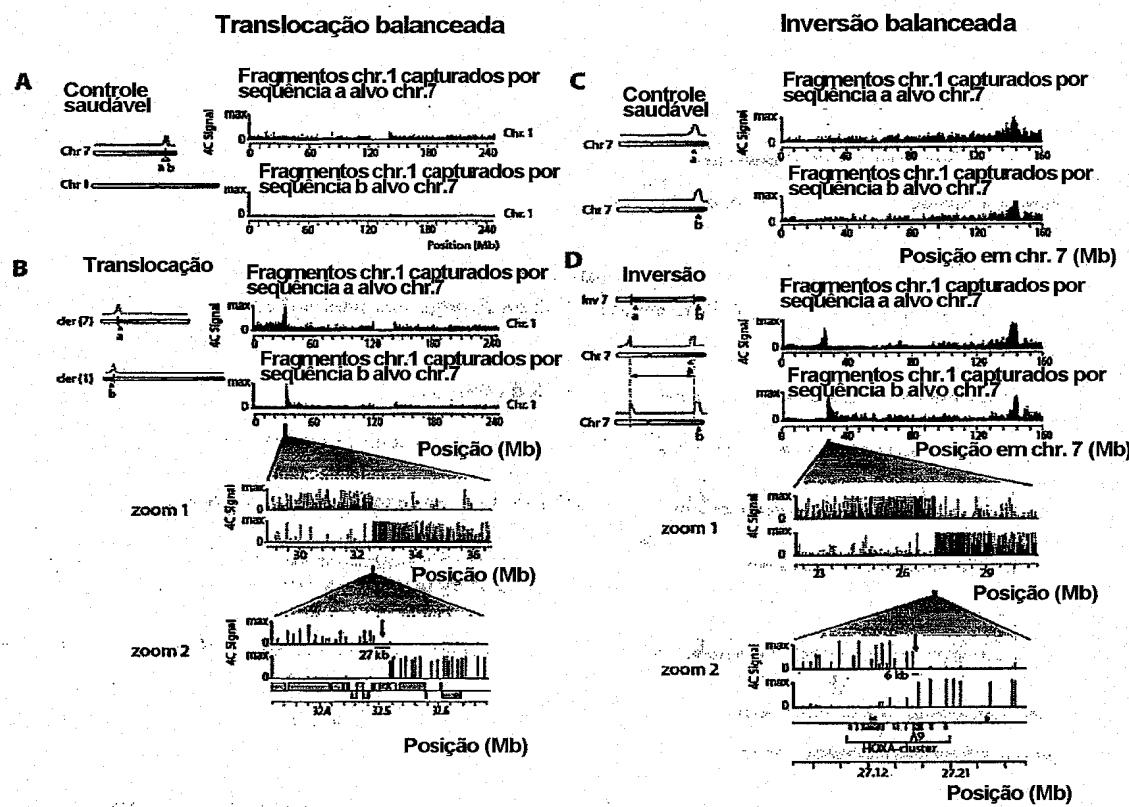
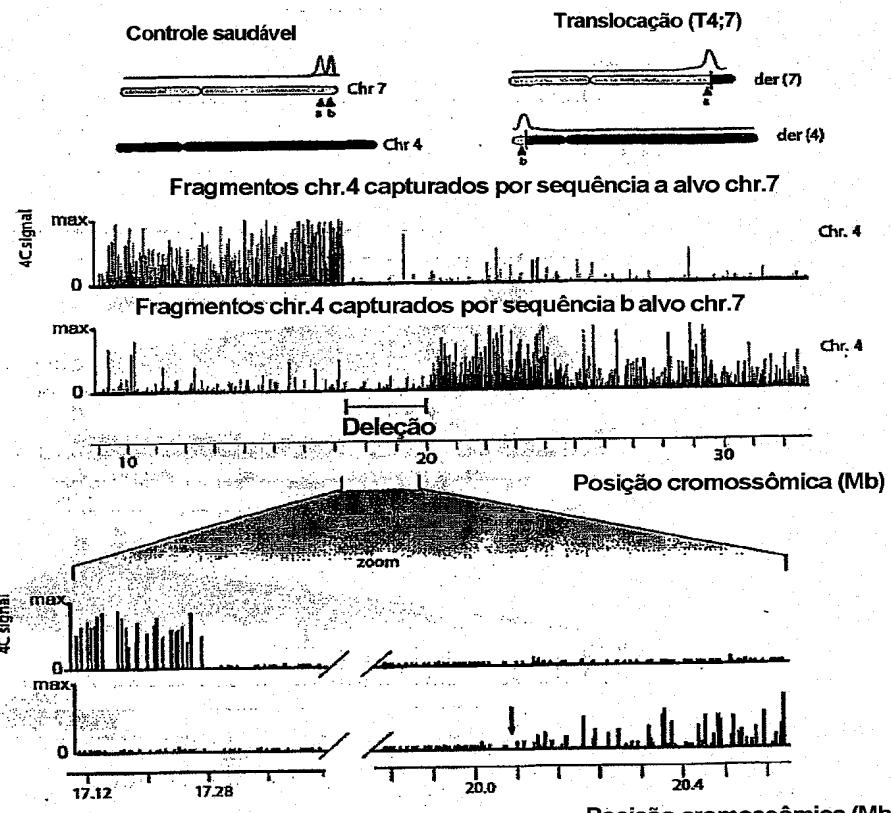


FIGURA 31

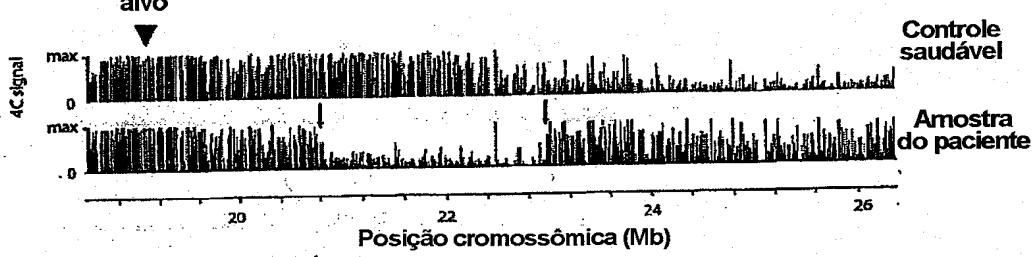
## Translocação não balanceada

**A****B**

aagagccagaaaacacagactcttgtGGAACCATGACACAACAGTGCTTGGTATTATTT der(4)

## Deleção

## Sequência alvo

**C****D**

aaaccccatgccataagacgtcactATTCTGAACCTCTGTTTTTTT

FIGURA 32

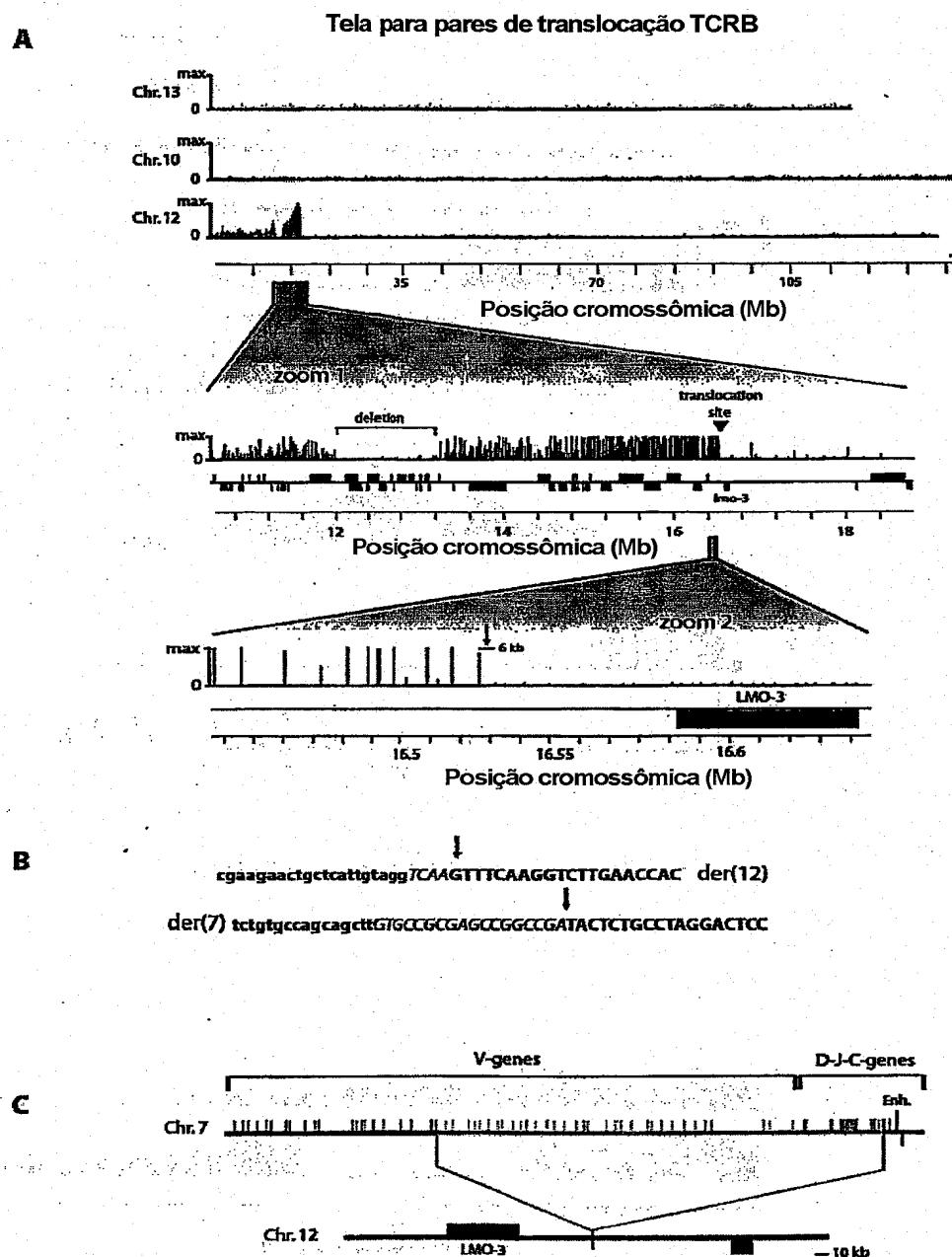


FIGURA 33

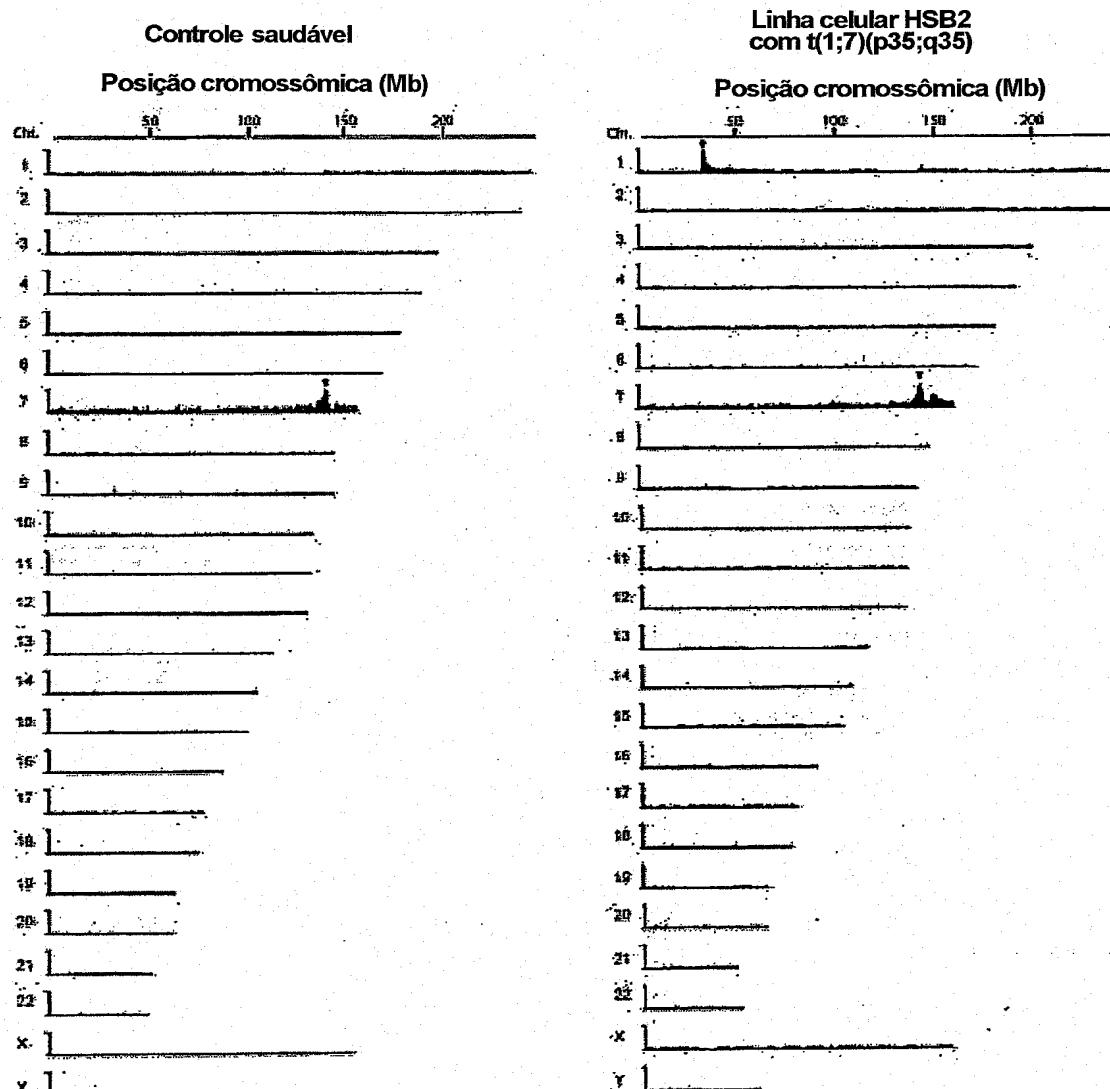


FIGURA 34

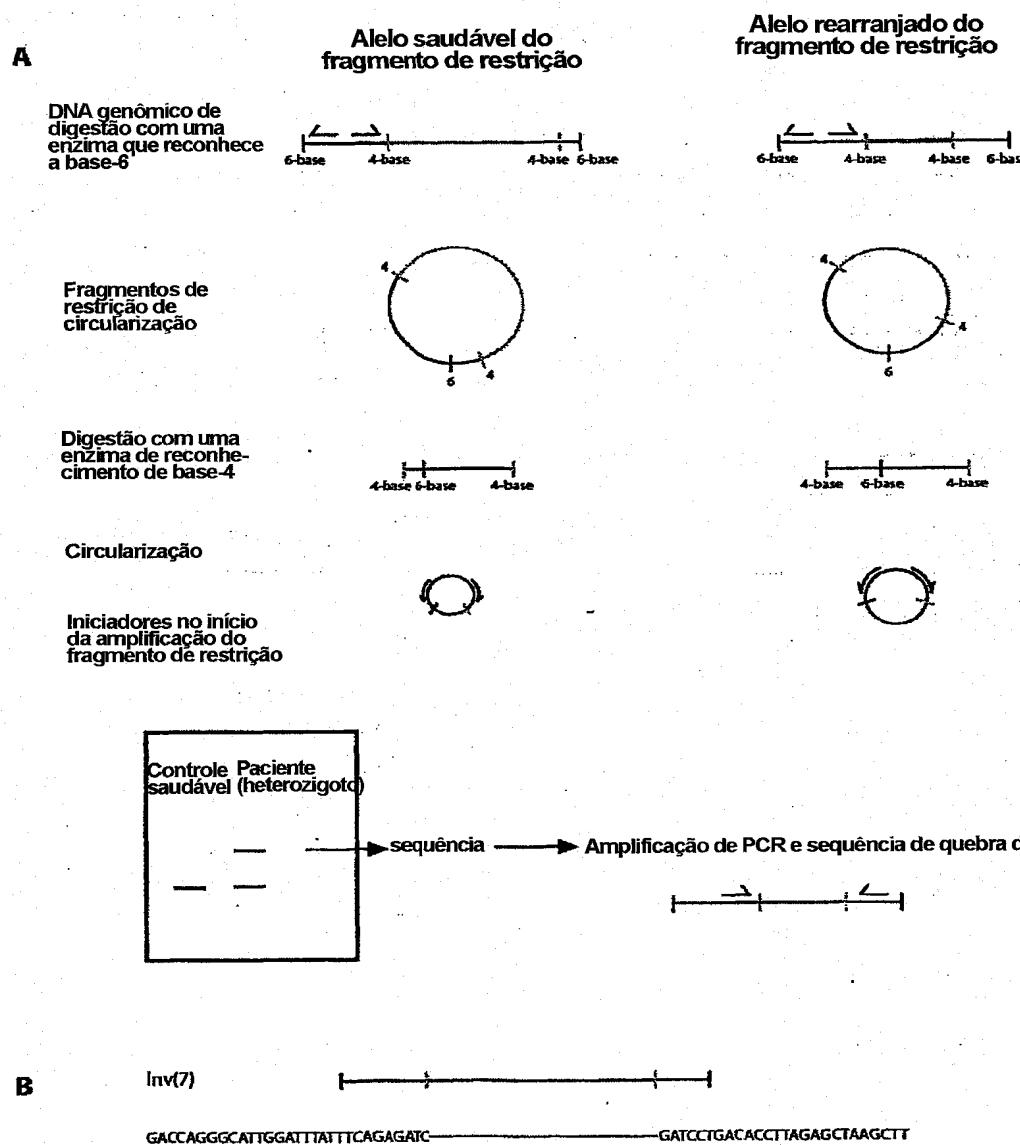


FIGURA 35

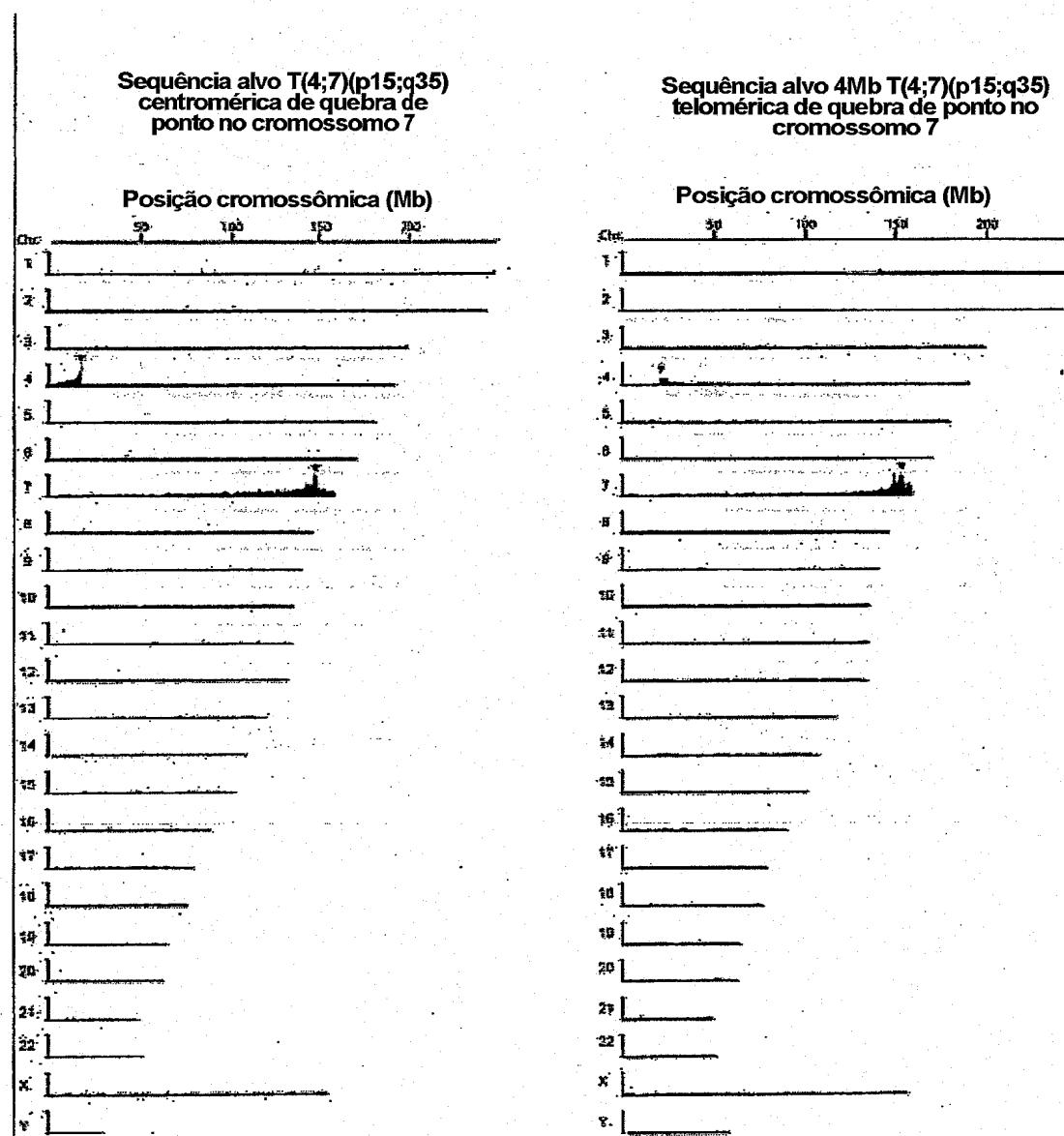


FIGURA 36

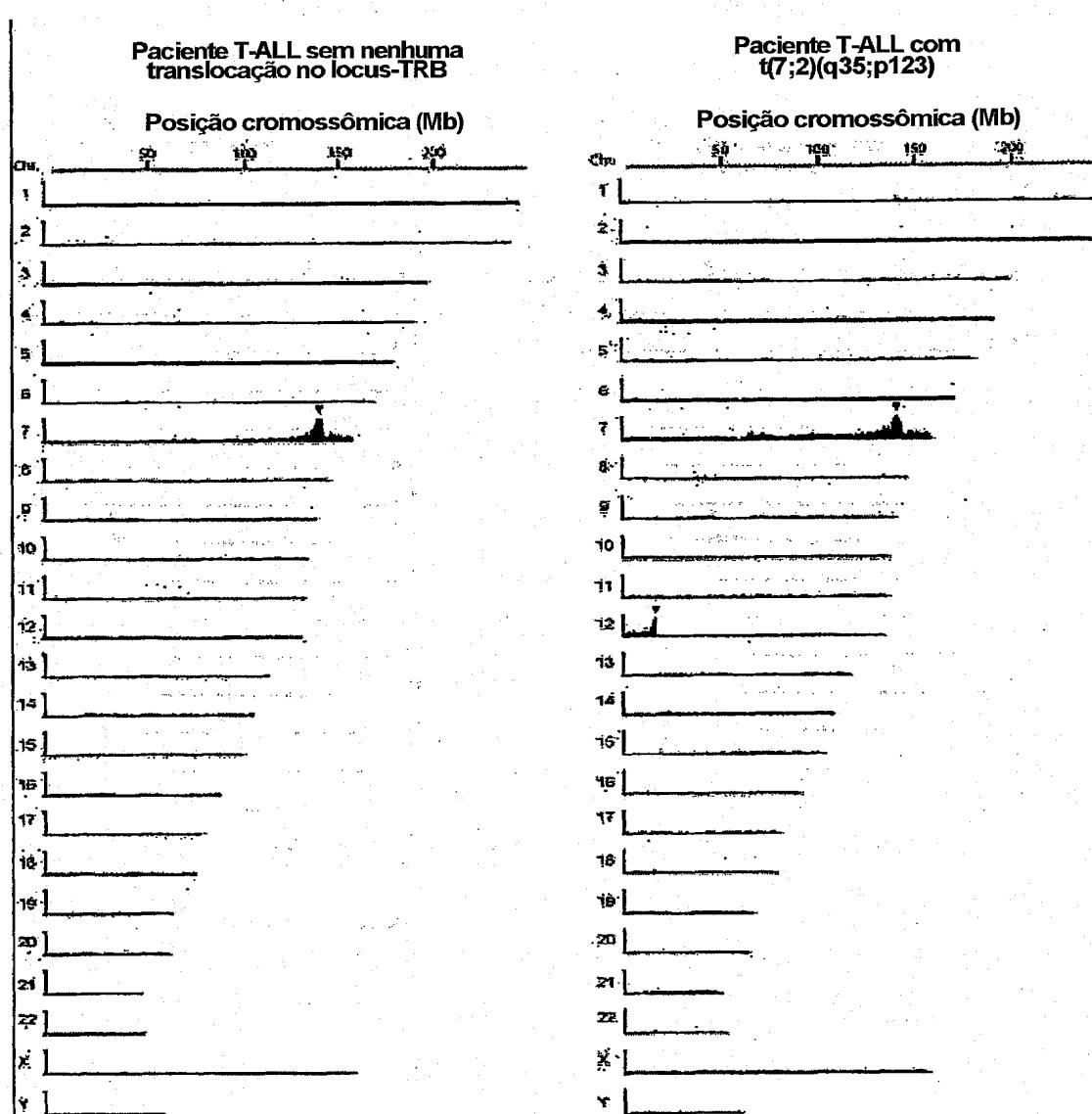


FIGURA 37

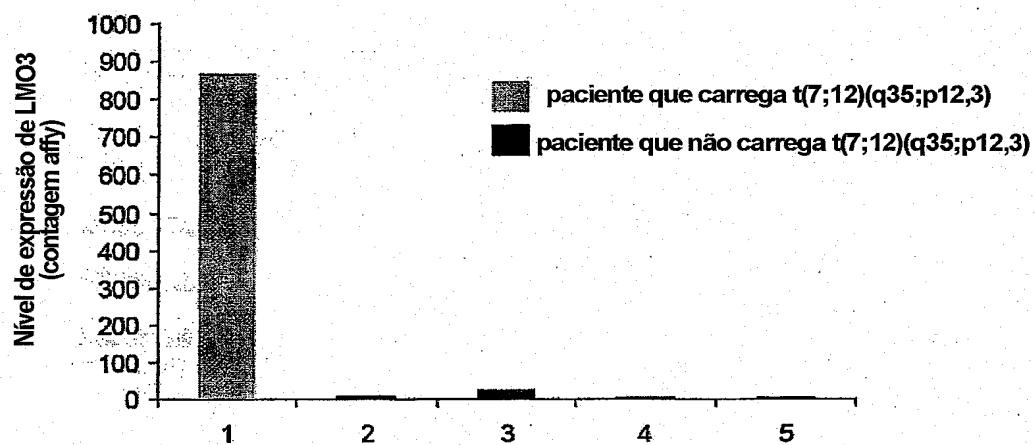


FIGURA 38

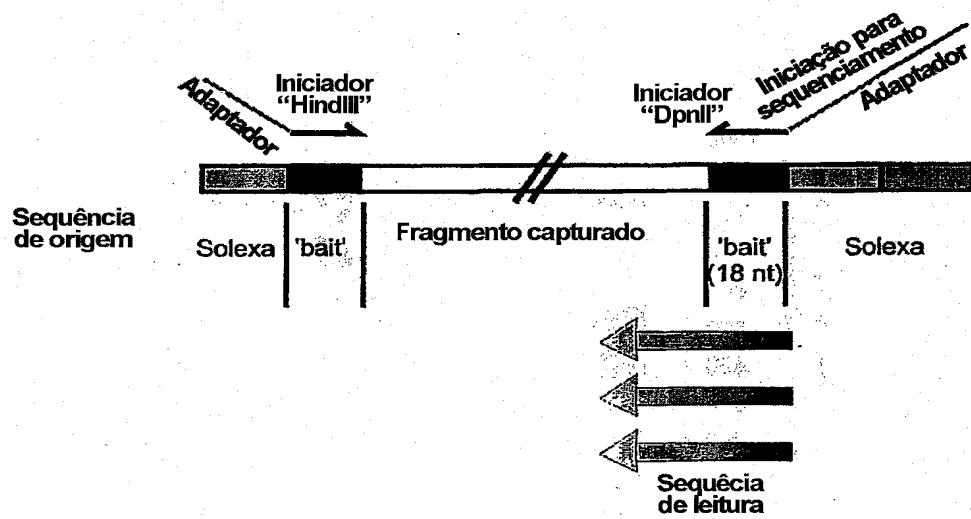


FIGURA 39

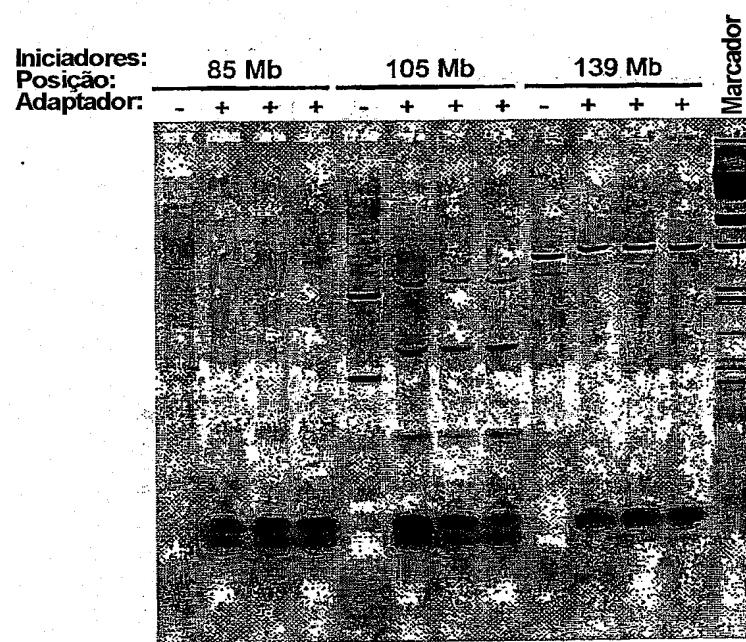


FIGURA 40

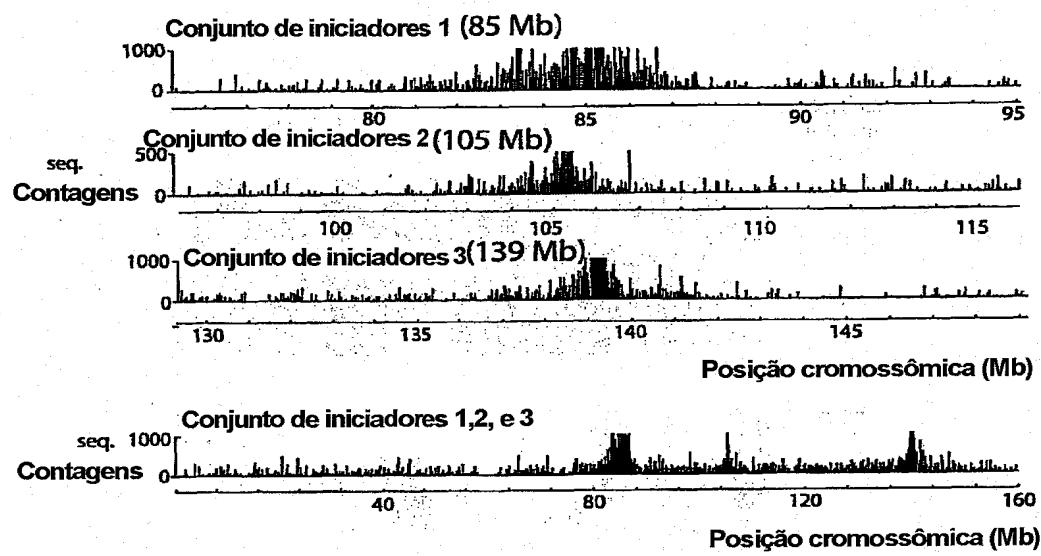
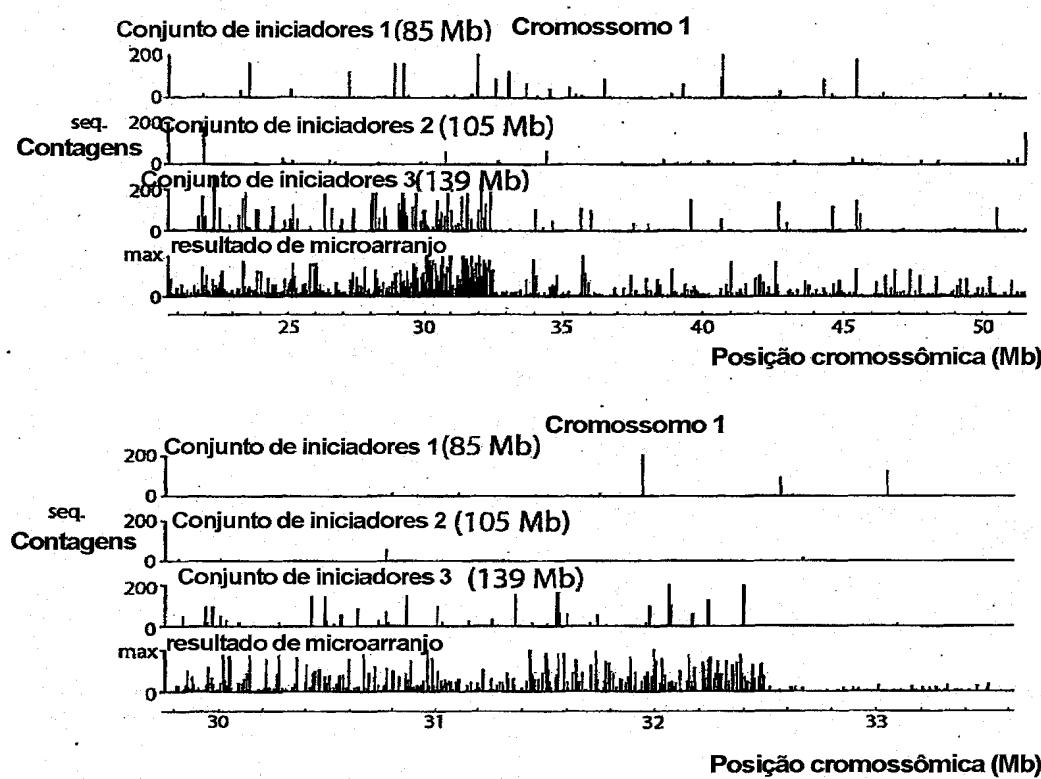


FIGURA 41



## RESUMO

### **“CAPTURA DE CONFORMAÇÃO DE CROMOSSOMO CIRCULAR”**

A presente invenção diz respeito em um aspecto a um método para analisar a frequência de interação de uma sequência de nucleotídeos alvo com uma ou mais sequências de nucleotídeos de interesse (por exemplo, um ou mais loci genômicos) compreendendo as etapas de: (a) fornecer uma amostra de DNA reticulado; (b) digerir o DNA reticulado com uma enzima de restrição primária; (c) ligar as sequências de nucleotídeos reticulados; (d) reverter a reticulação; (e) digerir opcionalmente as sequências de nucleotídeos com uma enzima de restrição secundária; (f) ligar opcionalmente uma ou mais sequências de DNA de composição de nucleotídeo conhecida no(s) sítio(s) de reconhecimento de enzima de restrição secundária disponível(s) que flanqueiam uma ou mais sequências de nucleotídeos de interesse; (g) amplificar uma ou mais sequências de nucleotídeos interesse usando pelo menos dois oligonucleotídeos iniciadores, em que cada iniciador hibridiza nas sequências que flanqueiam as sequências de nucleotídeos de interesse; (h) hibridizar a(s) sequência(s) em uma arranjo; e (i) determinar a frequência de interação entre as sequências de DNA.