Gil Monteiro Novo Filho

**Sequenciamento de nova geração dos pontos de
quebra do DNA para investigação dos mecanismos
de formação em rearranjos genômicos**

Tese apresentada à Faculdade de Medicina
da Universidade de São Paulo para a
obtenção do título de Doutor em Ciências.

Programa de Patologia

Orientadora: Prof^a. Dr^a. Leslie Domenici
Kulikowski

Co-orientadora: Prof^a. Dra. Chong Ae Kim

2018

Gil Monteiro Novo Filho

**Sequenciamento de nova geração dos pontos de
quebra do DNA para investigação dos mecanismos de
formação em rearranjos genômicos**

Tese apresentada à Faculdade de Medicina da
Universidade de São Paulo para a obtenção do
título de Doutor em Ciências.

Programa de Patologia

Orientadora: Prof^a. Dr^a. Leslie Domenici
Kulikowski

Co-orientadora: Prof^a. Dra. Chong Ae Kim

2018

Dados Internacionais de Catalogação na Publicação (CIP)

Preparada pela Biblioteca da
Faculdade de Medicina da Universidade de São Paulo

©reprodução autorizada pelo autor

Novo Filho, Gil Monteiro
Sequenciamento de nova geração dos pontos de
quebra do DNA para investigação dos mecanismos de
formação em rearranjos genômicos / Gil Monteiro Novo
Filho. -- São Paulo, 2018.
Tese (doutorado) -- Faculdade de Medicina da
Universidade de São Paulo.
Programa de Patologia.
Orientadora: Leslie Domenici Kulikowski.
Coorientadora: Chong Ae Kim.

Descritores: 1. Pontos de quebra do cromossomo
2. Rearranjo gênico 3. Análise de sequência 4. DNA
5. Variação estrutural do genoma 6. Sequenciamento de
nucleotídeos em larga escala

USP/FM/DBD-513/18

Responsável: Erinalva da Conceição Batista, CRB-8 6755

*Dedico esse trabalho à minha mãe,
pelo seu amor, dedicação a mim, e
meu maior exemplo de caráter.*

Agradecimentos

À minha filha, Alice Cimatti Monteiro Novo, razão da minha vida, por ter proporcionado momentos de alívio e felicidade infinita pelo simples fato de existir e estar junto de mim.

À minha esposa Mariana Mahmoud Cimatti, por sempre ter estado ao meu lado em todos momentos, por ter me dado a Alice, maior presente, me apoiando e incentivando, e também se sacrificando para me permitir finalizar esse trabalho, fazendo com que cada dia fosse muito mais belo.

Ao meu pai Gil Monteiro Novo, por ter me incentivado e aconselhado nos momentos difíceis, por ter sempre me incentivado aos estudos, ao ceticismo e à leitura e por ter me apoiado em minhas escolhas.

Às minhas irmãs Roberta Monteiro Novo e Flávia Monteiro Novo, por todo companheirismo, suporte e todos os momentos de alegria que me proporcionaram.

À minha tia, Eliane Victório, por ter me acolhido durante todos esses anos, me ajudado em tudo que eu precisei, me feito rir. Por ter sido uma das maravilhosas peças fundamentais da minha formação.

Aos meus sogros, Selma Rita Mahmoud Cimatti e Anselmo Cimatti, e à Fayez Assaad Mahmoud e Celina da Conceição Mahmoud, por todo tipo de apoio que me deram, pelo carinho demonstrado em todos os momentos, especialmente os mais difíceis.

À minha orientadora Prof^a. Dr^a. Leslie Domenici Kulikowski, por ter acreditado e confiado em mim, por sua dedicação, seus conselhos e por sua exímia capacidade de despertar otimismo ante às maiores dificuldades.

À Prof^a. Dr^a. Chong Ae Kim, por ter me acolhido e me dado todo suporte necessário para desenvolvimento e conclusão deste trabalho.

À Marília Montenegro, por toda sua amizade, pelas horas de discussões genéticas intensas, calorosas e, principalmente, edificantes; ao Amom Nascimento, pelas conversas, discussões e elucubrações acerca da vida e da ciência; à Évelin Zanardo, por todo o trabalho imprescindível realizado e pelo companheirismo, a Alexandre Dias, pelos conselhos, amizade e confiança, a Jullian Damasceno, pelas pacientes e altruístas instruções com a bioinformática, a Yanca Gasparini, pela amizade e companheirismo por todos esses anos.

Aos meus amigos do Laboratório de Citogenômica, Lucas Liro, Amanda Brasil, Thaís Moura, Samar Chehimi, Fabrícia Madia, Gleyson, Cristian e Mayara por todo o suporte em todos os sentidos, que foram responsáveis diretos da realização desse trabalho.

Aos médicos da Unidade de Genética Instituto da Criança, Dra. Rachel Honjo e Dra. Débora Bertola, e aos residentes Gustavo, Santiago, Julia, Cristina, Christiane, Carol e Sofia pelo competente atendimento aos pacientes, imprescindível para a realização deste trabalho.

À Fernanda Macaferri, Andréia, Ana Lúcia, Laila, por todo apoio e amizade, tornando os dias de trabalhos definitivamente menos penosos e mais alegres.

Ao Prof. Dr. Raymundo e ao departamento de Patologia da FMUSP, por todo suporte e por tornar possível a realização de parte deste trabalho na Filadélfia.

À FAPESP, CAPES e CNPq por viabilizarem financeiramente este estudo.

Epígrafe

“O anti-intelectualismo tem sido uma ameaça constante, se insinuando na nossa vida política e cultural, alimentado pela falsa noção de que a democracia significa que ‘a minha ignorância é tão boa quanto o seu conhecimento’”.

Isaac Asimov

Sumário

SUMÁRIO

1. INTRODUÇÃO	1
2. OBJETIVOS.....	10
2.1. Objetivos específicos	10
3. MÉTODOS	12
3.1. Casuística.....	12
3.1.1. Desenho do estudo.....	12
3.2. Sequenciamento por “ <i>Sinthetic Long Read</i> ”	13
3.2.1. Análise da integridade das amostras	14
3.2.2. Quantificação e diluição das amostras	14
3.2.3. Reação do GemCode	15
3.2.4. Construção da Biblioteca.....	16
3.2.5. Sequenciamento por HiSeq	19
3.2.6. Processamento e análise dos dados.....	19
3.3. Análise dos <i>arrays</i>	20
3.4. Sequenciamento dos pontos de quebra.....	23
3.4.1. Análise dos dados brutos do sequenciamento Illumina.....	24
4. RESULTADOS E DISCUSSÃO.....	26
4.1. Avaliação por 10X GemCode	26
4.1.1. Avaliação de integridade das amostras para 10X GemCode	26
4.1.2. Análise por Loupe	29
4.1.3. Considerações sobre o sequenciamento utilizando Linked Reads	36
4.2. Análise dos <i>arrays</i> para sequenciamento	37
4.3. Análise do sequenciamento das regiões flanqueadas	38
4.3.1. Paciente G-191	38
4.3.2. Paciente G-396	40
4.3.3. Paciente G-524	40
4.3.4. Paciente G-602	43
4.3.5. Demais amostras e considerações	45
4.4. Mecanismos de Formação de Rearranjos.....	47
5. CONCLUSÕES.....	51

6. REFERÊNCIAS	53
----------------------	----

Resumo e Abstract

Resumo

Novo Filho GM. *Sequenciamento de nova geração dos pontos de quebra do DNA para investigação dos mecanismos de formação em rearranjos genômicos* [tese]. São Paulo: Faculdade de Medicina, Universidade de São Paulo; 2018.

Rearranjos genômicos são alterações estruturais na molécula de DNA e podem ser a causa de inúmeras doenças genéticas. O mecanismo gerador dessas alterações é bem variável. Ele pode ser recorrente, por intermédio de low copy repeats (LCRs), resultando num rearranjo causado por recombinação homóloga não-alélica (NAHR), ou não recorrente, ou seja, sem intermédio de um hotspot. Dentre os mecanismo não recorrentes temos: a junção das extremidades não-homólogas (NHEJ – non-homologous end joining) e a junção mediada por micro-homologia (MMEJ – microhomology-mediated end joining), a replicação em série por deslizamento (SRS), a SRS induzida por quebra (BISRS), a replicação induzida pela quebra de DNA por homologia (MMBIR - microhomology-mediated break induced replication), o enrolamento da forquilha de replicação e mudança de molde de DNA (FoSTeS - fork stalling and template switching). A análise dos pontos de quebra dos rearranjos genômicos pode fornecer informações importantes para uma maior compreensão da arquitetura genômica e seu papel na geração das anormalidades estruturais. O objetivo deste trabalho foi sequenciar os pontos de quebra genômicos a fim de identificar o mecanismo formador das alterações encontradas. Para isso, investigamos o panorama estrutural de 10 pacientes por sequenciamento por meio de linked reads (10X Genomics) e sequenciamos os pontos de quebra previamente identificados por array CytoSNP-12 (Illumina) de 12 pacientes com rearranjos genômicos estruturais por utilizando a captura por Nextera Rapid Capture (Illumina). A investigação por linked reads revelou rearranjos estruturais em 5 pacientes, destacando translocações encontradas em dois pacientes, impossíveis de serem detectadas por metodologias de sequenciamento que não envolva long reads. Foi possível sugerir os mecanismos causadores dessas alterações como NHEJ. O sequenciamento após a captura por Nextera foi capaz de identificar elementos que permitiram definir o mecanismo em três pacientes (NAHR E FoSTeS/MMBIR) e sugerir em mais dois pacientes (NHEJ). Com a estratégia utilizada foi possível sequenciar pontos de quebra por meio do flaqueamento das regiões identificadas por array, identificar os elementos genômicos presentes nos pontos de quebra e os mecanismos formadores dessas alterações.

Descritores: pontos de quebra do cromossomo; rearranjo gênico; análise de sequência de DNA; variação estrutural do genoma; sequenciamento de nucleotídeos em larga escala.

Abstract

Novo Filho GM. *Next Generation Sequencing of DNA breakpoints for investigation of formation mechanisms in genomic rearrangements* [thesis]. São Paulo: “Faculdade de Medicina, Universidade de São Paulo”; 2018.

Genomic rearrangements are structural changes in the DNA molecule and can be the cause of numerous genetic diseases. The mechanisms that generate these alterations can occur in different ways. It can be recurrent, mediated by low copy repeats (LCRs), resulting in a rearrangement caused by non-allelic homologous recombination (NAHR), or non-recurrent, without a hotspot. Among non-recurrent mechanisms there are: non-homologous end joining (NHEJ), microhomology-mediated end joining (MMEJ), serial replication slippage (SRS), break-induced serial replication slippage (BISRS), microhomology-mediated break induced replication (MMBIR) and fork stalling and template switching (FoSTeS). Analysis of the breakpoints of genomic rearrangements may provide important information for a better understanding of genomic architecture and its role in generating structural abnormalities. The aim of this work was to sequence the genomic breakpoints in order to identify the mechanism that formed the alterations found. To do this, we investigated the genomic structure of 10 patients by linked reads (10X Genomics) sequencing and sequenced the breakpoints previously identified by Illumina CytoSNP-12 array of 12 patients with structural genomic imbalances by using Nextera Rapid Capture (Illumina). The research by linked reads revealed structural rearrangements in 5 patients, highlighting translocations found in two patients, impossible to be detected by sequencing methodologies that did not involve long reads. It was possible to suggest the mechanisms causing these changes as NHEJ. The sequencing after capture by Nextera was able to identify elements that allowed us to determine the formation mechanism in three patients (NAHR and FoSTeS / MMBIR) and to suggest in two patients (NHEJ). With the approach employed here, it was possible to sequence breakpoints by flanking the regions identified by array, identifying the genomic elements present at breakpoints and the formation mechanisms of the alterations.

Descriptors: chromosome breakpoints; gene rearrangement; sequence analysis; DNA; genomic structural variation; high-throughput nucleotide sequencing.

Introdução

1. Introdução

Rearranjos genômicos são definidos como alterações estruturais na molécula de DNA, com extensão que pode variar de algumas centenas a milhões de pares de bases (1). Essas alterações contemplam deleções, duplicações, inversões, inserções e translocações e são responsáveis por grande parte das diferenças existentes entre os genomas humanos, podendo ser também a causa de inúmeras doenças genéticas (1-3). Esses rearranjos podem ser simples, quando há apenas um rearranjo com um ponto de quebra, ou complexos, quando há mais que um rearranjo e mais que duas junções de pontos de quebra (4). Os mecanismos causadores de rearranjos podem ser classificados como replicativos, quando são provenientes de um erro durante a replicação, e não-replicativos (5).

As características estruturais do genoma humano podem resultar numa suscetibilidade região-específica a rearranjos, levando então a instabilidade genômica (6). Rearranjos recorrentes ocorrem principalmente devido a presença de regiões conhecidas como *low copy repeats* (LCRs). As LCRs são sequências maiores genômicas de 1 kb de tamanho, com mais de 95% de similaridade, que aparecem mais de duas vezes no genoma (7, 8).

O principal mecanismo causador das alterações recorrentes é a recombinação homóloga não-alélica (NAHR). Esse mecanismo não-replicativo ocorre preferencialmente por intermédio das LCRs, mas pode também ser mediado por L1s, sequências *Alu* ou pseudogenes (9). Essas sequências, devido sua alta similaridade, podem eventualmente se alinhar durante a mitose ou a meiose. Esse alinhamento pode ocorrer tanto entre cromossomos distintos, sejam eles homólogos ou não, quanto entre diferentes *loci* do mesmo cromossomo. Desta maneira, esse alinhamento indevido pode propiciar um cruzamento entre esses dois *loci*, resultando num rearranjo genômico (Figura 1A). O NAHR pode levar a deleções, duplicações e inversões, quando as LCRs estão localizadas num mesmo cromossomo, ou resultar em translocações, quando as LCRs estão localizadas em cromossomos diferentes (1, 6).

Os rearranjos genômicos podem também ocorrer de forma não recorrente, apresentando extensões variáveis e pontos de quebra dispersos, sem que haja, aparentemente, um *hotspot* para que essa recombinação aconteça (9).

Dentre os mecanismos propostos para os rearranjos não recorrentes, estão a junção das extremidades não-homólogas (NHEJ – *non-homologous end joining*) e a junção mediada por micro-homologia (MMEJ – *microhomology-mediated end joining*), a replicação em série por deslizamento (SRS), a SRS induzida por quebra (BISRS), a replicação induzida pela quebra de DNA por homologia (MMBIR - *microhomology-mediated break induced replication*) e o enrolamento da forquilha de replicação e mudança de molde de DNA (FoSTeS - *fork stalling and template switching*) e retrotransposição (1, 9).

O mecanismo não-replicativo NHEJ ocorre quando há uma quebra na dupla fita de DNA sucedido de um reparo pela ligação de ambas as extremidades. Esse reparo não requer a presença de um molde homólogo e, em consequência disso, pode ser notada a presença de pequenas deleções (1-34 pb) e inserções no ponto de quebra. Sequências repetitivas podem proporcionar uma instabilidade genômica, tornando a região suscetível a uma quebra na dupla fita, culminando num reparo por NHEJ (Figura 1B). Já o mecanismo MMEJ se utiliza da excisão de bases para expor regiões de micro-homologia e proporciona o anelamento e posterior junção das extremidades. Esses mecanismos podem gerar deleções e translocações nas fitas envolvidas no rearranjo (10-14). No mecanismo replicativo SRS, o primer e a fita molde podem se dissociar e se reassociar posteriormente de maneira desalinhada, podendo gerar deleções e duplicações, dependendo do sentido em que ocorre o deslizamento (15). O BISRS ocorre quando esse deslizamento é ocasionado por uma quebra na dupla fita de DNA (16).

Figura 1. Comparação entre NAHR e NHEJ. **A** - Os blocos em cor azul representam LCRs que se alinham erroneamente devido à alta similaridade entre elas, resultando em rearranjos que podem causar deleção ou duplicação do segmento que é flanqueado por elas. **B** - Uma quebra na dupla fita entre duas sequências (representadas pelos blocos em azul e vermelho) sem homologia entre elas. O mecanismo de NHEJ modifica e realiza a junção das duas extremidades, causando deleção do segmento entre as duas regiões. (Adaptado de Gu, 2008)(1)

FoSTeS e MMBIR são dois mecanismos replicativos que ocorrem devido a erros na forquilha de replicação. No caso do FoSTeS, um erro na fita simples de DNA interrompe a forquilha de replicação. Para dar continuidade à replicação, a extremidade 3' se desvencilha da fita molde original e se anela em outra forquilha, podendo ser adjacente ou em proximidade tridimensional, que apresente micro-homologia à região. Desta maneira, a fita recomeça sua síntese na forquilha invadida. Este processo pode se repetir diversas vezes, gerando regiões deletadas (caso a outra forquilha de replicação invadida esteja posicionada posteriormente em relação à sequência original), duplicadas (caso a outra forquilha de replicação invadida esteja posicionada anteriormente em relação à sequência original) e invertidas (dependendo da direção da progressão da forquilha invadida). Esses três tipos de rearranjos podem estar presentes concomitantemente num único fragmento, iniciado por um único evento (1, 17, 18).

Já no MMBIR, quando ocorre o erro na dupla fita de DNA interrompendo a forquilha de replicação, esta entra em colapso. Isto resulta numa cauda 3' de

fita simples (ssDNA) gerada pela quebra da dupla fita. Essa extremidade 3' se anela por micro-homologia a outras ssDNAs expostas (2-5 nucleotídeos). A extensão por este primeiro molde ocorre em baixa capacidade de processamento, o que pode gerar um novo colapso, acarretando numa mudança do molde, gerando um ciclo. Finalmente, a fita rearranjada retorna à cromátide irmã original, restabelecendo a replicação. Como resultado final, temos uma translocação não recíproca complexa. (14, 17).

Figura 2. Comparação entre FoSTeS (A) e MMBIR(B). **A)** Uma quebra na fita lagging durante a replicação (a1) pode levar à formação de uma estrutura secundária (a2) que impede que a replicação prossiga. A extremidade 3' fica então livre (a3) podendo se alinhar com outra fita simples exposta numa forquilha de replicação próxima ou adjacente, que apresente micro-homologia. **B)** A forquilha de replicação entre em colapso (b1) e se separa da mesma (vermelha), expondo uma cauda na extremidade 3' (b2). Essa extremidade pode se anelar à fita *lagging* de outra forquilha de replicação por micro-homologia (azul) (b3) que pode ser estabilizada com a síntese de das duas fitas *leading* e *lagging* a partir da junção de micro-homologia (b4). A fita, que foi copiada de uma sequência diferente da sua (azul), se separa da forquilha que utilizou como molde e gera uma nova cauda 3' que irá então se anelar novamente a uma outra fita simples também por micro-homologia, repetindo o processo (b5). Esta fita estendida se separa e anela novamente com outra fita simples da molécula que estava anteriormente (vermelha) (b6). Uma nova forquilha se forma (b7), agora totalmente funcional que continua até o final da extensão da molécula (b8), carregando consigo sequências curtas de outras localidades genômicas. (adaptado de Hastings, 2009)(5).

Diante das suas características, podemos identificar os mecanismos formadores de rearranjos através do perfil das alterações genômicas estruturais. Para que se caracterize um rearranjo mediado por NAHR, elementos como LCRs, L1s, sequências *Alu* ou pseudogenes devem estar presentes próximos aos pontos de quebra. Os mediados por NHEJ deve apresentar uma “cicatriz” na junção do ponto de quebra, com a perda ou a inserção de randômica de alguns nucleotídeos. O MMEJ deve apresentar micro-homologia (5–25 bp), além dos mesmos elementos observados no ponto de quebra do NHEJ (1, 9, 14, 19-21). Os mecanismos replicativos preveem "assinaturas" típicas, como a presença de inserções e micro-homologias nas junções dos pontos de quebra. O rearranjo por FoSTeS pode utilizar moldes de mais de 100 kb distantes do ponto de quebra, causando rearranjos de grande escala enquanto o SRS utiliza moldes adjacentes e causa rearranjos menores. O MMBIR se distingue pela identificação da quebra na dupla fita de DNA que expõe uma cauda 3', a qual é utilizada para iniciar a síntese numa forquilha de replicação distante (1, 9, 14, 19, 20).

Nos últimos anos surgiram relatos de que os rearranjos genômicos podem ser originados por um único evento catastrófico, gerando de dezenas a centenas de translocações ou alterações no número de cópias. Esses eventos podem ser de três diferentes classes, conhecidas como *chromothripsis*, *chromoanasythesis* e *chromoplexy* (18).

Chromothripsis designa a quebra em pedaços de um local num cromossomo num único evento seguido pela junção desses pedaços novamente num cromossomo rearranjado. Essa junção de segmentos ocorre principalmente por NHEJ e MMEJ, gerando assim deleções e inversões, mas nunca duplicações (18, 22).

Chromoanasythesis se dá quando os rearranjos advindos deste evento são restritos a um único cromossomo. A análise dos pontos de quebra revela regiões de micro-homologias, além de inserções que podem variar de 100 a 1000 pb. A *chromoanasythesis* é causada principalmente por FoSTeS e MMBIR

podendo, portanto, resultar em deleções, duplicações, triplicações e inversões, gerando a síntese de um novo cromossomo (23).

Finalmente, o *chromoplexy* é caracterizado por uma cadeia de translocações envolvendo múltiplos cromossomos, resultando em pouca ou nenhuma alteração no número de cópias. Essas translocações são geradas por múltiplas quebras na dupla fita de DNA em um único evento. O mecanismo pelo qual o *chromoplexy* é formado ainda é desconhecido, tendo sido observado em câncer de próstata (18, 24, 25).

A identificação e determinação desses mecanismos geradores de rearranjos genômicos requer dados precisos e minuciosos. Técnicas de *arrays* de alta densidade, como *arrays* por hibridização genômica comparativa (CGH *arrays*) e SNP *microarrays*, são capazes de identificar diferenças de ganho e perda de número de cópias mesmo entre fragmentos muito próximos no genoma. Entretanto esses testes são incapazes de identificar translocações e inversões, dados importantes na caracterização do tipo de mecanismo gerador de uma determinada alteração (26). As técnicas provenientes da utilização de FISH (Hibridação *In Situ* por Fluorescência), apesar de capazes de identificar translocações e inversões, são laboriosas e demoradas. Outrossim, elas não possuem a acurácia necessária para o mapeamento adequado dos pontos de quebra, principalmente devido ao tamanho de suas sondas (27, 28).

Portanto, a investigação e o mapeamento dos pontos de quebra para identificação dos mecanismos causadores dos rearranjos é realizada por meio do sequenciamento dos pontos de quebra gênicos (2, 26, 27, 29). Apesar deste sequenciamento poder ser realizado por eletroforese capilar (26), técnicas de sequenciamento de nova geração (SNG) têm sido utilizadas com sucesso em diversos estudos para caracterização das alterações e mapeamento dos pontos de quebra para identificação dos mecanismos de formação dos rearranjos (30-32). As técnicas de SNG são mais econômicas, rápidas e menos trabalhosas quando comparadas ao sequenciamento por capilar. Esse fato faz com que seja possível sequenciar mais pares de base em menos tempo, diminuindo o custo (33, 34). Dentre essas técnicas, a tecnologia desenvolvida pela Illumina tem se mostrado muito eficiente, com maior capacidade e maior facilidade de análise.

Essa técnica consiste na fragmentação do DNA em pedaços de poucas centenas de pares de base (*short reads*) para posterior alinhamento desses fragmentos. Quanto menor o fragmento gerado anteriormente à corrida, mais complexo é o alinhamento, principalmente quando há grandes deleções, duplicações, inversões ou translocações envolvidas (35).

Entretanto, ao fragmentar o DNA, apesar de ser possível visualizar as *reads* quiméricas e identificar o ponto de quebra, corre-se o risco de não podermos identificar grandes rearranjos, principalmente translocações.

Para solucionar esse problema, há metodologias de sequenciamento de *long reads*, como a desenvolvida pela Pacific Bioscience, onde o DNA de alto peso molecular (até 40 kb) é sequenciado diretamente pela metodologia SMRT (do inglês, *Single Molecule Real Time*).

Entre outros processos recentemente desenvolvidos, podemos citar a tecnologia de *long reads* sintéticas, chamada de *linked reads*, pela 10X Genomics. Nessa metodologia, o DNA de alto peso molecular (até 40-100 kb) passa inicialmente por um processo de “etiquetagem” antes da fragmentação. Essas marcações moleculares perdurarão por todo processo posterior de construção da biblioteca de *short reads* pela metodologia SureSelect (Agilent) e, posteriormente, serão utilizadas para recompor a molécula original (anterior a fragmentação) para identificar sua estrutura (35, 36).

Assim, nosso estudo teve como principal objetivo investigar os mecanismos formadores de rearranjos utilizando novas metodologias de SNG em amostras de DNA de pacientes com alterações cromossômicas estruturais.

Para isso, utilizamos duas tecnologias distintas: *long reads* sintéticas para rearranjos cromossômicos identificados por citogenética clássica e/ou *arrays*, e captura de regiões alvo onde houve quebra identificadas por *array* para sequenciamento do ponto de quebra específico.

A identificação da estrutura da molécula do DNA por metodologias moleculares e o esclarecimento dos mecanismos responsáveis pelos rearranjos

estruturais são cruciais para a compreensão da influência da arquitetura genômica em estudos clínicos.

Objetivos

2. Objetivos

O presente estudo tem como objetivo geral:

- Investigar os mecanismos formadores de rearranjos utilizando novas metodologias de SNG em amostras de DNA de pacientes com alterações cromossômicas estruturais.

2.1. Objetivos específicos

- Identificar elementos genômicos presentes nos pontos de quebra e regiões com alterações.
- Relacionar os elementos encontrados com o tipo de mecanismo formador das alterações.
- Padronizar, implantar e desenvolver o método de análise de sequenciamento de nova geração no laboratório de Citogenômica.

Métodos

3. Métodos

3.1. Casuística

A casuística foi composta de 12 pacientes selecionados em acompanhamento no ambulatório de Genética do Instituto da Criança do HC–FMUSP portadores de rearranjos genômicos estruturais previamente avaliados pela metodologia de *array* genômico (Illumina CytoSNP-12, Affimetrix 6.0 ou Agilent 105K) e 21 pacientes com rearranjos genômicos estruturais previamente identificados por cariotipagem em bandamento G.

O laboratório possui os termos de consentimento previamente aprovado pela Comissão de Ética para Análise de Projetos de Pesquisa do Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo, CAPPesq (sob o nº 0718/11), autorizando o armazenamento e futura utilização do material genético.

E o nosso estudo subsequente foi também aprovado pela CAPPesq (sob o nº 881.661).

3.1.1. Desenho do estudo

Os 21 pacientes previamente analisados por cariotipagem por bandamento G foram pré-selecionados para investigação por *linked reads* (10X Genomics) e, destes, 10 foram selecionados para a realização do experimento, baseado na integridade das amostras e nas características estruturais apresentadas. O critério de seleção para essas 10 amostras está elucidado na seção dos resultados.

Paralelamente, os 12 pacientes pré-avaliados por *arrays*, tiveram seus pontos de quebra flanqueados para a captura por Nextera Rapid Capture Custom e foram sequenciados por MiSeq para avaliação dos pontos de quebra.

3.2. Sequenciamento por “*Sinthetic Long Read*”

O experimento foi realizado com auxílio do *Center for Applied Genomics* do *Children's Hospital of Philadelphia*, Filadélfia, PA, EUA, com o intuito de, utilizando um novo método de “etiquetagem” do DNA de alto peso molecular (com fragmentos maiores que 40 kb), compreender a estrutura genômica e cromossômica da molécula. Assim, visamos delimitar o ponto de quebra dos rearranjos identificados nos cariótipos e, ainda, buscamos visualizar possíveis rearranjos estruturais não identificados pela citogenética clássica e ocultos na metodologia de *oligoarrays*.

Para isso, utilizamos um protocolo inédito que permite “etiquetar” o genoma de modo que se saiba o *locus* de proveniência de uma *short read*.

Em uma breve descrição do método, utiliza-se 1 ng de DNA genômico de alto peso molecular (40~100 kb) distribuídos e particionados em mais de 100.000 gotículas, onde esses grandes fragmentos são etiquetados (marcados com uma sequência específica única de 16pb) e amplificados aleatoriamente. Essas moléculas, agora marcadas, são então liberadas das gotículas, podendo ser fragmentadas e sequenciadas normalmente, de acordo com o teste empregado, por exemplo, para sequenciamento por painéis, exoma ou genoma completo. As *short reads* que pertencem à mesma molécula e estão próximas umas às outras estarão com a mesma marcação (chamadas então de *linked reads*), tornando possível o alinhamento e a identificação de sua localização original no genoma (35, 36).

Iniciamos o procedimento de análise de integridade das amostras de todos os pacientes do estudo e, depois, selecionamos 10 amostras para seguir com a etiquetagem e posterior sequenciamento. Essa seleção foi feita visando garantir a maior heterogeneidade possível dos casos em busca de uma identificação de diferentes mecanismos causadores de alteração. Os detalhes dos critérios para seleção encontram-se nos resultados.

3.2.1. Análise da integridade das amostras

Em uma breve descrição do método, utiliza-se 1 ng de DNA genômico de alto peso molecular (40~100 kb) distribuídos e particionados em mais de 100.000 gotículas, onde esses grandes fragmentos são etiquetados (marcados com uma sequência específica única de 16pb) e amplificados aleatoriamente. Essas moléculas, agora marcadas, são então liberadas das gotículas, podendo ser fragmentadas e sequenciadas normalmente, de acordo com o teste empregado, por exemplo, para sequenciamento por painéis, exoma ou genoma completo. As *short reads* que pertencem à mesma molécula e estão próximas umas às outras estarão com a mesma marcação (chamadas então de *linked reads*), tornando possível o alinhamento e a identificação de sua localização original no genoma (35, 36).

A integridade das amostras foi analisada por meio do kit “screen tape high-sensitivity D1000” (Agilent) de acordo com as orientações do fabricante.

Inicialmente, os reagentes (High Sensitivity D1000 Ladder, High Sensitivity D1000 Sample Buffer) foram colocados à temperatura ambiente por 30 minutos e homogeneizados em vórtex. Em um microtubo de 200 µL foram adicionados 2 µL do High Sensitivity D1000 Sample Buffer e 2 µL High Sensitivity D1000 Ladder. Para cada uma das amostras, foram adicionados 2 µL do High Sensitivity D1000 Sample Buffer e 2 µL em tubos de 200 µL. Os tubos foram então homogeneizados em vórtex, por um minuto, a 2000 RPM e, depois, rapidamente centrifugados para que as amostras se posicionassem no fundo do tubo. Os tubos foram posicionados no 2200 TapeStation (Agilent) para a avaliação dos resultados.

3.2.2. Quantificação e diluição das amostras

Todas as amostras elegíveis para o experimento foram quantificadas por Qubit® 3.0 Fluorometer (Thermo Fisher Scientific) de acordo com as orientações do fabricante. Dessa quantificação, diluiu-se todas as amostras para 20 ng/µL e

o resultado dessa diluição foi quantificada por Qubit em duplicata. Os resultados dessa quantificação foram então diluídos para 1 ng/μL.

Primeiramente, foi preparada uma "solução de trabalho", diluindo-se em tubos cônicos de 15 mL, para cada 8 amostras, 10 μL do Qubit® dsDNA HS Reagent em 1990 μL do Qubit® dsDNA HS Buffer. Depois, adicionou-se 190 μL da solução de trabalho em 2 microtubos diferentes de 1,5 mL e então 10 μL do padrão 1 em um tubo e 10 μL do padrão 2 no outro. Homogeneizou-se em vortex. Independentemente, para cada amostra, foi separado um tubo onde adicionou-se 197 μL da solução de trabalho e 3 μL de amostra. Homogeneizou-se em vórtex por 3 segundos e incubou-se em temperatura ambiente. A leitura foi realizada no Qubit® 3.0 Fluorometer.

3.2.3. *Reação do GemCode*

A reação do GemCode foi realizada a partir do DNA a 1 ng/μL obtido por meio das diluições previamente descritas. Todas as pipetagens de homogeneização foram realizadas vagarosamente e com ponteiros de abertura larga para evitar fragmentação do DNA. Inicialmente, foi preparada uma solução "Master Mix" com 182 μL de água sem nuclease, 286 μL de reagente GemCode, 13 μL do "primer release agent", e 31 μL da polymerase GemCode. Homogeneizou-se em vórtex e centrifugou-se por 2 segundos. Separou-se 8 microtubos de 200 μL e adicionou-se 63,8 μL da solução "Master Mix" em cada um. Depois, transferiu-se 1,2 μL de amostra diluída a 1 ng/μL e homogeneizou-se por pipetagem.

Separadamente, uma "GemCode gel-bead strip" foi homogeneizada em vórtex por 25 segundos à velocidade máxima.

Foram transferidos 65 μL da amostra na fileira 1 do "GemCode Chip", 85 μL da suspensão "GemCode Gel-Bead" na fileira do do "GemCode Chip" e 150 μL do "Partitioning Oil" na fileira 3. Em seguida, o "GemCode Chip" foi processado no "GemCode Instrument". Após o procedimento, transferiu-se 115 μL do material processado pelo instrumento para uma placa de PCR de 96 poços, que

foi selada com o “Bio-Rad Plate sealer” a 185°C por 6 segundos. Posteriormente, foi realizada uma PCR no termociclador utilizando os seguintes parâmetros:

Estágio	Temperatura	Tempo	Rampagem
1	4°C		
2	98 °C	5:00	2,0 °C/Seg
3 Repetir 17x	4 °C	0:30	2,0 °C/Seg
	45 °C	0:01	0,1 °C/Seg
	70 °C	0:20	2,0 °C/Seg
	98 °C	0:30	2,0 °C/Seg
4	4°C		

Após a etapa de PCR, foi adicionado, em cada amostra, 1 µL do “Additive 1” e, posteriormente, 125 µL do “Recovery Agent”, homogeneizou-se e transferiu-se todo o conteúdo para tubos de 0,2 mL. As soluções foram então colocadas em vortex para homogeneização e centrifugadas, para subsequente retirada do “Recovery Agent”. A lavagem foi realizada com a adição de 200 µL do “DynaBead Clean-up Mix” para ligação das “beads” magnéticas, e então, na rack magnética, duas lavagens de etanol 80% (300 µL e 200 µL), eluição com 50,5 µL de “Elution Solution I” incubada a 65°C por 1 minuto, finalizando com o resgate de 50 µL do sobrenadante. Uma segunda lavagem foi realizada utilizando 30 µL do “SPRIselect Reagent” para ligação das “beads” magnéticas e duas lavagens de etanol 80% (125 µL), eluição com 52,5 µL de “Elution Solution II” a temperatura ambiente por 2 minutos, finalizando com o resgate de 52 µL do sobrenadante que foi transferido para um novo tubo de 0,6 mL.

3.2.4. Construção da Biblioteca

A construção da biblioteca de DNA foi realizada a partir de 51 µL do DNA obtido na etapa anterior foi fragmentado em 250 pb por meio do Covaris LE 220 de acordo com as especificações do fabricante. Subsequentemente, a reação de reparo da terminação e adição da cauda A foi realizada utilizando 50 µL de amostra e 10 µL do Mix para cada amostra e então incubados a 20°C e 65°C por 30 minutos cada e depois colocada a 4°C. Depois, a ligação do adaptador foi realizada adicionando, em cada amostra, 50 µL do Mix de ligação do adaptador (7,5 µL de água, 30 µL da solução tampão de ligação, 10 µL de DNA ligase e 2,5

µL do mix do adaptador) e homogeneizando por pipetagem e incubando por 15 minutos a 20°C.

A lavagem foi então realizada utilizando 88 µL do “SPRIselect Reagent” para ligação das “beads” magnéticas e duas lavagens de etanol 80% (200 µL), eluição com 40,5 µL de tampão de eluição a temperatura ambiente por 2 minutos, finalizando com o resgate de 40 µL do sobrenadante que foi transferido para um novo tubo de 0,6 mL.

A adição do index foi então realizada com a adição de 55 µL da “Sample Index PCR Master Mix” (50 µL de PCR Master Mix e 5 µL de P5 Primer) e 5 µL da GemCode Sample Index (10 µM) em cada amostra. A reação de PCR foi realizada com os seguintes parâmetros:

Estágio	Temperatura	Tempo	Rampagem
1	98°C	0:45	2,0 °C/Seg
2 Repetir 12x	98 °C	0:20	2,0 °C/Seg
	60 °C	0:30	2,0 °C/Seg
	72 °C	0:20	2,0 °C/Seg
3	72°C	1:00	2,0 °C/Seg
4	4 °C		2,0 °C/Seg

A lavagem foi realizada utilizando 100 µL do “SPRIselect Reagent” para ligação das “beads” magnéticas e duas lavagens de etanol 80% (200 µL), eluição com 30,5 µL de “Elution Buffer III” a temperatura ambiente por 2 minutos, finalizando com o resgate de 30 µL do sobrenadante que foi transferido para um novo tubo de 0,6 mL.

Por fim, foi realizada a quantificação por qPCR por meio do “KAPA Illumina Library Quantification Kit-Universal (PN KK4824) de acordo com as instruções do fabricante.

Exome Target Enrichment

A seleção de exons foi realizada por meio do “Agilent SureSelectXT Target Enrichment System” utilizando 5 µg da biblioteca de DNA preparada anteriormente, com algumas modificações de protocolo. Brevemente, foi adicionado 1 µL do “xGen Universal Blocking Oligo – TS HT-i5” e 1 µL do “xGen Universal Blocking Oligo – TS HT-i7” à biblioteca. O tubo foi então centrifugado

a 45°C por uma hora a -0,5 Barr e, depois, ressuspensionado em 3,4 µL de água. Em seguida as amostras foram transferidas para uma placa de PCR de 96 poços, e adicionou-se 5,6 µL do “blocking mix” à cada uma das amostras. Na mesma placa, adicionou-se 20 µL do tampão de hibridação em poços suprajacentes e 7 µL do “capture baits mix” nos poços subjacentes aos da amostra.

A placa foi posicionada no termociclador com as seguintes condições:

Estágio	Temperatura	Tempo
1	95 °C	5:00
2	70 °C	5:00
3	70 °C	2:00
4	70 °C	

Após a passagem do estágio 1, o programa foi pausado e 13 µL do tampão de hibridação de cada poço foi transferido para os poços com o “capture baits mix”. Após a passagem do estágio 2, o programa foi novamente pausado e todo o conteúdo dos poços com as amostras mais o “blocking mix” foram adicionados aos poços com o “capture baits mix” e o tampão de hibridação e homogeneizados. As amostras permaneceram incubadas por 16 horas a 70 °C.

Posteriormente, transferiu-se todo o conteúdo de cada um dos poços com amostras para tubos de 1,5 mL com 200 µL de “DynaBeads MyOne Streptavidin T1” e então homogeneizados e incubados por 30 minutos a 22°C e 1500 RPM. As amostras foram colocadas em rack magnético para retirada do sobrenadante, ressuspensionadas em 500 µL de solução “SureSelect Wash 1” e incubadas novamente por 15 minutos a 22°C e 1500 RPM. As *beads* foram ressuspensionadas em 500 µL do tampão “SureSelect Wash 2” pré-aquecida a 70°C e incubadas por 10 minutos (repetido 3 vezes). Depois, adicionou-se 20 µL de água para ressuspender as *beads*, que foram transferidas para novos tubos de 0,6 mL.

Para a realização da PCR pós captura, adicionou-se 30 µL do “post Capture PCR mix” em cada um dos tubos, homogeneizou-se e a reação se deu nas seguintes condições:

Estágio	Temperatura	Tempo	Rampagem
1	98 °C	0:45	
2 Repetir 9x	98 °C	0:15	2,0 °C/Seg
	60 °C	0:30	2,0 °C/Seg
	72 °C	0:30	2,0 °C/Seg
3	72°C	2:00	2,0 °C/Seg
4	4°C		

Então, o tubo foi transferido para rack magnética e o sobrenadante foi transferido para novos tubos de 0,6 mL. A lavagem foi então realizada utilizando 40 µL do “SPRIselect Reagent” para ligação das “beads” magnéticas e duas lavagens de etanol 80% (125 µL), eluição com 20,5 µL de tampão de eluição a temperatura ambiente por 2 minutos, finalizando com o resgate de 20 µL do sobrenadante que foi transferido para um novo tubo de 0,6 mL para subsequente sequenciamento.

3.2.5. Sequenciamento por HiSeq

O sequenciamento foi realizado por HiSeq 2500 (Illumina) de acordo com instruções do fabricante.

3.2.6. Processamento e análise dos dados

Os dados brutos foram processados por meio do software Long Ranger - 2.1.2 (10X genomics) para alinhamento (hg19), posicionamento dos *barcodes*, e identificação de indels, SNPs e variantes estruturais, gerando arquivos “.loupe”. A análise foi realizada por meio do software Loupe 2.1.1 (10X genomics) comparando com dados de SNP *arrays* e cariótipos. Nessa análise, gera-se um *heat map* pelo cruzamento de duas posições no genoma, no qual quanto mais próximas duas regiões se encontram, maior a sobreposição de *barcodes* mostrada.

O procedimento de *linked reads* foi realizado nas amostras G-6, G-68, G-101, G-149, G-195, G-239, G-415, G-670, G-928 e G-2070. Os critérios e razões de inclusão dessas amostras estão elucidados na seção de resultados.

3.3. Análise dos *arrays*

A análise inicial dos resultados de *array* da plataforma Illumina nos quais nos baseamos para determinar as regiões a serem sequenciadas foi realizada com o software KaryoStudio. Entretanto, devido ao surgimento de um novo software de análise de SNP *arrays* da Illumina, o BlueFuse Multi 4.3 (Illumina®), e da necessidade de precisão em relação aos pontos de quebra preditos pelo exame, reanalisamos todos os *arrays* desta plataforma no novo software e encontramos discrepâncias em relação a diversas estimativas.

A omissão de uma análise aprofundada nessas regiões poderia inviabilizar o sequenciamento dos pontos de quebra. Para solucionar esse novo problema, entramos em contato com o fabricante e expusemos as divergências entre as análises. Tais divergências não eram do conhecimento da empresa e, após diversas considerações, decidimos por basear a análise no software BlueFuse Multi em detrimento do KaryoStudio, visto que o fabricante o considera mais preciso e confiável. Para garantir que as regiões compreendessem os pontos de quebra, levamos em consideração a coordenada da última sonda sem alteração no número de cópias dada pelo software, analisamos manualmente a estrutura da gráfica da escala logarítmica do posicionamento da sonda e estimamos uma margem segura e economicamente viável para garantir o sequenciamento de cada ponto de quebra.

Como exemplo, na figura 3, na análise do *array* da deleção no braço longo do cromossomo 22 do paciente G-101, consideramos o posicionamento da sonda no eixo X e desconsideramos as sondas dadas como alteradas que não apresentavam variação significativa nesse eixo (nesse caso, três sondas). Dessa forma, decidimos como seguro determinar o sequenciamento dessa região entre chr22:18.640.300-18.877.787. Esse procedimento foi repetido e analisado para cada alteração encontrada.

Figura 3: Análise das sondas consideradas deletadas no braço longo do cromossomo 22 do paciente G-101 por meio do software BlueFuse Multi (Illumina®) e determinação das coordenadas para a confecção de sondas para o sequenciamento.

Como resultado, o sequenciamento compreendeu os pacientes e as regiões presentes na tabela I.

Tabela I: Regiões para sequenciamento determinadas após reanálise manual por meio do software BlueFuse Multi (regiões definidas em hg19).

REGISTRO	Chr	Start	End	Chr	Start	End
G-101	15	84041308	84047583	15	84106195	84112710
	22	18640300	18877787	22	21445064	21798907
G-191	22	41025934	41036329	22	41640297	41651034
G-239	9	-	-	9	13014232	13044407
	15	24346226	24352956	15	24472002	24521698
	18	70612590	70657389	18	-	-
G-242	7	38280817	38298285	7	38358412	38361999
	18	14090271	14090403	18	14283101	14341954
	22	25650648	25732697	22	25905922	25931372
	X	45532458	45541623	X	45605273	45610875
G-291	5	59174103	59178177	5	59522613	59544653
G-390	14	20217188	20227986	14	20427142	20436017
	15	20408183	20481802	15	22509154	22558856
	17	116443	148192	17	2310471	2322773
	17	74289668	74307123	17	80943089	80969524
G-396	X	18151389	18179714	X	19719264	19729105
G-425	1	23438088	23441018	1	23714665	23724183
	2	113342566	113344997	2	113429752	113431092
	2	139848110	139866554	2	139988961	139994754
	3	122437834	122445389	3	122658417	122664852
	3	182084954	182096311	3	182178216	182194305
	3	194924130	194935314	3	195050247	195057001
	4	112927422	112942442	4	113244313	113251247
	4	118756418	118777687	4	-	-
	7	14423100	14436385	7	14838576	14840654
	7	103058720	103071193	7	103183170	103190358
	7	158493873	158504967	7	-	-
	8	71100634	71106705	8	71224998	71229377
	9	90073584	90095024	9	90202007	90206006
	12	130054905	130069008	12	130138215	130144736
	X	31387752	31394057	X	31534414	31535731
	X	101132504	101132883	X	101248915	101253185
G-524	9	13466310	13468616	9	14566406	14589694
G-602	2	143335573	143387612	2	144967929	145079065
	3	102201538	102215225	3	102607793	102609928
	5	36875575	36900843	5	37561120	37567267
	6	24342371	24342612	6	50197539	50203633
	11	31150302	31159553	11	31385819	31389267
	12	59023563	59055325	12	59295453	59316881
	14	64131064	64157117	14	64306714	64313112
	17	27973332	27978455	17	28335949	28348047
	X	91063807	91158935	X	91536873	91577490
G-928	16	-	-	16	828466	831496
	16	85763447	85817324	16	-	-
	22	22296119	22302762	22	22548831	22561973
G-1517	2	-	-	2	14800749	14805647
	2	63482704	63496998	2	63678069	63686428
	4	186457759	186463750	4	-	-
	X	42341441	42360230	X	42560611	42567603

As regiões foram então definitivamente determinadas e o desenho das sondas foi realizado pelo software design studio (Illumina) para confecção do kit de captura utilizando a metodologia NEXTERA rapid capture.

3.4. Sequenciamento dos pontos de quebra

Após a definição das regiões para sequenciamento serem definidas (hg19), elas foram submetidas para a confecção do kit de captura no software Illumina Design Studio utilizando a metodologia NEXTERA Rapid Capture Enrichment (Illumina Inc., San Diego, CA).

O ensaio foi realizado de acordo com instruções do fabricante. Brevemente, o DNA foi quantificado por Qubit 2.0 e normalizado a 10 ng/uL. Posteriormente, foi requantificado pelo mesmo método e normalizado a 5 ng/uL. O DNA foi fragmentado e etiquetado com adaptadores, utilizando 10 uL de cada amostra do DNA normalizado foi disposta cada uma em um poço de uma placa de PCR, adicionou-se 25 uL da solução TD e 15 uL da solução TDE1, e então a placa foi aquecida a 58°C por 10 minutos. Depois, adicionou-se 15 uL de ST em cada poço e incubou-se a temperatura ambiente por 4 minutos.

As amostras foram então lavadas utilizando as *beads* magnéticas SPB inicialmente, seguindo a duas lavagens de etanol 80%, secadas a temperatura ambiente e ressuspensas em água e resgatadas para uma nova placa. Posteriormente, às amostras foram adicionados os index i7 e i5 e elas foram amplificadas por 10 ciclos de PCR. Em seguida, as amostras foram lavadas da mesma maneira que anteriormente e misturadas num único tubo. Foram então hibridadas com os oligos de captura que foram confeccionados exclusivamente para este experimento. O tubo foi então aquecido no termociclador a 95°C com a temperatura regredindo 1°C por minuto até atingir 58°C. Iniciou-se o processo de captura com utilização das *beads* de streptavidina, posterior lavagem com o tampão EWS, seguida de eluição com o *premix* contendo EE1 e HP3, sendo resgatada para um novo tubo. O mesmo processo foi repetido novamente para a segunda hibridação, captura e eluição.

Em seguida, a biblioteca foi finalmente lavada utilizando as *beads* magnéticas e etanol 80%, eluída em água e resgatada. Posteriormente a biblioteca foi amplificada por 12 ciclos em uma reação de PCR e finalmente lavada novamente.

Foi feita uma análise final do tamanho do fragmento da biblioteca por TapeStation (Agilent), conforme instruções do fabricante, e a quantificação por Qubit.

A biblioteca foi sequenciada no MiSeq (Illumina), kit v2 de 300 ciclos, conforme instruções do fabricante.

3.4.1. Análise dos dados brutos do sequenciamento Illumina

Os arquivos fastq gerados pelo sequenciamento foram alinhados por meio do software SureCall (Agilent Technologies). A análise das *reads* químéricas foi realizada por meio do software IGV 2.4.14 (Broad Institute and the Regents of the University of California).

Resultados e Discussão

4. Resultados e Discussão

4.1. Avaliação por 10X GemCode

4.1.1. Avaliação de integridade das amostras para 10X GemCode

A avaliação da qualidade das 15 amostras previamente analisadas por *array*, revelou 7 amostras consideradas elegíveis para o procedimento de marcação por 10X GemCode (fragmentos de tamanho superior a 40 kb) (Figura 4, Tabela II).

Figura 4. Análise da integridade das amostras previamente analisadas por *array* por Tape Station. MPM: Marcador de peso molecular.

Tabela II: Concentração e integridade das amostras de DNA previamente analisadas por *array* e elegibilidade para prosseguimento.

Paciente	Concentração [ng/μL]	Integridade (pb)	Elegível?
G-6	387	>60.000	SIM
G-68	60,6	49.821	SIM
G-101	93,3	54.528	SIM
G-191	68,9	50.353	SIM
G-239	339	>60.000	SIM
G-242	70,2	26.401	NÃO
G-291	44,9	16.200	NÃO
G-317	54	35.173	NÃO
G-390	9,55	23.842	NÃO
G-396	218	25.413	NÃO
G-425	54,6	36.184	NÃO
G-524	358	>60.000	SIM
G-602	32,8	23.097	NÃO
G-928	223	58.672	SIM
G-1517	127	21.210	NÃO

A avaliação das 21 amostras previamente cariotipadas revelou 11 amostras consideradas elegíveis para o procedimento de marcação por 10X GemCode (fragmentos de tamanho superior a 40 kb) (Figura 5, Tabela III).

Figura 5. Análise da integridade das amostras previamente cariotipadas por Tape Station. MPM: Marcador de peso molecular.

Tabela III: Concentração e integridade das amostras de DNA previamente cariotipadas e elegibilidade para prosseguimento.

Paciente	Concentração [ng/μL]	Integridade (pb)	Elegível?
G-10	34	>60.000	SIM
G-46	111	>60.000	SIM
G-149	48,6	49.069	SIM
G-195	374	>60.000	SIM
G-308	339	22.434	NÃO
G-312	39,2	52.136	SIM
G-415	104	>60.000	SIM
G-427	57,9	24.001	NÃO
G-512	75,0	30.403	NÃO
G-539	351	>60.000	SIM
G-589	7.82	19.711	NÃO
G-629	196	32.774	NÃO
G-670	2030	>60.000	SIM
G-1103	6,31	16.619	NÃO
G-1141	125	55.761	SIM
G-1217	22,2	35.589	NÃO
G-1373	29,1	9.154	NÃO
G-1533	31,1	57.096	SIM
G-1996	75,4	18.684	NÃO
G-1997	206	19.592	NÃO
G-2070	47,9	>60.000	SIM

É essencial que as amostras estejam com uma integridade superior a 40 kb para que se garanta que um grande fragmento de DNA obtenha a mesma marcação, possibilitando a identificação das alterações estruturais.

Dentre as amostras elegíveis, selecionamos 10 para seguir o protocolo 10X GemCode: G-6, G-68, G-101, G-149, G-195, G-239, G-415, G-670, G-928 e G-2070. O critério de seleção de cada amostra encontra-se na seção destinada aos resultados encontrados em cada uma delas.

Essas 10 amostras foram etiquetadas e, posteriormente, fragmentadas por sonicação pelo equipamento LE220 (Covaris, Inc.) para captura dos éxons de acordo com o protocolo SureSelect V5 (Agilent Technologies) e sequenciadas por HiSeq (Illumina). Por fim, elas foram analisadas pelo software Loupe (10X Genomics) para avaliação do panorama estrutural e dos mecanismos de formação.

4.1.2. Análise por Loupe

Dentre essas amostras selecionadas e analisadas, a G-670 foi selecionada por apresentar um cariótipo 46,X,t(X;16)(p11.1;q11.2), rearranjo único, e a G-928 por apresentar 7 CNVs envolvendo diversos cromossomos, sendo 3 CNVs no cromossomo 16. Entretanto não obtivemos um bom resultado durante o procedimento de marcação por GemCode, resultando em apenas 45,4% (G-670) e 41,9% (G-928) dos genes *phased*. Isso impossibilita a análise, uma vez que mais da metade dos genes podem não estar marcados, ocasionando uma perda significativa de material para análise. As amostras dos pacientes G-101 e G-2070 não apresentaram alterações que ratificassem as alterações encontradas por *array*.

4.1.2.1. Paciente G-6

O sequenciamento e posterior análise por Loupe resultou em 92,5% dos genes *phased*. A amostra deste paciente foi selecionada para a análise por *linked reads* por apresentar duplicações e deleção por *array* em regiões muito próximas no mesmo cromossomo, a saber: duas duplicações em Chr9:41.692.304-44.244.868 e Chr9:44.864.687-45.723.022 e uma deleção em Chr9:44.727.846-44.824.251. Uma vez que o *array* identifica apenas desequilíbrios, nosso intuito foi averiguar se as duplicações encontravam-se no mesmo cromossomo ou teriam sido adicionadas em outros cromossomos ou outra região do mesmo cromossomo. Entretanto, a análise por *linked reads* revelou apenas uma proximidade da região chr9:41.962.006-66.517.203, indicando uma possível inversão pericêntrica do cromossomo 9, inversão não ratificada pelo cariótipo prévio do paciente.

Não foi possível identificar as alterações encontradas por *array* neste paciente, esse fato se deve, provavelmente, ao fato de utilizarmos o sequenciamento do exoma para alinhamento das *linked reads*, o qual apresenta *gaps*, que dificultaram a identificação da deleção e duplicações desta região. Contudo, a ausência de proximidade dessa região com outras regiões do genoma na análise por Loupe, pode indicar que a duplicação não se encontra

em outro cromossomo, e esteja próxima ou subsequente à própria região de origem, embora não possamos afirmar essa possibilidade indubitavelmente.

4.1.2.2. Paciente G-68

O sequenciamento e posterior análise por Loupe resultou em 87,1% dos genes *phased*. A amostra deste paciente foi selecionada para a análise por *linked reads* por apresentar deleções e duplicações em cromossomos diferentes, a saber: uma deleção em Chr2:239.550.182-243.029.573 e duplicações em Chr5:172.246.068-180.705.539 e Chr18:27.778.530-28.050.968. Devido à limitação do *array* de não identificar rearranjos, havia uma possibilidade de haver um rearranjo envolvendo essas regiões (37, 38). Por loupe foi possível identificar a duplicação no cromossomo 18, definida entre Chr18:27.809.927-28.051.890 (Figura 6).

Figura 6. Heatmap da mostrando a duplicação no cromossomo 18.

Esse gráfico mostra que, pelo menos a duplicação do cromossomo 18 ocorreu imediatamente junto com a sequência original, descartando a

possibilidade de rearranjo envolvendo essa região. Isso prova que esta duplicação encontrada ocorreu de maneira independente, sem relação com as outras alterações.

Apesar de não identificarmos as outras duas alterações, como no caso G-6, também não há indicação de proximidade entre as regiões envolvidas em perda e ganho dos cromossomos 2 e 5, contudo, isso não significa de que não haja um rearranjo envolvendo essas regiões, apenas uma suspeita.

Não obstante, o fato de a duplicação do cromossomo 18 ser independente, afasta a possibilidade de um evento único envolvendo os desequilíbrios genômicos encontrados neste paciente, afastando, assim, mecanismos de catástrofes cromossômicas. Esta avaliação também permitiu um melhor delineamento do ponto de quebra dessa região.

A análise das regiões envolvidas no ponto de quebra, revelou a presença de elementos repetitivos na região (como LINEs, SINEs e LTRs), o que, unido ao fato de ser uma duplicação e ocorrer de maneira isolada, sugere que o mecanismo causador desta alteração tenha sido um NAHR.

4.1.2.3. Paciente G-149

O sequenciamento e posterior análise por Loupe resultou em 95,6% dos genes *phased*. A amostra deste paciente foi selecionada para a análise por *linked reads* por possuir uma translocação identificada por citogenética clássica, com cariótipo prévio 46,XY,t(7;14)(q36;q22), e por não possuir *array*. Com isso, poderíamos delimitar de maneira muito mais precisa o ponto de quebra e compreender se a translocação havia sido única.

A translocação foi confirmada, mostrando uma ligação da região chr7:138.399.664 (em 7q34) com a chr14:46.624.191 (em 14q21.2) (Figura 7). Uma vez que as regiões cruzadas não se encontram no mesmo cromossomo, a presença de marcação entre elas significa que estão muito próximas na mesma molécula, evidenciando, assim, a translocação (Figura 8). O fato da localização não se encontrar exatamente onde descrito no cariótipo justifica-se pela capacidade de resolução limitada a ao menos 5 Mb e, portanto, seu resultado final ser aproximado. Com isso, detectamos o local exato da translocação e definimos o cariótipo como 46,XY,t(7;14)(7pter->q34::14q21.2->qter).

Figura 7. *Heatmap* da interseção entre os cromossomos 7 e 14 mostrando a ligação nas regiões indicadas.

Figura 8. A. Visão linear mostrando o ponto de quebra onde ocorreu a translocação entre o cromossomo 7 e o cromossomo 14 (barra verde). Em azul, posição e os nomes dos genes nas regiões envolvidas. B. Ideograma da estrutura cromossômica resultante da translocação.

Ademais, foi possível encontrar outras alterações não identificadas por cariótipo, que podem ser importantes para a elucidação dos mecanismos de formação envolvidos nas alterações encontradas no paciente: uma deleção homozigota em 13q e uma 2q, uma inversão em 3q e uma duplicação em 9q.

Entretanto, a presença de alterações em heterozigose como inversão de 3q e a duplicação de 9q, poderiam significar que as alterações encontradas não ocorreram de maneira isolada, mas sim são consequências de um único evento que envolveu todo o genoma.

Entretanto, como não foi detectada nenhuma outra translocação envolvendo regiões próximas das translocadas nos cromossomos envolvidos, exclui-se a probabilidade de um mecanismo replicativo, mas sim um não-replicativo. Uma vez que as regiões não estão próximas a regiões repetitivas (como LCRs), podemos inferir que o provável mecanismo causador da translocação foi NHEJ.

4.1.2.4. Paciente G-195

O sequenciamento e posterior análise por Loupe resultou em 95,1% dos genes *phased*. A amostra deste paciente foi selecionada para a análise por *linked reads* por, assim como no caso G-149, possuir uma translocação identificada por citogenética clássica, com cariótipo prévio 46,XY,der(9)t(9;18)(p22;q12.2), e por não possuir *array*. Com isso, poderíamos delimitar de maneira muito mais precisa o ponto de quebra e compreender se a translocação havia sido única, inferindo o mecanismo do ponto de quebra. Também, por se tratar de um cariótipo, seria inviável desenhar sondas que compreendessem toda a região para subsequente sequenciamento, sendo esta a única possibilidade de obtermos uma análise mais refinada.

Foi possível confirmar a translocação, mostrando uma ligação da região chr9:4.368.704 (em9p24.1) com a chr18:39.163.156 (em 18q12.3) (Figura 9). Como no caso do paciente G-149, as regiões identificadas diferem ligeiramente das apresentadas no cariótipo, devido sua baixa resolução. Com isso, conseguimos identificar com grande precisão o local da translocação e o cariótipo como 46,XY,der(9)t(9;18)(9qter->p24.1::18q12.3->qter) (Figura 10).

Figura 9. Heatmap da interseção entre os cromossomos 9 e 18 mostrando a ligação nas regiões indicadas.

Figura 10. A. Visão linear mostrando o ponto de quebra onde ocorreu a translocação entre o cromossomo 9 e o cromossomo 18 (barra verde). Em azul, posição e os nomes dos genes nas regiões envolvidas. B. Ideograma da estrutura cromossômica resultante da translocação.

Dada a ausência de alterações envolvendo regiões próximas aos cromossomos envolvidos na translocação, e de ela ser uma translocação única,

sem elementos repetitivos nas regiões do ponto de quebra, descartamos a possibilidade de um mecanismo recorrente, como NAHR e os mecanismos replicativos. Assim, inferimos que o rearranjo observado, assim como no caso G-149, é advindo de NHEJ.

4.1.2.5. Paciente G-239

A amostra deste paciente foi selecionada para a análise por *linked reads* por apresentar perdas e ganhos de regiões de diferentes cromossomos, visando identificar se as duplicações apresentadas encontravam-se em outras regiões e se haveria um rearranjo envolvido. O sequenciamento e posterior análise por Loupe resultou em 92,7% dos genes *phased*, entretanto o tamanho médio da molécula de DNA “etiquetado” foi de 28,8kb, inferior ao tamanho ideal recomendado (acima de 40kb). Esse fato prejudica a análise, pois não consegue garantir um alinhamento perfeito de dezenas de megabases. O paciente G-239 possuía diagnóstico prévio por *array* (CytoSNP-12 - Illumina), que evidenciava uma duplicação no cromossomo 9 e uma no X, além de uma deleção no cromossomo 18, conforme mostra a tabela IV.

Tabela III. Alterações previamente identificadas por *array* no paciente G-239

Paciente	PLATAFORMA	Alter.	Chr	START	END
G-239	Illumina CytoSNP-12	dup	9	46587	13014232
		del	18	24352956	24480615
		dup	X	70642373	78014582

A análise revelou uma deleção em homozigose em chr4:69.262.505-69.594.276, de aproximadamente 331kb, não detectada por *array*, devido à ausência de sondas para a região (Figura 9). Não foi possível a identificação de todas as alterações encontradas por *array*, porém foi identificada uma alteração na região da duplicação de X por *array*. A qualidade desses dados ainda não garante que é uma duplicação. Isso se deve provavelmente ao diminuto tamanho das regiões etiquetadas, principalmente nas alterações dos cromossomos 9 e 18, com mais de 8 megabases entre elas. No entanto esse resultado nos permite inferir que quanto maior a alteração, mais importante é um tamanho adequado das regiões marcadas.

Figura 11. Acima: Heatmap mostrando a deleção em homozigose do cromossomo 4. Abaixo: Visão linear mostrando o ponto de quebra e a região deletada (arco verde). Em azul, posição e os nomes dos genes nas regiões envolvidas.

4.1.2.6. Paciente G-415

O sequenciamento e posterior análise por Loupe resultou em 81,6% dos genes *phased*. A amostra deste paciente foi selecionada para a análise por *linked reads* por apresentar um cariótipo prévio 46,XY,der15,t(14;15)(q11.3;q11). Assim, seria possível analisar a que ponto próximo ao centrômero foi ocorreu o a quebra, sendo este o único caso com regiões centroméricas envolvidas. Entretanto, não foi possível detectar essa translocação, pois ela envolve os centrômeros, região repetitiva onde não se é possível sequenciar utilizando a captura do exoma. Ainda assim, conseguimos detectar uma pequena deleção heterozigota no braço curto do cromossomo 1, de 113 kb, assim como nos pacientes G-149 e G-195, provavelmente mediadas por NHEJ.

4.1.3. Considerações sobre o sequenciamento utilizando Linked Reads

O sequenciamento por *linked reads* permitiu investigarmos com sucesso diversas regiões do genoma, identificando os rearranjos e conferindo uma visualização real da configuração da molécula. O fato da não identificação de

algumas regiões pode ser explicado por termos utilizado o sequenciamento do exoma para identificar a estrutura. Com isso, regiões de “desertos gênicos” não foram identificadas, o que dificulta a montagem e posterior alinhamento para definição de algumas partes da estrutura. Uma forma de sanar esse problema seria utilizar o sequenciamento do genoma completo com as *linked reads*, mas isso elevaria enormemente o custo dos experimentos, o que inviabilizaria o estudo completamente.

Ressaltamos que a abordagem aqui utilizada é uma novidade e demonstrou um resultado eficiente, possibilitando definir com boa precisão os rearranjos detectados de uma maneira totalmente molecular. Estudos utilizando a técnica vêm sendo publicados recentemente com sucesso em humanos e diversas espécies (39-42). Ainda, seria impossível identificar os rearranjos estruturais por meio unicamente do sequenciamento do exoma, sem a utilização das *linked reads*. O próprio sequenciamento do genoma completo com *short reads* tornaria muito mais dificultosa a tarefa da identificação dos rearranjos, além de mais custosa, sendo virtualmente inviável (43). Também foi possível propor os mecanismos de geração de pontos de quebra das amostras analisadas pela análise do o panorama estrutural mesmo que não tenhamos sequenciado o ponto de quebra em si.

4.2. Análise dos *arrays* para sequenciamento

A análise realizada pelo software BlueFuse resultou nos seguintes desequilíbrios genômicos mostrados na tabela IV.

Tabela IV. Coordenadas genômicas dos desequilíbrios encontrados nas amostras dos pacientes após avaliação por BlueFuse (hg19).

Paciente	Alteração	Paciente	Alteração	Paciente	Alteração
G-101	dup Chr15:84047483-84106195 del Chr22:18646514-21445064		del Chr1:23441018-23714665 del Chr2:113344997-113429752	G-524	del Chr9:13466329-14566406 dup Chr2:143387612-144967929
G-191	del Chr22:41032863-41640297		del Chr2:139866554-139988961 del Chr3:122445389-122658417		del Chr3:102215225-102607793 del Chr5:36895270-37561120
G-239	dup Chr9:46587-13014232 del Chr15:24352956-24480615 del Chr18:70642373-78014582		del Chr3:182096311-182183458 del Chr3:194935314-195050247	G-602	dup Chr6:24342612-50197539 del Chr11:31159553-31385819
	del Chr7:38292663-38358956 del Chr18:14090403-14308620	G-425	del Chr4:112929836-113244313 dup Chr4:118777687-190880409		del Chr12:59055325-59295453 del Chr14:64149889-64306714
G-242	del Chr22:25711667-25910879 del ChrX:45541623-45605273		del Chr7:14436385-14838576 del Chr7:103071193-103183170		del Chr17:27978455-28335949 del ChrX:91158935-91536873
G-291	del Chr5:59178177-59522613 del Chr14:20227886-20427242		del Chr7:158504967-159119486 del Chr8:71106705-71224998	G-928	del Chr16:105320-828466 dup Chr16:85768265-90148796
G-390	dup Chr15:20481702-22509254 del Chr17:148092-2310571 dup Chr17:74307023-80943189		del Chr9:90095024-90202007 del Chr12:130069008-130138215 del ChrX:31394057-31534414		dup Chr22:22302762-22550078 dup Chr2:72184-14800749
G-396	del ChrX:18179714-19719264		del ChrX:31394057-31534414 del ChrX:101132883-101248915	G-1517	del Chr2:63496998-63678069 del Chr4:186463750-190880409 del ChrX:42347568-42560611

4.3. Análise do sequenciamento das regiões flanqueadas

Os arquivos fastq gerados pelo sequenciamento foram alinhados de modo mais permissivo, uma vez que a finalidade foi encontrar *reads* quiméricas que revelassem o ponto de quebra.

4.3.1. Paciente G-191

O paciente G-191 apresentava, por *array* uma única deleção no cromossomo 22. A análise por IGV revelou uma *read* quimérica, com o início alinhando por volta da região Chr22:41.026.100 e o final alinhando por volta da região chr22:41.650.750. Com essa análise foi possível estabelecer a região exata desta deleção: Chr22:41.026.112-41.650.780 (hg19), o que configura uma deleção de 624.668 pb (Figura 12).

Referência Chr22:41.026.090	...GCGGATCACCTGAGGTCAGGAG	TTT GAGACCAG	CCTGGCCAACA...
Read Chr22:41.026.090	...GCGGATCACCTGAGGTCAGGAG	ATC GAGACCAG	CCTGGCTAACG...
Referência Chr22:41.650.751	...TGGGTGGATCATGAGGTCAGGA	ATC GAGACCAT	CCTGGCTAACG...

Figura 12. Sequência do ponto de quebra: na primeira e na última linha estão as sequências de referência (hg19) das regiões indicadas, na segunda linha a sequência do paciente obtida pelo sequenciamento. Destacado em amarelo, o ponto de quebra.

A deleção ocorreu em íntrons de dois genes distintos, unindo o gene *MKL1* em seu início e o gene *RANGAP1* em sua porção intermediária. A investigação da região utilizando o *genome browser* revelou que ambas se encontram dentro de sequências Alu, sendo a região Chr22:41.026.112 em uma sequência AluSx e a região Chr22:41.650.780 em uma sequência AluY.

Sequências Alu são SINEs (do inglês *Short Interspersed Elements*), retrotransposons originados possivelmente de um DNA parasita. Os elementos Alu são as SINEs mais frequentes, compreendendo mais de 10% de todo o genoma humano, e existem apenas em primatas. Esses elementos têm por volta de 300 pb e são classificados em diversas subfamílias. AluSx é uma subfamília (AluS) de origem relativamente intermediária na escala evolutiva, e não é mais ativa, ou seja, não possui mais a capacidade de se replicar e reposicionar no genoma. AluY corresponde a uma subfamília que era classificada como AluS, mas foi recentemente considerada uma subfamília separada, devido sua origem mais recente e o fato de ainda serem ativas (44-53).

A análise da sequência de ambas SINEs, obtidas no Dfam (54), revelou que elas possuem alto grau de similaridade. No local exato onde ocorreu a quebra, observamos uma região de homologia total e, subsequentemente, a sequência do segmento à frente. Todas essas características (homologia junto ao ponto de quebra, presença de sequências repetitivas e o fato de ocorrerem em regiões de alta similaridade) são típicas de um mecanismo de NAHR (55, 56).

É a primeira vez que um rearranjo por NAHR nessa região é sequenciado e também há poucos relatos de sequenciamento de pontos de quebra envolvendo sequências Alu. Apesar da existência de relatos de rearranjos por NAHR envolvendo essas sequências, poucos observaram na resolução de par de base.

4.3.2. Paciente G-396

Com a análise do sequenciamento do paciente G-396 no IGV, levando em consideração o direcionamento das *reads* e a posição, ratificamos a deleção encontrada por *array* no cromossomo X. Esta era a única região com desequilíbrio encontrada por *array* e foi possível não apenas identificá-la, como delimitá-la. Após o alinhamento, não foi possível identificar *reads* quiméricas, o que impossibilitou o sequenciamento exato do ponto de quebra, mas foi possível identificar duas *reads* cujo par se encontrava na outra região sequenciada, o que permitiu a delimitação, definindo o tamanho da deleção em chrX:18.173.514-19.725.291. Foi realizada uma busca pelas *reads* que não haviam sido alinhadas, separando-as em um arquivo bam à parte e convertendo o arquivo em fastq novamente, mas não foi possível identificar *reads* quiméricas. O início da deleção se encontra em região intergênica e o final no íntron do gene SH3KBP1, em sua região intermediária.

Não há, nas regiões próximas, elementos repetitivos. Como essa é a única alteração encontrada, podemos inferir que o rearranjo foi causado por NHEJ.

4.3.3. Paciente G-524

A análise do sequenciamento no IGV revelou diversas *reads* quiméricas nas regiões sequenciadas do cromossomo 9 (Figuras 13 e 14).

Figura 13. Imagem do IGV mostrando a região envolvida na quebra, no cromossomo 9. Cada barra representa uma *read*. As *reads* em vermelho são as que seu par se encontram em outro lugar do genoma. As bases destacadas são as que diferem do genoma de referência, indicando que elas não estão alinhadas com a região posterior a elas.

Referência Chr9:13.468.318	...AAATATTTTAATTCTTAATAAAATATTA	GC	ACTTAACCTTTGCTTTAG...
Read chr9:13.468.318	...AAATATTTTAATTCTTAATAAAATATTA	GC	TCCATAAAGG ← início
Read chr9:14.584.890	← início	TAATAAAATATTA	GC TCCATAAAGGTTAGATTTA...
Referência chr9:14.584.890	...TAAAGCAGCT	GC	TCCATAAAGGTTAGATTTA...

Figura 14. Sequência do ponto de quebra: na primeira e na última linha estão as sequências de referência (hg19) das regiões indicadas. Na segunda e na terceira linha, a sequência do paciente obtida pelo sequenciamento alinhadas em diferentes partes do genoma. Destacado em amarelo, a micro-homologia e o ponto de quebra.

Essas *reads* foram alinhadas imediatamente no início ou no final da deleção. Por esse motivo, foi possível avaliá-la no IGV. Com o dado preciso do local da deleção, localizamos diretamente no arquivo fastq uma *read* quimérica que não havia sido alinhada, que também evidencia a deleção e o ponto de quebra (Figura 15).

Referência Chr9:13.468.318...TCCAAATATTTTAATTCTTAATAAAATATTA GC ACTTAACCTTTGCTTAG...
 Read ...TCACCTTGGCTGCCCTGATCCAAATATTTTAATTCTTAATAAAATATTA GC TCATAAAGGTTAGATTTATTTCCCTTATTTCTCCTTTTACCAAAAGTGCCTCAGTCTAGTAC
 Referência chr9:14.584.890 ...TAAAGCAGCT GC TCATAAAGGTTAGATTTATTTCCCTTATTTCTCCTTTTACCAAAAGTGCCTCAGTCTAGTAC

Figura 15. Sequência do ponto de quebra: na primeira e na última linha estão as sequências de referência (hg19) das regiões indicadas. Na segunda linha, a sequência do paciente obtida pelo sequenciamento que não foi alinhada ao genoma. Destacado em amarelo, a micro-homologia e o ponto de quebra.

Com essa análise foi possível detectar a região exata da deleção: chr9:13.468.456-14.584.904 (hg19), o que configura uma deleção de 1.116.448 pb. Isso mostra que a imprecisão do *array* em definir o ponto de quebra nesse caso específico foi de, aproximadamente, 2 kb a mais no início da deleção e 18 kb a menos no final da deleção, sendo que a deleção real teve uma discrepância de 16 kb.

As duas regiões envolvidas com o ponto de quebra estão localizadas em regiões intergênicas e foi possível identificar uma região de micro-homologia de dois pares de base (GC) exatamente no ponto de quebra.

Inferimos que essa micro-homologia foi provavelmente a responsável pelo rearranjo observado. Por se tratar de um rearranjo único no paciente, a suspeita de mecanismo causador dessa alteração seria o NHEJ. Entretanto, esse mecanismo tem como característica a adição e a remoção de alguns pares de base para a cicatrização, elementos que não se encontram presentes no *locus* observado.

Outrossim, analisamos a região deletada no *genome browser* e identificamos que ambas regiões encontram-se em elementos repetitivos do genoma. O *locus* chr9:14,584,903 (hg19) está em uma SINE (do inglês *short interspersed element*) MIRb, que pode desempenhar papel de *enhancer* (53). O *locus* chr9:13,468,456 (hg19) está em uma LTR (do inglês *long terminal repeat*), a LTR16D, um retrotransposon da superfamília ERVL (do inglês *endogenous retroviruses*), derivado de retrovírus (54).

A presença de regiões repetitivas próximas ao ponto de quebra pode sugerir ação do mecanismo NAHR, entretanto, os dois elementos repetitivos encontrados não possuem similaridade entre si e a ausência de LCRs próximas à deleção descarta essa possibilidade. O ponto de quebra encontrado alinha-se então com a característica de MMBIR ou FoSTeS que, apesar de ter gerado uma

única deleção, apresenta a região de micro-homologia peculiar destes mecanismos.

4.3.4. Paciente G-602

O paciente G-602 apresentava 9 regiões de perdas ou ganhos por *array*, esse resultado sugere a possibilidade de que um evento único tenha causado as diversas alterações encontradas. Com o sequenciamento dessas regiões e subsequente análise no IGV, levando em consideração o direcionamento das *reads* e a posição, ratificamos a duplicação encontrada por *array* no cromossomo 2 e delimitamos o tamanho do desequilíbrio em chr2:143.378.126-144.973.394. A figura 16 revela as *reads* invertidas, indicando que a região entre as duas *reads* encontram-se duplicadas.

Figura 16. Imagem do IGV mostrando as regiões envolvidas na quebra, no cromossomo 12. Cada barra representa uma *read*. As *reads* em verde são as que seu par se encontram em outro lugar do genoma, indicando também que elas estão em posição invertida, denotando uma duplicação.

Entretanto, não foi possível identificar *reads* quiméricas pelo IGV. Cientes da possibilidade de as *reads* quiméricas não terem sido alinhadas, construímos

um arquivo BAM apenas com as *reads* não alinhadas e convertemos em fastq para realizarmos a busca por sequências próximas ao ponto de quebra em busca de uma *read* quimérica. Contudo, essa busca não revelou resultados, o que indica que o ponto de quebra, em si, não foi sequenciado.

No cromossomo 11 não foi possível identificar a deleção observada por SNP *array*, entretanto, descobrimos uma pequena deleção em homozigose entre Chr11:31.152.447-31.152.881 (hg19) de 434 pb. Junto ao ponto de quebra encontramos a inserção de alguns pares de bases e a substituição de outros (Figura 17). Ressalta-se que a região onde está presente a deleção é muito repetitiva, sendo que parte dela integra uma SINE AluYb8, o que a torna mais propensa a quebras. Esta avaliação é corroborada pela consulta ao DGV, que revela que são comuns deleções envolvendo essas regiões. Por ser uma região de SINE, com intensa repetição de nucleotídeos, e se tratar de uma deleção recorrente, o mecanismo responsável foi, provavelmente o NAHR. A busca pelas deleções e duplicações das outras regiões encontradas por *array* não surtiu resultados.

Figura 17. Imagem do IGV mostrando a região envolvida na quebra, no cromossomo 11. Acima, a imagem mais ampla mostrando toda a região deletada. Cada barra representa uma *read*. As *reads* em vermelho são as que seu par se encontram em outro lugar do genoma. As linhas verticais cortando as barras horizontais (*reads*) de diversas cores representam as bases que diferem do genoma de referência. Abaixo, a imagem mais aproximada, mostrando as bases que diferem do genoma de referência

4.3.5. Demais amostras e considerações

Os pacientes G-101, G-239, G-242, G-291, G-390, G-425, G-928 e G-1517 não apresentaram *reads* quiméricas ou com seu par em outra região que

fosse confiável definirmos onde o ponto de quebra estava. Apesar de haver diversas regiões sequenciadas, foi possível sequenciar apenas alguns pontos.

A estratégia de flanquear os pontos de quebra previamente identificados por *array* surgiu como forma de realizarmos o estudo de forma eficiente e menos custosa. Ao mesmo tempo em que fomos criteriosos ao definir o ponto de quebra, e termos dado uma margem segura para garantir o sequenciamento do ponto de quebra com exatidão, nós intencionalmente fomos mais permissivos em relação à quantidade de alterações observadas em um único paciente. Nós excluimos as regiões mais propensas a serem identificadas como rearranjos inexistentes e também as regiões em que a chamada da alteração era ruído da técnica. Entretanto, consideramos as regiões em que não estavam totalmente esclarecidas quanto se a alteração era real ou não. Fizemos isso com o intuito de avaliar cada região, sem correr o risco de perder uma alteração que tenha sido subestimada.

Ademais, os pontos de quebra haviam sido previamente identificados por plataformas de *array* que não possuíam uma grande resolução. Com isso, a estimativa do ponto de quebra pode se tornar mais imprecisa, ou é necessário definir uma região maior para sequenciar. *Arrays* com maior capacidade de resolução, com maior número de sondas, poderiam facilitar e aumentar a taxa de sucesso para essa detecção. A alternativa de customizar um segundo *array* de alta densidade nas regiões definidas pelo *array* que já possuíamos foi aventada, mas após diversas considerações foi descartada, pois elevaria o custo total do experimento de maneira que o tornaria inviável.

O sequenciamento dos pontos de quebra por NGS utilizando short *reads* permitiu-nos sequenciar o ponto de quebra exato em 3 pacientes e uma grande aproximação em um paciente. Como isso, foi possível determinar os mecanismos produtores dessas quebras e provou ser um planejamento viável para investigar os rearranjos.

4.4. Mecanismos de Formação de Rearranjos

Por meio da estratégia utilizada, conseguimos resultados surpreendentes ao observar alterações importantes e definidoras dos pontos de quebra. Foi possível identificar com precisão o mecanismo de formação dos rearranjos e observar a variedade entre os diferentes pacientes.

Com as *linked reads*, foi possível inferir mecanismos baseado no panorama estrutural cromossômico. O mecanismo NHEJ, por ser advindo de uma quebra na dupla fita e posterior junção das mesmas, justifica uma perda de material genético, mas não uma duplicação dos mesmos. Portanto, os casos em que há apenas deleção simples e não há duplicação, são sugestivos de serem decorrentes deste mecanismo. A utilização de *linked reads*, por meio do GemCode, foi de grande ajuda na identificação e confirmação das alterações previamente detectadas, inclusive identificando novas alterações para o sequenciamento.

Também é preciso considerar o fato de trabalhar com DNA previamente extraído impediu a realização do procedimento em algumas amostras, pois a extração para os procedimentos até então utilizados não primavam por uma obtenção de fragmentos de alto peso molecular. Situações que poderão ser evitadas em experimentos futuros.

O mecanismo de NAHR implica numa recorrência dessas CNVs, mediadas principalmente por LCRs. Os mecanismos FoSTeS ou MMBIR podem acarretar em duplicações e deleções concomitantes, pois utilizam como molde segmentos que podem estar localizados em outras regiões do genoma. Portanto, os casos que apresentam esses rearranjos complexos, podem ser decorrentes de um destes tipos de mecanismo. Alinhamentos incorretos durante a replicação devido às sequências repetitivas nestes loci podem acarretar MMBIR e gerar rearranjos complexos dup/del.

Com o sequenciamento dos pontos de quebra por flaqueamento das regiões identificadas por *array*, foi possível identificar com precisão os mecanismos. Com isso, observamos as micro-homologias típicas do FoSTeS e MMBIR, bem como a ausência dessa micro-homologia no rearranjo por NAHR.

Por se tratar de um evento único, a deleção em G-524 foi inicialmente esperada de ser por NHEJ, porém o sequenciamento não revelou a excisão e inserção de bases típicas desse mecanismo. Esse caso mostra que um mecanismo replicativo como FoSTeS e MMBIR podem gerar um pequeno rearranjo sem levar a consequências mais drásticas. Isso ocorre provavelmente com o retorno a fita molde original após a troca de fita molde do início, possivelmente devido à proximidade territorial entre as regiões envolvidas.

Esse estudo revelou que a inovação da detecção da estrutura cromossômica por *linked reads* é muito importante na compreensão da estrutura genômica e pode substituir metodologias de *long reads* de maneira menos custosa e com similar eficiência, sendo útil como ferramenta complementar na identificação de mecanismos causadores de rearranjos genômicos. Em relação ao flanqueamento dos pontos de quebra baseado em *arrays*, a estratégia utilizada foi eficiente em alguns casos, mas deve ser considerado um *array* com maior quantidade de sondas para definir uma região a ser sequenciada.

4.5. Considerações Finais

Como contribuições desse trabalho para a investigação dos mecanismos de formação em rearranjos genômicos utilizando estratégias de sequenciamento de nova geração ressaltamos os pontos abaixo:

- Utilizando a abordagem de flanquear os pontos de quebra por *arrays*, fomos capazes de sequenciar os pontos de quebra de 3 amostras com precisão e identificar por meio de identificação dos mates o ponto de quebra de uma amostra. Também inovamos a aplicar uma nova tecnologia de *linked reads* e identificamos os rearranjos estruturais, delimitando os pontos de quebra e sugerindo mecanismos.
- O sequenciamento dos pontos de quebra genômicos revelou estruturas capazes típicas de mecanismo que foram identificados. Fomos capazes de identificar mecanismos como NAHR envolvendo duas sequências Alu distintas, mas com grande homologia e FoSTeS/MMBIR, mostrando claramente a região de micro-homologia onde a quebra ocorreu. Identificamos ainda mecanismos

como NHEJ pela tecnologia de *linked reads*, que se mostrou também eficiente em revelar rearranjos estruturais.

- A metodologia do sequenciamento foi implantada com sucesso, sendo possível de ser reproduzida e utilizada para futuras pesquisas de ponto de quebra. Ainda identificamos os aspectos que atenção na análise dos arrays para definir a região com maior precisão.

Conclusões

5. Conclusões

Pelo presente estudo:

- Foi possível investigar os mecanismos formadores de rearranjos utilizando duas novas metodologias de SNG em amostras de DNA de pacientes com alterações cromossômicas estruturais.
- Foi possível identificar elementos genômicos presentes nos pontos de quebra e regiões com alterações.
- Foi possível relacionar os elementos encontrados com o tipo de mecanismo formador das alterações, na maioria das amostras estudadas.
- Foi possível padronizar, implantar e desenvolver o método de análise de sequenciamento de nova geração no laboratório de Citogenômica.

Referências

6. Referências

1. Gu W, Zhang F, Lupski JR. Mechanisms for human genomic rearrangements. *PathoGenetics*. 2008;1(1):4.
2. Chen JM, Cooper DN, Ferec C, Kehrer-Sawatzki H, Patrinos GP. Genomic rearrangements in inherited disease and cancer. *Seminars in cancer biology*. 2010;20(4):222-33.
3. Onishi-Seebacher M, Korbel JO. Challenges in studying genomic structural variant formation mechanisms: the short-read dilemma and beyond. *BioEssays : news and reviews in molecular, cellular and developmental biology*. 2011;33(11):840-50.
4. Liu P, Carvalho CM, Hastings PJ, Lupski JR. Mechanisms for recurrent and complex human genomic rearrangements. *Current opinion in genetics & development*. 2012;22(3):211-20.
5. Hastings PJ, Lupski JR, Rosenberg SM, Ira G. Mechanisms of change in gene copy number. *Nature reviews Genetics*. 2009;10(8):551-64.
6. Lupski JR, Stankiewicz P. Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. *PLoS genetics*. 2005;1(6):e49.
7. Stankiewicz P, Lupski JR. Genome architecture, rearrangements and genomic disorders. *Trends in genetics : TIG*. 2002;18(2):74-82.
8. Arlt MF, Wilson TE, Glover TW. Replication stress and mechanisms of CNV formation. *Current opinion in genetics & development*. 2012;22(3):204-10.
9. Verdin H, D'Haene B, Beysen D, Novikova Y, Menten B, Sante T, et al. Microhomology-mediated mechanisms underlie non-recurrent disease-causing microdeletions of the FOXL2 gene or its regulatory domain. *PLoS genetics*. 2013;9(3):e1003358.
10. Weterings E, van Gent DC. The mechanism of non-homologous end-joining: a synopsis of synapsis. *DNA repair*. 2004;3(11):1425-35.
11. Lee JA, Lupski JR. Genomic rearrangements and gene copy-number alterations as a cause of nervous system disorders. *Neuron*. 2006;52(1):103-21.
12. Burma S, Chen BP, Chen DJ. Role of non-homologous end joining (NHEJ) in maintaining genomic integrity. *DNA repair*. 2006;5(9-10):1042-8.
13. Lieber MR. The mechanism of human nonhomologous DNA end joining. *The Journal of biological chemistry*. 2008;283(1):1-5.

14. Jones MJ, Jallepalli PV. Chromothripsis: chromosomes in crisis. *Developmental cell*. 2012;23(5):908-17.
15. Chen JM, Chuzhanova N, Stenson PD, Ferec C, Cooper DN. Complex gene rearrangements caused by serial replication slippage. *Human mutation*. 2005;26(2):125-34.
16. Sheen CR, Jewell UR, Morris CM, Brennan SO, Ferec C, George PM, et al. Double complex mutations involving F8 and FUNDC2 caused by distinct break-induced replication. *Human mutation*. 2007;28(12):1198-206.
17. Yatsenko SA, Hixson P, Roney EK, Scott DA, Schaaf CP, Ng YT, et al. Human subtelomeric copy number gains suggest a DNA replication mechanism for formation: beyond breakage-fusion-bridge for telomere stabilization. *Human genetics*. 2012;131(12):1895-910.
18. Zhang CZ, Leibowitz ML, Pellman D. Chromothripsis and beyond: rapid genome evolution from complex chromosomal rearrangements. *Genes & development*. 2013;27(23):2513-30.
19. Carvalho CM, Pehlivan D, Ramocki MB, Fang P, Alleva B, Franco LM, et al. Replicative mechanisms for CNV formation are error prone. *Nature genetics*. 2013;45(11):1319-26.
20. Zhang F, Carvalho CM, Lupski JR. Complex human chromosomal and genomic rearrangements. *Trends in genetics : TIG*. 2009;25(7):298-307.
21. Spagnolo L, Rivera-Calzada A, Pearl LH, Llorca O. Three-dimensional structure of the human DNA-PKcs/Ku70/Ku80 complex assembled on DNA and its implications for DNA DSB repair. *Molecular cell*. 2006;22(4):511-9.
22. Stephens PJ, Greenman CD, Fu B, Yang F, Bignell GR, Mudie LJ, et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell*. 2011;144(1):27-40.
23. Liu P, Erez A, Nagamani SC, Dhar SU, Kolodziejska KE, Dharmadhikari AV, et al. Chromosome catastrophes involve replication mechanisms generating complex genomic rearrangements. *Cell*. 2011;146(6):889-903.
24. Baca SC, Prandi D, Lawrence MS, Mosquera JM, Romanel A, Drier Y, et al. Punctuated evolution of prostate cancer genomes. *Cell*. 2013;153(3):666-77.
25. Wang K, Wang Y, Collins CC. Chromoplexy: a new paradigm in genome remodeling and evolution. *Asian journal of andrology*. 2013;15(6):711-2.
26. Kidd JM, Graves T, Newman TL, Fulton R, Hayden HS, Malig M, et al. A human genome structural variation sequencing resource reveals insights into mutational mechanisms. *Cell*. 2010;143(5):837-47.

27. Schluth-Bolard C, Labalme A, Cordier MP, Till M, Nadeau G, Tevissen H, et al. Breakpoint mapping by next generation sequencing reveals causative gene disruption in patients carrying apparently balanced chromosome rearrangements with intellectual deficiency and/or congenital malformations. *Journal of medical genetics*. 2013;50(3):144-50.
28. Pang AW, Migita O, Macdonald JR, Feuk L, Scherer SW. Mechanisms of formation of structural variation in a fully sequenced human genome. *Human mutation*. 2013;34(2):345-54.
29. Conrad DF, Bird C, Blackburne B, Lindsay S, Mamanova L, Lee C, et al. Mutation spectrum revealed by breakpoint sequencing of human germline CNVs. *Nature genetics*. 2010;42(5):385-91.
30. Sobreira NL, Gnanakkan V, Walsh M, Marosy B, Wohler E, Thomas G, et al. Characterization of complex chromosomal rearrangements by targeted capture and next-generation sequencing. *Genome research*. 2011;21(10):1720-7.
31. Hermetz KE, Newman S, Conneely KN, Martin CL, Ballif BC, Shaffer LG, et al. Large inverted duplications in the human genome form via a fold-back mechanism. *PLoS genetics*. 2014;10(1):e1004139.
32. Utami KH, Hillmer AM, Aksoy I, Chew EG, Teo AS, Zhang Z, et al. Detection of chromosomal breakpoints in patients with developmental delay and speech disorders. *PloS one*. 2014;9(3):e90852.
33. Liu L, Li Y, Li S, Hu N, He Y, Pong R, et al. Comparison of next-generation sequencing systems. *Journal of biomedicine & biotechnology*. 2012;2012:251364.
34. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, et al. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC genomics*. 2012;13:341.
35. Ambardar S, Gupta R, Trakroo D, Lal R, Vakhlu J. High Throughput Sequencing: An Overview of Sequencing Chemistry. *Indian journal of microbiology*. 2016;56(4):394-404.
36. Zheng GX, Lau BT, Schnall-Levin M, Jarosz M, Bell JM, Hindson CM, et al. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nature biotechnology*. 2016;34(3):303-11.
37. Schaaf CP, Wiszniewska J, Beaudet AL. Copy number and SNP arrays in clinical diagnostics. *Annual review of genomics and human genetics*. 2011;12:25-51.
38. Manning M, Hudgins L, Professional P, Guidelines C. Array-based technology and recommendations for utilization in medical genetics practice for

detection of chromosomal abnormalities. *Genetics in medicine : official journal of the American College of Medical Genetics*. 2010;12(11):742-5.

39. Armstrong J, Fiddes IT, Diekhans M, Paten B. Whole-Genome Alignment and Comparative Annotation. *Annual review of animal biosciences*. 2018.

40. Wong KHY, Levy-Sakin M, Kwok PY. De novo human genome assemblies reveal spectrum of alternative haplotypes in diverse populations. *Nature communications*. 2018;9(1):3040.

41. Liu H, Liu Q, Chen Z, Liu Y, Zhou C, Liang Q, et al. Draft genome of *Glyptosternon maculatum*, an endemic fish from Tibet Plateau. *GigaScience*. 2018;7(9).

42. Richards S. Arthropod Genome Sequencing and Assembly Strategies. *Methods in molecular biology*. 2019;1858:1-14.

43. Stapleton JA, Kim J, Hamilton JP, Wu M, Irber LC, Maddamsetti R, et al. Haplotype-Phased Synthetic Long Reads from Short-Read Sequencing. *PloS one*. 2016;11(1):e0147229.

44. Schmid CW, Deininger PL. Sequence organization of the human genome. *Cell*. 1975;6(3):345-58.

45. Jurka J, Smith T. A fundamental division in the Alu family of repeated sequences. *Proceedings of the National Academy of Sciences of the United States of America*. 1988;85(13):4775-8.

46. Roy-Engel AM, Carroll ML, Vogel E, Garber RK, Nguyen SV, Salem AH, et al. Alu insertion polymorphisms for the study of human genomic diversity. *Genetics*. 2001;159(1):279-90.

47. Lebedev YB, Amosova AL, Mamedov IZ, Fisunov GY, Sverdlov ED. Most recent AluY insertions in human gene introns reduce the content of the primary transcripts in a cell type specific manner. *Gene*. 2007;390(1-2):122-9.

48. Nozu K, Iijima K, Ohtsuka Y, Fu XJ, Kaito H, Nakanishi K, et al. Alport syndrome caused by a COL4A5 deletion and exonization of an adjacent AluY. *Molecular genetics & genomic medicine*. 2014;2(5):451-3.

49. Kurosaki T, Ueda S, Ishida T, Abe K, Ohno K, Matsuura T. The unstable CCTG repeat responsible for myotonic dystrophy type 2 originates from an AluSx element insertion into an early primate genome. *PloS one*. 2012;7(6):e38379.

50. Costa E, Duque F, Oliveira J, Garcia P, Goncalves I, Diogo L, et al. Identification of a novel AluSx-mediated deletion of exon 3 in the SBDS gene in a patient with Shwachman-Diamond syndrome. *Blood cells, molecules & diseases*. 2007;39(1):96-101.

51. Bennett EA, Keller H, Mills RE, Schmidt S, Moran JV, Weichenrieder O, et al. Active Alu retrotransposons in the human genome. *Genome research*. 2008;18(12):1875-83.
52. Elliott B, Richardson C, Jasin M. Chromosomal translocation mechanisms at intronic alu elements in mammalian cells. *Molecular cell*. 2005;17(6):885-94.
53. Jjingo D, Conley AB, Wang J, Marino-Ramirez L, Lunyak VV, Jordan IK. Mammalian-wide interspersed repeat (MIR)-derived enhancers and the regulation of human gene expression. *Mobile DNA*. 2014;5:14.
54. Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, et al. The Dfam database of repetitive DNA families. *Nucleic acids research*. 2016;44(D1):D81-9.
55. Robberecht C, Voet T, Zamani Esteki M, Nowakowska BA, Vermeesch JR. Nonallelic homologous recombination between retrotransposable elements is a driver of de novo unbalanced translocations. *Genome research*. 2013;23(3):411-8.
56. Ou Z, Stankiewicz P, Xia Z, Breman AM, Dawson B, Wiszniewska J, et al. Observation and prediction of recurrent human translocations mediated by NAHR between nonhomologous chromosomes. *Genome research*. 2011;21(1):33-46.