UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE QUÍMICA
Programa de Pós-Graduação em Ciências Biológicas (Bioquímica)

ELISA RENNÓ DONNARD MOREIRA

Estudo de Variações Genômicas para a Identificação de Biomarcadores Personalizados e Novos Alvos Terapêuticos em Tumores Colorretais

Versão corrigida da Tese defendida

São Paulo
Data do Depósito na SPG:
04/06/2014

ELISA RENNÓ DONNARD MOREIRA

Estudo de Variações Genômicas para a Identificação de Biomarcadores Personalizados e Novos Alvos Terapêuticos em Tumores Colorretais

Tese apresentada ao Instituto de Química da Universidade de São Paulo para obtenção do Título de Doutor em Ciências (Bioquímica)

Orientadora: Dra. Anamaria Aranha Camargo
Co-Orientador: Dr. Pedro A. F. Galante

São Paulo
2014

Ficha Catalográfica
Elaborada pela Divisão de Biblioteca e
Documentação do Conjunto das Químicas da USP.

Moreira, Elisa Rennó Donnard
M838e Estudo de variações genômicas para a identificação de biomarcadores personalizados e novos alvos terapêuticos em tumores colorretais / Elisa Rennó Donnard Moreira. -- São Paulo, 2014.
114p.

Tese (doutorado) - Instituto de Química da Universidade de São Paulo. Departamento de Biquímica.
Orientador: Camargo, Anamaria Aranha
Co-orientador: Galante, Pedro Alexandre Favoretto

1. Genoma : Biologia molecular 2. Bioinformática 3. Neoplasias I. T. II. Camargo, Anamaria Aranha, orientador. III, Galante, Pedro Alexandre Favoretto.

574.88 CDD

À Marie Yvonne Donnard

AGRADECIMENTOS

Agradeço ao meu pai e minha mãe pelo apoio constante, pelo laboratório que despertou minha fascinação por experimentos aos sete anos de idade e as aventuras que me formaram como pessoa. Agradeço a minha irmã por ter me treinado com sua experiência e me ensinado até mesmo a escrever meu primeiro caderno de laboratório.

Agradeço à Anamaria Camargo pela acolhida e orientação. À FAPESP pelo apoio financeiro durante o doutorado. Ao Hospital Sírio-Libanês e ao Instituto Ludwig de Pesquisa sobre o Câncer por tornarem possível o desenvolvimento deste trabalho e a todos os colaboradores dos projetos que contribuíram para os resultados aqui apresentados.

Agradeço a todos os amigos e mentores que me acompanharam de alguma forma no meu caminho científico e que me ensinaram tanto. Concluo mais uma etapa com o apoio de todos vocês.

Por fim, agradeço especialmente aos orientadores que me ensinaram tanto e continuam me tornando uma melhor cientista, Miguel Ortega, Daniel Gietz e Pedro Galante. Muito obrigada!

“Coming back to where you started is not the same as never leaving.”

— Terry Pratchett, A Hat Full of Sky

RESUMO

DONNARD, E.R. **Estudo de Variações Genômicas para a Identificação de Biomarcadores Personalizados e Novos Alvos Terapêuticos em Tumores Colorretais**. 2014. 109p. Tese - Programa de Pós-Graduação em Bioquímica. Instituto de Química, Universidade de São Paulo, São Paulo.

O câncer colorretal é um dos tipos de tumores mais frequentes no mundo. A atual dificuldade na avaliação correta da resposta ao tratamento torna necessário o desenvolvimento de novas abordagens de detecção tumoral. Atualmente, o sequenciamento genômico em larga escala permite um estudo mais compreensivo das alterações estruturais e de sequência presentes no tumor. A aplicação destas abordagens de maneira personalizada permite o desenvolvimento de biomarcadores tumor específicos que podem facilitar a avaliação de resposta ao tratamento e a presença de doença residual, bem como revelar alterações de sequência em genes capazes de servir de novos alvos terapêuticos. Neste estudo foi desenvolvida uma metodologia eficiente para a identificação de biomarcadores baseados na existência de variações estruturais em genomas de tumores de reto, eliminando a necessidade de sequenciamento do genoma normal do mesmo paciente e diminuindo portanto o custo da abordagem. Os biomarcadores encontrados para cada um dos seis pacientes foram utilizados para avaliar a presença de doença residual após o tratamento através da detecção de DNA tumoral circulante nas amostras de plasma coletadas em momentos diferentes do tratamento. O sequenciamento em baixa cobertura personalizado é portanto uma alternativa viável e promissora para avaliar a resposta ao tratamento em pacientes com tumores de reto. Na segunda parte do estudo, a análise de linhagens celulares de tumores colorretais revelou uma grande quantidade de mutações pontuais somáticas (SNVs e InDels) em genes codificadores para proteínas de superfície celular (surfaceoma). Estas alterações no surfaceoma indicam potenciais novos alvos para drogas e vias regulatórias alteradas neste tipo de tumor. Além disso, estas mutações pontuais também são responsáveis pela geração de epítopos com potencial imunogênico e estes novos epítopos podem ser aplicados como vacinas antitumorais personalizadas e já haviam sido propostos como uma alternativa terapêutica. A presença de novos epítopos, principalmente nas linhagens com elevadas taxas de mutação (resultante da instabilidade de microssatélites e mutações em genes de reparo de DNA tipo *mismatch* ou POLE), sugerem também um potencial uso de drogas moduladoras do sistema imune em pacientes com tumores que apresentam estas mesmas características. Portanto, o estudo de alterações genômicas em tumores primários e linhagens de câncer colorretal permitiu a detecção de variações estruturais que foram utilizadas como biomarcadores personalizados em pacientes com tumores de reto assim como a identificação de genes contendo mutações pontuais em linhagens celulares de câncer colorretal, que revelam potenciais novos alvos terapêuticos a serem explorados na clínica.

Palavras-chave: câncer, bioinformática genoma, mutações, biomarcadores

ABSTRACT

DONNARD, E.R. **Study of Genomic Variation to Identify Biomarkers and Novel Therapeutic Targets in Colorectal Tumors.** 2014. 109p. PhD Thesis - Graduate Program in Biochemistry. Instituto de Química, Universidade de São Paulo, São Paulo.

Colorectal cancer is one of the more frequent tumor types in the world. To select the appropriate treatment course, it is necessary to develop more precise diagnostic approaches. The current availability of high throughput genome sequencing methods allows for a comprehensive characterization of the structural and sequence alterations present in each tumor. The use of tumor genome sequencing in a personalized setting can result in tumor specific biomarkers that help evaluate response to treatment and the presence of residual disease, improving the clinical management of these patients, as well as reveal sequence alterations in genes capable of serving as new therapeutic targets. In this study we developed an efficient bioinformatics pipeline to identify biomarkers based on the existing structural alterations in rectal tumor genomes, eliminating the need to sequence the matched normal genome and therefore reducing the cost for this approach. The biomarkers found for each of the six patients were used to evaluate the presence of residual disease after treatment through the detection of circulating tumor DNA in plasma samples collected at different points during the treatment. Sequencing tumor genomes with low coverage is therefore a viable and promising alternative to follow up rectal cancer patient's response to treatment. In the second part of this study, the analysis of colorectal cancer cell lines revealed a large quantity of point mutations (SNVs and InDels) in genes coding for proteins located in the cell surface (surfaceome). These alterations in the surfaceome indicate potential new drug targets and altered pathways in this type of tumor. Furthermore, these point mutations are also responsible for the generation of new epitopes with immunogenic potential and these new epitopes can be applied as personalized tumor vaccines and had previously been proposed as a therapeutic alternative. The presence of new epitopes, especially in the cell lines with elevated mutation rates (resulting from MSI and mutations in DNA mismatch-repair genes or POLE), suggests a potential use of immune checkpoint target drugs in patients with tumors that share these genetic characteristics. With a large-scale bioinformatics approach, we detected new tumor epitopes resulting from point mutations, present in most of the cell lines used. The analysis of gene expression data puts into perspective both the somatic mutations found and which targets are promising as well as the development of therapies based on vaccines derived from tumor epitopes. In conclusion, the study of genomic alterations in primary tumors and colorectal cancer cell lines allowed the detection of structural variations that were used as personalized biomarkers in patients with rectal tumors as well as the identification of genes containing point mutations in colorectal cancer cell lines, that reveal potential new therapeutic targets to be explored in the clinical setting.

Keywords: cancer, bioinformatics, genome, mutations, biomarkers

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS

5-FU – 5-Fluorouracil
cDNA – DNA complementar
CNA – *Copy Number Alteration*
ctDNA – *Circulating Tumor* DNA
dbSNP – SNP *Database*
DGIdb – *Drug Gene Interaction Database*
EGFR – *Epidermal Growth Factor Receptor*
FGFR – *Fibroblast Growth Factor Receptor*
FPKM – *Fragments per Kilobase of Exon per Million Fragments Mapped*
HLA – *Human Leukocyte Antigen*
InDel – Inserção ou Deleção
KEGG – *Kyoto Enciclopedia of Genes and Genomes*
KRAS – *Kirsten rat sarcoma viral oncogene homolog*
LINE – *Long Interspersed Element*
MHC – *Major Histocompatibility Complex*
MMR – *Mismatch Repair*
mRNA – RNA mensageiro
MSI – *Microssatellite Instability*
NCBI – *National Center for Biotechnology Information*
PCR – *Polimerase Chain Reaction*
PARE – Personalized Analysis of Rearranged Ends
POLE – DNA Polimerase Epsilon
QRT – Radioquimioterapia
rRNA – RNA ribossomal
SINE – *Short Interpersed Element*
SNP – *Single Nucleotide Polymorphism*
SNV – *Single Nucleotide Variation*
TCGA – *The Cancer Genome Atlas*
TME – *Total Mesorectal Excision*
TP53 – *Tumor Protein* 53
VEGF – *Vascular Endothelial Growth Factor*

SUMÁRIO

1 INTRODUÇÃO

1.1 Câncer Colorretal

O câncer colorretal está entre os tipos de tumores mais frequentes no mundo, correspondendo a 9,7% do total ou, em números absolutos, 1,4 milhão de novos casos estimados por ano (Ferlay *et al.*, 2013). Especificamente, o câncer do reto parece ter contribuição significativa nestes números, correspondendo a 40 mil novos casos diagnosticados por ano nos Estados Unidos (American Cancer Society, 2014). No Brasil, são cerca de 30 mil novos casos anuais de tumores colorretais (INCA, 2012) e não existem dados específicos para tumores de reto. Outro fator importante é a mortalidade deste tipo de câncer, que chega a 700.000 casos por ano no mundo (Ferlay *et al.*, 2013). No Brasil a mortalidade é estimada em cerca de 7 mil pacientes por ano, correspondendo a 5ª causa de morte por câncer em nosso país (INCA, 2012).

Os tumores colorretais em sua maioria são adenocarcinomas e se desenvolvem (~95% das vezes) de forma lenta a partir de um pólipo adenomatoso, considerado uma condição pré-cancerosa (Kang *et al.*, 2007).

Os sintomas apresentados em casos de câncer colorretal incluem sangramento gastrointestinal, mudança de hábitos intestinais, dor abdominal, obstrução intestinal, perda de peso, mudança de apetite e fraqueza (Stein *et al.*, 1993). Os sintomas apresentados não indicam o estádio da doença ou significam um prognóstico específico (Majumdar, Fletcher e Evans, 1999). Um método de detecção precoce para câncer colorretal é o exame de sangue oculto nas fezes, que pode ser uma ferramenta útil na triagem de populações, mas ainda não se sabe avaliar corretamente o impacto desta triagem na redução da mortalidade (Towler *et al.*, 1998). Ao menos na metade dos casos

de câncer colorretal, os pacientes manifestam somente sintomas de baixo-risco como constipação ou dor abdominal e não existe um teste intermediário para identificar aquelas pessoas que provavelmente apresentam um tumor. O principal meio diagnóstico para uma suspeita de câncer colorretal é a colonoscopia para detecção dos pólipos e lesões malignas (Hamilton *et al.*, 2009).

O cólon é composto de quatro seções, são elas cólon ascendente, transverso, descendente e sigmoide (ilustradas na Figura 1A). A definição do limite entre cólon e reto é somente anatômica (Figura 1A e B), sendo considerado parte do reto todo o canal de até 12cm a partir da borda anal (Nelson *et al.*, 2001). As abordagens terapêuticas para tumores localizados no cólon ou no reto diferem, mas a detecção e avaliação do grau de evolução do tumor são similares.

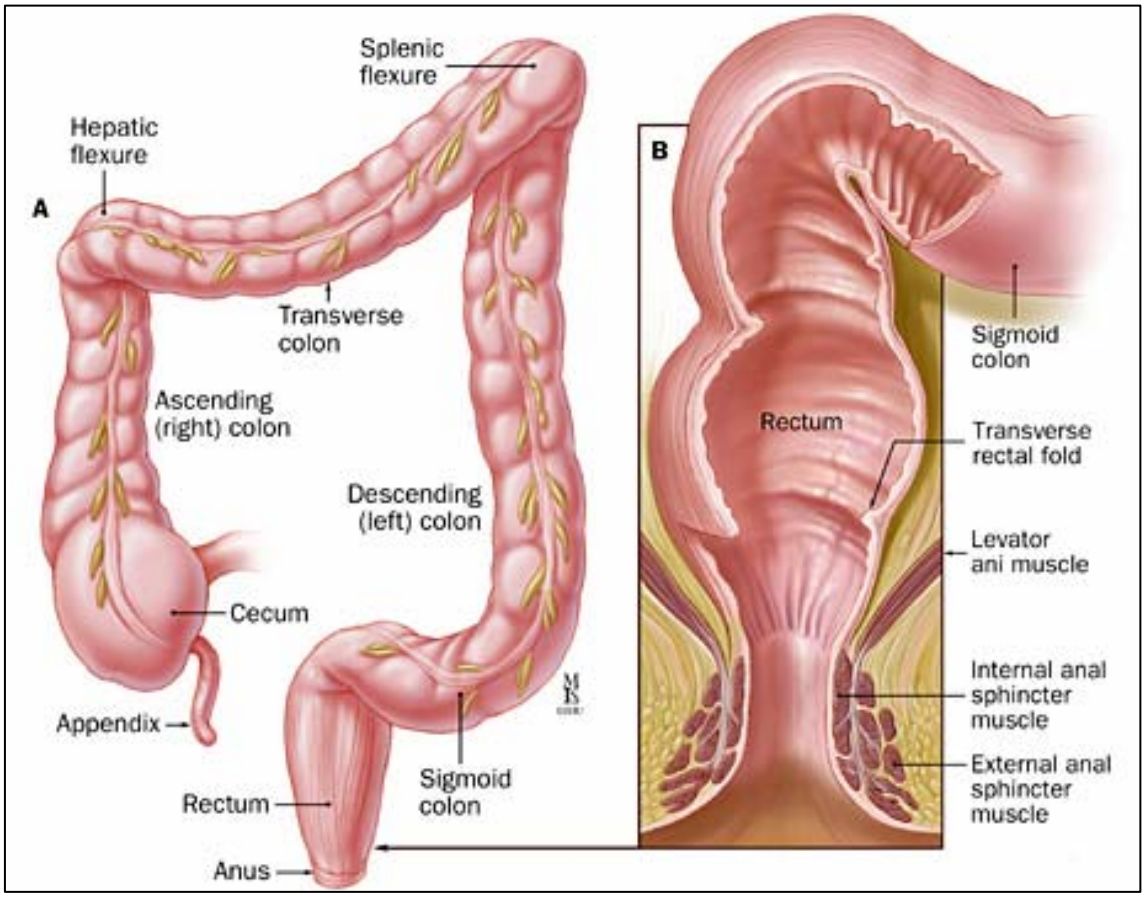


Figura 1 – Anatomia normal do cólon (A) e reto (B). (Figura: Johns Hopkins Medicine)

O estadiamento correto da doença auxilia na determinação do curso de terapia e, para tumores de reto, impacta diretamente a seleção da terapia neoadjuvante e abordagem cirúrgica (Schmidt, Gollub e Weiser, 2007). O estadiamento atualmente considerado mais específico é chamado de TNM e leva em consideração três parâmetros: o grau de penetração do tumor na parede do cólon ou reto (T), a presença de metástases linfonodais (N) ou ainda a presença de metástases (M) à distância detectadas através de exames de tomografia (CT) (Edge e Compton, 2010). A Figura 2 ilustra os diferentes estádios da doença, que também pode ter os estádios nomeados numericamente (0, I, II, III, IV), representando diferentes quadros da classificação TNM. O grau de penetração do tumor na parede do reto corresponde a um maior risco de envolvimento de linfonodos e recorrência local (Kosinski *et al.*, 2012).

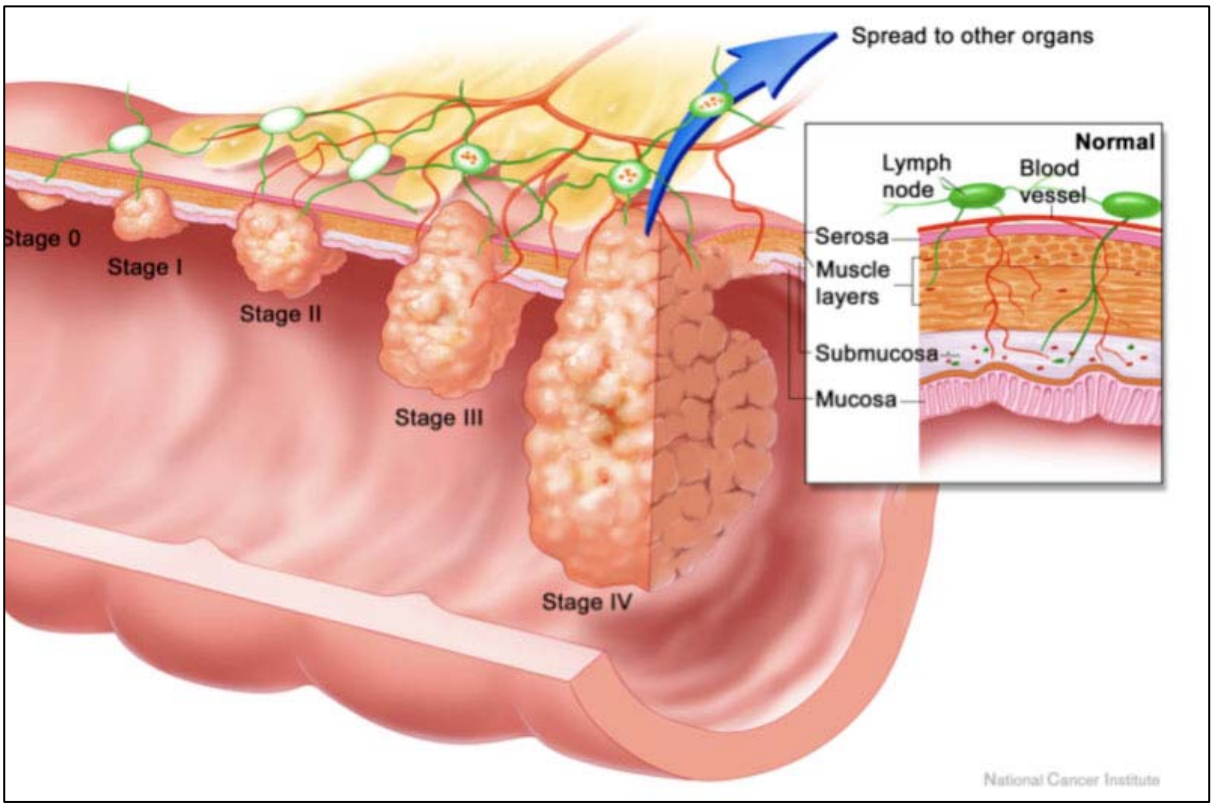


Figura 2 – Estadiamento de tumores colorretais. O desenho ilustra uma seção do cólon com tumores em diferentes estadiamentos (0-IV; Figura: National Cancer Institute, EUA).

1.2 Tratamento

Devido às diferenças anatômicas, o tratamento para tumores de reto e de cólon é distinto. A localização do reto torna complicada a remoção cirúrgica do tecido afetado com boas margens de ressecção, e mesmo quando são adotadas práticas cirúrgicas menos radicais as consequências pós-operatórias podem ser debilitantes. A cirurgia de tumores de reto está associada a elevada morbidade, mortalidade e pode resultar na implementação de um estoma permanente (Habr-Gama e Perez, 2009). No caso dos tumores de cólon a abordagem cirúrgica para tumores localizados é mais simples e na maioria das vezes não traz consequências tão prejudiciais ao paciente. As considerações específicas de cada tratamento estão detalhadas abaixo.

1.2.1 Tumores de reto

A abordagem terapêutica atual para o câncer de reto é complexa e pode envolver diversas modalidades como a cirurgia, a radioterapia e a combinação de radio e quimioterapia. A ressecção cirúrgica do tumor primário é o tratamento de escolha para os tumores de reto e quando aliada a excisão total do mesorreto (TME) é muito eficiente no controle local da doença (Kosinski *et al.*, 2012). No entanto, em pacientes com tumores com invasão da parede do reto além da camada muscular sem linfonodos comprometidos (estádio II) ou ainda com a presença de metástases linfonodais (estádio III) o uso de terapia adjuvante, envolvendo radio e quimioterapia, é recomendado para reduzir os riscos de recorrência local e sistêmica da doença. Além disso, o tratamento não-cirúrgico para alguns casos de tumores de reto começa a ser visto como uma

abordagem promissora (Habr-Gama e Perez, 2009; Habr-Gama *et al.*, 1998; Habr-Gama, Perez, Nadalin, *et al.*, 2004; Kosinski *et al.*, 2012).

A combinação de radio e quimioterapia tem sido amplamente utilizada para complementar o tratamento cirúrgico dos pacientes com tumor de reto e pode ser administrada tanto antes (tratamento neoadjuvante) como após (tratamento adjuvante) a cirurgia. A utilização de radioquimioterapia (QRT) neoadjuvante tem se mostrado mais eficiente do que a radioquimioterapia adjuvante no controle local da doença de acordo com estudos randomizados, e parece ter também maior impacto na sobrevida global dos pacientes quando comparada ao uso do tratamento adjuvante (Habr-Gama *et al.*, 2008; Habr-Gama, Perez, Kiss, *et al.*, 2004; Kosinski *et al.*, 2012; Sauer *et al.*, 2004). Além disso, o uso de radioquimioterapia neoadjuvante pode levar a uma redução significativa da massa tumoral permitindo a ressecção total do tumor e aumentando as chances de preservação da função esfincteriana (Habr-Gama *et al.*, 2008; Habr-Gama, Perez, Kiss, *et al.*, 2004; Kosinski *et al.*, 2012; Sauer *et al.*, 2004). A terapia neoadjuvante é administrada em regime curto ou longo aos pacientes e seis a oito semanas após o curso normal de tratamento é realizada a cirurgia radical (Figura 3A). Atualmente, foi proposto pelo grupo da Dra. Angelita Habr-Gama uma abordagem diferente no tratamento de tumores de reto (Figura 3B) que procura explorar estratégias cirúrgicas menos radicais. A QRT neoadjuvante é aplicada em regime longo e após o intervalo de descanso é feita uma reavaliação do paciente. Esta reavaliação pode ser feita por exames de toque retal ou de imagem por ultrassom ou tomografia. O resultado deste exame indica se o paciente deve passar por uma abordagem cirúrgica ou se deve ser mantido sob monitoramento frequente, estratégia chamada de "*Watch and Wait*". Infelizmente, a imprecisão nos métodos utilizados para fazer esta reavaliação torna difícil sua

implementação e a maioria dos grupos clínicos continuam optando pela cirurgia mesmo nos casos em que não existem evidências da presença tumoral.

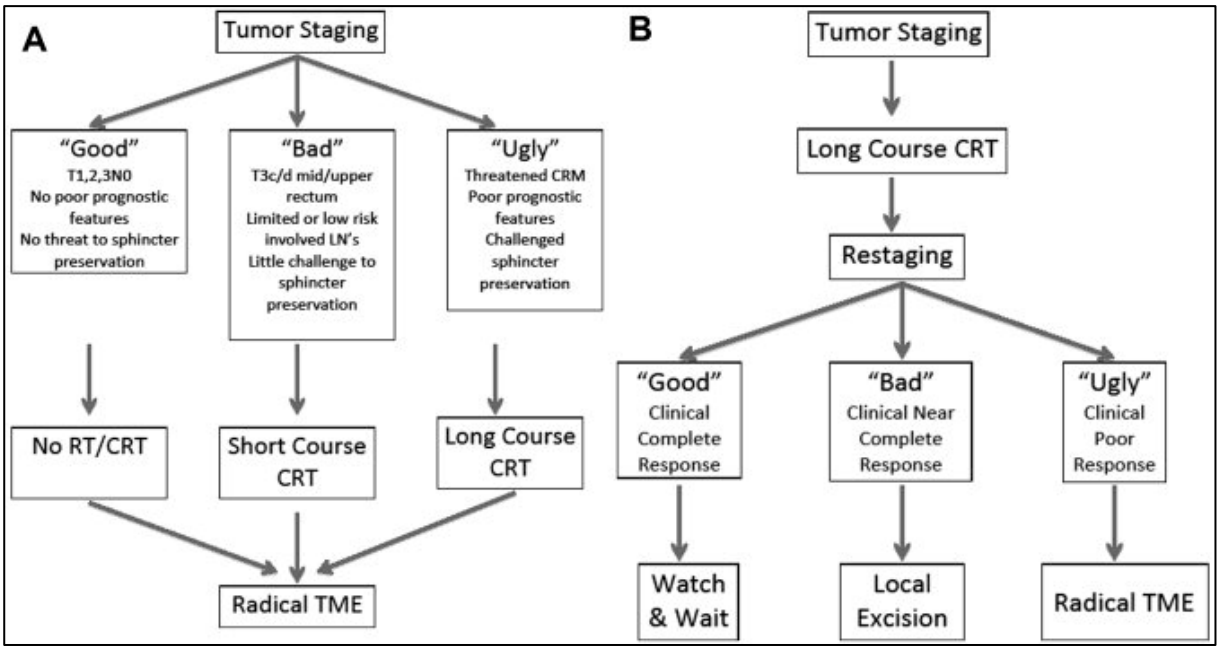


Figura 3 – Estratégias de tratamento para tumores de reto. (A) Estratégia de seleção do regime neoadjuvante, todos os pacientes são operados no final. (B) Estratégia de seleção da abordagem cirúrgica após reavaliação da resposta tumoral ao tratamento QRT. LN = linfonodos; CRM = margens de ressecção circunferenciais; RT = radioterapia; CRT = radioquimioterapia; TME = excisão total do mesorreto. (Figura: Kozinski et al., 2012)

A resposta ao tratamento neoadjuvante varia significativamente de paciente para paciente mesmo dentro de uma mesma instituição e apenas o grupo de pacientes que respondem ao tratamento neoadjuvante apresenta benefícios comprovados no que diz respeito ao aumento da sobrevida global (Habr-Gama *et al.*, 2005; Habr-Gama, Perez, Kiss, *et al.*, 2004; Sauer *et al.*, 2004). O leque de resposta observado entre os pacientes com tumores de reto inclui desde aqueles com resposta completa ao tratamento até outros com tumores considerados radioquimio-resistentes (Kosinski *et al.*, 2012). Este fato cria um grande dilema na escolha da melhor opção terapêutica para ser aplicada nesses pacientes. Se por um lado os pacientes resistentes ao tratamento poderiam ser poupados dos efeitos colaterais indesejáveis da radioquimioterapia e submetidos

imediatamente a cirurgia, reduzindo o risco de disseminação da doença, os pacientes que respondem ao tratamento se beneficiam com a redução da massa tumoral e consequentemente com uma maior chance de ressecção completa do tumor e preservação da função esfincteriana. Além disso existem pacientes com o quadro de resposta chamado completa patológica (de 5 a 42%) para os quais os exames clínicos e de imagem não são sensíveis e específicos o suficiente para descartar a presença de doença residual, e o paciente é submetido à cirurgia, mas uma análise histopatológica posterior da peça cirúrgica revela tecido necrótico e fibrótico e a ausência de células tumorais viáveis, e o paciente passou portanto por uma intervenção cirúrgica desnecessária (Kosinski *et al.*, 2012). A fração de pacientes que apresentam resposta clínica completa após o tratamento neoadjuvante é significativa (10-30%), e para a maioria destes, a intervenção cirúrgica não seria indicada devido aos seus riscos e co-morbidades (Habr-Gama *et al.*, 2008). O desenvolvimento de uma abordagem capaz de avaliar a resposta ao tratamento e a presença de doença residual com maior precisão, para excluir estes pacientes de uma cirurgia desnecessária, é de grande interesse para um tratamento mais adequado e eficiente. Esta questão é uma dos principais focos desta tese.

1.2.2 Tumores de cólon

No caso do câncer de cólon em estádios iniciais (0-II) a ressecção cirúrgica é eficaz para a quase totalidade dos casos, enquanto que para o tratamento de estádios mais avançados (III-IV) é necessário incluir a quimioterapia sistêmica para aumentar as chances de controle local (André *et al.*, 2004). As principais drogas usadas no

tratamento quimioterápico do câncer colorretal são os antimetabólitos 5-FU (fluorouracila) e oxaliplatina. O mecanismo de ação da 5-FU é baseado em sua conversão pela célula à fluorouridina trifosfato (FUTP) ou fluorodeoxiuridina trifosfato (FdUTP) que inibem a enzima timidilato sintase ou ainda são incorporadas ao DNA e RNA celular, respectivamente, inibindo a sua função normal (Longley, Harkin e Johnston, 2003). A oxaliplatina é um composto que forma ligações na mesma fita do DNA entre dois resíduos adjacentes de guanina ou entre um resíduo de guanina e adenina adjacentes (Fink *et al.*, 1997), impedindo os processos de replicação e transcrição. Os dois compostos levam as células tumorais à apoptose, mas apresentam alta toxicidade às células normais e as taxas de resposta ao tratamento nos casos de câncer colorretal avançado são de 10-15% para 5-FU (Johnston e Kaye, 2001). A combinação dos quimioterápicos aumenta a eficácia do tratamento para 40-50% mas ainda assim novas estratégias terapêuticas são necessárias (Longley, Harkin e Johnston, 2003). Em muitos casos os tumores apresentam resistência aos dois compostos (Arango *et al.*, 2004; Mariadason *et al.*, 2003), e boa parte dos mecanismos por trás da resistência ainda não são bem compreendidos (Longley, Harkin e Johnston, 2003). É bem conhecido que as células tumorais possuem fenótipos complexos resultantes da grande quantidade de variações presentes no seu genoma e consequentemente das alterações na expressão e regulação de diversos genes. Por exemplo, mutações em genes que fazem parte do sistema de reparo do tipo *mismatch* (MMR), muito comuns em tumores colorretais, resultam em resistência muito maior ao tratamento com 5-FU, provavelmente devido a uma maior tolerância a danos no DNA (Meyers *et al.*, 2001).

As drogas mencionadas acima já são aplicadas no tratamento de tumores colorretais há cerca de 20 anos, com eficiência limitada. Para o tratamento de metástases, é comum

atualmente o uso de anticorpos ou moléculas anti-VEGF (*Vascular Endothelial Growth Factor*) com benefícios modestos (Cutsem, Van *et al.*, 2012; Hurwitz *et al.*, 2004; Troiani *et al.*, 2013). O presente trabalho pretende abordar o problema da escassez de novas drogas para o tratamento de tumores colorretais, assim como das metástases originadas destes tumores.

1.3 O genoma tumoral

A instabilidade genômica em células tumorais, uma das principais características comuns a todos os tipos de câncer, é considerada um dos fatores que permitem o desenvolvimento do tumor, favorecendo a aquisição das modificações necessárias para o escape do sistema imune e da apoptose, por exemplo (Hanahan e Weinberg, 2011).

As tecnologias de sequenciamento de segunda geração permitiram uma melhor caracterização das neoplasias malignas, tornando possível o sequenciamento em larga escala de genes expressos (transcriptomas), regiões codificadoras (exomas) e genomas tumorais completos (Meyerson, Gabriel e Getz, 2010). A redução no custo de sequenciamentos em larga escala torna possível o estudo detalhado de diversos tipos tumorais. Consequentemente, a caracterização das mudanças genéticas associadas com a iniciação e progressão do câncer colorretal tem sido explorada por diferentes grupos de pesquisa (Cunha, da *et al.*, 2009; Katkoori *et al.*, 2009; Starr *et al.*, 2009; TCGA, 2012; Wood *et al.*, 2007), permitindo a identificação de marcadores biológicos para esta doença e possíveis alvos moleculares para sua detecção e tratamento. Embora a redução no custo seja favorável para o desenvolvimento de pesquisas, o uso dessa tecnologia na prática clínica permanece restrito devido ao seu alto valor.

A análise das sequências geradas em larga escala para detectar rearranjos estruturais e mutações pontuais no genoma tumoral depende do seu alinhamento ao genoma humano referência. Contudo, devido à natureza repetitiva do genoma (Lander *et al.*, 2001) assim como a presença de alterações polimórficas presentes nos genomas das células normais (Abecasis *et al.*, 2010), determinar as posições corretas de mapeamento para cada sequência e identificar os casos somáticos de rearranjos estruturais e mutações pontuais não é um processo simples (Garraway e Lander, 2013; Meyerson, Gabriel e Getz, 2010). Para a detecção destas variações genômicas, faz-se necessária a utilização de um conjuntos de ferramentas computacionais (*pipelines*) para analisar os resultados do sequenciamento em larga escala. Normalmente um número grande de alinhamentos falso-positivos é identificado, diminuindo assim a eficiência e a precisão da técnica (Bass *et al.*, 2011). Para reduzir a complexidade da análise e o número de rearranjos falsos, é considerado essencial o sequenciamento do genoma normal do mesmo paciente (Chen *et al.*, 2009; Drier *et al.*, 2013; Meyerson, Gabriel e Getz, 2010). Infelizmente, esta prática pode aumentar muito o custo de sequenciamento e de processamento computacional, limitando o uso desta técnica no cenário clínico.

A maioria dos tumores sólidos apresenta uma grande quantidade de rearranjos cromossômicos resultantes da instabilidade cromossômica das células tumorais. Estas alterações estruturais podem incluir variações do número de cópias, inversões, inserções, deleções e translocações (Mitelman, Johansson e Mertens, 2007). Estes rearranjos ocorrem nos estágios iniciais da tumorigênese, e persistem durante todo o desenvolvimento do tumor (Leary *et al.*, 2010). O uso de sequenciadores de nova geração possibilita a identificação destas alterações cromossômicas somáticas, através da análise do mapeamento das sequências geradas no genoma humano referência. A

Figura 4 ilustra os diversos padrões de mapeamento que podem ser encontrados nas análises de sequenciamento indicando rearranjos estruturais no genoma de estudo, quando comparado ao genoma referência.

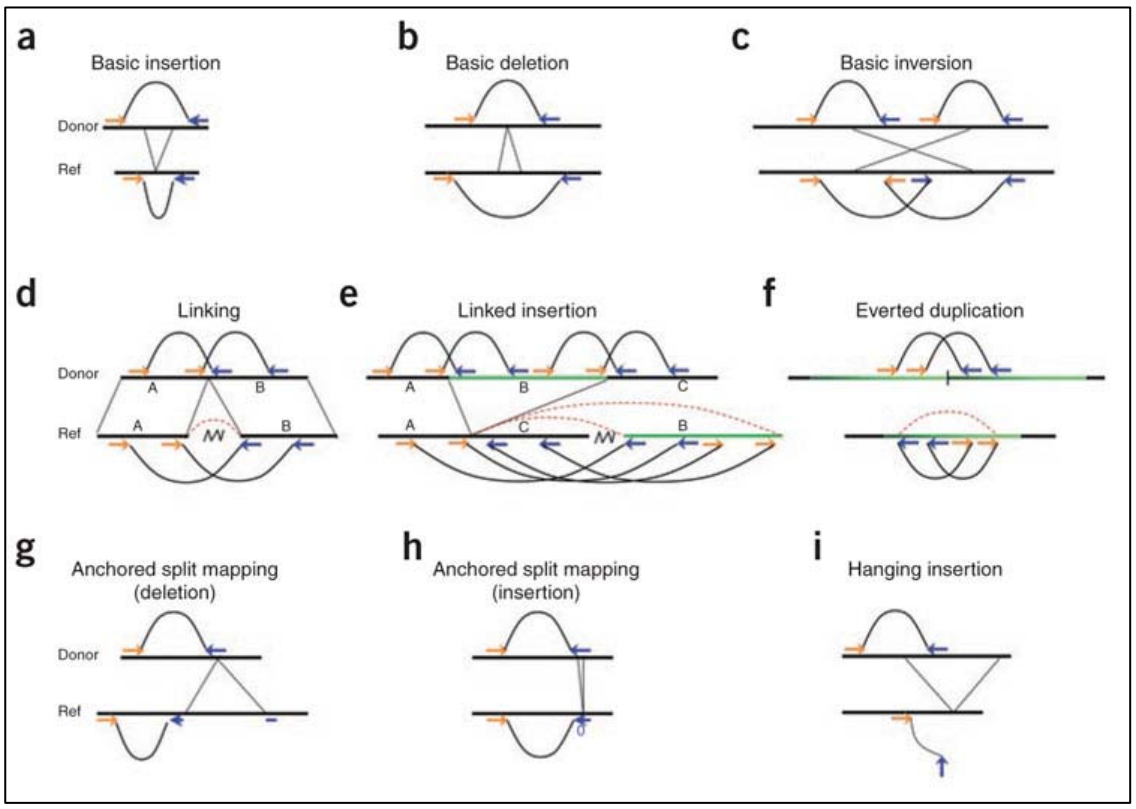


Figura 4 - Padrões de alinhamento de sequências indicando variações estruturais no genoma. Ilustração dos diversos e complexos padrões de alinhamento que podem ser observados na análise de sequenciamento em larga escala indicando alterações estruturais no genoma com relação ao genoma referência. As setas amarelas e azuis indicam sequências alinhadas no genoma e a conexão (traço preto) indica que elas pertencem a um mesmo par. Donor = genoma tumoral; Ref = genoma referência (Figura: Medvedev et al., 2009).

Diversos grupos de pesquisa desenvolveram recentemente pipelines que identificam variações estruturais somáticas (Chen *et al.*, 2009; Malhotra *et al.*, 2013; Medvedev, Stanciu e Brudno, 2009; Talkowski *et al.*, 2011). Estas abordagens são baseadas na análise comparativa do genoma tumoral e normal e diferem nos filtros aplicados e no rigor com o qual os candidatos finais são considerados, e ainda dependem da geração de maiores coberturas de sequência. Para a implementação imediata nos diagnósticos e

acompanhamentos de rotina, adaptações a coberturas menores de sequenciamento e à ausência da comparação com o genoma normal pareado ainda são necessárias à esta técnica promissora.

Em tumores sólidos comuns como os de mama e colorretais existem em média 33 e 66 mutações pontuais somáticas respectivamente (Figura 5A), que são capazes de alterar o produto proteico (Vogelstein *et al.*, 2013). Dentre as mutações pontuais mencionadas incluem-se as variações de um único nucleotídeo (*Single Nucleotide Variations* ou SNVs) e as pequenas inserções e deleções, geralmente menores do que 50nt, chamadas InDels (Vogelstein *et al.*, 2013). Alguns tipos tumorais apresentam um número muito maior de mutações, dependendo do envolvimento de agentes mutagênicos (cigarro no caso de câncer de pulmão) e defeitos no sistema de reparo de DNA (Figura 5B).

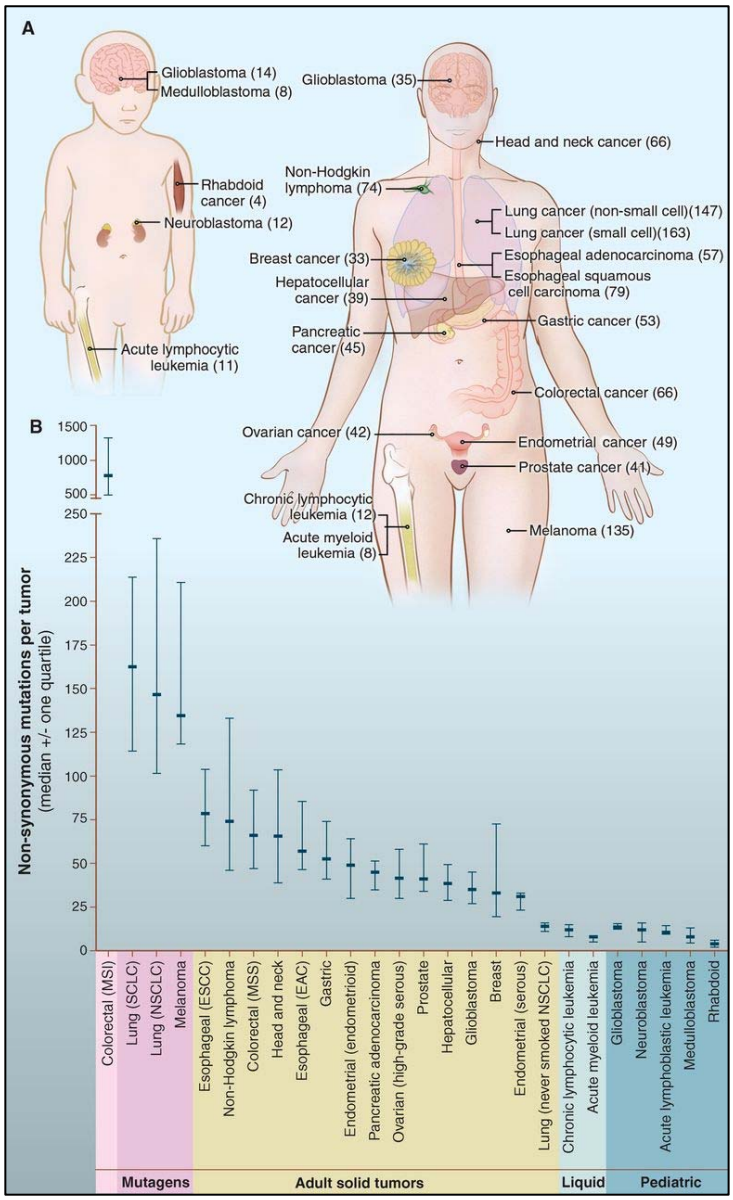


Figura 5 – Mutações pontuais somáticas em tumores humanos. Média de mutações que alteram a sequência de aminoácidos em tumores humanos. MSI = *Microssatellite Instability*; MSS = *Microssatellite Stable*; SCLC = *Small Cell Lung Cancer*; NSCLC = *Non-Small Cell Lung Cancer*; ESCC = *Esophageal Squamous Cell Carcinomas*; EAC = *Esophageal Adenocarcinomas* (Figura: Vogelstein et al., 2013)

Um dos maiores projetos atuais para identificar as alterações moleculares em tumores através de sequenciamento em larga escala é o TCGA (*The Cancer Genome Atlas*). Em 2012, o projeto publicou os resultados de uma análise molecular compreensiva em tumores de cólon e reto utilizando amostras de 276 pacientes (TCGA, 2012). A análise revela que 16% dos tumores apresentam o fenótipo de hipermutação, e

também identifica genes frequentemente mutados que podem servir de alvo para novas terapias. O conhecimento completo dos genes alterados nos diversos tipos de câncer é fundamental para estabelecer diagnósticos, terapias e seleção de combinações de drogas (Lawrence *et al.*, 2014). Acredita-se que a identificação dessas alterações permitirá um melhor monitoramento da doença e a escolha de um tratamento mais eficiente e adequado para o paciente em questão (Lawrence *et al.*, 2014). As descobertas em diversos estudos já proporcionaram avanços que podem ser aplicados à medicina clínica atual, como por exemplo a manipulação da via de VEGF em tumores colorretais e anticorpos contra o receptor EGFR (Epidermal Growth Factor Receptor) que podem ser eficientes em tumores sem mutações em KRAS (Troiani *et al.*, 2013).

O conhecimento das alterações genômicas tumorais tem portanto um imenso potencial para o estabelecimento de novas terapias, que serão abordadas a seguir. As variações estruturais somáticas podem servir como biomarcadores altamente sensíveis de detecção tumoral, e as mutações pontuais revelam os melhores alvos terapêuticos para cada tumor.

1.4 Variações estruturais no genoma como biomarcadores

Em tumores hematológicos, rearranjos estruturais tem sido utilizados historicamente como marcadores tumorais para detectar doença residual mínima (Flohr *et al.*, 2008) ou monitorar a resposta ao tratamento, por exemplo a detecção da fusão BCR-ABL (também conhecida como cromossomo Filadélfia), que indica a presença de leucemia mielóide crônica e a resistência tumoral ao tratamento com imatinib (Branford, 2007).

Para tumores sólidos, este monitoramento pode ser estabelecido com a análise de fluídos corpóreos, uma vez que os tumores liberam fragmentos de DNA na circulação. No sangue de um indivíduo estão contidas milhões de cópias do genoma, fragmentadas em pequenas sequências (~140pb-170pb). Em pacientes com câncer, uma pequena fração destas (cerca de milhares de cópias por mL de sangue) pode corresponder ao DNA tumoral circulante (ctDNA) (Forshew *et al.*, 2012). O DNA circulante de células tumorais possui uma fragmentação mais acentuada do que os derivados de células normais (Mouliere *et al.*, 2011; Thierry *et al.*, 2010). A presença do ctDNA está diretamente relacionada à "carga tumoral" e pode ser usada para acompanhar a dinâmica tumoral de pacientes com tumores colorretais em resposta a tratamento cirúrgico ou por quimioterapia (Diehl *et al.*, 2008), através da detecção de mutações tumor específicas em oncogenes (Diehl *et al.*, 2005; Holdhoff *et al.*, 2009). Devido também à dificuldade de obtenção de tecido tumoral pós-tratamento, estudos recentes propõe a utilização do ctDNA para analisar a resposta tumoral a diversos medicamentos, uma forma de "biópsia liquida" (Diaz Jr. *et al.*, 2012; Misale *et al.*, 2012), assim como uma possível antecipação do diagnóstico tumoral (Leary *et al.*, 2012). Os resultados mostram que o ctDNA permite a detecção da presença tumoral pós tratamento e muito antes do que possível com exames radiológicos, portanto possibilitando antecipar o início de uma segunda abordagem de tratamento (Diaz Jr. *et al.*, 2012; Misale *et al.*, 2012).

Ao contrário da detecção de uma fusão recorrente como nos tumores hematológicos, as abordagens de detecção de biomarcadores em tumores sólidos devem ser personalizadas, já que na maioria dos tumores sólidos, como os tumores colorretais, os rearranjos encontrados não são recorrentes (Bass *et al.*, 2011), mas sim tumor-

específicos, tornando mais difícil sua implementação na clínica. Atualmente foram desenvolvidas abordagens para identificar biomarcadores específicos para cada paciente, utilizando sequenciamento de segunda geração (Leary *et al.*, 2010). Construindo bibliotecas "*mate-pair*" de DNA genômico e submetendo as mesmas ao sequenciamento pela plataforma SOLiD (Applied Biosystems), Leary e colaboradores desenvolveram a abordagem PARE (Personalized Analysis of Rearranged Ends), ilustrada na Figura 6, identificando com sucesso alterações de número de cópias e rearranjos cromossômicos específicos do paciente em tumores de cólon e mama. Os rearranjos podem ser confirmados posteriormente por PCR utilizando *primers* compatíveis com as extremidades das regiões de rearranjo, sendo que a amplificação do DNA só é possível em tumores contendo este rearranjo específico, pois no tecido normal os segmentos de DNA correspondentes aos *primers* estão em regiões distantes do genoma. Esta amplificação por PCR é muito mais simples do que a necessária para detectar alterações de uma única base, e permite ainda que estes segmentos de DNA sejam identificados em concentração muito pequena, tornando-a uma excelente ferramenta para identificar o DNA tumoral circulante no sangue de pacientes. A amplificação pode ser suficientemente sensível para indicar se o paciente apresenta ou não tecido tumoral após um tratamento ou cirurgia. Os rearranjos estruturais personalizados são portanto marcadores ideais para identificar a resposta tumoral a um tratamento, detectar a presença de doença residual após cirurgia e monitoramento clínico a longo prazo, incluindo a detecção de metástases.

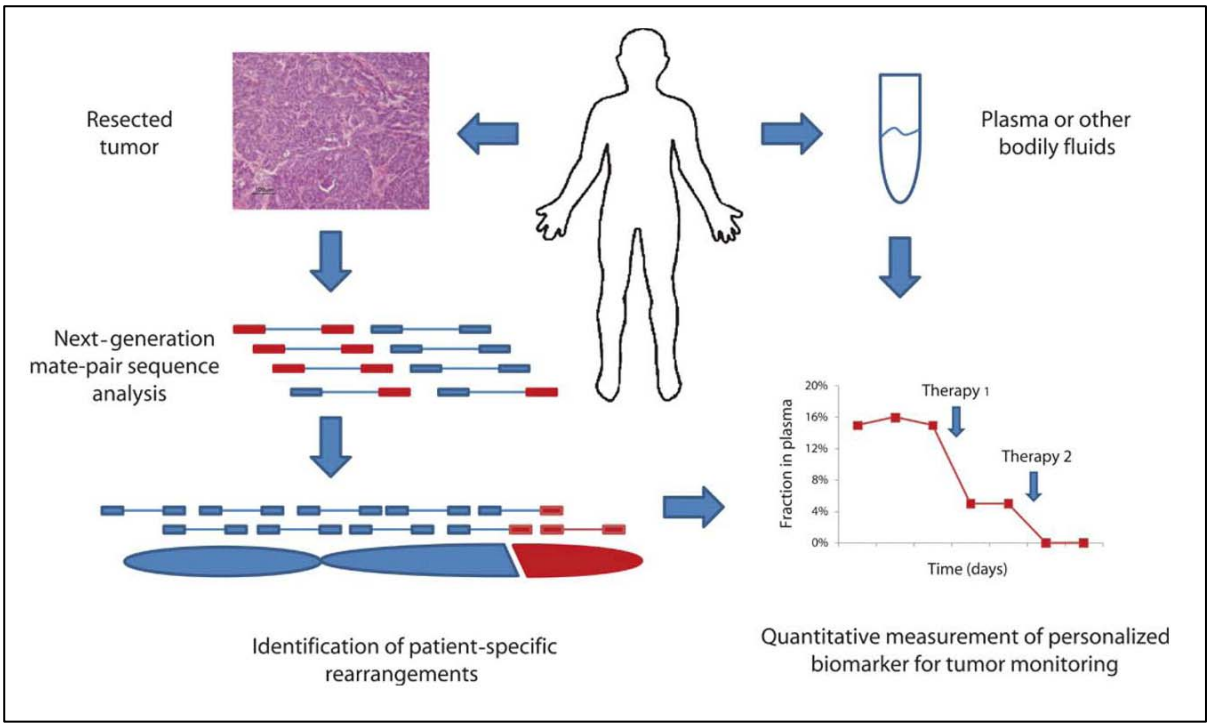


Figura 6 – Abordagem PARE (Figura: Leary et al., 2010)

Diversos estudos confirmam o potencial de alterações genéticas tumorais como biomarcadores. He e colaboradores procuraram rearranjos em genes de imunoglobulinas (IgH) que pudessem servir como biomarcadores para acompanhar o curso de tratamento de pacientes com linfomas do tipo não-Hodgkin (He *et al.*, 2011). A abordagem consistiu na captura direcionada de sequências referentes aos genes IgH seguida por sequenciamento em larga escala. Neste trabalho, os autores mostraram que foi possível identificar rearranjos específicos das células tumorais para quase todos os pacientes do estudo por este método e todos os rearranjos identificados foram amplificados com sucesso por PCR a partir do plasma destes pacientes. No ano passado, um grupo da Universidade de Cambridge descreveu uma comparação entre os diversos diagnósticos existentes para avaliar a carga tumoral em pacientes com câncer de mama metastático (Dawson *et al.*, 2013), incluindo imagem radiográfica, detecção do antígeno

CA 15-3, detecção de células tumorais circulantes e detecção de DNA tumoral circulante. A detecção do DNA tumoral circulante teve sucesso em 97% das pacientes nas quais alterações genômicas somáticas foram identificadas. Para o CA 15-3 e células tumorais circulantes a detecção foi menos eficiente (78% e 87% respectivamente). O DNA tumoral circulante possibilitou também a detecção mais inicial da resposta ao tratamento. A comparação mostrou que o DNA tumoral circulante é um biomarcador específico, informativo e altamente sensível para o câncer de mama metastático (Dawson *et al.*, 2013). O estudo mais recente de Bettegowda e colaboradores mostra que a análise do sangue dos pacientes revela alterações genéticas tumorais até mesmo para tumores em estádios iniciais, sem a necessidade de um procedimento invasivo para coletar tecido tumoral. No estudo os pesquisadores conseguiram identificar a presença de DNA tumoral circulante em 50% dos pacientes de 14 tipos tumorais diferentes em estádios iniciais, confirmando o potencial da abordagem para o monitoramento tumoral (Bettegowda *et al.*, 2014).

1.5 Mutações pontuais como potenciais novos alvos terapêuticos

A instabilidade genômica que resulta no grande número de mutações em tumores é uma característica que confere às células a capacidade de adquirir os principais atributos de um câncer (*Hallmarks of Cancer*), através da expansão seletiva de populações clonais com alterações vantajosas (Hanahan e Weinberg, 2011). Nas células normais, diversos mecanismos de reparo são extremamente eficientes em manter as taxas de mutação controladas durante cada geração. Para escapar deste controle, as células tumorais frequentemente sofrem uma seleção para aumentar suas taxas de mutação após adquirirem alterações que afetam os principais genes que monitoram a

integridade genômica, como TP53 (Negrini, Gorgoulis e Halazonetis, 2010; Salk, Fox e Loeb, 2010).

Estudos sistemáticos com tumores tem catalogado esta grande quantidade de genes alterados, os quais participam de uma variedade de processos celulares como sinalização, regulação epigenética, processamento de RNAs, entre outros. Estima-se que o catálogo destas mutações está longe de ser completado, e o conhecimento destas alterações e vias regulatórias será capaz de iluminar as vulnerabilidades celulares, guiando o desenvolvimento e aplicação de terapias (Garraway e Lander, 2013). A abordagem de sequenciamento em larga escala para detectar alterações genéticas que indiquem o melhor tratamento para os pacientes já foi explorada por vários grupos. Um estudo da Universidade de Michigan desenvolveu um protocolo abrangente de sequenciamento de genoma completo, sequenciamento de exoma e sequenciamento de transcriptoma (RNA-Seq) e aplicou o mesmo a dois pacientes, o primeiro com câncer colorretal e o segundo com melanoma (Roychowdhury *et al.*, 2011). O estudo piloto obteve informações de rearranjos estruturais, alterações de número de cópia (CNAs), mutações pontuais, inserções, deleções e expressão gênica para cada um dos pacientes. O quadro completo de alterações genéticas de cada paciente foi analisado por um conselho multidisciplinar que buscou definir para cada paciente qual seria a melhor estratégia de tratamento. Algumas terapias baseadas em alterações conhecidas em vias regulatórias nos tumores já estão em uso na clínica, como a droga trastuzumab para o tratamento de tumores de mama com superexpressão de ERBB2 ou também inibidores de BRAF (vemurafenib, por exemplo) para pacientes com melanoma que apresentam a mutação V600 no gene BRAF.

Existem genes com propriedades estruturais em suas proteínas que proporcionam melhor interação com compostos químicos, este conjunto é chamado de genoma drogável (*druggable*) e estima-se que fazem parte dele cerca de 3.000 genes humanos (Hopkins e Groom, 2002). Não só as características estruturais das proteínas são levadas em consideração para serem incluídas nesse grupo de alvos terapêuticos potenciais, mas também suas características funcionais, localização celular e se as modulações destas funções podem trazer benefícios. Dentre estas categorias funcionais com maior potencial encontram-se transportadores, canais iônicos, proteases, quinases e receptores acoplados à proteína G (Russ e Lampel, 2005).

Um estudo prévio do nosso grupo de pesquisa identificou o conjunto de genes codificadores das proteínas de superfície celular, chamado de surfaceoma (Cunha, da *et al.*, 2009). O conjunto foi definido pela identificação de domínios transmembrana na sequência de todos os genes conhecidos, e posteriormente foram eliminados os genes correspondentes a proteínas secretadas e localizadas em outros compartimentos celulares. No conjunto final do surfaceoma são encontradas proteínas pertencentes à maioria das categorias mencionadas anteriormente. Além disso, a localização das proteínas na superfície celular é ideal para intervenções terapêuticas. Este conjunto de genes pode portanto ser uma fonte interessante de potenciais novos alvos para o tratamento do câncer colorretal.

Outra consequência importante das alterações de sequência presentes no DNA tumoral é o surgimento de novos epítopos capazes de desencadear uma resposta imune contra o próprio tumor (Castle *et al.*, 2012). A degradação natural das proteínas celulares pelo proteassomo gera peptídeos curtos que dependendo de sua sequência de aminoácidos podem se ligar com maior afinidade à molécula de MHC de classe I, mas

poucos peptídeos se ligam de maneira eficiente e são apresentados na superfície da célula como epítopos para serem reconhecidos pelos linfócitos T e desencadear uma resposta contra a célula apresentadora (Yewdell e Bennink, 1999). Normalmente, as proteínas celulares não geram epítopos antigênicos, devido ao processo de tolerância central durante o desenvolvimento dos linfócitos. As mutações presentes em um câncer podem gerar epítopos novos imunodominantes, que não são reconhecidos como próprios, desencadeando uma resposta imune espontânea. A ativação de linfócitos contra células tumorais já foi observada em pacientes com melanoma, por exemplo, nos quais a resposta imune era direcionada contra epítopos resultantes das mutações tumorais (Lennerz *et al.*, 2005).

Atualmente, este cenário de antígenos tumorais vem gerando interesse constante e o potencial imunogênico destes epítopos mutantes já foi avaliado em modelos animais (Castle *et al.*, 2012) e nos conjuntos de alterações somáticas encontrados em outros estudos de tumores humanos (Khalili, Hanson e Szallasi, 2012; Warren e Holt, 2010). A predição de epítopos derivados de mutações em tumores de cólon e mama revelou uma média de 7 ou 10 novos epítopos com potencial imunogênico por amostra, respectivamente (Segal *et al.*, 2008). A existência destes epítopos abre caminho para abordagens terapêuticas baseadas em vacinas mais específicas derivadas dos antígenos do próprio paciente identificados, que funcionariam inclusive para a eliminação de metástases. Abordagens de modulação do sistema imune também estão sendo exploradas como alternativas viáveis como resultado da presença destas mutações imunogênicas (Khalili, Hanson e Szallasi, 2012). Acredita-se que a aplicação deste tipo de terapia na prática clínica pode beneficiar o tratamento de pacientes com tumores

colorretais, dentre outros. Portanto, estabelecer protocolos otimizados para a detecção destes epítopos é essencial para o desenvolvimento da técnica.

2 OBJETIVOS

- Desenvolver e implementar um protocolo eficiente para detecção de variações estruturais no genoma tumoral através do sequenciamento em larga escala de tumores (de reto), sem a necessidade do sequenciamento paralelo do genoma normal de cada paciente, visando a aplicação destas variações na clínica como biomarcadores personalizados capazes de monitorar a resposta ao tratamento e a presença de doença residual.

- Identificar mutações pontuais somáticas no surfaceoma de linhagens celulares de tumores colorretais, para a identificação de potenciais novos alvos terapêuticos.

- Detectar potenciais novos epítopos tumorais derivados da grande quantidade de alterações na sequência do DNA das linhagens de tumores colorretais, os quais podem desencadear uma resposta imune e serem aplicados como vacinas anti-tumorais.

3 MATERIAIS E METODOS

3.1 Escolha de amostras de tumores retais

Para este trabalho, foram escolhidas 2 amostras de pacientes com tumores retais que foram diagnosticados clínica e patologicamente com resposta incompleta ao tratamento neoadjuvante, 1 amostra de um paciente que foi inicialmente diagnosticado como resposta incompleta e um diagnóstico patológico posterior revelou que não havia mais tecido tumoral (resposta completa patológica) e 3 amostras de pacientes cuja resposta ao tratamento neoadjuvante foi considerada completa após a avaliação clínica. Para a inclusão do paciente no estudo, o resultado final do tratamento já era conhecido e, portanto, todo o período de tratamento estava finalizado. As amostras escolhidas são provenientes de pacientes que concordaram com o termo de consentimento de doação para o banco de tumores do Hospital Alemão Oswaldo Cruz (HAOC) e este projeto foi aprovado pelo comitê de ética da mesma Instituição.

Tabela 1 - Pacientes e estadiamento. Resumo dos pacientes incluídos no estudo. *cT e cN se referem ao estadiamento inicial do tumor e indicam o grau de extensão (cT) e envolvimento de linfonodos (cN) (Fischer *et al.*, 2010). **ypTNM indica que a classificação patológica (p) foi determinada após a terapia pré-operatória (Brierley *et al.*, 2006). *** O paciente P2 apresentou metástase no fígado (semana 32). O paciente P6 apresentou uma recorrência local e foi então diagnosticado como ypT2N0 (semana 51). cCR = *Clinical Complete Response*.

PACIENTE	sexo	cT*	cN*	ypTNM**	Localização	Resposta
P1	F	3	1	ypT3N0	Reto Baixo	Incompleta
P2	M	3	1	ypT2N0	Mesorreto	Incompleta***
P3	F	3	1	ypT0N0	Reto Baixo	Completa Patológica
P4	M	3	1	cCR	Reto Baixo	Clínica Completa
P5	M	3	0	cCR	Reto Baixo	Clínica Completa
P6	F	3	1	cCR	Reto Baixo	Clínica Completa***

3.2 Construção de bibliotecas mate-pair

As amostras escolhidas foram submetidas à extração de DNA genômico por um protocolo baseado em Trizol para extração simultânea de DNA e RNA (Chevillard, 1993). A qualidade do DNA obtido foi analisada para garantir um bom resultado na etapa posterior de sequenciamento. As bibliotecas *mate-pair* foram geradas com o *SOLiD™ Long Mate Paired Library Construction Kit* (Life Technologies). Este tipo de biblioteca de DNA permite a obtenção de pares de sequências com 50pb cada (SOLiD 3 e 4; 60pb para SOLiD 5500) e a distância entre as duas no genoma sequenciado é definida de acordo com o interesse do projeto. Resumidamente, foi realizada uma fragmentação aleatória do DNA tumoral (após amplificação completa do genoma) seguida pela ligação de adaptadores às extremidades dos fragmentos. Os fragmentos de DNA foram submetidos a um gel de agarose *Low Melting* 0.8% e aqueles de tamanho entre 600 a 1000 pares de base foram cortados e purificados do gel. Este tamanho dos fragmentos será exatamente a faixa de tamanho esperada dos insertos contidos entre cada par de sequências *mate-paired*. O DNA purificado é tratado, circularizado e digerido com diferentes enzimas de restrição e modificação de acordo com o protocolo do fabricante (ilustrado na Figura 7). Como última etapa, a biblioteca foi amplificada através da PCR convencional e purificadas utilizando *beads Agencourt AMPure XP* (Beckman Coulter) e eluídas em 25 µL de TE, e quantificadas através do fluorômetro Qubit 2.0 (Invitrogen), utilizando o ensaio *dsDNA High Sensitivity*.

Ao final do protocolo, foi obtida uma biblioteca de DNA para cada amostra na qual somente as extremidades (~50pb) dos fragmentos isolados inicialmente estão ligadas a adaptadores específicos e necessários para o PCR em emulsão e sequenciamento no SOLiD.

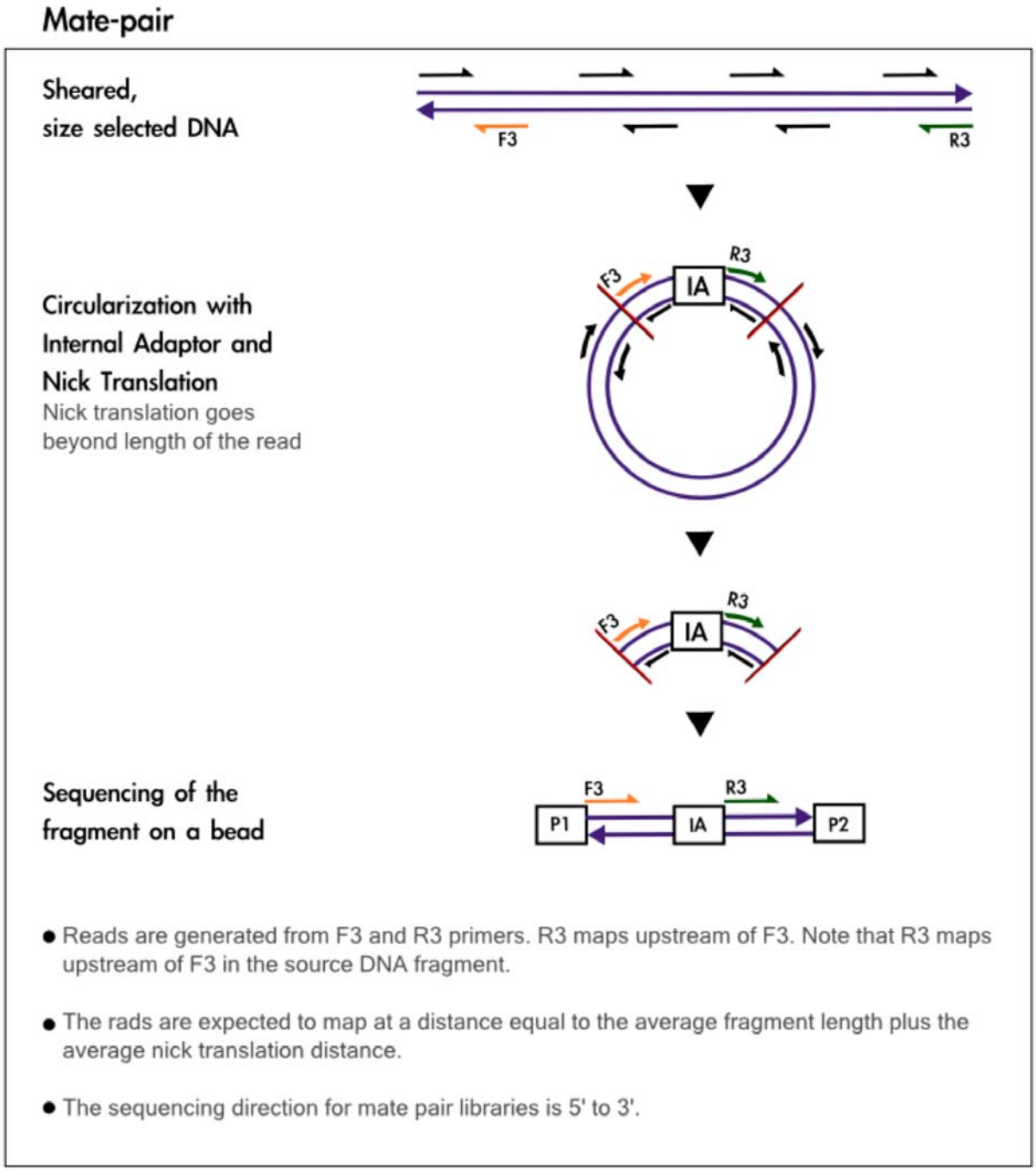


Figura 7 – Construção da biblioteca mate-pair. A figura ilustra as etapas da construção de uma biblioteca *mate-pair* para sequenciamento. A informação de que a posição genômica da sequência R3 é localizada a montante da sequência F3 é de extrema importância para a identificação do rearranjo e para o desenho de *primers* na validação (Figura: Life Technologies).

A etapa 3.2 descrita acima foi executada pela aluna de doutorado Paola A. Carpinetti, participante deste projeto e orientanda da Dra. Anamaria A. Camargo.

3.3 Sequenciamento em larga escala

Os sequenciamentos foram feitos através da plataforma SOLiD da Applied Biosystems (hoje Life Technology) disponível nas dependências do Instituto Ludwig de Pesquisa sobre o Câncer. Duas técnicas de nível superior com treinamento específico são responsáveis por todos os sequenciamentos nesta plataforma, incluindo os deste projeto.

A técnica de sequenciamento da plataforma SOLiD 4 é baseada no isolamento de moléculas únicas de DNA através de *beads* que são recobertos com sequências específicas e complementares aos respectivos adaptadores utilizados na construção das bibliotecas de DNA que serão sequenciadas. A ligação de uma única molécula de DNA a um *bead* é o ponto crucial no protocolo. Para tanto, após serem homogeneizados em fase aquosa, a biblioteca e os *beads* são adicionados a uma fase oleosa e posteriormente submetidos à agitação constante. Ao final, é obtida uma emulsão com microbolhas (microrreatores) cujo objetivo é que cada uma contenha um *bead* e uma molécula molde. Além disso, nos microrreatores também estão contidos os reagentes necessários para uma reação de PCR convencional, os quais foram adicionados durante a preparação da fase aquosa. A emulsão então é submetida à amplificação em um termociclador convencional, de maneira que em cada microrreator ocorre a amplificação clonal de uma única molécula. Após a amplificação, os *beads* recobertos por milhares de cópias de uma mesma molécula são recuperados, enriquecidos e submetidos ao sequenciamento.

A técnica de sequenciamento da plataforma SOLiD é baseada na ligação sucessiva de oligos contendo 8 nucleotídeos ao DNA de interesse. Os oligos são fabricados com todas as combinações possíveis de nucleotídeos nas 5 primeiras posições (4^5 = 1024 oligos diferentes). As últimas três bases são nucleotídeos modificados que fazem o pareamento

com qualquer outro nucleotídeo presente na fita molde. Além disso, os oligos são marcados com diferentes corantes. A reação inicia-se com a utilização de um *primer* específico que se anela à fita molde e posteriormente apenas um oligo liga-se ao mesmo. O sinal do corante presente em tal oligo é captado e posteriormente as 3 bases marcadas da porção 5' do oligo são removidas enzimaticamente, restando assim apenas os 5 nucleotídeos que apresentam complementaridade com a fita molde. O processo se repete com novas ligações até um número de máximo de 10 ligações por iniciador. Uma vez que o corante utilizado na marcação dos oligos é definido apenas pelos 2 primeiros nucleotídeos de cada oligo, ao final deste primeiro ciclo de ligações, o que se obtém são informações apenas dos nucleotídeos: 1 e 2; 6 e 7; 11 e 12; 16 e 17; e assim sucessivamente até os nucleotídeos 46 e 47. Para a obtenção da sequência de todos os nucleotídeos, múltiplos ciclos de ligação, detecção e clivagem são feitos, utilizando-se iniciadores que diferem na posição em que se anelam ao adaptador utilizado na construção da biblioteca que está sendo sequenciada. No total são utilizados 5 iniciadores diferentes com os quais é possível obter as informações sobre todos os nucleotídeos da fita molde. O número de ciclos é correspondente ao tamanho de sequência gerada. As sequências são geradas em formato *color-space* (0,1,2,3), que indica a base de um nucleotídeo com referência ao nucleotídeo seguinte.

3.4 Alinhamento das sequências geradas

As sequências geradas foram alinhadas ao genoma referência utilizando o algoritmo *mapreads* do pacote BioScope (Applied Biosystems) que alinha sequências curtas em formato *color-space* (como as geradas pela plataforma SOLiD) ao genoma humano referência (versão GRCh37/hg19). Foram mantidos os parâmetros padrão do programa

com as seguintes modificações: mapping.mismatch.penalty = -2.0; mapping.qual.filter.cutoff = 0; clear.zone = 5. Após o mapeamento em *color-space* é feita uma conversão para o formato *base-space* (A, C, T, G) e os resultados finais são armazenados no formato BAM (equivalente binário do formato *Sequence Alignment/Map*) (Li *et al.*, 2009).

3.5 Análise do alinhamento e identificação de rearranjos intercromossomais

Depois do alinhamento de todas as sequências contra o genoma humano, foi feito processamento e análise em larga escala dos alinhamentos de cada amostra. O primeiro passo foi a remoção das sequências sem ao menos um alinhamento confiável. Para isto, todas as sequências sem um alinhamento de qualidade maior ou igual a 20 são filtradas. Essa qualidade é dada a cada alinhamento pelo *mapreads*/BioScope e utiliza como parâmetro o número de vezes que a sequência se alinha contra a sequência de referência (genoma humano, neste caso) e a sua composição de nucleotídeos (ACTG). Esta qualidade segue a escala Phred (Ewing *et al.*, 1998), o que significa que uma qualidade maior ou igual a 20 dá ao menos 99% de chance do alinhamento estar correto. Isto garante, inclusive, que o alinhamento foi único e evita que um falso alinhamento (por exemplo devido a regiões repetitivas do genoma) seja considerado nas análises posteriores. No final deste passo restam apenas as sequências alinhadas com confiabilidade.

As sequências foram então pareadas através de algoritmos locais e o próximo passo é a remoção dos pares de sequências alinhados dentro da distância e orientação esperadas (Figura 8a). Como estas sequências representam as regiões genômicas sem

nenhuma grande variação, elas são excluídas. Também não são analisadas neste momento as sequências alinhadas no mesmo cromossomo mas fora da distância e/ou orientação esperada (Figura 8b e 8c). Estas sequências podem indicar regiões genômicas relacionadas a deleções, inserções ou inversões. Posteriormente, as regiões que indicam uma deleção no genoma tumoral (Figura 8b) foram selecionadas e analisadas para selecionar biomarcadores. Esta etapa será descrita na seção 'Análise de deleções'.

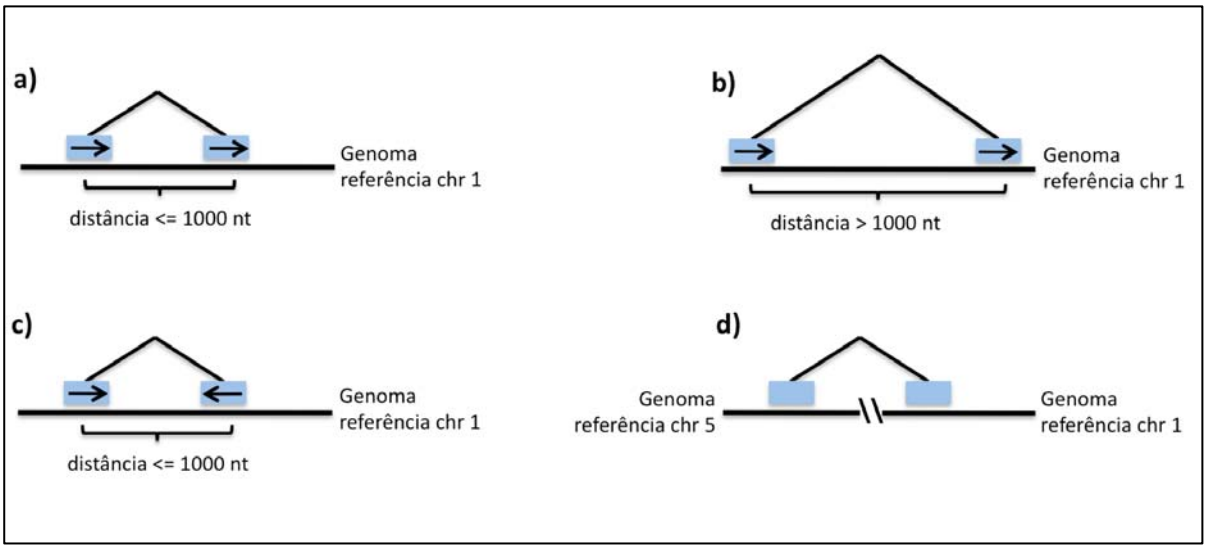


Figura 8 - Padrões de alinhamentos, contra o genoma de referência, encontrados nos resultados do sequenciamento.

Selecionando portanto somente os pares de sequências onde os membros estão alinhados em cromossomos diferentes (Figura 8d), foi possível analisar a existência de rearranjos intercromossomais. Para tanto, algoritmos locais foram desenvolvidos em Perl (www.perl.org) e os arquivos de pares de sequências foram processados.

Uma segunda etapa de mapeamento foi implementada, utilizando como referência três montagens alternativas do genoma humano. São estas i) HuRef: Sequência diploide completa de um indivíduo (J. Craig Venter Institute) (Levy *et al.*, 2007); ii) GRC37_alt:

Contém porções do genoma com representações alternativas, não é uma referência de toda a sequência genômica (Genome Reference Consortium); iii) CRA: Sequência completa do cromossomo 7 humano (The Center for Applied Genomics) (Scherer *et al.*, 2003). A necessidade deste mapeamento alternativo foi observada após uma análise manual de alguns pares de sequências com o programa BLAST (Altschul *et al.*, 1990), que mostrou alguns casos nos quais os as sequências de um mesmo par alinhavam no mesmo cromossomo de acordo com outra montagem do genoma mas não segundo a referência (hg19). Todos os casos que apresentaram um alinhamento no mesmo cromossomo foram portanto removidos antes das próximas análises.

Sequências distintas com mesma posição de alinhamento foram removidas por serem, em sua maioria, resultado de amplificação por PCR do mesmo fragmento genômico geradas durante a construção da biblioteca. Com este filtro evita-se que estas duplicações proporcionem suporte falso para uma variação estrutural (Handsaker *et al.*, 2011; Kozarewa *et al.*, 2009; Nielsen *et al.*, 2011). A posição do alinhamento das sequencias restantes foi avaliada para identificar sobreposições a regiões de certos elementos repetitivos definidos pelo RepeatMasker (Smit Hubley, R & Green, P, 2010) (LINEs, SINEs, *low complexity*, satélites e rRNA) ou regiões correspondentes a centrômeros (+1M) e telômeros (+1M), e estas foram removidas da análise para minimizar a interferência de alinhamentos incorretos na identificação de rearranjos. Sequências mapeadas em regiões de duplicações de segmento (Bailey *et al.*, 2001) também foram removidas para evitar rearranjos falsos devido à similaridade entre a região parental e a região duplicada. Por fim, foram removidas as sequências mapeadas no genoma mitocondrial pois foi observado que a presença de cópias mitocondriais

(*numts*) no genoma nuclear (Hazkani-Covo, Zeller e Martin, 2010) geram alinhamentos incorretos e levam à identificação de falso-positivos.

Após os filtros descritos de regiões repetitivas as sequências foram utilizadas para gerar agrupamentos (*clusters*) com pares cujas sequências estavam alinhadas nos mesmos dois cromossomos (exemplo: par1 - sequência 1a alinhada no cromossomo 1 e sequência 1b alinhada no cromossomo 5; par2 - sequência 2a alinhada no cromossomo 5 e sequência 2b alinhada no cromossomo 1) e ao mesmo tempo avaliando se as sequências de dois pares alinhadas no mesmo cromossomo (exemplo: sequência 1a e sequência 2b) estavam em posições distantes não mais do que 1000 pares de bases (considerando o tamanho médio de inserto somado a dois desvios-padrão). Como não é possível utilizar a informação da distância entre as sequências de um mesmo par no caso de alinhamentos em cromossomos diferentes, a análise de distância esperada deve ser feita desta maneira. No genoma tumoral, um rearranjo significa que as regiões de dois cromossomos diferentes foram fusionadas e a construção das bibliotecas de *mate-pair* leva à geração de sequências dentro da distância esperada para esta região, que após o alinhamento perdem esta informação por pertencerem a cromossomos diferentes no genoma referência. Considerando o tamanho médio dos insertos, as sequências de um *cluster* mapeadas em um mesmo cromossomo devem obedecer esta disposição para que o rearranjo seja verdadeiro. Os *clusters* formados foram então avaliados quanto ao número de sequências reportando o rearranjo. O corte utilizado foi de no mínimo de 3 e máximo de 80 pares inicialmente, reduzindo o impacto de alinhamentos incorretos e outros artefatos. Estas duas etapas (clusterização e limites de suporte) reduzem substancialmente o número de reads indicando rearranjos e

simplificam muito as últimas etapas do *pipeline*, uma consideração importante quando a análise envolve dados em larga escala.

Por último, as sequências restantes foram submetidas a outro alinhamento, usando o programa BLAT (Kent, 2002) e a referência hg19. Como este programa utiliza uma heurística diferente para mapeamento e também já parte do resultado encontrado pelo *mapreads* do BioScope (sequências em formato FASTA), foi possível obter para alguns casos uma posição de mapeamento diferente da encontrada originalmente. Neste caso, se as duas sequências de um mesmo *mate-pair* são mapeadas em um mesmo cromossomo (independente da qualidade deste alinhamento), as mesmas são então removidas do conjunto final. Os parâmetros utilizados para o alinhamento BLAT foram estes: -stepSize=5 -repMatch=2253 -minScore=24 -minIdentity=80 -noTrimA –fine; sendo os mesmos indicados para reproduzir os resultados da ferramenta na página da internet descritos no *Genome Browser* (Kent *et al.*, 2002; Rhead *et al.*, 2010).

Com os *clusters* restantes é possível avaliar a orientação indicada pelas sequências, que podem se encaixar dentro de quatro casos (a, b, c e d) como ilustra a Figura 9. Nesta etapa um mínimo de dois pares de sequências para cada evento de rearranjo é exigido, uma cobertura já testada por outros grupos de pesquisa (Talkowski *et al.*, 2011). Após este filtro de cobertura os candidatos são encaminhados para a validação experimental por PCR. Os principais passos do sequenciamento e *pipeline* de bioinformática estão esquematizados na Figura 10.

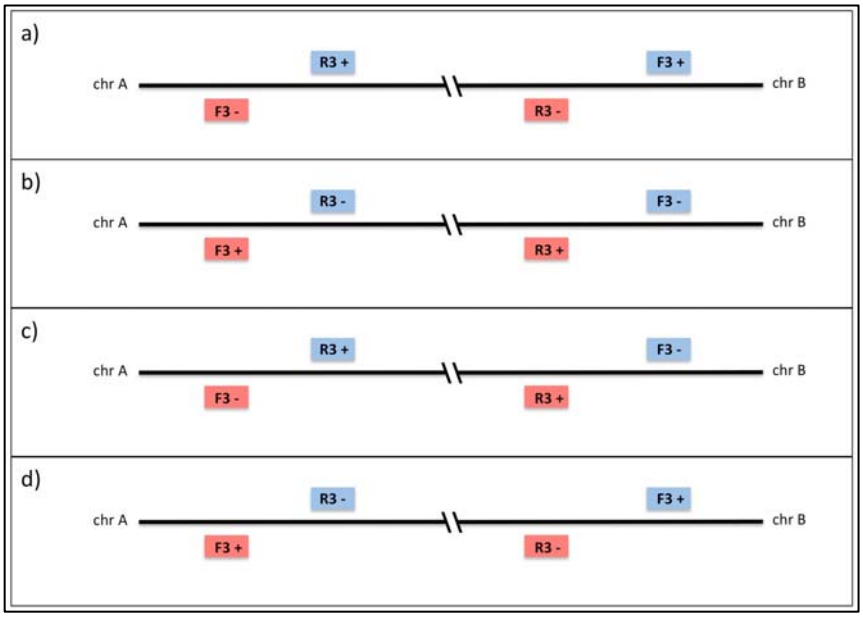


Figura 9 – Padrões de orientação das sequências indicando eventos de rearranjo. As sequências com a mesma cor (azul ou vermelha) fazem parte de um mesmo par, ou seja, são provenientes de um mesmo fragmento genômico. (Figura: Donnard et al., em revisão)

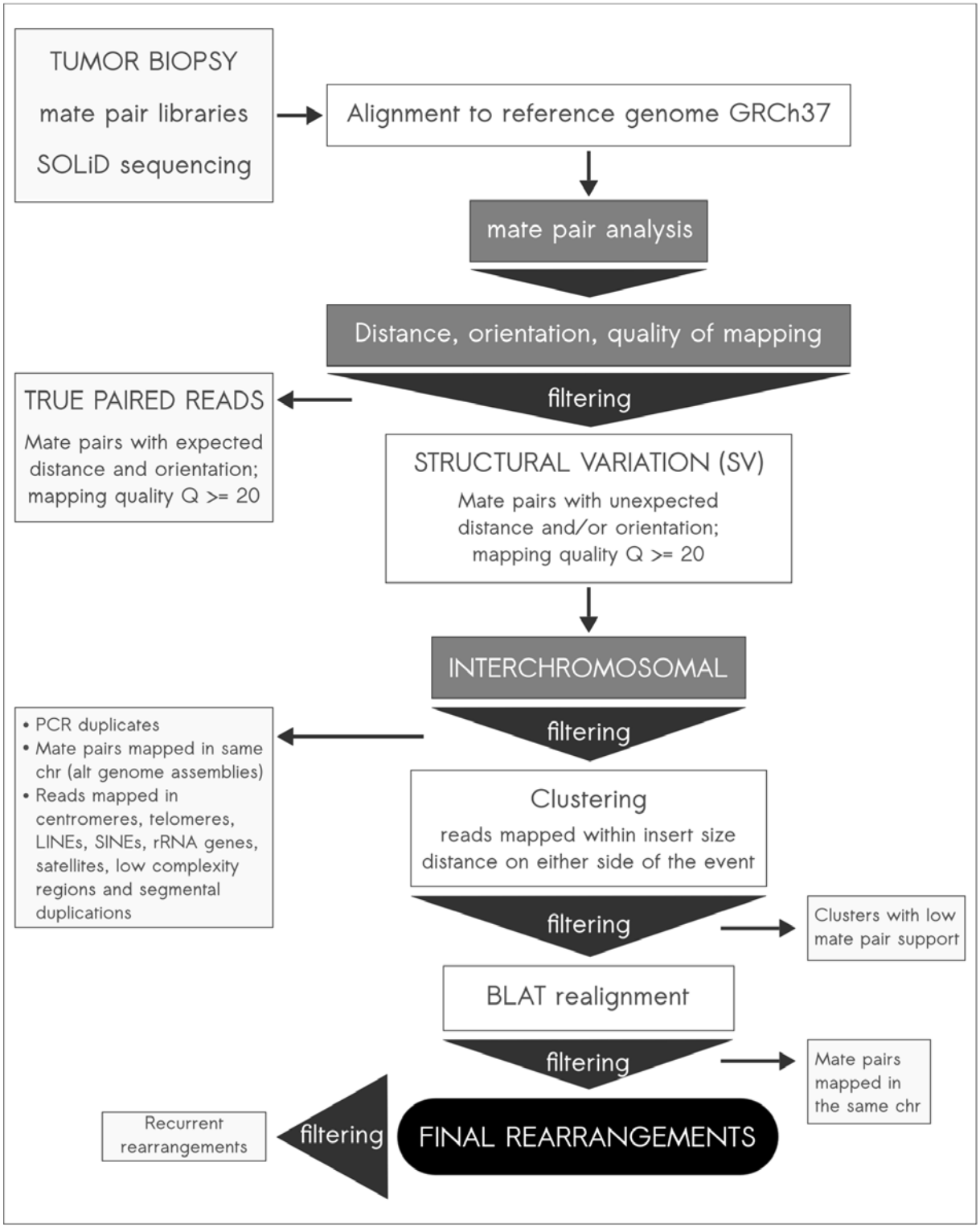


Figura 10 - *Pipeline* de bioinformática. A figura resume as principais etapas do *pipeline* desenvolvido para detectar rearranjos intercromossomais no sequenciamento. (Figura: Donnard et al., em revisão)

3.6 Análise de deleções

Deleções maiores representam outra possível fonte de variações estruturais específicas de células tumorais. Para analisar estes eventos com os dados genômicos foi utilizado o mesmo *pipeline* desenvolvido para buscar os rearranjos intercromossomais, com algumas modificações descritas a seguir. Foram selecionadas somente as sequências pareadas com orientação esperada e um tamanho de inserto maior do que 4kb mais o tamanho médio da biblioteca (padrão de mapeamento representado na Figura 8b) e com boa qualidade de mapeamento (Q>=20). Assim como no *pipeline* de rearranjos, as sequências duplicadas foram removidas, e as sequências mapeadas em cada borda da região deletada a uma distância menor ou igual a 1000pb (indicando a mesma deleção) foram identificadas e agrupadas em *clusters*. Foram incluídos os mesmos filtros de regiões repetitivas nesta análise, e também identificados os casos cujo tamanho da deleção (~6kb) indica que a variação estrutural pode corresponder a uma inserção de um LINE (*Long Interspersed Nuclear Element*) no genoma referência, que acabam mostrando o mesmo padrão de mapeamento típico de deleções. Usando o tamanho específico de todos os LINEs conhecidos no genoma referência atual, os casos identificados como deleções foram analisados e deles foram removidos os que a região genômica e o tamanho do LINE correspondem ao evento de alteração estrutural. O filtro de artefatos recorrentes não foi necessário na análise de deleções.

3.7 Confirmação da presença de variações estruturais por PCR e identificação em amostras de plasma sanguíneo

Amplificações por PCR foram feitas com as amostras tumorais utilizando *primers* desenhados para cada extremidade dos rearranjos e das deleções. Após a amplificação os fragmentos foram sequenciados por Sanger, confirmando (ou não) a presença da variação estrutural no genoma tumoral. A PCR foi feita utilizando como DNA molde o DNA da biópsia tumoral (mesmo utilizado para a construção das bibliotecas para sequenciamento) e também o DNA extraído da porção celular do sangue do mesmo paciente (chamada de *Buffy Coat*), que representa o genoma das células normais. No caso de rearranjos intercromossomais, se o evento testado for germinativo, espera-se amplificação nas duas reações e, portanto, este rearranjo não poderá ser utilizado como biomarcador tumoral por estar presente em todas as células do indivíduo. No caso de amplificação específica na reação com DNA tumoral o rearranjo é considerado somático e é selecionado para a próxima fase de testes de detecção no plasma sanguíneo. Para uma validação completa, faz-se também o sequenciamento Sanger do produto da amplificação pela PCR. Como resultado, espera-se uma sequência composta de regiões dos dois cromossomos do rearranjo e é possível determinar precisamente o ponto de quebra do evento. No caso das deleções, a reação de PCR também irá mostrar amplificação apenas no genoma tumoral, devido à distância entre cada borda do evento, que não permitiria uma amplificação no genoma normal. Com o sequenciamento do fragmento amplificado para confirmação, a sequência resultante quando alinhada ao genoma referência deve se mostrar partida, indicando a ausência de um fragmento deletado do genoma tumoral.

Para possibilitar a utilização dos rearranjos como marcadores de resposta ao tratamento, o plasma sanguíneo dos pacientes cujas amostras tumorais foram sequenciadas podem ser utilizadas em reações de PCR para verificar a presença dos biomarcadores em diferentes momentos do tratamento, possibilitando sua utilização como biomarcador personalizado. Foram coletados 3 mL de plasma em diferentes momentos do tratamento de cada paciente. O DNA circulante foi extraído através do kit *QIAamp MinElute Virus Kit* (QIAGEN). O DNA extraído do plasma sanguíneo foi utilizado como molde para uma PCR digital (Vogelstein e Kinzler, 1999). A técnica de PCR digital é muito utilizada atualmente na detecção de moléculas de DNA tumoral circulante por sua alta sensibilidade (Bettegowda *et al.*, 2014; Dawson *et al.*, 2013). Na PCR digital, o DNA amplificado é marcado com fluorescência e cada microrreator é avaliado após a reação para detectar a presença de amplificação.

As etapas de validação na bancada descritas acima foram executadas pela aluna de doutorado Paola A. Carpinetti.

3.8 Simulação de genomas com rearranjos

Com base na sequência do genoma humano de referência (hg19), três conjuntos (GR1, GR2 e GR3) de rearranjos foram gerados randomicamente utilizando programas escritos em Perl. Resumidamente, foi sorteado um primeiro cromossomo e uma posição genômica dentro dele, seguido por um segundo cromossomo e outra posição de ponto de quebra. Com os pares de cromossomos e pontos de quebra escolhidos, a sequência FASTA de cada par de cromossomos foi unida nas posições definidas, criando um novo cromossomo para cada rearranjo e simulando uma translocação. Este processo foi utilizado para gerar 20 rearranjos para o conjunto GR1, 30 para o conjunto GR2 e 40

para o conjunto GR3. Os cromossomos que não estavam envolvidos em nenhum rearranjo foram mantidos intactos no arquivo final. Com outro programa, estes genomas foram usados para simular pares de sequências *color-space* de 50nt cada com tamanho de inserto de 700nt entre elas. Para simular um sequenciamento real, as sequências geradas contém 1% de erro. O número de sequências geradas foi calculado com base na cobertura de sequência e na cobertura física desejada para cada simulação com relação ao tamanho do genoma humano referência.

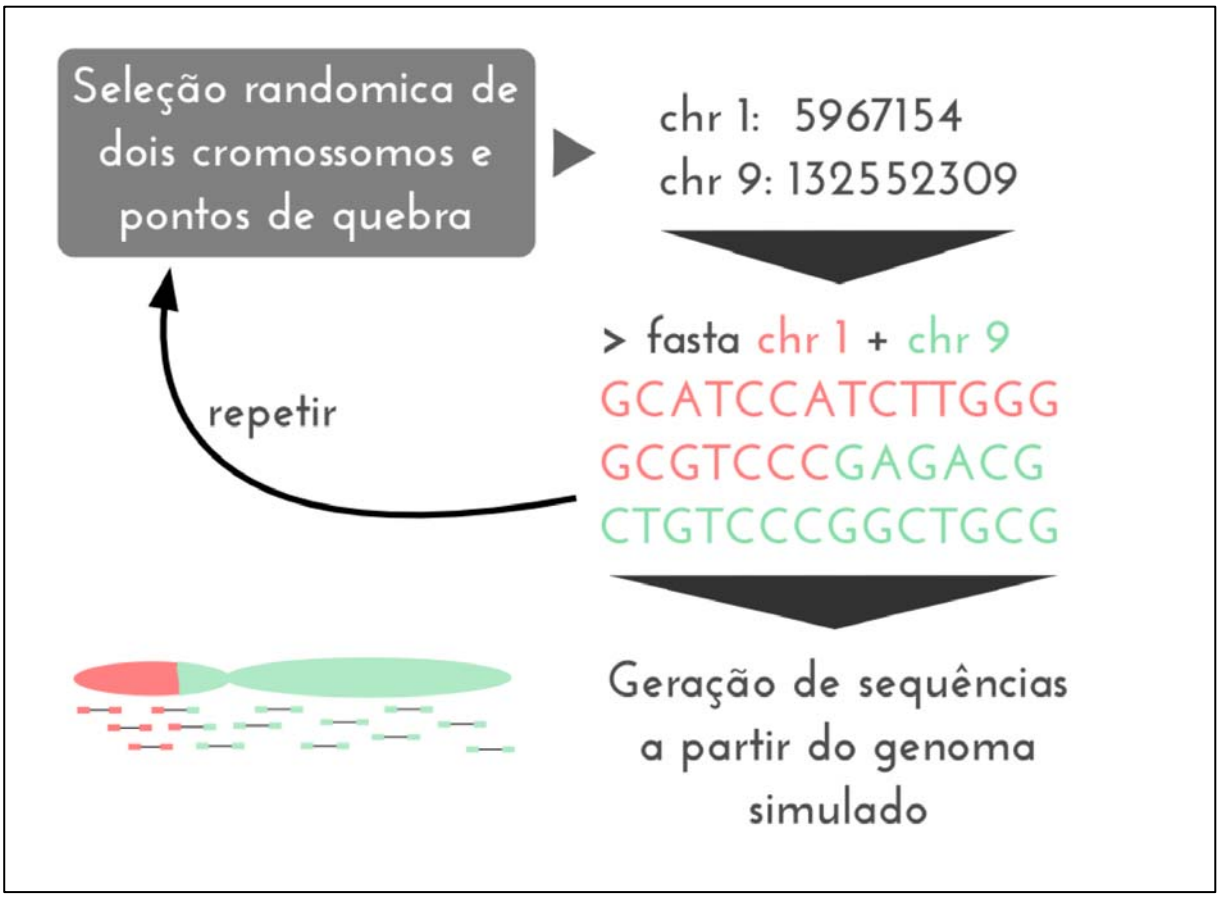


Figura 11 – Simulação de genomas com rearranjos. O esquema ilustra como foi feita a geração de genomas contendo rearranjos entre dois cromossomos e a simulação do sequenciamento destes genomas.

3.9 Aplicação do pipeline em indivíduos do projeto 1000 Genomes

Os dados de sequenciamento para 2.362 indivíduos disponíveis na estrutura do Open Science Data Cloud (www.opensciencedatacloud.org) foram submetidos ao *pipeline* descrito anteriormente para detectar rearranjos intercromossomais. As modificações feitas para o processamento destes dados foram no suporte de sequências indicando o evento (utilizamos um mínimo de 5 pares) e no tamanho de inserto (~300nt) (Abecasis *et al.*, 2010), já que estas bibliotecas foram geradas na plataforma Illumina.

3.10 Linhagens no estudo de alvos terapêuticos

23 linhagens celulares estabelecidas a partir de tumores colorretais foram selecionadas para o estudo de variações de sequência no surfaceoma. São estas: CACO2, COLO205, COLO320, HCT116, HCT15, HT29, LOVO, RKO, SKCO1, SW1116, SW403, SW48, SW480, SW620, SW837, SW948 e T84 (*American Type Culture Collection, Manassas, VA*); LIM1215 e LIM2405 [*Ludwig Institute for Cancer Research* (Devine *et al.*, 1992; Whitehead *et al.*, 1985)]; HCC2998 e KM12 (National Cancer Institute-Frederick Cancer DCT Tumor Repository); e RW2982 e RW7213 (Tibbetts *et al.*, 1988).

3.11 Captura do exoma e sequenciamento

Após a extração do DNA genômico de cada linhagem como descrito previamente na seção 3.2 foram construídas bibliotecas do tipo *paired-end* para cada amostra com fragmentos do DNA de acordo com o protocolo da Life Technologies. As bibliotecas *paired-end* permitem o sequenciamento na plataforma SOLiD de uma sequência de 50 nt (F3 representada na Figura 12). Para a maioria das linhagens só foi feito o

sequenciamento deste fragmento, em alguns casos foi sequenciado também o fragmento F5 como indicado na Figura 12. Cada amostra foi identificada no sequenciamento com o auxílio de uma sequência *barcode* (BC). Após a fragmentação do DNA genômico, foi feita a captura dos fragmentos de DNA correspondentes às regiões codificadoras pelo sistema SureSelect da Agilent para enriquecimento das regiões alvo de acordo com o manual do fabricante. Neste sistema, os fragmentos são hibridizados com sondas de RNA biotiniladas que correspondem às sequências das regiões codificadoras (50Mpb do genoma humano). A solução é então exposta à *beads* magnéticos cobertos com moléculas de streptavidina, que capturam apenas os fragmentos hibridizados com as sondas. As moléculas de RNA são então digeridas e os fragmentos da biblioteca correspondentes à região alvo são sequenciados. O sequenciamento foi feito na plataforma SOLiD, descrito previamente na seção 3.3.

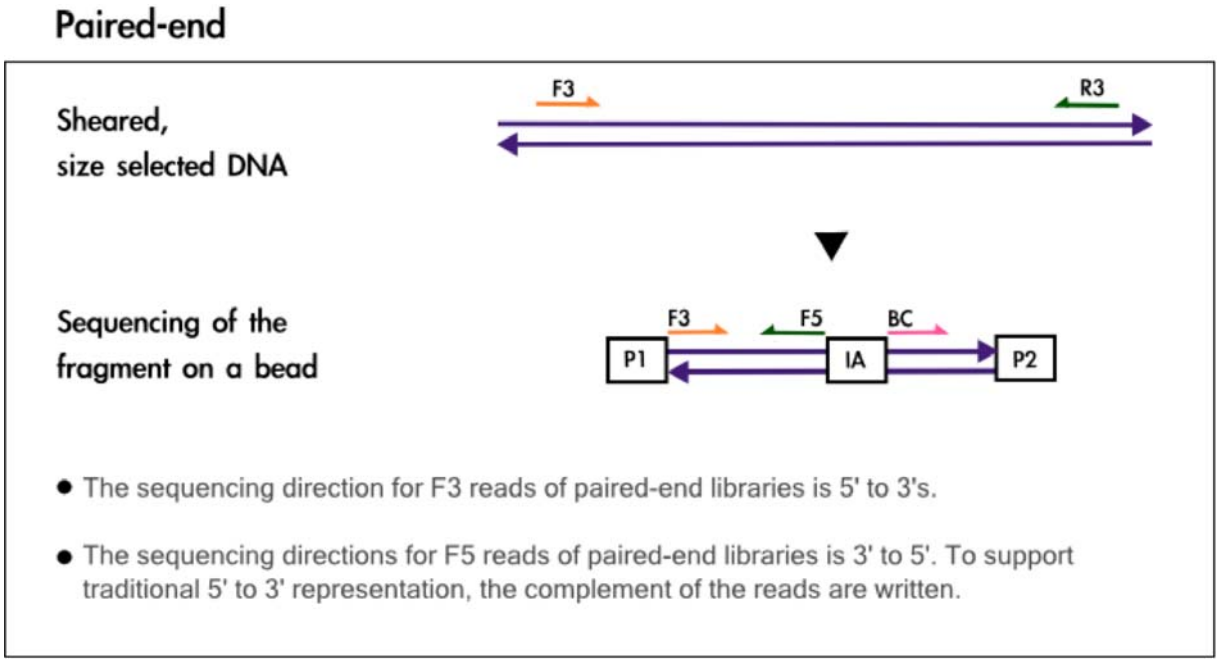


Figura 12 - Construção da biblioteca paired-end. A figura ilustra as etapas da construção de uma biblioteca *paired-end* para sequenciamento. Na maioria das amostras, somente a sequência F3 foi gerada. O BC representa o *barcode* uma sequência conhecida ligada ao fragmento para permitir a identificação da amostra quando é feito o sequenciamento de várias amostras em uma mesma corrida (Figura: Life Technologies).

3.12 Alinhamento para detecção de mutações pontuais

O alinhamento das sequências para a identificação de variações de um único nucleotídeo (*Single Nucleotide Variations* ou SNVs) foi feito como descrito na seção 3.4 com o algoritmo *mapreads* do pacote BioScope. Este algoritmo não é capaz de alinhar sequências com interrupções (*gaps*) e portanto, para identificar InDels (definidos como inserções e deleções com menos de 50nt) nas amostras foi também utilizado o programa NovoAlignCS da NovoCraft. Este algoritmo é capaz de alinhar pequenos trechos de uma sequência e permitir no resultado final que a sequência seja mapeada em uma região de forma descontínua. Os parâmetros do NovoAlignCS alterados foram: –o Softclip –r All –e 10 –p 20,10 0.8,10 –s 5 –K.

3.13 Detecção de variações de sequência

O pipeline para detecção de SNVs e InDels é igual na maioria das etapas, mas executado de forma independente para os resultados do alinhamento BioScope (SNVs) ou do alinhamento NovoAlignCS (InDels). Inicialmente os arquivos são processados para remover sequências mapeadas nas mesmas posições resultantes da duplicação por PCR na etapa de construção da biblioteca. Em seguida, submetemos as sequências ao algoritmo *mpileup* do pacote Samtools (samtools.sourceforge.net). O *mpileup* faz um empilhamento das sequências mapeadas em regiões com sobreposição junto com o genoma referência (hg19) e o parâmetro –B foi utilizado. Após o empilhamento, o programa *bcftools* é usado para selecionar as posições variantes entre o genoma e as sequências da amostra. Com as posições variantes, o programa vcfutils.pl (samtools) é utilizado para aplicar alguns filtros ao resultado, como número mínimo de sequências na

posição variante (–d 3) e qualidade mínima da base avaliada (–Q 25). Estes programas geram como resultado arquivos no formato VCF (samtools) para SNVs e InDels.

Aos arquivos VCF de SNVs e InDels são aplicados outros filtros locais e anotações com programas escritos na linguagem Perl. Foram selecionados somente os casos de SNVs ou InDels com no mínimo 3 sequências indicando a variação do genoma e as sequências que indicam a alteração devem ser provenientes de ambas as fitas do genoma fragmentado, para evitar um viés que resulta em falso positivos (Koboldt *et al.*, 2012). Além disso são anotados os SNVs e InDels correspondentes à variações já descritas na dbSNP (NCBI) e são anotados os SNVs e InDels que estão dentro da região alvo do estudo (no caso a região codificadora do surfaceoma dentro das regiões de captura SureSelect). Outras informações importantes adicionadas com estes programas são em quais genes as variações estão localizadas, assim como em qual posição do gene estão localizadas e que tipo de alteração será gerada (sinônima, não sinônima, *in frame*, ou *frameshift*). No caso de SNVs não sinônimos, também é anotada a troca de códon na sequência e a troca de aminoácido na proteína.

3.14 Análise do impacto funcional

Nem todas as mutações encontradas são variantes genéticas funcionais. Para detectar um possível impacto funcional da mutação na proteína correspondente foram utilizados (com os parâmetros originais) três programas: SIFT (Sorting Intolerant From Tolerant - sift.bii.a-star.edu.sg), PolyPhen-2 (Polymorphism Phenotyping v2 - genetics.bwh.harvard.edu/pph2) e MutationAssessor (mutationassessor.org). As

predições feitas pelos três programas foram combinadas e foram anotadas as mutações que, segundo pelo menos dois dos três programas tem um impacto funcional.

3.15 Análise de epítopos

As mutações não sinônimas e *frameshift* foram avaliadas quanto à geração de novos epítopos nas proteínas correspondentes de forma semelhante à descrita por Segal e colaboradores (Segal *et al.*, 2008). Os algoritmos disponíveis para predição de ligação ao MHC não são adaptados para análises em larga escala como esta, e por isto foi necessário estabelecer um *pipeline* local para processar estes dados. Com programas escritos em Perl, as sequências de cada variante flanqueadas por 10nt adjacentes foram concatenadas em uma única sequência FASTA, incluindo entre elas caracteres para espaçamento. O mesmo foi feito para a sequência referência na mesma posição. Após a geração destas sequências contendo as variações e referências, dois algoritmos de predição de epítopos foram utilizados para detectar nestas sequências peptídeos de 9 aminoácidos com afinidade pela molécula de MHC classe I mais comum na população: HLA- A*0201.

O programa RANKPEP (imed.med.ucm.es/Tools/rankpep.html) usa matrizes de pontuação específicas de conjuntos de peptídeos com ligação conhecida ao MHC como preditores. Os peptídeos testados são considerados epítopos se a porcentagem da afinidade ótima for maior ou igual a 50%. O programa também avalia se o peptídeo testado é formado naturalmente pelo processo de clivagem conhecido do proteassomo. Somente estes peptídeos com clivagem conhecida foram considerados.

O programa NetMHC (www.cbs.dtu.dk/services/NetMHC) utiliza redes neurais artificiais (ANN) para predizer afinidade a moléculas de MHC. Os peptídeos são

considerados epítopos se o valor de afinidade IC_{50} (metade da concentração máxima inibitória) é menor ou igual a 500nM. Para verificar se a clivagem do epítopo ocorre foi utilizado conjuntamente o programa NetChop (www.cbs.dtu.dk/services/NetChop) e somente os peptídeos com clivagem predita foram considerados.

Os resultados dos dois programas foram processados pelo pipeline estabelecido localmente e os epítopos resultantes de artefatos da concatenação das sequências referência ou variante foram removidos através das posições conhecidas no arquivo. Os epítopos mutantes incluídos no resultado final foram aqueles únicos à sequência variante com os valores de ligação ao MHC exigidos para os dois programas e também os epítopos mutantes que quando comparados ao mesmo epítopo referência apresentavam uma afinidade maior ao MHC (pelo menos 20% maior no valor RANKPEP e 20% menor no valor NetMHC).

3.16 Análise de expressão

Para 12 linhagens (CACO2, COLO205, COLO320, HCT116, HCT15, HT29, KM12, LIM1215, LIM2405, RKO, SW480, SW948) também obtivemos dados de RNAseq. Resumidamente, o mRNA total foi extraído para cada linhagem e utilizado para gerar bibliotecas de cDNA. As bibliotecas foram sequenciadas na plataforma SOLiD como descrito em 3.3. O alinhamento e análise dos dados de RNAseq foi feito com um *pipeline* do laboratório baseado nos programas TopHat (tophat.cbcb.umd.edu) e Cufflinks (cufflinks.cbcb.umd.edu). O resultado desta análise estimou a abundância de cada transcrito baseando-se no número de sequências dando suporte a cada genes. Para tanto, o número de sequências por gene é convertido para FPKM (Fragmentos por

kilobase de transcrito por milhão de fragmentos mapeados), um valor normalizado que leva em consideração o tamanho do transcrito e também o rendimento total obtido no sequenciamento. Os genes foram considerados expressos quando o valor FPKM era maior ou igual a 3.0 em pelo menos uma das linhagens estudadas, representando a presença de aproximadamente uma molécula do transcrito por célula, segundo estimativas definidas previamente na literatura (Bakel, van *et al.*, 2011; Marinov *et al.*, 2014; Mortazavi *et al.*, 2008).

3.17 Dados públicos utilizados

3.17.1 Genoma humano de referência

A sequência completa do genoma humano (versão GRCh37/hg19) foi obtida no UCSC (University of California Santa Cruz) *Genome Browser* (Rhead *et al.*, 2010).

3.17.2 Anotações de genes conhecidos

Foram obtidas do projeto RefSeq (Pruitt, Tatusova e Maglott, 2007) todas as sequências correspondendo aos genes codificadores (identificadas por NM_*) e não codificadores (identificadas por NR_*). Também utilizamos as sequências do Ensembl (Flicek *et al.*, 2010). Todas estas sequências foram alinhadas contra o genoma humano (utilizando o BLAT (Kent, 2002)), agrupadas em conjuntos (*clusters*) correspondendo a genes e organizadas em um banco de dados relacional (MySQL – www.mysql.com).

3.17.3 Elementos repetitivos

Regiões do genoma correspondentes a elementos repetitivos e de baixa complexidade definidas pelo RepeatMasker (Smit Hubley, R & Green, P, 2010) foram obtidas através do UCSC *Genome Browser* (Rhead *et al.*, 2010). Para cada uma destas regiões é definida uma posição de início e fim e o tipo de elemento repetitivo presente na mesma. Os limites de centrômeros e telômeros definidos para cada cromossomo também foram obtidos no UCSC *Genome Browser*. As informações necessárias para este projeto foram selecionadas e armazenadas em arquivos simples de texto na forma de tabelas.

3.17.4 Duplicações de Segmento

Regiões do genoma com alta similaridade entre cromossomos são chamadas de duplicações de segmento - *Segmental Duplications* (Bailey *et al.*, 2001) e foram obtidas através do UCSC *Genome Browser* (Rhead *et al.*, 2010). Utilizamos a posição de início e fim para cada uma destas regiões e os cromossomos que apresentam a similaridade.

3.17.5 dbSNP

O dbSNP (www.ncbi.nlm.nih.gov/snp) é o banco de dados de polimorfismos de um único nucleotídeo e deleções/inserções pequenas para diversos organismos. A versão usada para identificação das variantes conhecidas foi a 135. Nesta versão há 44.467.332 entradas para SNPs e 548.248 para InDels com mapeamento único.

3.17.6 DGIdb

Para anotar os genes que possuem interações conhecidas com drogas foi utilizado o banco DGIdb (dgidb.genome.wustl.edu) que mantém o conjunto mais completo de interações conhecidas provenientes da literatura e outros bancos de dados previamente estabelecidos (Griffith *et al.*, 2013). A versão usada foi a de Outubro de 2013.

3.17.7 Kinome

Para identificar os genes do surfaceoma que tem atividade de quinase foi utilizada a lista completa publicada pelo projeto *Human Kinome* (kinase.com/human/kinome).

3.17.8 TCGA

Os genes com mutações pontuais anotados pelo TCGA no estudo de tumores de cólon e reto (TCGA, 2012) foram obtidos do material suplementar da mesma publicação.

3.17.9 KEGG Pathway

O envolvimento dos genes do surfaceoma em vias regulatórias foi avaliado pelo banco KEGG Pathway através da ferramenta KEGG Mapper (www.kegg.jp/kegg/tool/map_pathway1.html).

3.17.10 Dados de Expressão por *Microarray*

Os dados de *microarray* para 8 linhagens (LOVO, SKCO1, SW1116, SW403, SW48, SW620, SW837, e T84) utilizadas no projeto Cancer Cell Line Encyclopedia (www.broadinstitute.org/ccle) foram obtidos do GEO (www.ncbi.nlm.nih.gov/geo) com o *accession*: GSE36133. A média por gene foi obtida para cada linhagem e os genes foram

considerados expressos se o valor era maior ou igual a 5.5 em pelo menos uma das linhagens. Este valor exclui um percentil de ~45% dos genes.

3.18 Estrutura do laboratório de Bioinformática

Todas as análises de bioinformática foram feitas no laboratório de Bioinformática do Instituto de Ensino e Pesquisa do Hospital Sírio Libanês. A estrutura do laboratório conta com 16 servidores de processamento (8 nós de processamento físicos e 8 nós de processamento virtualizados), que agregam 244 unidades de processamento (CPUs) e 1,1 Tb de memória RAM (sendo variável de 23 Gb a 252 Gb por servidor). Ao todo existem 60 TB para armazenamento de dados. Existem também dois servidores para bancos de dados e um servidor para ferramentas web.

4 RESULTADOS

4.1 Biomarcadores em tumores de reto

4.1.1 Sequenciamento das bibliotecas *mate-pair*

Os pacientes selecionados para o estudo representam diferentes respostas obtidas ao tratamento neoadjuvante. O DNA genômico tumoral foi extraído de amostras da biópsia e utilizado para construir bibliotecas pareadas (*mate-pair*) com tamanho médio de inserto de ~700pb. O sequenciamento foi realizado na plataforma SOLiD com cobertura variável para que fosse possível avaliar posteriormente a cobertura real necessária para encontrar as variações estruturais. Também foi feito o sequenciamento de DNA genômico normal (proveniente da camada celular do sangue coletado ou *Buffy Coat*) para três dos pacientes incluídos no estudo (P1, P5 e P6), com o objetivo de comparar as variações estruturais encontradas no tecido tumoral e normal e avaliar a eficiência do pipeline para excluir casos não somáticos. Os dados brutos gerados no sequenciamento estão resumidos na Tabela 2.

Tabela 2 – Sequenciamento dos tumores de reto. Dados brutos de sequenciamento para todas as amostras.

PACIENTE	Sequências Geradas	Nucleotídeos
P1	1.035.604.016	51.780.200.800
P2	393.756.912	19.687.845.600
P3	385.789.584	19.289.479.200
P4	398.436.826	19.921.841.300
P5	425.460.416	29.782.229.120
P6	991.625.036	49.581.251.800
P1 normal	728.574.754	43.714.485.240
P5 normal	682.517.714	40.951.062.840
P6 normal	305.101.394	15.255.069.700

4.1.2 Alinhamento das sequências geradas

As sequências geradas foram alinhadas ao genoma humano referência (hg19) com o auxílio do software *mapreads* do pacote BioScope (Life Technologies). A quantidade de sequências alinhadas com qualidade maior do que 20 para cada amostra pode ser vista na Tabela 3 e corresponde a ~60% do total de sequências geradas para cada amostra. Esta porcentagem de sequências com boa qualidade de alinhamento está dentro do esperado para os resultados da plataforma SOLiD (Harismendy *et al.*, 2009; Suzuki *et al.*, 2011). Estas sequências mapeadas com confiabilidade são usadas para as análises subsequentes. Os dados de mapeamento estão resumidos na Tabela 3.

Tabela 3 – Mapeamento das amostras de tumores de reto. Dados de mapeamento para as amostras com tamanho de inserto ~700pb. Foram utilizadas apenas as sequências com mapeamento confiável (Q >= 20).

PACIENTE	Sequências Mapeadas	Nucleotídeos Mapeados
P1	560.355.119 (54%)	25.967.977.794 (50%)
P2	256.652.253 (65%)	12.232.074.899 (62%)
P3	242.963.348 (63%)	11.616.923.871 (60%)
P4	256.206.645 (64%)	12.231.994.959 (61%)
P5	247.758.837 (58%)	15.150.461.585 (51%)
P6	459.743.724 (46%)	20.921.867.961 (42%)
P1 normal	401.979.094 (55%)	21.532.212.957 (49%)
P5 normal	491.520.545 (72%)	26.824.738.972 (65%)
P6 normal	191.096.606 (63%)	9.129.663.072 (60%)

A cobertura de sequência é um parâmetro importante quando o objetivo do trabalho é detectar alterações de nucleotídeo com alta sensibilidade. Para isto alguns autores sugerem que cada base do genoma deve ser sequenciada, em média, cerca de 30 vezes para garantir confiabilidade (Meyerson, Gabriel e Getz, 2010). Na análise de sequências

pareadas (*mate-pairs*) para detectar variações estruturais, a cobertura física é o mais importante. A cobertura física é calculada somando o tamanho de cada inserto (distância entre sequências de um mesmo *mate-pair*) com o número de bases mapeadas de cada sequência. Para este cálculo, são considerados somente *mate-pairs* que mostram orientação e distância esperadas (média do tamanho de inserto +/- dois desvios padrão). Segundo a literatura não é necessário obter coberturas físicas maiores do que 20x para analisar estas variações estruturais (Roychowdhury *et al.*, 2011). Para as amostras sequenciadas, a cobertura de sequência obtida varia de 4 a 9x, e a cobertura física calculada com base no tamanho médio de inserto dos fragmentos sequenciados varia de 15 a 62x. Os dados de cobertura estão resumidos na Tabela 4. Para as amostras de tecido normal, as coberturas de sequência e física variam de 3,2 a 9,3x e de 15 a 32x, respectivamente.

Tabela 4 - Cobertura. Dados de cobertura de sequência e física para as amostras com tamanho de inserto ~700pb. Tamanho do genoma referência (regiões mapeáveis): 2.897.310.462 bases.

PACIENTE	Cobertura de Sequência do Genoma	Cobertura Física do Genoma
P1	9,0x	180.593.981.498 (62x)
P2	4,2x	57.714.657.702 (20x)
P3	4,0x	53.821.366.172 (19x)
P4	4,2x	59.355.360.167 (20x)
P5	5,2x	42.211.056.664 (15x)
P6	7,2x	77.491.127.440 (27x)
P1 normal	7,4x	78.716.757.426 (27x)
P5 normal	9,3x	94.020.024.357 (32x)
P6 normal	3,2x	43.929.159.747 (15x)

4.1.3 Desenvolvimento e aplicação do *pipeline* de bioinformática

O *pipeline* foi desenvolvido com o objetivo de identificar com eficiência um conjunto mínimo de variações estruturais a partir de genomas tumorais com baixa cobertura de sequenciamento e sem a necessidade de sequenciar o genoma normal pareado. Desta forma buscamos reduzir o custo de sequenciamento e criar uma oportunidade para implementar o uso de biomarcadores personalizados na rotina clínica de tratamento de, neste caso, tumores de reto. As etapas do *pipeline* final utilizado estão descritas detalhadamente em Materiais e Métodos. É importante ressaltar que, inicialmente foram testadas outras combinações de filtros e as etapas que compõe o método apresentado foram adicionadas após avaliações do desenho de *primers* e validações por PCR e sequenciamento mostrarem que seu impacto no conjunto final de variações estruturais encontradas era positivo.

A partir de arquivos em formato bed (BEDTools (Quinlan e Hall, 2010)), foram selecionados os pares de sequências com padrão de alinhamento correspondente ao esperado para rearranjos intercromossomais (Figura 8d). A porcentagem de pares com sequências mapeadas em cromossomos diferentes é semelhante para as amostras (~5%), correspondendo a uma porção esperada do total de sequências segundo a literatura (Campbell *et al.*, 2008). Os pares selecionados foram submetidos aos filtros descritos para remover alinhamentos em regiões repetitivas. Tanto as regiões de baixa complexidade (satélites, low complexity) quanto os elementos transponíveis podem apresentar alta similaridade entre regiões diferentes do genoma. Devido à natureza não exaustiva dos algoritmos de mapeamento, estas regiões podem ocasionar mapeamentos com alta qualidade embora incorretos, e prejudicar a análise de variações estruturais. Da mesma forma, os genes de RNA ribossomal estão dispersos em grande quantidade

pelo genoma e possuem uma alta similaridade, gerando o mesmo tipo de mapeamento incorreto.

As regiões de centrômeros e telômeros são também responsáveis por originar mapeamentos de baixa confiabilidade. Removemos, portanto, sequências localizadas dentro dos limites definidos pela UCSC destas regiões. Removemos também sequências mapeadas em regiões de grande similaridade entre dois cromossomos, compreendendo mais de 1000pb, chamadas de duplicações de segmento - *Segmental duplications* (Bailey *et al.*, 2001). Estas regiões dão origem a mapeamentos em cromossomos diferentes para sequências de um *mate-pair* e induzem nosso *pipeline* a reportar rearranjos falsos.

Após os filtros de cobertura e do alinhamento com o programa BLAT, as sequências foram avaliadas quanto ao padrão de orientação e exigimos que ao menos dois pares reportassem um mesmo rearranjo para mantê-los no conjunto final. Como esperado, um número diverso de eventos indicando rearranjos intercromossomais foi encontrado para cada amostra (variando de 9 a 105, média 32). Quando comparamos os conjuntos de candidatos para as diferentes amostras, notamos que existiam vários eventos recorrentes (presentes em duas ou mais amostras), representados na Figura 13. Eventos em duas amostras diferentes foram considerados iguais e portanto recorrentes quando as sequências alinhadas em cada um dos lados do rearranjo se localizavam na mesma região, de forma semelhante à geração dos clusters em uma mesma amostra. Como tumores sólidos são conhecidos por sua falta de rearranjos tumor-específicos recorrentes (Bass *et al.*, 2011), nós estipulamos que estes eventos eram artefatos e poderiam ser removidos com a comparação com as amostras normais correspondentes (Galante *et al.*, 2011; Ray *et al.*, 2013; Yang *et al.*, 2013). Nós selecionamos oito destes eventos recorrentes para validação por PCR com *primers* desenhados para cada uma das

regiões definidas pelas sequências dos pares indicando o evento. Como esperado para tais falso-positivos, a amplificação com os *primers* específicos ocorreu utilizando como molde tanto o DNA tumoral quanto o DNA normal do mesmo indivíduo (indicado na Figura 13). Estes casos portanto, não são candidatos para rearranjos tumor-específicos e foram removidos de cada um dos conjuntos encontrados. A presença de variações estruturais recorrentes foi observada recentemente em um estudo de rearranjos genômicos complexos (Malhotra *et al.*, 2013) e como estratégia de validação, somente eventos presentes em um único tumor foram considerados somáticos.

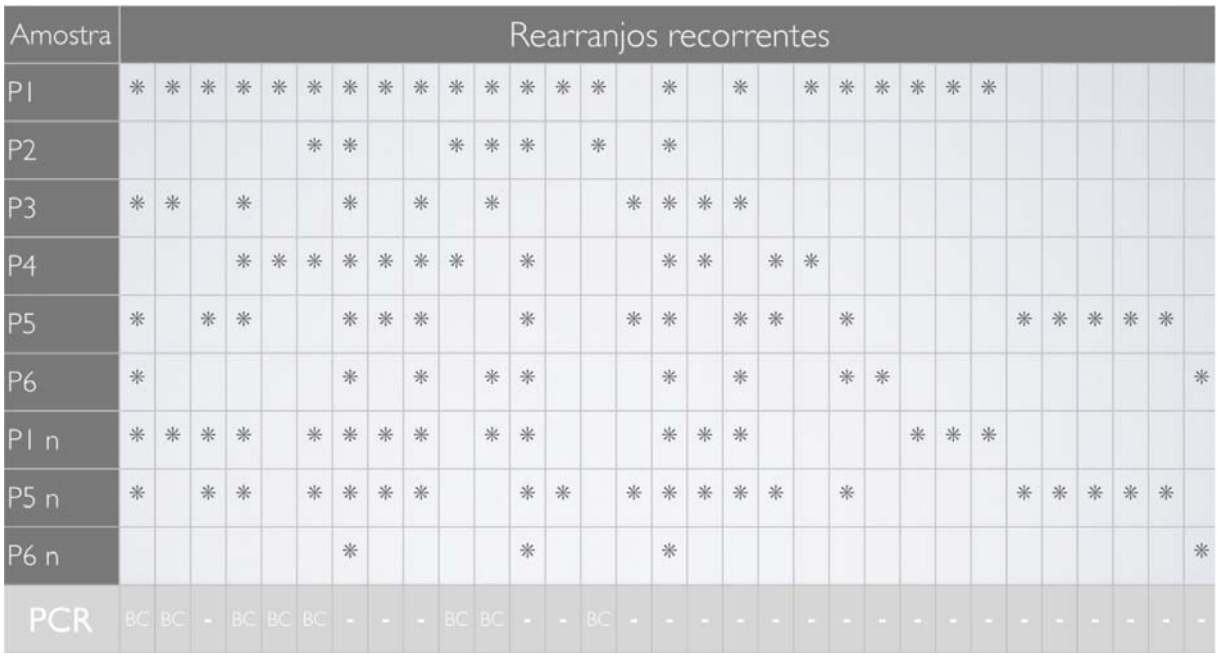


Figura 13- Rearranjos recorrentes. Representação esquemática mostrando os rearranjos encontrados em duas ou mais amostras. Cada coluna corresponde a um rearranjo intercromossomal e o asterisco representa a presença deste rearranjo na amostra indicada pela linha. Estes rearranjos provavelmente não são bons candidatos a biomarcadores e devem corresponder a polimorfismos da população (ou alguma complexidade relativa ao alinhamento das seqüências). BC= Amplificação na PCR também com DNA do *Buffy Coat*. Os demais artefatos não foram submetidos à validação (-). (Figura: Donnard et al., em revisão)

O número encontrado para cada uma das amostras passa a ser de 2 a 96 eventos, com uma média de 23 rearranjos por paciente (detalhado na Tabela 5), sendo este

número e a grande variação observada esperados de acordo com a literatura de alterações estruturais em tumores colorretais (Bass *et al.*, 2011; Leary *et al.*, 2010; Roychowdhury *et al.*, 2011).

Tabela 5 – Rearranjos encontrados em tumores de reto e validações.

PACIENTE	Número de clusters	Rearranjos após filtro recorrentes	Rearranjos testados	Validados como tumor específicos
P1	27	10	6	3 (50%)
P2	18	15	10	9 (90%)
P3	9	4	4	3 (75%)
P4	105	96	11	1 (9%)
P5	14	2	1	1 (100%)
P6	19	14	4	1 (25%)

4.1.4 Comparação com o sequenciamento do tecido normal pareado

Para avaliar a eficiência de detecção de rearranjos tumor-específicos e a importância do sequenciamento do tecido normal pareado, foi sequenciado o DNA de células não tumorais (chamado de DNA normal) para três amostras (P1, P5 e P6). O sequenciamento foi feito com cobertura variada para avaliar também a possibilidade de minimizar o custo e manter os benefícios desta abordagem (Tabela 4). Após a comparação dos rearranjos encontrados nestas amostras e nas tumorais, observamos que a maioria dos casos (62%) de eventos que poderiam ser eliminados pela comparação com o tecido normal pareado também podem ser evitados pela comparação entre amostras tumorais (Figura 13). Por exemplo, dos 13 eventos detectados na amostra P1 e também em seu tecido normal, 10 estavam presentes em algum outro genoma tumoral. Mais do que isto, para algumas amostras o número de eventos removidos pela comparação entre tumores foi maior do que o obtido pela comparação somente com o tecido normal pareado. Três eventos que não foram encontrados no genoma normal do paciente P1 foram removidos

pela comparação com outros tumores e um evento adicional foi removido ao compararmos o conjunto encontrado para o paciente P1 com os eventos no genoma normal de outro paciente. Resultados semelhantes foram obtidos para os outros pacientes com tecido normal sequenciado como mostrado na Figura 13. Já no caso do tecido normal sequenciado com baixa cobertura (P6 normal), a comparação com o tumoral removeu um número muito pequeno de rearranjos, sendo que só um deles foi encontrado exclusivamente pela comparação com o tecido normal, sugerindo que o sequenciamento do tecido normal pareado com baixa cobertura não é uma alternativa eficaz para reduzir o número de falso-positivos. A contribuição do tecido normal pareado não justifica portanto a grande despesa gerada pelo seu sequenciamento e processamento dos dados, sendo uma comparação entre as amostras tumorais capaz de identificar grande parte dos casos de polimorfismos e artefatos que não são bons candidatos para a validação.

4.1.5 Deleções encontradas

Os rearranjos intercromossomais são eventos típicos de uma célula com uma grande instabilidade genômica, sendo muito pouco frequentes em células normais. Como os biomarcadores devem ser tumor específicos, estes eventos apresentam uma melhor oportunidade de evitar os casos germinativos como candidatos a validação. Deleções maiores (acima de 4kb) são também eventos com maior frequência em células com genoma instável e também foram detectados para a obtenção de candidatos a biomarcadores tumorais. Numa análise inicial de deleções para todos os pacientes, utilizando a maioria dos filtros do *pipeline* de rearranjos e a mesma lógica nas outras etapas como descrito em Materiais e Métodos, obteve-se um conjunto final de

candidatos resumidos na Tabela 6. Além de detectar deleções no genoma tumoral, esta análise também acaba identificando inserções de LINEs no genoma referência, por seguirem o mesmo padrão de variação no mapeamento. Procuramos separar estes casos dos eventos verdadeiros de deleção pelo tamanho associado a estes elementos repetitivos, cerca de 6Kb, e a posição genômica identificada para o evento e para todos os LINEs conhecidos no genoma referência. O conjunto final de deleções e rearranjos intercromossomais de cada paciente está representado de forma gráfica na Figura 14.

Tabela 6 – Deleções encontradas em tumores de reto e validações.

PACIENTE	Número de clusters	Deleções após remoção de LINEs	Deleções testadas	Validadas como tumor específicas
P1	36	19	7	4 (57%)
P2	11	5	0	-
P3	22	6	0	-
P4	24	15	10	5 (50%)
P5	22	3	3	2 (66%)
P6	35	10	10	2 (20%)

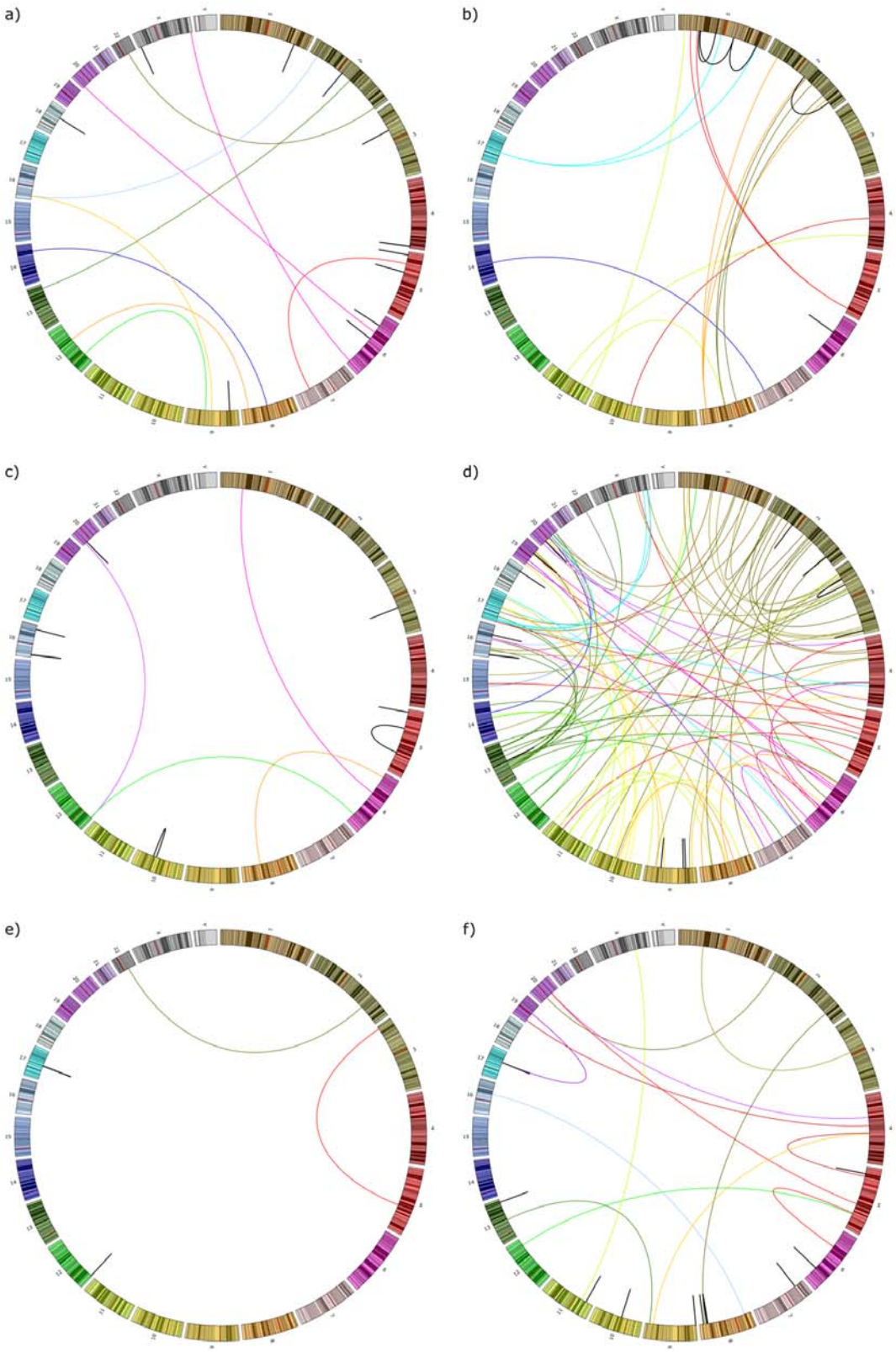


Figura 14 - Variações estruturais nas amostras sequenciadas. Gráfico circos representando os rearranjos intercromossomais (linhas longas coloridas) e as deleções (linhas curtas pretas) encontradas para as amostras P1(a), P2(b), P3(c), P4(d), P5(e) e P6(f).

4.1.6 Validações por PCR e detecção nas amostras de plasma

As etapas de validação foram executadas pela aluna de doutorado Paola A. Carpinetti. Variações estruturais de cada amostra foram selecionadas para validação por PCR. Para isto, *primers* específicos foram desenhados para amplificação da porção genômica contendo o rearranjo ou a deleção. Para o desenho destes *primers* é necessário identificar exatamente a organização do rearranjo e isto é possível devido à lógica de construção da biblioteca. Partindo de um fragmento de DNA são geradas duas sequências no *mate-pair*, F3 e R3. Como foi indicado na Figura 7, a sequência R3 está originalmente localizada a montante (*upstream*) da sequência F3 no genoma. Com esta informação, dependendo da orientação e localização cromossomal das sequências R3 e F3 indicando o rearranjo, sabemos que é necessário desenhar um primer na fita "+" do cromossomo A e um *primer* na fita "-" do cromossomo B, por exemplo. Para deleções o desenho dos *primers* é feito de forma semelhante, observando as orientações das sequências que indicam o evento, porém no mesmo cromossomo.

Os rearranjos e deleções de interesse para utilização como biomarcadores são os que apresentam amplificação somente com o DNA molde tumoral. Obtivemos pelo menos três variações estruturais tumor específicas para cada paciente para que nas etapas posteriores fosse possível identificar ao menos uma das variações nas amostras coletadas em diferentes momentos do tratamento. Nove rearranjos do paciente P2 foram identificados especificamente no tumor assim como três rearranjos do paciente P3 e três do P1 (Figura 15). Ao todo, de 36 rearranjos intercromossomais testados foram validados 18 (Tabela 5), e de 30 deleções testadas foram validadas 13 (Tabela 6). No caso do paciente P4, a grande quantidade de candidatos no final do *pipeline* prejudica a escolha de bons rearranjos para a validação. Já observamos que alguns padrões de

orientação de sequências dos mate-pairs indicando o evento informam que a alteração estrutural é muito complexa, portanto o desenho de *primers* não é tão trivial como os casos que indicam uma translocação simples, e os *primers* utilizados podem estar posicionados de forma incorreta para a amplificação, resultando na baixa taxa de validação obtida para rearranjos deste paciente (Tabela 5). De 10 rearranjos escolhidos para o paciente P4, 9 apresentam este padrão. As deleções por sua vez, proporcionaram ótimos candidatos e das 10 testadas foram validadas 6 como tumor específicas. Para as variações com amplificação por PCR, o sequenciamento do amplicon revela na maioria das vezes o local exato do ponto de quebra. Esta determinação é importante para o desenho de novos *primers* otimizados, mais próximos ao ponto de quebra, que são utilizados para a amplificação do DNA circulante a partir do plasma sanguíneo.

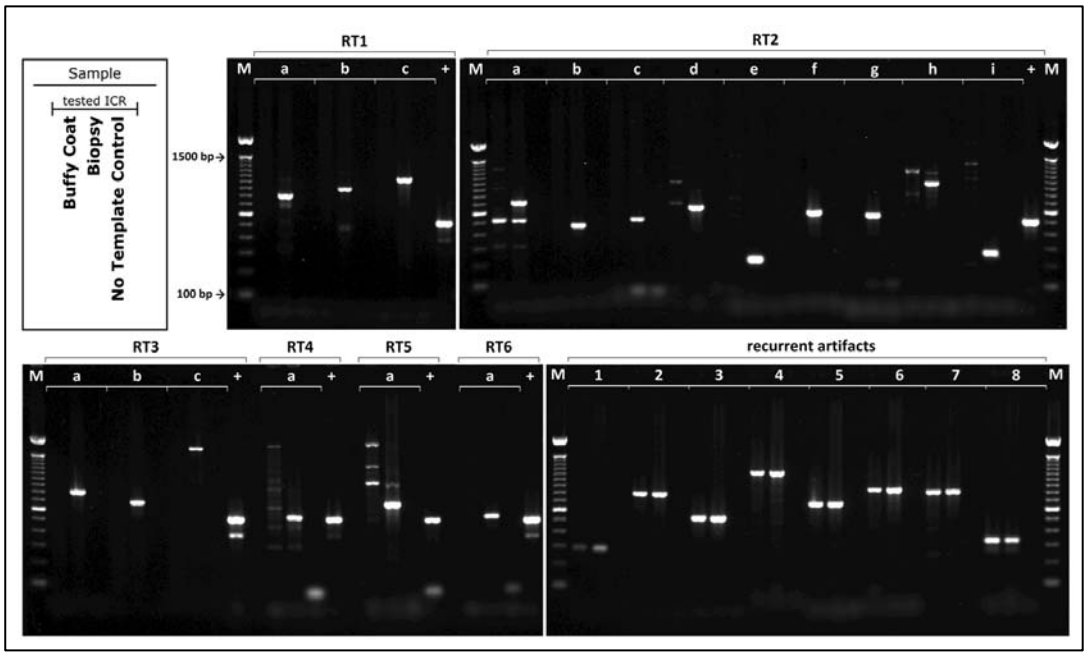


Figura 15 – Validações por PCR. Resultado das amplificações por PCR para rearranjos intercromossomais validados (topo e esquerda inferior) para cada paciente (RT1-6 identificado no topo e rearranjos diferentes identificados por letras logo abaixo) e para os oito artefatos recorrentes testados (direita inferior, identificados por números 1-8). Rearranjos intercromossomais validados apresentam amplificação específica na biópsia. Artefatos recorrentes testados apresentam amplificação tanto no DNA normal (*Buffy Coat*) quanto no DNA tumoral (biópsia). M=marcador; +=Controle positivo de DNA molde BC (locus MLH1 chr3:37001673-37002152); RT1=P1; RT2=P2; RT3=P3; RT4=P4; RT5=P5; RT6=P6. (Figura: Donnard et al., em revisão)

As alterações estruturais tumor-específicas foram testadas em cada um dos pacientes como biomarcadores no plasma. A amplificação da variação foi feita através da PCR digital, um ensaio com sensibilidade alta e necessário para detectar um número muito pequeno de moléculas de DNA tumoral presentes nos 3 mL de plasma coletados em diferentes momentos do tratamento dos pacientes.

Primeiramente, o objetivo das validações no plasma foi detectar as variações estruturais do tumor para todos os pacientes na etapa de coleta da biópsia, antes de qualquer tratamento. Para os pacientes P5, P1 e P3 foi possível detectar um ou mais biomarcadores no diagnóstico inicial, como indicado na Figura 16.

Figura 16 – Detecção de biomarcadores no plasma de pacientes ao longo do tratamento. Cada cor representa um paciente diferente do estudo, as tonalidades em cada quadro representam a detecção de um, dois, ou três dos biomarcadores, do mais claro para o mais escuro, respectivamente. Somente os pacientes com biomarcadores detectados em algum momento foram representados. O asterisco representa o momento da cirurgia. QRT= radioquimioterapia; (0) nenhum biomarcador detectado; (-) plasma não testado.

As detecções observadas na Figura 16 confirmam o quadro clínico para os pacientes P1, P3 e P5. No caso do paciente P5, a recidiva sistêmica da doença foi detectada pelo biomarcador no plasma antes mesmo da detecção pelo exame clínico (semana 61), o que é extremamente interessante e mostra o poder desta metodologia. O Paciente P3, com resposta patológica completa, teria sido poupado da cirurgia já que o exame dos

biomarcadores indica ausência de DNA tumoral após o tratamento de QRT. Para os pacientes P6 e P4, nenhum dos biomarcadores foi detectado em nenhuma amostra de plasma coletada. Este resultado está de acordo com a observação recente de que o DNA tumoral circulante só está presente em cerca de 60% dos pacientes com tumores em estádios não-metastáticos (Bettegowda *et al.*, 2014).

4.1.7 Simulação de genomas com rearranjos intercromossomais

De forma geral, o *pipeline* desenvolvido foi eficaz quando aplicado aos genomas tumorais sequenciados e permitiu a identificação de um conjunto mínimo de variações estruturais personalizadas que puderam ser utilizadas para a detecção de DNA tumoral circulante em amostras de plasma de pacientes com câncer de reto. A heterogeneidade tumoral e a própria instabilidade genômica impedem tanto a identificação de todas as alterações estruturais presentes no tumor quanto uma estimativa verdadeira da acurácia do método descrito. De forma a avaliar melhor o *pipeline* desenvolvido e a cobertura necessária para a detecção das variações estruturais, foi criado um conjunto formado por três genomas simulados computacionalmente. De forma aleatória, foram gerados 20, 30 e 40 rearranjos intercromossomais para cada um dos genomas (GR1, GR2 e GR3, respectivamente). Os detalhes estão descritos em Materiais e Métodos. A partir de cada um destes genomas, foram também gerados três conjuntos de sequências aleatórias, com 50nt de tamanho e inserto entre sequências de um mesmo par de aproximadamente 700nt, simulando, portanto, um sequenciamento em larga escala como na plataforma SOLiD. O resultado representa uma cobertura física simulada de 44x, 25x e 13x, respectivamente (Tabela 7). Usando o *pipeline* e aumentando o filtro de

cobertura final para no mínimo cinco pares indicando o evento, foram identificados 42 dos 90 rearranjos intercromossomais simulados nos três genomas (especificidade de 47%) além de oito falso-positivos (acurácia de 84%; Tabela 8). Como esperado, para as simulações com baixa cobertura de sequência (1,9x e 3,8x; Tabela 7), o resultado obtido teve menor sensibilidade (42,5%) mas uma boa acurácia (surpreendente 100%). Se for reduzida a cobertura mínima para identificar o evento (três pares de sequências) nas simulações GR2 e GR3, um número maior de rearranjos verdadeiros é identificado (18/30 para GR2 e 20/40 para GR3) mas também aparecem alguns falso-positivos no conjunto final (14 para GR2 e 3 para GR3). Os resultados mostrados na Tabela 8 sugerem que o *pipeline* atingiu uma acurácia e sensibilidade aceitável nos experimentos para detecção de rearranjos intercromossomais e cumpriu os objetivos mesmo em condições de baixa cobertura de sequência (1,9x) e física (13x) como no caso do GR3. Os resultados obtidos para as três simulações estão representados de forma gráfica na Figura 17.

Tabela 7 – Genomas com rearranjos (GRs). Sequências geradas, mapeamento e cobertura.

Genomas simulados	Reads Generated	Mapped reads	Cobertura de Sequência	Cobertura Física
GR1	402.771.620	336.548.722 (83,5%)	5,8x	44x
GR2	263.391.883	221.636.923 (84,1%)	3,8x	25x
GR3	131.695.960	110.878.325 (84,2%)	1,9x	13x

Tabela 8 – Rearranjos simulados encontrados

Genomas simulados	Rearranjos simulados	Rearranjos encontrados	Positivos verdadeiros*	Sensibilidade
GR1	20	21	13 (62%)	65% (13/20)
GR2	30	15	15 (100%)	50% (15/30)
GR3	40	14	14 (100%)	35% (15/40)
Total	90	50	42 (84%)	47% (42/90)

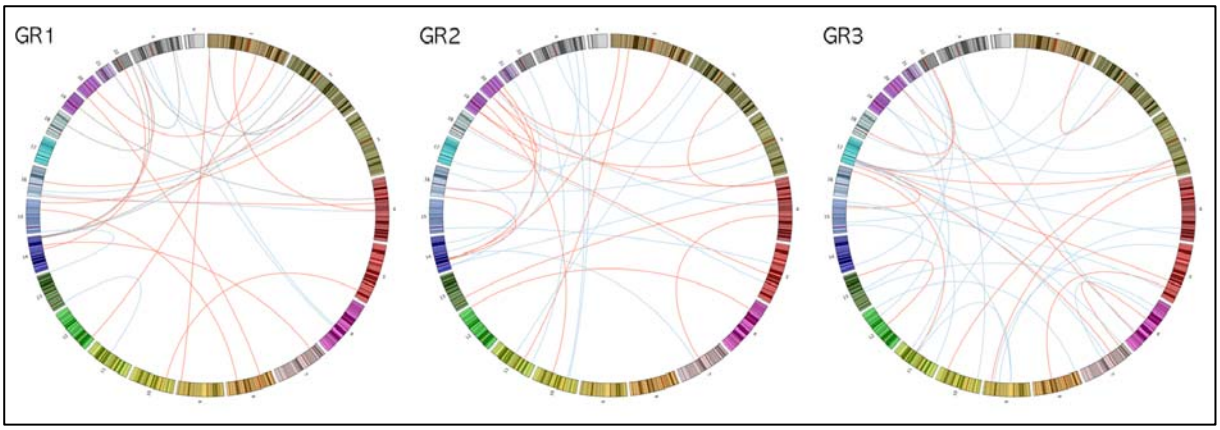


Figura 17 – Resultado da simulação de genomas com rearranjos. A figura mostra o resultado das três simulações feitas indicando os rearranjos corretos encontrados (em vermelho), artefatos (em cinza), e rearranjos corretos não encontrados (em azul).

4.1.8 Expansão da lista de artefatos recorrentes

Uma das principais conclusões após a análise de variações estruturais no genoma de tumores de reto foi que a remoção de casos recorrentes reduz de forma considerável o custo e o tempo de análise das validações na abordagem personalizada, dispensando a necessidade de sequenciamento do genoma normal pareado. Atualmente estão disponíveis diversos dados de projetos de sequenciamento de genomas (tumorais e normais). Com o objetivo de expandir a lista de artefatos e melhorar ainda mais a capacidade de exclusão destes casos em análises futuras, o *pipeline* foi aplicado ao resultado de sequenciamento de 2.362 indivíduos normais provenientes do projeto *1000 Genomes* (www.1000genomes.org) o primeiro a sequenciar o genoma de um grande número de indivíduos normais, com o objetivo de gerar um recurso compreensivo de variações genéticas humanas (Abecasis *et al.*, 2010). Os resultados desta análise revelam um total de 2.800 eventos indicando rearranjos intercromossomais recorrentes em três ou mais indivíduos. Dos 31 casos de artefatos recorrentes identificados na comparação inter-tumoral e com genomas normais

pareados dos pacientes, 16 estão presentes no conjunto encontrado para os indivíduos do 1000 Genomes, Existem também três novos artefatos recorrentes presentes apenas no conjunto do 1000 Genomes que foram identificados no conjunto final de 141 rearranjos dos 6 pacientes avaliados previamente. Dois destes três novos artefatos recorrentes (1 no paciente P6 e 1 no paciente P1) haviam sido testados na validação por PCR e resultaram em amplificação inespecífica e germinativa, respectivamente.

4.2 Novos alvos terapêuticos para o câncer colorretal

4.2.1 Linhagens selecionadas

Existem atualmente diversas linhagens celulares estabelecidas a partir de tumores colorretais, mantendo características moleculares e fisiológicas do câncer das quais se originaram. Para este trabalho, foi selecionado um painel de 23 linhagens (detalhadas em Materiais e Métodos) que possuem características distintas com relação à instabilidade de microssatélites e alterações em genes de reparo (refletindo em diferentes perfis mutacionais) além de respostas variadas aos quimioterápicos 5-FU e oxaliplatina (Arango *et al.*, 2004; Mariadason *et al.*, 2003).

4.2.2 Captura e sequenciamento do exoma

Para cada uma das linhagens selecionadas foi realizada a captura do exoma (porção genômica correspondente aos genes codificantes do organismo) com sondas direcionadas e as bibliotecas *paired-end* geradas foram submetidas ao sequenciamento na plataforma SOLiD.

Das regiões do exoma, selecionamos computacionalmente aquelas correspondentes aos genes codificadores de proteínas de superfície previamente identificadas pelo nosso grupo (Cunha, da *et al.*, 2009), ou surfaceoma, uma região total de 6.097.710 pares de bases do genoma humano, correspondendo a 33.405 éxons de 3.594 genes. Todos os cálculos de cobertura e das análises posteriores foram gerados com relação a este conjunto de genes do surfaceoma. Em média, foi gerado ~1,2G de bases mapeadas na região alvo do estudo levando a uma cobertura de sequência por linhagem de 30x (Tabela 9), sendo que em média 75% da região possui cobertura de pelo menos 10x.

Tabela 9 – Sequenciamento Surfaceoma. Dados de sequências geradas e cobertura para cada linhagem.

LINHAGEM	Bases surfaceoma	Cobertura	%	>10x	>20x
CACO2	858.621.967	21,69	94,01	72,18	65,71
COLO205	964.932.858	23,32	93,53	75,95	69,91
HCC2998	776.194.619	19,59	91,86	67,79	61,35
HCT116	865.230.723	19,63	87,46	70,78	65,22
HCT15	883.598.719	22,35	91,01	73,73	68,32
HT29	834.430.884	19,06	89,45	64,41	58,94
KM12	800.347.001	19,91	88,00	71,21	65,34
LIM1215	780.155.130	20,64	87,64	71,25	65,61
LIM2405	826.244.327	20,01	89,67	67,98	62,62
RKO	826.244.327	20,34	91,89	68,02	61,95
SW480	828.577.905	21,45	89,64	72,90	67,18
SW620	870.322.345	20,79	88,02	71,47	66,15
COLO320	1.422.843.956	37,44	96,67	78,90	64,70
LOVO	1.484.687.708	34,87	96,77	80,80	63,80
RW2982	1.481.300.560	39,87	96,56	75,90	59,00
RW7213	1.609.474.378	43,33	96,77	77,80	63,20
SKCO1	1.564.643.937	41,67	96,88	80,90	68,70
SW1116	1.668.214.550	44,23	96,77	79,90	65,70
SW403	1.980.147.215	49,36	96,77	85,40	75,60
SW48	1.816.359.174	45,51	96,67	84,90	74,20
SW837	1.753.017.836	44,74	96,77	84,60	75,30
SW948	710.553.204	18,46	96,35	51,20	19,30
T84	1.761.549.149	43,46	96,67	86,10	77,20

4.2.3 Detecção de alterações de sequência no surfaceoma

Com as sequências mapeadas para cada linhagem, foi utilizado um *pipeline* de detecção de variações de um único nucleotídeo (SNV – *Single Nucleotide Variation*) e de pequenas inserções e deleções (InDel) baseado no algoritmo *mpileup* do pacote samtools e descrito detalhadamente em Materiais e Métodos. As variações detectadas foram comparadas com anotações disponíveis no dbSNP (NCBI) para que fosse possível eliminar polimorfismos conhecidos do conjunto final de mutações somáticas tumorais, já que não foi sequenciado o DNA normal correspondente de cada linhagem. Como esperado ao comparar as sequências geradas com um único genoma referência, o resultado total dos SNVs compreende uma alta porcentagem de polimorfismos anotados

(Tabela 10), exceto para as linhagens com altas taxas mutacionais, indicando uma boa qualidade no *pipeline* utilizado. No caso de InDels, a porcentagem encontrada com anotação no banco de dados é muito menor, consistente com o fato de que poucas InDels estão anotadas no dbSNP (~11% do banco de dados).

Para gerar a lista final de mutações somáticas, foram removidas as variações recorrentes em mais de 3 linhagens, variações que provavelmente correspondem a polimorfismos não anotados no banco dbSNP, como já foi observado em outros estudos com tumores nos quais genes frequentemente alterados mostravam muitos candidatos somáticos falsos (Kumar *et al.*, 2011). As linhagens apresentam de 41 a 1.071 SNVs somáticas, dependendo de sua característica de instabilidade de microssatélites e mutações em genes de reparo ou no gene POLE (Garraway e Lander, 2013; Heitzer e Tomlinson, 2014; Loeb, 2001; Vogelstein *et al.*, 2013), resultando em uma grande variação na taxa de mutação para cada linhagem (de 1 por 10^6 bases a 100 por 10^6 bases). No total, foram encontradas 3.944 SNVs não sinônimas distintas (Tabela 10) e 595 InDels distintas (Tabela 11) afetando 2.061 genes do surfaceoma (57%), sendo em média 174 SNVs não sinônimas e 26 InDels por linhagem.

Tabela 10 – SNVs no surfaceoma.

LINHAGEM	Total SNVs	% dbSNP	Somáticas	Sinônimas	Não sinônimas
CACO2	2572	97,59	41	17	24
COLO205	2753	97,82	43	10	33
HCC2998	3391	77,65	738	169	569
HCT116	3193	88,41	357	120	237
HCT15	3959	72,57	1071	296	775
HT29	2373	96,29	66	17	49
KM12	2978	85,53	422	138	284
LIM1215	2841	94,65	146	43	103
LIM2405	2731	91,94	207	59	148
RKO	3500	85,66	482	138	344
SW480	2440	95,86	83	23	60
SW620	2535	95,42	99	30	69
COLO320	3813	96,51	86	21	65
LOVO	4591	87,65	495	143	352
RW2982	3545	96,53	80	23	57
RW7213	3763	97,66	62	22	40
SKCO1	4055	95,59	118	39	79
SW1116	3719	96,88	89	25	64
SW403	4196	94,26	165	42	123
SW48	4706	88,02	530	165	365
SW837	3912	96,96	62	18	44
SW948	2820	97,80	56	21	35
T84	4089	95,89	118	32	86

Tabela 11 - InDels

LINHAGEM	Total InDels	% dbSNP	Somáticas	*In frame*	*Frameshift*
CACO2	51	33,33	12	4	8
COLO205	64	29,69	15	4	11
HCC2998	52	36,54	16	2	14
HCT116	106	17,92	61	7	54
HCT15	77	25,97	31	4	27
HT29	49	32,65	9	1	8
KM12	117	15,38	74	9	65
LIM1215	81	20,99	41	2	39
LIM2405	91	17,58	53	0	53
RKO	123	16,26	83	8	75
SW480	43	27,91	8	0	8
SW620	49	28,57	13	4	9
COLO320	43	53,49	2	0	2
LOVO	130	22,31	82	8	74
RW2982	32	50,00	8	3	5
RW7213	44	47,73	7	6	1
SKCO1	43	48,84	5	0	5
SW1116	47	51,06	6	2	4
SW403	48	58,33	4	0	4
SW48	177	22,60	96	4	92
SW837	42	50,00	4	1	3
SW948	34	52,94	3	2	1
T84	51	50,98	5	3	2

4.2.4 Impacto funcional das mutações somáticas encontradas

Para verificar se as SNVs não-sinônimas resultam em um impacto na função proteica foram utilizados três algoritmos (SIFT, PolyPhen2 e MutationAssessor) para estimar o impacto das substituições de aminoácidos baseado em dados de conservação da sequência de DNA, estrutura da proteína e domínios funcionais. Posições com alta conservação tendem a ser intolerantes a substituições (Kumar, Henikoff e Ng, 2009), assim como certas regiões na estrutura da proteína, por exemplo hélices transmembrana (Adzhubei *et al.*, 2010). As mutações encontradas podem resultar em perda de função, ativação, resistência a drogas ou troca de função nas proteínas (Reva, Antipin e Sander, 2011), mas somente ensaios funcionais podem revelar qual destes efeitos realmente ocorre. Ao todo, foram identificadas 1434 SNVs não sinônimas distintas (36% do total) que provavelmente tem impacto nas proteínas respectivas, de acordo com a predição de pelo menos dois dos três programas. Em média, cada linhagem apresenta 62 alterações com impacto predito (Tabela 12).

No caso das InDels, o único algoritmo dos três mencionados capaz de avaliar o impacto funcional é o SIFT. Todas as InDels foram avaliadas e um total de 474 InDels distintas (80% do total) possuem algum impacto funcional predito, sendo em média 21 por linhagem (Tabela 13).

A Tabela 12 resume a análise feita das SNVs não sinônimas encontradas no surfaceoma das 23 linhagens, indicando quantas delas apresentam impacto funcional, quantas são localizadas em genes alvo de drogas conhecidas ou localizadas em quinases. Os resultados também incluem as mutações responsáveis por gerar novos epítopos. Na Tabela 13 os mesmos resultados são mostrados para as InDels. Estas análises serão detalhadas nas seções a seguir.

Tabela 12 –Análise das SNVs encontradas no surfaceoma.

LINHAGEM	Não sinônimas	Impacto funcional	Drogáveis	Kinases	Epítopos
CACO2	24	8	4	2	-
COLO205	33	10	7	-	-
HCC2998	569	196	125	13	11
HCT116	237	81	52	10	6
HCT15	775	287	177	34	19
HT29	49	10	14	3	2
KM12	284	91	61	6	5
LIM1215	102	42	30	4	-
LIM2405	148	48	43	3	5
RKO	344	131	81	7	5
SW480	60	26	11	-	3
SW620	69	26	16	-	-
COLO320	65	24	15	-	2
LOVO	352	160	63	13	6
RW2982	56	17	9	3	2
RW7213	40	13	9	2	-
SKCO1	79	24	9	4	1
SW1116	64	29	9	1	2
SW403	123	36	22	-	-
SW48	365	141	80	7	4
SW837	44	9	6	-	1
SW948	35	14	8	2	-
T84	86	27	14	1	1

Tabela 13 – Análise das InDels encontradas no surfaceoma.

LINHAGEM	InDels	Impacto funcional	Drogáveis	Kinases	Epítopos
CACO2	12	6	-	-	-
COLO205	15	11	2	1	1
HCC2998	16	9	3	-	-
HCT116	61	45	14	1	1
HCT15	31	20	7	1	-
HT29	9	4	2	-	-
KM12	74	54	13	4	1
LIM1215	41	33	10	3	-
LIM2405	53	46	13	3	2
RKO	83	72	14	2	2
SW480	8	7	2	-	-
SW620	13	11	2	-	1
COLO320	2	1	-	-	-
LOVO	82	57	6	3	2
RW2982	8	3	-	-	1
RW7213	7	4	-	-	-
SKCO1	5	3	1	-	-
SW1116	6	3	-	-	-
SW403	4	4	-	-	-
SW48	96	75	11	3	-
SW837	4	2	-	-	-
SW948	3	3	-	-	-
T84	5	4	1	-	-

4.2.5 Genes frequentemente mutados e vias regulatórias

Embora alguns genes implicados no câncer sejam mutados em alta frequência, a maioria das alterações ocorrem em frequências baixas e intermediárias. Estes genes mutados em menor frequência são ainda assim importantes para o tratamento do câncer, visto que as alterações podem ocasionar mudanças regulatórias nas células tumorais que devem ser consideradas na escolha de um tratamento ideal (Lawrence *et al.*, 2014).

Recentemente foi feito um estudo compreensivo de mutações em tumores colorretais (TCGA, 2012), que catalogou um número extenso de mutações em 824 genes com frequências variadas nas 276 amostras analisadas. Mesmo assim, o presente estudo com 23 linhagens celulares de câncer colorretal revela a diversidade de mutações que podem ainda ser descobertas. Foram encontradas mutações pontuais em 171 genes que na análise do TCGA não apresentaram mutações em nenhuma amostra. Representando um acréscimo de 21% no conjunto de genes alterados no câncer colorretal. Alguns destes genes não previamente descritos são expressos nas linhagens estudadas e estão mutados em mais de uma amostra, como por exemplo nove genes mutados em mais de 10% das linhagens (SEMA4C, SLC36A1, FAM38A, FGFRL1, PKD1, TMEM136, WDR81, SLC26A6 e IGFLR1).

Os novos genes mutados estão envolvidos em diversos processos celulares, e alguns deles são membros de vias regulatórias conhecidas por seu envolvimento no câncer. Foram encontradas vias importantes frequentemente alteradas (mais de um gene participante alterado ou um mesmo gene alterado em mais de uma linhagem), como as vias de WNT, EGFR, RAS e PI3K-AKT. As vias regulatórias e genes mutados encontrados

são essenciais para expandir as linhas de tratamento atualmente existentes para este tipo de câncer.

Dentre os nove genes expressos mencionados com expressão em mais de 10% das linhagens, dois alvos interessantes se destacam (SEMA4C e FGFRL1) e são descritos a seguir. Para os demais sete genes (SLC36A1, FAM38A, FGFRL1, PKD1, TMEM136, WDR81 e SLC26A6), não foi encontrado na literatura alguma evidência de papel funcional ou potencial terapêutico para o câncer colorretal. Contudo, mutações recorrentes em FAM38A (proteína envolvida em comunicação celular), SLC36A1 (permeasse capaz de transportar aminoácidos) e WDR81 (proteína envolvida em transdução de sinal) foram observadas em outros tumores primários analisados pelo TCGA, e outros estudos funcionais serão necessários para avaliar o seu envolvimento na tumorigênese colorretal.

SEMA4C codifica um receptor de superfície com atividade sinalizadora da família das semaforinas. As mutações em SEMA4C foram encontradas em 4 das linhagens estudadas (HCT15, KM12, RW2982 e T84) e mutações recorrentes em genes pertencentes a via de sinalização das semaforinas também foram observadas, incluindo mutações em SEMA4D e SEMA4G, que já haviam sido descritas previamente em tumores colorretais (TCGA, 2012). As semaforinas são uma família de proteínas conservadas evolutivamente que recentemente foram envolvidas na progressão do câncer e angiogênese nos tumores (Serini *et al.*, 2009). O papel na angiogênese tumoral foi bem estabelecido para a proteína SEMA4D, e mutações em SEMA4D foram encontradas em 5 linhagens (HCT116, HCT15, LIM1215, LOVO, RKO). No total, mutações em membros da família SEMA4 foram encontrados em 13 das 23 linhagens, sugerindo novos alvos para a terapia antiangiogênica em tumores colorretais.

O gene FGFRL1 está mutado em 4 das linhagens celulares (LOVO, KM12, LIM1215, RKO), três linhagens apresentam InDels com alteração do quadro de leitura da proteína (*frameshift*) e a linhagem restante apresenta uma mutação não-sinônima com impacto funcional predito, indicando uma possível perda de função da proteína FGFRL1 no câncer colorretal. FGFRL1 codifica um receptor de superfície da família FGFR incompleto (não contém o domínio tirosina quinase), que atua como um regulador negativo da sinalização por FGFR ao interferir com a dimerização do FGFR ou pelo sequestro de ligantes (Steinberg *et al.*, 2010). A perda de função de FGFRL1 pode ser portanto um novo mecanismo no câncer colorretal de manutenção da via de FGFR ativa. A amplificação ou superexpressão de receptores de FGF já foi observada em tumores colorretais e está associada com a presença de metástases hepáticas (Sato *et al.*, 2009). Alterações nos genes da família FGFR foram observadas em 8 das 23 linhagens estudadas (35%) indicando um papel importante para esta via no câncer colorretal e sugerindo alternativas de tratamento, especialmente para o câncer metastático. A molécula Regorafenib foi aprovada recentemente para tratamento de câncer colorretal metastático e um de seus alvos é a proteína FGFR1 (Troiani *et al.*, 2013). Estudos pré-clínicos estão avaliando outros inibidores específicos de FGFR1 e os resultados obtidos no estudo das linhagens indica que o uso destes inibidores em tumores colorretais merece atenção especial.

Uma categoria funcional relevante para o estudo do câncer são as quinases, já que a fosforilação é responsável por regular a maioria dos processos celulares (Cohen, 2002). Justamente, o domínio tipo quinase foi o mais representativo no conjunto de genes associados a diversos tipos de câncer definido por Futreal e colaboradores (Futreal *et al.*, 2004). Dos genes que fazem parte do surfaceoma, 71 deles são classificados como

quinases segundo o projeto Human Kinome (Manning *et al.*, 2002). Destes, 54 (76%) apresentam uma ou mais mutações em pelo menos uma das 23 linhagens. Cerca de metade destas quinases (25) estão expressas em pelo menos uma das linhagens, e as mutações encontradas nas quinases expressas podem ser vistas nos anexos A e B.

4.2.6 Mutações em genes drogáveis

As proteínas do surfaceoma são excelentes alvos para intervenções terapêuticas (Cunha, da *et al.*, 2009) e por isto o foco deste estudo foi nas mutações que afetam este conjunto de genes. Para muitos genes, interações conhecidas com drogas estão disponíveis em diversos bancos de dados e podem ser consideradas como alternativas de tratamento para tumores que apresentam mutações nos mesmos. O DGIdb (Drug-Gene Interaction database) integra dados de 13 fontes, incluindo literatura e bancos de drogas, previamente estabelecidos (Griffith *et al.*, 2013). Para os 2.061 genes com mutações não sinônimas ou *frameshift* no surfaceoma, um total de 409 genes (20%) são alvos de drogas conhecidas. Das quinases mutadas mencionadas na seção anterior, 29 são drogáveis e estão entre os alvos mais atraentes para estabelecer novos tratamentos para este câncer.

A tirosina quinase AXL está mutada em cinco das linhagens estudadas (COLO205, KM12, HCT116, HCT15 e LOVO), algumas das mutações estão localizadas no domínio funcional da proteína e tem impacto funcional predito. AXL foi associada a diversos processos celulares incluindo crescimento, migração, agregação e anti-inflamação em diversos tipos celulares. Esta quinase foi também associada à progressão do câncer e resistência ao tratamento quimioterápico em diversos outros tumores (Paccez *et al.*, 2014), o que a torna um candidato muito interessante para um alvo terapêutico em

tumores colorretais. A subfamília TAM de receptores do tipo tirosina quinases inclui também os genes TYRO3 e MERTK, mutados em três das linhagens estudadas. As vias de sinalização reguladas por estas proteínas induzem a ativação de genes associados a fatores de crescimento, como PI3K, RAS e ERK (Verma *et al.*, 2011). Pequenas moléculas inibitórias já existem para as quinases AXL e MERTK e estão em várias fases de testes clínicos e desenvolvimento, e os resultados deste estudo indicam que esta família de quinases deve ser melhor explorada como alvo terapêutico em tumores colorretais também.

4.2.7 Análise de novos epítopos

Além de potencialmente afetar a função e regulação das proteínas, as mutações somáticas que geram modificações na composição de aminoácidos podem gerar novos epítopos reconhecidos pelo MHC de classe I. Estes epítopos são interessantes por serem únicos ao tumor e diferenciarem estas células das células normais, permitindo abordagens específicas para estes antígenos como por exemplo a vacinação antitumoral. As mutações não sinônimas e *frameshift* encontradas nas 23 linhagens foram avaliadas em larga escala com um *pipeline* local, utilizando dois algoritmos de predição de epítopos (NetMHC e RANKPEP) e novos epítopos foram encontrados na maioria das amostras (Tabela 12 e Tabela 13). Partindo de 15.814 peptídeos iniciais contendo mutações foram encontrados ao todo 82 novos epítopos distintos, resultantes de mutações com afinidade predita para a molécula HLA-A*0201 de acordo com ambos os programas (~3 por amostra). A maioria dos peptídeos é descartada pois não existe afinidade predita ao MHC ou não há predição de clivagem para gerar o peptídeo a partir

da sequência proteica. Além disso, só foram considerados novos epítopos com potencial de resposta imune os casos de peptídeos cuja sequência referência não gerasse também um epítopo, ou os casos de peptídeos referência com afinidade muito menor pelo MHC (ver Material e Métodos).

4.2.8 Análise de expressão

A análise do surfaceoma em linhagens de tumores colorretais nos mostra a complexidade e abundância de alterações de sequência nas proteínas de superfície. Estas alterações podem afetar a função e regulação das proteínas em questão, assim como gerar novos epítopos tumor específicos. Contudo, a alteração só terá um impacto ou será responsável pela presença de um novo antígeno se o gene em questão for expresso nesta célula. Para 12 linhagens, estão disponíveis dados obtidos de sequenciamento de mRNA em larga escala (RNA-Seq) em nosso laboratório. A expressão gênica de outras 8 linhagens foi avaliada com dados de experimentos de *microarray* do projeto Cancer Cell Line Encyclopedia (Barretina *et al.*, 2012). Os resultados da análise de expressão mostram que 17% do surfaceoma (624 genes) está expresso em pelo menos uma das linhagens estudadas. Para as SNVs não sinônimas, 1273 de 3.944 estão em genes expressos. No conjunto de 595 InDels encontradas, 206 estão em genes expressos. As mutações em genes expressos estão representadas na Figura 18 e podem ser vistas nos anexos A e B.

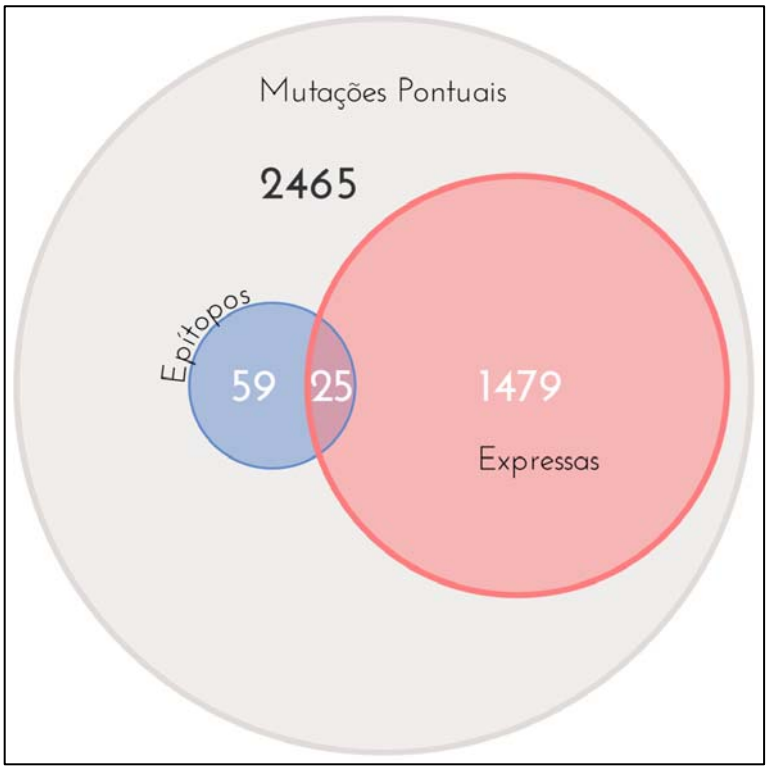


Figura 18 – Mutações expressas nas linhagens colorretais.

Surpreendentemente, um número muito baixo dos epítopos tumorais encontrados provém de genes expressos. Estudos anteriores sugeriram que estes epítopos seriam uma fonte de novos alvos terapêuticos para desenvolvimento de vacinas tumor específicas e encontraram cerca de 7 epítopos por amostra de tumor colorretal (Segal *et al.*, 2008). Como o conjunto de genes estudados neste trabalho é muito menor (~20% dos genes codificadores) já esperava-se encontrar um menor número médio de novos epítopos por linhagem. Porém, esta análise dos dados de expressão mostra que a porcentagem dos epítopos encontrados expressos é muito baixa (apenas 25 de 84; detalhados na Tabela 14), e portanto as abordagens de terapia imunogênica anteriormente consideradas promissoras podem não ser tão eficazes. Por outro lado, foi demonstrado recentemente que pacientes com tumores que apresentam naturalmente números elevados de mutações imunogênicas apresentam maior infiltrado de linfócitos

e maior sobrevida global (Brown *et al.*, 2014). Foi proposto que pacientes com mutações imunogênicas são bons candidatos para anticorpos monoclonais moduladores do sistema imune como anti-CTLA4 e anti-PD-1 (Khalili, Hanson e Szallasi, 2012). Os dados de uso das terapias moduladoras do sistema imune para tumores colorretais são limitados, mas recentemente foi publicado o resultado de um estudo de seguimento de longo prazo do primeiro teste clínico baseado no anticorpo monoclonal contra PD-1. Este estudo incluiu um paciente com tumor colorretal que obteve uma resposta completa e sustentada (>4 anos) ao tratamento (Lipson *et al.*, 2013). Portanto, embora o baixo número de mutações imunogênicas expressas possa ser problemático para a implementação de terapias imunogênicas personalizadas, o uso de drogas moduladoras do sistema imune em pacientes com tumores colorretais contendo taxas elevadas de mutação deve ser considerado como uma alternativa terapêutica.

Tabela 14 – Epítopos expressos gerados pelas mutações pontuais. AA ref= aminoácido na proteína não mutada; AA var = aminoácido na proteína mutada.

LINHAGEM	Chr	Posição	Referência	Variante	Gene	AA ref	AA var	Epítopos
HCC2998	CHR7	38256675	C	T	STARD3NL	S	L	FLLAKVILL
HCC2998	CHR19	10738426	T	A	SLC44A2	F	I	IICCVILLL
HCC2998	CHR7	47937617	A	C	PKD1L1	L	V	TLILPSHTV
HCT116	CHR15	40662233	C	T	DISP2	A	V	CLSTSEPSV
HCT116	CHR2	120003091	A	G	STEAP3	K	E	EMDEPLISL
HCT116	CHR1	11346051	T	G	UBIAD1	F	V	PLLTIPMAV
HCT116	CHR12	57606243	GG	G	LRP1	-	-	TLYMGAMAV
HCT15	CHR1	116605413	A	T	SLC22A15	H	L	GVFAVVNSL
HCT15	CHR12	51868929	C	T	SLC4A8	T	I	ILFFITFIL
HCT15	CHR5	76129424	G	A	F2RL1	C	Y	IVALYLSTL
HCT15	CHR4	6303369	C	T	WFS1	A	V	LLLRWWTKV
HCT15	CHR12	7302164	G	A	CLSTN3	G	D	NLDDCEISL
HCT15	CHR21	45786740	G	T	TRPM2	G	V	NLLISVTGV
HCT15	CHR3	9974300	A	G	IL17RC	K	E	YIHERWALV
HT29	CHR2	95947686	C	T	PROM2	P	L	ELFEFADTL
KM12	CHR4	79400768	C	T	FRAS1	A	V	ALADASDNV
KM12	CHR1	11888183	C	T	CLCN6	H	Y	ALLFHVCYL
LIM2405	CHR6	111587368	TT	T	KIAA1919	-	-	FLVSVIFFV
LIM2405	CHR1	11333834	G	T	UBIAD1	R	S	GVLDPSLLV
LOVO	CHR6	33541938	C	T	BAK1	G	D	ALLGFDYRL
LOVO	CHR6	111587360	ATTTTTTTTT	ATTTTTTTT	KIAA1919	-	-	FLVSVIFFV
RKO	CHR19	14938468	G	C	OR7A5	H	D	FLNDMVIYF
RKO	CHR12	3387704	G	A	TSPAN9	G	S	LVIAISTIV
SW1116	CHR20	50286589	C	T	ATP9A	G	S	HLGTVAYSL
SW48	CHR12	13220116	A	T	KIAA1467	D	V	AVALRVIFV
SW48	CHR5	1403077	G	A	SLC6A3	A	V	SMAMVPIYV
SW480	CHR3	9974300	A	G	IL17RC	K	E	YIHERWALV

5 DISCUSSÃO

5.1 Biomarcadores em tumores de reto

O tratamento do câncer de reto atualmente é resultado de uma grande evolução nos diagnósticos por imagem, estadiamento preciso do tumor, combinações de regimes radioquimioterápicos e precisão cirúrgica. Ainda assim, para a implementação de estratégias de tratamento personalizadas, que permitam poupar os pacientes com resposta completa ao tratamento neoadjuvante de uma abordagem cirúrgica, os métodos de avaliação de resposta ao tratamento e detecção de doença residual necessitam de maior precisão. Poucos cirurgiões se arriscam a seguir a estratégia proposta de "*Watch and Wait*", já que em alguns casos os pacientes parecem ter resposta clínica completa após o tratamento mas o curso não cirúrgico de tratamento abre espaço para uma recidiva local. Já os pacientes com resposta completa patológica dependem completamente de um diagnóstico mais preciso, pois são encaminhados para a cirurgia desnecessária de qualquer forma e estão sujeitos portanto aos riscos e morbidades associados a mesma (Kosinski *et al.*, 2012).

As variações genômicas tem sido cada vez mais estudadas como biomarcadores, para desenvolver abordagens diagnósticas menos invasivas e que permitem o monitoramento da carga tumoral de forma sensível e específica (Leary *et al.*, 2010). A eficiência dessa abordagem já foi comprovada por diversos estudos, mas a implementação do método no cenário clínico ainda depende de adaptações como menores coberturas de sequência e eliminação do sequenciamento do genoma normal pareado (Bettegowda *et al.*, 2014). Com o objetivo de aperfeiçoar esta metodologia e avaliar a possibilidade de sua implementação para guiar a escolha do tratamento em

paciente com câncer de reto, desenvolvemos o protocolo (*pipeline* de bioinformática) descrito de forma a minimizar a cobertura genômica necessária para detectar de forma eficiente um número mínimo de variações estruturais sem a comparação com o genoma normal do mesmo paciente.

A eficiência do *pipeline* obtido pode ser calculada com o auxílio das simulações de genomas com rearranjos, resultando em uma acurácia de 84%. A maior acurácia no conjunto final de candidatos implica em um menor custo e tempo de análises de validação, que podem ser um fator limitante para a aplicação clínica. A acurácia pode ainda ser melhorada, a medida em que são sequenciados mais genomas de indivíduos normais (como os do projeto 1000 Genomes) que permitem a detecção de eventos polimórficos que podem ser confundidos com alterações somáticas quando se analisa somente o genoma tumoral. Embora a sensibilidade do *pipeline* para detectar alterações estruturais não seja alta (47%), no caso da aplicação desejada isto não consiste um empecilho. O grande número de rearranjos presentes na maioria dos tumores sólidos possibilita que um número mínimo de eventos seja validado por paciente. A conhecida heterogeneidade tumoral poderia ser um problema se o diagnóstico dependesse de um único biomarcador, mas os três eventos selecionados garantem que seja possível detectar a presença de tecido tumoral com pelo menos um deles em momentos diferentes, como demostrado pelos nossos resultados e por outros resultados do nosso grupo (dado não mostrado).

O maior obstáculo no momento para a implementação prática desta técnica é a escassez do DNA tumoral circulante (ctDNA), principalmente em tumores localizados. Os resultados obtidos mostram que foi possível detectar o ctDNA em apenas metade dos pacientes, mesmo quando os exames clínicos do paciente em questão indicavam a

presença do tumor residual, e isto provavelmente reflete a ausência de ctDNA ou uma falta de sensibilidade nos métodos de detecção e não uma ausência dos biomarcadores. Os estudos anteriores mostrando sucesso na detecção de DNA tumoral no plasma observaram estes resultados em pacientes com tumores mais avançados, normalmente em estádio IV (metastático). Este ano, uma avaliação mais geral com pacientes em diferentes graus de evolução tumoral mostra um cenário um pouco diferente (Bettegowda *et al.*, 2014). Em pacientes com tumores sólidos avançados de vários tipos, o DNA tumoral circulante foi observado em 82% dos casos. Já nos casos de doença localizada (estádios I-III), a detecção de ctDNA foi possível em 47-69% dos casos. É importante lembrar que para biomarcadores atualmente usados na prática clínica a detecção também não é possível em um grande número de casos (Dawson *et al.*, 2013). Embora a técnica seja mais promissora para casos clínicos metastáticos, cerca de 50% dos pacientes com doença localizada podem se beneficiar do diagnóstico por ctDNA (Bettegowda *et al.*, 2014), dado confirmado pelos nossos resultados.

Uma preocupação comum no caso de abordagens de sequenciamento em larga escala personalizadas é o alto custo da técnica e o tempo de análise proibitivo. Baseados nos dados obtidos, foi feita uma estimativa de custo final associado com a aplicação imediata deste protocolo na clínica. Nas instalações de pesquisa do Hospital Sírio Libanês, o custo de sequenciamento na plataforma SOLiD 5500XL é de aproximadamente US$1 (um dólar) por 20 milhões de bases. Para obter uma cobertura de sequenciamento de 4x para cada amostra de paciente, é preciso gerar ~20 bilhões de bases (400 milhões de sequências com 50nt cada), totalizando ~US$1.000 (mil dólares) por paciente. Para cada paciente o custo final de desenvolvimento de biomarcadores personalizados baseados em alterações estruturais deve incluir custos adicionais para o processamento de

amostras e construção da biblioteca de sequenciamento (~US$470 por paciente) assim como a validação por PCR. O teste de 10 candidatos a biomarcador por paciente implica em um custo de ~US$200 de síntese de *primers* por paciente e ~US$10 por todas as reações de PCR. Assumindo uma taxa de validação de 50%, o sequenciamento por Sanger dos cinco rearranjos validados por PCR para a determinação do ponto de quebra adiciona ~US$50, gerando um custo final de ~US$1730 por paciente. É importante ressaltar que o custo inicial é o maior, e para acompanhar a evolução tumoral do paciente, o custos posteriores caem para menos de 20% do custo inicial. Com relação ao tempo de análise, para detecção das variações estruturais após o sequenciamento, foi estimado que para cada paciente será necessário o período de, no máximo, uma semana. Com a inclusão da etapa de sequenciamento na estimativa, há um aumento no tempo gasto, mas certamente o período de administração da terapia neoadjuvante, cerca de 3 meses, é mais do que o suficiente para a conclusão de todo o processo.

O aumento da acurácia nas análises de sequenciamento irá garantir a eficiência do protocolo, diminuindo ainda mais o custo da abordagem de sequenciamento personalizada e possibilitando sua implementação como um diagnóstico clínico de rotina. Os resultados obtidos neste projeto são motivadores, assim como diversos estudos publicados nos últimos anos (Bass *et al.*, 2011; Bettegowda *et al.*, 2014; Cronin e Ross, 2011; Dawson *et al.*, 2013; Dennis Lo e Chiu, 2011; He *et al.*, 2011; Roychowdhury *et al.*, 2011) destacando o uso do sequenciamento em larga escala para abordagens diagnósticas mais precisas e de maior rendimento de dados.

5.2 Novos alvos terapêuticos para o câncer colorretal

A grande quantidade de mutações presentes nos genomas tumorais resulta na alteração de diversos produtos proteicos e modificações em vias regulatórias das quais estas proteínas participam (Vogelstein *et al.*, 2013). A maioria das alterações somáticas são na verdade inofensivas (passageiras) mas muitas podem ter efeitos oncogênicos ou que alteram a resposta tumoral ao tratamento (Lawrence *et al.*, 2014; Vogelstein *et al.*, 2013). A causa das diferenças na resposta aos medicamentos mais utilizados ainda estão sendo elucidadas mas as alterações em genes diferentes em cada tumor provavelmente tem um papel fundamental para esta diversidade fenotípica (Loeb, 2001).

Seguramente, em um futuro próximo a medicina será mais precisa e tratamentos melhores serão selecionados para cada paciente com a compreensão dos genes e vias alterados no seu próprio tumor (Lawrence *et al.*, 2014). A identificação dos genes alterados em frequências intermediárias na maioria dos tumores e quais destes genes exibem um potencial como alvo de drogas irá facilitar o desenvolvimento destas terapias (Lawrence *et al.*, 2014).

As linhagens celulares estabelecidas a partir de tumores são ótimas ferramentas para o estudo do câncer (Barretina *et al.*, 2012). As linhagens escolhidas de tumores colorretais já foram estudadas previamente por diversos grupos, com perspectivas diferentes. Ao avaliar a sensibilidade destas linhagens aos principais quimioterápicos 5-FU e oxaliplatina, perfis de expressão gênica diferentes foram observados e correlacionados com a resistência à droga (Arango *et al.*, 2004; Mariadason *et al.*, 2003). As diferenças genômicas entre as linhagens, incluindo perfis mutacionais diferentes, despertaram o interesse em detectar as alterações que podem estar afetando o funcionamento das respectivas proteínas.

Nem todas as proteínas contendo mutações com impacto funcional são de interesse terapêutico, e muitas vezes podem tornar difícil a interpretação dos resultados. O conjunto de proteínas localizadas na superfície celular foi escolhido para detectar potenciais novos alvos terapêuticos devido ao seu potencial de resposta à drogas (Cunha, da *et al.*, 2009). Foram encontradas mutações pontuais somáticas em grande parte do surfaceoma (57%) e um terço destas mutações provavelmente tem algum impacto funcional nas respectivas proteínas. Uma grande quantidade dos genes alterados tem interações conhecidas com drogas (409 genes de 2.061). Estudos anteriores mostraram que dos 31 genes mutados em pelo menos 2,5% dos pacientes com câncer de mama, apenas 6 são alvos de drogas conhecidos (AKT1, CDH1, LRP2, PIK3CA, RYR2 e TP53) e portanto é interessante expandir as possibilidades terapêuticas visando os genes mutados em menor frequência (Griffith *et al.*, 2013). Os dados de expressão incluem uma informação valiosa para selecionar melhor os possíveis alvos, já que grande parte das mutações está localizada em genes não expressos nas linhagens.

Algumas mutações pontuais que alteram a sequência proteica acabam gerando novos epítopos imunogênicos específicos do tumor. O estudo destes epítopos em trabalhos anteriores estimou um grande potencial terapêutico através da ativação dos linfócitos contra as células tumorais (Segal *et al.*, 2008). Mesmo trabalhando com um conjunto de genes menor (20% dos genes humanos codificadores), novos epítopos foram identificados na maioria das linhagens com predição de forte ligação à molécula de MHC. Os dados de expressão contudo, mostram que a maioria destes epítopos estão presentes em genes não expressos nas linhagens, diminuindo a expectativa do potencial terapêutico. O trabalho mais recente com epítopos também analisa dados de expressão gênica (Brown *et al.*, 2014), correlacionando a expressão dos genes contendo mutações,

expressão elevada de HLA-A e predição de ligação do novo epítopo ao MHC. Considerando todos estes fatores, apenas 35% dos pacientes analisados possuem em média 3 mutações imunogênicas (de 1 a 147), semelhante ao encontrado nas linhagens analisadas. Os pacientes com mutações imunogênicas mostravam maior expressão de CD8A, indicando a presença de linfócitos T citotóxicos, e apresentaram maior sobrevida.

O estudo das mutações pontuais nas linhagens colorretais revela potenciais novos alvos no surfaceoma para drogas conhecidas e a presença de novos epítopos tumorais com potencial imunogênico. Para o melhor aproveitamento dos dados, a expansão da análise de epítopos para todas as proteínas mutadas, não só do surfaceoma, pode resultar em uma maior proporção de epítopos expressos.

6 CONCLUSÕES

- O protocolo computacional aqui desenvolvido para detecção de variações estruturais elimina a necessidade de sequenciamento tanto do genoma tumoral quanto do genoma normal do mesmo paciente, diminuindo o custo da técnica e o tempo de análise, e possibilitando sua aplicação imediata na prática clínica. Um conjunto eficiente de filtros de mapeamento e comparação entre eventos encontrados em outros genomas tumorais e normais foi implementado para reduzir o número de falso positivos. A abordagem descrita foi capaz de identificar um conjunto mínimo de variações estruturais específicas do tumor para cada um dos pacientes. Os biomarcadores foram detectados no plasma dos pacientes em diversos momentos do tratamento e, na maioria dos casos em que o paciente apresenta DNA tumoral circulante, capazes de avaliar corretamente a resposta tumoral. Os resultados obtidos são promissores para a implementação clínica deste protocolo a curto prazo.

- Foi possível identificar em 23 linhagens celulares derivadas de tumores colorretais novos alvos mutados nos genes de proteínas localizadas na superfície celular (surfaceoma), muitos deles com interações já conhecidas com drogas. A análise de vias contendo os genes mutados também revela alterações importantes neste tipo tumoral, incluindo novos membros mutados em vias envolvidas no câncer descritas previamente.

- As mutações pontuais também são responsáveis pelo surgimento de novos epítopos imunogênicos exclusivos do tumor, e o *pipeline* utilizado é capaz de detectar estes antígenos em larga escala. Os epítopos estão presentes na maioria das linhagens, porém localizados em genes com baixa ou nenhuma expressão, sugerindo uma baixa eficiência se utilizados como vacinas tumorais personalizadas na prática clínica.

7 REFERÊNCIAS

ABECASIS, G. R. *et al.* A map of human genome variation from population-scale sequencing. **Nature**, v. 467, n. 7319, p. 1061–73, 28 out. 2010.

ADZHUBEI, I. A. *et al.* A method and server for predicting damaging missense mutations. **Nature methods**, v. 7, n. 4, p. 248–9, abr. 2010.

ALTSCHUL, S. F. *et al.* Basic local alignment search tool. **Journal of molecular biology**, v. 215, n. 3, p. 403–10, 5 out. 1990.

AMERICAN CANCER SOCIETY. **ACS**. Disponível em: <http://www.cancer.org/>.

ANDRÉ, T. *et al.* Oxaliplatin, fluorouracil, and leucovorin as adjuvant treatment for colon cancer. **The New England journal of medicine**, v. 350, n. 23, p. 2343–51, 3 jun. 2004.

ARANGO, D. *et al.* Molecular mechanisms of action and prediction of response to oxaliplatin in colorectal cancer cells. **British journal of cancer**, v. 91, n. 11, p. 1931–46, 29 nov. 2004.

BAILEY, J. A. *et al.* Segmental duplications: organization and impact within the current human genome project assembly. **Genome Res**, v. 11, n. 6, p. 1005–1017, 2001.

BAKEL, H. VAN *et al.* Response to "The Reality of Pervasive Transcription". **PLoS Biology**, v. 9, n. 7, p. e1001102, 12 jul. 2011.

BARRETINA, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. **Nature**, v. 483, n. 7391, p. 603–7, 29 mar. 2012.

BASS, A. J. *et al.* Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 fusion. **Nature Genetics**, v. 43, n. 10, p. 964–8, 2011.

BETTEGOWDA, C. *et al.* Detection of circulating tumor DNA in early- and late-stage human malignancies. **Science translational medicine**, v. 6, n. 224, p. 224ra24, 19 fev. 2014.

BRANFORD, S. Chronic myeloid leukemia: molecular monitoring in clinical practice. **Hematology / the Education Program of the American Society of Hematology. American Society of Hematology. Education Program**, p. 376–83, jan. 2007.

BRIERLEY, J. D. *et al.* The "y" symbol: an important classification tool for neoadjuvant cancer treatment. **Cancer**, v. 106, n. 11, p. 2526–2527, 2006.

BROWN, S. D. *et al.* Neo-antigens predicted by tumor genome meta-analysis correlate with increased patient survival. **Genome Research**, v. 24, n. 5, p. 743–750, 29 abr. 2014.

CAMPBELL, P. J. *et al.* Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. **Nat Genet**, v. 40, n. 6, p. 722–729, 2008.

CASTLE, J. C. *et al.* Exploiting the mutanome for tumor vaccination. **Cancer research**, v. 72, n. 5, p. 1081–91, 1 mar. 2012.

CHEN, K. *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. **Nature methods**, v. 6, n. 9, p. 677–81, set. 2009.

CHEVILLARD, S. A method for sequential extraction of RNA and DNA from the same sample, specially designed for a limited supply of biological material. **Biotechniques**, v. 15, n. 1, p. 22–24, 1993.

COHEN, P. Protein kinases--the major drug targets of the twenty-first century? **Nature reviews. Drug discovery**, v. 1, p. 309–315, 2002.

CRONIN, M.; ROSS, J. S. Comprehensive next-generation cancer genome sequencing in the era of targeted therapy and personalized oncology. **Biomark Med**, v. 5, n. 3, p. 293–305, 2011.

CUNHA, J. P. DA *et al.* Bioinformatics construction of the human cell surfaceome. **Proc Natl Acad Sci U S A**, v. 106, n. 39, p. 16752–16757, 2009.

CUTSEM, E. VAN *et al.* Addition of aflibercept to fluorouracil, leucovorin, and irinotecan improves survival in a phase III randomized trial in patients with metastatic colorectal cancer previously treated with an oxaliplatin-based regimen. **Journal of clinical oncology : official journal of the American Society of Clinical Oncology**, v. 30, n. 28, p. 3499–506, 1 out. 2012.

DAWSON, S. J. *et al.* Analysis of circulating tumor DNA to monitor metastatic breast cancer. **N Engl J Med**, v. 368, n. 13, p. 1199–1209, 2013.

DENNIS LO, Y.; CHIU, R. W. Plasma nucleic acid analysis by massively parallel sequencing: pathological insights and diagnostic implications. **J Pathol**, v. 225, n. 3, p. 318–323, 2011.

DEVINE, P. L. *et al.* Expression of MUC1 and MUC2 mucins by human tumor cell lines. **Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine**, v. 13, n. 5-6, p. 268–77, jan. 1992.

DIAZ JR., L. A. *et al.* The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers. **Nature**, v. 486, n. 7404, p. 537–540, 2012.

DIEHL, F. *et al.* Detection and quantification of mutations in the plasma of patients with colorectal tumors. **Proc Natl Acad Sci U S A**, v. 102, n. 45, p. 16368–16373, 2005.

DIEHL, F. *et al* . Circulating mutant DNA to assess tumor dynamics. **Nat Med**, v. 14, n. 9, p. 985–990, 2008.

DRIER, Y. *et al.* Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. **Genome Res**, v. 23, n. 2, p. 228–235, 2013.

EDGE, S. B.; COMPTON, C. C. **The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM.Annals of surgical oncology**, 2010.

EWING, B. *et al.* Base-calling of automated sequencer traces using phred. I. Accuracy assessment. **Genome Res**, v. 8, n. 3, p. 175–185, 1998.

FERLAY, J. *et al.* **GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11**. Lyon, France: [s.n.]. Disponível em: <http://globocan.iarc.fr>.

FINK, D. *et al.* In vitro and in vivo resistance to cisplatin in cells that have lost DNA mismatch repair. **Cancer research**, v. 57, n. 10, p. 1841–5, 15 maio 1997.

FISCHER, C. A. *et al.* p16 expression in oropharyngeal cancer: its impact on staging and prognosis compared with the conventional clinical staging parameters. **Ann Oncol**, v. 21, n. 10, p. 1961–1966, 2010.

FLICEK, P. *et al.* Ensembl's 10th year. **Nucleic Acids Res**, v. 38, n. Database issue, p. D557–62, 2010.

FLOHR, T. *et al.* Minimal residual disease-directed risk stratification using real-time quantitative PCR analysis of immunoglobulin and T-cell receptor gene rearrangements in the international multicenter trial AIEOP-BFM ALL 2000 for childhood acute lymphoblastic leukemia. **Leukemia**, v. 22, n. 4, p. 771–82, abr. 2008.

FORSHEW, T. *et al.* Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. **Sci Transl Med**, v. 4, n. 136, p. 136ra68, 2012.

FUTREAL, P. A. *et al.* A census of human cancer genes. **Nature reviews. Cancer**, v. 4, n. 3, p. 177–83, mar. 2004.

GALANTE, P. A. F. *et al.* Distinct patterns of somatic alterations in a lymphoblastoid and a tumor genome derived from the same individual. **Nucleic acids research**, v. 39, n. 14, p. 6056–68, ago. 2011.

GARRAWAY, L. A.; LANDER, E. S. Lessons from the cancer genome. **Cell**, v. 153, n. 1, p. 17–37, 28 mar. 2013.

GRIFFITH, M. *et al.* DGIdb: mining the druggable genome. **Nature methods**, v. 10, n. 12, p. 1209–10, 13 dez. 2013.

HABR-GAMA, A. *et al.* Low rectal cancer: impact of radiation and chemotherapy on surgical treatment. **Dis Colon Rectum**, v. 41, n. 9, p. 1087–1096, 1998.

HABR-GAMA, A.; PEREZ, R. O.; NADALIN, W.; *et al.* Operative versus nonoperative treatment for stage 0 distal rectal cancer following chemoradiation therapy: long-term results. **Ann Surg**, v. 240, n. 4, p. 711–718, 2004.

HABR-GAMA, A.; PEREZ, R. O.; KISS, D. R.; *et al.* Preoperative chemoradiation therapy for low rectal cancer. Impact on downstaging and sphincter-saving operations. **Hepatogastroenterology**, v. 51, n. 60, p. 1703–1707, 2004.

HABR-GAMA, A. *et al.* Long-term results of preoperative chemoradiation for distal rectal cancer correlation between final stage and survival. **J Gastrointest Surg**, v. 9, n. 1, p. 90–101, 2005.

HABR-GAMA, A. *et al* . Interval between surgery and neoadjuvant chemoradiation therapy for distal rectal cancer: does delayed surgery have an impact on outcome? **Int J Radiat Oncol Biol Phys**, v. 71, n. 4, p. 1181–1188, 2008.

HABR-GAMA, A.; PEREZ, R. O. Non-operative management of rectal cancer after neoadjuvant chemoradiation. **Br J Surg**, v. 96, n. 2, p. 125–127, 2009.

HAMILTON, W. *et al.* The risk of colorectal cancer with symptoms at different ages and between the sexes: a case-control study. **BMC Med**, v. 7, p. 17, 2009.

HANAHAN, D.; WEINBERG, R. A. Hallmarks of cancer: the next generation. **Cell**, v. 144, n. 5, p. 646–674, 2011.

HANDSAKER, R. E. *et al.* Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. **Nature genetics**, v. 43, n. 3, p. 269–76, mar. 2011.

HARISMENDY, O. *et al.* Evaluation of next generation sequencing platforms for population targeted sequencing studies. **Genome Biol**, v. 10, n. 3, p. R32, 2009.

HAZKANI-COVO, E.; ZELLER, R. M.; MARTIN, W. Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. **PLoS genetics**, v. 6, n. 2, p. e1000834, fev. 2010.

HE, J. *et al.* IgH gene rearrangements as plasma biomarkers in Non- Hodgkin's lymphoma patients. **Oncotarget**, v. 2, n. 3, p. 178–185, 2011.

HEITZER, E.; TOMLINSON, I. Replicative DNA polymerase mutations in cancer. **Current opinion in genetics & development**, v. 24C, p. 107–113, fev. 2014.

HOLDHOFF, M. *et al.* Analysis of circulating tumor DNA to confirm somatic KRAS mutations. **J Natl Cancer Inst**, v. 101, n. 18, p. 1284–1285, 2009.

HOPKINS, A. L.; GROOM, C. R. The druggable genome. **Nature reviews. Drug discovery**, v. 1, n. 9, p. 727–30, set. 2002.

HURWITZ, H. *et al.* Bevacizumab plus irinotecan, fluorouracil, and leucovorin for metastatic colorectal cancer. **The New England journal of medicine**, v. 350, n. 23, p. 2335–42, 3 jul. 2004.

INCA. **Estimativa 2012: incidência de câncer no Brasil**, 2012.

JOHNSTON, P. G.; KAYE, S. Capecitabine: a novel agent for the treatment of solid tumors. **Anti-cancer drugs**, v. 12, n. 8, p. 639–46, set. 2001.

KANG, H. *et al.* Rare tumors of the colon and rectum: a national review. **Int J Colorectal Dis**, v. 22, n. 2, p. 183–189, 2007.

KATKOORI, V. R. *et al.* Prognostic significance of p53 codon 72 polymorphism differs with race in colorectal adenocarcinoma. **Clin Cancer Res**, v. 15, n. 7, p. 2406–2416, 2009.

KENT, W. J. BLAT--the BLAST-like alignment tool. **Genome Res**, v. 12, n. 4, p. 656–664, 2002.

KENT, W. J. *et al.* The human genome browser at UCSC. **Genome research**, v. 12, n. 6, p. 996–1006, jun. 2002.

KHALILI, J. S.; HANSON, R. W.; SZALLASI, Z. In silico prediction of tumor antigens derived from functional missense mutations of the cancer gene census. **Oncoimmunology**, v. 1, n. 8, p. 1281–1289, 1 nov. 2012.

KOBOLDT, D. C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. **Genome research**, v. 22, n. 3, p. 568–76, 2 mar. 2012.

KOSINSKI, L. *et al.* Shifting concepts in rectal cancer management: a review of contemporary primary rectal cancer treatment strategies. **CA Cancer J Clin**, v. 62, n. 3, p. 173–202, 2012.

KOZAREWA, I. *et al.* Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. **Nature methods**, v. 6, n. 4, p. 291–5, abr. 2009.

KUMAR, A. *et al.* Exome sequencing identifies a spectrum of mutation frequencies in advanced and lethal prostate cancers. **Proceedings of the National Academy of Sciences of the United States of America**, v. 108, n. 41, p. 17087–92, 11 out. 2011.

KUMAR, P.; HENIKOFF, S.; NG, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. **Nature protocols**, v. 4, p. 1073–1081, 2009.

LANDER, E. S. *et al.* Initial sequencing and analysis of the human genome. **Nature**, v. 409, n. 6822, p. 860–921, 15 fev. 2001.

LAWRENCE, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. **Nature**, v. 505, n. 7484, p. 495–501, 23 jan. 2014.

LEARY, R. J. *et al.* Development of personalized tumor biomarkers using massively parallel sequencing. **Sci Transl Med**, v. 2, n. 20, p. 20ra14, 2010.

LEARY, R. J. *et al* . Detection of chromosomal alterations in the circulation of cancer patients with whole-genome sequencing. **Sci Transl Med**, v. 4, n. 162, p. 162ra154, 2012.

LENNERZ, V. *et al.* The response of autologous T cells to a human melanoma is dominated by mutated neoantigens. **Proceedings of the National Academy of Sciences of the United States of America**, v. 102, n. 44, p. 16013–8, 1 nov. 2005.

LEVY, S. *et al.* The diploid genome sequence of an individual human. **PLoS Biol**, v. 5, n. 10, p. e254, 2007.

LI, H. *et al.* The Sequence Alignment/Map format and SAMtools. **Bioinformatics**, v. 25, n. 16, p. 2078–2079, 2009.

LIPSON, E. J. *et al.* Durable cancer regression off-treatment and effective reinduction therapy with an anti-PD-1 antibody. **Clinical cancer research : an official journal of the American Association for Cancer Research**, v. 19, n. 2, p. 462–8, 15 jan. 2013.

LOEB, L. A. A mutator phenotype in cancer. **Cancer research**, v. 61, n. 8, p. 3230–9, 15 abr. 2001.

LONGLEY, D. B.; HARKIN, D. P.; JOHNSTON, P. G. 5-fluorouracil: mechanisms of action and clinical strategies. **Nature reviews. Cancer**, v. 3, n. 5, p. 330–8, maio 2003.

MAJUMDAR, S. R.; FLETCHER, R. H.; EVANS, A. T. How does colorectal cancer present? Symptoms, duration, and clues to location. **The American journal of gastroenterology**, v. 94, n. 10, p. 3039–45, out. 1999.

MALHOTRA, A. *et al.* Breakpoint profiling of 64 cancer genomes reveals numerous complex rearrangements spawned by homology-independent mechanisms. **Genome research**, v. 23, n. 5, p. 762–76, maio 2013.

MANNING, G. *et al.* The protein kinase complement of the human genome. **Science**, v. 298, n. 5600, p. 1912–34, 6 dez. 2002.

MARIADASON, J. M. *et al.* Gene expression profiling-based prediction of response of colon carcinoma cells to 5-fluorouracil and camptothecin. **Cancer research**, v. 63, n. 24, p. 8791–812, 15 dez. 2003.

MARINOV, G. K. *et al.* From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. **Genome research**, v. 24, n. 3, p. 496–510, 1 mar. 2014.

MEDVEDEV, P.; STANCIU, M.; BRUDNO, M. Computational methods for discovering structural variation with next-generation sequencing. **Nat Methods**, v. 6, n. 11 Suppl, p. S13–20, 2009.

MEYERS, M. *et al.* Role of the hMLH1 DNA Mismatch Repair Protein in Fluoropyrimidine-mediated Cell Death and Cell Cycle Responses. **Cancer Res.**, v. 61, n. 13, p. 5193–5201, 1 jul. 2001.

MEYERSON, M.; GABRIEL, S.; GETZ, G. Advances in understanding cancer genomes through second-generation sequencing. **Nat Rev Genet**, v. 11, n. 10, p. 685–696, 2010.

MISALE, S. *et al.* Emergence of KRAS mutations and acquired resistance to anti-EGFR therapy in colorectal cancer. **Nature**, v. 486, n. 7404, p. 532–536, 2012.

MITELMAN, F.; JOHANSSON, B.; MERTENS, F. The impact of translocations and gene fusions on cancer causation. **Nat Rev Cancer**, v. 7, n. 4, p. 233–245, 2007.

MORTAZAVI, A. *et al.* Mapping and quantifying mammalian transcriptomes by RNA-Seq. **Nature methods**, v. 5, n. 7, p. 621–8, jul. 2008.

MOULIERE, F. *et al.* High fragmentation characterizes tumour-derived circulating DNA. **PLoS ONE**, v. 6, n. 9, p. e23418, 2011.

NEGRINI, S.; GORGOULIS, V. G.; HALAZONETIS, T. D. Genomic instability--an evolving hallmark of cancer. **Nature reviews. Molecular cell biology**, v. 11, n. 3, p. 220–8, mar. 2010.

NELSON, H. *et al.* Guidelines 2000 for Colon and Rectal Cancer Surgery. **JNCI Journal of the National Cancer Institute**, v. 93, n. 8, p. 583–596, 18 abr. 2001.

NIELSEN, R. *et al.* Genotype and SNP calling from next-generation sequencing data. **Nature reviews. Genetics**, v. 12, n. 6, p. 443–51, jun. 2011.

PACCEZ, J. D. *et al.* The receptor tyrosine kinase Axl in cancer: biological functions and therapeutic implications. **International journal of cancer. Journal international du cancer**, v. 134, n. 5, p. 1024–33, 1 mar. 2014.

PRUITT, K. D.; TATUSOVA, T.; MAGLOTT, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. **Nucleic Acids Res**, v. 35, n. Database issue, p. D61–5, 2007.

QUINLAN, A. R.; HALL, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. **Bioinformatics (Oxford, England)**, v. 26, n. 6, p. 841–2, 15 mar. 2010.

RAY, M. *et al.* Discovery of structural alterations in solid tumor oligodendroglioma by single molecule analysis. **BMC genomics**, v. 14, p. 505, jan. 2013.

REVA, B.; ANTIPIN, Y.; SANDER, C. Predicting the functional impact of protein mutations: application to cancer genomics. **Nucleic acids research**, v. 39, n. 17, p. e118, 1 set. 2011.

RHEAD, B. *et al.* The UCSC Genome Browser database: update 2010. **Nucleic Acids Res**, v. 38, n. Database issue, p. D613–9, 2010.

ROYCHOWDHURY, S. *et al.* Personalized oncology through integrative high-throughput sequencing: a pilot study. **Sci Transl Med**, v. 3, n. 111, p. 111ra121, 2011.

RUSS, A. P.; LAMPEL, S. The druggable genome: an update. **Drug discovery today**, v. 10, n. 23-24, p. 1607–10, dez. 2005.

SALK, J. J.; FOX, E. J.; LOEB, L. A. Mutational heterogeneity in human cancers: origin and consequences. **Annual review of pathology**, v. 5, p. 51–75, jan. 2010.

SATO, T. *et al.* Overexpression of the fibroblast growth factor receptor-1 gene correlates with liver metastasis in colorectal cancer. **Oncology Reports**, v. 21, n. 1, p. 211–216, 1 jan. 2009.

SAUER, R. *et al.* Preoperative versus postoperative chemoradiotherapy for rectal cancer. **N Engl J Med**, v. 351, n. 17, p. 1731–1740, 2004.

SCHERER, S. W. *et al.* Human chromosome 7: DNA sequence and biology. **Science**, v. 300, n. 5620, p. 767–772, 2003.

SCHMIDT, C. R.; GOLLUB, M. J.; WEISER, M. R. Contemporary imaging for colorectal cancer. **Surgical oncology clinics of North America**, v. 16, n. 2, p. 369–88, abr. 2007.

SEGAL, N. H. *et al.* Epitope landscape in breast and colorectal cancer. **Cancer research**, v. 68, n. 3, p. 889–92, 1 fev. 2008.

SERINI, G. *et al.* Semaphorins and tumor angiogenesis. **Angiogenesis**, v. 12, n. 2, p. 187–93, jan. 2009.

SMIT HUBLEY, R & GREEN, P, A. F. A. **RepeatMasker Open-3.0**. Disponível em: <www.repeatmasker.org>.

STARR, T. K. *et al.* A transposon-based genetic screen in mice identifies genes altered in colorectal cancer. **Science**, v. 323, n. 5922, p. 1747–1750, 2009.

STEIN, W. *et al.* Characteristics of colon cancer at time of presentation. **Family practice research journal**, v. 13, n. 4, p. 355–63, dez. 1993.

STEINBERG, F. *et al.* The FGFRL1 receptor is shed from cell membranes, binds fibroblast growth factors (FGFs), and antagonizes FGF signaling in Xenopus embryos. **The Journal of biological chemistry**, v. 285, n. 3, p. 2193–202, 15 jan. 2010.

SUZUKI, S. *et al.* Comparison of sequence reads obtained from three next-generation sequencing platforms. **PLoS ONE**, v. 6, n. 5, p. e19534, 2011.

TALKOWSKI, M. E. *et al.* Next-generation sequencing strategies enable routine detection of balanced chromosome rearrangements for clinical diagnostics and genetic research. **Am J Hum Genet**, v. 88, n. 4, p. 469–481, 2011.

TCGA. Comprehensive molecular characterization of human colon and rectal cancer. **Nature**, v. 487, n. 7407, p. 330–7, 19 jul. 2012.

THIERRY, A. R. *et al.* Origin and quantification of circulating DNA in mice with human colorectal cancer xenografts. **Nucleic Acids Res**, v. 38, n. 18, p. 6159–6175, 2010.

TIBBETTS, L. M. *et al.* Cell culture of the mucinous variant of human colorectal carcinoma. **Cancer research**, v. 48, n. 13, p. 3751–9, 1 jul. 1988.

TOWLER, B. *et al.* A systematic review of the effects of screening for colorectal cancer using the faecal occult blood test, hemoccult. **BMJ**, v. 317, n. 7158, p. 559–565, 1998.

TROIANI, T. *et al.* Targeted approach to metastatic colorectal cancer: what comes beyond epidermal growth factor receptor antibodies and bevacizumab? **Therapeutic advances in medical oncology**, v. 5, n. 1, p. 51–72, jan. 2013.

VERMA, A. *et al.* Targeting Axl and Mer kinases in cancer. **Molecular cancer therapeutics**, v. 10, n. 10, p. 1763–73, 1 out. 2011.

VOGELSTEIN, B. *et al.* Cancer genome landscapes. **Science**, v. 339, n. 6127, p. 1546–1558, 2013.

VOGELSTEIN, B.; KINZLER, K. W. Digital PCR. **Proceedings of the National Academy of Sciences of the United States of America**, v. 96, n. 16, p. 9236–41, 3 ago. 1999.

WARREN, R. L.; HOLT, R. A. A census of predicted mutational epitopes suitable for immunologic cancer control. **Human immunology**, v. 71, n. 3, p. 245–54, mar. 2010.

WHITEHEAD, R. H. *et al.* A colon cancer cell line (LIM1215) derived from a patient with inherited nonpolyposis colorectal cancer. **Journal of the National Cancer Institute**, v. 74, n. 4, p. 759–65, abr. 1985.

WOOD, L. D. *et al.* The genomic landscapes of human breast and colorectal cancers. **Science**, v. 318, n. 5853, p. 1108–1113, 2007.

YANG, L. *et al.* Diverse mechanisms of somatic structural variations in human cancer genomes. **Cell**, v. 153, n. 4, p. 919–29, 9 maio 2013.

YEWDELL, J. W.; BENNINK, J. R. Immunodominance in major histocompatibility complex class I-restricted T lymphocyte responses. **Annual review of immunology**, v. 17, p. 51–88, jan. 1999.

LISTA DE ANEXOS

ANEXO A: Tabela com SNVs em genes expressos

SNVs não-sinônimos encontrados em genes expressos nas linhagens										
Linhagem	Chr	Posição	Ref	Var	Gene	SIFT	PolyPhen2	MutAcessor	Kinoma	Drogável
CACO2	CHR1	153751933	G	A	SLC27A3	DAMAGING	-	low	no	no
CACO2	CHR1	208255817	C	T	PLXNA2	DAMAGING	-	medium	no	no
CACO2	CHR13	95735444	C	T	ABCC4	-	-	neutral	no	yes
CACO2	CHR17	48755267	A	G	ABCC3	DAMAGING	DAMAGING	high	no	yes
CACO2	CHR22	19055679	G	A	DGCR2	-	-	medium	no	no
CACO2	CHR3	45132683	G	A	CDCP1	-	-	low	no	no
CACO2	CHR6	117674165	G	T	ROS1	-	-	low	yes	yes
CACO2	CHR7	75513072	G	A	RHBDD2	-	-	low	no	no
COLO205	CHR10	102743263	G	T	SEMA4G	-	-	neutral	no	no
COLO205	CHR11	57268637	C	T	SLC43A1	-	-	medium	no	no
COLO205	CHR11	68213980	C	T	LRP5	-	DAMAGING	medium	no	no
COLO205	CHR1	209933481	C	T	TRAF3IP3	DAMAGING	-	medium	no	no
COLO205	CHR16	16232225	G	A	ABCC1	DAMAGING	DAMAGING	medium	no	yes
COLO205	CHR20	44806828	C	T	CDH22	DAMAGING	DAMAGING	medium	no	no
COLO205	CHR5	140256399	G	A	PCDHA12	DAMAGING	-	medium	no	no
COLO205	CHR6	149285664	C	T	UST	DAMAGING	DAMAGING	medium	no	yes
COLO205	CHR6	167549743	A	G	CCR6	DAMAGING	-	neutral	no	no
COLO205	CHR6	26446030	A	C	BTN3A3	-	-	neutral	no	no
COLO205	CHR6	26446042	A	G	BTN3A3	-	-	neutral	no	no
COLO205	CHR9	119976858	C	T	ASTN2	DAMAGING	-	neutral	no	no
COLO320	CHR1	109714594	G	A	KIAA1324	-	-	low	no	no
COLO320	CHR11	14825503	C	G	PDE3B	-	-	low	no	yes
COLO320	CHR11	66136079	C	G	SLC29A2	-	-	neutral	no	no
COLO320	CHR14	71524364	A	G	PCNX	DAMAGING	DAMAGING	medium	no	no
COLO320	CHR15	68500659	C	T	CLN6	-	-	low	no	no
COLO320	CHR15	90768389	G	A	SEMA4B	-	-	low	no	no
COLO320	CHR16	2158596	C	T	PKD1	DAMAGING	DAMAGING	medium	no	no
COLO320	CHR18	29793156	C	T	MEP1B	DAMAGING	-	low	no	no
COLO320	CHR19	36435593	C	T	LRFN3	DAMAGING	DAMAGING	medium	no	no
COLO320	CHR21	30953829	A	G	GRIK1	DAMAGING	DAMAGING	medium	no	yes
COLO320	CHR21	45798949	G	A	TRPM2	-	-	neutral	no	no
COLO320	CHR2	160761171	G	C	LY75	-	-	medium	no	no
COLO320	CHR5	140255988	G	A	PCDHA12	DAMAGING	-	high	no	no
COLO320	CHR5	140741831	C	A	PCDHGB2	DAMAGING	-	medium	no	no
COLO320	CHR5	140772601	T	A	PCDHGA8	-	-	low	no	no
COLO320	CHRX	152960008	G	A	SLC6A8	-	-	neutral	no	yes
COLO320	CHRX	49126913	A	C	PPP1R3F	-	-	neutral	no	no
HCC2998	CHR10	105209640	T	C	CALHM2	-	-	low	no	no
HCC2998	CHR10	95185861	C	A	MYOF	DAMAGING	-	medium	no	no
HCC2998	CHR1	109730918	C	A	KIAA1324	-	-	medium	no	no
HCC2998	CHR1	110466024	T	G	CSF1	-	-	neutral	no	yes
HCC2998	CHR11	113073162	C	A	NCAM1	-	-	-	no	no
HCC2998	CHR11	117165945	G	A	BACE1	DAMAGING	DAMAGING	medium	no	yes
HCC2998	CHR11	120200748	G	A	TMEM136	DAMAGING	DAMAGING	medium	no	no
HCC2998	CHR11	125830883	G	T	CDON	-	-	medium	no	no
HCC2998	CHR11	4112604	A	G	STIM1	-	-	low	no	no
HCC2998	CHR1	153748997	T	G	SLC27A3	DAMAGING	DAMAGING	neutral	no	no
HCC2998	CHR11	57193087	T	G	SLC43A3	DAMAGING	DAMAGING	medium	no	no
HCC2998	CHR1	158057815	T	G	KIRREL	-	-	neutral	no	no
HCC2998	CHR11	59559682	G	A	STX3	-	DAMAGING	medium	no	no
HCC2998	CHR11	68177402	G	T	LRP5	-	DAMAGING	neutral	no	no
HCC2998	CHR1	19175987	C	T	TAS1R2	-	-	low	no	yes
HCC2998	CHR12	13215874	C	T	KIAA1467	-	-	-	no	no
HCC2998	CHR1	21563331	C	T	ECE1	-	-	low	no	yes
HCC2998	CHR1	223177420	T	G	DISP1	DAMAGING	-	low	no	no
HCC2998	CHR1	223438053	T	G	SUSD4	-	DAMAGING	neutral	no	no
HCC2998	CHR1	226065189	T	C	TMEM63A	-	DAMAGING	medium	no	no
HCC2998	CHR12	26816628	C	A	ITPR2	DAMAGING	DAMAGING	medium	no	no
HCC2998	CHR1	236332034	C	A	GPR137B	DAMAGING	-	medium	no	no
HCC2998	CHR1	24485998	G	T	IL28RA	DAMAGING	-	medium	no	no
HCC2998	CHR12	56351386	C	A	PMEL	DAMAGING	DAMAGING	medium	no	no

HCC2998	CHR12	7585167	G	T	CD163L1	-	-	neutral	no	no
HCC2998	CHR1	32198122	A	C	BAI2	-	-	low	no	no
HCC2998	CHR1	32203095	A	C	BAI2	DAMAGING	DAMAGING	low	no	no
HCC2998	CHR1	32207291	G	T	BAI2	DAMAGING	-	neutral	no	no
HCC2998	CHR1	40432547	C	A	MFSD2A	-	-	neutral	no	no
HCC2998	CHR14	23245454	A	C	SLC7A7	DAMAGING	DAMAGING	medium	no	no
HCC2998	CHR1	45292607	C	T	PTCH2	-	-	low	no	no
HCC2998	CHR14	58605919	T	C	C14ORF37	-	-	medium	no	no
HCC2998	CHR1	46650467	T	C	TSPAN1	-	-	low	no	no
HCC2998	CHR14	95670487	T	C	CLMN	DAMAGING	-	low	no	no
HCC2998	CHR15	63631072	A	C	CA12	-	DAMAGING	low	no	yes
HCC2998	CHR15	65943014	A	C	SLC24A1	-	-	neutral	no	no
HCC2998	CHR16	19501839	G	A	TMC5	-	-	neutral	no	no
HCC2998	CHR16	4164499	T	G	ADCY9	DAMAGING	DAMAGING	medium	no	no
HCC2998	CHR16	50348970	A	C	ADCY7	DAMAGING	DAMAGING	medium	no	no
HCC2998	CHR16	68842729	G	T	CDH1	DAMAGING	DAMAGING	high	no	yes
HCC2998	CHR17	26684860	C	A	TMEM199	DAMAGING	-	neutral	no	no
HCC2998	CHR17	43216468	A	G	ACBD4	-	-	-	no	yes
HCC2998	CHR17	65026687	C	A	CACNG4	DAMAGING	DAMAGING	medium	no	no
HCC2998	CHR17	7318876	G	A	NLGN2	DAMAGING	DAMAGING	medium	no	no
HCC2998	CHR19	10738426	T	A	SLC44A2	-	-	neutral	no	no
HCC2998	CHR1	92200359	C	A	TGFBR3	DAMAGING	-	low	no	no
HCC2998	CHR19	38848953	G	A	CATSPERG	-	-	medium	no	no
HCC2998	CHR19	38853070	G	T	CATSPERG	-	-	-	no	no
HCC2998	CHR19	42485962	G	A	ATP1A3	-	DAMAGING	low	no	no
HCC2998	CHR19	51133101	G	T	SYT3	-	-	neutral	no	no
HCC2998	CHR19	55710235	C	T	PTPRH	-	-	low	no	no
HCC2998	CHR19	55715272	C	T	PTPRH	-	-	low	no	no
HCC2998	CHR19	7964040	G	T	LRRC8E	-	-	low	no	no
HCC2998	CHR20	3003380	G	A	PTPRA	DAMAGING	-	low	no	no
HCC2998	CHR20	30738616	T	G	TM9SF4	DAMAGING	-	medium	no	no
HCC2998	CHR21	31062065	A	G	GRIK1	DAMAGING	-	medium	no	yes
HCC2998	CHR21	45837898	G	A	TRPM2	-	-	neutral	no	no
HCC2998	CHR2	202900778	C	T	FZD7	DAMAGING	DAMAGING	high	no	no
HCC2998	CHR2	219137385	A	G	PNKD	-	-	low	no	no
HCC2998	CHR2	220028161	C	T	SLC23A3	-	-	medium	no	no
HCC2998	CHR2	220092646	T	C	ATG9A	-	DAMAGING	low	no	no
HCC2998	CHR2	220502893	G	A	SLC4A3	DAMAGING	DAMAGING	medium	no	no
HCC2998	CHR2	230911135	G	T	SLC16A14	-	-	low	no	no
HCC2998	CHR22	46932200	C	T	CELSR1	-	-	low	no	no
HCC2998	CHR2	27430382	T	G	SLC5A6	DAMAGING	-	medium	no	yes
HCC2998	CHR2	8871839	G	A	KIDINS220	-	-	neutral	no	no
HCC2998	CHR2	97463322	A	G	CNNM4	-	-	neutral	no	no
HCC2998	CHR3	124527903	T	G	ITGB5	-	-	low	no	yes
HCC2998	CHR3	126708064	G	A	PLXNA1	-	DAMAGING	low	no	no
HCC2998	CHR3	183995145	C	A	ECE2	DAMAGING	-	medium	no	no
HCC2998	CHR3	37774271	C	A	ITGA9	-	-	medium	no	no
HCC2998	CHR3	49569259	A	G	DAG1	-	-	neutral	no	no
HCC2998	CHR3	53845183	A	C	CACNA1D	DAMAGING	-	medium	no	yes
HCC2998	CHR3	66449409	T	G	LRIG1	DAMAGING	-	neutral	no	no
HCC2998	CHR4	16024954	G	A	PROM1	-	-	medium	no	no
HCC2998	CHR4	170913056	C	T	MFAP3L	-	DAMAGING	low	no	no
HCC2998	CHR4	55161339	C	A	PDGFRA	-	DAMAGING	low	yes	yes
HCC2998	CHR5	140248752	G	A	PCDHA11	-	-	low	no	no
HCC2998	CHR5	140262322	G	A	PCDHA13	DAMAGING	-	high	no	no
HCC2998	CHR5	140753655	C	T	PCDHGA6	-	-	low	no	no
HCC2998	CHR5	140754398	A	C	PCDHGA6	-	-	low	no	no
HCC2998	CHR5	140754506	G	T	PCDHGA6	-	-	-	no	no
HCC2998	CHR5	140782910	C	T	PCDHGA9	DAMAGING	-	high	no	no
HCC2998	CHR5	140865688	T	G	PCDHGC4	-	-	-	no	no
HCC2998	CHR5	36679758	T	G	SLC1A3	DAMAGING	DAMAGING	medium	no	yes
HCC2998	CHR6	170594398	C	T	DLL1	DAMAGING	-	medium	no	no
HCC2998	CHR6	26413436	T	C	BTN3A1	DAMAGING	DAMAGING	medium	no	no

HCC2998	CHR6	43109672	T	G	PTK7	-	-	low	yes	no
HCC2998	CHR7	142574519	C	T	TRPV6	DAMAGING	DAMAGING	medium	no	no
HCC2998	CHR7	38256675	C	T	STARD3NL	-	DAMAGING	medium	no	no
HCC2998	CHR7	47937617	A	C	PKD1L1	DAMAGING	DAMAGING	low	no	no
HCC2998	CHR7	99474317	G	T	OR2AE1	-	DAMAGING	medium	no	no
HCC2998	CHR8	125499050	G	A	RNF139	-	-	low	no	no
HCC2998	CHR8	17611890	T	G	MTUS1	-	-	low	no	no
HCC2998	CHR8	95164141	T	G	CDH17	-	-	neutral	no	no
HCC2998	CHR9	131107816	G	A	SLC27A4	-	-	medium	no	no
HCC2998	CHR9	34657070	G	A	IL11RA	-	DAMAGING	low	no	yes
HCC2998	CHR9	34658520	G	A	IL11RA	DAMAGING	DAMAGING	medium	no	yes
HCC2998	CHR9	74360124	T	G	TMEM2	-	-	low	no	no
HCC2998	CHRX	109247278	G	T	TMEM164	DAMAGING	DAMAGING	medium	no	no
HCC2998	CHRX	77243968	A	C	ATP7A	DAMAGING	-	low	no	yes
HCT116	CHR10	102738986	G	A	SEMA4G	-	-	neutral	no	no
HCT116	CHR10	105798278	A	T	COL17A1	-	-	low	no	no
HCT116	CHR10	125804291	C	T	CHST15	-	-	neutral	no	no
HCT116	CHR1	109793495	G	A	CELSR2	DAMAGING	DAMAGING	medium	no	no
HCT116	CHR1	109794814	G	A	CELSR2	DAMAGING	DAMAGING	neutral	no	no
HCT116	CHR1	109814024	T	A	CELSR2	DAMAGING	-	low	no	no
HCT116	CHR11	113563829	C	T	TMPRSS5	DAMAGING	DAMAGING	medium	no	no
HCT116	CHR11	125830892	G	A	CDON	-	-	medium	no	no
HCT116	CHR1	11346051	T	G	UBIAD1	DAMAGING	-	low	no	no
HCT116	CHR11	36248721	G	A	LDLRAD3	-	-	neutral	no	no
HCT116	CHR11	3838610	G	A	PGAP2	-	-	neutral	no	no
HCT116	CHR11	4112916	G	A	STIM1	-	-	neutral	no	no
HCT116	CHR1	158057650	G	T	KIRREL	DAMAGING	DAMAGING	medium	no	no
HCT116	CHR11	62655936	G	A	SLC3A2	DAMAGING	DAMAGING	medium	no	no
HCT116	CHR1	16458244	C	T	EPHA2	DAMAGING	DAMAGING	neutral	yes	yes
HCT116	CHR11	76370985	A	C	LRRC32	-	-	high	no	no
HCT116	CHR12	12334095	A	G	LRP6	DAMAGING	DAMAGING	high	no	no
HCT116	CHR1	223178359	A	C	DISP1	-	-	low	no	no
HCT116	CHR1	226050192	C	T	TMEM63A	DAMAGING	DAMAGING	medium	no	no
HCT116	CHR1	24483983	C	A	IL28RA	DAMAGING	-	neutral	no	no
HCT116	CHR12	49165675	C	T	ADCY6	DAMAGING	-	medium	no	no
HCT116	CHR12	56121015	C	T	CD63	-	-	low	no	no
HCT116	CHR12	56349328	C	T	PMEL	-	DAMAGING	low	no	no
HCT116	CHR12	56481853	C	T	ERBB3	-	-	-	yes	yes
HCT116	CHR12	57581173	G	A	LRP1	-	DAMAGING	medium	no	yes
HCT116	CHR12	7287939	C	T	CLSTN3	DAMAGING	DAMAGING	low	no	no
HCT116	CHR12	7303172	C	A	CLSTN3	DAMAGING	DAMAGING	medium	no	no
HCT116	CHR1	27427705	C	T	SLC9A1	DAMAGING	-	medium	no	yes
HCT116	CHR12	7526008	G	A	CD163L1	-	-	low	no	no
HCT116	CHR13	113481084	C	T	ATP11A	DAMAGING	DAMAGING	high	no	no
HCT116	CHR14	24788331	A	G	ADCY4	DAMAGING	DAMAGING	low	no	no
HCT116	CHR1	45125892	C	T	TMEM53	DAMAGING	DAMAGING	medium	no	no
HCT116	CHR14	58599878	C	T	C14ORF37	-	-	-	no	no
HCT116	CHR1	46650979	A	C	TSPAN1	DAMAGING	DAMAGING	medium	no	no
HCT116	CHR14	77706926	T	C	TMEM63C	-	-	low	no	no
HCT116	CHR15	40661500	A	T	DISP2	DAMAGING	DAMAGING	medium	no	no
HCT116	CHR15	40662233	C	T	DISP2	-	-	neutral	no	no
HCT116	CHR15	73590836	C	T	NEO1	DAMAGING	-	medium	no	no
HCT116	CHR15	75137864	C	T	SCAMP2	DAMAGING	-	low	no	no
HCT116	CHR15	79614369	A	T	TMED3	DAMAGING	-	medium	no	no
HCT116	CHR15	95022245	G	T	MCTP2	-	-	low	no	no
HCT116	CHR16	16101760	C	A	ABCC1	DAMAGING	DAMAGING	medium	no	yes
HCT116	CHR16	4042192	C	T	ADCY9	DAMAGING	DAMAGING	medium	no	no
HCT116	CHR16	50338385	C	T	ADCY7	-	-	medium	no	no
HCT116	CHR16	66600540	C	T	CMTM1	-	-	neutral	no	no
HCT116	CHR16	67289823	G	A	SLC9A5	-	-	medium	no	no
HCT116	CHR17	43214437	T	G	ACBD4	-	-	medium	no	yes
HCT116	CHR17	48765012	G	A	ABCC3	-	-	low	no	yes
HCT116	CHR17	60767623	G	T	MRC2	-	-	neutral	no	no

HCT116	CHR17	70845793	A	G	SLC39A11	DAMAGING	DAMAGING	high	no	no
HCT116	CHR18	13682013	T	C	C18ORF19	-	-	neutral	no	no
HCT116	CHR19	1042802	G	A	ABCA7	-	-	neutral	no	no
HCT116	CHR1	92185542	G	A	TGFBR3	-	DAMAGING	low	no	no
HCT116	CHR19	36230811	G	A	IGFLR1	-	-	neutral	no	no
HCT116	CHR19	38834946	A	G	CATSPERG	-	DAMAGING	medium	no	no
HCT116	CHR19	41758310	T	A	AXL	DAMAGING	-	high	yes	no
HCT116	CHR19	4550898	G	A	SEMA6B	-	DAMAGING	medium	no	no
HCT116	CHR19	48836576	C	T	TMEM143	-	-	neutral	no	no
HCT116	CHR19	49246702	C	T	IZUMO1	-	-	low	no	no
HCT116	CHR1	95001617	G	A	F3	DAMAGING	DAMAGING	medium	no	yes
HCT116	CHR20	3214917	C	T	SLC4A11	-	DAMAGING	medium	no	no
HCT116	CHR20	3564702	G	T	ATRN	-	-	low	no	no
HCT116	CHR20	35835700	G	A	RPN2	-	-	low	no	no
HCT116	CHR20	50290777	T	C	ATP9A	DAMAGING	-	low	no	no
HCT116	CHR2	102793131	A	G	IL1R1	-	-	neutral	no	yes
HCT116	CHR2	120003091	A	G	STEAP3	-	-	low	no	no
HCT116	CHR21	45798938	G	A	TRPM2	-	-	medium	no	no
HCT116	CHR22	20127112	G	A	ZDHHC8	-	-	high	no	no
HCT116	CHR22	24224959	C	T	SLC2A11	DAMAGING	-	medium	no	no
HCT116	CHR2	230910754	T	C	SLC16A14	-	DAMAGING	low	no	no
HCT116	CHR2	230911291	G	A	SLC16A14	-	-	low	no	no
HCT116	CHR2	8871397	G	A	KIDINS220	-	DAMAGING	low	no	no
HCT116	CHR2	97464925	C	T	CNNM4	-	-	-	no	no
HCT116	CHR3	126748855	G	T	PLXNA1	DAMAGING	DAMAGING	medium	no	no
HCT116	CHR3	19575058	C	T	KCNH8	-	-	low	no	no
HCT116	CHR3	37670700	G	T	ITGA9	DAMAGING	-	low	no	no
HCT116	CHR3	48696585	G	C	CELSR3	DAMAGING	DAMAGING	high	no	no
HCT116	CHR3	49568662	G	A	DAG1	-	DAMAGING	medium	no	no
HCT116	CHR3	53700457	G	T	CACNA1D	DAMAGING	DAMAGING	high	no	yes
HCT116	CHR3	57743437	G	A	SLMAP	DAMAGING	DAMAGING	medium	no	no
HCT116	CHR5	140215961	G	A	PCDHA7	DAMAGING	-	medium	no	no
HCT116	CHR5	140750794	T	C	PCDHGB3	-	-	neutral	no	no
HCT116	CHR5	140753844	G	T	PCDHGA6	DAMAGING	-	high	no	no
HCT116	CHR5	140773957	G	A	PCDHGA8	-	-	low	no	no
HCT116	CHR5	141248471	A	T	PCDH1	DAMAGING	DAMAGING	medium	no	no
HCT116	CHR5	141248764	C	A	PCDH1	-	-	medium	no	no
HCT116	CHR5	159344255	C	T	ADRA1B	DAMAGING	DAMAGING	medium	no	yes
HCT116	CHR5	36677150	A	G	SLC1A3	DAMAGING	DAMAGING	medium	no	yes
HCT116	CHR6	26459903	C	T	BTN2A1	-	-	-	no	no
HCT116	CHR6	26468259	A	G	BTN2A1	-	-	low	no	no
HCT116	CHR7	100226953	G	A	TFR2	DAMAGING	DAMAGING	low	no	no
HCT116	CHR7	150767407	G	A	SLC4A2	-	-	neutral	no	no
HCT116	CHR7	47970788	G	A	PKD1L1	-	-	neutral	no	no
HCT116	CHR7	99474469	A	T	OR2AE1	DAMAGING	DAMAGING	high	no	no
HCT116	CHR8	125487484	G	A	RNF139	-	-	neutral	no	no
HCT116	CHR8	145640451	G	T	SLC39A4	-	-	-	no	no
HCT116	CHR8	38279368	G	A	FGFR1	-	-	low	yes	yes
HCT116	CHR8	74209486	C	T	RDH10	-	-	low	no	no
HCT116	CHR9	113704219	G	A	LPAR1	DAMAGING	DAMAGING	high	no	no
HCT116	CHR9	92020267	T	A	SEMA4D	-	-	low	no	no
HCT116	CHRX	109416555	C	T	TMEM164	-	-	neutral	no	no
HCT15	CHR10	102796287	C	T	SFXN3	-	-	low	no	no
HCT15	CHR10	103609579	T	C	C10ORF76	DAMAGING	DAMAGING	medium	no	no
HCT15	CHR10	104836858	C	T	CNNM2	-	-	low	no	no
HCT15	CHR10	105207018	C	T	CALHM2	-	-	medium	no	no
HCT15	CHR10	15256331	T	C	FAM171A1	-	-	low	no	no
HCT15	CHR10	47087098	C	A	PPYR1	DAMAGING	DAMAGING	low	no	yes
HCT15	CHR10	47087852	G	C	PPYR1	-	-	low	no	yes
HCT15	CHR10	95079703	C	A	MYOF	DAMAGING	DAMAGING	high	no	no
HCT15	CHR10	95121223	G	A	MYOF	DAMAGING	DAMAGING	medium	no	no
HCT15	CHR10	95168557	G	A	MYOF	DAMAGING	DAMAGING	medium	no	no
HCT15	CHR1	109735299	G	A	KIAA1324	DAMAGING	DAMAGING	medium	no	no

HCT15	CHR1	109793815	C	A	CELSR2	-	-	low	no	no
HCT15	CHR1	109794148	C	A	CELSR2	DAMAGING	DAMAGING	low	no	no
HCT15	CHR1	109803885	T	C	CELSR2	-	-	medium	no	no
HCT15	CHR1	110464558	G	T	CSF1	-	-	low	no	yes
HCT15	CHR11	113102911	C	A	NCAM1	-	-	-	no	no
HCT15	CHR11	119181103	G	T	MCAM	DAMAGING	DAMAGING	low	no	no
HCT15	CHR11	120200994	G	T	TMEM136	-	-	-	no	no
HCT15	CHR11	121429379	C	A	SORL1	DAMAGING	-	medium	no	no
HCT15	CHR11	121440952	G	A	SORL1	-	-	medium	no	no
HCT15	CHR11	124747068	T	C	ROBO3	DAMAGING	DAMAGING	low	no	no
HCT15	CHR11	129722528	G	A	TMEM45B	-	-	low	no	no
HCT15	CHR11	130079714	T	C	ST14	DAMAGING	-	low	no	yes
HCT15	CHR11	130784232	T	C	SNX19	-	-	low	no	no
HCT15	CHR1	113459995	T	C	SLC16A1	-	-	neutral	no	yes
HCT15	CHR1	116233809	C	T	VANGL1	-	-	-	no	no
HCT15	CHR1	116605413	A	T	SLC22A15	-	-	neutral	no	no
HCT15	CHR1	117127352	C	A	IGSF3	-	-	low	no	no
HCT15	CHR1	117509698	C	T	PTGFRN	DAMAGING	DAMAGING	low	no	no
HCT15	CHR1	120497754	C	T	NOTCH2	-	DAMAGING	low	no	yes
HCT15	CHR11	36119975	G	A	LDLRAD3	-	DAMAGING	medium	no	no
HCT15	CHR11	3846353	C	T	PGAP2	-	-	low	no	no
HCT15	CHR11	3988796	G	A	STIM1	-	DAMAGING	low	no	no
HCT15	CHR11	407948	C	G	SIGIRR	-	-	neutral	no	no
HCT15	CHR11	4112826	C	T	STIM1	-	-	neutral	no	no
HCT15	CHR11	46900496	C	A	LRP4	DAMAGING	DAMAGING	medium	no	no
HCT15	CHR11	47436384	G	T	SLC39A13	-	DAMAGING	medium	no	no
HCT15	CHR1	153948354	C	A	JTB	-	-	-	no	no
HCT15	CHR1	155162039	C	A	MUC1	-	DAMAGING	neutral	no	yes
HCT15	CHR1	156216025	C	A	PAQR6	-	-	medium	no	no
HCT15	CHR11	57258830	A	C	SLC43A1	-	-	low	no	no
HCT15	CHR11	57268302	G	T	SLC43A1	DAMAGING	DAMAGING	low	no	no
HCT15	CHR1	158061202	G	A	KIRREL	DAMAGING	DAMAGING	medium	no	no
HCT15	CHR1	160389346	G	T	VANGL2	-	-	medium	no	no
HCT15	CHR11	62655873	G	A	SLC3A2	DAMAGING	DAMAGING	medium	no	no
HCT15	CHR1	16458741	G	A	EPHA2	-	-	-	yes	yes
HCT15	CHR1	16475436	T	C	EPHA2	-	-	medium	yes	yes
HCT15	CHR1	165532813	C	G	LRRC52	-	-	medium	no	no
HCT15	CHR1	169446700	A	G	SLC19A2	-	-	low	no	yes
HCT15	CHR11	73101936	G	A	RELT	-	-	neutral	no	no
HCT15	CHR1	17332225	G	A	ATP13A2	-	-	neutral	no	no
HCT15	CHR1	19166988	T	A	TAS1R2	-	-	medium	no	yes
HCT15	CHR1	19175961	G	C	TAS1R2	-	-	low	no	yes
HCT15	CHR11	93913399	G	A	PANX1	-	-	low	no	yes
HCT15	CHR1	208234098	C	T	PLXNA2	-	-	low	no	no
HCT15	CHR12	118506289	C	G	VSIG10	-	-	neutral	no	no
HCT15	CHR12	12278291	G	A	LRP6	-	-	low	no	no
HCT15	CHR12	125302247	G	A	SCARB1	DAMAGING	-	medium	no	yes
HCT15	CHR12	129283823	C	T	SLC15A4	-	-	neutral	no	no
HCT15	CHR12	1893145	G	A	ADIPOR2	-	-	-	no	no
HCT15	CHR12	26572040	G	A	ITPR2	-	-	low	no	no
HCT15	CHR12	26816653	G	T	ITPR2	DAMAGING	DAMAGING	medium	no	no
HCT15	CHR1	227069634	G	A	PSEN2	DAMAGING	-	low	no	yes
HCT15	CHR1	246810865	G	T	CNST	-	-	medium	no	no
HCT15	CHR12	49742912	T	C	DNAJC22	-	-	low	no	no
HCT15	CHR12	51868929	C	T	SLC4A8	DAMAGING	-	medium	no	no
HCT15	CHR12	52377808	G	A	ACVR1B	-	-	-	yes	yes
HCT15	CHR12	52377857	C	G	ACVR1B	-	-	low	yes	yes
HCT15	CHR12	52385722	T	A	ACVR1B	DAMAGING	DAMAGING	medium	yes	yes
HCT15	CHR12	56350787	G	A	PMEL	-	-	neutral	no	no
HCT15	CHR12	56478922	C	A	ERBB3	-	-	medium	yes	yes
HCT15	CHR12	56489535	G	A	ERBB3	DAMAGING	DAMAGING	medium	yes	yes
HCT15	CHR12	56495068	C	A	ERBB3	-	DAMAGING	neutral	yes	yes
HCT15	CHR12	56536498	G	A	ESYT1	-	DAMAGING	low	no	no

HCT15	CHR12	57552211	T	C	LRP1	DAMAGING	DAMAGING	medium	no	yes
HCT15	CHR12	57573228	G	T	LRP1	-	-	low	no	yes
HCT15	CHR12	7302164	G	A	CLSTN3	-	-	low	no	no
HCT15	CHR1	27427020	T	G	SLC9A1	-	-	neutral	no	yes
HCT15	CHR1	27429017	T	C	SLC9A1	-	-	low	no	yes
HCT15	CHR1	29650155	C	T	PTPRU	-	-	neutral	no	no
HCT15	CHR13	114150033	G	T	TMCO3	-	-	neutral	no	no
HCT15	CHR13	114202628	C	T	TMCO3	DAMAGING	-	low	no	no
HCT15	CHR13	39262883	T	G	FREM2	DAMAGING	-	low	no	no
HCT15	CHR13	39262976	A	G	FREM2	-	-	neutral	no	no
HCT15	CHR13	39358842	G	C	FREM2	DAMAGING	DAMAGING	high	no	no
HCT15	CHR13	52542743	C	T	ATP7B	DAMAGING	DAMAGING	high	no	no
HCT15	CHR13	52548178	G	T	ATP7B	-	-	medium	no	no
HCT15	CHR13	75873578	G	T	TBC1D4	DAMAGING	DAMAGING	high	no	no
HCT15	CHR1	38227575	C	T	EPHA10	-	-	neutral	yes	no
HCT15	CHR1	43396500	C	T	SLC2A1	DAMAGING	-	medium	no	no
HCT15	CHR14	45605562	T	C	FANCM	DAMAGING	DAMAGING	medium	no	no
HCT15	CHR1	45120593	G	T	TMEM53	-	DAMAGING	low	no	no
HCT15	CHR14	77709313	T	G	TMEM63C	-	-	neutral	no	no
HCT15	CHR14	95662892	C	T	CLMN	-	-	neutral	no	no
HCT15	CHR14	96707737	T	C	BDKRB2	-	-	low	no	yes
HCT15	CHR15	34657280	A	T	LPCAT4	DAMAGING	DAMAGING	medium	no	no
HCT15	CHR15	41870266	C	T	TYRO3	-	-	neutral	yes	no
HCT15	CHR15	63631123	G	T	CA12	DAMAGING	DAMAGING	medium	no	yes
HCT15	CHR15	99465549	A	G	IGF1R	-	-	medium	yes	yes
HCT15	CHR16	16208844	G	T	ABCC1	DAMAGING	DAMAGING	medium	no	yes
HCT15	CHR16	16232381	G	A	ABCC1	DAMAGING	DAMAGING	low	no	yes
HCT15	CHR16	19483425	C	A	TMC5	-	-	neutral	no	no
HCT15	CHR16	2373553	T	C	ABCA3	-	-	medium	no	no
HCT15	CHR16	2376256	G	A	ABCA3	-	DAMAGING	medium	no	no
HCT15	CHR16	2569335	G	A	ATP6V0C	DAMAGING	DAMAGING	high	no	no
HCT15	CHR16	50327332	G	A	ADCY7	DAMAGING	-	low	no	no
HCT15	CHR16	57693321	G	A	GPR56	-	-	neutral	no	no
HCT15	CHR16	57693399	T	C	GPR56	DAMAGING	DAMAGING	medium	no	no
HCT15	CHR16	67979022	C	T	SLC12A4	DAMAGING	-	medium	no	yes
HCT15	CHR16	67988649	C	T	SLC12A4	DAMAGING	DAMAGING	medium	no	yes
HCT15	CHR16	68710372	A	G	CDH3	-	-	low	no	no
HCT15	CHR16	88801377	G	A	FAM38A	DAMAGING	-	medium	no	no
HCT15	CHR17	1640879	C	A	WDR81	DAMAGING	DAMAGING	medium	no	no
HCT15	CHR17	26687771	G	T	TMEM199	-	DAMAGING	low	no	no
HCT15	CHR17	30376225	T	C	LRRC37B	-	-	medium	no	no
HCT15	CHR17	38080399	G	A	ORMDL3	DAMAGING	DAMAGING	medium	no	no
HCT15	CHR17	40734031	C	T	FAM134C	-	-	neutral	no	no
HCT15	CHR17	42153062	C	T	G6PC3	DAMAGING	DAMAGING	medium	no	no
HCT15	CHR17	42266652	C	A	TMUB2	-	-	neutral	no	no
HCT15	CHR17	48470176	C	T	LRRC59	-	-	neutral	no	no
HCT15	CHR17	56439952	G	T	RNF43	DAMAGING	-	low	no	no
HCT15	CHR17	66267544	T	C	SLC16A6	-	-	neutral	no	yes
HCT15	CHR17	71080954	G	T	SLC39A11	DAMAGING	DAMAGING	high	no	no
HCT15	CHR17	7165273	C	T	CLDN7	-	-	-	no	no
HCT15	CHR17	73486414	G	T	KIAA0195	-	-	low	no	no
HCT15	CHR17	7460209	T	C	TNFSF12	DAMAGING	DAMAGING	low	no	no
HCT15	CHR18	71816078	C	T	C18ORF55	-	-	low	no	no
HCT15	CHR19	1005405	G	T	GRIN3B	DAMAGING	DAMAGING	medium	no	yes
HCT15	CHR19	1051287	C	A	ABCA7	-	-	medium	no	no
HCT15	CHR19	11408976	G	A	TSPAN16	-	-	-	no	no
HCT15	CHR19	16664645	G	A	SLC35E1	-	-	neutral	no	no
HCT15	CHR19	19765505	C	T	ATP13A1	-	-	low	no	no
HCT15	CHR19	3542949	C	T	C19ORF28	-	-	neutral	no	no
HCT15	CHR19	35646499	A	G	FXYD5	DAMAGING	-	low	no	no
HCT15	CHR19	38850118	A	G	CATSPERG	-	-	neutral	no	no
HCT15	CHR19	38857901	T	G	CATSPERG	DAMAGING	-	medium	no	no
HCT15	CHR19	38861248	C	T	CATSPERG	DAMAGING	-	neutral	no	no

HCT15	CHR19	41762398	T	C	AXL	DAMAGING	DAMAGING	high	yes	no
HCT15	CHR19	49464127	C	A	BAX	-	-	neutral	no	no
HCT15	CHR19	49713557	C	A	TRPM4	DAMAGING	DAMAGING	medium	no	no
HCT15	CHR19	55693185	C	A	PTPRH	DAMAGING	-	neutral	no	no
HCT15	CHR19	55693220	C	T	PTPRH	-	-	neutral	no	no
HCT15	CHR19	7174706	C	A	INSR	DAMAGING	-	medium	yes	yes
HCT15	CHR19	7184528	A	G	INSR	-	DAMAGING	medium	yes	yes
HCT15	CHR19	7965515	A	G	LRRC8E	-	-	neutral	no	no
HCT15	CHR20	10633240	C	T	JAG1	-	-	-	no	no
HCT15	CHR20	14307053	G	A	FLRT3	-	-	low	no	no
HCT15	CHR20	30136867	G	T	HM13	-	-	low	no	no
HCT15	CHR20	4864322	C	T	SLC23A2	-	-	low	no	yes
HCT15	CHR20	49196319	G	A	PTPN1	-	-	low	no	yes
HCT15	CHR20	50287767	A	G	ATP9A	-	-	neutral	no	no
HCT15	CHR20	50346505	C	T	ATP9A	-	-	-	no	no
HCT15	CHR2	103324788	C	T	SLC9A2	-	-	neutral	no	no
HCT15	CHR21	27423425	C	T	APP	DAMAGING	DAMAGING	medium	no	yes
HCT15	CHR21	34635390	C	T	IFNAR2	-	-	neutral	no	yes
HCT15	CHR21	45656879	G	A	ICOSLG	DAMAGING	DAMAGING	low	no	no
HCT15	CHR21	45786740	G	T	TRPM2	DAMAGING	DAMAGING	medium	no	no
HCT15	CHR21	45825109	G	A	TRPM2	-	-	medium	no	no
HCT15	CHR2	187466795	C	A	ITGAV	-	-	medium	no	yes
HCT15	CHR2	219205490	C	A	PNKD	DAMAGING	DAMAGING	medium	no	no
HCT15	CHR2	220087443	G	T	ATG9A	-	-	low	no	no
HCT15	CHR22	20126806	C	T	ZDHHC8	DAMAGING	DAMAGING	medium	no	no
HCT15	CHR2	227779053	G	A	RHBDD1	-	DAMAGING	medium	no	no
HCT15	CHR22	47022779	C	T	GRAMD4	DAMAGING	DAMAGING	low	no	no
HCT15	CHR22	50942315	C	A	LMF2	-	-	low	no	no
HCT15	CHR2	25141496	G	T	ADCY3	DAMAGING	-	medium	no	no
HCT15	CHR2	26998021	A	T	C2ORF18	DAMAGING	-	medium	no	no
HCT15	CHR2	27001159	G	A	C2ORF18	DAMAGING	DAMAGING	medium	no	no
HCT15	CHR2	27426107	C	A	SLC5A6	-	-	low	no	yes
HCT15	CHR2	86071595	G	A	ST3GAL5	DAMAGING	DAMAGING	medium	no	no
HCT15	CHR2	97529491	C	A	SEMA4C	-	DAMAGING	low	no	no
HCT15	CHR3	105252484	G	A	ALCAM	DAMAGING	DAMAGING	medium	no	no
HCT15	CHR3	121725969	C	T	ILDR1	DAMAGING	DAMAGING	low	no	no
HCT15	CHR3	126708506	T	G	PLXNA1	DAMAGING	DAMAGING	medium	no	no
HCT15	CHR3	129289658	C	A	PLXND1	DAMAGING	-	low	no	no
HCT15	CHR3	129289968	T	C	PLXND1	-	-	medium	no	no
HCT15	CHR3	19432129	T	C	KCNH8	-	-	low	no	no
HCT15	CHR3	196387040	G	A	LRRC33	-	-	neutral	no	no
HCT15	CHR3	30715697	T	C	TGFBR2	DAMAGING	DAMAGING	medium	yes	yes
HCT15	CHR3	38520668	T	C	ACVR2B	DAMAGING	DAMAGING	low	no	no
HCT15	CHR3	48666074	C	A	SLC26A6	-	-	low	no	no
HCT15	CHR3	48691087	C	A	CELSR3	-	-	neutral	no	no
HCT15	CHR3	49153201	G	T	USP19	-	-	low	no	no
HCT15	CHR3	49569080	C	T	DAG1	DAMAGING	DAMAGING	medium	no	no
HCT15	CHR3	49570030	C	A	DAG1	-	-	low	no	no
HCT15	CHR3	49570279	C	T	DAG1	DAMAGING	DAMAGING	medium	no	no
HCT15	CHR3	49933245	C	A	MST1R	-	-	medium	yes	yes
HCT15	CHR3	49936038	G	T	MST1R	-	-	-	yes	yes
HCT15	CHR3	53157825	G	T	RFT1	-	-	medium	no	no
HCT15	CHR3	53757914	T	C	CACNA1D	DAMAGING	DAMAGING	high	no	yes
HCT15	CHR3	62142767	G	A	PTPRG	-	-	neutral	no	no
HCT15	CHR3	62177218	C	T	PTPRG	DAMAGING	DAMAGING	low	no	no
HCT15	CHR3	66436657	G	A	LRIG1	DAMAGING	-	medium	no	no
HCT15	CHR3	9974300	A	G	IL17RC	-	-	neutral	no	no
HCT15	CHR4	110972565	G	T	ELOVL6	DAMAGING	-	medium	no	no
HCT15	CHR4	110972736	C	T	ELOVL6	DAMAGING	-	medium	no	no
HCT15	CHR4	170913205	G	A	MFAP3L	-	-	low	no	no
HCT15	CHR4	1816168	C	T	LETM1	-	-	medium	no	no
HCT15	CHR4	1824845	T	C	LETM1	-	-	neutral	no	no
HCT15	CHR4	1836606	C	T	LETM1	-	-	neutral	no	no

HCT15	CHR4	187521417	G	A	FAT1	-	-	neutral	no	no
HCT15	CHR4	30724591	C	G	PCDH7	DAMAGING	DAMAGING	medium	no	no
HCT15	CHR4	42466966	T	C	ATP8A1	DAMAGING	-	medium	no	yes
HCT15	CHR4	47570940	A	C	ATP10D	-	-	neutral	no	no
HCT15	CHR4	6303369	C	T	WFS1	-	-	low	no	no
HCT15	CHR4	79188420	G	T	FRAS1	DAMAGING	-	high	no	no
HCT15	CHR4	79393458	G	A	FRAS1	-	-	medium	no	no
HCT15	CHR5	1318506	G	A	CLPTM1L	-	-	low	no	no
HCT15	CHR5	140215031	C	A	PCDHA7	DAMAGING	-	medium	no	no
HCT15	CHR5	140250915	G	A	PCDHA11	-	-	medium	no	no
HCT15	CHR5	140262472	G	A	PCDHA13	-	-	low	no	no
HCT15	CHR5	140347684	G	A	PCDHAC2	DAMAGING	DAMAGING	high	no	no
HCT15	CHR5	140719131	G	T	PCDHGA2	DAMAGING	-	medium	no	no
HCT15	CHR5	140720117	G	A	PCDHGA2	-	-	low	no	no
HCT15	CHR5	140740619	A	T	PCDHGB2	-	-	neutral	no	no
HCT15	CHR5	140753671	T	A	PCDHGA6	-	-	neutral	no	no
HCT15	CHR5	140762912	G	T	PCDHGA7	DAMAGING	-	high	no	no
HCT15	CHR5	140810777	C	T	PCDHGA12	DAMAGING	-	high	no	no
HCT15	CHR5	140869595	G	A	PCDHGC5	-	-	low	no	no
HCT15	CHR5	140870720	A	G	PCDHGC5	DAMAGING	-	medium	no	no
HCT15	CHR5	150838393	T	A	SLC36A1	-	-	neutral	no	yes
HCT15	CHR5	179393916	G	A	RNF130	-	-	neutral	no	no
HCT15	CHR5	36608541	G	A	SLC1A3	-	-	neutral	no	yes
HCT15	CHR5	76129322	T	C	F2RL1	DAMAGING	DAMAGING	medium	no	no
HCT15	CHR5	76129424	G	A	F2RL1	DAMAGING	DAMAGING	medium	no	no
HCT15	CHR5	96314958	T	C	LNPEP	DAMAGING	DAMAGING	low	no	no
HCT15	CHR6	160492969	G	A	IGF2R	DAMAGING	DAMAGING	medium	no	yes
HCT15	CHR6	170155446	T	A	C6ORF70	DAMAGING	DAMAGING	medium	no	no
HCT15	CHR6	26468604	C	T	BTN2A1	DAMAGING	DAMAGING	medium	no	no
HCT15	CHR6	29694749	C	A	HLA-F	-	-	low	no	no
HCT15	CHR6	44116103	A	G	TMEM63B	-	-	neutral	no	no
HCT15	CHR7	100173498	G	A	LRCH4	DAMAGING	DAMAGING	medium	no	no
HCT15	CHR7	100463406	C	A	SLC12A9	-	-	medium	no	no
HCT15	CHR7	105636785	T	A	CDHR3	-	-	neutral	no	no
HCT15	CHR7	128445829	T	G	CCDC136	-	-	low	no	no
HCT15	CHR7	128802333	C	A	TSPAN33	DAMAGING	DAMAGING	low	no	no
HCT15	CHR7	131194268	C	A	PODXL	-	-	neutral	no	no
HCT15	CHR7	143091961	C	A	EPHA1	DAMAGING	DAMAGING	high	yes	no
HCT15	CHR7	143095114	T	G	EPHA1	DAMAGING	-	low	yes	no
HCT15	CHR7	147844694	G	T	CNTNAP2	DAMAGING	DAMAGING	high	no	no
HCT15	CHR7	155100034	C	A	INSIG1	-	-	neutral	no	no
HCT15	CHR7	90895145	G	A	FZD1	-	-	low	no	no
HCT15	CHR7	99956639	A	G	PILRB	-	DAMAGING	low	no	no
HCT15	CHR7	99987615	T	G	PILRA	-	-	low	no	no
HCT15	CHR8	145140286	C	T	GPAA1	-	DAMAGING	low	no	no
HCT15	CHR8	145584086	C	A	GPR172A	DAMAGING	DAMAGING	medium	no	no
HCT15	CHR8	27509144	G	T	SCARA3	-	DAMAGING	low	no	no
HCT15	CHR8	38282161	C	A	FGFR1	DAMAGING	-	neutral	yes	yes
HCT15	CHR9	111795623	G	A	C9ORF5	DAMAGING	DAMAGING	neutral	no	no
HCT15	CHR9	114840940	C	T	SUSD1	DAMAGING	-	neutral	no	no
HCT15	CHR9	119413904	C	T	ASTN2	-	-	neutral	no	no
HCT15	CHR9	130169493	C	A	SLC2A8	DAMAGING	DAMAGING	medium	no	no
HCT15	CHR9	131670378	A	G	LRRC8A	-	-	low	no	no
HCT15	CHR9	2641415	G	A	VLDLR	-	-	neutral	no	no
HCT15	CHR9	34635670	G	A	SIGMAR1	-	-	neutral	no	yes
HCT15	CHR9	86924601	T	C	SLC28A3	-	-	low	no	yes
HCT15	CHR9	91978345	C	T	SEMA4D	-	-	-	no	no
HCT15	CHRX	153132187	G	T	L1CAM	DAMAGING	DAMAGING	medium	no	no
HCT15	CHRX	37586986	A	C	XK	-	-	low	no	no
HCT15	CHRX	48318212	C	T	SLC38A5	-	-	-	no	no
HCT15	CHRX	77264669	T	C	ATP7A	-	-	medium	no	yes
HT29	CHR12	117187658	G	T	RNFT2	-	-	low	no	no
HT29	CHR1	223178961	G	A	DISP1	-	-	low	no	no

HT29	CHR12	56092319	C	T	ITGA7	DAMAGING	DAMAGING	medium	no	no
HT29	CHR16	88801402	C	T	FAM38A	-	-	neutral	no	no
HT29	CHR2	95947686	C	T	PROM2	DAMAGING	-	medium	no	no
HT29	CHR5	140215257	G	A	PCDHA7	-	-	neutral	no	no
HT29	CHR5	140740472	C	T	PCDHGB2	-	-	low	no	no
HT29	CHR5	140774391	G	C	PCDHGA8	-	-	low	no	no
HT29	CHR7	100404997	G	A	EPHB4	DAMAGING	DAMAGING	medium	yes	yes
KM12	CHR10	104228871	C	A	TMEM180	-	DAMAGING	low	no	no
KM12	CHR10	129867964	T	C	PTPRE	DAMAGING	DAMAGING	medium	no	yes
KM12	CHR1	109793324	C	A	CELSR2	DAMAGING	DAMAGING	medium	no	no
KM12	CHR1	110466805	G	T	CSF1	-	DAMAGING	medium	no	yes
KM12	CHR11	117164643	C	T	BACE1	-	DAMAGING	low	no	yes
KM12	CHR1	117142946	G	A	IGSF3	-	DAMAGING	low	no	no
KM12	CHR1	117504248	G	A	PTGFRN	DAMAGING	-	low	no	no
KM12	CHR1	11888183	C	T	CLCN6	-	-	-	no	yes
KM12	CHR1	120458899	G	A	NOTCH2	DAMAGING	DAMAGING	medium	no	yes
KM12	CHR1	120468105	C	T	NOTCH2	-	-	neutral	no	yes
KM12	CHR11	5475494	C	T	OR51I2	-	DAMAGING	low	no	no
KM12	CHR1	155221343	C	T	FAM189B	DAMAGING	DAMAGING	low	no	no
KM12	CHR1	156339125	C	T	RHBG	DAMAGING	-	medium	no	no
KM12	CHR11	57456067	G	A	ZDHHC5	-	-	-	no	no
KM12	CHR1	160064880	G	A	IGSF8	-	-	neutral	no	no
KM12	CHR1	160388883	T	C	VANGL2	-	-	low	no	no
KM12	CHR11	61120500	C	T	CYBASC3	-	-	medium	no	no
KM12	CHR11	61632664	G	A	FADS2	-	DAMAGING	medium	no	yes
KM12	CHR11	61632694	C	T	FADS2	-	-	neutral	no	yes
KM12	CHR11	63342488	C	T	PLA2G16	DAMAGING	-	low	no	no
KM12	CHR1	165532832	C	T	LRRC52	-	-	low	no	no
KM12	CHR11	66134962	C	T	SLC29A2	-	-	low	no	no
KM12	CHR11	68171055	G	A	LRP5	-	-	-	no	no
KM12	CHR11	68839463	G	A	TPCN2	-	-	neutral	no	no
KM12	CHR12	118520169	C	T	VSIG10	-	-	low	no	no
KM12	CHR1	246811121	C	T	CNST	DAMAGING	-	low	no	no
KM12	CHR12	4919564	G	T	KCNA6	-	-	neutral	no	no
KM12	CHR12	50357920	C	T	AQP5	-	-	medium	no	yes
KM12	CHR12	52370122	G	A	ACVR1B	-	-	neutral	yes	yes
KM12	CHR12	56536172	G	A	ESYT1	-	-	neutral	no	no
KM12	CHR12	6342605	G	A	CD9	DAMAGING	-	medium	no	no
KM12	CHR1	28293200	C	T	XKR8	-	DAMAGING	medium	no	no
KM12	CHR1	29644308	G	A	PTPRU	DAMAGING	-	medium	no	no
KM12	CHR1	32205164	C	T	BAI2	DAMAGING	DAMAGING	medium	no	no
KM12	CHR1	32542909	C	T	TMEM39B	DAMAGING	-	low	no	no
KM12	CHR13	39264464	T	G	FREM2	-	-	low	no	no
KM12	CHR1	35908603	C	T	KIAA0319L	DAMAGING	DAMAGING	medium	no	no
KM12	CHR1	39340468	G	A	GJA9	DAMAGING	-	neutral	no	no
KM12	CHR13	95886946	A	G	ABCC4	-	-	low	no	yes
KM12	CHR1	40434318	C	T	MFSD2A	-	-	neutral	no	no
KM12	CHR14	23312957	C	T	MMP14	DAMAGING	-	neutral	no	yes
KM12	CHR14	70171407	G	A	KIAA0247	DAMAGING	DAMAGING	medium	no	no
KM12	CHR14	94578041	G	A	IFI27	-	-	neutral	no	no
KM12	CHR14	95670218	C	T	CLMN	-	-	neutral	no	no
KM12	CHR15	63631125	G	A	CA12	-	-	low	no	yes
KM12	CHR15	65916511	G	A	SLC24A1	-	-	low	no	no
KM12	CHR15	90771576	C	T	SEMA4B	DAMAGING	DAMAGING	medium	no	no
KM12	CHR16	2376220	C	T	ABCA3	-	-	low	no	no
KM12	CHR16	4016669	C	T	ADCY9	DAMAGING	DAMAGING	neutral	no	no
KM12	CHR1	6524697	C	A	TNFRSF25	-	-	neutral	no	no
KM12	CHR16	68732277	A	G	CDH3	DAMAGING	-	medium	no	no
KM12	CHR16	85024191	C	T	ZDHHC7	-	DAMAGING	medium	no	no
KM12	CHR17	37866384	C	T	ERBB2	DAMAGING	-	medium	yes	yes
KM12	CHR17	4337456	T	C	SPNS3	DAMAGING	-	medium	no	no
KM12	CHR17	7318466	C	T	NLGN2	DAMAGING	-	medium	no	no
KM12	CHR17	7318864	G	A	NLGN2	DAMAGING	-	medium	no	no

KM12	CHR17	73482016	C	T	KIAA0195	DAMAGING	DAMAGING	neutral	no	no
KM12	CHR17	73486315	G	A	KIAA0195	-	-	low	no	no
KM12	CHR17	73488819	G	A	KIAA0195	-	-	low	no	no
KM12	CHR17	74469770	C	T	RHBDF2	-	-	low	no	no
KM12	CHR19	10395094	A	G	ICAM1	-	-	low	no	yes
KM12	CHR19	1049360	G	A	ABCA7	-	-	medium	no	no
KM12	CHR19	19737432	C	T	LPAR2	-	-	medium	no	no
KM12	CHR19	19766709	C	T	ATP13A1	DAMAGING	DAMAGING	low	no	no
KM12	CHR1	92224237	G	A	TGFBR3	DAMAGING	DAMAGING	medium	no	no
KM12	CHR19	42867254	G	A	MEGF8	-	-	low	no	no
KM12	CHR19	48836515	G	C	TMEM143	-	-	-	no	no
KM12	CHR19	54675679	G	A	TMC4	DAMAGING	DAMAGING	medium	no	no
KM12	CHR19	55708607	C	T	PTPRH	-	-	neutral	no	no
KM12	CHR20	3209557	G	A	SLC4A11	DAMAGING	DAMAGING	medium	no	no
KM12	CHR20	3627468	C	T	ATRN	-	-	neutral	no	no
KM12	CHR2	220032762	A	G	SLC23A3	-	DAMAGING	medium	no	no
KM12	CHR2	220047181	A	G	FAM134A	-	-	low	no	no
KM12	CHR2	220087012	C	T	ATG9A	-	-	low	no	no
KM12	CHR2	220496785	C	T	SLC4A3	-	-	low	no	no
KM12	CHR2	223423150	G	A	SGPP2	-	-	neutral	no	no
KM12	CHR22	46859778	G	A	CELSR1	DAMAGING	DAMAGING	medium	no	no
KM12	CHR2	675581	G	A	TMEM18	-	-	neutral	no	no
KM12	CHR2	97531449	A	G	SEMA4C	DAMAGING	DAMAGING	medium	no	no
KM12	CHR3	119177019	T	C	TMEM39A	-	-	medium	no	no
KM12	CHR3	125786983	C	T	SLC41A3	-	-	low	no	no
KM12	CHR3	129370532	A	T	TMCC1	DAMAGING	DAMAGING	medium	no	no
KM12	CHR3	14526442	C	A	SLC6A6	DAMAGING	-	low	no	no
KM12	CHR3	184295458	G	A	EPHB3	-	-	low	yes	no
KM12	CHR3	19575016	C	T	KCNH8	-	-	low	no	no
KM12	CHR3	30715703	T	C	TGFBR2	DAMAGING	DAMAGING	high	yes	yes
KM12	CHR3	48697208	C	T	CELSR3	DAMAGING	DAMAGING	medium	no	no
KM12	CHR3	49570391	C	T	DAG1	DAMAGING	DAMAGING	medium	no	no
KM12	CHR3	49756861	A	G	AMIGO3	-	-	low	no	no
KM12	CHR3	53764553	C	T	CACNA1D	-	-	low	no	yes
KM12	CHR4	187540637	G	A	FAT1	-	-	medium	no	no
KM12	CHR4	187549655	G	A	FAT1	-	-	neutral	no	no
KM12	CHR4	79351571	G	T	FRAS1	DAMAGING	DAMAGING	medium	no	no
KM12	CHR4	79400768	C	T	FRAS1	DAMAGING	-	medium	no	no
KM12	CHR5	138716294	G	A	SLC23A1	-	-	low	no	yes
KM12	CHR5	140216218	C	A	PCDHA7	-	-	neutral	no	no
KM12	CHR5	140250297	C	T	PCDHA11	-	-	low	no	no
KM12	CHR5	140256507	G	A	PCDHA12	-	-	neutral	no	no
KM12	CHR5	140347204	G	A	PCDHAC2	DAMAGING	DAMAGING	high	no	no
KM12	CHR5	140720342	G	A	PCDHGA2	DAMAGING	-	medium	no	no
KM12	CHR5	140774386	C	T	PCDHGA8	DAMAGING	-	medium	no	no
KM12	CHR5	140782901	G	A	PCDHGA9	DAMAGING	-	high	no	no
KM12	CHR5	140856299	C	T	PCDHGC3	DAMAGING	-	medium	no	no
KM12	CHR5	140868986	T	A	PCDHGC5	-	-	-	no	no
KM12	CHR5	72419784	C	T	TMEM171	-	-	low	no	no
KM12	CHR6	158014175	G	A	ZDHHC14	-	DAMAGING	low	no	no
KM12	CHR6	160464261	G	T	IGF2R	-	-	medium	no	yes
KM12	CHR6	160471623	G	A	IGF2R	-	-	medium	no	yes
KM12	CHR6	43098081	G	A	PTK7	-	-	low	yes	no
KM12	CHR7	100456752	C	T	SLC12A9	DAMAGING	DAMAGING	low	no	no
KM12	CHR7	105660937	C	T	CDHR3	-	-	low	no	no
KM12	CHR7	147600710	T	C	CNTNAP2	DAMAGING	DAMAGING	high	no	no
KM12	CHR7	55231511	G	T	EGFR	-	-	-	yes	yes
KM12	CHR7	99474206	C	T	OR2AE1	-	-	medium	no	no
KM12	CHR8	145139357	G	T	GPAA1	-	-	neutral	no	no
KM12	CHR8	22269696	G	A	SLC39A14	DAMAGING	DAMAGING	low	no	yes
KM12	CHR8	38853825	G	A	TM2D2	-	-	neutral	no	no
KM12	CHR9	125797574	G	T	GPR21	-	-	low	no	no
KM12	CHR9	131708619	G	A	DOLK	DAMAGING	DAMAGING	medium	no	no

KM12	CHR9	74360108	G	A	TMEM2	-	DAMAGING	medium	no	no
KM12	CHR9	86955512	C	T	SLC28A3	-	-	low	no	yes
LIM1215	CHR11	124745090	G	T	ROBO3	-	-	low	no	no
LIM1215	CHR11	62656157	G	A	SLC3A2	-	-	neutral	no	no
LIM1215	CHR1	16458244	C	T	EPHA2	DAMAGING	DAMAGING	neutral	yes	yes
LIM1215	CHR1	19644220	G	A	PQLC2	DAMAGING	-	medium	no	no
LIM1215	CHR12	106633854	C	T	CKAP4	-	DAMAGING	medium	no	no
LIM1215	CHR12	56480408	C	T	ERBB3	-	-	medium	yes	yes
LIM1215	CHR12	57581173	G	A	LRP1	-	DAMAGING	medium	no	yes
LIM1215	CHR12	57588410	C	T	LRP1	-	-	medium	no	yes
LIM1215	CHR1	27436120	C	T	SLC9A1	DAMAGING	DAMAGING	medium	no	yes
LIM1215	CHR16	2143546	C	T	PKD1	-	-	low	no	no
LIM1215	CHR16	68716239	A	G	CDH3	-	-	low	no	no
LIM1215	CHR19	10385665	G	A	ICAM1	-	-	low	no	yes
LIM1215	CHR19	1051038	G	A	ABCA7	DAMAGING	DAMAGING	low	no	no
LIM1215	CHR1	9164542	G	A	GPR157	-	-	neutral	no	no
LIM1215	CHR19	32968478	C	T	DPY19L3	DAMAGING	DAMAGING	medium	no	no
LIM1215	CHR19	35741473	T	C	LSR	-	-	low	no	no
LIM1215	CHR19	36230811	G	A	IGFLR1	-	-	neutral	no	no
LIM1215	CHR19	579545	T	C	BSG	-	-	low	no	no
LIM1215	CHR19	7965163	G	A	LRRC8E	-	-	low	no	no
LIM1215	CHR20	30729696	A	G	TM9SF4	-	-	low	no	no
LIM1215	CHR20	4843508	C	T	SLC23A2	DAMAGING	-	neutral	no	yes
LIM1215	CHR2	220033503	A	C	SLC23A3	-	DAMAGING	medium	no	no
LIM1215	CHR22	46932005	G	A	CELSR1	-	-	-	no	no
LIM1215	CHR22	47073102	A	G	GRAMD4	-	DAMAGING	low	no	no
LIM1215	CHR3	14509377	G	T	SLC6A6	-	DAMAGING	low	no	no
LIM1215	CHR3	48475196	A	G	CCDC51	DAMAGING	-	medium	no	no
LIM1215	CHR3	49568662	G	A	DAG1	-	DAMAGING	medium	no	no
LIM1215	CHR3	66457811	G	A	LRIG1	DAMAGING	DAMAGING	neutral	no	no
LIM1215	CHR5	140347624	G	A	PCDHAC2	-	-	low	no	no
LIM1215	CHR5	140753844	G	T	PCDHGA6	DAMAGING	-	high	no	no
LIM1215	CHR5	140764756	C	T	PCDHGA7	DAMAGING	-	medium	no	no
LIM1215	CHR5	140811210	A	C	PCDHGA12	DAMAGING	-	high	no	no
LIM1215	CHR5	1422074	G	A	SLC6A3	DAMAGING	DAMAGING	high	no	yes
LIM1215	CHR5	150867777	G	A	SLC36A1	-	-	neutral	no	yes
LIM1215	CHR6	109763488	G	A	SMPD2	-	-	neutral	no	no
LIM1215	CHR6	128319988	A	C	PTPRK	-	DAMAGING	low	no	no
LIM1215	CHR6	26468259	A	G	BTN2A1	-	DAMAGING	low	no	no
LIM1215	CHR6	41160563	A	C	TREML2	-	-	medium	no	no
LIM1215	CHR7	143092588	C	A	EPHA1	DAMAGING	-	high	yes	no
LIM1215	CHR8	22277108	G	A	SLC39A14	-	-	low	no	yes
LIM1215	CHR9	119976799	G	A	ASTN2	DAMAGING	DAMAGING	neutral	no	no
LIM1215	CHR9	92020267	T	A	SEMA4D	-	-	low	no	no
LIM1215	CHRX	109416555	C	T	TMEM164	-	-	neutral	no	no
LIM1215	CHRX	70327701	A	G	IL2RG	-	DAMAGING	medium	no	yes
LIM2405	CHR10	105209392	G	T	CALHM2	-	-	low	no	no
LIM2405	CHR10	123263358	C	G	FGFR2	-	-	medium	yes	yes
LIM2405	CHR10	125769704	C	A	CHST15	DAMAGING	-	low	no	no
LIM2405	CHR11	113075150	C	T	NCAM1	-	-	-	no	no
LIM2405	CHR11	113140998	C	T	NCAM1	-	-	-	no	no
LIM2405	CHR1	11333834	G	T	UBIAD1	-	-	neutral	no	no
LIM2405	CHR1	116206573	G	T	VANGL1	DAMAGING	DAMAGING	medium	no	no
LIM2405	CHR1	120459086	G	A	NOTCH2	-	-	low	no	yes
LIM2405	CHR1	153750812	G	A	SLC27A3	DAMAGING	DAMAGING	low	no	no
LIM2405	CHR11	57265257	G	T	SLC43A1	-	DAMAGING	low	no	no
LIM2405	CHR11	64065651	C	T	KCNK4	-	DAMAGING	medium	no	no
LIM2405	CHR1	16458881	A	G	EPHA2	DAMAGING	-	neutral	yes	yes
LIM2405	CHR12	56525318	G	T	ESYT1	DAMAGING	DAMAGING	medium	no	no
LIM2405	CHR12	57577933	C	T	LRP1	DAMAGING	DAMAGING	medium	no	yes
LIM2405	CHR13	95861803	C	T	ABCC4	-	-	low	no	yes
LIM2405	CHR1	44087648	G	A	PTPRF	-	DAMAGING	neutral	no	no
LIM2405	CHR15	94841705	C	T	MCTP2	-	-	neutral	no	no

LIM2405	CHR16	19501841	A	G	TMC5	-	DAMAGING	medium	no	no
LIM2405	CHR16	2376023	T	C	ABCA3	-	-	neutral	no	no
LIM2405	CHR17	1628915	C	A	WDR81	DAMAGING	-	-	no	no
LIM2405	CHR17	37809884	T	C	STARD3	-	-	neutral	no	no
LIM2405	CHR17	37840938	C	T	PGAP3	DAMAGING	DAMAGING	medium	no	no
LIM2405	CHR17	73484958	G	A	KIAA0195	DAMAGING	DAMAGING	low	no	no
LIM2405	CHR19	11217291	A	G	LDLR	-	-	neutral	no	yes
LIM2405	CHR19	11492608	A	G	EPOR	DAMAGING	DAMAGING	medium	no	yes
LIM2405	CHR19	17441024	C	T	ANO8	-	-	neutral	no	no
LIM2405	CHR19	38849104	A	T	CATSPERG	-	-	medium	no	no
LIM2405	CHR19	52033086	C	T	SIGLEC6	DAMAGING	-	low	no	no
LIM2405	CHR20	3842084	G	A	MAVS	-	-	neutral	no	no
LIM2405	CHR21	34635228	G	T	IFNAR2	-	-	neutral	no	yes
LIM2405	CHR21	43716452	C	G	ABCG1	DAMAGING	DAMAGING	medium	no	yes
LIM2405	CHR21	45815315	C	T	TRPM2	DAMAGING	-	medium	no	no
LIM2405	CHR2	219204817	G	T	PNKD	-	-	neutral	no	no
LIM2405	CHR2	220085922	C	G	ATG9A	-	-	neutral	no	no
LIM2405	CHR3	124635290	C	T	MUC13	-	-	neutral	no	no
LIM2405	CHR3	45152150	C	T	CDCP1	-	DAMAGING	medium	no	no
LIM2405	CHR4	187524677	C	T	FAT1	-	DAMAGING	medium	no	no
LIM2405	CHR4	187540974	C	T	FAT1	DAMAGING	DAMAGING	high	no	no
LIM2405	CHR4	79432468	G	A	FRAS1	DAMAGING	DAMAGING	medium	no	no
LIM2405	CHR5	1057650	C	T	SLC12A7	-	-	low	no	yes
LIM2405	CHR5	140755029	A	G	PCDHGA6	DAMAGING	-	medium	no	no
LIM2405	CHR6	160481628	C	A	IGF2R	-	-	-	no	yes
LIM2405	CHR6	170592574	C	T	DLL1	-	DAMAGING	medium	no	no
LIM2405	CHR6	31322427	C	T	HLA-B	-	-	medium	no	yes
LIM2405	CHR7	128415166	C	T	OPN1SW	DAMAGING	DAMAGING	high	no	no
LIM2405	CHR7	94285380	C	T	SGCE	-	-	neutral	no	no
LIM2405	CHR8	27516475	G	A	SCARA3	-	-	low	no	no
LIM2405	CHR8	29923585	T	C	TMEM66	-	-	low	no	no
LIM2405	CHR9	125797349	G	C	GPR21	DAMAGING	DAMAGING	medium	no	no
LIM2405	CHR9	98241300	C	T	PTCH1	-	-	-	no	yes
LOVO	CHR10	104232618	A	C	TMEM180	DAMAGING	DAMAGING	medium	no	no
LOVO	CHR10	135080895	G	T	ADAM8	-	-	-	no	no
LOVO	CHR1	100546163	T	C	HIAT1	-	-	neutral	no	no
LOVO	CHR1	109795250	T	G	CELSR2	DAMAGING	DAMAGING	high	no	no
LOVO	CHR11	117166020	A	G	BACE1	DAMAGING	DAMAGING	medium	no	yes
LOVO	CHR11	430132	T	G	ANO9	DAMAGING	-	medium	no	no
LOVO	CHR11	46920222	T	G	LRP4	-	-	neutral	no	no
LOVO	CHR11	47436649	G	A	SLC39A13	-	DAMAGING	low	no	no
LOVO	CHR1	17320215	C	T	ATP13A2	-	-	neutral	no	no
LOVO	CHR1	197479912	T	A	DENND1B	-	-	neutral	no	no
LOVO	CHR1	20097931	T	C	TMCO4	-	-	low	no	no
LOVO	CHR1	207930431	G	A	CD46	DAMAGING	DAMAGING	medium	no	yes
LOVO	CHR1	208202230	A	G	PLXNA2	DAMAGING	DAMAGING	medium	no	no
LOVO	CHR12	110783175	A	G	ATP2A2	DAMAGING	-	high	no	yes
LOVO	CHR12	12336940	A	G	LRP6	-	-	medium	no	no
LOVO	CHR12	12397441	A	C	LRP6	DAMAGING	-	medium	no	no
LOVO	CHR1	22047625	G	A	USP48	DAMAGING	DAMAGING	medium	no	no
LOVO	CHR12	21926489	T	G	KCNJ8	-	-	low	no	yes
LOVO	CHR12	56120970	C	T	CD63	DAMAGING	DAMAGING	high	no	no
LOVO	CHR12	56350994	G	A	PMEL	-	-	-	no	no
LOVO	CHR12	57598907	A	C	LRP1	-	-	neutral	no	yes
LOVO	CHR12	6497984	G	A	LTBR	-	-	neutral	no	no
LOVO	CHR12	7586227	A	C	CD163L1	DAMAGING	DAMAGING	high	no	no
LOVO	CHR1	28293284	A	C	XKR8	-	-	medium	no	no
LOVO	CHR12	89984781	T	C	ATP2B1	-	DAMAGING	medium	no	no
LOVO	CHR12	9096398	G	A	M6PR	-	-	low	no	yes
LOVO	CHR12	94697763	A	G	PLXNC1	-	DAMAGING	neutral	no	no
LOVO	CHR1	29641997	G	A	PTPRU	DAMAGING	DAMAGING	high	no	no
LOVO	CHR12	98987839	A	C	SLC25A3	-	-	neutral	no	no
LOVO	CHR13	113975734	C	T	LAMP1	DAMAGING	DAMAGING	medium	no	no

LOVO	CHR13	114152825	G	A	TMCO3	-	-	neutral	no	no
LOVO	CHR13	39266025	G	A	FREM2	-	-	neutral	no	no
LOVO	CHR1	35250967	A	C	GJB3	DAMAGING	-	medium	no	no
LOVO	CHR13	52520478	A	C	ATP7B	DAMAGING	DAMAGING	low	no	no
LOVO	CHR14	23600683	G	A	SLC7A8	DAMAGING	DAMAGING	medium	no	yes
LOVO	CHR14	24788600	T	G	ADCY4	DAMAGING	DAMAGING	high	no	no
LOVO	CHR14	71445069	A	C	PCNX	-	-	low	no	no
LOVO	CHR14	77714802	A	C	TMEM63C	DAMAGING	-	low	no	no
LOVO	CHR15	23049280	G	A	NIPA1	DAMAGING	-	low	no	no
LOVO	CHR1	53724095	T	G	LRP8	DAMAGING	DAMAGING	medium	no	no
LOVO	CHR15	40660468	T	C	DISP2	DAMAGING	DAMAGING	medium	no	no
LOVO	CHR15	40661537	T	G	DISP2	DAMAGING	DAMAGING	neutral	no	no
LOVO	CHR15	41870355	A	C	TYRO3	-	-	neutral	yes	no
LOVO	CHR15	55652665	G	A	CCPG1	DAMAGING	-	neutral	no	no
LOVO	CHR15	68500525	G	A	CLN6	-	DAMAGING	medium	no	no
LOVO	CHR15	73866007	G	T	NPTN	-	-	low	no	no
LOVO	CHR15	90768257	C	T	SEMA4B	DAMAGING	DAMAGING	medium	no	no
LOVO	CHR16	2140103	G	A	PKD1	-	-	-	no	no
LOVO	CHR16	2164370	C	T	PKD1	DAMAGING	-	low	yes	no
LOVO	CHR16	29996576	C	T	TAOK2	DAMAGING	DAMAGING	medium	yes	yes
LOVO	CHR16	29999191	C	T	TAOK2	-	-	neutral	yes	yes
LOVO	CHR16	4033342	T	G	ADCY9	-	-	low	no	no
LOVO	CHR16	67283045	A	G	SLC9A5	-	-	neutral	no	no
LOVO	CHR16	88802618	C	T	FAM38A	-	-	low	no	no
LOVO	CHR17	37829899	C	T	PGAP3	-	-	neutral	no	no
LOVO	CHR17	57812701	C	A	VMP1	DAMAGING	-	low	no	no
LOVO	CHR17	73495049	A	C	KIAA0195	-	-	neutral	no	no
LOVO	CHR17	74471144	C	T	RHBDF2	DAMAGING	-	low	no	no
LOVO	CHR17	74473316	C	T	RHBDF2	-	-	low	no	no
LOVO	CHR17	78197095	A	C	SLC26A11	DAMAGING	DAMAGING	high	no	no
LOVO	CHR18	29784272	C	T	MEP1B	DAMAGING	DAMAGING	high	no	no
LOVO	CHR18	55351387	T	C	ATP8B1	-	-	low	no	no
LOVO	CHR18	66365210	G	A	TMX3	-	-	neutral	no	no
LOVO	CHR18	77107881	G	A	ATP9B	-	-	neutral	no	no
LOVO	CHR19	1005130	T	C	GRIN3B	DAMAGING	DAMAGING	high	no	yes
LOVO	CHR19	1041420	A	C	ABCA7	DAMAGING	DAMAGING	low	no	no
LOVO	CHR19	1051532	T	G	ABCA7	DAMAGING	DAMAGING	medium	no	no
LOVO	CHR19	10747397	G	A	SLC44A2	DAMAGING	DAMAGING	high	no	no
LOVO	CHR19	14281571	G	A	LPHN1	-	-	low	no	no
LOVO	CHR19	17330035	T	G	USE1	-	-	neutral	no	no
LOVO	CHR19	35506803	G	A	GRAMD1A	DAMAGING	-	low	no	no
LOVO	CHR19	38780809	C	T	SPINT2	DAMAGING	DAMAGING	high	no	no
LOVO	CHR19	41726597	C	T	AXL	DAMAGING	DAMAGING	low	yes	no
LOVO	CHR19	42874398	A	C	MEGF8	-	-	neutral	no	no
LOVO	CHR19	47127190	G	T	PTGIR	DAMAGING	DAMAGING	medium	no	yes
LOVO	CHR19	52034878	C	T	SIGLEC6	-	-	neutral	no	no
LOVO	CHR19	54666855	G	A	TMC4	-	DAMAGING	medium	no	no
LOVO	CHR19	7965646	C	A	LRRC8E	DAMAGING	DAMAGING	medium	no	no
LOVO	CHR20	3844977	A	C	MAVS	DAMAGING	-	low	no	no
LOVO	CHR20	7980500	C	T	TMX4	-	-	low	no	no
LOVO	CHR21	45655287	C	T	ICOSLG	-	-	low	no	no
LOVO	CHR21	45789120	T	C	TRPM2	DAMAGING	DAMAGING	medium	no	no
LOVO	CHR2	206659717	G	A	NRP2	DAMAGING	DAMAGING	medium	no	yes
LOVO	CHR2	220089023	C	T	ATG9A	DAMAGING	DAMAGING	medium	no	no
LOVO	CHR2	231742121	A	G	ITM2C	-	-	medium	no	no
LOVO	CHR22	46761563	C	T	CELSR1	-	-	neutral	no	no
LOVO	CHR22	46787138	C	T	CELSR1	-	-	-	no	no
LOVO	CHR2	32396375	G	A	SLC30A6	-	-	low	no	no
LOVO	CHR2	62728423	A	G	TMEM17	DAMAGING	DAMAGING	low	no	no
LOVO	CHR2	70683600	C	T	TGFA	DAMAGING	DAMAGING	medium	no	no
LOVO	CHR2	74900656	G	A	SEMA4F	-	-	neutral	no	no
LOVO	CHR2	97427562	A	C	CNNM4	DAMAGING	DAMAGING	high	no	no
LOVO	CHR3	124492643	C	T	ITGB5	DAMAGING	-	neutral	no	yes

LOVO	CHR3	124854538	A	G	SLC12A8	-	-	-	no	no
LOVO	CHR3	125786918	T	G	SLC41A3	-	-	neutral	no	no
LOVO	CHR3	129290023	C	T	PLXND1	-	-	neutral	no	no
LOVO	CHR3	183585757	C	T	PARL	-	-	low	no	no
LOVO	CHR3	196387946	C	G	LRRC33	-	-	low	no	no
LOVO	CHR3	27427493	G	A	SLC4A7	DAMAGING	DAMAGING	medium	no	no
LOVO	CHR3	48699902	T	C	CELSR3	-	-	neutral	no	no
LOVO	CHR3	49935049	C	A	MST1R	-	-	-	yes	yes
LOVO	CHR3	62257058	G	T	PTPRG	DAMAGING	DAMAGING	medium	no	no
LOVO	CHR4	1018197	G	A	FGFRL1	-	DAMAGING	neutral	no	no
LOVO	CHR4	1807345	A	G	FGFR3	DAMAGING	DAMAGING	neutral	yes	yes
LOVO	CHR4	187004485	G	A	TLR3	-	-	low	no	yes
LOVO	CHR4	187531023	C	T	FAT1	-	-	medium	no	no
LOVO	CHR4	527664	A	C	PIGG	DAMAGING	-	low	no	no
LOVO	CHR4	6293694	C	T	WFS1	-	DAMAGING	low	no	no
LOVO	CHR4	6302694	A	G	WFS1	DAMAGING	-	medium	no	no
LOVO	CHR4	75937991	G	A	PARM1	-	-	neutral	no	no
LOVO	CHR4	79429993	C	T	FRAS1	DAMAGING	DAMAGING	low	no	no
LOVO	CHR5	140256391	T	G	PCDHA12	DAMAGING	-	high	no	no
LOVO	CHR5	140256553	T	G	PCDHA12	DAMAGING	-	neutral	no	no
LOVO	CHR5	140263772	T	G	PCDHA13	-	-	neutral	no	no
LOVO	CHR5	140802818	A	C	PCDHGA11	DAMAGING	-	medium	no	no
LOVO	CHR5	140870675	C	T	PCDHGC5	-	-	low	no	no
LOVO	CHR5	52356848	C	T	ITGA2	DAMAGING	-	medium	no	no
LOVO	CHR6	128294289	G	C	PTPRK	DAMAGING	DAMAGING	high	no	no
LOVO	CHR6	26410240	G	A	BTN3A1	-	-	neutral	no	no
LOVO	CHR6	26448676	G	T	BTN3A3	-	-	-	no	no
LOVO	CHR6	33541938	C	T	BAK1	DAMAGING	DAMAGING	medium	no	no
LOVO	CHR6	44108011	A	C	TMEM63B	DAMAGING	DAMAGING	medium	no	no
LOVO	CHR6	44108014	T	C	TMEM63B	DAMAGING	-	neutral	no	no
LOVO	CHR6	70500336	A	G	LMBRD1	DAMAGING	-	low	no	no
LOVO	CHR7	100172869	G	A	LRCH4	-	-	neutral	no	no
LOVO	CHR7	100225424	C	T	TFR2	-	-	low	no	no
LOVO	CHR7	142574205	C	T	TRPV6	DAMAGING	DAMAGING	medium	no	no
LOVO	CHR7	143097119	A	C	EPHA1	DAMAGING	DAMAGING	medium	yes	no
LOVO	CHR7	150778821	A	G	TMUB1	-	-	low	no	no
LOVO	CHR7	17913195	G	A	SNX13	DAMAGING	DAMAGING	low	no	no
LOVO	CHR7	2698630	G	A	TTYH3	DAMAGING	DAMAGING	medium	no	no
LOVO	CHR7	38256841	T	C	STARD3NL	DAMAGING	DAMAGING	medium	no	no
LOVO	CHR7	47945458	C	T	PKD1L1	DAMAGING	-	low	no	no
LOVO	CHR7	73097245	G	A	DNAJC30	DAMAGING	DAMAGING	medium	no	no
LOVO	CHR7	86537827	T	A	KIAA1324L	DAMAGING	-	low	no	no
LOVO	CHR7	94259085	G	A	SGCE	-	-	low	no	no
LOVO	CHR7	97822901	T	G	LMTK2	DAMAGING	DAMAGING	medium	yes	no
LOVO	CHR9	34659772	T	C	IL11RA	-	DAMAGING	low	no	yes
LOVO	CHR9	92011729	G	A	SEMA4D	DAMAGING	DAMAGING	high	no	no
RKO	CHR10	102740081	A	G	SEMA4G	-	-	medium	no	no
RKO	CHR10	104679142	C	T	CNNM2	DAMAGING	DAMAGING	high	no	no
RKO	CHR10	95168603	C	T	MYOF	-	-	low	no	no
RKO	CHR11	113075210	C	G	NCAM1	-	-	-	no	no
RKO	CHR11	117061379	T	C	SIDT2	DAMAGING	DAMAGING	medium	no	no
RKO	CHR11	124747437	G	A	ROBO3	-	-	low	no	no
RKO	CHR1	113460571	C	A	SLC16A1	DAMAGING	DAMAGING	high	no	yes
RKO	CHR1	116206372	A	G	VANGL1	-	-	low	no	no
RKO	CHR1	116228128	A	T	VANGL1	-	-	low	no	no
RKO	CHR1	117142745	C	T	IGSF3	-	DAMAGING	low	no	no
RKO	CHR1	117158816	T	G	IGSF3	DAMAGING	-	neutral	no	no
RKO	CHR1	11898608	T	C	CLCN6	DAMAGING	-	high	no	yes
RKO	CHR11	2418040	G	A	CD81	DAMAGING	DAMAGING	medium	no	no
RKO	CHR11	3846602	A	C	PGAP2	DAMAGING	DAMAGING	medium	no	no
RKO	CHR1	156255211	G	T	TMEM79	DAMAGING	DAMAGING	low	no	no
RKO	CHR11	57193479	T	G	SLC43A3	-	DAMAGING	low	no	no
RKO	CHR1	160321569	G	A	NCSTN	-	-	neutral	no	yes

RKO	CHR11	64852670	C	A	ZFPL1	-	-	low	no	no
RKO	CHR11	66052344	A	C	YIF1A	-	-	neutral	no	no
RKO	CHR11	67765177	G	A	UNC93B1	-	DAMAGING	-	no	no
RKO	CHR11	68030155	G	A	C11ORF24	-	-	neutral	no	no
RKO	CHR1	168073766	A	G	GPR161	DAMAGING	-	medium	no	no
RKO	CHR11	68846041	T	C	TPCN2	DAMAGING	-	low	no	no
RKO	CHR1	17328812	C	A	ATP13A2	-	-	medium	no	no
RKO	CHR12	117217062	G	T	RNFT2	DAMAGING	DAMAGING	low	no	no
RKO	CHR12	125298860	T	C	SCARB1	-	-	low	no	yes
RKO	CHR1	223178790	C	A	DISP1	-	-	low	no	no
RKO	CHR12	3387704	G	A	TSPAN9	DAMAGING	DAMAGING	medium	no	no
RKO	CHR1	24768554	C	T	NIPAL3	DAMAGING	DAMAGING	medium	no	no
RKO	CHR12	56527209	T	G	ESYT1	DAMAGING	-	medium	no	no
RKO	CHR12	57553743	G	T	LRP1	DAMAGING	DAMAGING	low	no	yes
RKO	CHR12	57575050	C	T	LRP1	DAMAGING	-	medium	no	yes
RKO	CHR12	57579569	C	T	LRP1	DAMAGING	-	low	no	yes
RKO	CHR12	6483859	G	T	SCNN1A	-	-	low	no	yes
RKO	CHR12	9096441	G	A	M6PR	DAMAGING	DAMAGING	medium	no	yes
RKO	CHR13	101287114	C	T	TMTC4	-	-	neutral	no	no
RKO	CHR13	113508637	C	T	ATP11A	DAMAGING	DAMAGING	high	no	no
RKO	CHR13	113530127	A	G	ATP11A	-	-	low	no	no
RKO	CHR1	32204498	C	T	BAI2	DAMAGING	DAMAGING	medium	no	no
RKO	CHR1	32568228	C	T	TMEM39B	-	DAMAGING	medium	no	no
RKO	CHR13	30091863	C	T	SLC7A1	-	-	neutral	no	yes
RKO	CHR1	35915572	A	G	KIAA0319L	DAMAGING	DAMAGING	medium	no	no
RKO	CHR14	23346509	G	T	LRP10	DAMAGING	-	low	no	no
RKO	CHR1	45120334	G	T	TMEM53	DAMAGING	-	medium	no	no
RKO	CHR1	45289032	C	T	PTCH2	DAMAGING	-	medium	no	no
RKO	CHR1	45293664	T	C	PTCH2	-	-	neutral	no	no
RKO	CHR15	73528735	G	C	NEO1	-	-	low	no	no
RKO	CHR15	99251062	G	C	IGF1R	DAMAGING	DAMAGING	medium	yes	yes
RKO	CHR16	16225773	C	G	ABCC1	-	-	low	no	yes
RKO	CHR16	2569378	T	C	ATP6V0C	DAMAGING	-	medium	no	no
RKO	CHR16	31049475	G	A	STX4	DAMAGING	-	medium	no	yes
RKO	CHR16	67300025	G	T	SLC9A5	-	-	low	no	no
RKO	CHR16	67979098	G	A	SLC12A4	DAMAGING	-	medium	no	yes
RKO	CHR16	68712439	C	A	CDH3	-	-	medium	no	no
RKO	CHR17	30349379	A	G	LRRC37B	-	-	medium	no	no
RKO	CHR17	56435492	G	T	RNF43	-	DAMAGING	low	no	no
RKO	CHR17	60744769	G	T	MRC2	-	-	neutral	no	no
RKO	CHR17	65014379	T	A	CACNG4	-	DAMAGING	low	no	no
RKO	CHR17	7324449	A	T	SPEM1	-	-	neutral	no	no
RKO	CHR17	7460614	C	A	TNFSF12	-	-	neutral	no	no
RKO	CHR17	7464137	T	C	TNFSF12-TNFSF13	DAMAGING	-	low	no	no
RKO	CHR19	10394319	A	G	ICAM1	-	-	medium	no	yes
RKO	CHR19	10748932	C	T	SLC44A2	DAMAGING	-	medium	no	no
RKO	CHR19	11038608	G	A	YIPF2	DAMAGING	DAMAGING	medium	no	no
RKO	CHR19	14938468	G	C	OR7A5	-	-	neutral	no	no
RKO	CHR19	35501037	C	T	GRAMD1A	DAMAGING	-	medium	no	no
RKO	CHR19	35524464	T	C	SCN1B	-	DAMAGING	low	no	yes
RKO	CHR19	35655100	T	C	FXYD5	-	-	low	no	no
RKO	CHR19	38845419	G	A	CATSPERG	-	DAMAGING	medium	no	no
RKO	CHR19	38847382	T	C	CATSPERG	-	-	low	no	no
RKO	CHR1	9661436	C	T	TMEM201	-	-	-	no	no
RKO	CHR19	7132268	C	T	INSR	-	-	medium	yes	yes
RKO	CHR20	33764148	A	G	PROCR	-	DAMAGING	medium	no	yes
RKO	CHR20	50241871	G	A	ATP9A	-	-	low	no	no
RKO	CHR20	50305627	A	G	ATP9A	DAMAGING	DAMAGING	medium	no	no
RKO	CHR2	103281806	T	C	SLC9A2	DAMAGING	-	medium	no	no
RKO	CHR2	103324613	A	T	SLC9A2	-	-	neutral	no	no
RKO	CHR21	42629197	A	G	BACE2	DAMAGING	DAMAGING	medium	no	no
RKO	CHR2	158626966	T	C	ACVR1	DAMAGING	DAMAGING	high	yes	yes
RKO	CHR2	191364873	C	T	MFSD6	-	-	neutral	no	no

RKO	CHR2	206607948	T	A	NRP2	-	DAMAGING	medium	no	yes
RKO	CHR22	46792574	C	T	CELSR1	-	-	neutral	no	no
RKO	CHR22	46806300	C	T	CELSR1	-	-	neutral	no	no
RKO	CHR2	25095489	G	A	ADCY3	DAMAGING	-	medium	no	no
RKO	CHR2	27001017	G	A	C2ORF18	-	-	low	no	no
RKO	CHR3	126748314	C	A	PLXNA1	DAMAGING	DAMAGING	medium	no	no
RKO	CHR3	183707111	A	G	ABCC5	-	-	low	no	yes
RKO	CHR3	183995790	C	T	ECE2	DAMAGING	DAMAGING	low	no	no
RKO	CHR3	196674848	G	A	PIGZ	DAMAGING	DAMAGING	medium	no	no
RKO	CHR3	48473904	T	C	CCDC51	-	-	medium	no	no
RKO	CHR3	48686177	G	A	CELSR3	-	DAMAGING	medium	no	no
RKO	CHR3	9959265	G	A	IL17RC	-	-	neutral	no	no
RKO	CHR4	151520252	A	G	LRBA	DAMAGING	DAMAGING	medium	no	no
RKO	CHR4	520968	C	T	PIGG	-	-	medium	no	no
RKO	CHR5	140755224	T	C	PCDHGA6	-	-	neutral	no	no
RKO	CHR5	140794385	G	A	PCDHGA10	DAMAGING	-	medium	no	no
RKO	CHR5	179221104	T	C	LTC4S	DAMAGING	DAMAGING	low	no	yes
RKO	CHR5	76128829	G	A	F2RL1	-	-	low	no	no
RKO	CHR5	94230493	A	T	MCTP1	DAMAGING	-	medium	no	no
RKO	CHR6	160510968	A	T	IGF2R	DAMAGING	DAMAGING	medium	no	yes
RKO	CHR6	43127684	C	T	PTK7	-	-	neutral	yes	no
RKO	CHR6	44120367	C	T	TMEM63B	DAMAGING	DAMAGING	medium	no	no
RKO	CHR7	131196094	G	T	PODXL	-	-	neutral	no	no
RKO	CHR7	133979776	C	T	SLC35B4	DAMAGING	DAMAGING	medium	no	no
RKO	CHR7	150762019	G	A	SLC4A2	-	-	neutral	no	no
RKO	CHR7	77572031	G	C	PHTF2	DAMAGING	DAMAGING	medium	no	no
RKO	CHR7	99474065	T	C	OR2AE1	-	-	neutral	no	no
RKO	CHR8	38285611	G	A	FGFR1	-	-	neutral	yes	yes
RKO	CHR9	139571926	C	G	AGPAT2	-	-	neutral	no	no
RKO	CHR9	86893153	C	A	SLC28A3	-	-	neutral	no	yes
RKO	CHR9	91993695	T	C	SEMA4D	-	-	neutral	no	no
RKO	CHR9	95887195	G	A	NINJ1	-	-	-	no	no
RW2982	CHR14	45628379	G	T	FANCM	DAMAGING	-	low	no	no
RW2982	CHR20	3838304	C	T	MAVS	-	-	neutral	no	no
RW2982	CHR2	158617452	C	T	ACVR1	DAMAGING	-	low	yes	yes
RW2982	CHR2	97527502	G	A	SEMA4C	-	-	medium	no	no
RW2982	CHR4	151770012	T	C	LRBA	-	-	neutral	no	no
RW2982	CHR5	159344562	C	T	ADRA1B	-	-	neutral	no	yes
RW2982	CHR7	100183660	C	T	LRCH4	-	-	low	no	no
RW2982	CHR7	105615475	T	A	CDHR3	-	-	low	no	no
RW2982	CHR7	97821335	G	A	LMTK2	-	-	low	yes	no
RW2982	CHR9	74344988	A	T	TMEM2	DAMAGING	-	medium	no	no
RW7213	CHR1	109795012	G	A	CELSR2	-	-	low	no	no
RW7213	CHR12	52374851	A	G	ACVR1B	DAMAGING	DAMAGING	medium	yes	yes
RW7213	CHR16	19483485	A	G	TMC5	-	-	neutral	no	no
RW7213	CHR16	726120	G	A	RHBDL1	-	-	neutral	no	no
RW7213	CHR17	48750402	T	A	ABCC3	DAMAGING	DAMAGING	high	no	yes
RW7213	CHR19	19764782	G	A	ATP13A1	DAMAGING	DAMAGING	medium	no	no
RW7213	CHR3	129305085	C	T	PLXND1	-	-	low	no	no
RW7213	CHR3	184001720	G	A	ECE2	DAMAGING	DAMAGING	medium	no	no
RW7213	CHR3	49548142	G	A	DAG1	-	-	low	no	no
SKCO1	CHR10	135081560	C	T	ADAM8	-	-	neutral	no	no
SKCO1	CHR11	113102986	G	A	NCAM1	-	-	-	no	no
SKCO1	CHR1	17316695	T	G	ATP13A2	-	-	neutral	no	no
SKCO1	CHR12	51868953	T	A	SLC4A8	-	-	-	no	no
SKCO1	CHR16	2369721	A	G	ABCA3	DAMAGING	DAMAGING	medium	no	no
SKCO1	CHR16	74761211	T	G	FA2H	DAMAGING	DAMAGING	medium	no	no
SKCO1	CHR17	1630507	C	T	WDR81	-	-	-	no	no
SKCO1	CHR19	18092608	G	A	KCNN1	DAMAGING	DAMAGING	medium	no	yes
SKCO1	CHR19	48845878	T	C	TMEM143	-	DAMAGING	low	no	no
SKCO1	CHR20	25290161	C	A	ABHD12	DAMAGING	-	medium	no	no
SKCO1	CHR22	43524524	G	T	BIK	-	-	neutral	no	no
SKCO1	CHR22	46835249	C	T	CELSR1	DAMAGING	DAMAGING	medium	no	no

SKCO1	CHR3	38518787	G	A	ACVR2B	-	-	neutral	no	no
SKCO1	CHR3	49933270	C	T	MST1R	-	-	low	yes	yes
SKCO1	CHR4	84205934	G	A	COQ2	-	-	-	no	yes
SKCO1	CHR5	140256607	T	G	PCDHA12	DAMAGING	-	medium	no	no
SKCO1	CHR6	128718787	C	A	PTPRK	-	-	medium	no	no
SKCO1	CHR6	41166123	C	T	TREML2	DAMAGING	DAMAGING	medium	no	no
SKCO1	CHR7	97822800	C	T	LMTK2	-	-	neutral	yes	no
SW1116	CHR10	102740703	C	A	SEMA4G	-	-	neutral	no	no
SW1116	CHR1	29637287	C	A	PTPRU	DAMAGING	DAMAGING	high	no	no
SW1116	CHR16	2369767	C	T	ABCA3	-	-	low	no	no
SW1116	CHR20	3209797	C	A	SLC4A11	-	-	neutral	no	no
SW1116	CHR20	50286589	C	T	ATP9A	-	-	low	no	no
SW1116	CHR2	206605246	T	A	NRP2	DAMAGING	DAMAGING	medium	no	yes
SW1116	CHR2	230910527	C	T	SLC16A14	-	-	neutral	no	no
SW1116	CHR4	166021848	C	T	TMEM192	-	-	low	no	no
SW1116	CHR4	187541255	G	A	FAT1	DAMAGING	DAMAGING	medium	no	no
SW1116	CHR5	102361004	A	G	PAM	-	-	neutral	no	yes
SW1116	CHR7	129842422	C	A	TMEM209	DAMAGING	-	low	no	no
SW1116	CHR7	55238079	G	A	EGFR	-	-	neutral	yes	yes
SW1116	CHR8	38848906	A	G	TM2D2	DAMAGING	DAMAGING	medium	no	no
SW403	CHR10	102732779	C	G	SEMA4G	DAMAGING	-	low	no	no
SW403	CHR10	114158758	G	A	ACSL5	DAMAGING	-	low	no	no
SW403	CHR1	109737169	G	T	KIAA1324	DAMAGING	DAMAGING	medium	no	no
SW403	CHR11	120200709	G	T	TMEM136	DAMAGING	DAMAGING	medium	no	no
SW403	CHR1	120612013	G	A	NOTCH2	-	-	neutral	no	yes
SW403	CHR11	5474897	A	G	OR51I2	DAMAGING	DAMAGING	medium	no	no
SW403	CHR11	59556388	A	T	STX3	DAMAGING	-	medium	no	no
SW403	CHR11	60694795	C	A	TMEM132A	DAMAGING	DAMAGING	medium	no	no
SW403	CHR1	208391152	C	A	PLXNA2	-	-	neutral	no	no
SW403	CHR1	211966472	T	C	LPGAT1	DAMAGING	DAMAGING	medium	no	no
SW403	CHR12	124221722	G	T	ATP6V0A2	DAMAGING	-	medium	no	no
SW403	CHR12	56091027	C	T	ITGA7	-	-	low	no	no
SW403	CHR1	32201230	G	C	BAI2	DAMAGING	DAMAGING	low	no	no
SW403	CHR13	39266354	C	A	FREM2	-	DAMAGING	low	no	no
SW403	CHR14	23243206	G	T	SLC7A7	-	-	-	no	no
SW403	CHR14	77722980	G	T	TMEM63C	-	-	-	no	no
SW403	CHR16	66657381	C	A	CMTM4	-	-	low	no	no
SW403	CHR19	3193423	G	T	NCLN	-	-	-	no	no
SW403	CHR19	49244571	C	T	IZUMO1	-	-	neutral	no	no
SW403	CHR20	56227606	G	T	PMEPA1	-	-	low	no	no
SW403	CHR20	7964486	C	A	TMX4	-	-	-	no	no
SW403	CHR2	36691772	G	T	CRIM1	DAMAGING	DAMAGING	medium	no	no
SW403	CHR3	141626072	C	T	ATP1B3	-	-	low	no	no
SW403	CHR3	196674036	C	A	PIGZ	-	-	-	no	no
SW403	CHR3	31641900	C	G	STT3B	DAMAGING	DAMAGING	high	no	no
SW403	CHR3	53835189	G	T	CACNA1D	-	-	low	no	yes
SW403	CHR4	151829958	T	A	LRBA	-	-	-	no	no
SW403	CHR4	99428851	C	T	TSPAN5	DAMAGING	-	high	no	no
SW403	CHR5	1085384	C	A	SLC12A7	DAMAGING	DAMAGING	high	no	yes
SW403	CHR5	131309087	C	T	ACSL6	-	-	low	no	no
SW403	CHR5	140763515	C	T	PCDHGA7	DAMAGING	-	medium	no	no
SW403	CHR6	117866651	G	T	DCBLD1	DAMAGING	DAMAGING	low	no	no
SW403	CHR6	46761170	T	G	MEP1A	DAMAGING	-	low	no	no
SW403	CHR7	117188797	A	G	CFTR	-	-	neutral	no	yes
SW403	CHR9	131483519	C	T	ZDHHC12	-	-	neutral	no	no
SW403	CHR9	139568239	C	T	AGPAT2	-	-	low	no	no
SW480	CHR10	74100566	C	T	DNAJB12	-	DAMAGING	low	no	no
SW480	CHR11	113105855	C	T	NCAM1	-	-	-	no	no
SW480	CHR1	158063168	G	A	KIRREL	DAMAGING	DAMAGING	low	no	no
SW480	CHR1	21548274	G	A	ECE1	DAMAGING	-	high	no	yes
SW480	CHR1	27440424	C	T	SLC9A1	DAMAGING	DAMAGING	high	no	yes
SW480	CHR13	39262935	C	A	FREM2	-	-	neutral	no	no
SW480	CHR16	19451553	T	A	TMC5	-	-	medium	no	no

SW480	CHR19	14273439	C	T	LPHN1	DAMAGING	DAMAGING	low	no	no
SW480	CHR20	49196439	A	G	PTPN1	-	-	neutral	no	yes
SW480	CHR2	54081568	G	A	GPR75	-	-	neutral	no	no
SW480	CHR3	9974300	A	G	IL17RC	-	-	neutral	no	no
SW480	CHR4	6290834	C	T	WFS1	-	-	medium	no	no
SW480	CHR5	140256021	G	A	PCDHA12	DAMAGING	-	high	no	no
SW480	CHR5	140750356	C	A	PCDHGB3	-	-	low	no	no
SW480	CHR5	140866239	G	A	PCDHGC4	-	-	neutral	no	no
SW480	CHR5	140866458	G	A	PCDHGC4	-	-	neutral	no	no
SW480	CHR5	150844730	G	T	SLC36A1	DAMAGING	DAMAGING	medium	no	yes
SW480	CHR6	149285598	C	T	UST	DAMAGING	-	medium	no	yes
SW480	CHRX	48319067	C	T	SLC38A5	DAMAGING	DAMAGING	medium	no	no
SW48	CHR10	104816683	C	T	CNNM2	DAMAGING	DAMAGING	medium	no	no
SW48	CHR10	105209329	G	A	CALHM2	DAMAGING	-	medium	no	no
SW48	CHR10	17659199	C	T	PTPLA	-	-	low	no	no
SW48	CHR10	85901284	C	T	GHITM	DAMAGING	-	medium	no	no
SW48	CHR10	95111323	T	A	MYOF	-	DAMAGING	medium	no	no
SW48	CHR11	124743754	T	G	ROBO3	DAMAGING	DAMAGING	medium	no	no
SW48	CHR11	124747950	G	A	ROBO3	-	-	medium	no	no
SW48	CHR11	124749361	G	A	ROBO3	-	-	low	no	no
SW48	CHR11	14865528	C	T	PDE3B	DAMAGING	DAMAGING	medium	no	yes
SW48	CHR1	116941271	G	A	ATP1A1	DAMAGING	DAMAGING	high	no	yes
SW48	CHR1	117127416	A	C	IGSF3	-	-	low	no	no
SW48	CHR1	117146318	G	A	IGSF3	DAMAGING	DAMAGING	neutral	no	no
SW48	CHR1	117509724	G	A	PTGFRN	DAMAGING	DAMAGING	low	no	no
SW48	CHR1	120471800	G	A	NOTCH2	DAMAGING	-	neutral	no	yes
SW48	CHR1	165533006	G	A	LRRC52	-	-	low	no	no
SW48	CHR11	66083901	C	T	CD248	-	-	neutral	no	no
SW48	CHR11	67811057	T	C	TCIRG1	-	-	neutral	no	no
SW48	CHR1	168201106	T	C	SFT2D2	-	-	low	no	no
SW48	CHR1	208391105	C	T	PLXNA2	-	-	medium	no	no
SW48	CHR12	13220116	A	T	KIAA1467	DAMAGING	DAMAGING	medium	no	no
SW48	CHR1	223396855	A	C	SUSD4	-	-	low	no	no
SW48	CHR1	236347106	C	T	GPR137B	-	-	low	no	no
SW48	CHR1	241803346	G	A	OPN3	DAMAGING	DAMAGING	medium	no	no
SW48	CHR12	44781990	A	C	TMEM117	-	-	low	no	no
SW48	CHR12	49165507	C	T	ADCY6	DAMAGING	DAMAGING	low	no	no
SW48	CHR12	52385715	C	T	ACVR1B	-	-	-	yes	yes
SW48	CHR12	57575015	C	T	LRP1	DAMAGING	-	medium	no	yes
SW48	CHR12	57605360	G	A	LRP1	-	-	neutral	no	yes
SW48	CHR12	71971740	C	T	LGR5	DAMAGING	-	medium	no	no
SW48	CHR1	27427705	C	T	SLC9A1	DAMAGING	-	medium	no	yes
SW48	CHR1	27480699	C	T	SLC9A1	-	-	low	no	yes
SW48	CHR12	7527197	T	C	CD163L1	-	-	neutral	no	no
SW48	CHR12	94965381	T	C	TMCC3	-	-	neutral	no	no
SW48	CHR1	29611356	G	T	PTPRU	-	DAMAGING	low	no	no
SW48	CHR14	105612974	C	T	JAG2	-	DAMAGING	medium	no	no
SW48	CHR14	23310721	T	C	MMP14	DAMAGING	DAMAGING	medium	no	yes
SW48	CHR14	23344638	G	A	LRP10	DAMAGING	-	medium	no	no
SW48	CHR14	24658853	C	T	TM9SF1	-	-	low	no	no
SW48	CHR1	44069698	A	G	PTPRF	-	-	low	no	no
SW48	CHR1	44086778	G	A	PTPRF	DAMAGING	DAMAGING	high	no	no
SW48	CHR1	45297396	C	A	PTCH2	DAMAGING	-	medium	no	no
SW48	CHR14	96706990	G	A	BDKRB2	-	DAMAGING	medium	no	yes
SW48	CHR15	50897127	C	T	TRPM7	DAMAGING	DAMAGING	medium	yes	yes
SW48	CHR1	57002662	G	A	PPAP2B	DAMAGING	DAMAGING	medium	no	no
SW48	CHR15	73552705	T	A	NEO1	DAMAGING	DAMAGING	low	no	no
SW48	CHR15	73994743	T	C	CD276	DAMAGING	-	medium	no	no
SW48	CHR16	2327650	T	C	ABCA3	DAMAGING	DAMAGING	medium	no	no
SW48	CHR16	50324539	G	A	ADCY7	-	-	neutral	no	no
SW48	CHR16	68713872	C	T	CDH3	DAMAGING	DAMAGING	high	no	no
SW48	CHR16	726995	G	C	RHBDL1	-	-	neutral	no	no
SW48	CHR16	88782225	C	A	FAM38A	DAMAGING	DAMAGING	medium	no	no

SW48	CHR16	919988	C	A	LMF1	-	-	medium	no	no
SW48	CHR17	48141461	G	A	ITGA3	-	-	neutral	no	no
SW48	CHR17	65026704	C	A	CACNG4	DAMAGING	DAMAGING	medium	no	no
SW48	CHR18	55398934	G	A	ATP8B1	-	-	-	no	no
SW48	CHR18	77089301	G	A	ATP9B	-	-	low	no	no
SW48	CHR19	11034847	G	A	YIPF2	DAMAGING	DAMAGING	medium	no	no
SW48	CHR19	14938998	T	G	OR7A5	DAMAGING	-	neutral	no	no
SW48	CHR19	17330093	A	G	USE1	DAMAGING	DAMAGING	medium	no	no
SW48	CHR19	19762564	C	T	ATP13A1	DAMAGING	DAMAGING	high	no	no
SW48	CHR19	42489532	C	T	ATP1A3	DAMAGING	DAMAGING	high	no	no
SW48	CHR19	4294607	C	T	TMIGD2	-	-	-	no	no
SW48	CHR19	7594587	T	C	MCOLN1	-	-	low	no	no
SW48	CHR20	14306391	C	T	FLRT3	DAMAGING	DAMAGING	low	no	no
SW48	CHR20	3017913	G	T	PTPRA	DAMAGING	DAMAGING	high	no	no
SW48	CHR20	44828182	C	T	CDH22	-	-	neutral	no	no
SW48	CHR20	4893561	C	T	SLC23A2	-	DAMAGING	neutral	no	yes
SW48	CHR2	103321078	C	T	SLC9A2	-	-	-	no	no
SW48	CHR21	42622828	G	T	BACE2	DAMAGING	-	medium	no	no
SW48	CHR2	220085879	C	A	ATG9A	-	-	neutral	no	no
SW48	CHR2	220502919	A	G	SLC4A3	DAMAGING	DAMAGING	medium	no	no
SW48	CHR2	223423282	A	G	SGPP2	-	-	neutral	no	no
SW48	CHR2	74902493	G	A	SEMA4F	DAMAGING	-	medium	no	no
SW48	CHR2	9008616	C	A	MBOAT2	-	-	low	no	no
SW48	CHR2	9661445	T	A	ADAM17	DAMAGING	DAMAGING	medium	no	yes
SW48	CHR2	99154426	G	A	INPP4A	-	-	low	no	no
SW48	CHR3	121720670	C	A	ILDR1	-	-	-	no	no
SW48	CHR3	14526474	C	T	SLC6A6	-	-	low	no	no
SW48	CHR3	172428858	G	A	NCEH1	-	-	neutral	no	no
SW48	CHR3	19491679	G	A	KCNH8	-	-	neutral	no	no
SW48	CHR3	196674344	A	G	PIGZ	DAMAGING	DAMAGING	low	no	no
SW48	CHR3	48665467	G	A	SLC26A6	-	-	low	no	no
SW48	CHR3	48685756	C	T	CELSR3	-	-	medium	no	no
SW48	CHR3	49755908	C	T	AMIGO3	-	-	medium	no	no
SW48	CHR3	49936300	C	A	MST1R	DAMAGING	DAMAGING	low	yes	yes
SW48	CHR3	53137969	C	T	RFT1	-	DAMAGING	medium	no	no
SW48	CHR4	16010663	C	T	PROM1	-	-	low	no	no
SW48	CHR4	48035037	C	T	NIPAL1	DAMAGING	DAMAGING	low	no	no
SW48	CHR4	6303551	G	A	WFS1	-	-	medium	no	no
SW48	CHR4	84511381	G	A	AGPAT9	DAMAGING	-	low	no	no
SW48	CHR4	84511382	T	G	AGPAT9	DAMAGING	-	medium	no	no
SW48	CHR5	140264014	C	A	PCDHA13	DAMAGING	-	medium	no	no
SW48	CHR5	1403077	G	A	SLC6A3	-	-	low	no	yes
SW48	CHR5	140719392	C	T	PCDHGA2	-	-	low	no	no
SW48	CHR5	140741197	C	T	PCDHGB2	-	-	medium	no	no
SW48	CHR5	140741758	G	A	PCDHGB2	-	-	low	no	no
SW48	CHR5	140751666	C	T	PCDHGB3	-	-	neutral	no	no
SW48	CHR5	140801351	G	T	PCDHGA11	DAMAGING	-	medium	no	no
SW48	CHR5	55237601	G	A	IL6ST	-	DAMAGING	low	no	no
SW48	CHR6	122766202	C	A	SERINC1	DAMAGING	DAMAGING	high	no	no
SW48	CHR6	14118286	C	T	CD83	-	-	low	no	no
SW48	CHR6	46800989	C	A	MEP1A	DAMAGING	-	medium	no	no
SW48	CHR6	70459254	C	T	LMBRD1	-	-	low	no	no
SW48	CHR7	100175513	A	C	LRCH4	-	-	low	no	no
SW48	CHR7	100179785	T	C	LRCH4	DAMAGING	DAMAGING	high	no	no
SW48	CHR7	100410758	A	G	EPHB4	DAMAGING	DAMAGING	medium	yes	yes
SW48	CHR7	105636730	G	T	CDHR3	DAMAGING	DAMAGING	medium	no	no
SW48	CHR7	17930085	T	C	SNX13	DAMAGING	DAMAGING	medium	no	no
SW48	CHR7	47851597	G	A	PKD1L1	-	-	neutral	no	no
SW48	CHR7	47971561	G	A	PKD1L1	-	-	neutral	no	no
SW48	CHR7	55241707	G	A	EGFR	DAMAGING	DAMAGING	high	yes	yes
SW48	CHR7	73183733	C	T	CLDN3	-	-	medium	no	no
SW48	CHR7	92848737	A	C	HEPACAM2	DAMAGING	DAMAGING	neutral	no	no
SW48	CHR8	17507364	C	T	MTUS1	-	-	low	no	no

SW48	CHR8	28420410	G	A	FZD3	-	-	low	no	no
SW48	CHR8	29924342	T	C	TMEM66	DAMAGING	-	medium	no	no
SW48	CHR8	30623872	A	G	UBXN8	-	-	-	no	no
SW48	CHR8	37688390	C	T	GPR124	-	-	low	no	no
SW48	CHR8	38275434	C	A	FGFR1	DAMAGING	-	low	yes	yes
SW48	CHR9	108127907	A	C	SLC44A1	-	DAMAGING	medium	no	no
SW48	CHR9	131105562	C	T	SLC27A4	DAMAGING	DAMAGING	medium	no	no
SW48	CHR9	98239971	C	T	PTCH1	DAMAGING	-	medium	no	yes
SW48	CHR9	98278975	C	T	PTCH1	-	-	-	no	yes
SW48	CHRX	37587576	G	A	XK	-	-	low	no	no
SW620	CHR1	110464481	A	C	CSF1	DAMAGING	DAMAGING	medium	no	yes
SW620	CHR11	113105855	C	T	NCAM1	-	-	-	no	no
SW620	CHR1	116932195	G	A	ATP1A1	-	-	low	no	yes
SW620	CHR1	27440424	C	T	SLC9A1	DAMAGING	DAMAGING	high	no	yes
SW620	CHR15	79614468	C	A	TMED3	DAMAGING	DAMAGING	medium	no	no
SW620	CHR16	2161275	C	A	PKD1	-	-	low	no	no
SW620	CHR17	37815309	G	A	STARD3	-	-	low	no	no
SW620	CHR19	49713497	G	A	TRPM4	DAMAGING	DAMAGING	low	no	no
SW620	CHR3	37574866	G	C	ITGA9	-	-	low	no	no
SW620	CHR3	45132911	C	G	CDCP1	-	-	neutral	no	no
SW620	CHR4	170912813	C	A	MFAP3L	-	DAMAGING	low	no	no
SW620	CHR4	30725149	G	A	PCDH7	DAMAGING	DAMAGING	medium	no	no
SW620	CHR4	6290834	C	T	WFS1	DAMAGING	-	medium	no	no
SW620	CHR5	140774131	G	A	PCDHGA8	DAMAGING	-	low	no	no
SW620	CHR5	140866239	G	A	PCDHGC4	-	-	neutral	no	no
SW620	CHR5	140866458	G	A	PCDHGC4	-	-	neutral	no	no
SW620	CHR6	149285598	C	T	UST	DAMAGING	-	medium	no	yes
SW620	CHR6	29692844	C	A	HLA-F	-	-	high	no	no
SW620	CHR7	47917167	C	T	PKD1L1	-	-	low	no	no
SW620	CHRX	109247337	C	G	TMEM164	DAMAGING	DAMAGING	medium	no	no
SW837	CHR11	119183296	C	T	MCAM	DAMAGING	DAMAGING	medium	no	no
SW837	CHR11	76371747	C	T	LRRC32	-	-	-	no	no
SW837	CHR1	26189336	A	G	PAQR7	-	-	low	no	no
SW837	CHR1	27434256	C	T	SLC9A1	-	-	low	no	yes
SW837	CHR14	77686380	G	T	TMEM63C	-	-	-	no	no
SW837	CHR16	19126638	G	C	ITPRIPL2	-	-	neutral	no	no
SW837	CHR17	4342997	C	T	SPNS3	DAMAGING	-	neutral	no	no
SW837	CHR17	66274385	G	A	SLC16A6	-	-	neutral	no	yes
SW837	CHR7	90894227	G	C	FZD1	-	-	neutral	no	no
SW837	CHR8	145066637	G	A	GRINA	DAMAGING	DAMAGING	low	no	no
SW837	CHR8	38884212	T	C	ADAM9	DAMAGING	-	low	no	no
SW948	CHR12	57593066	G	T	LRP1	-	-	-	no	yes
SW948	CHR15	73547132	C	T	NEO1	-	-	neutral	no	no
SW948	CHR16	74753022	C	T	FA2H	-	-	low	no	no
SW948	CHR3	52294451	C	T	WDR82	DAMAGING	-	low	no	no
SW948	CHR5	140800862	C	T	PCDHGA11	-	-	low	no	no
SW948	CHR7	129841781	C	G	TMEM209	-	DAMAGING	medium	no	no
SW948	CHR7	142569515	C	T	TRPV6	-	-	low	no	no
SW948	CHR7	86542399	C	A	KIAA1324L	DAMAGING	DAMAGING	medium	no	no
SW948	CHR7	92844905	C	A	HEPACAM2	DAMAGING	DAMAGING	low	no	no
SW948	CHR8	37693087	G	A	GPR124	DAMAGING	DAMAGING	medium	no	no
T84	CHR1	109803761	G	C	CELSR2	-	-	low	no	no
T84	CHR1	117487543	C	T	PTGFRN	DAMAGING	DAMAGING	medium	no	no
T84	CHR1	160320003	T	G	NCSTN	-	-	medium	no	yes
T84	CHR11	60694754	G	T	TMEM132A	-	-	low	no	no
T84	CHR11	68133078	A	C	LRP5	-	-	high	no	no
T84	CHR11	70033939	G	A	ANO1	-	-	neutral	no	no
T84	CHR11	70033988	A	G	ANO1	-	-	low	no	no
T84	CHR1	180135658	G	T	QSOX1	DAMAGING	DAMAGING	high	no	no
T84	CHR11	824794	G	C	PNPLA2	-	-	neutral	no	no
T84	CHR1	46158981	C	T	TMEM69	-	-	neutral	no	no
T84	CHR14	94582130	T	C	IFI27	-	-	neutral	no	no
T84	CHR16	16208860	T	C	ABCC1	DAMAGING	-	low	no	yes

T84	CHR16	3021780	T	G	PAQR4	DAMAGING	DAMAGING	medium	no	no
T84	CHR16	3071795	G	C	TNFRSF12A	-	-	low	no	no
T84	CHR1	6522200	G	A	TNFRSF25	DAMAGING	-	medium	no	no
T84	CHR16	626214	G	A	PIGQ	DAMAGING	-	low	no	no
T84	CHR16	88782205	G	C	FAM38A	-	-	medium	no	no
T84	CHR17	60749449	G	T	MRC2	DAMAGING	DAMAGING	medium	no	no
T84	CHR19	7592459	C	T	MCOLN1	-	-	neutral	no	no
T84	CHR20	4880328	C	T	SLC23A2	DAMAGING	DAMAGING	low	no	yes
T84	CHR2	97526695	G	T	SEMA4C	-	-	neutral	no	no
T84	CHR3	37818946	G	T	ITGA9	-	-	-	no	no
T84	CHR3	49153185	G	T	USP19	-	-	neutral	no	no
T84	CHR4	47593282	A	G	ATP10D	-	-	neutral	no	no
T84	CHR4	79173594	C	T	FRAS1	-	-	medium	no	no
T84	CHR6	157963727	T	C	ZDHHC14	DAMAGING	-	low	no	no
T84	CHR6	88218823	A	C	SLC35A1	DAMAGING	DAMAGING	high	no	no

Anexo B: Tabela com InDels em genes expressos

InDels encontradas em genes expressos nas linhagens									
Linhagem	Chr	Posição	Referência	Variante	Gene	Classificação	SIFT	Kinoma	Drogável
CACO2	CHR8	95189865	ACAG	A	CDH17	InFrame	-	no	no
CACO2	CHR14	23524363	G	GG	CDH24	Frameshift	DAMAGING	no	no
CACO2	CHR9	5811126	G	GAGC	ERMP1	InFrame	DAMAGING	no	no
CACO2	CHR8	27516295	A	AGA	SCARA3	Frameshift	DAMAGING	no	no
CACO2	CHR19	54676754	G	GGGAA	TMC4	Frameshift	DAMAGING	no	no
COLO205	CHR19	41743939	C	CC	AXL	Frameshift	DAMAGING	yes	no
COLO205	CHR9	111853309	T	TCT	C9orf5	Frameshift	DAMAGING	no	no
COLO205	CHR4	30725829	G	GCATG	PCDH7	Frameshift	DAMAGING	no	no
COLO205	CHR3	125725274	ACAAA	A	SLC41A3	Frameshift	-	no	no
COLO205	CHR11	57268783	G	GTG	SLC43A1	Frameshift	DAMAGING	no	no
HCC2998	CHR10	105209356	CAG	C	CALHM2	Frameshift	DAMAGING	no	no
HCC2998	CHR1	223178805	TT	T	DISP1	Frameshift	-	no	no
HCC2998	CHR12	9098999	A	AACA	M6PR	InFrame	-	no	yes
HCC2998	CHR5	140795219	A	AA	PCDHGA10	Frameshift	-	no	no
HCC2998	CHR3	125725274	ACAAA	A	SLC41A3	Frameshift	-	no	no
HCT116	CHR3	53760966	CC	C	CACNA1D	Frameshift	DAMAGING	no	yes
HCT116	CHR16	68835768	GG	G	CDH1	Frameshift	DAMAGING	no	yes
HCT116	CHR14	23523734	G	GG	CDH24	Frameshift	DAMAGING	no	no
HCT116	CHR12	7301643	CC	C	CLSTN3	Frameshift	DAMAGING	no	no
HCT116	CHR1	35250498	GG	G	GJB3	Frameshift	DAMAGING	no	no
HCT116	CHR12	57567701	C	CC	LRP1	Frameshift	DAMAGING	no	yes
HCT116	CHR12	57606243	GG	G	LRP1	Frameshift	-	no	yes
HCT116	CHR12	9098999	A	AACA	M6PR	InFrame	-	no	yes
HCT116	CHR4	31144325	AA	A	PCDH7	Frameshift	-	no	no
HCT116	CHR5	140812784	TTTT	T	PCDHGA12	InFrame	-	no	no
HCT116	CHR7	47869147	AA	A	PKD1L1	Frameshift	DAMAGING	no	no
HCT116	CHR1	44084781	A	AA	PTPRF	Frameshift	DAMAGING	no	no
HCT116	CHR1	156354355	C	CC	RHBG	Frameshift	DAMAGING	no	no
HCT116	CHR17	56448302	GG	G	RNF43	Frameshift	DAMAGING	no	no
HCT116	CHR19	52034556	G	GTG	SIGLEC6	Frameshift	DAMAGING	no	no
HCT116	CHR2	230911099	TCTTCT	T	SLC16A14	Frameshift	DAMAGING	no	no
HCT116	CHR5	1403170	GG	G	SLC6A3	Frameshift	DAMAGING	no	yes
HCT116	CHR11	60703977	CC	C	TMEM132A	Frameshift	DAMAGING	no	no
HCT116	CHR9	74360136	T	TT	TMEM2	Frameshift	-	no	no
HCT15	CHR3	184071145	T	TGGGCT	CLCN2	Frameshift	DAMAGING	no	no
HCT15	CHR5	76128977	GG	G	F2RL1	Frameshift	DAMAGING	no	no
HCT15	CHR17	42153280	CC	C	G6PC3	Frameshift	DAMAGING	no	no
HCT15	CHR12	57603947	C	CC	LRP1	Frameshift	-	no	yes
HCT15	CHR20	4850575	GGG	G	SLC23A2	Frameshift	-	no	yes
HCT15	CHR3	14513791	TT	T	SLC6A6	Frameshift	DAMAGING	no	no
HCT15	CHR6	149395147	CC	C	UST	Frameshift	DAMAGING	no	yes
HCT15	CHR17	1631225	T	TT	WDR81	Frameshift	DAMAGING	no	no
HT29	CHR16	88782164	G	GCG	FAM38A	Frameshift	DAMAGING	no	no
HT29	CHR19	42879966	T	TTGTGC	MEGF8	Frameshift	DAMAGING	no	no
KM12	CHR19	41744389	CC	C	AXL	Frameshift	DAMAGING	yes	no
KM12	CHR19	49458978	GA	GGA	BAX	Frameshift	DAMAGING	no	no
KM12	CHR2	203407127	AA	A	BMPR2	Frameshift	DAMAGING	no	no
KM12	CHR3	184073514	GG	G	CLCN2	Frameshift	DAMAGING	no	no
KM12	CHR7	73184149	C	CGTCAA	CLDN3	Frameshift	DAMAGING	no	no
KM12	CHR14	39819375	A	AACCCT	CTAGE5	Frameshift	-	no	no
KM12	CHR8	145689666	CC	C	CYHR1	Frameshift	DAMAGING	no	no
KM12	CHR4	1019073	CAC	C	FGFRL1	Frameshift	-	no	no
KM12	CHR6	160485495	G	GG	IGF2R	Frameshift	DAMAGING	no	yes
KM12	CHR12	57603947	C	CC	LRP1	Frameshift	DAMAGING	no	yes
KM12	CHR1	160324016	TTT	TCGCTT	NCSTN	InFrame	-	no	yes
KM12	CHR5	140263185	GG	G	PCDHA13	Frameshift	DAMAGING	no	no
KM12	CHR5	140764780	C	CC	PCDHGA7	Frameshift	DAMAGING	no	no
KM12	CHR14	71443764	GT	G	PCNX	Frameshift	DAMAGING	no	no
KM12	CHR12	94691191	T	TA	PLXNC1	Frameshift	-	no	no
KM12	CHRX	37312618	C	CCC	PRRG1	Frameshift	DAMAGING	no	no
KM12	CHR1	227071476	G	GCGGCG	PSEN2	Frameshift	DAMAGING	no	yes

KM12	CHR1	45292631	GG	G	PTCH2	Frameshift	DAMAGING	no	no
KM12	CHR6	128302318	A	AA	PTPRK	Frameshift	DAMAGING	no	no
KM12	CHR11	66133905	CT	C	SLC29A2	Frameshift	DAMAGING	no	no
KM12	CHR12	94976263	C	CC	TMCC3	Frameshift	DAMAGING	no	no
KM12	CHR1	9667818	T	TCT	TMEM201	Frameshift	DAMAGING	no	no
LIM1215	CHR14	95669435	TT	T	CLMN	Frameshift	DAMAGING	no	no
LIM1215	CHR7	143095446	T	TTCCCC	EPHA1	Frameshift	-	yes	no
LIM1215	CHR4	1019075	C	CAC	FGFRL1	Frameshift	DAMAGING	no	no
LIM1215	CHR15	65703481	CT	C	IGDCC4	Frameshift	DAMAGING	no	no
LIM1215	CHR12	57603946	CC	C	LRP1	Frameshift	DAMAGING	no	yes
LIM1215	CHR4	30726094	AA	A	PCDH7	Frameshift	DAMAGING	no	no
LIM1215	CHR4	55156530	G	GCTGTG	PDGFRA	Frameshift	DAMAGING	yes	yes
LIM1215	CHR1	208315793	G	GG	PLXNA2	Frameshift	DAMAGING	no	no
LIM1215	CHR3	48666133	CC	C	SLC26A6	Frameshift	DAMAGING	no	no
LIM1215	CHR9	74360136	T	TT	TMEM2	Frameshift	DAMAGING	no	no
LIM1215	CHR17	42268039	CC	C	TMUB2	Frameshift	DAMAGING	no	no
LIM2405	CHR1	116930023	G	GG	ATP1A1	Frameshift	DAMAGING	no	yes
LIM2405	CHR10	125805505	TT	T	CHST15	Frameshift	DAMAGING	no	no
LIM2405	CHR10	125805650	GG	G	CHST15	Frameshift	DAMAGING	no	no
LIM2405	CHR3	172046795	CA	CACTGCA	FNDC3B	Frameshift	DAMAGING	no	no
LIM2405	CHR8	28385236	G	GG	FZD3	Frameshift	DAMAGING	no	no
LIM2405	CHR2	202900599	TG	T	FZD7	Frameshift	DAMAGING	no	no
LIM2405	CHR2	187521091	G	GG	ITGAV	Frameshift	DAMAGING	no	yes
LIM2405	CHR20	10625848	CC	C	JAG1	Frameshift	DAMAGING	no	no
LIM2405	CHR6	111587368	TT	T	KIAA1919	Frameshift	DAMAGING	no	no
LIM2405	CHR3	49935638	GG	G	MST1R	Frameshift	DAMAGING	yes	yes
LIM2405	CHR15	73581576	C	CC	NEO1	Frameshift	DAMAGING	no	no
LIM2405	CHR20	30737497	C	CC	TM9SF4	Frameshift	DAMAGING	no	no
LOVO	CHR2	148683685	TAAAAAAAA	TAAAAAAA	ACVR2A	Frameshift	-	no	no
LOVO	CHR17	29185288	TAAAAAAA	TAAAAAA	ATAD5	Frameshift	DAMAGING	no	no
LOVO	CHR18	55328619	TTTCTTCTTCT	TTTCTTCT	ATP8B1	InFrame	-	no	no
LOVO	CHR2	203420129	GAAAAAAA	GAAAAAA	BMPR2	Frameshift	DAMAGING	no	no
LOVO	CHR16	66603929	GTTTTTTTT	GTTTTTTT	CMTM1	Frameshift	DAMAGING	no	no
LOVO	CHR14	39819357	ATTTTT	ATTTT	CTAGE5	Frameshift	-	no	no
LOVO	CHR3	49568919	GCCCCC	GCCCC	DAG1	Frameshift	-	no	no
LOVO	CHR9	5801291	GTTTTTTTT	GTTTTTTT	ERMP1	Frameshift	DAMAGING	no	no
LOVO	CHR3	172114964	GAGAAGAAG	GAGAAG	FNDC3B	InFrame	DAMAGING	no	no
LOVO	CHR6	111587360	ATTTTTTTTT	ATTTTTTTT	KIAA1919	Frameshift	DAMAGING	no	no
LOVO	CHR19	11221403	TG	T	LDLR	Frameshift	DAMAGING	no	yes
LOVO	CHR7	97766671	CTGTGTGTGTG	CTGTGTGTG	LMTK2	Frameshift	DAMAGING	yes	no
LOVO	CHR4	151509210	ATTTTTTT	ATTTTTT	LRBA	Frameshift	DAMAGING	no	no
LOVO	CHR5	140720904	ATTTTTT	ATTTTT	PCDHGA2	Frameshift	DAMAGING	no	no
LOVO	CHR11	14880753	TAAAAA	TAAAA	PDE3B	Frameshift	DAMAGING	no	yes
LOVO	CHRX	37312605	ACCCC	ACCCCCC	PRRG1	Frameshift	DAMAGING	no	no
LOVO	CHRX	105937255	GTTTTTTTT	GTTTTTTT	RNF128	Frameshift	-	no	no
LOVO	CHR5	158630629	GTTTTTTTTTTCTTTTTTTTTT	GTTTTTTTTCTTTTTTTTTT	RNF145	Frameshift	-	no	no
LOVO	CHR9	131112834	TCCCCC	TCCCC	SLC27A4	Frameshift	DAMAGING	no	no
LOVO	CHR11	57185345	CTTTTT	CTTTTTT	SLC43A3	Frameshift	DAMAGING	no	no
LOVO	CHR20	3215424	CAAAAAAA	CAAAAAA	SLC4A11	Frameshift	DAMAGING	no	no
LOVO	CHR1	212548596	CTTTTTTT	CTTTTTTTT	TMEM206	Frameshift	DAMAGING	no	no
LOVO	CHR12	83444731	CAAAAAAA	CAAAAAA	TMTC2	Frameshift	DAMAGING	no	no
LOVO	CHR21	45837906	GCCCCCC	GCCCCC	TRPM2	Frameshift	DAMAGING	no	no
LOVO	CHR12	58140476	TCAACAA	TCAA	TSPAN31	InFrame	DAMAGING	no	no
LOVO	CHR1	160389124	GCCCC	GCCC	VANGL2	Frameshift	DAMAGING	no	no
LOVO	CHR22	20130521	GCCCCCC	GCCCCC	ZDHHC8	Frameshift	DAMAGING	no	no
LOVO	CHR22	20132875	TAAGAAG	TAAG	ZDHHC8	InFrame	DAMAGING	no	no
RKO	CHR3	183669316	GGA	G	ABCC5	Frameshift	DAMAGING	no	yes
RKO	CHR21	43697013	A	AATTA	ABCG1	Frameshift	DAMAGING	no	yes
RKO	CHR11	69934049	CC	C	ANO1	Frameshift	DAMAGING	no	no
RKO	CHR2	203420135	AA	A	BMPR2	Frameshift	DAMAGING	no	no
RKO	CHR11	125880259	TG	T	CDON	Frameshift	DAMAGING	no	no
RKO	CHR3	48698230	TCT	T	CELSR3	Frameshift	DAMAGING	no	no
RKO	CHR8	145689667	C	CC	CYHR1	Frameshift	DAMAGING	no	no

RKO	CHR4	187532799	A	AA	FAT1	Frameshift	DAMAGING	no	no
RKO	CHR4	1019073	CAC	C	FGFRL1	Frameshift	DAMAGING	no	no
RKO	CHR1	160063819	GTG	G	IGSF8	Frameshift	DAMAGING	no	no
RKO	CHR20	10630972	CC	C	JAG1	Frameshift	DAMAGING	no	no
RKO	CHR12	12303944	TT	T	LRP6	Frameshift	DAMAGING	no	no
RKO	CHR3	172365832	T	TT	NCEH1	Frameshift	DAMAGING	no	no
RKO	CHR5	140812785	TTT	T	PCDHGA12	Frameshift	-	no	no
RKO	CHR20	4850575	GGG	G	SLC23A2	Frameshift	DAMAGING	no	yes
RKO	CHR8	22262236	CTGCTGC	C	SLC39A14	InFrame	DAMAGING	no	yes
RKO	CHR1	27428611	GG	G	SLC9A1	Frameshift	DAMAGING	no	yes
RKO	CHR17	7324623	C	CC	SPEM1	Frameshift	DAMAGING	no	no
RKO	CHR19	51135703	CTGC	C	SYT3	InFrame	DAMAGING	no	no
RKO	CHR1	160393932	GG	G	VANGL2	Frameshift	DAMAGING	no	no
RW2982	CHR16	88787607	CCTTCTTC	CCTTC	FAM38A	InFrame	-	no	no
RW2982	CHR16	19126043	TCTCCTC	TCTC	ITPRIPL2	InFrame	DAMAGING	no	no
RW2982	CHR11	5475021	GTTT	GT	OR51I2	Frameshift	DAMAGING	no	no
RW2982	CHR1	156354347	TCCCCCCCC	TCCCCCCC	RHBG	Frameshift	-	no	no
RW7213	CHR16	88787607	CCTTCTTC	CCTTC	FAM38A	InFrame	DAMAGING	no	no
SKCO1	CHR19	36230499	CAG	C	IGFLR1	Frameshift	DAMAGING	no	no
SW1116	CHR3	184071131	CCGGCGG	CCGGCGGCGG	CLCN2	InFrame	-	no	no
SW1116	CHR5	140772898	GCCCC	GCCC	PCDHGA8	Frameshift	-	no	no
SW480	CHR5	159344459	G	GAA	ADRA1B	Frameshift	DAMAGING	no	yes
SW480	CHR19	17441667	TT	T	ANO8	Frameshift	DAMAGING	no	no
SW480	CHR1	160327026	G	GAATCG	NCSTN	Frameshift	DAMAGING	no	yes
SW480	CHR11	5474927	A	AC	OR51I2	Frameshift	DAMAGING	no	no
SW480	CHR1	44086146	C	CA	PTPRF	Frameshift	DAMAGING	no	no
SW480	CHR1	29631319	A	AC	PTPRU	Frameshift	DAMAGING	no	no
SW48	CHR2	148683685	TAAAAAAAA	TAAAAAAA	ACVR2A	Frameshift	DAMAGING	no	no
SW48	CHR3	105258916	TCCCCC	TCCCC	ALCAM	Frameshift	DAMAGING	no	no
SW48	CHR19	17441667	TT	T	ANO8	Frameshift	DAMAGING	no	no
SW48	CHR4	47559697	TGGGGGGG	TGGGGGG	ATP10D	Frameshift	DAMAGING	no	no
SW48	CHR18	55365039	ATTTTTTTT	ATTTTTTT	ATP8B1	Frameshift	DAMAGING	no	no
SW48	CHR20	3564725	GCCCCC	GCCCC	ATRN	Frameshift	DAMAGING	no	no
SW48	CHR1	32202283	AAAGAAGAAG	AAAGAAG	BAI2	InFrame	DAMAGING	no	no
SW48	CHR1	32222151	CGGGGGG	CGGGGG	BAI2	Frameshift	DAMAGING	no	no
SW48	CHR14	23517513	CGGGGGGG	CGGGGG	CDH24	Frameshift	DAMAGING	no	no
SW48	CHR1	109801512	GCCTCCTCCTCC	GCCTCCTCC	CELSR2	InFrame	DAMAGING	no	no
SW48	CHR12	102108337	GTTTTTTTT	GTTTTTTT	CHPT1	Frameshift	DAMAGING	no	no
SW48	CHR8	145689658	GCCCCCCCCC	GCCCCCCCC	CYHR1	Frameshift	DAMAGING	no	no
SW48	CHR6	170594673	GCCCCCCC	GCCCCCC	DLL1	Frameshift	DAMAGING	no	no
SW48	CHR1	16462198	CGGGGGG	CGGGGG	EPHA2	Frameshift	DAMAGING	yes	yes
SW48	CHR4	1801137	TGGGG	TGGG	FGFR3	Frameshift	DAMAGING	yes	yes
SW48	CHR13	20717071	GCTCTCTCTCT	GCTCTCTCT	GJA3	Frameshift	DAMAGING	no	no
SW48	CHR1	100534121	CTTTTTTTT	CTTTTTTT	HIAT1	Frameshift	DAMAGING	no	no
SW48	CHR15	65676731	TGGGGG	TGGGG	IGDCC4	Frameshift	DAMAGING	no	no
SW48	CHRX	70327613	TGGGGGGG	TGGGGGG	IL2RG	Frameshift	DAMAGING	no	yes
SW48	CHR12	26553091	CA	CAA	ITPR2	Frameshift	DAMAGING	no	no
SW48	CHR1	202183357	TG	T	LGR6	Frameshift	-	no	no
SW48	CHR11	76372197	TCCCCC	TCCCCCC	LRRC32	Frameshift	DAMAGING	no	no
SW48	CHR1	90400066	CGGGGG	CGGGG	LRRC8D	Frameshift	DAMAGING	no	no
SW48	CHR5	94050577	CTTTTTTT	CTTTTTT	MCTP1	Frameshift	DAMAGING	no	no
SW48	CHR17	60755932	TGGGGG	TGGGG	MRC2	Frameshift	DAMAGING	no	no
SW48	CHR11	5474927	A	AC	OR51I2	Frameshift	DAMAGING	no	no
SW48	CHR5	140751068	CAAAA	CAAA	PCDHGB3	Frameshift	-	no	no
SW48	CHR9	98211548	TGGGGGG	TGGGGG	PTCH1	Frameshift	DAMAGING	no	yes
SW48	CHR11	73101831	GCCCCCC	GCCCCC	RELT	Frameshift	DAMAGING	no	no
SW48	CHR8	101300215	CTTTTTTT	CTTTTTT	RNF19A	Frameshift	DAMAGING	no	no
SW48	CHR17	56435160	ACCCCCCC	ACCCCCC	RNF43	Frameshift	DAMAGING	no	no
SW48	CHR17	56437564	CCACACA	CCACA	RNF43	Frameshift	DAMAGING	no	no
SW48	CHR5	149357646	CTTTTTTT	CTTTTTT	SLC26A2	Frameshift	DAMAGING	no	no
SW48	CHR17	79219741	TCCCCCCC	TCCCCCC	SLC38A10	Frameshift	DAMAGING	no	no
SW48	CHR2	220502412	GCCCCCCC	GCCCCCCCC	SLC4A3	Frameshift	DAMAGING	no	no
SW48	CHR12	21355480	TAAAAAA	TAAAAA	SLCO1B1	Frameshift	DAMAGING	no	no

SW48	CHR6	132793429	TGGGGG	TGGGG	STX7	Frameshift	DAMAGING	no	no
SW48	CHR11	85342821	GAAAAAAA	GAAAAAA	TMEM126B	Frameshift	DAMAGING	no	no
SW48	CHR6	75994292	GCCCCCC	GCCCCC	TMEM30A	Frameshift	-	no	no
SW48	CHR7	77423459	CTTTTTTTTT	CTTTTTTTT	TMEM60	Frameshift	DAMAGING	no	no
SW48	CHR1	156255497	GCCCCCCC	GCCCCCC	TMEM79	Frameshift	DAMAGING	no	no
SW48	CHR8	30601786	CGG	CG	UBXN8	Frameshift	-	no	no
SW48	CHR3	49147679	TGGGGGG	TGGGGG	USP19	Frameshift	DAMAGING	no	no
SW48	CHR12	6574048	ATACTTACTTAC	ATACTTAC	VAMP1	Frameshift	-	no	no
SW48	CHR17	1631340	TGAGGAGGAGGAG	TGAGGAGGAG	WDR81	InFrame	-	no	no
SW620	CHR9	111853308	C	CCTC	C9orf5	InFrame	DAMAGING	no	no
SW620	CHR12	94691191	T	TA	PLXNC1	Frameshift	DAMAGING	no	no
SW620	CHR17	79219504	ATGA	A	SLC38A10	InFrame	DAMAGING	no	no
SW620	CHR17	1636866	A	ACC	WDR81	Frameshift	DAMAGING	no	no
SW837	CHR8	120220771	GCCCC	GCCC	MAL2	Frameshift	-	no	no
SW837	CHR2	197729757	ATT	ATTT	PGAP1	Frameshift	DAMAGING	no	no
SW837	CHR1	156354347	TCCCCCCCC	TCCCCCCC	RHBG	Frameshift	-	no	no
T84	CHR5	140772898	GCCCC	GCCC	PCDHGA8	Frameshift	DAMAGING	no	no
T84	CHR11	130784840	CCTTCTTC	CCTTC	SNX19	InFrame	DAMAGING	no	no

ANEXO C: Artigo submetido para publicação: ICRmax: an optimized approach to detect tumor-specific InterChromosomal Rearrangements for Clinical Application

Title: ICRmax: an optimized approach to detect tumor-specific **I**nter**C**hromosomal **R**earrangements for Clinical Application

Elisa R. Donnard[1,2], Paola A. Carpinetti[1,2], Fábio C. P. Navarro[1,2], Rodrigo O. Perez[3,4], Angelita Habr-Gama[3], Raphael B. Parmigiani[1], Anamaria A. Camargo[1,4], Pedro A. F. Galante[1,#].

1- Centro de Oncologia Molecular - Hospital Sírio-Libanês – São Paulo – Brazil

2- Departamento de Bioquímica, Instituto de Química - Universidade de São Paulo – São Paulo – Brazil

3- Instituto Angelita & Joaquim Gama – São Paulo – Brazil

4- Ludwig Institute for Cancer Research – São Paulo - Brazil

- To whom the correspondence should be addressed: pgalante@mochsl.org.br

Author emails:

ERD: edonnard@mochsl.org.br

PAC: pcarpinetti@mochsl.org.br

FCPN: fnavarro@mochsl.org.br

ROP: rodrigo.operez@gmail.com

AH-G: gamange@uol.com.br

RBP: rparmigiani@mochsl.org.br

AAC: aacamargo@mochsl.org.br

PAFG: pgalante@mochsl.org.br

Keywords: cancer biomarkers, chromosomal rearrangements, clinical application.

ABSTRACT

BACKGROUND: Somatically acquired chromosomal rearrangements occur at early stages during tumor formation and can be used to indirectly detect tumor cells, serving as highly sensitive and specific tumor biomarkers. Recent advances in high throughput sequencing technology have allowed the genome-wide identification of patient-specific chromosomal rearrangements that can be used as personalized biomarkers to efficiently assess response to treatment, detect residual disease and monitor disease recurrence. However, sequencing and data processing costs still represent major obstacles for the widespread application of personalized biomarkers in oncology.

RESULTS: We developed a computational pipeline (ICRmax) designed for the cost-effective identification of a minimal set of tumor-specific interchromosomal rearrangements (ICRs) for clinical application. We examined ICRmax performance on our own sequencing data derived from six rectal tumors and simulated data achieving an average accuracy of 68% for ICR identification.

CONCLUSIONS: ICRmax allows the identification of interchromosomal translocations from low-coverage (3x) sequenced tumor genomes, eliminates the need to sequence a matched normal tissue genome and significantly reduces the costs that currently limit the utilization of personalized biomarkers in the clinical setting to approximately US$1,700 per patient.

INTRODUCTION

Somatically acquired chromosomal rearrangements, including duplications, inversions, deletions, insertions and translocations, constitute a key feature of tumor genomes (1-3). They occur at early stages during tumor formation and persist throughout tumor progression (4). These rearrangements are not present in normal cells from cancer patients and can thus be used to unequivocally detect tumor cells, serving as highly sensitive and specific tumor biomarkers to approach clinically relevant endpoints, such as assessment of response to therapy and detection of disease recurrence (5-7).

Technically, PCR detection of translocated DNA sequences that originate from different chromosomes or thousands of base pairs apart is a straightforward process when compared to PCR discrimination of single-base alterations and, therefore, tumor-specific chromosomal rearrangements represent ideal markers for monitoring tumor burden (5,8). Highly sensitive and specific assays developed to detect recurrent chromosomal translocations in hematological tumors have become standard practice to monitor residual disease and predict relapse to targeted therapy, allowing individualized therapeutic choices (9,10).

Unfortunately, a similar use of known chromosomal rearrangements in solid tumors has been hampered by the absence of recurrent rearrangements in these tumors (3). Recently, however, whole-genome sequencing has become efficient and affordable enough to allow genome-wide identification of patient-specific somatic chromosomal rearrangements (Campbell, et al., 2008; Mardis, et al., 2009; Stephens, et al., 2009; Stratton, et al., 2009; Pleasance, et al., 2010a,b) that could be used as personalized biomarkers to detect tumor cells (5).

Currently, analysis of short read paired-end or mate-pair sequences aligned against the human reference genome is the most efficient strategy to detect somatic chromosomal rearrangements present in tumor genomes. However, due to the repetitive nature of the human genome and the presence of structural variations, assigning the correct mapping positions to short reads and calling somatic chromosomal rearrangements are not straightforward processes (1,11), and a large number of false positive candidates are usually identified (3), generating unacceptable levels of noise. In order to reduce the high number of false positives candidates and analysis complexity, sequencing of a matched normal tissue DNA sample is usually required (12,13), increasing both the sequencing and computational costs, and further limiting the use of this strategy in actual clinical practice.

Here we present a pipeline (ICRmax) that allows the efficient identification of tumor-specific interchromosomal rearrangements. ICRmax has sets of filters that efficiently eliminate the need to sequence a matched normal genome and to allow the efficient identification of interchromosomal rearrangements from low-coverage (~3x) sequenced tumor genomes. The use of ICRmax significantly reduces the sequencing and computational costs associated with the identification of tumor-specific chromosomal rearrangements and will certainly contribute to the widespread use of personalized biomarkers in the routine clinical practice of patients with solid tumors.

RESULTS AND DISCUSSION

ICRmax implementation and filters

ICRmax was developed to efficiently identify a minimal set of reliable interchromosomal rearrangements from low coverage sequenced tumor genomes without the need to sequence a matched normal genome, reducing the sequencing cost and creating an opportunity to implement the use of personalized biomarkers in the routine clinical management of solid tumors.

ICRmax starts from sequence alignment data against the human reference genome in .bam format allowing users to apply their algorithm of choice for the alignment of mate-pair or pair-end reads. First, all reads with identical mapping coordinates are removed from further analysis, since they likely result from PCR duplicates generated during the library construction and amplification steps and can provide sequencing support to a false positive candidate (14-16). Next, .bam files are processed using BEDTools (17) and read pairs presenting an alignment against the human reference genome in the expected orientation and distance are discarded while aberrant pairs with reads mapping on different chromosomes are retained in the analysis.

In order to minimize reference genome assembly errors and structural polymorphisms, selected reads pairs are submitted through a second mapping step using alternative human genome assemblies and read pairs mapping on the same chromosome are excluded. Subsequently, read pairs mapping inside centromeric (+1Mb) and telomeric regions (+1Mb), and read pairs mapping to some of the regions defined by RepeatMasker (18) that may result in ambiguous mapping are removed. We also removed read pairs mapped in regions corresponding to segmental duplications (19) in order to avoid false-

positive rearrangements caused by the misalignment of read pairs in the parental and duplicated regions. Lastly we removed mate pairs with reads mapping to the mitochondrial genome since we observed that the presence of nuclear copies of mitochondrial DNA (*numts*) in the nuclear genome (20) leads to misalignments and the identification of false-positives.

Finally, read pairs are grouped (clustered) based on their genomic coordinates. Reads from different pairs mapped within a minimal distance corresponding to the average library insert size + 2 standard deviation (s.d.) on either side of the putative rearrangement are grouped. Putative ICRs are then ranked by the number of supporting pairs and we select those clusters composed of a minimum of three and a maximum of 80 pairs, reducing the impact of incorrect mapping and other artifacts. Both the clustering and this cutoff reduce substantially the number of reads indicating rearrangement events and simplify the last steps of the pipeline. For SOLiD platform data, the small number of remaining reads is then realigned using BLAT to remove further unreliable read mapping that may result from the initial alignment step for color space reads (see Methods). After removing these mate pairs, the minimum read pair support should once again be evaluated and only clusters with at least two reads respecting the same orientation pattern should be maintained (see Supplementary Figure 1). ICRmax pipeline is summarized in Figure 1 and a step-by-step command line is provided in the supplementary material and also available at http://www.bioinfo.mochsl.org.br/ICRmax/downloads (file ICRmax_pipeline.htm).

Identification of interchromosomal rearrangements in rectal tumors with ICRmax.

In order to show how our pipeline works, we used ICRmax to identify interchromosomal rearrangements in six rectal tumor genomes (RT1-RT6). For each sample, mate-pair libraries with average insert size of 600bp were generated using tumor genomic DNA and sequenced in a SOLiD sequencing platform to varying depths (see Table 1). Sequenced reads were aligned against the human genome reference sequence using Bioscope (Applied Biosystems) and only high quality (Q>20) alignments were selected for further analysis. Sequence coverage varied from 4 to 9x and the calculated physical coverage based on the average insert size of sequenced fragments varied between 15 to 62x. Sequencing and coverage information for each tumor sample is presented in Table 1. We also sequenced three matched normal genomes (N1, N5 and N6) and submitted the sequenced reads through the same pipeline to evaluate their contribution to exclude false positive events that were not filtered out by ICRmax. Sequence and physical coverage for the matched normal genomes varied from 3.2 to 9.3x and from 15 to 32x, respectively (Table 1).

As expected, a variable number of putative interchromosomal rearrangements were identified in each tumor genome (ranging from 9 to 105 events per sample, average of 32). Interestingly, when we compared the set of ICR candidates in each tumor genome we found a significant number of recurrent events (Table S1). Since solid tumors are well known for their lack of recurrent tumor-specific rearrangements (3) we assumed that these recurrent events (Table S1) might correspond to artifacts that could be removed by comparison with the matched normal samples (21-23). The presence of recurrent structural variation breakpoints was recently observed in a study of complex genomic

rearrangements (24) and as a validation strategy only breakpoints present in a single tumor were considered somatic. Accordingly, we selected eight tumor-recurrent events for validation by PCR amplification with primers flanking the putative breakpoint region determined by the alignment of mate pair reads. As expected for such false positive somatic events, amplification with specific primers was obtained when using both the tumor and matched normal DNA (Figure S2) and these cases are therefore not candidates for tumor-specific rearrangements. The list of recurrent artifacts is available as supplementary data (Table S2), or directly at http://www.bioinfo.mochsl.org.br/ICRmax/downloads (file recurrent_artifacts.bed).

Thus, ICRmax does not rely on the need to sequence matched normal/tumor genomes due to the implementation of additional filters to remove recurrent events and to eliminate false-positive rearrangements, reducing by at least half the sequencing cost, but still allowing the identification of a reliable minimal set of tumor specific ICRs for clinical application.

After removing recurrent artifacts and false-positive candidates, the number of identified events for each sample varied from 2 to 96, with an average of 23 per sample (detailed in Table 2). Rearrangements present in each sample are graphically represented in Figure 2 and Supplementary Figure 3. A subset of theses putative rearrangements was then selected for PCR validation followed by Sanger sequencing (Table 2). Of the 36 putative interchromossomal rearrangements selected for validation, 18 (50%) were confirmed as somatic tumor-specific events and their exact breakpoints were determined for most cases after sequencing (PCR amplification results are shown in Supplementary Figure 2). Validation efficiency between different samples varied from 9 to 90%, and

some patients, like RT4, provided more difficulty in the validation stage, possibly due to the presence of different levels of genetic instability and tumor heterogeneity. Figure 3 illustrates the genomic region, as well as the sequence coverage, of a confirmed rearrangement between chromosomes 1 and 17 detected in sample RT2. The exact breakpoint was determined by PCR amplification followed by Sanger sequencing of the amplified fragment (see Supplementary Figure 4). Interestingly, this rearrangement is located in the *TP53BP2* gene region, and loss of this functional gene product has been associated with gastric and other cancer susceptibility (25).

These validated rearrangements can be used in further investigation as biomarkers for their corresponding patients. The validation stage is essential in distinguishing tumor specific events from germline polymorphisms or false positives candidates. Furthermore, one could argue that studies can rely mainly on this step to select for tumor-specific events. However, the costs of primer design and the time spent on breakpoint amplification can be a limiting factor when identifying personalized chromosomal rearrangements in the clinical setting. The optimization of the bioinformatics analysis is therefore critical for defining a concise set of reliable rearrangements to substantially increase the efficiency of the entire process.

Comparison to a tumor-normal paired sequencing

As previously mentioned, to assess the contribution of sequencing matched normal genomes when using ICRmax, we sequenced matched normal DNA for three of our tumor samples. Sequencing and coverage information for all normal samples is presented in Table 1. We found that most (62%) of the ICR candidates detected in both

normal and tumor DNA were also detected in other tumor samples and therefore present in our list of recurrent artifacts. For instance, 13 events were detected in both N1 and RT1 samples derived from the same patient, but 10 of these rearrangements were also present in our list of tumor-recurrent artifacts. Interestingly the number of events filtered by the recurrent artifact list was for some of the samples greater than the number of events removed exclusively by comparison to the matched normal genomes. For example, for patient #1 a total of 3 rearrangements that were not identified by sequencing the matched normal genome were filtered out using the list of tumor-recurrent rearrangements and one extra event was removed by comparison with one of the other normal tissues sequenced. Similar results were observed for the other two patients for whom we have sequenced the matched normal DNA (Table S1). Noteworthy, for patient #6, sequencing the matched normal DNA (N6) removed only four rearrangements found in the tumor genome (and only one of these could not be removed by the tumor-recurrent list), likely due to the low coverage obtained for the normal sample (Table 1), suggesting that sequencing paired normal tissue with lower coverage is not an effective option to remove false-positive candidates. The matched normal tissue contribution therefore does not justify the increase in the sequencing cost, since a search for recurrent events in different tumor genomes allows the efficient identification of artifactual events.

Simulated rearrangement datasets

Overall, our pipeline was effective when applied to the tumor genomes sequenced and allowed the identification of a minimal subset of personalized interchromosomal rearrangements. However, tumor heterogeneity and genetic instability prevents both the

complete identification of rearrangements present in the tumor genome as well as a true estimate of our method's accuracy. In order to better evaluate our pipeline we simulated three human genomes (based on hg19) containing different numbers of ICRs. We randomly created 20, 30 and 40 ICR events for each genome (RG1, RG2 and RG3, respectively; for details, see Methods and Supplementary data). From each genome (RG1, RG2 and RG3) we also generated three sets of randomly sampled reads (mate pairs of 50nt and insert size of 700nt, on average), which represent a simulated physical coverage of 44x, 25x and 13x, respectively (see Table 3). Using ICRmax and increasing the minimal mate pair support to five reads, we were able to correctly identify 42 out of the 90 simulated rearrangements (47% specificity) and eight additional false positive candidates (84% accuracy; Table 4). As expected, for the low sequence coverage genomes (1.9x and 3.8x; Table 3), ICRmax performed with a lower sensitivity (42.5%), but a remarkable accuracy (100%). When we decrease the mate pair support requirement to three reads in the RG2 and RG3 simulated genomes, we detected a higher number of true simulated rearrangements (18/30 for RG2 and 20/40 for RG3) but a number of false positives also appear in the sets (14 for RG2 and 3 for RG3). The results presented in Table 4 suggest that our pipeline achieved an experimentally acceptable accuracy and sensitivity for ICR detection and performed well even under low sequencing (1.9x) and physical (13x) coverage (for RG3, for instance). Further details on the simulated reads can be found in the Methods section and results regarding our simulated data are presented in Tables 3 and 4 and Supplementary Figure 3.

Cost Analysis

Based on the data presented here, we estimated the final cost associated with the immediate application of this protocol in the clinical setting. At our facilities, the sequencing cost in the SOLiD 5500XL platform is approximately ~US$1 per 20 million bases. To obtain a 4X sequencing coverage for each patient sample, one needs to generate ~20 billion bases (400 million reads with 50nt), totalizing ~US$1,000 per patient. For each patient the final cost for the development of personalized chromosomal rearrangements should include additional costs for sample processing and mate-pair library construction (~US$470 per patient) as well as for PCR validation. Testing 10 putative rearrangements will cost ~US$200 in primer synthesis and ~US$10 for all PCR reactions. Assuming a 50% validation rate, subsequent Sanger sequencing of the five PCR-validated rearrangements for breakpoint determination would add ~US$50 reaching the final cost of ~US$1730 per patient.

CONCLUSION

Recent advances in high throughput sequencing technology have allowed the genome-wide identification of patient-specific chromosomal rearrangements (2,3,5,6,13). These personalized biomarkers are especially useful to assess response to treatment, detect residual disease and monitor disease recurrence and are expected to have a widespread use in clinical oncology (5,6). Unfortunately, current methods for chromosomal rearrangement detection require the sequencing of both tumor and matched normal genomes increasing the sequencing and computational costs and precluding the implementation of personalized biomarkers in clinical practice (12,13). ICRmax aims to

eliminate this need, relying on a set of strict mapping filters and inter-tumoral comparison to greatly reduce false positive candidates. However it is important to emphasize that the main goal is not to obtain the full set of ICR events present in the tumor genome, but to efficiently identify a minimal set of rearrangements which in any case can only represent part of a much larger set of tumor alterations due to the percentage of a solid tumor collected in a biopsy, for example. The minimal set of identified rearrangements can be validated and included into practice, reducing the cost and time required the identification of clinically useful personalized biomarkers.

METHODS

Sample preparation and whole genome sequencing

Biopsies from six locally advanced rectal tumors were obtained from patients treated at the Instituto Angelita & Joaquim Gama/Hospital Alemão Oswaldo Cruz (HAOC – São Paulo, Brazil; http://www.hospitalalemao.org.br). Tumor samples were collected after informed consent and approval of the HAOC Ethics Committee in Research. Blood samples from the same patients were also collected and circulating leucocytes were used as a source of normal matched germline DNA. Genomic DNA was extracted using a Trizol based protocol for simultaneous DNA/RNA extraction (26). Mate-pair libraries were generated for the SOLiD platform according to the manufacturer's instructions. In brief, 5ug of whole genome amplified (WGA) DNA samples were randomly sheared into 0.6-1.0 kb fragments using the Covaris S2 System. Fragmented DNA was size selected on 0.8% agarose gels and used as template in emulsion PCR. DNA fragments were coupled to bead via an adapter sequence and clonally amplified. Amplified DNA fragments were then covalently attached to a glass slide. Sequencing primers hybridized to the adapter sequence and fluorescently labeled di-base probes were used in ligation-based sequencing generating 50nt mate-pair reads.

Read alignment and selection of discordant mate pairs

All color space reads were aligned using standard BioScope mapreads (Applied Biosystems) algorithm against hg19/GRCh37 reference sequence downloaded from UCSC Genome Browser (http://genome.ucsc.edu/) (27). Only sequences with unique mapping and mapping quality greater than or equal to 20 (Q>=20) are used in subsequent

steps. Once the reads are paired, their orientation and mapping position are analyzed. Mate pairs showing same orientation and mapping within expected distance (mean insert size +/- 2s.d.) are discarded. For interchromosomal rearrangements, paired reads mapped in different chromosomes are selected and further investigated. Other mate pairs with anomalous mapping patterns can be selected to search for interchromosomal rearrangements such as insertions, deletions and inversions (28).

Selected reads presenting discordant interchromosomal mapping were submitted through a second alignment step using a set of three alternative genome assemblies: HuRef (J. Craig Venter Institute) (29), GRCh37_alt (partial reference genome with alternative representations - Genome Reference Consortium) and CRA (human chr7 complete sequence - The Center for Applied Genomics) (30). After mapping, all cases where both reads in a mate pair map in the same chromosome were removed from further analysis.

Finally, sequences with the same alignment (identical start or end positions) putatively resulting from PCR amplification in the library construction step are removed from the set.

Rearrangement identification

Sequences mapped on certain repetitive regions defined by RepeatMasker (18) (LINEs, SINEs, low complexity regions, satellites and rRNA genes) are removed, as well as reads mapped within segmental duplications (19). Centromeric and telomeric regions (downloaded from UCSC Genome Browser – http://genome.ucsc.edu) are expanded by

1Mb and reads mapped within those regions are also removed. Reads mapped to mitochondrial chromosome are also excluded.

Remaining mate pairs are clustered based on mapping positions and indication of the same rearrangement. For interchromosomal rearrangements, mate pairs with reads mapping in the same two chromosomes are compared and reads from each pair should map within a predetermined window to indicate the same event (windows are calculated by adding 2s.d. to mean insert size).

Once clusters are formed, the number of mate pairs supporting the event is evaluated. At this point all clusters supported by less than 3 and more than 80 mate pairs are removed. For the evaluation of the recurrent events, we lowered the mate pair support to two reads for the comparison between patients, keeping only events in the final set that could not be found in other patients even with the lower support.

For SOLiD sequencing data, remaining reads are submitted through a third and final mapping step. The FASTA sequences resulting from the initial alignment are used as input for BLAT (31). BLAT parameters used were the same indicated for reproducing the webtool results and can be found in Genome Browser (27,32). Similarly, reads from a single mate pair showing other mapping possibilities in the same chromosome and within expected distance from their mate pairs are removed from the final set. After removing these reads, support for the clusters is reevaluated and once again maintained only cases with 3 or more mate pairs indicating the event. The final sets of rearrangements found are represented in Figure 2 and Figure S3 using Circos (33).

Simulated datasets

Based on the human genome sequence (hg19), three sets (RG1, RG2 and RG3) of ICR were randomly generated by using Perl scripts. Briefly, we first randomly selected a chromosome and a genomic coordinate, followed by a second selection of another breakpoint in a different chromosome. The FASTA sequences of both chromosomes involved in the rearrangements were joined, creating a new chromosome. This process was repeated 20x for RG1, 30x for RG2 and 40x for RG3 (Supplementary Table 3). Chromosomes that were not involved in any simulated rearrangement were also kept in the final file. Finally, these genomes were then used to simulate color space sequenced reads with 50bp and 700bp insert size between mate pair reads. In order to simulate the real sequencing, the generated reads contain 1% sequencing errors. The number of reads generated was calculated based on the sequence and physical coverage we wished to obtain for each simulation relative to the size of the reference human genome (hg19).

Orientation patterns and primer design

Read orientation is also considered when defining the candidates for interchromosomal rearrangements. For the true positive candidates, mate pairs spanning a single breakpoint will show equal patterns of orientation. In case of SOLiD sequencing, mate pairs indicating the same breakpoint can be classified into four different patterns detailed in Figure S1. After dividing remaining mate pairs into different orientation patterns clusters still containing two or more mate pairs are chosen for validation. Clusters can end up divided into two sets of orientations, which most likely represent two separate breakpoints for an insertion or translocation. After observing read orientation,

support for the final clusters is lowered to at least two mate pairs indicating the rearrangement.

Breakpoint validation by PCR and sequencing

Primers were designed for a randomly chosen subset of the identified rearrangements for each patient. Mate-pair sequences flanking predicted breakpoints were used as target sequences for primer design using Primer3 when possible, with further manual evaluations necessary to adjust the sequences to the target regions. Mate-pair sequence order and orientation was used to guide primer design. When primers could not be designed from mate-pair sequences, upstream genomic sequence up to 200 bp was used for primer design. Primers were used for PCR on tumor and matched normal samples to confirm the somatic origin and tumor-specificity of the rearranged fragments. Sanger sequencing of PCR products was used confirm the specificity of the PCR amplification and to map breakpoint sequences.

AVAILABILITY AND REQUIREMENTS

The step-by-step command line and necessary files to implement the ICRmax pipeline are available at http://www.bioinfo.mochsl.org.br/ICRmax/downloads. The sequences used in this analysis are available at: The European Nucleotide Archive (ENA; www.ebi.ac.uk/ena) under accession number PRJEB4781.

AUTHORS' CONTRIBUTIONS

ERD designed and executed the pipeline, analyzed the data and wrote the manuscript. PAFG and AAC led the project, contributed to the concept, designed of the pipeline and wrote the manuscript. FCPN helped with the bioinformatics analysis. ROP and AH-G collected the samples. PAC and RBP performed the sample sequencing and validations. All the authors critically discussed the pipeline, read and approved the final manuscript.

ACKNOWLEDGMENTS

We thank Fabiana Bettoni, Paula Asprino, Fernanda Koyama, Bruna S de Quevedo, and Natalia Felício for their technical supports on the sample collecting, sample preparation or on the DNA sequencing. We also thank Daniel T. Ohara for technical support on the bioinformatics pipelines. AAC and ED are supported by FAPESP (11/50684-8 and 10/12658-2, respectively). This work was also supported by the Ludwig Institute for Cancer Research.

COMPETING INTERESTS

The authors declare they have no competing interests.

FIGURE LEGENDS

Figure 1. Bioinformatics pipeline showing the main steps for whole tumor genome analysis and selection of sequences indicating the presence of structural variation such as interchromosomal rearrangements.

Figure 2. Circos representation of the interchromosomal rearrangements found in 3 tumor genomes. a) RT1; 10 rearrangements b) RT2; 15 rearrangements c) RT3; 4 rearrangements.

Figure 3. Genome browser view of reads (green) indicating one of RT2 rearrangements between chromosome 17 (left) and chromosome 1 (right). Region shown for chromosome 17 ends in the breakpoint found after Sanger sequencing. Region shown for chromosome 1 also starts in the exact breakpoint nucleotide. Read orientation indicates that the rearrangement is structured as shown and confirmed through Sanger sequencing. The nucleotide sequence of the breakpoint region is shown above. Last nucleotide of chromosome 17 and first nucleotide of chromosome 1 are marked in red. Sequence coverage is shown in dark blue, only reads with mapping quality >= 20 are used.

TABLES

Table 1. Mapping results for rectal tumor samples and paired normal tissue samples submitted through whole genome sequencing.

Sample	Total reads	Mapped Nucleotides	Physical Coverage	Sequence Coverage
RT1	1,035,604,016	25,967,977,794 (50%)	180,593,981,498 (62x)	9.0x
RT2	393,756,912	12,232,074,899 (62%)	57,714,657,702 (20x)	4.2x
RT3	385,789,584	11,616,923,871 (60%)	53,821,366,172 (19x)	4.0x
RT4	398,436,826	12,231,994,959 (61%)	59,355,360,167 (20x)	4.2x
RT5	425,460,416	15,150,461,585 (51%)	42,211,056,664 (15x)	5.2x
RT6	991,625,036	20,921,867,961 (42%)	77,491,127,440 (27x)	7.2x
N1	728,574,754	21,532,212,957 (49%)	78,716,757,426 (27x)	7.4x
N5	682,517,714	26,824,738,972 (65%)	94,020,024,357 (32x)	9.3x
N6	305,101,394	9,129,663,072 (60%)	43,929,159,747 (15x)	3.2x

Table 2. Validation results

Sample	Number of clusters	Rearrangements after recurrent filter	Tested rearrangements	Validated as tumor specific
RT1	27	10	6	3 (50.0%)
RT2	18	15	10	9 (90.0%)
RT3	9	4	4	3 (75.0%)
RT4	105	96	11	1 (9.0%)
RT5	14	2	1	1 (100.0%)
RT6	19	14	4	1 (25.0%)

Table 3. Mapping results for genomes containing randomly simulated ICRs.

Simulated Genomes	Reads Generated	Mapped Reads	Sequence Coverage	Physical Coverage
RG1	402,771,620	336,548,722 (83.5%)	5.8x	44x
RG2	263,391,883	221,636,923 (84.1%)	3.8x	25x
RG3	131,695,960	110,878,325 (84.2%)	1.9x	13x

Table 4. Identification of simulated events by ICRmax.

Simulated Genomes	Simulated ICRs	Found ICRs (final set)	True positives (accuracy)	Sensitivity
RG1	20	21	13 (62%)	65% (13/20)
RG2	30	15	15 (100%)	50% (15/30)
RG3	40	14	14 (100%)	35% (14/40)
Total	90	50	42 (84%)	47% (42/90)

REFERENCES

1. Meyerson, M., Gabriel, S. and Getz, G. (2010) Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet*, **11**, 685-696.
2. Campbell, P.J., Stephens, P.J., Pleasance, E.D., O'Meara, S., Li, H., Santarius, T., Stebbings, L.A., Leroy, C., Edkins, S., Hardy, C. *et al.* (2008) Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat Genet*, **40**, 722-729.
3. Bass, A.J., Lawrence, M.S., Brace, L.E., Ramos, A.H., Drier, Y., Cibulskis, K., Sougnez, C., Voet, D., Saksena, G., Sivachenko, A. *et al.* (2011) Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 fusion. *Nat Genet*, **43**, 964-968.
4. Hanahan, D. and Weinberg, R.A. (2011) Hallmarks of cancer: the next generation. *Cell*, **144**, 646-674.
5. Leary, R.J., Kinde, I., Diehl, F., Schmidt, K., Clouser, C., Duncan, C., Antipova, A., Lee, C., McKernan, K., De La Vega, F.M. *et al.* (2010) Development of personalized tumor biomarkers using massively parallel sequencing. *Sci Transl Med*, **2**, 20ra14.
6. Dawson, S.J., Tsui, D.W., Murtaza, M., Biggs, H., Rueda, O.M., Chin, S.F., Dunning, M.J., Gale, D., Forshew, T., Mahler-Araujo, B. *et al.* (2013) Analysis of circulating tumor DNA to monitor metastatic breast cancer. *N Engl J Med*, **368**, 1199-1209.
7. Diehl, F., Schmidt, K., Choti, M.A., Romans, K., Goodman, S., Li, M., Thornton, K., Agrawal, N., Sokoll, L., Szabo, S.A. *et al.* (2008) Circulating mutant DNA to assess tumor dynamics. *Nat Med*, **14**, 985-990.
8. Talkowski, M.E., Ernst, C., Heilbut, A., Chiang, C., Hanscom, C., Lindgren, A., Kirby, A., Liu, S., Muddukrishna, B., Ohsumi, T.K. *et al.* (2011) Next-generation sequencing strategies enable routine detection of balanced chromosome rearrangements for clinical diagnostics and genetic research. *Am J Hum Genet*, **88**, 469-481.
9. Branford, S. (2007) Chronic myeloid leukemia: molecular monitoring in clinical practice. *Hematology Am Soc Hematol Educ Program*, 376-383.
10. Flohr, T., Schrauder, A., Cazzaniga, G., Panzer-Grumayer, R., van der Velden, V., Fischer, S., Stanulla, M., Basso, G., Niggli, F.K., Schafer, B.W. *et al.* (2008) Minimal residual disease-directed risk stratification using real-time quantitative PCR analysis of immunoglobulin and T-cell receptor gene rearrangements in the international multicenter trial AIEOP-BFM ALL 2000 for childhood acute lymphoblastic leukemia. *Leukemia*, **22**, 771-782.
11. Garraway, L.A. and Lander, E.S. (2013) Lessons from the cancer genome. *Cell*, **153**, 17-37.
12. Chen, K., Wallis, J.W., McLellan, M.D., Larson, D.E., Kalicki, J.M., Pohl, C.S., McGrath, S.D., Wendl, M.C., Zhang, Q., Locke, D.P. *et al.* (2009) BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods*, **6**, 677-681.
13. Drier, Y., Lawrence, M.S., Carter, S.L., Stewart, C., Gabriel, S.B., Lander, E.S., Meyerson, M., Beroukhim, R. and Getz, G. (2013) Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. *Genome Res*, **23**, 228-235.

14. Kozarewa, I., Ning, Z., Quail, M.A., Sanders, M.J., Berriman, M. and Turner, D.J. (2009) Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat Methods*, **6**, 291-295.
15. Nielsen, R., Paul, J.S., Albrechtsen, A. and Song, Y.S. (2011) Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet*, **12**, 443-451.
16. Handsaker, R.E., Korn, J.M., Nemesh, J. and McCarroll, S.A. (2011) Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat Genet*, **43**, 269-276.
17. Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841-842.
18. Smit, A., Hubley, R & Green, P. (1996-2010).
19. Bailey, J.A., Yavor, A.M., Massa, H.F., Trask, B.J. and Eichler, E.E. (2001) Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res*, **11**, 1005-1017.
20. Hazkani-Covo, E., Zeller, R.M. and Martin, W. (2010) Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. *PLoS Genet*, **6**, e1000834.
21. Galante, P.A., Parmigiani, R.B., Zhao, Q., Caballero, O.L., de Souza, J.E., Navarro, F.C., Gerber, A.L., Nicolas, M.F., Salim, A.C., Silva, A.P. *et al.* (2011) Distinct patterns of somatic alterations in a lymphoblastoid and a tumor genome derived from the same individual. *Nucleic Acids Res*, **39**, 6056-6068.
22. Ray, M., Goldstein, S., Zhou, S., Potamousis, K., Sarkar, D., Newton, M.A., Esterberg, E., Kendziorski, C., Bogler, O. and Schwartz, D.C. (2013) Discovery of structural alterations in solid tumor oligodendroglioma by single molecule analysis. *BMC Genomics*, **14**, 505.
23. Yang, L., Luquette, L.J., Gehlenborg, N., Xi, R., Haseley, P.S., Hsieh, C.H., Zhang, C., Ren, X., Protopopov, A., Chin, L. *et al.* (2013) Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell*, **153**, 919-929.
24. Malhotra, A., Lindberg, M., Faust, G.G., Leibowitz, M.L., Clark, R.A., Layer, R.M., Quinlan, A.R. and Hall, I.M. (2013) Breakpoint profiling of 64 cancer genomes reveals numerous complex rearrangements spawned by homology-independent mechanisms. *Genome Res*, **23**, 762-776.
25. Ju, H., Lee, K.A., Yang, M., Kim, H.J., Kang, C.P., Sohn, T.S., Rhee, J.C., Kang, C. and Kim, J.W. (2005) TP53BP2 locus is associated with gastric cancer susceptibility. *Int J Cancer*, **117**, 957-960.
26. Chevillard, S. (1993) A method for sequential extraction of RNA and DNA from the same sample, specially designed for a limited supply of biological material. *Biotechniques*, **15**, 22-24.
27. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M. and Haussler, D. (2002) The human genome browser at UCSC. *Genome Res*, **12**, 996-1006.
28. Medvedev, P., Stanciu, M. and Brudno, M. (2009) Computational methods for discovering structural variation with next-generation sequencing. *Nat Methods*, **6**, S13-20.
29. Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., Walenz, B.P., Axelrod, N., Huang, J., Kirkness, E.F., Denisov, G. *et al.* (2007) The diploid genome sequence of an individual human. *PLoS Biol*, **5**, e254.
30. Scherer, S.W., Cheung, J., MacDonald, J.R., Osborne, L.R., Nakabayashi, K., Herbrick, J.A., Carson, A.R., Parker-Katiraee, L., Skaug, J., Khaja, R. *et al.*

(2003) Human chromosome 7: DNA sequence and biology. *Science*, **300**, 767-772.
31. Kent, W.J. (2002) BLAT--the BLAST-like alignment tool. *Genome Res*, **12**, 656-664.
32. Rhead, B., Karolchik, D., Kuhn, R.M., Hinrichs, A.S., Zweig, A.S., Fujita, P.A., Diekhans, M., Smith, K.E., Rosenbloom, K.R., Raney, B.J. *et al.* (2010) The UCSC Genome Browser database: update 2010. *Nucleic Acids Res*, **38**, D613-619.
33. Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J. and Marra, M.A. (2009) Circos: an information aesthetic for comparative genomics. *Genome Res*, **19**, 1639-1645.

Fig1

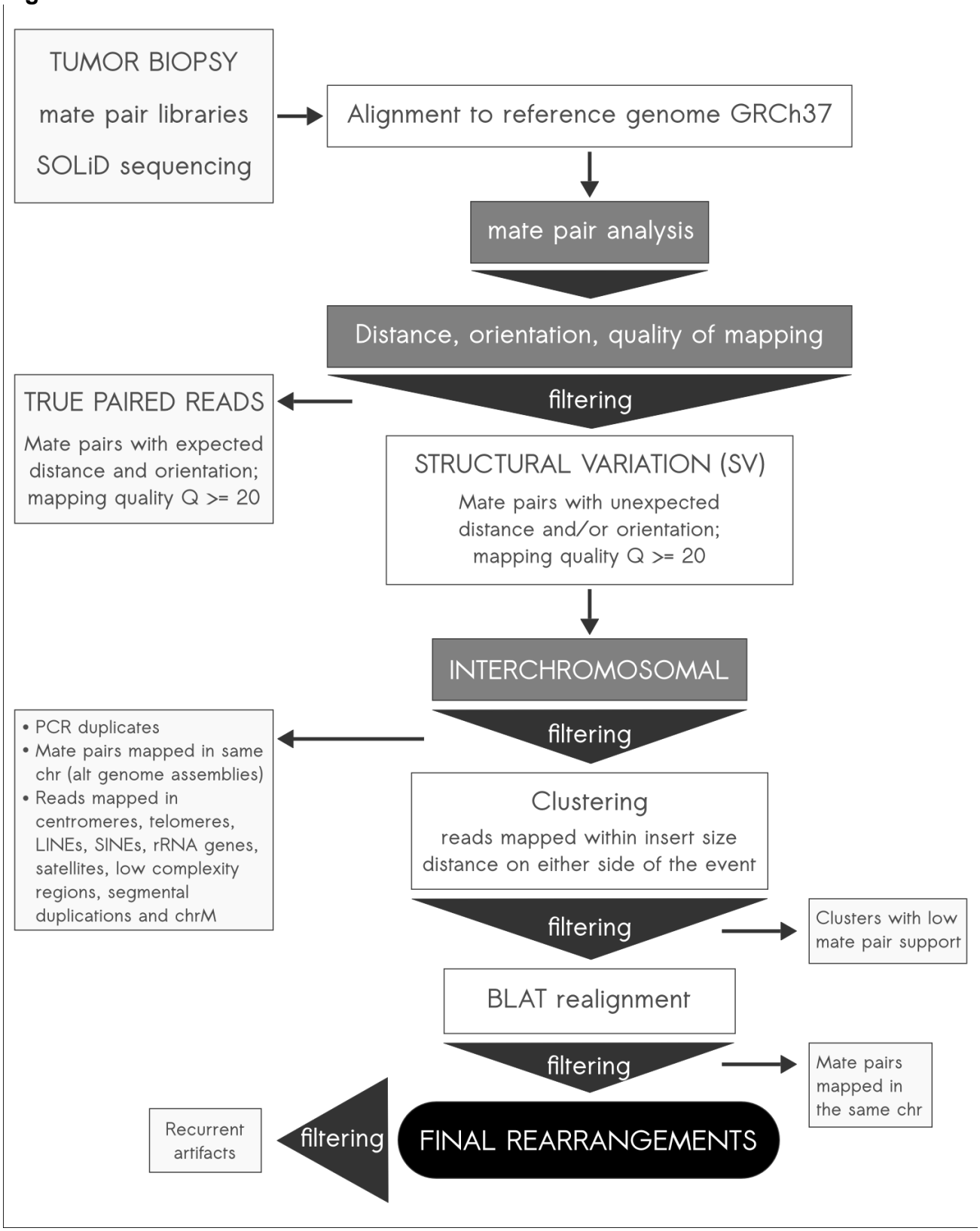
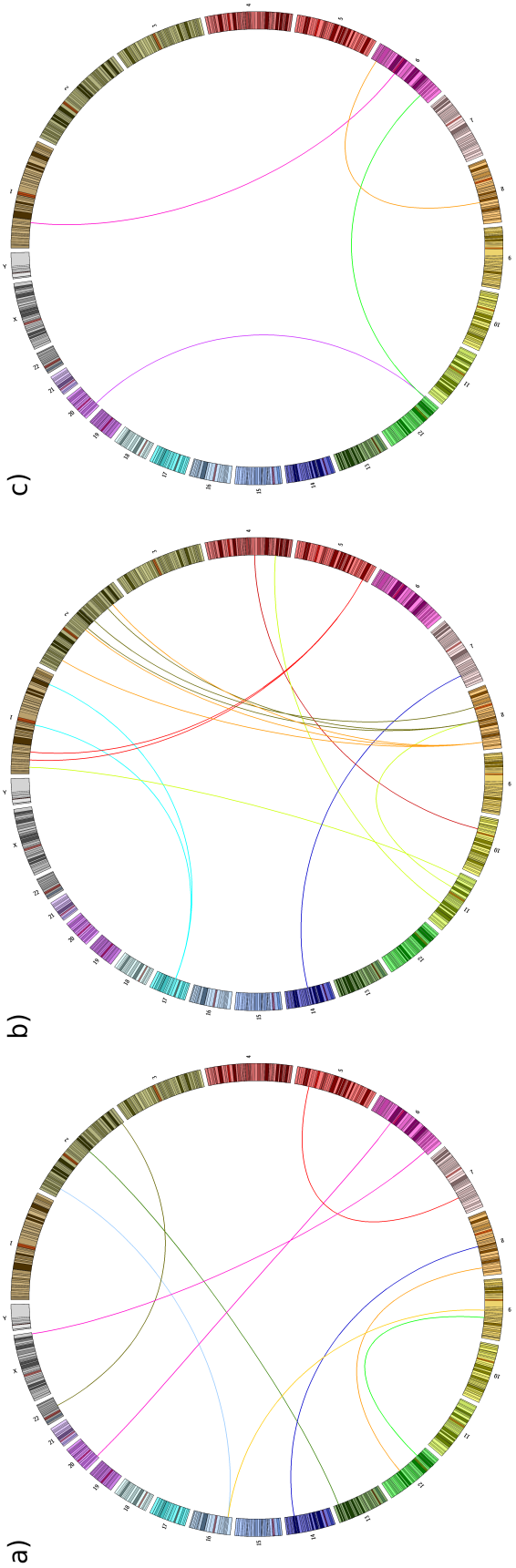

TUMOR BIOPSY
mate pair libraries
SOLiD sequencing
Alignment to reference genome GRCh37
mate pair analysis
Distance, orientation, quality of mapping
filtering
TRUE PAIRED READS
Mate pairs with expected distance and orientation; mapping quality Q >= 20
STRUCTURAL VARIATION (SV)
Mate pairs with unexpected distance and/or orientation; mapping quality Q >= 20
INTERCHROMOSOMAL
filtering
• PCR duplicates
• Mate pairs mapped in same chr (alt genome assemblies)
• Reads mapped in centromeres, telomeres, LINEs, SINEs, rRNA genes, satellites, low complexity regions, segmental duplications and chrM
Clustering
reads mapped within insert size distance on either side of the event
filtering
Clusters with low mate pair support
BLAT realignment
filtering
Mate pairs mapped in the same chr
Recurrent artifacts
filtering
FINAL REARRANGEMENTS

Fig2



CAGCCAGAAGGTCCCAGACACTGGGATGGTCTGAGTATACAGAAATGGGTGAGTGCTGGAGGCAGCATTGTGAAGTAAATATGTGGATCCTTAATCCAGGACTTCAACTGCACTCTTTCACAAATCAGTATATTCAGGTTTTTTCTAC

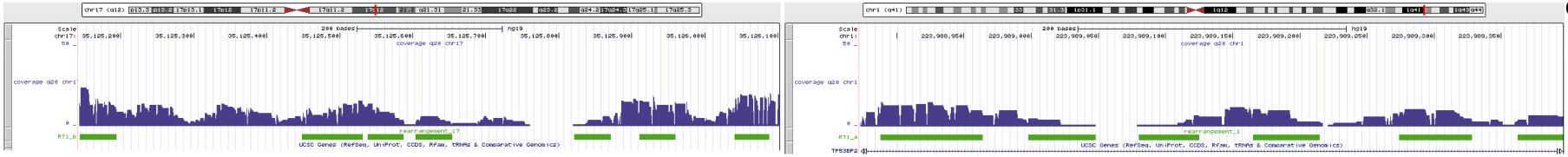


Fig3

Additional files provided with this submission:

Additional file 1: SupplementaryICRmax.doc, 8371K
http://www.biomedcentral.com/imedia/8198585861142758/supp1.doc

Supplementary Information

ICRmax: an optimized approach to detect tumor-specific InterChromosomal Rearrangements for Clinical Application

Elisa R. Donnard[1,2], Paola A. Carpinetti[1,2], Fábio C. P. Navarro[1,2], Rodrigo O. Perez[3,4], Angelita Habr-Gama[3], Raphael B. Parmigiani[1], Anamaria A. Camargo[1,4], Pedro A. F. Galante[1,#].

1- Centro de Oncologia Molecular - Hospital Sírio-Libanês – São Paulo – Brazil

2- Departamento de Bioquímica, Instituto de Química - Universidade de São Paulo – São Paulo – Brazil

3- Instituto Angelita & Joaquim Gama – São Paulo – Brazil

4- Ludwig Institute for Cancer Research – São Paulo - Brasil

 # - To whom the correspondence should be addressed: pgalante@mochsl.org.br

SUPPLEMENTARY TABLES

Supplementary Table 1 – Recurrent artifacts. BC=Buffy Coat amplification shown in figure S2.

Sample	Recurrent Artifacts																														
RT1	*	*	*	*	*	*	*	*	*	*	*	*	*	*		*		*		*	*	*	*	*	*						
RT2						*	*			*	*	*		*		*															
RT3	*	*		*			*		*		*				*	*	*	*													
RT4				*	*	*	*	*	*	*		*				*	*		*	*												
RT5	*		*	*			*	*	*			*			*	*		*	*		*					*	*	*	*	*	
RT6	*						*		*		*	*				*		*			*	*									*
N1	*	*	*	*		*	*	*	*		*	*				*	*	*					*	*	*						
N5	*		*	*		*	*	*	*			*	*		*	*	*	*	*		*					*	*	*	*	*	
N6							*					*				*															*
PCR	BC	BC	-	BC	BC	BC	-	-	-	BC	BC	-	-	BC	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Supplementary Table 2 - List of Recurrent Artifacts Identified

Chromosome	Start	End	Artifact_Id
chr11	108584189	108586766	89614_96597
chr13	21726906	21729281	89614_96597
chr11	108584788	108587264	89620_96597
chr13	21741127	21743464	89620_96597
chr13	105796631	105800272	132773_162053

chr15	90327152	90329234	132773_162053
chr12	73238274	73241432	134053_128214
chr15	94885320	94889540	134053_128214
chr1	182273321	182275468	184306_39495
chr17	74318711	74320960	184306_39495
chr1	37937257	37939377	243905_10777
chr2	68913270	68916895	243905_10777
chr1	119400044	119402193	261947_31449
chr2	141851156	141854725	261947_31449
chr17	74376274	74378425	265514_246449
chr2	153995541	153999677	265514_246449
chr14	65446302	65448678	277955_172280
chr2	194542895	194545899	277955_172280
chr10	35955348	35957410	308075_61447
chr20	60396725	60400323	308075_61447
chr18	36903625	36905797	332944_234051
chr3	3638027	3641216	332944_234051
chr1	218882476	218885378	337297_42691
chr3	25098294	25100912	337297_42691
chr17	41399252	41402014	341745_237366
chr3	46876431	46878494	341745_237366
chr11	8837834	8840151	393090_84435
chr4	109328905	109330958	393090_84435
chr4	80893096	80895435	406835_434722
chr5	21206859	21209840	406835_434722
chr13	61460610	61462902	437966_150506
chr5	21898980	21901255	437966_150506
chr22	32926900	32929511	438497_357381
chr6	24682294	24684997	438497_357381
chr3	187727662	187730166	448852_412995
chr6	93095822	93098009	448852_412995
chr17	64636488	64639044	480508_243791
chr6	13190048	13192272	480508_243791
chr19	34958601	34961602	491188_264861
chr6	142382498	142386242	491188_264861
chr12	129348944	129352162	509380_131398
chr7	64085163	64087486	509380_131398
chr7	129355731	129358254	518341_554542
chr8	96786681	96789830	518341_554542

chr5	37708752	37710828	519752_509914
chr7	8661351	8665285	519752_509914
chr1	168419966	168423607	543490_37263
chr8	30104410	30106823	543490_37263
chr8	30144402	30146626	520417_254001
chr17	7165565	7169441	520417_254001
chr11	38810960	38814409	563868_91918
chr8	52729384	52731790	563868_91918
chrX	6136090	6139710	579571_585981
chrY	16641626	16643936	579571_585981
chr17	41462768	41467792	349866_237387
chr3	73159060	73161128	349866_237387
chr18	69615587	69617907	269626_243213
chr2	144010082	144012186	269626_243213
chr11	62608096	62610525	159670_101207
chr17	41462768	41467230	159670_101207
chr13	90584185	90586218	159670_168661
chr17	41465151	41467239	159670_168661

Supplementary Table 3 - Simulated Rearranged Genomes

RG1 Rearrangements	RG2 Rearrangements	RG3 Rearrangements
1 chr1 5967154	1 chr19 16145942	1 chr13 99313058
1 chr9 132552309	1 chr20 10316088	1 chr12 62898082
2 chr4 99190688	2 chr19 17550368	2 chr22 27728600
2 chr16 10583406	2 chr3 88342067	2 chr17 17054965
3 chr2 157052395	3 chr13 24498602	3 chr17 21161135
3 chr16 20564272	3 chr4 62772317	3 chr5 124398911
4 chr1 92160215	4 chrX 94200100	4 chr3 143721961
4 chr4 96086097	4 chr10 36494648	4 chr10 112595223
5 chr13 7056402	5 chr10 85282775	5 chr6 48101937
5 chr11 78511841	5 chr19 47026802	5 chr7 82819179
6 chr4 75278019	6 chr1 193129800	6 chr3 16864669
6 chr16 3864870	6 chr20 38935527	6 chrX 55966944
7 chr1 221326909	7 chr5 131277311	7 chr22 23196028
7 chr20 40038782	7 chr18 37667584	7 chr14 8722577
8 chr8 36692250	8 chr14 10259236	8 chr17 12776619
8 chr20 4574603	8 chrX 122182020	8 chr11 12042766
9 chr2 202754912	9 chr14 9334101	9 chr15 31304144
9 chr15 95565379	9 chr3 195239030	9 chr3 183308460

10 chr12 10090207	10 chr17 15428849	10 chr17 19192213
10 chr1 155896653	10 chr2 19402593	10 chr7 4376601
11 chr14 75116791	11 chr14 35627881	11 chr3 193690685
11 chr7 26022571	11 chr7 95468121	11 chr9 119329369
12 chr16 50403890	12 chr6 36417132	12 chr5 148028181
12 chr2 138307643	12 chr12 131746448	12 chr8 84824378
13 chrX 34607425	13 chr16 31918313	13 chr19 36497196
13 chr2 10671534	13 chr3 126519428	13 chr3 141059015
14 chrX 4357428	14 chr20 31731275	14 chr20 42287075
14 chr14 99112591	14 chr4 79044069	14 chr12 99153280
15 chr13 113470086	15 chr5 68070400	15 chr6 132236187
15 chr15 1407346	15 chr15 11819673	15 chrX 67254446
16 chr6 100356908	16 chr21 21526994	16 chr20 43537346
16 chrX 132773677	16 chr14 55412783	16 chr6 127455466
17 chr14 63804958	17 chr4 7078779	17 chr17 14217930
17 chr22 21157826	17 chr2 134634744	17 chr15 80324145
18 chr6 106355647	18 chr20 10259276	18 chr21 42632094
18 chrX 47650389	18 chr4 3677150	18 chr17 20332313
19 chr10 29293916	19 chr5 102032871	19 chr17 17443518
19 chr5 158682739	19 chr18 47884662	19 chr3 159511236
20 chr8 111816948	20 chr20 56739411	20 chr7 13970235
20 chr15 72730220	20 chr16 28163826	20 chr6 74233668
	21 chrX 144458641	21 chr18 14269931
	21 chr10 31663026	21 chr2 115932833
	22 chrX 57483753	22 chr22 27467657
	22 chr2 211914206	22 chr18 25612658
	23 chr11 72791775	23 chr9 58055373
	23 chr2 52469082	23 chr11 76223695
	24 chr1 26715498	24 chr13 55487635
	24 chr14 33620656	24 chr15 97179693
	25 chr4 114650909	25 chr17 13998622
	25 chr7 81686153	25 chr1 234588536
	26 chr22 24626489	26 chr19 35450393
	26 chr11 13459873	26 chr11 100198960
	27 chr16 77790570	27 chr1 41470080
	27 chr14 83261902	27 chr21 6680057
	28 chr11 39536531	28 chr1 214112454
	28 chr1 57303536	28 chr2 104202591
	29 chr14 29184928	29 chr16 83969989
	29 chr16 1349126	29 chr11 81160042
	30 chr20 41624822	30 chr15 85239175
	30 chr14 31881671	30 chr5 155565902

		31 chr17 8167243 31 chr4 173428557 32 chr5 15277572 32 chr9 112001540 33 chr9 58220360 33 chr15 23346838 34 chr10 14437461 34 chr22 8705078 35 chr17 19289868 35 chr7 65304537 36 chr4 153358108 36 chr9 95296962 37 chr6 27160174 37 chr13 44547386 38 chr2 64436242 38 chr11 82911249 39 chr12 126775967 39 chr8 5905290 40 chr7 47388266 40 chr17 1871452

STEP-BY-STEP COMMAND LINE FOR ICRMAX FILTERS

At this step you should have a BED file containing the aligned mate pair reads mapped in different chromosomes with mapping quality greater than or equal to 20, after the reference genome mapping and mapping to alternative assemblies. The duplicate reads should also have been removed. For that, samtools rmdup is a good option but we prefer a local script to remove reads with the same start OR end positions, keeping the read with the highest mapping quality. You should have a recent version of Bedtools installed (v2.17.0 or higher see bedtools.readthedocs.org/en/latest/) and BLAT (source at users.soe.ucsc.edu/~kent/src/).

Example input:

$1 == read1_chromosome
$2 == read1_start_position
$3 == read1_end_position
$4 == read2_chromosome
$5 == read2_start_position
$6 == read2_end_position
$7 == matepair_id
$8 == qual
$9 == read1_strand
$10 == read2_strand

$1 $2 $3 $4 $5 $6 $7 $8 $9 $10

```
chr1    54832  54882  chr19  95964        96014        872_1929_510_21      1      +      +

chr1    88252  88299  chr8   125193040    125193089    1591_1759_1257_21    1      -      -

chr1    96549  96599  chr6   123975472    123975522    1207_623_783_21      1      -      -
```

1. Remove reads mapped in the mitochondrial chromosome and order the bed file:

```
$ grep -v 'chrM' input.bed | sortBed > step1_woM.bed
```

2. Remove reads mapped in centromere and telomere regions. To do that on both reads in the mate pair you must invert the file and repeat the command. Access the file with centromere end telomere positions at http://www.bioinfo.mochsl.org.br/ICRmax/downloads file centr_and_tel.bed:

```
$ bedtools subtract -A -a step1_woM.bed -b centr_and_tel.bed >
step2.wo_centr_tel.bed

$ awk '{print $4,$5,$6,$1,$2,$3,$7,$8,$10,$9}' step2.wo_centr_tel.bed | sed
"s/\s/\t/g" | sortBed > step2.wo_centr_tel.inv.bed

$ bedtools subtract -A -a step2.wo_centr_tel.inv.bed -b centr_and_tel.bed >
step2.wo_centr_tel.final.bed
```

3. Remove reads mapped in masked regions. Access the file with regions to mask at http://www.bioinfo.mochsl.org.br/ICRmax/downloads file all_to_mask.bed:

```
$ bedtools subtract -A -f 1.0 -a step2.wo_centr_tel.final.bed
 -b all_to_mask.bed > step3.masked.bed


$ awk '{print $4,$5,$6,$1,$2,$3,$7,$8,$10,$9}' step3.masked.bed
 | sed "s/\s/\t/g" | sortBed > step3.masked.inv.bed

$ bedtools subtract -A -f 1.0 -a step3.masked.inv.bed -b all_to_mask.bed | sortBed
> step3.masked.final.bed
```

4. Cluster the reads from different mate pairs mapped in the same chromosome. Use mean insert size +2s.d. as cluster distance (-d **size**, e.g. 1000). At this point observe that a cluster number will be generated and the file will have an extra column ($11):

```
$ bedtools cluster -i step3.masked.final.bed -d 1000 > step4.cluster.bed

$ awk '{print $4,$5,$6,$1,$2,$3,$7,$8,$10,$9,$11}' step4.cluster.bed | sed
"s/\s/\t/g" | sortBed >
step4.cluster.inv.bed

$ bedtools cluster -i step4.cluster.inv.bed -d 1000 > step4.cluster.final.bed
```

5. Join cluster numbers generated for both sides and select only clusters with 3 or more reads:

```
$ sed -i "s/\t/_/11"
step4.cluster.final.bed

$ awk '{print $11}' step4.cluster.final.bed | nsort | uniq -c | awk '{if ($1>=3)
print $2}' > clusters_over_3_reads


$ fgrep -w -f clusters_over_3_reads step4.cluster.final.bed > step5_cutoff3.bed
```

6. For SOLiD platform we suggest realigning the reads with BLAT using as input the sequences resulting from the initial alignment, BLAT parameters are the same used as default in the webtool:

```
$ blat -stepSize=5 -repMatch=2253 -minScore=24 -minIdentity=80 -noTrimA -fine -
out=pslx genome.2bit input.fa
```

7. Parse BLAT results and remove reads mapped in the same chromosome after alignment.

8. Invert the read order in the BED file once again and recluster the reads once more, this step should remove any clusters containing large gaps from reads that were removed by the BLAT filter. After that select only clusters still represented by at least 3 reads, at this point there are three numbers in the cluster id:

```
$ awk '{print $4,$5,$6,$1,$2,$3,$7,$8,$10,$9,$11}' step7_BLAT_filter.bed | sed
"s/\s/\t/g" | sortBed | bedtools cluster -d 1000 > step8_recluster.bed

$sed -i "s/\t/_/11" step8_recluster.bed

$awk '{print $11}' step8_recluster.bed | nsort | uniq -c | awk '{if ($1>=3) print
$2}' > clusters_over_3_reads

$fgrep -w -f clusters_over_3_reads3 step8_recluster.bed >
step8_recluster_cutoff3.bed
```

Example final output:

$1 == read1_chromosome
$2 == read1_start_position
$3 == read1_end_position
$4 == read2_chromosome
$5 == read2_start_position
$6 == read2_end_position
$7 == matepair_id
$8 == qual
$9 == read1_strand
$10 == read2_strand
$11 == cluster_id

$1	$2	$3	$4	$5	$6	$7	$8	$9	$10	$11
chr11	10847236	10847277	chr1	16503168	16503205	929_1168_1869_21	1	-	+	14774_4005_1
chr11	10847381	10847431	chr1	16502755	16502805	1503_783_582_21	1	-	+	14774_4005_1
chr11	10847383	10847425	chr1	16502701	16502751	990_1616_1004_21	1	+	+	14774_4005_1

```
chr11 10847439 10847488 chr1 16502752 16502799 1441_1504_612_21 1 + - 14774_4005_1
chr11 10847443 10847493 chr1 16502691 16502741 1368_1680_1079_21 1 - + 14774_4005_1
chr11 10847474 10847523 chr1 16502875 16502920 973_233_660_21 1 - + 14774_4005_1
chr11 10847479 10847526 chr1 16502664 16502713 1140_774_769_21 1 + - 14774_4005_1
chr11 10847518 10847566 chr1 16502744 16502794 1608_1480_1404_21 1 + + 14774_4005_1
chr11 10847555 10847605 chr1 16502700 16502750 1496_361_1184_21 1 - + 14774_4005_1
```

9. Remove recurrent artifacts. To compare final clusters between different tumor samples it is necessary to first merge all reads and get the rearrangements span on each chromosome. This can be done with a local script or with the Bedtools merge command. Example for the rearrangement shown above:

```
$ awk '{print $1"\t"$2"\t"$3"\t"$11"\n"$4"\t"$5"\t"$6"\t"$11}'
step8_recluster_cutoff3.bed | sortBed | bedtools merge -d 1000 -nms >
step9_merged.bed

chr1    16502664    16503205    14774_4005_1
chr11   1084723     10847605    14774_4005_1
```

Comparison between rearrangements from different samples can be easily done with the bedtools merge command as used above, make sure to allow for a distance similar to the clustering distance used (-d 1000) outside of the read span and alter the cluster names to include the patient identification (example: 14774_4005_1_RT2). This way, after the bedtools merge command using the parameter **-nms** you should have a single cluster and the different cluster names separated by a semicolon.

The list of recurrent artifacts found in our tumor samples can be accessed at http://www.bioinfo.mochsl.org.br/ICRmax/downloads file recurrent_artifacts.bed:

SUPPLEMENTARY FIGURES

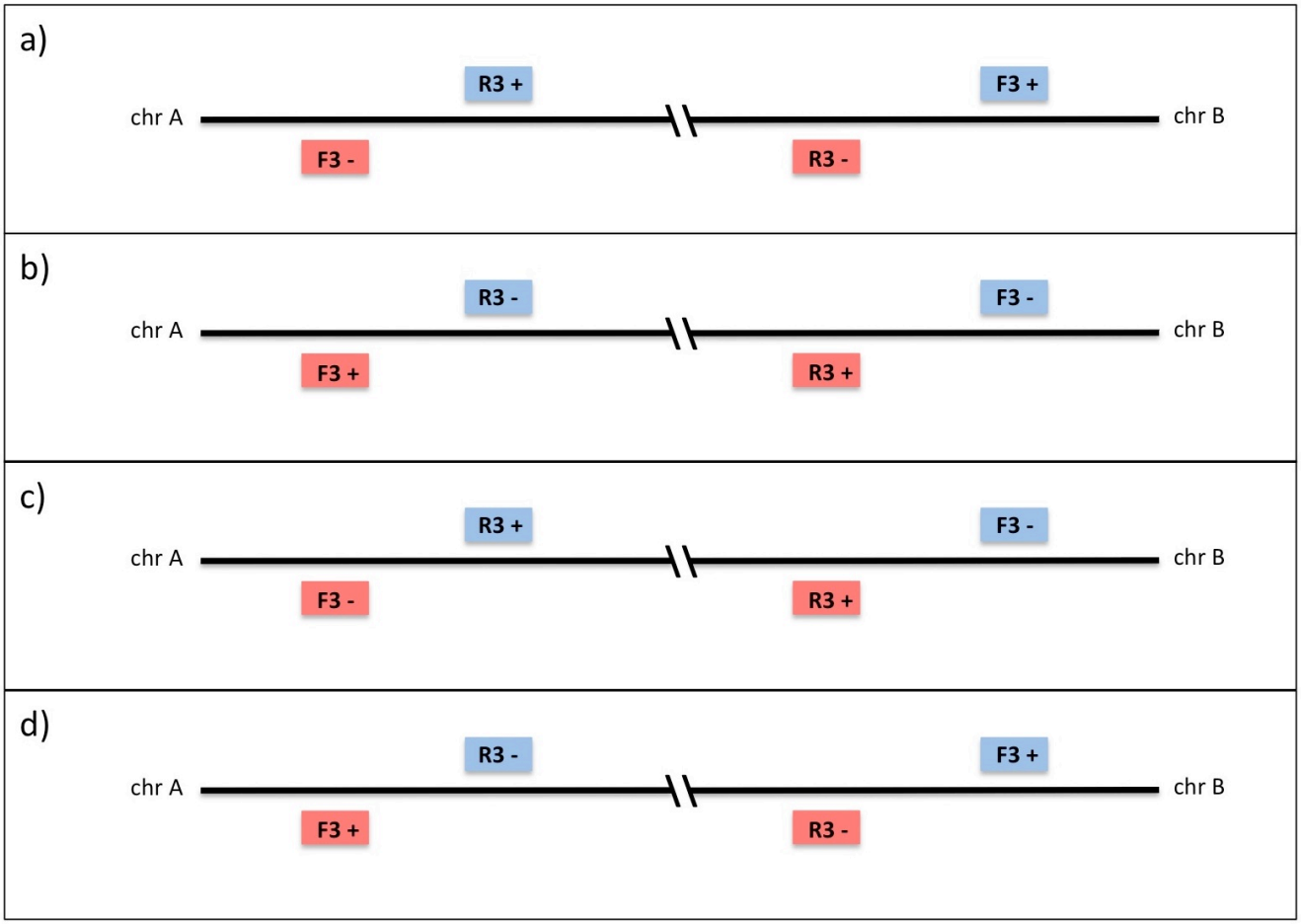


Figure S1 – Read orientation patterns. For primer design and orientation, it is important to evaluate the read pattern in the rearrangements found, the figure shows the 4 possible **mate-pair** read patterns to be considered.

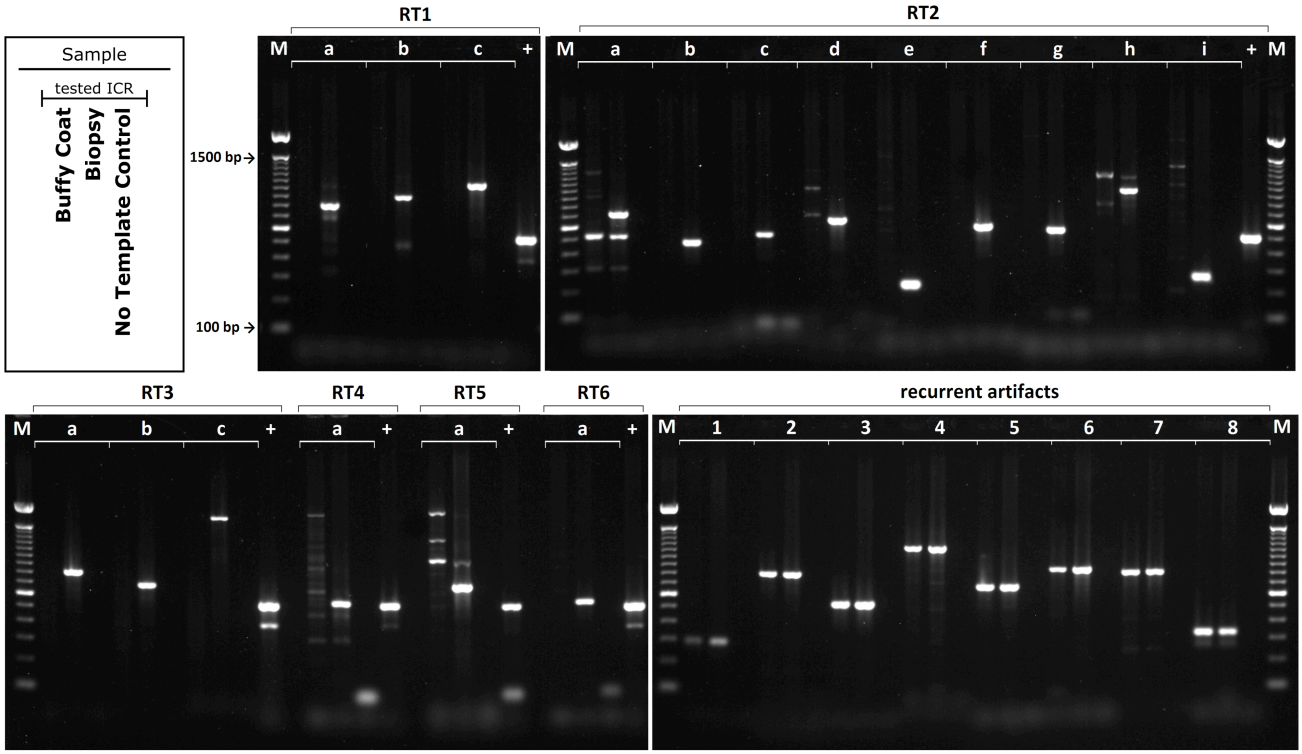


Figure S2 - PCR validations. PCR amplification results for all tumor-specific validated rearrangements (top and bottom left) for each patient (RT1-6 identified at the top and different ICRs are identified by letters below) and for the 8 recurrent artifacts tested (bottom right, identified by numbers 1-8). Validated interchromosomal rearrangement (ICR) candidates show biopsy-specific amplification. Recurrent artifacts tested result in amplification both in normal DNA (Buffy Coat = BC) and in biopsy DNA. M = marker; + = positive BC template control (MLH1 locus chr3:37001673-37002152).

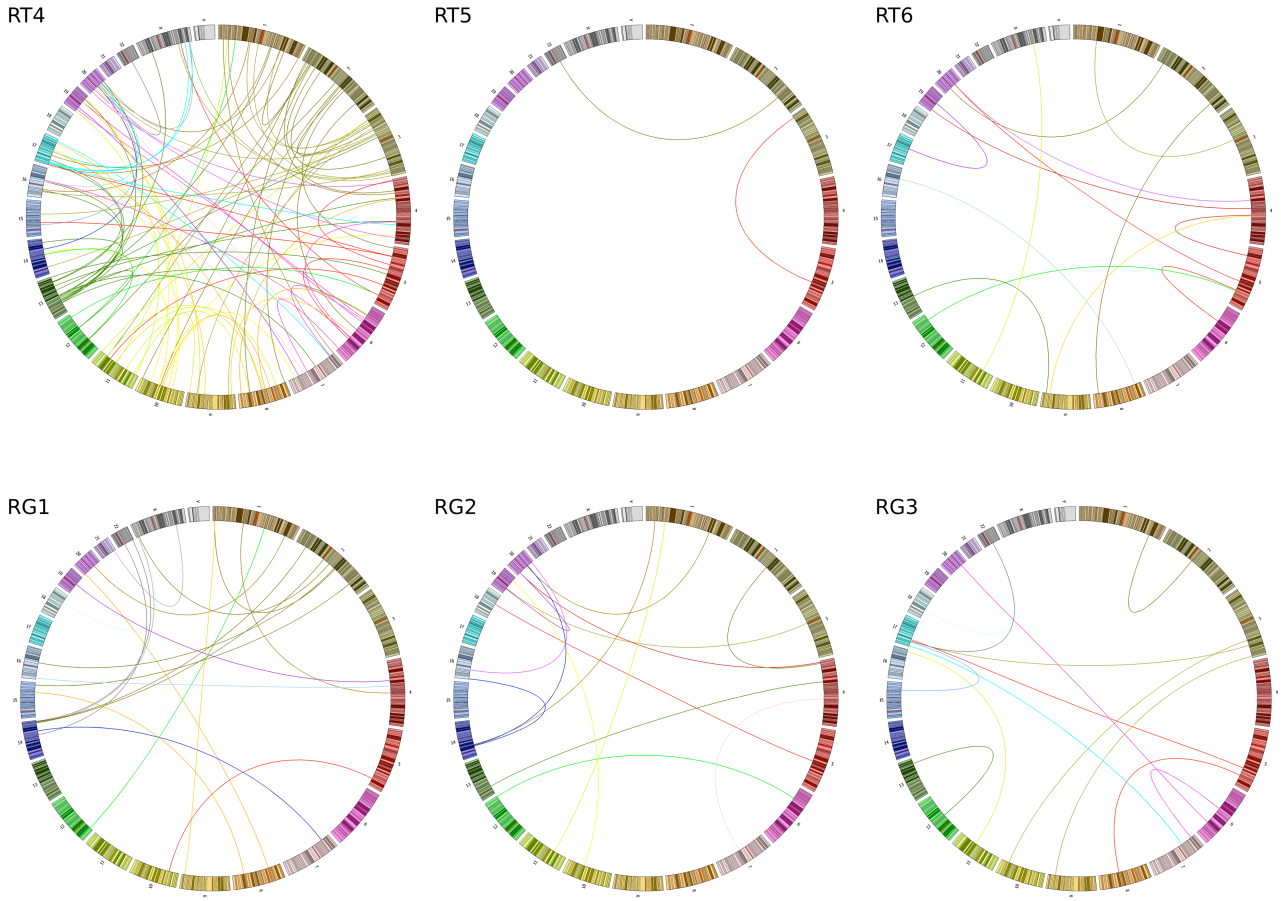


Figure S3 – Rearrangements found. Circos plot representation for the patient samples 4-6 and for simulated datasets.

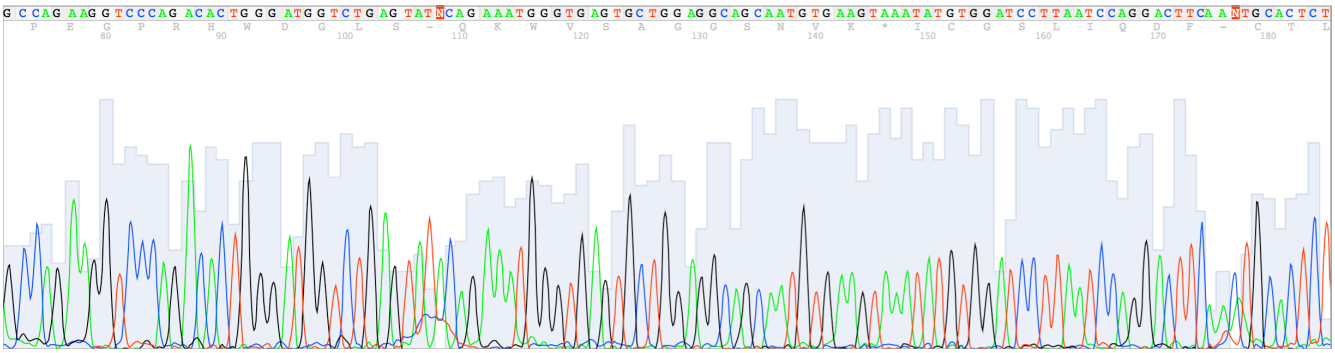


Figure S4 - Sanger sequencing result for a RT2 rearrangement between chr1 and chr17. Chromatogram image for the rearrangement shown in Figure 3 identified and validated for sample RT2.

ANEXO D: Artigo publicado: Mutation analysis of genes coding for cell surface proteins in colorectal cancer cell lines reveal new altered pathways, drugable mutations and mutated epitopes for target therapy

Mutational analysis of genes coding for cell surface proteins in colorectal cancer cell lines reveal novel altered pathways, druggable mutations and mutated epitopes for targeted therapy

Elisa Donnard[1,2,§], Paula F. Asprino[1,§], Bruna R. Correa[1], Fabiana Bettoni[1], Fernanda C. Koyama[1,3], Fabio C.P. Navarro[1,2], Rodrigo O. Perez[3,4], John Mariadason[5], Oliver M. Sieber[6,7], Robert L. Strausberg[8], Andrew J.G. Simpson[8], Denis L.F. Jardim[1], Luiz Fernando L. Reis[1], Raphael B. Parmigiani[1], Pedro A.F. Galante[1], Anamaria A. Camargo[1,3]

[1] Centro de Oncologia Molecular, Hospital Sírio-Libanês, São Paulo, Brazil.

[2] Programa de Pós Graduação do Departamento de Bioquímica, Instituto de Química, Universidade de São Paulo, São Paulo, Brazil.

[3] Laboratory of Molecular Biology and Genomics, Ludwig Institute for Cancer Research, São Paulo, Brazil.

[4] Instituto Angelita & Joaquim Gama, São Paulo, Brazil.

[5] Oncogenic Transcription Laboratory, Ludwig Institute for Cancer Research, Melbourne, Australia.

[6] Colorectal Cancer Genetics Laboratory, Systems Biology and Personalised Medicine Division, Walter and Eliza Hall Institute of Medical Research, Parkville, Australia.

[7] Faculty of Medicine, Dentistry and Health Sciences, Department of Medical Biology, University of Melbourne, Parkville, Australia

[8] Ludwig Institute for Cancer Research, New York, USA.

[§] These Authors contributed equally to this work

Correspondence to:
Dr. Anamaria A. Camargo, **email**: aacamargo@mochsl.org.br

Keywords: colorectal cancer, targeted therapy, cell surface proteins, somatic mutations.

Received: July 03, 2014 ***Accepted:*** August 20, 2014 ***Published:*** August 25, 2014

ABSTRACT

We carried out a mutational analysis of 3,594 genes coding for cell surface proteins (Surfaceome) in 23 colorectal cancer cell lines, searching for new altered pathways, druggable mutations and mutated epitopes for targeted therapy in colorectal cancer. A total of 3,944 somatic non-synonymous substitutions and 595 InDels, occurring in 2,061 (57%) Surfaceome genes were catalogued. We identified 48 genes not previously described as mutated in colorectal tumors in the TCGA database, including genes that are mutated and expressed in >10% of the cell lines (SEMA4C, FGFRL1, PKD1, FAM38A, WDR81, TMEM136, SLC36A1, SLC26A6, IGFLR1). Analysis of these genes uncovered important roles for FGF and SEMA4 signaling in colorectal cancer with possible therapeutic implications. We also found that cell lines express on average 11 druggable mutations, including frequent mutations (>20%) in the receptor tyrosine kinases AXL and EPHA2, which have not been previously considered as potential targets for colorectal cancer. Finally, we identified 82 cell surface mutated epitopes, however expression of only 30% of these epitopes was detected in our cell lines. Notwithstanding, 92% of these epitopes were expressed in cell lines with the mutator phenotype, opening new venues for the use of "general" immune checkpoint drugs in this subset of patients.

INTRODUCTION

Colorectal cancer is the most common gastrointestinal cancer in the world, with approximately one million new cases being diagnosed and more than 500,000 deaths occurring yearly. Approximately, one in five patients is diagnosed with metastatic disease, and an additional 30%–40% develop metastasis during the

course of their disease. Unfortunately, only a minority of the patients with metastatic disease is amenable to curative resection and remains free of disease recurrence [1]. Even though survival for patients with unresectable metastatic colorectal cancer has improved over the past decade, due to the introduction of agents targeting the Epidermal Growth Factor Receptor (EGFR) and the Vascular Endothelial Growth Factor (VEGF), these treatments are often not curative, and intrinsic and acquired drug resistance is frequently observed in the clinical practice [2]. Therefore, the identification of altered pathways and new therapeutic targets is critical to improve the management of a significant proportion of colorectal cancer patients.

Genetic analysis of colorectal tumors over the past 30 years allowed the characterization of distinct molecular pathways altered during the development and progression of this disease [3]. Initial whole-exome screenings using colorectal cancer cell lines detected an average of 80 point mutations in coding regions of the genome and a small number of frequently mutated cancer genes [4]. More recently, in a major effort to dissect the genetic basis of colorectal cancer, the TCGA released the results of a comprehensive and integrated genome-scale analysis of 276 tumors. No significant genetic differences were observed between rectal and colon tumors, and twenty–four genes were identified as frequently mutated in colorectal cancer, including several novel cancer genes such as SOX9, ARID1A, ATM, TCF7L2 and FAM123B. Most importantly, new potentially druggable targets were identified, including amplifications in the ERBB2 and IGF2 genes [5]. Despite this massive sequencing effort, a recent mutation saturation analysis of 4,742 tumors, across 21 cancer types, revealed that the cancer gene catalogue is far from complete, and that many more mutated genes with putative druggable mutations remain to be discovered [6].

Cell surface proteins are involved in a variety of cellular functions, including nutrient and ion transport, adhesion and signaling. These proteins also play important roles in pathological conditions such as diabetes, neurological disorders and cancer. They represent approximately 18% of all protein-coding genes in the human genome [7] and, due to their accessibility on the cell surface, they constitute optimal targets for directed therapies [8]. We have recently generated a catalog of genes coding for transmembrane proteins located at the surface of human cells (Surfaceome), and by integrating publically available gene expression data from a variety of sources, we searched for altered pathways, new therapeutic targets and tumor antigens in gliomas, colorectal and breast tumors [9, 10]. In the present work, we carried out a systematic mutational analysis of the Surfaceome in a panel of 23 representative colorectal cancer cell lines, searching for novel altered pathways, druggable mutations and mutated epitopes for targeted therapy in colorectal cancer. Collectively, our results point towards the potential use of FDA (U.S. Food and Drug Administration) approved RTK inhibitors and immune checkpoint target drugs in specific subsets of colorectal cancer patients.

RESULTS

Targeted sequencing the Surfaceome in colorectal cancer

We have recently used a combined bioinformatics approach to generate a catalog of genes coding for transmembrane proteins located on the surface of human cells [9]. Briefly, we searched the complete set of protein-coding genes for an annotated and/or predicted transmembrane domain and eliminated false positive candidates containing a signal peptide or known to be located on the membrane of other intracellular compartments. An updated list of genes coding for cell surface proteins was generated for this study (Supplementary information Table S1).

To define the mutational profile of the Surfaceome in colorectal cancer, we target sequenced the coding regions of the 3,594 cell surface protein genes in a panel of 23 tumor cell lines (Supplementary information Table S2) that altogether are representative of the main subtypes of primary colorectal tumors at the genomic level [11]. A total of 33,405 exons, covering ~6Mb of the human genome, were screened for the presence of somatic point mutations (nucleotide non-synonymous substitutions and InDels). For each cell line we analyzed approximately 1.2 Gb of on target sequences, with an average coverage of 30X (Table 1).

Somatic mutations in the colorectal cancer Surfaceome

Somatic point mutations were detected using an in house computational pipeline based on SAMtools mpileup calling (Figure 1). As matched normal tissue for these cell lines was not available, putative somatic mutations were identified by annotation against databases of known human germline variants (Table 2). A total of 3,944 putative somatic non-synonymous substitutions and 595 InDels were catalogued affecting 2,061 (57%) Surfaceome genes (Supplementary information Table S3). We identified an average of 174 putative non-synonymous substitutions and 28 InDels per cell line (Table 2). Mutation rates for genes coding for cell surface proteins varied significantly across cell lines and were similar to those previously reported for the entire set of protein-coding genes in colorectal tumors (Table 2) [5]. As expected, higher mutation rates (mutator phenotype) were observed in cell lines with microsatellite instability (MSI) and mutations in the

Table 1: Sequencing and coverage data of the Surfaceome in colorectal cancer cell lines.

Cell lines	Sequenced bases on target	Targeted region coverage	% of the target region covered	% of the target region covered >10X
CACO2	858,621,967	21.69 x	94	72
COLO205	964,932,858	23.32 x	94	76
COLO320	1,422,843,956	37.44 x	97	79
HCC2998	776,194,619	19.59 x	92	68
HCT116	865,230,723	19.63 x	87	71
HCT15	883,598,719	22.35 x	91	74
HT29	834,430,884	19.06 x	89	64
KM12	800,347,001	19.91 x	88	71
LIM1215	780,155,130	20.64 x	88	71
LIM2405	826,244,327	20.01 x	90	68
LOVO	1,484,687,708	34.87 x	97	81
RKO	826,244,327	20.34 x	92	68
RW2982	1,481,300,560	39.87 x	97	76
RW7213	1,609,474,378	43.33 x	97	78
SKCO1	1,564,643,937	41.67 x	97	81
SW1116	1,668,214,550	44.23 x	97	80
SW403	1,980,147,215	49.36 x	97	85
SW48	1,816,359,174	45.51 x	97	85
SW480	828,577,905	21.45 x	90	73
SW620	870,322,345	20.79 x	88	71
SW837	1,753,017,836	44.74 x	97	85
SW948	710,553,204	18.46 x	96	51
T84	1,761,549,149	43.46 x	97	86

DNA mismatch-repair genes or POLε (Supplementary information Table S2).

A total of 184 (5%) putative non-synonymous substitutions were nonsense, and 529 (89%) of the InDels introduced a frame-shift alteration in the mutated protein (Table 2 and Supplementary information Table S4). To further identify substitutions that may impact protein function, we used three different algorithms (PolyPhen, SIFT and Mutation Assessor) to estimate the impact of amino acid substitutions using information from DNA sequence, evolutionary conservation and structural data. A total of 1,434 (36%) putative non-synonymous substitutions and 474 (80%) InDels were classified as having an impact on protein function, and colorectal cancer cell lines harbor on average 85 putative point mutations (non-synonymous substitutions and indels) with a predicted impact on protein function (Table 3 and Supplementary information Table S5).

Novel mutated cell surface proteins and altered pathways in colorectal cancer

To further address the biological significance of the uncovered point mutations, we have incorporated gene expression data available for the cell lines (RNAseq and microarray) and restricted our downstream analysis to mutated and expressed genes. A list of genes coding for cell surface proteins that are mutated and expressed in >10% of the 23 cell lines analyzed is provided in Supplementary information Table S6. Analysis of

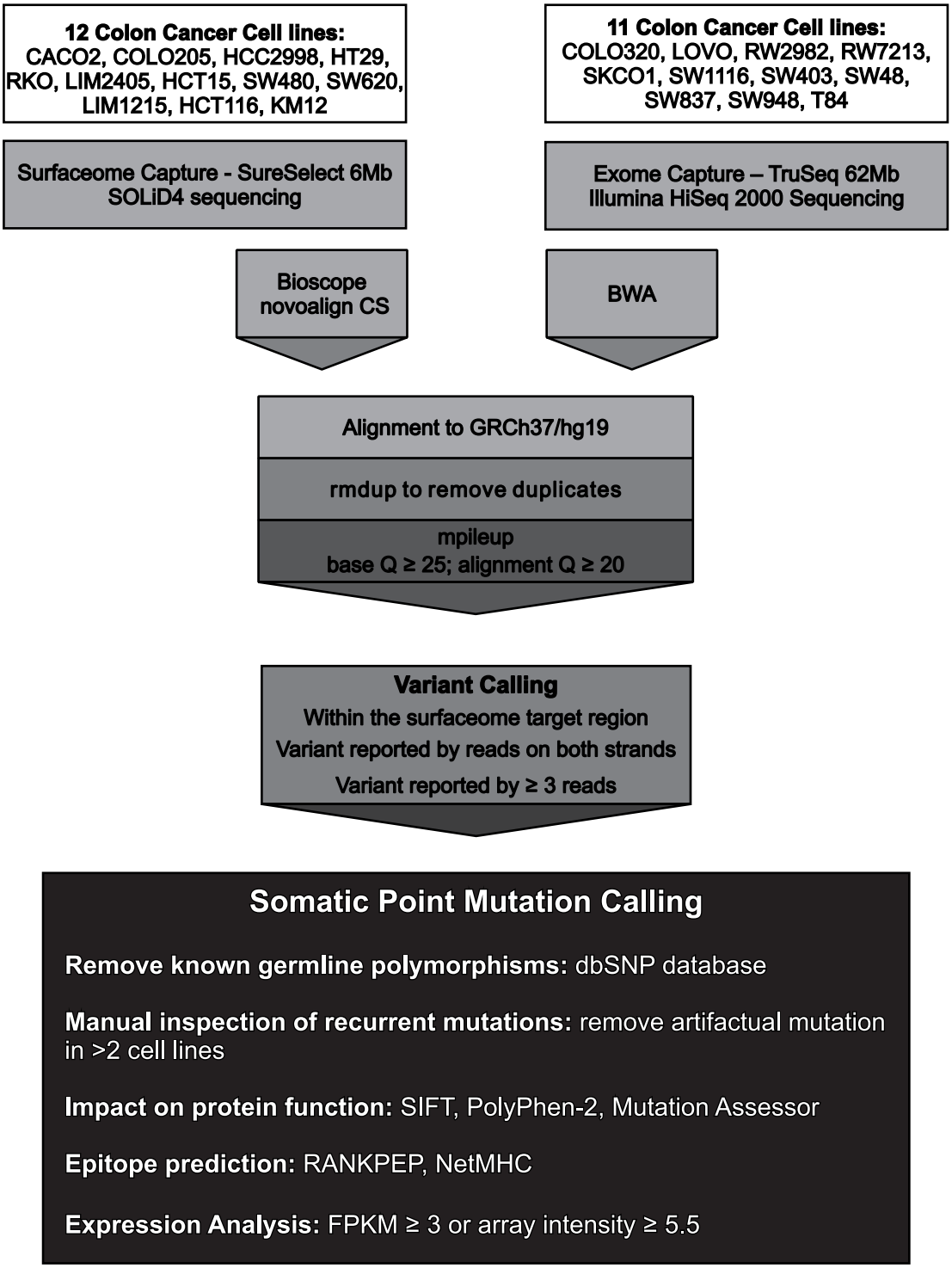


Figure 1: Sequencing strategy and computational pipeline used for the detection of somatic point mutations in the Surfaceome of colorectal cancer cell lines. The coding regions of 3,594 cell surface proteins were screened for the presence of somatic point mutations in 23 colorectal cancer cell lines. Genomic sequences were generated using either a SOLiD4 or a HiSeq 2000 sequencing platform. Sequences were aligned against the human genome reference sequence (GRCh37/hg19) using Bioscope and NovoAlign CS for SOLiD4 sequences and BWA for HiSeq 2000 sequences. Variant calling was performed using samtools mpileup and requiring at least 3 high quality reads (Q≥25; q≥20) on both strands supporting the variant. Known germline polymorphisms were removed and recurrent mutations were manually inspected to remove alignment artifacts. SIFT, PolyPhen-2 and Mutation Assessor were used to predict the functional impact of non-synonymous substitutions on protein function. RANKPEP and NetMHC were used for epitope prediction. Gene expression data was obtained from RNASeq (FPKM>3) or microarray (hybridization intensity≥5.5) experiments.

Table 2: SNV and InDel calling in the surfaceome of colorectal cancer cell lines.

Cell line	SNVs	% of SNVs in dbSNP	Somatic SNVs	Somatic non-synonymous SNVs	Somatic nonsense mutations	InDels	% of InDels in dbSNP	Somatic InDels	Somatic Frameshift InDels	Mutation rate	Mutator Phenotype
CACO2	2572	98	41	24	-	51	33	12	8	6.72E-06	No
COLO205	2753	98	43	33	-	64	30	15	11	7.05E-06	No
COLO320	3813	97	86	65	4	43	53	2	2	1.21E-04	No
HCC2998	3391	78	738	569	44	52	37	16	14	5.85E-05	Yes
HCT116	3193	88	357	237	10	106	18	61	54	1.76E-04	Yes
HCT15	3959	73	1071	775	34	77	26	31	27	1.08E-05	Yes
HT29	2373	96	66	49	3	49	33	9	8	6.92E-05	No
KM12	2978	86	422	284	12	117	15	74	65	2.39E-05	Yes
LIM1215	2841	95	146	103	2	81	21	41	39	3.39E-05	Yes
LIM2405	2731	92	207	148	6	91	18	53	53	7.90E-05	Yes
LOVO	4591	88	495	352	11	130	22	82	74	1.36E-05	Yes
RKO	3500	86	482	344	15	123	16	83	75	1.62E-05	Yes
RW2982	3545	97	80	57	2	32	50	8	5	1.41E-05	No
RW7213	3763	98	62	40	-	44	48	7	1	8.12E-05	No
SKCO1	4055	96	118	79	5	43	49	5	5	1.31E-05	No
SW1116	3719	97	89	64	-	47	51	6	4	1.02E-05	No
SW403	4196	94	165	123	9	48	58	4	4	1.94E-05	No
SW48	4706	88	530	365	15	177	23	96	92	1.46E-05	Yes
SW480	2440	96	83	60	3	43	28	8	8	2.71E-05	No
SW620	2535	95	99	69	3	49	29	13	9	8.69E-05	No
SW837	3912	97	62	44	3	42	50	4	3	1.02E-05	No
SW948	2820	98	56	35	2	34	53	3	1	9.18E-06	No
T84	4089	96	118	86	2	51	51	5	2	1.94E-05	No

expressed mutated surface genes revealed recurrent mutations in genes belonging to pathways known to be involved in colorectal cancer, including the WNT (LRP5 and FZD10), TGFβ (TGFBR3 and ACVR1B) and RTK-Ras (EGFR and ERBB3) signaling pathways [5]. Our analysis also identified 48 expressed genes that were not previously described as mutated in primary colorectal tumors in the TCGA database [5] (Supplementary information Table S7). This list includes mutations in 9 genes (SEMA4C, FGFRL1, PKD1, FAM38A, WDR81, TMEM136, SLC36A1, SLC26A6, IGFLR1) that occur in >10% of the cell lines and were confirmed by Sanger sequencing.

Semaphorin 4C (SEMA4C) mutations were detected and validated by Sanger sequencing in 4 cell lines (HCT15, KM12, RW2982, T84). Two of these mutations occur in the SEMA domain, a highly conserved sequence of approximately 500 amino acids critical for inducing targets of Semaphorin signaling. A third mutation occurs in the plexin-semaphorin-integrin (PSI) domain, another highly conserved domain, enriched in cysteine residues (Figure 2). Recurrent mutations in other genes belonging to the Semaphorin signaling pathway were also observed, including frequent mutations (>20%) in SEMA4G and SEMA4D, some of which also occurring in the SEMA and PSI domains (Figure 2). Semaphorins are an evolutionarily

Table 3: Analysis of somatic point mutations present in the surfaceome of colorectal cancer cell lines.

Cell line	Mutator phenotype	Non-synonymous mutations with predicted functional impact	InDels with predicted functional impact	Druggable mutations	Expressed druggable mutations	Mutated epitopes	Expressed mutated epitopes
CACO2	No	8	6	4	3	-	-
COLO205	No	10	11	9	2	1	0
COLO320	No	24	1	15	3	2	0
HCC2998	Yes	196	9	128	17	11	3
HCT116	Yes	81	45	66	19	7	4
HCT15	Yes	287	20	184	57	19	7
HT29	No	10	4	16	1	2	1
KM12	Yes	91	54	74	27	6	2
LIM1215	Yes	42	33	40	13	-	-
LIM2405	Yes	48	46	56	16	7	2
LOVO	Yes	160	57	69	19	8	2
RKO	Yes	131	72	95	30	7	2
RW2982	No	17	3	9	2	3	0
RW7213	No	13	4	10	2	-	-
SKCO1	No	24	3	10	3	1	0
SW1116	No	29	3	9	3	2	1
SW403	No	36	4	22	4	-	-
SW48	Yes	141	75	91	24	4	2
SW480	No	26	7	13	7	3	1
SW620	No	26	11	18	4	1	0
SW837	No	9	2	6	2	1	0
SW948	No	14	3	8	1	-	-
T84	No	27	4	15	3	1	0

conserved family of proteins that have been initially implicated in nervous system development and, more recently, in cancer progression and tumor angiogenesis [12, 13]. SEMA4C expression is significantly down-regulated during stem cell differentiation [14] and plays an important role in TGFβ-1 induced epithelial-mesenchymal transition [15]. To date, there is no published evidence of the direct involvement of SEMA4C in cancer, but somatic point mutations in SEMA4C were also reported by TCGA in 4% of the cutaneous melanomas. Conversely, an important role of the SEMA4D-Plexin-B1 interaction in regulating different aspects of tumor progression and angiogenesis is well established [16]. In all, alterations in SEMA4 family members were detected in 56% (13/23) of the cell lines, indicating an important role of SEMA4 signaling in colorectal cancer.

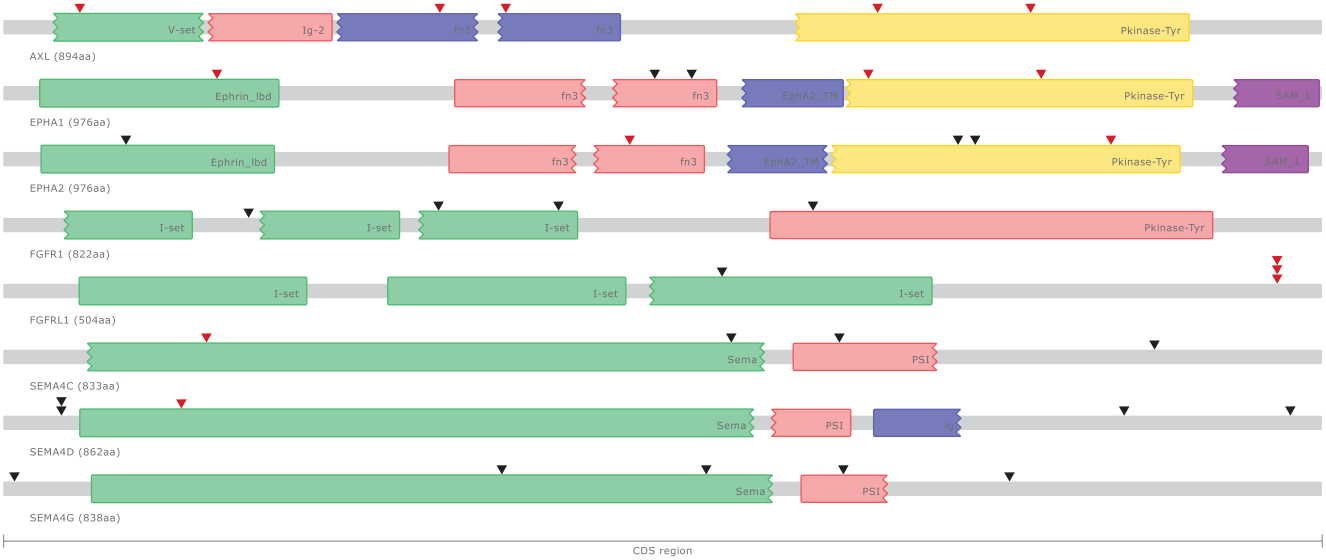


Figure 2: Schematic representation of somatic point mutations affecting the coding regions of putative druggable genes in colorectal cancer cell lines. Known protein domains are represented using different colors. Somatic point mutations occurring in different colorectal cancer cell lines are indicated (▾) and highlighted in red when predicted to have an impact on protein function.

Fibroblast Growth Factor Receptor Like Protein 1 (FGFRL1) alterations were detected and validated by Sanger sequencing in 4 cell lines (LOVO, KM12, LIM1215, RKO). Three cell lines carry frameshift mutations and the remaining cell line carries a non-synonymous point mutation with a predicted damaging effect, indicating a loss of function of the FGFRL1 protein in colorectal cancer (Figure 2). Recurrent (>10%) FGFRL1 somatic mutations were also reported in bladder tumors [17]. FGFRL1 acts as a negative regulator of Fibroblast Growth Factor Receptor 1 (FGFR1) signaling either by interfering with FGFR1 dimerization and phosphorylation or by sequestering FGFR1 ligands [18]. FGFR1 amplification and overexpression has been reported in colorectal cancer and associated with the presence of liver metastasis [19]. Indeed, in our study we also detected and validated by Sanger sequencing somatic mutations in FGFR1 in 4 cell lines (HCT116, HCT15, RKO, SW48), including a non-synonymous substitution in the tyrosine kinase domain (Figure 2). Mutations in FGFR2 (LIM2405) and FGFR3 (LOVO, SW48) were also observed at a lower frequency. In all, alterations in FGFR family members were detected in 35% (8/23) of the cell lines, suggesting an important role of FGF signaling in colorectal cancer.

Although the remaining 7 genes (PKD1, FAM38A, WDR81, TMEM136, SLC36A1, SLC26A6, IGFLR1) are mutated in >10% of the colorectal cancer cell lines, literature searches did not reveal evidence of the functional role or therapeutic potential of these genes in colorectal cancer. Nevertheless, recurrent mutations (>3%) in FAM38A, SLC36A1 and WDR81 have been reported for other primary tumors in the TCGA database, and further functional studies will be necessary to address their involvement in colorectal tumorigenesis.

Druggable mutations in cell surface proteins for targeted therapy in colorectal cancer

In order to identify putative druggable mutations in cell surface proteins, we searched for mutated genes present in the Drug-Gene Interaction Database (DGIdb), which integrates drug-target information from 13 different sources, including the literature and previously established databases [20]. We generated a catalogue of point mutations in druggable genes, and we found that colorectal cell lines harbor on average 11 mutations in druggable expressed genes (Table 3 and Supplementary information Table S8).

A significant fraction (34%) of these mutations occurred in membrane transporters. Membrane transporters, including solute carriers (SLCs) and ABC transporters, control the uptake and efflux of amino acids, sugars, lipids and vitamins, and their expression and activity are frequently altered in cancer as a consequence of the higher energy and nutritional requirements of the tumor cells [21]. Membrane transporters represent potential targets for cancer therapy and blocking their activity could be one way to interfere with tumor progression. In addition, membrane transporters can also serve as chemo-sensitizing targets, since they actively participate in drug delivery and resistance [21, 22].

Mutations in ABCA3, ABCA7, ABCC1, SLC23A2 and SLC9A1 were each observed in >20% of the cell lines.

We then focused on expressed genes with putative druggable mutations that were not previously considered as potential therapeutic targets for colorectal cancer, but for which specific inhibitors have previously been developed. We particularly focused on surface proteins with kinase activity, as they represent a significant fraction of the genes mutated in cancer and are highly amenable to targeting by rationally designed small molecule inhibitors. Two druggable RTKs (AXL and EPHA2) were found to be frequently (>20%) altered in our cell lines.

Five point mutations in the kinase domain and/or with predicted functional impact in the AXL Receptor Tyrosine Kinase (AXL) were detected and validated by Sanger Sequencing in 22% (5/23) of our cell lines (COLO205, KM12, HCT116, HCT15 and LOVO) (Figure 2). One of these mutations (g.chr19:41726597 C>T) occurring in the GAS6-ligand binding domain was also observed in a uterine corpus endometrioid carcinoma and in a glioblastoma. AXL is a member of the TAM family of RTKs, which also includes Mer and Tyro-3 [23]. Mutations in Mer (SW48) and Tyro-3 (HCT15) were also observed. Point mutations in AXL have not been specifically described in the literature for colorectal cancer, and lower mutation frequencies (3.5%) were reported for primary colorectal cancers in the TCGA database [5]. Overexpression of AXL in colorectal tumors was reported in metastatic lesions [24] and AXL was recently characterized as a poor prognostic marker in early stage colorectal tumors, and as an important mediator of basal and 5-FU induced EMT and invasiveness [25].

Point mutations in the EPH Receptor A2 gene (EPHA2) were detected and validated by Sanger sequencing in 3 cell lines (HCT15, LIM1215, LIM2405). Three of these mutations are located in the tyrosine kinase domain and one in the Ephrin-ligand binding domain (Figure 2). Mutations with predicted functional impact in the Ephrin-ligand binding domain and in the tyrosine kinase domain of the EPHA1 gene were also observed in 3 cell lines (HCT15, LIM1215 and LOVO) (Figure 2). Point mutations in EPHA2 and EPHA1 have not been specifically described in the literature for primary colorectal tumors, and lower mutation frequencies for these genes (4.4% EPHA1 and 2.6% EPHA2) were reported for primary colorectal tumors in the TCGA database [5]. EPHA2 is overexpressed in tumor cells and in tumor blood vessels in different types of cancer [26]. In colorectal tumors, EPHA2 overexpression was detected in approximately half of the samples and higher expression was associated with advanced stage tumors, metastatic disease and higher microvessels counts [27, 28]. Moreover, loss of EPHA2 reduced tumor formation in Apc Min/+ mice [29]. Conversely, elevated levels of EPHA1 were observed in early stage compared to late stage colorectal tumors. Reduced EPHA1 expression was associated with poorly differentiated and invasive tumors and poor overall survival, indicating that EPHA1 may play different roles during different stages of colorectal carcinoma progression [30, 31].

Mutated epitopes exposed on the cell surface of colorectal cancer

Non-synonymous and frameshift mutations in the Surfaceome of the 23 colorectal cancer cell lines were used to identify mutated epitopes with differential binding affinity to HLA when compared to epitopes generated by the corresponding non-mutated (reference) sequences. Our local pipeline for immunogenic epitope prediction was based on two algorithms RANKPEP and NetMHC as described in Materials and Methods. Mutated epitopes were required to have a binding affinity to the HLA*0201 molecule that was at least 20% higher than the reference epitope as predicted by both algorithms. A total of 82 putative mutated epitopes were identified (73 epitopes from non-synonymous mutations and 9 epitopes from frameshift mutations). However, when we combined gene expression data with epitope prediction analysis, we found that only 30% (25/82) of the predicted epitopes are expressed, and that 92% (23/25) of these epitopes are expressed in a subset of the cell lines with the mutator phenotype. These results suggest that the use of potentially immunogenic mutations in cell surface proteins for personalized T-cell based immunotherapy in colorectal cancer is limited, as only 30% of the mutated epitopes are expressed and less than half (11/23) of the tumors cell lines express mutated epitopes.

Discussion and Therapeutic Implications

One of the major objectives of cancer genome sequencing projects is to identify therapeutically targetable mutations. This objective has been achieved with repeated success in cancer therapy, resulting in the introduction of new treatment protocols in the clinical practice. The use of Imatinib, for chronic myeloid leukemia and other solid tumors, of Trastuzumab and Lapatinib, for ERBB2 positive breast cancer, and of Vemurafenib, for BRAF mutant melanomas, are emblematic examples of how genomic alterations can be used to target cancer cells [32]. Over the past years, these sequencing projects have revealed many new cancer genes, most of which are mutated at intermediate frequencies (2–20%) or lower, uncovering an unprecedented level of genetic heterogeneity in human cancers and establishing the need for a continued effort to determine the functional significance of these mutations and to translate these findings to the bedside [33].

Cell surface proteins constitute optimal targets for directed therapies and represent two-thirds of the protein-based drug targets [34, 35]. Surface proteins are also excellent targets for antibody-based therapies and vaccine

development since they are exposed on the cell surface and, therefore, have the highest chances to be recognized as antigens [36]. In the present work, we carried out a systematic mutational analysis of human genes coding for cell surface proteins, aiming to uncover novel altered pathways, druggable mutations and mutated epitopes for targeted therapy in colorectal cancer. We target sequenced the coding regions of cell surface protein genes in a panel of 23 tumor cell lines that altogether are representative of the main subtypes of primary colorectal tumors at the genomic level [11]. We opted to use cell lines in this study, instead of primary tumors, to overcome limitations imposed by the high level of colorectal intratumoral genetic heterogeneity in the mutation detection efficiency and to have straightforward cell models to further address the therapeutic potential of the uncovered altered pathways and druggable mutations.

We found that a significant (57%) fraction of the Surfaceome is reshaped by somatic point mutations in colorectal cancer cell lines. Our analysis identified 48 genes coding for cell surface proteins that were not previously described as mutated in primary colorectal tumors in the TCGA database [5], including mutations in SEMA4C and FGFRL1 which have not been previously considered as potential therapeutic targets for colorectal cancer. Although we cannot exclude the possibility that some of these alterations correspond to mutations acquired during *in vitro* propagation of the cell lines, our results are in agreement with a recent mutation saturation analysis of 4,742 sequenced tumors, across 21 cancer types [6]. This study revealed that the discovery of cancer genes mutated at frequencies of 5–10% in colorectal tumors is increasing linearly in relation to the number of tumor genomes sequenced, and that the current collection of sequenced colorectal tumors lacks the desired power to detect genes mutated at frequencies of 5% above the background rate [6].

SEMA4C mutations were found in 17% of the cell lines and recurrent mutations in SEMA4G (17%) and SEMA4D (22%) were also observed. The effects of Semaphorins and their receptors in cancer are broad, context dependent and complex [37]. SEMA4C is expressed in neural stem cells and its expression is downregulated during stem cell differentiation [14]. SEMA4C expression is induced by TGFβ-1 in renal epithelial cells and plays and important role in TGFβ-1 induced epithelial-mesenchymal transition [15]. In addition, an important role of SEMA4D-Plexin-B1 interaction in regulating different aspects leading to tumor progression, including invasive growth and angiogenesis, is well established [16]. The pro-angiogenic effect of SEMA4D was demonstrated both *in vitro* and *in vivo* and is comparable to that elicited by other well-known angiogenic molecules, such as VEGF-A, HGF and bFGF [38, 39]. Our results suggest that SEMA4 signaling is activated by point mutations in a significant fraction of colorectal tumors, and although specific inhibitors targeting SEMA4 proteins are not currently available, several biological process driven by SEMA4 signaling, such as angiogenesis and invasiveness, could be targeted with FDA approved drugs, including anti-angiogenic agents and MET inhibitors.

Inactivating mutations in FGFRL1, the most recently discovered member of the FGFR family, were detected in 17% of our cell lines. FGFRL1 binds with high affinity to heparin and FGF ligands, but it does not possess an intracellular protein kinase domain and, therefore, cannot signal by trans-auto-phosphorylation [18]. FGFRL1 thus acts as a negative regulator of FGFR1 signaling and loss of function mutations described here may represent a novel mechanism of FGF signaling activation in colorectal cancer. Alterations in FGFR1, FGFR2 and FGFR3 were also observed at a lower frequency, and 35% of the cell lines harbored somatic mutations in members of the FGF signaling pathway. Different FGFR specific inhibitors are currently under development [40], and further evaluation of their activity in the subset of colorectal cancer with FGFR/FGFRL1 alterations should be pursued. Moreover, Regorafenib, a multi-kinase inhibitor that targets FGFR1 among other RTKs, was recently approved by the FDA for the treatment of advanced colorectal cancer [41], but predictive biomarkers for this indication are not yet currently available.

Higher mutation frequencies in the RTKs AXL (22%) and EPHA2 (17%) were detected in our panel compared to those reported in the TCGA database for primary colorectal tumors (3.51% AXL and 2.63% EPHA2) [5]. Both RTKs have not been considered as potential therapeutic targets for colorectal cancer, however the availability of specific inhibitors and pre-clinical data support their potential use for therapeutic intervention. The oncogenic properties of AXL were initially described in patients with chronic myelogenous and lymphoblastic leukemia (CML), but overexpression of AXL have also been detected in many solid tumors and associated with poor prognosis [23]. AXL has a well established oncogenic role in survival, proliferation and migration of cancer cells *in vitro*, as well as in tumor angiogenesis and metastasis *in vivo* [23]. Moreover, recent studies have uncovered a major role of AXL in primary and acquired resistance to several anticancer therapies. AXL overexpression has been linked to Imatinib-resistance in gastrointestinal stromal tumors [42], Nilotinib-resistance in CML [43] and Lapatinib-resistance in HER-2 positive breast tumor cells [44]. In lung cancer, AXL was identified as a potential target for overcoming EGFR inhibitor resistance and combination of an AXL specific inhibitor (SGI-7079) with Erlotinib reversed Erlotinib resistance in a xenograft model of mesenchymal non-small cell lung cancer [45].

In colorectal cancer, AXL expression is associated to increased invasiveness of tumor cell lines with overexpression of the chemokine receptors CXCR4

and CXCR7, and AXL knock-down in these cell lines significantly hampered tumor cell invasion [46]. Considering that many multi-kinase inhibitors under development have AXL as one of their targets, further exploration of the pharmacologic inhibition of this pathway in pre-clinical models, including tumor cells lines with resistance to anti-EGFR drugs, should be pursued. In addition, monoclonal antibodies and small-molecule tyrosine kinase inhibitors specifically targeting AXL are currently in development and their use in colorectal cancer patients should also be further explored [47]. Noteworthy, some of the cell lines analyzed herein presented concomitant mutations in AXL and FGFR or FGFRL1 (HCT116, HCT15, LOVO, KM12), which suggests that these mutations are not mutually exclusive. In this setting, it will be important to explore the interdependence of both pathways, specially considering that some multi-kinase inhibitors under development are capable of blocking AXL and FGFR concomitantly [48]. Indeed, combination of these multi-kinase inhibitors with bevacizumab led to near total inhibition of tumor growth in colon carcinoma xenograft models and caused tumor growth arrest in bevacizumab-resistant tumors [48].

Somatic alterations in EPH receptors were also frequently observed in our cell lines, including frequent mutations in EPHA1 and EPHA2. Point mutations in EPHA2 and EPHA1 have not so far been described in the literature for colorectal cancer. Nevertheless, mutations in the kinase domain of EPHA3 was reported in 5% of colorectal cancer cell lines [49] and EPHA3 was listed among the top 3 cancer genes in a large-scale screening for somatic mutations in colorectal cancer [4]. EPH receptors play critical roles in embryonic development and their expression is frequently altered in a variety of cancers and tumor cell lines [50]. They comprise the largest family of RTKs and bind to ephrins (EFN) available on the surface of neighboring cells. Unlike others RTKs, EPH-EFN signaling is unique, since it triggers a bi-directional signal that affects both receptor and EFN expressing cells [50]. EPH receptors are thus important mediators of tumor cell interactions with the tumor stroma and tumor vasculature, and have been proposed as promising targets for cancer therapy, since targeting these receptors could simultaneously inhibit several aspects of tumor progression [26, 50]. EPHA2 overexpression in colorectal cancer is associated with advanced stage tumors, metastatic disease and higher microvessel counts [27, 28]. Moreover, loss of EPHA2 was shown to reduce Apc Min/+ tumorigenesis [29]. Confirmation of the activation of EPH signaling mediated by EPHA2 point mutations in colorectal cancer is of upmost importance considering the availability of FDA approved drugs targeting this receptor, such as Dasatinib [51]. In addition, EPHA2-FC soluble receptors were shown to significantly reduce tumor volume and overall metastatic burden in pre-clinical models of breast [52] and pancreatic tumors [53], but have not been evaluated in colorectal cancer models. Finally, receptor endocytosis promoted by anti-EPHA2 monoclonal antibodies has also been used to reduce EPHA2 activity and inhibit malignant cell behavior *in vitro* [54]. On the other hand, therapies targeting EPHA1 in colorectal cancer should be carefully evaluated since this gene seems to play different roles during disease progression [30, 31].

Non-synonymous and frameshift mutations in tumor cells can generate unique T-cell mutated epitopes and induce tumor antigen-specific immune response [55]. There is evidence supporting the efficacy of vaccination strategies using mutated epitopes [56] and the use of personalized peptide vaccines and adoptive T-cell transfer protocols based on patient-specific mutated epitopes holds great promise in cancer therapy [57]. Unfortunately, combining epitope prediction algorithms and gene expression data, we found that the use of potentially immunogenic mutations in cell surface proteins for personalized immunotherapy in colorectal cancer is limited, since the expression of approximately 70% of these epitopes was not detected in the tumor cells. However, additional studies including mutated epitopes present in intracellular proteins will be required to further address the applicability of personalized vaccines in colorectal patients.

Notwithstanding, we observed that mutated expressed epitopes are predominantly found in colorectal cell lines presenting a mutator phenotype and that this specific subset of cell lines express a total of 23 mutated epitopes. In this context, it was recently demonstrated that patients with tumors showing naturally occurring immunogenic mutations presented higher cytotoxic T-cell infiltration and improved overall survival and, based on these observations, the use general immune modulators that block immune regulatory checkpoints such as anti-CTLA4 and anti-PD1 was proposed as a treatment strategy for patients with immunogenic mutations [58]. Accordingly, tumors with a high level of mutations as revealed by the TCGA [59], such as melanoma and non-small cell lung cancer, are currently deriving striking benefits with immune checkpoint blockage drugs [60, 61]. Although our results do not support the use of personalized T-cell based immunotherapy in colorectal cancer, they suggest that colorectal cancer patients harboring tumors with a mutator phenotype could be more responsive to immune checkpoint blockage. Indeed, increased counts of CD8+ T-cells were observed in colorectal cancer tumors with high mutational loads [58] and microsatellite instability [62]. Data on the use of immune checkpoint target drugs in colorectal cancer are still limited, but the results of the first long term follow-up study from the first clinical trial based on the PD1-targeting monoclonal antibody have recently been reported. This study included a 71-year-old patient with colorectal cancer who attained

a complete and durable (>4 years) response to anti-PD1 treatment [63].

To the best of our knowledge, this is the first systematic and focused screen of point mutations in genes coding for cell surface proteins in colorectal cancer. By combining high-throughput sequencing, bioinformatics tools, data integration and literature searches, we have successfully discovered novel altered pathways and druggable mutations for targeted therapy in colorectal cancer. We have also uncovered the potential use of existing RTK inhibitors and immune checkpoint target drugs in specific subsets of colorectal cancer patients. Results presented here are encouraging, however our study also presents some limitations.

First, although we have described novel druggable mutations occurring in a representative panel of colorectal cancer cell lines, it will be important to confirm the prevalence of these alterations in clinical samples matched with normal tissue. At present, we cannot completely exclude the possibility that some of the alterations reported in this study correspond to mutations acquired during *in vitro* propagation of the cell lines or to very rare germline polymorphisms not represented in public databases, nor in individuals sequenced by the 1000 Genomes Project. However, we believe that these possibilities do not significantly affect our results, since we have previously shown that the rate of mutation accumulation during *in vitro* propagation is not significant [11] and stringent bioinformatics cut-offs were implemented to filter most, if not all, non-clonal mutations eventually introduced during *in vitro* growth. Second, the functional consequences of the uncovered genetic alterations were predicted primarily using computational tools, and confirmation with functional *in vitro* assays is further required. Similarly, additional experiments to evaluate the effects of pharmacologic inhibition of the altered pathways using pre-clinical models are compulsory to translate our findings to the bedside. Finally, although we suggest potential molecular therapeutic targets in colon cancer, it is important to recognize that a recent study matching targeted therapy with specific molecular abnormalities for patients with advanced colorectal cancer failed to confer significant clinical benefit [64]. We believe that a diversification of potential targets, including those proposed by our study, could bring new opportunities to change this paradigm.

MATERIALS AND METHODS

Colorectal cancer cell lines

The panel of 23 colorectal cancer cell lines used in this study was obtained from different sources (Supplementary information Table S2). CACO2, COLO205, COLO320-DM, HCT116, HCT15, HT29, LOVO, RKO, SKCO-1, SW1116, SW403, SW48, SW480, SW620, SW837, SW948 and T84 were obtained from the American Type Culture Collection (Manassas, VA). LIM1215 and LIM2405 were generated by the Ludwig Institute for Cancer Research. HCC2998 and KM12 were obtained from the National Cancer Institute-Frederick Cancer DCT Tumor Repository. RW2982 and RW7213 were provided by Dr. P Calabresi from Roger Williams General Hospital. Cells were cultured with Dulbecco's Modified Eagle Medium and 10% FBS at 37ºC and 5% CO_2. Cell lines were authenticated and tested for *Mycoplasma* contamination as previously described [11].

Target sequencing the human surfaceome

We target-sequenced the coding regions of 3,594 cell surface proteins in 12 cell lines (CACO2, COLO205, HCT116, HCT15, HT29, RKO, SW480, SW620, LIM1215, LIM2405, HCC2998, KM12). Surfaceome-capture and sequencing were performed using Sure Select Target Enrichment baits (Agilent Technologies) and the SOLiD 4.0 sequencing platform (Life Technologies), respectively. For the remaining cell lines (COLO320, LOVO, SKCO1, SW1116, SW403, SW48, SW837, SW948, T84, RW2982 and RW7213) whole-exome capture was performed using the TruSeq Exome Enrichment Kit (Illumina) and paired-end sequencing was performed using Illumina HiSeq 2000. A local pipeline was then developed to extract the genomic sequences corresponding to the Surfaceome targeted region from whole-exome data.

Public Data and Databases

Exome-capture sequencing data on colorectal tumors were retrieved from TCGA and used to identify novel mutated genes and to determine mutation frequencies in colorectal cancer primary tumors. The DGIdb [20] was used to identify druggable mutated genes and the gene list provided by the Human Kinome project [65] (kinase.com/human/kinome) was used to identify genes coding for cell surface proteins with kinase activity.

Somatic mutation detection, validation and functional analysis

For single nucleotide variations (SNVs) detection, SOLiD 4.0 and Illumina reads were aligned to the human reference genome sequence (GRCh37/hg19) using BioScope (Life Technologies) and BWA [66], respectively. For InDels detection, alignments were performed using NovoAlignCS (www.novocraft.com). A local pipeline for point mutations was developed using Samtools mpileup and bcftools [67]. Duplicated reads were removed with rmdup (Samtools) to avoid potential PCR duplicates generated during library construction. Variants were filtered against known germline variations

annotated in dbSNP (version #135) and variations present in more than 3 cell lines were manually inspected to distinguish recurrent mutations (eg. EGFR mutations) from false positive mutations due to alignment artifacts. Somatic mutations were validated using PCR amplification and Sanger sequencing using standard protocols (Supplementary information Table S9, S10). SIFT [68], PolyPhen-2 [69] and Mutation Assessor [70] were used to evaluate the impact of non-synonymous substitutions and InDels on protein function. Mutations were annotated as having an impact on protein function when predicted by at least two of these algorithms in the case of non-synonymous substitutions and by SIFT in the case of InDels.

Gene expression data

RNA-Seq data was generated for 12 cell lines (CACO2, COLO205, COLO320, HCT116, HCT15, HT29, KM12, LIM1215, LIM2405, RKO, SW480, SW948) using the 5500XL sequencing platform to a depth of >100 million reads. Sequences were aligned to the human reference genome sequence (GRCh37/hg19) using TopHat [71] with standard parameters for color space reads. Isoform assembly and transcript relative abundance was determined using Cufflinks [72]. Genes were considered expressed when FPKM [72, 73] was ≥ 3 in at least one of the cell lines [73, 74]. For the remaining cell lines (LOVO, SKCO1, SW1116, SW403, SW48, SW620, SW837, e T84) microarray expression data was extracted from GEO [75]; Accession GSE36133) and genes were considered expressed when the array values were ≥ 5.5.

Epitope prediction

Peptide sequences corresponding to non-synonymous mutations and InDels, flanked by 10 aminoacids on either side, were used for epitope prediction by applying a similar approach to that described by Segal et al. 2008 [76]. The same process was performed for peptide sequences corresponding to the non-altered (reference) sequences. Concatamers of these peptides were analyzed by RANKPEP [77] and NetMHC [78] to identify 9 aa peptide sequences with binding affinity to the class I MHC molecule HLA-A*0201. RANKPEP predicts binding based on scoring matrices from known peptides that bind to MHC molecules. Peptides were considered immunogenic if the percentage optimum was ≥ 50%. RANKPEP also evaluates if the peptide tested results from a known cleavage process and therefore only predicted cleaved peptides were analyzed. NetMHC uses artificial neural networks to predict binding to the MHC molecule. The peptides were considered immunogenic if the IC_{50} was ≤ 500nM. To check for predicted cleavage, sequences were then analyzed using the NetChop algorithm [79], and only peptides with predicted cleavage were selected. Results from both algorithms were processed using a local pipeline and epitopes resulting from sequence concatenation artifacts were excluded. Mutated epitopes were defined as those predicted by both algorithms and that were unique to the variant sequence or showing an increase in MHC biding affinity by >20% when compared to the reference peptide.

ACKNOWLEDGEMENTS

The authors declare no conflict of interest. This study was financed by the Ludwig Institute for Cancer Research. O.M.S. is a NHMRC R.D. Wright Biomedical Career Development Fellow. E.D and B.R.C were supported by fellowships from Fundação de Amparo a Pesquisa do Estado de São Paulo - FAPESP. F.C.N was supported by a fellowship from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES.

REFERENCES

1. Siegel R, Desantis C and Jemal A. Colorectal cancer statistics, 2014. CA. Cancer J. Clin. 2014; 64:104–17.
2. Arnold D and Seufferlein T. Targeted treatments in colorectal cancer: state of the art and future perspectives. Gut. 2010; 59:838–58.
3. Fearon ER. Molecular genetics of colorectal cancer. Annu. Rev. Pathol. 2011; 6:479–507.
4. Sjöblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, Mandelker D, Leary RJ, Ptak J, Silliman N, Szabo S, Buckhaults P and Farrell C. et al. The consensus coding sequences of human breast and colorectal cancers. Science. 2006; 314:268–74.
5. TCGA, Cancer T, Atlas G: Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012; 487:330–7.
6. Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, Meyerson M, Gabriel SB, Lander ES and Getz G. Discovery and saturation analysis of cancer genes across 21 tumour types. Nature. 2014; 505:495–501.
7. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, Boychenko V and et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 2012; 22:1760–74.
8. Rask-Andersen M, Almén MS and Schiöth HB. Trends in the exploitation of novel drug targets. Nat. Rev. Drug Discov. 2011; 10:579–90.
9. Da Cunha JP, Galante PA, de Souza JE, de Souza RF, Carvalho PM, Ohara DT, Moura RP, Oba-Shinja SM, Marie SK, Silva WA Jr, Perez RO, Stransky B, Pieprzyk M and et al. Bioinformatics construction of the human

cell surfaceome. Proc Natl Acad Sci U S A. 2009; 106: 16752–16757.

10. Da Cunha JPC, Galante PAF, de Souza JES, Pieprzyk M, Carraro DM, Old LJ, Camargo AA and de Souza SJ. The human cell surfaceome of breast tumors. Biomed Res. Int. 2013; 2013:976816.
11. Mouradov D, Sloggett C, Jorissen RN, Love CG, Li S, Burgess AW, Arango D, Strausberg RL, Buchanan D, Wormald S, O'Connor L, Wilding JL, Bicknell D and et al. Colorectal cancer cell lines are representative models of the main molecular subtypes of primary cancer. Cancer Res. 2014; 74:3238–47.
12. Flannery E and Duman-Scheel M. Semaphorins at the interface of development and cancer. Curr. Drug Targets. 2009; 10:611–9.
13. Gu C and Giraudo E. The role of semaphorins and their receptors in vascular development and cancer. Exp. Cell Res. 2013; 319:1306–16.
14. Wu H, Fan J, Zhu L, Liu S, Wu Y, Zhao T, Wu Y, Ding X, Fan W and Fan M. Sema4C expression in neural stem/progenitor cells and in adult neurogenesis induced by cerebral ischemia. J. Mol. Neurosci. 2009; 39:27–39.
15. Zeng R, Han M, Luo Y, Li C, Pei G, Liao W, Bai S, Ge S, Liu X and Xu G. Role of Sema4C in TGF-β1-induced mitogen-activated protein kinase activation and epithelial-mesenchymal transition in renal tubular epithelial cells. Nephrol. Dial. Transplant. 2011; 26:1149–56.
16. Ch'ng ES and Kumanogoh A. Roles of Sema4D and Plexin-B1 in tumor progression. Mol. Cancer. 2010; 9:251.
17. Guo G, Sun X, Chen C, Wu S, Huang P, Li Z, Dean M, Huang Y, Jia W, Zhou Q, Tang A, Yang Z and Li X. Whole-genome and whole-exome sequencing of bladder cancer identifies frequent alterations in genes involved in sister chromatid cohesion and segregation. Nat. Genet. 2013; 45:1459–63.
18. Trueb B. Biology of FGFRL1, the fifth fibroblast growth factor receptor. Cell. Mol. Life Sci. 2011; 68:951–64.
19. Sato T, Oshima T, Yoshihara K, Yamamoto N, Yamada R, Nagano Y, Fujii S, Kunisaki C, Shiozawa M, Akaike M, Rino Y, Tanaka K and Masuda M. Overexpression of the fibroblast growth factor receptor-1 gene correlates with liver metastasis in colorectal cancer. Oncol. Rep. 2009; 21:211–216.
20. Griffith M, Griffith OL, Coffman AC, Weible J V, McMichael JF, Spies NC, Koval J, Das I, Callaway MB, Eldred JM, Miller CA, Subramanian J, Govindan R and et al. DGIdb: mining the druggable genome. Nat. Methods. 2013; 10:1209–10.
21. El-Gebali S, Bentz S, Hediger M a, Anderle P and et al. Solute carriers (SLCs) in cancer. Mol. Aspects Med. 2013; 34:719–34.
22. Kunická T SP. Importance of ABCC1 for cancer therapy and prognosis. Drug Metab. Rev. 2014.
23. Paccez JD, Vogelsang M, Parker MI and Zerbini LF. The receptor tyrosine kinase Axl in cancer: biological functions and therapeutic implications. Int. J. Cancer. 2014; 134:1024–33.
24. Craven RJ and X L. Receptor tyrosine kinases expressed in metastatic colon cancer. Int. J. Cancer. 1995; 60:791–797.
25. Dunne PD, McArt DG, Blayney JK, Kalimutho M, Greer S, Wang T, Srivastava S, Ong CW, Arthur K, Loughrey M, Redmond K, Longley DB, Salto-Tellez M and et al. AXL is a key regulator of inherent and chemotherapy-induced invasion and predicts a poor clinical outcome in early-stage colon cancer. Clin. Cancer Res. 2014; 20:164–75.
26. Ireton RC and Chen J. EphA2 receptor tyrosine kinase as a promising target for cancer therapeutics. Curr. Cancer Drug Targets. 2005; 5:149–57.
27. Saito T, Masuda N, Miyazaki T, Kanoh K, Suzuki H, Shimura T, Asao T and Kuwano H. Expression of EphA2 and E-cadherin in colorectal cancer: Correlation with cancer metastasis. Oncol. Rep. 2004; 11:605–611.
28. Kataoka H, Igarashi H, Kanamori M, Ihara M, Wang J-D, Wang Y-J, Li Z-Y, Shimamura T, Kobayashi T, Maruyama K, Nakamura T, Arai H, Kajimura M and et al. Correlation of EPHA2 overexpression with high microvessel count in human primary colorectal cancer. Cancer Sci. 2004; 95:136–41.
29. Bogan C, Chen J, O'Sullivan MG and Cormier RT. Loss of EphA2 receptor tyrosine kinase reduces ApcMin/+ tumorigenesis. Int. J. Cancer. 2009; 124:1366–71.
30. Herath NI, Doecke J, Spanevello MD, Leggett Ba and Boyd a W. Epigenetic silencing of EphA1 expression in colorectal cancer is correlated with poor survival. Br. J. Cancer. 2009; 100:1095–102.
31. Dong Y, Wang J, Sheng Z, Li G, Ma H, Wang X, Zhang R, Lu G, Hu Q, Sugimura H and Zhou X. Downregulation of EphA1 in colorectal carcinomas correlates with invasion and metastasis. Mod. Pathol. 2009; 22:151–60.
32. McDermott U, Downing JR and Stratton MR. Genomics and the continuum of cancer care. N. Engl. J. Med. 2011; 364:340–50.
33. Chang K, Creighton CJ, Davis C, Donehower L, Drummond J, Wheeler D, Ally A, Balasundaram M, Birol I, Butterfield YSN, Chu A, Chuah E, Chun H-JE and et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat. Genet. 2013; 45:1113–20.
34. Hopkins AL and Groom CR. The druggable genome. Nat. Rev. Drug Discov. 2002; 1:727–30.
35. Overington JP, Al-Lazikani B and Hopkins AL. How many drug targets are thereŒ. Nat. Rev. Drug Discov. 2006; 5:993–6.
36. Olaya-Abril A, Jiménez-Munguía I, Gómez-Gascón L and Rodríguez-Ortega MJ. Surfomics: shaving live organisms for a fast proteomic identification of surface proteins. J. Proteomics. 2014; 97:164–76.

37. Rehman M and Tamagnone L. Semaphorins in cancer: biological mechanisms and therapeutic approaches. Semin. Cell Dev. Biol. 2013; 24:179–89.

38. Conrotto P, Valdembri D, Corso S, Serini G, Tamagnone L, Comoglio PM, Bussolino F and Giordano S. Sema4D induces angiogenesis through Met recruitment by Plexin B1. Blood. 2005; 105:4321–9.

39. Basile JR, Castilho RM, Williams VP and Gutkind JS. Semaphorin 4D provides a link between axon guidance processes and tumor-induced angiogenesis. Proc. Natl. Acad. Sci. U. S. A. 2006; 103:9017–22.

40. Dieci MV, Arnedos M, Andre F and Soria JC. Fibroblast growth factor receptor inhibitors as a cancer treatment: from a biologic rationale to medical perspectives. Cancer Discov. 2013; 3:264–79.

41. Khan G, Moss R a, Braiteh F and Saltzman M. Proactive strategies for regorafenib in metastatic colorectal cancer: implications for optimal patient management. Cancer Manag. Res. 2014; 6:93–103.

42. Mahadevan D, Cooke L, Riley C, Swart R, Simons B, Della Croce K, Wisner L, Iorio M, Shakalya K, Garewal H, Nagle R and Bearss D. A novel tyrosine kinase switch is a mechanism of imatinib resistance in gastrointestinal stromal tumors. Oncogene. 2007; 26:3909–19.

43. Gioia R, Leroy C, Drullion C, Lagarde V, Etienne G, Dulucq S, Lippert E, Roche S, Mahon F-X and Pasquet J-M. Quantitative phosphoproteomics revealed interplay between Syk and Lyn in the resistance to nilotinib in chronic myeloid leukemia cells. Blood. 2011; 118:2211–21.

44. Liu L, Greger J, Shi H, Liu Y, Greshock J, Annan R, Halsey W, Sathe GM, Martin A-M and Gilmer TM. Novel mechanism of lapatinib resistance in HER2-positive breast tumor cells: activation of AXL. Cancer Res. 2009; 69:6871–8.

45. Byers LA, Diao L, Wang J, Saintigny P, Girard L, Peyton M, Shen L, Fan Y, Giri U, Tumula PK, Nilsson MB, Gudikote J and Tran H. An epithelial-mesenchymal transition gene signature predicts resistance to EGFR and PI3K inhibitors and identifies Axl as a therapeutic target for overcoming EGFR inhibitor resistance. Clin. Cancer Res. 2013; 19:279–90.

46. Heckmann D, Maier P, Laufs S, Li L, Sleeman JP, Trunk MJ, Leupold JH, Wenz F, Zeller WJ, Fruehauf S and Allgayer H. The disparate twins: a comparative study of CXCR4 and CXCR7 in SDF-1α-induced gene expression, invasion and chemosensitivity of colon cancer. Clin. Cancer Res. 2014; 20:604–16.

47. Verma A, Warner SL, Vankayalapati H, Bearss DJ and Sharma S. Targeting Axl and Mer kinases in cancer. Mol. Cancer Ther. 2011; 10:1763–73.

48. Burbridge MF, Bossard CJ, Saunier C, Fejes I, Bruno A, Léonce S, Ferry G, Da Violante G, Bouzom F, Cattan V, Jacquet-Bescond A, Comoglio PM and Lockhart BP. S49076 is a novel kinase inhibitor of MET, AXL, and FGFR with strong preclinical activity alone and in association with bevacizumab. Mol. Cancer Ther. 2013; 12:1749–62.

49. Bardelli A, Parsons DW, Silliman N, Ptak J, Szabo S, Saha S, Markowitz S, Willson JK V, Parmigiani G, Kinzler KW, Vogelstein B, Velculescu VE and et al. Mutational analysis of the tyrosine kinome in colorectal cancers. Science. 2003; 300:949.

50. Boyd AW, Bartlett PF and Lackmann M. Therapeutic targeting of EPH receptors and their ligands. Nat. Rev. Drug Discov. 2014; 13:39–62.

51. Montero JC, Seoane S, Ocaña A and Pandiella A. Inhibition of SRC family kinases and receptor tyrosine kinases by dasatinib: possible combinations in solid tumors. Clin. Cancer Res. 2011; 17:5546–52.

52. Brantley DM, Cheng N, Thompson EJ, Lin Q, Brekken R a, Thorpe PE, Muraoka RS, Cerretti DP, Pozzi A, Jackson D, Lin C and Chen J. Soluble Eph A receptors inhibit tumor angiogenesis and progression in vivo. Oncogene. 2002; 21:7011–26.

53. Dobrzanski P, Hunter K, Jones-Bolin S, Chang H, Robinson C, Pritchard S, Zhao H and Ruggeri B. Antiangiogenic and antitumor efficacy of EphA2 receptor antagonist. Cancer Res. 2004; 64:910–919.

54. Carles-Kinch K, Kilpatrick KE, Stewart JC and Kinch MS. Antibody targeting of the EphA2 tyrosine kinase inhibits malignant cell behavior. Cancer Res. 2002; 62:2840–2847.

55. Heemskerk B, Kvistborg P and Schumacher TNM. The cancer antigenome. EMBO J. 2013; 32:194–203.

56. Gjertsen MK, Buanes T, Rosseland a R, Bakka a, Gladhaug I, Søreide O, Eriksen J a, Møller M, Baksaas I, Lothe R a, Saeterdal I and Gaudernack G. Intradermal ras peptide vaccination with granulocyte-macrophage colony-stimulating factor as adjuvant: Clinical and immunological responses in patients with pancreatic adenocarcinoma. Int. J. Cancer. 2001; 92:441–50.

57. Overwijk WW, Wang E, Marincola FM, Rammensee H-G and Restifo NP. Mining the mutanome: developing highly personalized Immunotherapies based on mutational analysis of tumors. J. Immunother. cancer. 2013; 1:11.

58. Brown SD, Warren RL, Gibb EA, Martin SD, Spinelli JJ, Nelson BH and Holt RA. Neo-antigens predicted by tumor genome meta-analysis correlate with increased patient survival. Genome Res. 2014; 24:743–750.

59. Lawrence MS, Stojanov P, Polak P, Kryukov G V, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts S a, Kiezun A, Hammerman PS, McKenna A and et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013; 499:214–8.

60. Brahmer JR. Immune checkpoint blockade: the hope for immunotherapy as a treatment of lung cancer? Semin. Oncol. 2014; 41:126–32.

61. Naidoo J, Page DB and Wolchok JD. Immune Checkpoint Blockade. Hematol. Oncol. Clin. North Am. 2014; 28: 585–600.

62. Dolcetti R, Viel a, Doglioni C, Russo a, Guidoboni M, Capozzi E, Vecchiato N, Macrì E, Fornasarig M and Boiocchi M. High prevalence of activated intraepithelial cytotoxic T lymphocytes and increased neoplastic cell apoptosis in colorectal carcinomas with microsatellite instability. Am. J. Pathol. 1999; 154:1805–13.

63. Lipson EJ. Re-orienting the immune system: Durable tumor regression and successful re-induction therapy using anti-PD1 antibodies. Oncoimmunology. 2013; 2:e23661.

64. Dienstmann R, Serpico D, Rodon J, Saura C, Macarulla T, Elez E, Alsina M, Capdevila J, Perez-Garcia J, Sánchez-Ollé G, Aura C, Prudkin L, Landolfi S and et al. Molecular profiling of patients with colorectal cancer and matched targeted therapy in phase I clinical trials. Mol. Cancer Ther. 2012; 11:2062–71.

65. Manning G, Whyte DB, Martinez R, Hunter T and Sudarsanam S. The protein kinase complement of the human genome. Science. 2002; 298:1912–1934.

66. Li H and Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25:1754–1760.

67. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G and Durbin R. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25:2078–2079.

68. Kumar P, Henikoff S and Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat. Protoc. 2009; 4:1073–1081.

69. Adzhubei I, Jordan DM and Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr. Protoc. Hum. Genet. 2013; Chapter 7: Unit 7.20.

70. Reva B, Antipin Y and Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. 2011; 39:e118.

71. Trapnell C, Pachter L and Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009; 25:1105–11.

72. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ and Pachter L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. 2010; 28:511–5.

73. Mortazavi A, Williams BA, McCue K, Schaeffer L and Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods. 2008; 5:621–8.

74. Marinov GK, Williams BA, McCue K, Schroth GP, Gertz J, Myers RM and Wold BJ. From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. Genome Res. 2014; 24:496–510.

75. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehár J, Kryukov G V, Sonkin D, Reddy A, Liu M, Murray L and et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012; 483:603–7.

76. Segal NH, Parsons DW, Peggs KS, Velculescu V, Kinzler KW, Vogelstein B and Allison JP. Epitope landscape in breast and colorectal cancer. Cancer Res. 2008; 68:889–92.

77. Reche PA and Reinherz EL. Prediction of peptide-MHC binding using profiles. Methods Mol. Biol. 2007; 409: 185–200.

78. Lundegaard C, Lamberth K, Harndahl M, Buus S, Lund O and Nielsen M. NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11. Nucleic Acids Res. 2008; 36: W509–12.

79. Nielsen M, Lundegaard C, Lund O and Keşmir C. The role of the proteasome in generating cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal cleavage. Immunogenetics. 2005; 57:33–41.

ANEXO E: Súmula Curricular

SÚMULA CURRICULAR

DADOS PESSOAIS

Nome: Elisa Rennó Donnard Moreira
Local e data de nascimento: Belo Horizonte – MG, 08/04/1987

EDUCAÇÃO

2010-2014
Doutorado no Departamento de Bioquímica do Instituto de Química
Universidade de São Paulo

2009-2010
Mestrado em Bioquímica e Imunologia
Universidade Federal de Minas Gerais

2005-2008
Bacharelado em Ciências Biológicas
Universidade Federal de Minas Gerais

FORMAÇÃO COMPLEMENTAR

Janeiro 2014 – Curso de Estatística ESALQ-USP
Brazilian Edition of the Summer School for Statistical Genetics

Janeiro-Março 2008 – Iniciação Científica no Exterior
University of Manitoba
Supervisor: R. Daniel Gietz

OCUPAÇÃO

Bolsista de Doutorado, FAPESP, 01/02/2011-31/07/2014
www.bv.fapesp.br/en/bolsas/116362

PUBLICAÇÕES

1. ICRmax: an optimized approach to detect tumor-specific InterChromosomal Rearrangements for Clinical Application
   Donnard ER, Carpinetti P, Navarro FCP, Perez RO, Habr-Gama A, Parmigiani RB, Camargo AA and Galante PAF
   [em revisão]

2. Mutational analysis of genes coding for cell surface proteins in colorectal cancer cell lines reveal novel altered pathways, druggable mutations and mutated epitopes for targeted therapy
   Donnard ER*, Asprino PF*, Correa BR, Bettoni F, Koyama F, Navarro FC, Perez RO, Mariadason J, Sieber OM, Straussberg RL, Simpson AJG, Jardim DLF, Reis LFR, Parmigiani RB, Galante PAF and Camargo AA
   Oncotarget, 2014

3. Yeast Two Hybrid Liquid Screening
   Donnard ER, Queiroz EM, Ortega JM, Gietz RD
   Methods in Molecular Biology (Yeast Protocols 3rd Ed., chapter 7), 2014

4. Preimplantation development regulatory pathway construction through a text-mining approach
   Donnard ER, Barbosa-Silva A, Guedes RLM, Fernandes GR, Velloso H, Kohn MJ, Andrade-Navarro MA, Ortega JM
   BMC Genomics, 2011

5. PESCADOR, a web-based tool to assist text-mining of biointeractions extracted from PubMed queries
   Barbosa-Silva A, Fontaine JF, Donnard ER, Stussi F, Ortega JM, Andrade-Navarro MA
   BMC Bioinformatics, 2011