Universidade Federal de Pernambuco – UFPE
Centro de Informática – CIn

LUCAS MENEZES VERAS

Análise de Abordagens para o Problema de Detecção de Copy Number Variations

Trabalho de Graduação

Recife, Agosto de 2014

Trabalho de Graduação

LUCAS MENEZES VERAS

Análise de Abordagens para o Problema de Detecção de Copy Number Variations

Trabalho apresentado ao Programa de Graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação.

Orientadora: Katia Silva Guimarães

AGRADECIMENTOS

Agradeço à minha família por todo o apoio e carinho que me dedicaram, em especial a minha irmã, que sempre me inspira a enfrentar os desafios da vida.

Agradeço à professora Katia Silva Guimarães pela sua orientação, paciência e dedicação ao longo de várias cadeiras e deste projeto.

Agradeço também à professora Liliane Salgado, que me permitiu fazer parte do time da Maratona de Programação por três anos, onde ganhei experiência e conheci amigos. Agradeço em especial àqueles que formaram time comigo: Rafael Marinheiro, Gustavo Stor, e Israel Batista, que muito me ensinaram.

RESUMO

CNVs são variações genéticas estruturais. Sua detecção é um problema de grande interesse moderno na área de bioinformática, devido, em grande parte, à sua relação (positiva ou negativa) comprovada com diversas doenças como câncer, AIDS e neuropsiquiátricas. Neste trabalho, são estudadas diversas abordagens e sua aplicabilidade, a fim de se obter uma compreensão sobre os desafios apresentados e a forma como são enfrentados.

Palavras-Chave: Biologia Computacional, Copy Number Variation, Next Generation Sequencing

ABSTRACT

CNVs are structural genetic variations. Their detection is currently a heavily discussed problem in the area of bioinformatics, in part due to their proved (positive or negative) association with diseases like cancer, AIDS and neuropsychiatric disorders. In this text, we will study several approaches and their applicability, in order to obtain an insight into the challenges presented and how they are confronted.

Keywords: Computational Biology, Copy Number Variation, Next Generation Sequencing

Lista de abreviaturas e siglas	
aCGH	*Array Comparative Genomic Hybridization*
AS	*Assembly*
CNV	*Copy Number Variation*
EM	*Expectation Maximization*
EMRC	*Exon Mean Read Count*
HSLM	*Heterogeneous Shifting Level Model*
HMM	*Hidden Markov Model*
m-HMM	*mixture-Hidden Markov Model*
NGS	*Next Generation Sequencing*
PEM	*Paired End Mapping*
RC	*Read Count*
RD	*Read Depth*
SNP	*Single Nucleotide Polymorphism*
SR	*Split Reads*
SV	*Structural Variation*
WES	*Whole Exome Sequencing*
WGS	*Whole Genome Sequencing*

SUMÁRIO

INTRODUÇÃO

Copy Number Variations, ou CNVs, são um tipo de variação estrutural genética que pode variar em tamanho de algumas bases a milhões. Assim como outros tipos de variações, CNVs são frequentemente associados a doenças genéticas, e portanto é de grande interesse a capacidade de detectar regiões afetadas por esse tipo de modificação.

Esta é uma área da bioinformática em rápido crescimento, com avanços tecnológicos na manipulação de dados abrindo as portas para novos algoritmos que aprimoram abordagens já conhecidas. Como é característico da área, o volume de dados é enorme e todas as técnicas de resolução possuem vantagens e desvantagens, que muitas vezes são complementares. Assim, a escolha de abordagem depende do caso específico do problema que está sendo tratado.

Este trabalho fornece uma introdução e examina em detalhes algumas técnicas de detecção de CNVs, conforme foram publicadas, examinando seu desempenho de diversas formas. Espera-se que seja adquirido um conhecimento das dificuldades do problema, e do que deve ser levado em conta na escolha de um método adequado para a realização de experimentos.

DESENVOLVIMENTO

1. Conceitos básicos

1.1. Conceitos biológicos

O DNA é uma molécula presente nas células dos seres vivos, e é utilizado na produção de proteínas no núcleo da célula, assim dizendo-se que possui *instruções* para a construção de proteínas. Ele também é capaz de se replicar, o que permite a passagem de informação de células para suas descendentes. A molécula possui formato de dupla hélice, sendo composta de duas fitas de bases nitrogenadas. A sequência de bases nitrogenadas é o que codifica a informação do DNA, e há quatro tipos de bases, cada uma representada por uma letra: *adenina (A)*, *citosina (C)*, *guanina (G)* e *timina (T)*. Cada base é pareada através de uma ligação de hidrogênio com a base correspondente na outra fita da molécula, sendo que a adenina sempre é ligada a uma timina, a citosina é sempre ligada a uma guanina, e vice-versa. Assim, é preciso conhecer apenas uma fita da molécula para se considerar sua informação, e uma fita pode ser representada por uma sequência de letras A, C, G, T. O tamanho de uma molécula ou região de DNA é dado pela quantidade desses *pares de bases (bp)* [1].

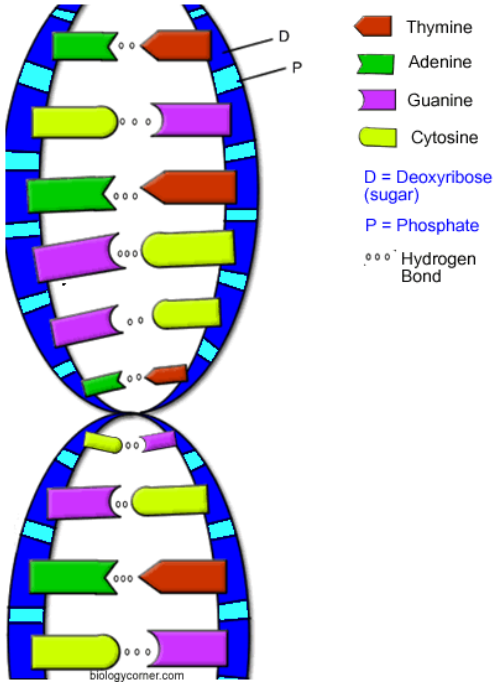

Figura 1. Representação gráfica do DNA, demonstrando as bases, suas ligações de hidrogênio, e a estrutura da fita.

Uma fita de DNA é usada, através de processos químicos, para se produzir proteínas. A proteína produzida depende da sequência de bases sendo lida, porém apenas algumas porções do DNA são utilizáveis. Assim, dizemos que existem regiões codificantes e regiões não codificantes da molécula. Nos humanos, a região codificante corresponde a menos de 2% do DNA [2].

O *genoma* é todo o material genético de um organismo, podendo ser codificado em DNA ou RNA (no caso de certos vírus). Ele inclui todos os genes do indivíduo, possuindo toda a informação necessária para construí-lo e mantê-lo [3]. O genoma humano consiste de mais de 3 bilhões de pares de bases, divididos em 23 pares de cromossomos [4].

Além de produzir moléculas, outra função vital do DNA é sua capacidade de se replicar, a fim de gerar novas células que possuam o mesmo material genético. Mas esta replicação não é perfeita, podendo ocorrer erros na cópia de certas bases ou regiões, o que é chamado de mutação. A maior parte das mutações são irrelevantes, mas outras são responsáveis por doenças genéticas. Elas também variam em tipo e raridade, com mutações presentes em mais de 1% da população sendo chamadas *polimorfismos* [5].

1.2. Sequenciamento de DNA

Foram desenvolvidas várias técnicas para obtenção de dados do DNA. O processo de se determinar a sequência de pares de bases é chamado de *sequenciamento*, com o maior sucesso do *Projeto Genoma Humano (HGP)* sendo o sequenciamento completo do DNA humano [6]. O processo de se sequenciar um genoma completo é chamado de *Whole Genome Sequencing (WGS)*, e há diversas técnicas modernas de sequenciamento como pirosequenciamento, Sanger, Maxam & Gilbert, sequenciamento de molécula única com exonuclease [7]. Apesar dos avanços tecnológicos, o custo de sequenciar um genoma inteiro ainda costuma ser proibitivo.

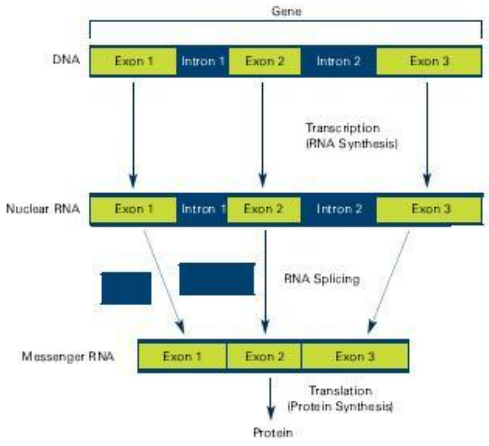

Figura 2. Representação gráfica de exons e introns. Um gene é uma região do genoma responsável por uma característica hereditária do organismo.

As instruções do DNA para a construção de uma proteína são "lidas" em duas etapas: transcrição e tradução. Na transcrição, uma fita da região do DNA a ser utilizada é replicada. Em um proceso chamado *splicing*, algumas partes da fita são removidas, e as restantes formam o mRNA que será traduzido. As partes que formam o mRNA são denominadas *exons*, pois serão expressas, e as descartadas são chamadadas *introns* [8].

O conjunto de todos os exons do DNA é denominado *exoma*. Por ser a região que efetivamente produz proteínas e ser consideravelmente menor que o DNA completo, é atraente a capacidade de economizar tempo e recursos realizando pesquisas apenas com o conhecimento do exoma. O processo de sequenciar o exoma é chamado de *Whole Exome Sequencing (WES)* [9].

1.3. CNVs

Há diversos tipos de variações genéticas, como polimorfismos de nucleotídeo único (SNPs), inserções, deleções, inversões, *variações do número de cópias (CNVs)*. Todas dão origem a seus próprios campos de estudo, com algoritmos de detecção e modelos estatísticos. Variações que afetam mais de 1 Kb (1000 bp) são denominadas *variações estruturais (SVs)*, como é o caso de CNVs [10].

CNVs foram definidos em 2006 por Redon et al. como um segmento de um Kb ou mais que está presente com *número de cópias* diferente em relação a um genoma de referência [11] [12]. Alguns CNVs não afetam a expressão dos genes, porém outros são relacionados a doenças [13]. Uma alteração no número de cópias significa que uma região do DNA aparece mais ou menos vezes no genoma sendo observado em relação ao de referência, resultado de uma deleção ou duplicação.

No caso de células diplóides, a ocorrência de uma região nos dois cromossomos de um par é contada como duas cópias. Assim, uma região presente nos dois cromossomos possui número de cópias igual a 2, uma região presente em apenas um cromossomo possui número de cópias 1, e uma região ausente em ambos possui número de cópias 0.

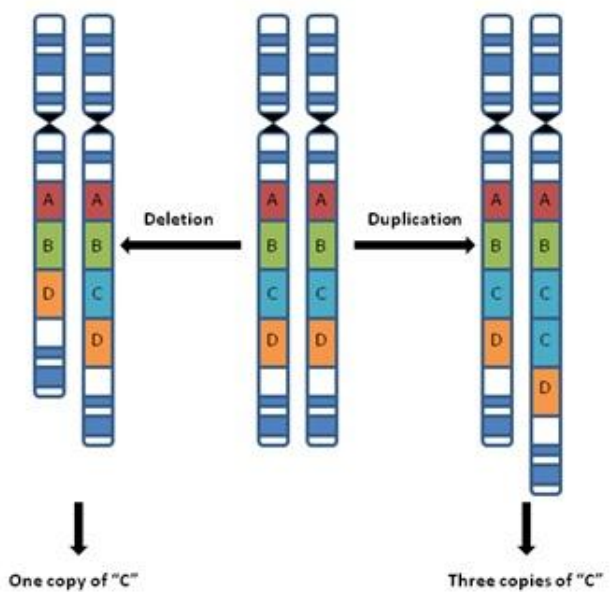

Figura 3 - Exemplo gráfico de uma variação no número de cópias da região C. Em relação ao cromossomo do meio, o da esquerda possui um número de cópias inferior, e o da direita possui um número de cópias superior.

2. Técnicas de detecção

Hibridização *in situ* fluorescente (FISH) é uma técnica de diversas aplicações, dentre estas a detecção de variações genéticas. Técnicas de hibridização molecular utilizam uma fita de DNA ou RNA (chamada de sonda) para identificar a posição *in situ* (natural) de seu complemento em uma amostra de DNA ou RNA. Maior parte do uso de hibridização *in situ* atualmente é utilizando procedimentos FISH [14].

Hibridização genômica comparativa com array (aCGH) é uma técnica específica para a detecção de variações do número de cópias, que combina os princípios de CGH com o uso de microarrays. A técnica consiste na coloração de um cromossomo de controle e um cromossomo de teste, e seu alinhamento com um microarray de sondas que representam diversas regiões do cromossomo. A razão entre o alinhamento de controle e o de teste é indicativa de ganho ou perda em certas regiões [15].

FISH e aCGH são consideradas técnicas tradicionais, que sofrem de baixa resolução (precisão na localização de eventos) das regiões genômicas. Apesar disso, a confiabilidade e natureza experimental do aCGH a tornam uma técnica bastante utilizada para validação dos resultados obtidos por outras técnicas. Com o surgimento de *sequenciamento de nova geração* (NGS), também chamado de *high-throughput sequencing* (HTS), novas técnicas têm sido desenvolvidas que utilizam dados de reads curtos. Métodos de NGS são capazes de sequenciar milhões e até bilhões de fitas de DNA em paralelo, obtendo um grande volume de dados num curto período de tempo. Estes dados vem na forma de *reads*, que são pequenas sequências de DNA de tamanho inferior a 1000 bp. A razão entre a quantidade de reads correspondentes a uma região e o tamanho desta região é chamada de *cobertura*, e uma razão maior tende a aumentar a capacidade dos algoritmos de analisar os dados. Devido à natureza recente dos procedimentos que utilizam estes dados para a detecção de CNVs, seu desempenho ainda não é muito bem entendido [16] [17].

2.1. Fontes de viés

Aplicações que usam dados de NGS tipicamente realizam um *alinhamento* dos reads com uma sequência de DNA, para encontrar uma localização na fita que corresponde a cada read, com certa margem de erro. Para isso são utilizadas ferramentas como BWA e Bowtie, sendo que há diferenças de implementação e tratamento de casos que devem ser levadas em consideração na escolha de um mapeador [18]. Como o DNA não é uma sequência aleatória, regiões similares ocorrem frequentemente, resultando em reads sendo mapeáveis para diversas localizações. Pode-se dizer que há reads que são *unicamente mapeáveis* e reads que são *multiplamente mapeáveis*. Isto define a mapeabilidade genômica, que pode ser controlada até certo ponto pelo tamanho dos reads e a margem de erro permitida durante o alinhamento. Alinhadores diferentes possuem

formas diferentes de tratar mapeabilidade, podendo retornar todas as localizações encontradas para um read, apenas uma, ou descartar reads que não possuem mapeamento único [19].

Conteúdo GC se refere à porcentagem de bases de uma região de DNA que são guanina ou citosina. O valor do conteúdo GC é correlacionado com a cobertura de reads da região, o que indica que ele é uma fonte de viés em experimentos que dependem da análise de reads mapeados. É essencial realizar processos de normalização do sinal que corrijam estas variações [20].

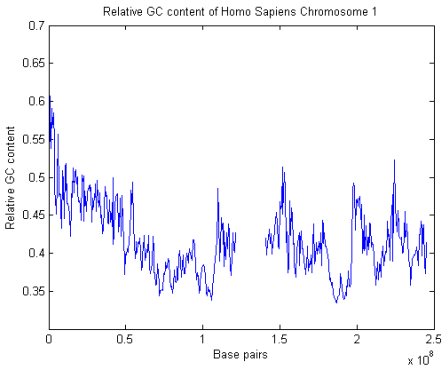

Figura 4. Variação do GC-content (eixo y) ao longo do cromossomo humano 1 (eixo x).

Processos que manipulam amostras fisicamente, como sequenciamento e microarrays, não são perfeitos e podem ter efeitos adversos nos dados. Estes efeitos dependem da técnica e tecnologia utilizadas, e por isso amostras que recebem o mesmo tratamento podem demonstrar *batch effects*, ou *efeitos de lote*. Esses efeitos podem, por exemplo, permitir a distinção de quais amostras foram sequenciadas por um mesmo laboratório dentre várias. Naturalmente, essa inserção de similaridades artificiais tem um impacto em experimentos realizados com os dados [21].

2.2. Abordagens gerais

Técnicas de sequenciamento de *extremidades pareadas* são capazes de retornar tanto o começo quanto o final de fragmentos de DNA, o que aumenta seu potencial de alinhamento e uso [22]. O primeiro uso de dados NGS para detecção de variações estruturais foi pelo método de *mapeamento de extremidades pareadas* (PEM). Nesta abordagem, variações são detectadas quando há discordância entre o mapeamento das extremidades de um read, tornando possível a detecção de inserções, deleções e inversões. Porém uma grande fraqueza da técnica é sua limitação à detecção de regiões de tamanho inferior ao comprimento médio dos reads. Para ser aplicável no estudo de CNVs e variações estruturais em geral, é preciso usar uma estratégia de clusterização ou baseada em modelos [20].

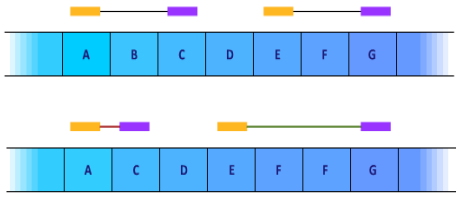

Figura 5. Exemplo de detecção de CNVs com mapeamento de extremidades pareadas.

Métodos de *Reads divididos* (SR) utilizam pares de reads, onde um read é alinhado ao genoma de referência enquanto o outro não conseguiu ser mapeado, ou foi mapeado parcialmente. Os reads com mapeamento incompleto são divididos em fragmentos, onde o primeiro e último são alinhados ao genoma independentemente. Este remapeamento permite localizar precisamente os locais de início e fim de eventos. Esta abordagem é altamente dependente do comprimento dos reads e só é aplicável em regiões unicamente mapeáveis [20].

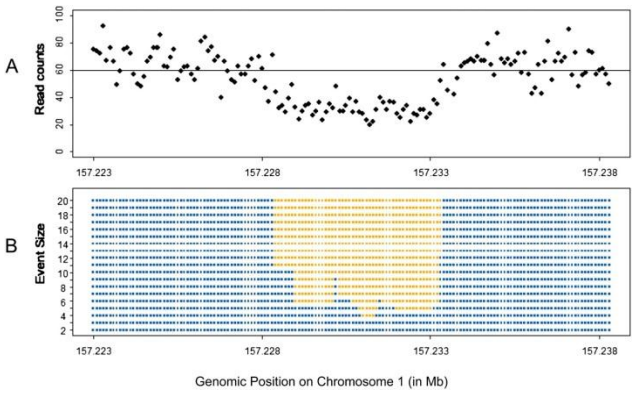

Figura 6. Exemplo de sinal de Read depth.

Métodos de *Profundeza de reads* (RD) consistem de quatro passos básicos: mapeamento, normalização, estimação do número de cópias e segmentação. Os reads são mapeados para suas localizações no genoma de referência, e são agrupados em "janelas" que contém um certo número de pares de bases. Assim temos um "sinal" de RD que é a sequência de números de reads mapeados para janelas consecutivas, porém este sinal é altamente volúvel devido a regiões similares de DNA e a forma como a coleta e alinhamento são realizados. Esta variância deve ser tratada por um método de normalização, após o qual é esperado encontrar um sinal alto em regiões onde ocorreu aumento do número de cópias, e um sinal baixo em regiões em que ocorreu redução do número de cópias. Após a previsão do número de cópias, as regiões alteradas são definidas a partir das janelas com números de cópias similares [20].

O uso de profundeza de reads tem crescido com o surgimento de dados NGS com alta cobertura, que resultam em uma melhor performance do método. Uma vantagem que ele tem em relação a PEM e SR é a capacidade de detectar o número de cópias exato de uma região, além de variações grandes em regiões complexas (que produzem mapeamentos únicos). Para os dois grandes tipos de sequenciamento, WGS e WES, há uma diversidade de métodos baseados em profundeza de reads [20].

Diferente dos outros métodos que começam com o alinhamento dos reads a um genoma de referência, métodos baseados em *Montagem* (AS) primeiro reconstroem fragmentos de DNA através da união de reads com sobreposição. A construção de regiões a partir de reads sem o uso

de uma referência é chamada montagem *de novo* (do começo), enquanto algumas técnicas podem usar uma referência como guia para aumentar a eficiência e qualidade da montagem. Estas regiões montadas são comparadas com o genoma de referência para detectar variações no número de cópias. Para satisfazer o critério de sobreposição, a abordagem depende de uma boa cobertura dos dados, enquanto que uma cobertura alta demais torna o processo mais lento, especialmente para *de novo* AS. Por não depender de alinhamento, estas técnicas têm potencial para detectar variações incomuns, porém não são muito adequadas para o estudo de CNVs [20].

Nenhuma das abordagens até então se mostrou capaz de detectar todos os tipos de CNVs com alta sensibilidade e especificidade. Analisando suas vantagens e desvantagens, elas se mostram complementares. PEM detecta CNVs de todos os tipos com alta precisão quanto à sua localização, porém não pode estimar o número de cópias e está limitado a regiões pequenas. RD também é capaz de detectar todos os tipos, pode identificar o número de cópias, e é aplicável a regiões grandes ao invés de pequenas, porém não é aplicável a variações que não mudam o número de cópias (como translações e inversões) e tem dificuldade em apontar o começo e fim de regiões. Técnicas de AS podem detectar variações incomuns, porém são computacionalmente caras e não são boas para certos tipos de variação. Técnicas de SR possuem altíssima resolução na detecção de regiões, porém sofrem com regiões de baixa complexidade que resultam em mapeamentos ambíguos. Como resultado destas diferenças, foram desenvolvidas ferramentas que buscam combinar métodos distintos a fim aproveitar os pontos fortes de cada um [20].

3. Artigo: m-HMM

Apesar da sua robustez, os métodos bem-desenvolvidos de análise de dados do aCGH não são capazes de tomar proveito total dos dados de NGS, é preciso desenvolver técnicas mais adequadas. Sendo uma área de pesquisa recente, a capacidade de detecção de CNVs usando NGS ainda é limitada pelo conhecimento de modelos estatísticos aplicáveis a estes dados. Em um artigo por Heng Wang, Dan Nettleton e Kai Ying, é proposta uma nova metodologia para tal [23].

Maioria dos métodos de detecção usando NGS podem ser classificados em duas categorias: *janela* e *modelo oculto de* Markov (HMM). Modelos de janela deslizante incluem *SegSeq*, *Event-wise testing*, *rSW-seq* e *jointSLM*. Um exemplo de método que usa um modelo oculto de Markov é CNAseg, que utiliza dois genomas adicionais para obter informações. O

método proposto, *m-HMM (mixture-HMM)*, utiliza apenas uma amostra sequenciada de cada genótipo, o que o torna mais viável quando não há amostras extras disponíveis.

3.1. Terminologia do problema

Genoma de referência se refere ao genoma do genótipo que foi completamente sequenciado através de WGS. *Genoma alvo* é o genoma do genótipo que não foi totalmente sequenciado. O objetivo é encontrar regiões do genoma alvo que sofreram alterações no seu número de cópias em relação ao genoma de referência. O local em que ocorre a mudança de um estado do número de cópias para outro é chamada de *ponto de mudança* ou *breakpoint*. Os dados do genoma alvo vem na forma de *reads*, que são *alinhados* ao genoma de referência para saber sua posição de origem, que é dada pela posição da primeira base do read. Uma posição que tenha uma contagem de reads positiva no genoma de referência ou no genoma alvo é chamada de *sítio*, assim os dados são valores não-negativos chamados *contagens de reads* (RC) com sítios associados no genoma de referência.

3.2. Processamento preliminar

Os valores originais de contagens de reads contém uma grande quantidade de zeros, o que torna mais prático agrupá-los em janelas. Cada janela corresponde a uma sequência de DNA do genoma de referência, e seu valor de RC é a soma dos RCs de suas bases individuais. Dois métodos são frequentemente utilizados para definir janelas: definir regiões de tamanho (em bp) fixo e RC variável, ou regiões que possuam um valor fixo de RC e tamanho variável. Devido à alta variação na densidade de sítios ao longo do genoma, as duas abordagens possuem fraquezas: Janelas com RC fixo podem agrupar sítios muito distantes, enquanto que janelas de tamanho fixo produzem resultam numa alta variância no número de reads.

Para obter um equilíbrio entre as abordagens, foi usado um método de *clusterização K-médias* para definir janelas levando em conta as distâncias e quantidades de sítios. Cada cromossomo foi dividido em 20 regiões (valor determinado pelos autores para uso em casos práticos) e em cada uma é realizada a clusterização, onde K é o número de sítios da região dividido por uma constante C (definida como 40 nos experimentos deste artigo). As janelas são definidas pelos sítios de um cluster. Agora temos duas sequências ($u_1^{[t]}$, $u_2^{[t]}$, …, $u_W^{[t]}$) e

($u_1^{[r]}$, $u_2^{[r]}$, …, $u_W^{[r]}$) de valores de RC para as W janelas do cromossomo nos genomas de alvo ([t]) e referência ([r]), além de conjuntos g_w de índices dos sítios da janela w. A localização da janela é definida como a mediana dos sítios da janela, então temos outra sequência de valores $(l_1, l_2, \dots, l_W)$. Assim como ocorre em outros métodos, é assumido que o número de cópias não muda dentro de janelas (devido à raridade de ocorrer mudanças), e é feita uma segmentação das janelas através do método m-HMM para determinar seus estados. Em cima deste resultado é feito um procedimento de ajuste que torna mais precisas as localizações dos breakpoints.

3.3. Modelo de Markov

O modelo oculto de Markov foi desenvolvido por Baum, et al. em diversos artigos [24] [25] [26] [27]. Ele consiste de um processo aleatório bivariado $\{S_w, U_w\}, w = 1, \dots, W$, onde o componente $\{S_w\}$ é uma cadeia de Markov finita não observada. Ou seja, o estado S_{w+1} é determinado somente pelo estado S_w, e os estados anteriores são irrelevantes. Os estados são ditos escondidos porque a sequência de estados que melhor representa os dados observados (no caso, o DNA) é desconhecida. No caso deste problema, os estados escondidos são os números de cópias das janelas, que podem assumir quatro valores: 1 = ganho, 2 = normal, 3 = perda, 4 = ausência. Resolver o modelo oculto de Markov é, portanto, determinar a sequência de estados que possui maior probabilidade de gerar a sequência observada, assim dizendo para cada janela seu número de cópias dentro de quatro possíveis. Modelos ocultos de Markov podem ser resolvidos eficientemente usando o algorito de Viterbi.

A probabilidade $P(S_{w+1}|S_w)$ de se ir para o estado S_{w+1} a partir do estado S_w é chamada de *probabilidade de transição*, e será uma matriz de 4x4 probabilidades para cada possível mudança de estado. O componente $\{U_w\}$ é uma sequência de observações, que corresponde aos valores de RC nos genomas de referência e alvo ($u_w^{[r]}$ e $u_w^{[t]}$). Para cada observação e cada possível estado da janela, existe uma probabilidade $P(U_w|S_w)$ chamada *probabilidade de emissão*, que é a chance de esta observação U_w ser gerada pelo estado S_w.

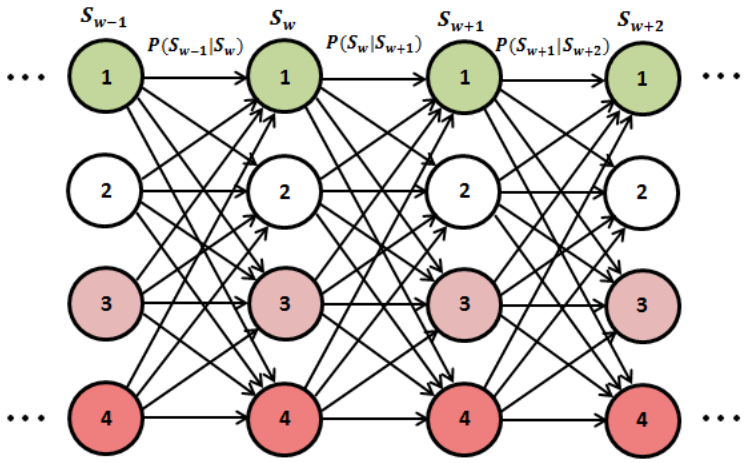

Figura 7. Representação em grafo de um modelo oculto de Markov. Cada estado S_w possui quatro possíveis números de cópia, e cada probabilidade de transição $P(S_w|S_{w+1})$ é uma matriz 4x4 de probabilidades $a_{kl}(w)$. É possível ver que todas as arestas de um estado S_w vêm do estado S_{w-1}.

Dado que uma janela w tem estado k, a probabilidade condicional da janela $w+1$ ter estado l é $a_{kl}(w), (k,l = 1,2,3,4)$. Para cada janela, temos uma matriz de probabilidades de transição, onde a probabilidade de ocorrer uma mudança no estado ($k \neq l$) aumenta conforme a distância entre as janelas w e $w+1$ aumenta. Mais detalhes da definição matemática da matriz são encontrados no artigo original.

As probabilidades de emissão $P(U_w|S_w)$ são determinadas por distribuições de emissão. A distribuição para as contagens de reads de referência é simplesmente uma distribuição de $Poisson(\lambda_w^{[r]})$, mas o uso da mesma para o alvo é problemático. A fim de definir o valor de contagem de reads esperado baseado no número de cópias da região, uma distribuição de $Poisson(K_k c_0 \lambda_w^{[r]})$ usaria o valor $(K_1 = 2,\ K_2 = 1,\ K_3 = 0.5,\ K_4 = 0)$ e um fator de normalização c_0 para dar o valor esperado de RC a partir do seu valor observado no genoma de referência $(\lambda_w^{[r]})$. Mas devido à alta variância nos valores de RC resultantes de diversos vieses e nuâncias dos dados e alinhamento, é bastante comum ocorrer regiões que apresentem esta variação na contagem de reads por fatores que não são uma mudança no número de cópias. Assim, este modelo ingênuo detectaria uma quantidade grande de falsos positivos.

Enquanto métodos similares definem procedimentos para combater estas variações, o m-HMM define a distribuição de probabilidade para as observações do alvo como sendo um *modelo mistura de Poisson*, que possui parâmetros para levar em conta erros diversos que estão misturados com a variação esperada. Neste modelo, o efeito do estado do número de cópias é dominante sobre efeitos de variações aleatórias. É preciso então encontrar uma estimação de máxima verossimilhança para o vetor de parâmetros do modelo, ou seja, o vetor que possui

maior probabilidade de gerar os valores observados de RC. Este vetor possui 26 dimensões, logo para maximizar a sua verossimilhança é utilizado um algoritmo de *maximização de esperança* (EM). A definição mais detalhada deste algoritmo e da inicialização de seus parâmetros está disponível no artigo. O uso deste modelo mistura é o que dá ao algoritmo o nome *mixture-HMM*.

Após a convergência dos parâmetros por maximização de esperança, é realizada a previsão dos estados escondidos resolvendo-se o modelo oculto. Então, tendo a probabilidade condicional para cada estado, a previsão para cada janela *w* é dada pelo estado que tem maior probabilidade condicional. Por fim, para levar em conta que o número de cópias pode mudar dentro de janelas, é utilizado um algoritmo que ajusta os breakpoints utilizando teste chi-quadrado de Pearson.

3.4. Testes de performance

Para avaliar o desempenho do método em relação a outras abordagens, foram realizados três testes com CNVs simulados. O primeiro teste foi utilizando dados reais de uma célula de câncer pulmonar. As posições dos sítios foram distribuídas aleatoriamente segundo uma distribuição uniforme, e as contagens de reads do genoma alvo foram geradas misturando as contagens de reads do cromossomo 4 de referência. Foram escolhidos aleatoriamente 90 CNVs no genoma de alvo, de tamanhos 10, 50 e 100 kb, e as suas contagens de reads foram modificadas de acordo com o tipo de CNV. A cobertura do sequenciamento é de apenas 2X. Os resultados são comparados com os do m-HMM sem a realização de ajuste, o HMM original que usa probabilidades de emissão de Poisson, e SegSeq. As medidas utilizadas são:

$$Sensitividade = \frac{número\ de\ CNVs\ identificados\ corretamente\ de\ um\ tipo}{número\ verdadeiro\ de\ CNVs\ de\ um\ tipo}$$

$$Especificidade = \frac{número\ de\ sítios\ normais\ identificados\ corretamente}{número\ verdadeiro\ de\ sítios\ normais}$$

$$EFPR\ (taxa\ empírica\ de\ falsos\ positivos) = \frac{número\ de\ CNVs\ incorretos\ de\ um\ tipo}{número\ de\ CNVs\ detectados\ de\ um\ tipo}$$

$$EFNR\ (taxa\ empírica\ de\ falsos\ negativos) = \frac{número\ de\ sítios\ normais\ incorretos}{número\ de\ sítios\ normais\ identificados}$$

		m-HMM	M-HMM sem	HMM	SegSeq
Ganho	Sensibilidade	0.917	0.915	0.863	0.000
	EFNR	0.097	0.174	0.751	0.000
Normal	Especificidade	0.996	0.994	0.945	0.998
	EFPR	0.002	0.002	0.003	0.019
Perda	Sensibilidade	0.858	0.831	0.826	-
	EFNR	0.309	0.434	0.846	-
Ausente	Sensibilidade	0.960	0.857	0.810	-
	EFNR	0.011	0.005	0.000	-
Perda/	Sensibilidade	0.898	-	-	0.076
Ausente	Especificidade	0.223	-	-	0.678

Tabela 1. Valores de sensibilidade, especificidade, falsos positivos e negativos para os quatro tipos de CNVs. SegSeq detecta ausência como perda, então estes tipos foram agrupados e comparados somente com o m-HMM. Ele também não detectou alterações de ganho, então obteve sensibilidade e EFNR 0 nesse tipo.

Quando comparados com o HMM normal, o m-HMM mesmo sem ajustes possui maior especificidade/sensibilidade, e taxas de falsos positivos/negativos mais baixas, com a única exceção sendo a EFNR para ausência. Com o ajuste, é percebido um aumento em ambos, com uma redução de EFNR e EFPR exceto no caso de ausência. Apesar de ser melhor na detecção de estados normais, o SegSeq teve performance bem pior na detecção de variações.

No segundo teste, foi estudada a capacidade de detecção de variações com diferentes comprimentos. Para cada tamanho dentre 10, 20 30, 50 e 100 kb, foram criadas 30 variações, com 10 de cada tipo (ganho, perda e ausência). O procedimento de simulação é o mesmo do anterior, e o critério para corretude da detecção é que haja uma interseção não-vazia entre um segmento identificado e um segmento criado com o mesmo tipo. Quase todos os segmentos de tamanho maior ou igual a 30 kb foram encontrados. Embora seja possível notar uma mudança no comportamento segundo o tamanho de CNVs, este teste não revelou a especificidade dos resultados, e o critério de corretude tal como descrito pode ser relaxado demais.

	10 kb	20 kb	30 kb	50 kb	100 kb
Perda	0	4	9	10	10
Ganho	3	5	10	10	10
Ausente	2	9	10	10	10

Tabela 2. Número de CNVs detectados de cada tipo e tamanho

Para o terceiro teste, foi utilizado o genoma de referência de milho versão 2, inserindo variações de tamanho 1, 5 e 10 kb no cromossomo 6 para produzir um novo genoma com CNVs, novamente com 10 de cada tipo. Foram criados reads aleatórios, desta vez com cobertura de 30X. Foi feita uma comparação entre m-HMM, SegSeq e CNVnator, incluindo a quantidade de regiões normais que foram incorretamente classificadas como CNVs pelos métodos. É interessante observar a melhoria na detecção de variações pequenas em relação ao teste anterior, provavelmente ocasionada pelo aumento na cobertura de 2X para 30X.

		m-HMM			SegSeq			CNVnator		
		1 kb	5 kb	10 kb	1 kb	5 kb	10 kb	1 kb	5 kb	10 kb
Ganho	Correto	10	10	10	7	10	10	6	9	10
Perda/ausência	Correto	9	10	10	7	4	6	8	10	10
Normal	Falsos pos.	0	0	3	1	7	6	436	466	399

Tabela 3. CNVs detectados corretamente e regiões normais marcadas incorretamente pelos três algoritmos. Tanto SegSeq quanto CNVnator não diferenciam perda de ausência, então os resultados foram agrupados.

Por fim, o m-HMM foi usado para comparar dois genótipos de milho: B73 como referência e Mo17 como alvo. Ao invés de reads e CNVs simulados, este foi um teste de aplicação usando dados reais. Foram obtidos 4.3 milhões de reads alinhados à referência e 1.54 milhões ao alvo, com 2.3 milhões de posições possuindo contagem de reads positiva na referência. Foram detectadas 1096 variações de tamanho maior que 2 kb, das quais 14 são de ganho, 835 são de perda e 247 são de ausência.

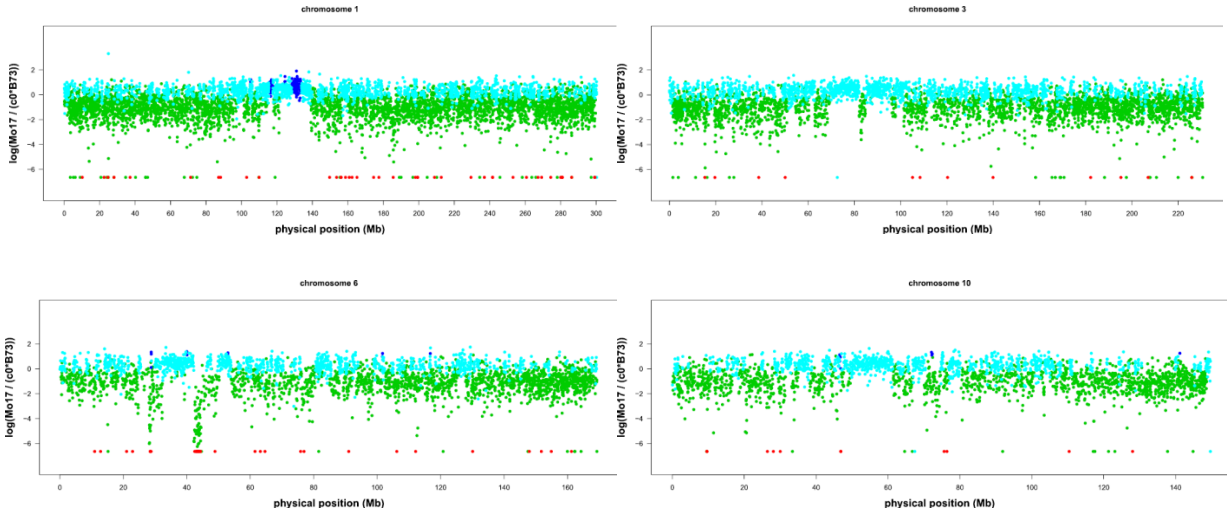

Figura 8. Resultados da aplicação do m-HMM aos cromossomos 1, 3, 6 e 10 do milho. O eixo horizontal representa a posição genômica, enquanto o vertical representa $log\left(\frac{Mo17}{c_0 B73}\right)$, setado para -7 em caso de ausência. Pontos azuis, azuis claro, verdes e vermelhos representam respectivamente regiões de ganho, normais, perda e ausência.

De acordo com resultados anteriores obtidos por aCGH. há diversos segmentos com poucas variações, notavelmente de 121.3 ~ 130 Mb no cromossomo 1, 69.2 ~ 82 Mb e 84.8 ~ 95.9 Mb no cromossomo 3, e 49.5 ~61.5 Mb no cromossomo 10. Também de acordo, a maior perda ocorre no cromossomo 6, de 42.2 a 46.2 Mb.

É importante notar que há uma grande divergência entre os genomas Mo17 e B73, o que dificulta o mapeamento de reads em certas regiões. Isso parcialmente justifica a alta taxa de alterações identificadas como perda ou ausência. O curto comprimento dos reads (44 bp) também facilita mapeamentos incorretos. A baixa cobertura (inferior a 1X) dificulta a identificação de CNVs pequenos, como foi observado nos testes. Estas dificuldades tendem a ser menos relevantes com o desenvolvimento de tecnologias que permitem obter coberturas de 30X, e reads de no mínimo 100 bp. Também é relevante notar que o viés de conteúdo GC não deve afetar o algoritmo desde que o conteúdo GC interno das janelas permaneça consistente entre genomas.

Os experimentos indicam que o m-HMM, aproveitando-se de novos modelos estatísticos em relação a outras técnicas, possui potencial de detecção escalável com o avanço de tecnologias de sequenciamento. A comparação direta entre os genomas de alvo e referência, sem a necessidade de amostras adicionais para remoção de ruído, tornam viável seu uso em casos de pouca disponibilidade de dados.

4. Artigo: Comparação de 4 métodos para WES

4.1. Detecção de CNVs usando WES

Dados de WES são o bastante para se estudar diversos tipos de variações genéticas, mas a natureza esparsa dos exons os torna inadequados para o estudo de variações estruturais, como é o caso de CNVs [9]. Dentre as principais abordagens para detectar CNVs a partir de paired-end reads vindos de NGS (profundeza de reads, mapeamento de extremidades pareadas, reads divididos, montagem, ou uma combinação destes), apenas a de profundeza de reads conseguiu ser adaptada para trabalhar com o exoma, as demais permanecendo restritas a WGS [18]. Ferramentas desenvolvidas para WES buscam reduzir os efeitos de vieses diversos [9].

Em um artigo por Renjie Tan, Yadong Wang, dentre outros, foram comparadas quatro técnicas de detecção de CNVs usando dados WES: XHMM (eXome-Hidden Markov Model), CoNIFER (Copy Number Inference From Exome Reads), ExomeDepth v0.9.7 e CONTRA (Copy Number Analysis for Targeted Resequencing) [18]. Através de seis métricas, concluiu-se que todas as técnicas possuem vantagens e desvantagens, e todos se mostraram inferiores a métodos que utilizaram dados de WGS. As métricas são tamanho/distribuição de CNVs, concordância com dados de WGS, descoberta em comum com tagSNPs, taxa de erro mendeliana, teste de heterozigose para deleções, e concordância entre algoritmos de detecção. Foram usadas 33 amostras sequenciadas, das quais 13 também foram sequenciadas usando WGS.

Há menos métodos que utilizam WES comparado a métodos que utilizam WGS, e todos são altamente diferentes em termos de modelo estatístico, linguagem, formato de entrada e saída. Alguns são específicos para detectar variações estruturais somáticas ou germinativas [18]. O trabalho de R. Tan avalia a detecção de variações germinativas, excluindo ferramentas especializadas e variações somáticas, resultando nos quatro métodos mencionados. Embora seja possível melhorar suas performances através de uma otimização de parâmetros, os testes foram realizados usando parâmetros padrão a fim de manter a imparcialidade.

4.2. Métricas

Medida 1: O tamanho de CNVs varia de poucos pares de bases a milhões, porém os CNVs mais abundantes são pequenos (~300 bp) e médios (~6 kb). Métodos baseados em RD

mapeiam reads para uma janela que consiste de uma sequência de pares de bases, e a confiabilidade do sinal aumenta com o tamanho da janela. Assim, a detecção de CNVs pequenos por esses métodos é um desafio, mas são os únicos aplicáveis a WES. CoNIFER detectou menos CNVs dentre os quatro métodos, uma mediana de 13 por exoma, devido a sua incapacidade de detectar CNVs menores que 1 kb. ExomeDepth e XHMM detectaram CNVs pequenos e grandes. CONTRA obteu o maior número de detecções, 811 por exoma, com 96.30% sendo menores que 1 kb. Em geral, poucos CNVs maiores que 10 kb foram detectados.

	< 1 kb	**1–10 kb**	**10–100 kb**	**≥ 100 kb**	**Geral**
XHMM	2 0.52	10 4.31	27 34.52	11 211.65	51 34.02
CoNIFER	0 NA	2 4.51	7 32.48	2 184.22	13 35.13
ExomeDepth	103 0.36	98 3.72	110 27.96	27 219.34	346 5.49
CONTRA	781 0.2	17 1.49	0 16.82	0 NA	811 0.22

Tabela 4. O primeiro valor de cada célula é a mediana de CNVs encontrados, enquanto o segundo é a mediana do tamanho dos CNVs.

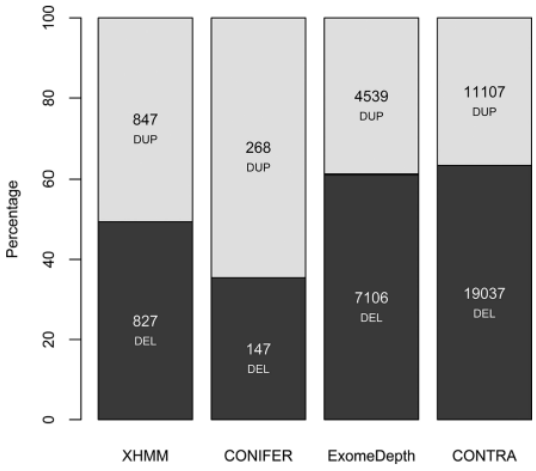

Figura 9. Quantidades totais de CNVs detectados nas 33 amostras, classificados entre deleções e duplicações.

Medida 2: Treze das amostras foram sequenciadas usando WGS, e utilizadas por dois algoritmos conhecidos de detecção: CNVnator e ERDS. Considerando apenas os CNVs identificados por ambos os métodos (a fim de aumentar a confiabilidade da validação), temos uma mediana de 2802 exons marcados com número de cópias alterado. Destes, temos que 509 exons (18.17%) foram marcados pelo ExomeDepth nas amostras WES correspondentes, 215 pelo XHMM (7.67%), 121 pelo CONTRA e 89 pelo CoNIFER (3.18%). Porém, os tamanhos das variações não foram incluídos na análise, o que seria útil para complementar o resultado anterior, embora seja possível inferir que maior parte das chamadas do CONTRA não foram validadas.

Medida 3: tagSNPs são SNPs representativos em certas regiões do genoma. Diversos CNVs comuns podem ser "marcados" por tagSNPs próximos, e um conjunto destes CNVs foi

identificado no 1000 Genomes Project. Como outra forma de validar os resultados, foi testada a proporção de CNVs marcados identificados pelos quatro algoritmos: Da mediana de 243 exons em 21 amostras, o ExomeDepth detectou 84 (34.57%), CONTRA detectou 38 (15.64%), XMM detectou 33 (13.58%) e CoNIFER detectou 4 (1.65%). ExomeDepth se mostra mais sensível para a detecção de CNVs comuns, o que é consistente com as outras métricas.

Medida 4: Não é viável avaliar exatamente a especificidade dos algoritmos através de validações experimentais dos CNVs encontrados. Dentre as 33 amostras, há nove trios de pais e filhos, e foi testada a regra de hereditariedade mendeliana nos CNVs dos filhos como outra forma de validação. CNVs que violem a regra são considerados falsos positivos, pois a taxa de surgimento de CNVs *de novo* (não presentes nos pais) é incrivelmente baixa. Assim, as taxas de erro Mendelianas foram usadas para inferir as taxas de falsas descobertas, que se mostraram altas para todos os algoritmos, mesmo com critérios relaxados. Para se reduzir o bias causado por baixa sensibilidade nos pais, foram removidos da análise CNVs presentes em outras amostras WES, levando em conta apenas as variações únicas ao trio. CONTRA obteve a menor taxa, de (10.68%), seguido por ExomeDepth (16.11%), XHMM (20%) e CoNIFER (56.25%).

Medida 5: Em regiões onde ocorreu uma deleção, há no máximo uma cópia da região no genoma. Assim, estas regiões só podem possuir SNVs (single nucleotide variants) homozigóticas, e a presença de SNVs heterozigóticas implica na região ser um falso positivo. Para levar em conta o grande ruído dos dados WES, o critério foi relaxado para: conter 2 ou mais SNVs heterozigóticas, e mais de 20% das SNVs da região serem heterozigóticas. Neste teste, apenas 3.77% das deleções do CONTRA são falsos positivos, 24.86% do ExomeDepth, 40% do CoNIFER e 54.35% do XHMM. Porém como grande parte dos CNVs do CONTRA são pequenos e não possuem SNVs, foi verificado que o teste restringido a deleções maiores que 1 kb aumenta as taxas para 18.75%, 34.02%, 40% e 64.10%, mantendo a mesma ordem dos algoritmos.

Medida 6: Boa parte dos CNVs identificados por um algoritmo costumam não ser encontrados por outros algoritmos. CNVs identificados simultaneamente por múltiplos algoritmos tendem a possuir maior especificidade, portanto foi feita a sobreposição dos resultados das quatro ferramentas. 78.10% dos exons com CNV detectados pelo CoNIFER

também foram detectados por outros, seguido por 48.54% do XHMM, 33.38% do CONTRA e 22.19% do ExomeDepth.

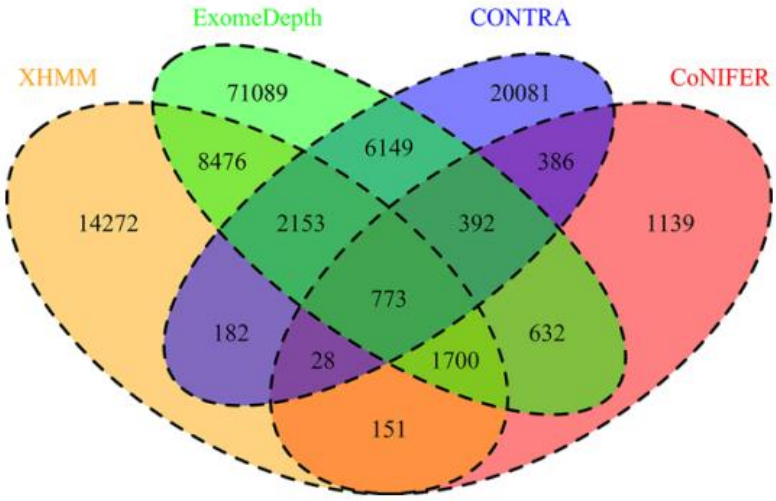

Figura 10. Diagrama de Venn descrevendo a sobreposição de exons com variação no número de cópias detectados pelas quatro técnicas ao longo de 33 amostras. O critério de sobreposição é que um exon tenha ao menos 50% de seu comprimento coberto por um CNV de outra ferramenta.

4.3. Conclusões

Uma diferença entre os quatro algoritmos que se tornou aparente nos resultados é o processo de normalização do sinal de read-depths. XMMM e CoNIFER usam métodos de PCA/SVD (Principal Component Analysis/Singular Value Decomposition), que utilizam um conjunto de amostras para remover ruído, enquanto ExomeDepth e CONTRA utilizam métodos baseados na razão do RD. A consequência de métodos PCA/SVD é que CNVs comuns aparecem frequentemente na população, e sua contribuição para o sinal de RD acaba sendo removida pela normalização. Porém, isto não afeta a detecção de CNVs raros, que são de grande interesse no estudo de doenças. A capacidade destes métodos de detectar variações raras é reafirmada no próximo artigo. Além disso, métodos de razão obtém uma maior taxa de falsos positivos em relação a PCA/SVD, como ilustrado pela mediana de CNVs detectados unicamente por um dos métodos: 3.23% a 9.09% para XHMM e CoNIFER, e 9.09% a 15.15% para ExomeDepth e CONTRA.

Em relação ao tamanho das variações detectadas, os modelos mostraram várias diferenças. CoNIFER possui um limiar rígido, aumentando sua especificidade mas reduzindo bastante sua

sensibilidade para CNVs pequenos, não fazendo chamadas menores que 1 kb nesse estudo. Em contrapartida, o CONTRA somente pode detectar variações em regiões pequenas, resultando num alto número de detecções, e usa uma heurística para prever variações grandes. O modelo de HMM usado pelo XHMM e ExomeDepth é capaz de detectar diversos tamanhos de CNV, e ExomeDepth possui a maior sensibilidade dentre os métodos, apesar de também possuir uma alta taxa de falsos positivos. Seria interessante realizar experimentos de validação similares distinguindo variações de diferentes tamanhos, porém o grande número de CNVs e a alta taxa de falsos positivos dificulta a validação a partir de métodos experimentais.

Foram realizados diversos tipos de validação, dentre os quais se destaca o uso de CNVs detectados por WGS. Seu uso no lugar de métodos experimentais (como aCGH ou SNP arrays, que não estão disponíveis para estas amostras) pode beneficiar a sensibilidade, principalmente para CNVs pequenos. A fim de melhorar a especificidade do conjunto de validação, foi utilizada a combinação de dois métodos de WGS, porém sua precisão ainda não é tão confiável.

Um fator que não foi avaliado inteiramente neste trabalho é o método de alinhamento de reads utilizado. Reads sequenciados de regiões repetitivas são tratados de forma diferente pelos métodos mrFAST e mrsFAST em relação aos métodos BWA e Bowtie. Neste estudo foi utilizado o Bowtie por seus méritos, porém o impacto das ferramentas de alinhamento na detecção de CNVs não foi avaliado.

Foi observado que nenhum método obteu boa performance em todas as métricas, e todos possuem vantagens e desvantagens bem distintas. Ainda há muito a ser explorado em termos de remoção de ruído em dados WES, e usuários devem fazer uma escolha cuidadosa de métodos caso seja desejem realizar experimentos com este tipo de dados.

5. Artigo: EXCAVATOR

Em um artigo por Alberto Magi, Lorenzo Tattini, dentre outros, foi desenvolvido um pacote de software para detecção de CNVs utilizando dados WES chamado EXCAVATOR (EXome Copy number Alterations/Variations annotATOR). A sua performance foi comparada com três outros algoritmos: XHMM, CoNIFER e ExomeCNV [9]. Os testes foram realizados com três conjuntos distintos de dados WES: Uma população gerada pelo 1000 Genomes Project,

e dois conjuntos gerados pelos autores contendo amostras de melanoma e deficiência intelectual. Os resultados foram validados utilizando SNP arrays.

ExomeCNV foi a primeira ferramenta implementada para detecção de CNVs usando WES. Ele usa um processo de normalização de dois passos para reduzir os efeitos de conteúdo GC e mapeabilidade, ainda sendo sucetível a efeitos de lote. Ele usa o algoritmo de *circular binary segmentation* (CBS) para detectar os limites das regiões alteradas, sendo que este não leva em conta as distâncias entre exons adjacentes, tornando-o inadequado para regiões esparsas. CoNIFER e XHMM usam SVD e PCA para remover as principais causas da uniformidade do sinal de RD. Porém estes procedimentos necessitam de um conjunto de amostras que sirva de população, o que restringe sua aplicabilidade a projetos que disponham destas amostras. Todas estas ferramentas classificam regiões em apenas três estados: redução, aumento e manutenção do número de cópias.

5.1. Estratégia

O EXCAVATOR também utiliza uma abordagem de contagem de reads, além de um processo de normalização de três passos e um algoritmo de segmentação para lidar com a esparsidade dos dados WES. O algoritmo de normalização foi desenvolvido a partir de um estudo dos vieses sistemáticos do conjunto de dados sequenciados. O EXCAVATOR consiste dos processos de normalização e segmentação, mais o processo que classifica de regiões pelo seu número de cópias.

Normalização: Para estudar os conjuntos de dados, foi utilizada a medida de média de contagem de reads por exon, ou *exon mean read count* (EMRC):

$$EMRC_e = \frac{RC_e}{L_e} \quad \begin{cases} RC_e = \text{ número de reads alinhados à região } e \\ L_e = \text{ comprimento da região em pares de bases} \end{cases}$$

As propriedades estatísticas do EMRC foram estudadas usando dados WES de oito indivíduos sequenciados pelo 1000 Genomes Project. Foi analisada sua relação com três fontes de viés: o percentual de conteúdo GC, a mapeabilidade genômica e o tamanho dos exons. Foi observado que há uma forte correlação com o conteúdo GC, uma mapeabilidade melhor reduz a

variância do EMRC, e que o valor cresce linearmente com o tamanho dos exons até 150 bp, onde ele fica estável.

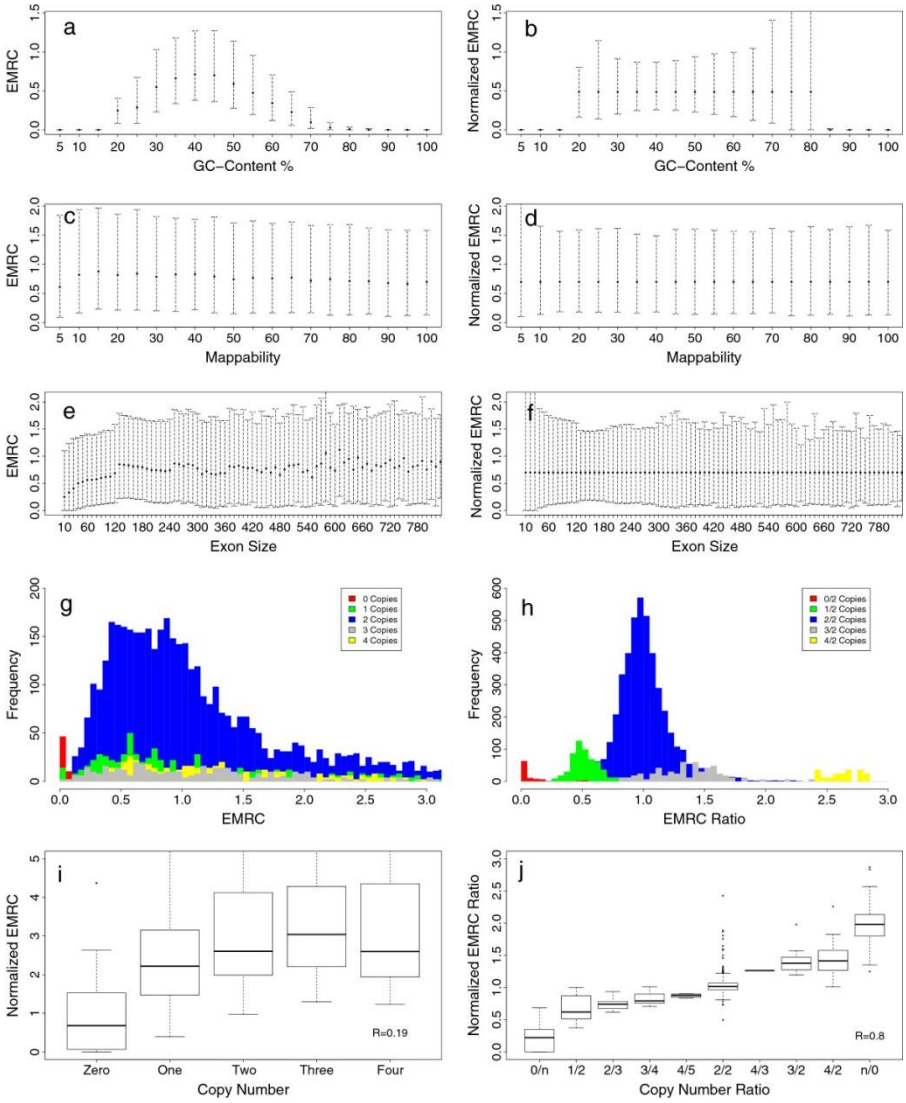

Figura 11. Resultados da análise do EMRC.

(a), (c), (e) Correlação entre EMRC com as três fontes de viés.

(b), (d), (f) mostram o efeito da normalização. A borda superior da linha é o 90° percentil e a inferior é o 10° percentil dos EMRCs.

(g), (h), (i), (j) demonstram o potencial de previsão do EMRC usando uma amostra e a razão do EMRC entre duas amostras. Devido ao bom resultado, nesse trabalho foi utilizada a razão entre amostras de teste e controle para detectar CNVs.

Após normalizar o EMRC e calcular a razão entre amostras de teste e controle, o sinal obtido é similar ao de contagem de reads. A diferença entre eles é que RCs são estimados para regiões disjuntas contíguas de tamanho fixo (janelas), enquanto EMRCs são calculados para regiões de tamanhos e distâncias variáveis. A distância entre exons varia de alguns pares de bases a 100 kb, com uma mediana de 1500 bp. Para tratar essa diferença, o algoritmo de segmentação *Shifting Level Model* (SLM) foi estendido para incluir a distância entre exons consecutivos. O resultado é chamado *Heterogeneous Shifting Level Model* (HSLM), em que a probabilidade de mudar de estado aumenta conforme a distância entre exons aumenta.

Para classificar as regiões em 5 estados quanto ao número de cópias (2-deleção, 1-deleção, normal, 1-duplicação e duplicação múltipla) foi utilizado o algoritmo FastCall, desenvolvido pelos próprios autores para detecção de variações em dados de aCGH [28].

5.2.Teste de performance

Para testar a performance do algoritmo, foram gerados cromossomos sintéticos a partir das 8 amostras, onde cada gene sintético tem um número aleatório de exons. Foram testadas várias combinações de parâmetros de geração, cada combinação dando origem a 1000 cromossomos. Para medir a precisão na detecção dos *breakpoints*, foi calculada a curva ROC (Receiver Operating Characteristic) e os resultados foram comparados com os do algoritmo *Circular Binary Segmentation* (CBS) utilizado em outros pacotes como ExomeCNV e VarScan2.

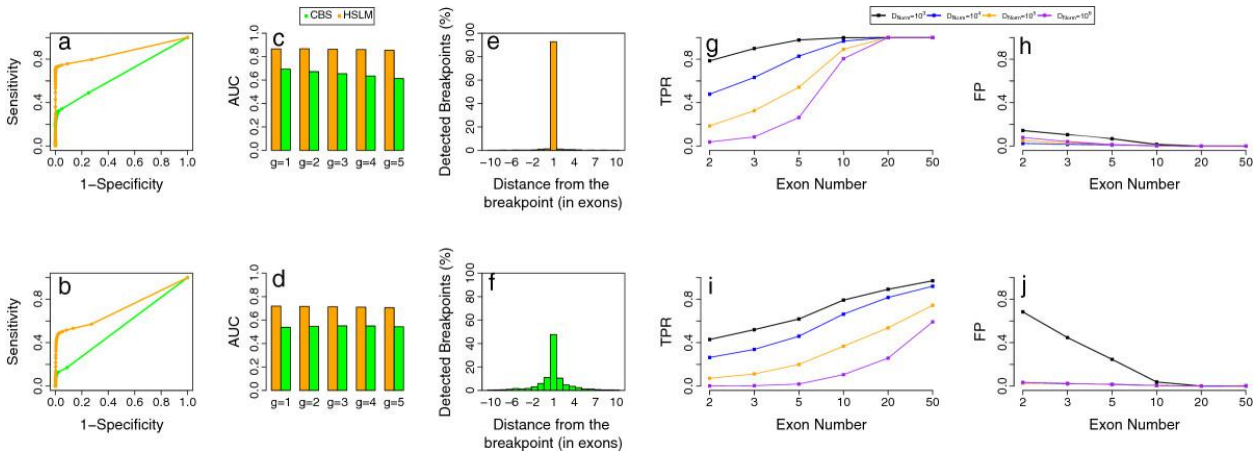

Figura 12. Avaliação de performance do HSLM em cromossomos sintéticos. (a), (b) Curvas ROC comparando sensibilidade e especificidade na detecção de uma cópia e três cópias. (c), (d) Comparação da área abaixo da curva quando se varia o parâmetro g, com uma cópia e três cópias. (e), (f) Precisão na detecção dos breakpoints. (g), (h), (i), (j) Taxa de verdadeiros positivos (TPR) e Falsos positivos (FP) para diferentes valores de D_{Norm} e número de exons da região. (g) e (h) são com regiões com número de cópias 1, (i) e (j) são com regiões com número de cópias 3.

Os resultados mostram melhor sensibilidade e especificidade em todos os casos. O aumento da quantidade de regiões alteradas reduz o desempenho do CBS, mas não do HSLM. Os testes com número de cópias 1 e 3 também mostram que é mais difícil classificar números de cópias maiores, devido a propriedades matemáticas do sinal. O HSLM detectou corretamente 94% dos breakpoints, enquanto CBS detectou 50%. Uma maior distância entre genes adjacentes também aumenta a sensibilidade do HSLM.

5.3. Estudo de populações

Para se testar o potencial do EXCAVATOR no estudo de populações, foram realizados testes comparativos com ExomeCNV, CoNIFER e XHMM usando dados WES de 20 indivíduos. Para CoNIFER e XHMM, que usam métodos de SVD/PCA para normalização, foram adicionadas 80 amostras extras. CoNIFER detectou 9 CNVs, XHMM detectou 55, ExomeCNV detectou 1791 e EXCAVATOR detectou 101. De acordo com os resultados do artigo por R. Tan visto anteriormente [18], XHMM e CoNIFER tiveram alto percentual de detecções de CNVs raros, enquanto maior parte dos detectados pelo EXCAVATOR e ExomeCNV são comuns.

As regiões marcadas pelos algoritmos foram validadas pela sobreposição com os CNVs conhecidos presentes no *Database of Genomic Variants* (DGV) e no *dbVAR* da NCBI, tratando variações comuns e raras separadamente. Dois critérios de igualdade diferentes foram usados: sobreposição de 10% e 50%. Nos testes de variações gerais e somente variações comuns, os melhores resultados foram obtidos pelo EXCAVATOR e CoNIFER, seguido por XHMM e ExomeCNV. No teste somente com variações raras, CoNIFER teve o melhor resultado, seguido por EXCAVATOR, XHMM e ExomeCNV.

Os CNVs também foram comparados com os obtidos por McCarroll e Conrad usando microarrays, usando medidas de *precisão* (detecções corretas/detecções totais) e *sensibilidade* (detecções corretas/total verdadeiro). ExomeCNV obteve alta sensibilidade, porém baixa precisão em todos os casos, devido ao alto número de detecções. XHMM e CoNIFER obtiveram baixa sensibilidade, devido a detectarem principalmente CNVs raros. CoNIFER obteve boa precisão nos dados de McCarroll, mas não no de Conrad. O EXCAVATOR manteve um bom equilíbrio entre os valores, só se mostrando inferior quando restrito aos CNVs raros de McCarroll.

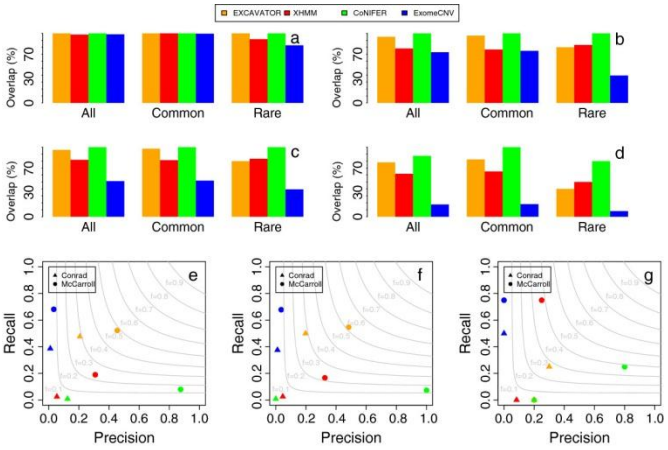

Figura 13. Comparação do EXCAVATOR, XHMM, CoNIFER e ExomeCNV. (a), (b) Sobreposição dos eventos detectados com os do DGV. (c), (d) Sobreposição dos eventos detectados com os do dbVar. (a), (c) usam critério de 10%. (b), (d) critério de 50%. (e), (f), (g) Mapas de precisão e recall para CNVs quaisquer, comuns e raros.

5.4.Teste com melanoma

Para avaliar o potencial do EXCAVATOR no estudo de câncer, ele foi testado com seis células de melanoma e seis células saudáveis como controle. O objetivo é testar a capacidade de detecção de alterações de diversos tipos e tamanhos, que são características de câncer. As amostras também foram analisadas usando SNP arrays, e foi calculada a correlação entre os resultados para diferentes tamanhos de regiões. Para regiões acima de 1 Mb, foi obtido um índice alto de 0.85, porém este valor decresceu consideravelmente para regiões menores, pois as distribuições de sondas SNP e exons tornaram-se mais inconsistentes, reduzindo a similaridade entre os dados analisados. Para regiões menores que 100 kb, o índice foi -0.02.

Embora ambos os métodos tenham retornado resultados indicativos de melanoma, em alguns cromossomos eles não foram tão consistentes. Foi observado também que os dados WES permitiram detectar um escopo maior de valores de número de cópias que os dados de SNP arrays, o que é de alta relevância nessa área de estudo. De fato, o EXCAVATOR detectou algumas alterações de impacto conhecido que não foram percebidas por SNP arrays.

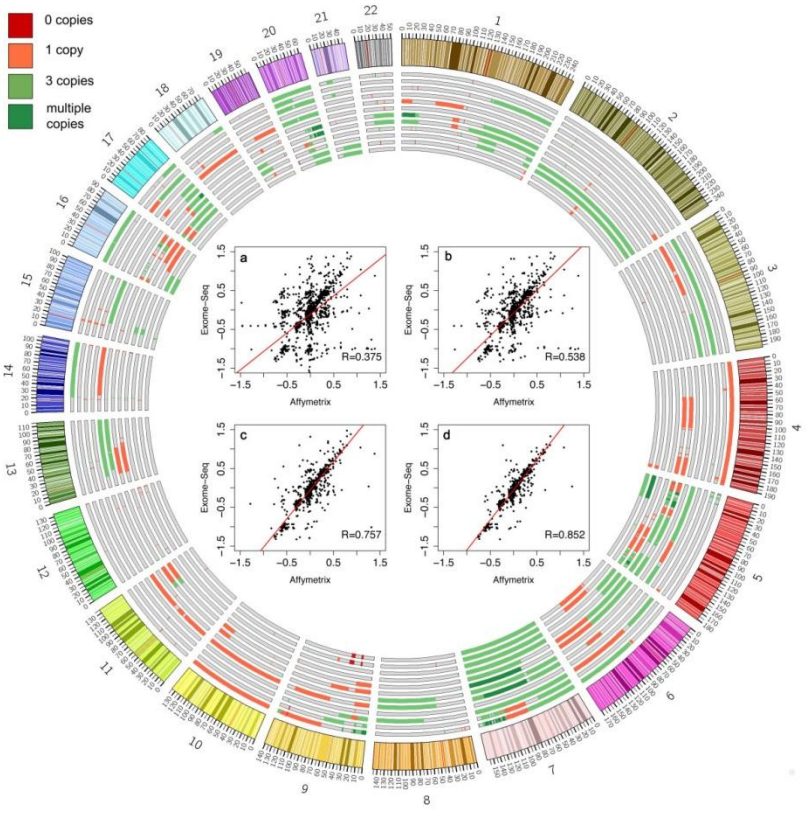

Figura 14. Alterações detectadas pelo EXCAVATOR e SNP arrays nos dados de melanoma. Em cada cromossomo, as amostras são alinhadas verticalmente, com duas linhas para cada, indicando os resultados de WES e SNP arrays. No centro, as correlações para todas as regiões (a), regiões maiores que 100 kb (b), regiões maiores que 500 kb (c) e regiões maiores que 1 Mb (d).

5.5. Teste com deficiência intelectual

Para demonstrar a capacidade do EXCAVATOR de detectar alterações envolvidas em retardo mental, foi realizado o sequenciamento do exoma de dois irmãos com a deficiência, e usado como controle os dados WES de um indivíduo saudável já sequenciado. Em um dos irmãos foram reconhecidos 29 CNVs, dos quais 22 possuem sobreposição com variações anotadas no DGV. No outro foram reconhecidos 24, dos quais 17 estão no DGV. Os tamanhos variam de 3 kb a 1 Mb. Dentre as regiões não presentes no DGV, há uma grande deleção compartilhada por ambos que foi confirmada usando SNP arrays. Uma pesquisa no banco de dados ISCA indica que problemas nessa região são frequentemente associados com deficiências deste tipo. Os outros algoritmos também foram usados, sendo que XHMM e CoNIFER não encontraram CNVs devido ao pequeno conjunto de amostras, enquanto ExomeCNV encontrou 342 Mb de regiões alteradas.

Por fim, para se testar os efeitos do tamanho dos reads e algoritmos de alinhamento na performance do EXCAVATOR, foram analisados dados WES de 4 indivíduos. Foram usados reads de extremidades pareadas com tamanhos 50 bp e 75 bp e três ferramentas de alinhamento: BWA, Bowtie2, SOAP2. Os reads foram mapeados para o genoma de referência humano (hg19) e pós-processados com algumas ferramentas. Para cada combinação de tamanho e ferramenta, foi usado o EXCAVATOR em três amostras com a quarta servindo de controle. Foi observada uma correlação entre os parâmetros e o número de reads mapeados, porém a capacidade de previsão do EMRC não foi afetada, indicando que a escolha do mapeador é pouco relevante para o desempenho do algoritmo.

Em conclusão, foi apresentado um novo método de detecção de CNVs que se aproveita do estudo de propriedades estatísticas e vieses dos conjuntos de dados. Ele obteve bons resultados comparado com outros algoritmos, mas o fator de destaque foi a riqueza de informações providas pelos experimentos. Foram usadas diversas formas de validação, e a distinção entre variações raras e comuns permite um entendimento mais minuncioso dos resultados. Foi demonstrada a aplicabilidade do EXCAVATOR em diferentes áreas de estudo de grande interesse: populações, câncer e deficiência intelectual. Assim, há um bom ponto de partida para futuras pesquisas envolvendo este pacote.

6. Considerações sobre os artigos

Em retrospectiva, os testes envolvendo o modelo m-HMM foram consideravelmente simples, e podem não fornecer informações o bastante para dar uma noção de sua aplicabilidade, sendo necessários mais testes como os presentes nos outros estudos. As informações referentes à detecção de diferentes tamanhos de CNVs comparadas a outros métodos são interessantes por se referirem à sensibilidade em casos particulares, porém não foi incluída a especificidade. A variação da performance segundo a cobertura é esperada neste problema, porém é importante o desenvolvimento de técnicas que façam melhor uso dos dados fornecidos por técnicas de sequenciamento mais avançadas. A capacidade de detecção de CNVs sem o uso de amostras adicionais além da referência e alvo também é um fator atrativo, pois o algoritmo é utilizável em estudos com recursos limitados.

O trabalho que realiza a comparação entre ExomeDepth, CONTRA, XHMM e CoNIFER realiza testes variados, especialmente na forma de validar chamadas. Ele também examina a aplicabilidade de métodos que usam dados WGS para validação no lugar de aCGH e SNP arrays, o que é interessante devido à baixa disponibilidade destes métodos experimentais. No entanto, a distinção entre diferentes tipos de CNVs não foi mantida durante as validações, o que reduziu o potencial dos dados obtidos. Mesmo assim, foi realizada uma boa análise comparativa dos algoritmos, e levantado outro ponto de interesse para estudos futuros: o impacto da ferramenta de mapeamento no desempenho dos algoritmos.

Combinando os pontos fortes de ambos os artigos, o EXCAVATOR foi comparado com outros algoritmos em termos de CNVs raros e comuns, o que deu uma boa perspectiva sobre seu uso. Sua aplicabilidade em casos diversos foi demonstrada através de testes minuciosos realizados com melanoma, deficiência intelectual e estudo de populações, o que deve incentivar a realização de novos experimentos.

Vale lembrar que, apesar do foco no estudo de abordagens individuais e suas diferenças, há potencial na combinação de abordagens complementares. Estes experimentos também foram ser realizados com os algoritmos em sua forma padrão, sendo que na prática é possível extrair mais performance através do ajuste de seus parâmetros para se adequar a casos específicos. Tais ganhos adicionais não foram explorados neste estudo.

CONCLUSÃO

Com o crescimento do conhecimento sobre o problema e aproveitando-se de avanços tecnológicos, trabalhos recentes desenvolvem uma variedade de ferramentas para detecção de CNVs, buscando novas formas de melhorar abordagens já conhecidas. Não é buscada uma solução universal, devido à natureza difícil dos problemas de bioinformática, que trabalham com grandes volumes de dados altamente variáveis, e por vezes incertos. As técnicas são testadas extensivamente a fim de se determinar em que situações seu uso é adequado e que fatores as prejudicam, o que são informações vitais tanto para pesquisadores quanto para usuários.

A detecção de variações pequenas continua sendo um desafio, assim como o uso de dados com baixa cobertura. Mesmo se aproveitando de avanços tecnológicos, técnicas de RD estão sujeitas a inúmeros vieses que restringem sua aplicabilidade. A influência do conteúdo GC na contagem de reads já é bastante estudada e tratada de diversas formas, enquanto que regiões com mapeamento ambíguo continuam sendo problemáticas. Técnicas de sequenciamento de terceira geração prometem dados com maior cobertura e comprimento de reads, que melhoram a qualidade do mapeamento e elevam o potencial das técnicas. Ainda assim, é essencial que seja feita uma boa escolha da ferramenta a ser utilizada através do estudo prévio de sua aplicabilidade, podendo inclusive ser desejável a combinação de abordagens distintas.

Os avanços desta área se dão através da combinação de avanços tecnológicos e algorítmicos, possíveis somente com esforço colaborativo nas áreas de pesquisa e indústria. O entendimento do papel dessas variações em doenças humanas, com seu potencial uso em diagnósticos e na produção de medicamentos personalizados, servem como incentivo para o contínuo crescimento da área.

GLOSSÁRIO

Array comparative genomic hybridization: Hibridização genômica comparativa com array

Assembly: Montagem

Batch effects: Efeitos de lote

Célula diplóide: Célula que possui duas cópias homólogas de cada cromossomo

Célula haplóide: Célula que possui um cromossomo de cada tipo

Copy number variation: Variação do número de cópias

Empirical false negative rate: Taxa empírica de falsos negativos

Empirical false positive rate: Taxa empírica de falsos positivos

Expectation maximization: Maximização de esperança

Exon mean read count: Média das contagens de reads do exon

Fluorescence *in situ* hybridization: Hibridização *no local* fluorescente

GC content: conteúdo GC

Gene: Região do genoma responsável por uma característica do organismo, sendo uma unidade hereditária

Heterozigose: Um organismo é heterozigoto em um locus quando ele possui alelos diferentes de um certo gene nos seus dois cromossomos homólogos

Hidden Markov model: Modelo oculto de Markov

K-means clustering: Clusterização K-médias

Locus: Uma localização específica em um cromossomo, onde pode estar presente um certo gene ou sequência de DNA

Next generation sequencing: Sequenciamento de nova geração

Paired end mapping: Mapeamento de fins pareados

Pyrosequencing: Pirosequenciamento

Read count: Contagem de reads

Read depth: Profundeza de reads

Sensitivity: Sensibilidade

Single-molecule sequencing with exonuclease: Sequenciamento de molécula única com exonuclease

Single nucleotide polymorphism: Polimorfismo de nucleotídeo único

Site: sítio

Sliding window: Janela deslizante

Specificity: Especificidade

Split reads: Reads divididos

Structural variation: Variação estrutural

Whole exome sequencing: Sequenciamento de exoma inteiro

Whole genome sequencing: Sequenciamento de genoma inteiro

REFERÊNCIAS

[1] "What is DNA?," Genetics Home Reference, [Online]. Available: http://ghr.nlm.nih.gov/handbook/basics/dna. [Accessed 10 Junho 2014].

[2] Elgar G and Vavouri T, "Tuning in to the signals: noncoding sequence conservation in vertebrate genomes," *Trends in Genetics,* pp. 344-352, 2008.

[3] "What is a genome?," Genetics Home Reference, [Online]. Available: http://ghr.nlm.nih.gov/handbook/hgp/genome. [Accessed 10 Junho 2014].

[4] "Whole genome," Ensembl, [Online]. Available: http://www.ensembl.org/Homo_sapiens/Location/Genome. [Accessed 10 Junho 2014].

[5] "What are gene mutations and how do mutations occur?," Genetics Home Reference, [Online]. Available: http://ghr.nlm.nih.gov/handbook/mutationsanddisorders/genemutation. [Accessed 10 Junho 2014].

[6] "DNA Sequencing," National Human Genome Research Institute, [Online]. Available: http://www.genome.gov/10001177. [Accessed 23 Junho 2014].

[7] Lilian T. C. França, Emanuel Carrilho and Tarso B. L., "A review of DNA sequencing techniques," *Quarterly Reviews of Biophysics 35, issue 2,* pp. 169-200, 2002.

[8] S. Clancy, "DNA transcription," Nature Education 1(1):41, 2008. [Online]. Available: http://www.nature.com/scitable/topicpage/dna-transcription-426. [Accessed 23 Junho 2014].

[9] "EXCAVATOR: detecting copy number variants from whole-exome sequencing data," *Genome Biology, Vol. 14, Iss. 10,* 2013.

[10] F. Lars, A. R. Carson and S. W. Scherer, "Structural variation in the human genome," *Nature Reviews Genetics 7 (2),* pp. 85-97, Fevereiro 2006.

[11] S. Clancy, "Copy number variation," Nature Education 1(1):95, 2008. [Online]. Available: http://www.nature.com/scitable/topicpage/copy-number-variation-445. [Accessed 5 Julho 2014].

[12] Redon, R. et al., "Global variation in copy number in the human genome," *Nature 444,* pp. 444-454, 2006.

[13] E. E. Eichler, "Copy Number Variation and Human Disease," Nature Education, [Online]. Available: http://www.nature.com/scitable/topicpage/copy-number-variation-and-human-disease-741737.

[Accessed 3 Agosto 2014].

[14] C. O'Connor, "Fluorescence in situ hybridization (FISH)," Nature Education 1(1):171, 2008. [Online]. Available: http://www.nature.com/scitable/topicpage/fluorescence-in-situ-hybridization-fish-327. [Accessed 6 Julho 2014].

[15] A. Theisen, "Microarray-based comparative genomic hybridization (aCGH)," Nature Education 1(1):45, 2008. [Online]. Available: http://www.nature.com/scitable/topicpage/microarray-based-comparative-genomic-hybridization-acgh-45432. [Accessed 6 Julho 2014].

[16] J. Duan, J.-G. Zhang, H.-W. Deng and Y.-P. Wang, "Comparative Studies of Copy Number Variation Detection Methods for Next-Generation Sequencing Technologies," Março 2013. [Online]. Available: http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0059128. [Accessed 6 Julho 2014].

[17] "Next-generation sequencing (definition)," [Online]. Available: http://www.nature.com/subjects/next-generation-sequencing. [Accessed 6 Julho 2014].

[18] R. Tan, Y. Wang, S. E. Kleinstein, Y. Liu, X. Zhu, H. Guo, Q. Jiang, A. S. Allen and M. Zhu, "An Evaluation of Copy Number Variation Detection Tools," *Human Mutation, 35,* pp. 899-907, Março 2014.

[19] T. Derrien, J. Estellé, S. M. Sola, D. G. Knowles, E. Raineri and et al, "Fast Computation and Applications of Genome Mappability," *PLoS ONE 7(1): e30377. doi:10.1371/journal.pone.0030377,* 19 Janeiro 2012.

[20] M. Zhao, Q. Wang, Q. Wang, P. Jia and Z. Zhao, "Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives," *BMC Bioinformatics, 14(Suppl 11):S1,* 13 Setembro 2013.

[21] J. T. Leek, R. B. Scharpf, H. C. Bravo, D. Simcha, B. Langmead, W. E. Johnson, D. Geman, K. Baggerly and R. A. Irizarry, "Tackling the widespread and critical impact of batch effects in high-throughput data," *Nature Reviews Genetics 11,* pp. 733-739, Outubro 2010.

[22] "Paired-End Sequencing," illumina, [Online]. Available: http://www.illumina.com/technology/next-generation-sequencing/paired-end-sequencing_assay.ilmn. [Accessed 3 Agosto 2014].

[23] H. Wang, D. Nettleton and K. Ying, "Copy number variation detection using next generation sequencing read counts," *BMC Bioinformatics,* 14 Abril 2014.

[24] L. E. Baum and T. Petrie, "Statistical Inference for Probabilistic Functions of Finite State Markov Chains," *The Annals of Mathematical Statistics 37, no. 6,* pp. 1554-1563, 1996.

[25] L. E. Baum and J. A. Eagon, "An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology," *Bulletin of the American Mathematical Society 73, no.3,* pp. 360-363, 1967.

[26] L. E. Baum, T. Petrie, G. Soules and N. Weiss, "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains," *The Annals of Mathematical Statistics 41, no.1,* pp. 164-171, 1970.

[27] L. E. Baum, "An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes," *Inequalities, Vol. 3,* pp. 1-8, 1972.

[28] M. Benelli, G. Marseglia, G. Nannetti, F. Zara, R. Paravidino, F. D. Bricarelli, F. Torricelli and A. Magi, "A very fast and accurate method for calling aberrations in array-CGH data.," *Biostatistics 2010,11,* pp. 515-518, Julho 2010.