



MURILO GUIMARÃES BORGES

APLICAÇÃO DE PROTOCOLOS E MÉTODOS EM
BIOINFORMÁTICA PARA ANÁLISE DE SEQUENCIAMENTO DE
EXOMAS HUMANOS

*APPLICATION OF BIOINFORMATICS PROTOCOLS AND METHODS
FOR HUMAN EXOME SEQUENCING ANALYSIS*

CAMPINAS

2015



UNIVERSIDADE ESTADUAL DE CAMPINAS

Faculdade de Ciências Médicas

MURILO GUIMARÃES BORGES

APLICAÇÃO DE PROTOCOLOS E MÉTODOS EM BIOINFORMÁTICA PARA
ANÁLISE DE SEQUENCIAMENTO DE EXOMAS HUMANOS

*APPLICATION OF BIOINFORMATICS PROTOCOLS AND METHODS FOR
HUMAN EXOME SEQUENCING ANALYSIS*

Dissertação apresentada à Faculdade de Ciências Médicas da
Universidade Estadual de Campinas como parte dos requisitos
exigidos para a obtenção do título de Mestre em Ciências.

*Dissertation submitted to School of Medical Sciences of the
University of Campinas as part of the requirements for obtaining the
title of Master in Sciences.*

ORIENTADORA: ISCIA TERESINHA LOPES CENDES

ESTE EXEMPLAR CORRESPONDE À VERSÃO
FINAL DA DISSERTAÇÃO DEFENDIDA PELO
ALUNO MURILO GUIMARÃES BORGES, E ORIENTADO PELA
PROF.^a DR.^a ISCIA TERESINHA LOPES CENDES.

CAMPINAS

2015

iii

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca da Faculdade de Ciências Médicas
Maristella Soares dos Santos - CRB 8/8402

B644a Borges, Murilo Guimarães, 1989-
Aplicação de protocolos e métodos em bioinformática para análise de sequenciamento de exomas humanos / Murilo Guimarães Borges. – Campinas, SP : [s.n.], 2015.

Orientador: Iscia Teresinha Lopes Cendes.
Dissertação (mestrado) – Universidade Estadual de Campinas, Faculdade de Ciências Médicas.

1. Exoma. 2. Genoma. 3. Biologia computacional. I. Lopes-Cendes, Iscia Teresinha, 1964-. II. Universidade Estadual de Campinas. Faculdade de Ciências Médicas. III. Título.

Informações para Biblioteca Digital

Título em outro idioma: Application of bioinformatics protocols and methods for human exome sequencing analysis

Palavras-chave em inglês:

Exome

Genome

Computational biology

Área de concentração: Fisiopatologia Médica

Titulação: Mestre em Ciências

Banca examinadora:

Ischia Teresinha Lopes Cendes [Orientador]

Marcondes Cavalcante França Júnior

Wilson Araújo da Silva Júnior

Data de defesa: 26-06-2015

Programa de Pós-Graduação: Fisiopatologia Médica

BANCA EXAMINADORA DA DEFESA DE MESTRADO

MURILO GUIMARÃES BORGES

Orientador (a) PROF(A). DR(A). ISCIA TERESINHA LOPES CENDES

MEMBROS:

1. PROF(A). DR(A). ISCIA TERESINHA LOPES CENDES



2. PROF(A). DR(A). MARCONDES CAVALCANTE FRANCA JUNIOR



3. PROF(A). DR(A). WILSON ARAÚJO DA SILVA JÚNIOR



Programa de Pós-Graduação em Fisiopatologia Médica da Faculdade de Ciências Médicas da Universidade Estadual de Campinas

Data: 26 de junho de 2015

RESUMO

Introdução: Os avanços técnicos em sequenciamento alcançados em menos de uma década, atrelados ao desenvolvimento e barateamento do sequenciamento de alto desempenho, oferecem-nos a possibilidade de aplicação dessas tecnologias na medicina genômica. Nesse contexto, surge o sequenciamento do exoma humano, constituído das regiões codificantes do genoma, menor que 2% de sua totalidade. O sequenciamento do exoma (WES) se estabelece hoje como uma ferramenta custo-efetiva com a finalidade de identificar variantes de sequência relacionadas a várias doenças humanas. A análise através da bioinformática é essencial para lidar com o alto volume de dados gerados e realizar a ligação entre o experimento biológico e os dados obtidos. **Objetivo:** Aplicar e avaliar protocolos e aplicações disponíveis na análise dos dados gerados pelo sequenciamento de exomas humanos, bem como aplicar e aperfeiçoar protocolos e aplicações disponíveis para prever variantes como potencialmente patológicas a partir de dados gerados pelo sequenciamento de exomas humanos. **Materiais e métodos:** Foram utilizadas as seguintes ferramentas: FastQC, Rqc, BWA, Picard, GATK e VEP. Estas foram então aplicadas às sequências do exoma humano possibilitando a identificação de variações nos perfis de qualidade das sequências, realinhamento local ao redor de inserções e deleções, recalibração da qualidade e posterior chamada das variantes potencialmente envolvidas nos fenótipos em estudo. No intuito de avaliar se a cobertura no exoma sofre variações mediante diferenças técnicas e étnicas, selecionamos amostras do Projeto 1000 Genomas. **Resultados:** A aplicação de nosso protocolo em 27 amostras WES resultou em gráficos de

controle de qualidade pré e pós-alinhamento, que nos permitiram avaliar de modo global os perfis de qualidade destas sequências; realinhamento ao redor de inserções e deleções que ocorreu em mais de 15% da definição do exoma, realinhando mais de 79% das sequências; recalibração da qualidade que nos permitiu minimizar sua variação por ciclo da reação. Das sequências empregadas, 72% foram pareadas ao genoma, contudo 46% se estendem para fora da definição do exoma, com uma cobertura média de 59x para o exoma estendido e 66x para o exoma restrito. Temos que a cobertura para WES possui uma tendência a variar de acordo com a metodologia de captura empregada e ao grupo étnico de onde as amostras foram obtidas. **Conclusão:** A aplicação de um *workflow* para interrogação de variantes que considera a qualidade das sequências fornecidas pelo sequenciador, o alinhamento contra o genoma, realinhamento ao redor de regiões sabidamente conhecidas como portadoras de variações, recalibração da qualidade e anotação permitiu identificar variantes de sequência. Além disso, através da cobertura obtida pelo sequenciamento do exoma foi possível perceber diferenças técnicas e populacionais, refletindo que a complexidade do genoma pode interferir na reação de captura das sequências, influenciando na efetividade da técnica empregada.

Palavras-chave: exoma, genoma, bioinformática.

ABSTRACT

Background: The technical advances in sequencing made in less than a decade associated with the development and low costs of high throughput sequencing techniques allow their application in genomic medicine. Therefore, Whole Exome Sequencing (WES), which corresponds to less than 2% of the entire genome, emerges as a cost-effective tool that aims to identify variants related to human diseases. Bioinformatics is fundamental to process the big volume of data and link the obtained results with the biology. **Objective:** We aim to apply and evaluate protocols and applications designed for WES data analysis on human subjects. We also intend to apply and enhance protocols and applications designed to predict variants as potentially pathological from WES data. **Materials and Methods:** We used the following tools: FastQC, Rqc, BWA, Picard, GATK e VEP. We applied them to exome data, determining variation in quality profiles, local realignment, quality recalibration and variant calls. We also evaluated whether or not technical and population differences affect the depth profiles of samples from the 1000 Genomes Project. **Results:** We applied our protocol on 27 samples, resulting in pre and post-alignment quality control charts. Local realignment took place at more than 15% of the exome definition, extending to more than 79% of sequences. Quality recalibration minimized per cycle variation. In total, 72% of the sequences were paired against the genome, nevertheless 46% extended off-target. The mean coverage was 59X for the exome. We also detected that depth tends to vary based on technical and population differences between samples. **Conclusion:** We applied

a variant-calling workflow that accounts for sequence quality, the alignment against the genome, local realignment, quality recalibration and annotation. In addition, we concluded that depth depends on technical and population differences, showing that genomic complexity may interfere with the capturing phase, affecting downstream analyses.

Keywords: exome, genome, bioinformatics.

SUMÁRIO

RESUMO	vii
ABSTRACT	ix
DEDICATÓRIA	xiii
AGRADECIMENTOS.....	xv
1. INTRODUÇÃO	1
Sequenciamento de alto desempenho.....	1
Captura do exoma.....	1
Aplicações em medicina genômica	5
Fenótipos estudados.....	6
Análises de bioinformática	7
2. OBJETIVOS	10
3. MATERIAIS E MÉTODOS	11
Perfil das amostras.....	11
Delineamento de experimentos.....	13
Análises de bioinformática	16
<i>Controle de qualidade pré alinhamento</i>	16
<i>Alinhamento das sequências</i>	16
<i>Processamento pós-alinhamento</i>	17
<i>Realinhamento local</i>	19
<i>Recalibração da qualidade</i>	20
<i>Controle de qualidade pós alinhamento</i>	20
<i>Descoberta de variantes</i>	21
<i>Anotação das variantes</i>	22
Avaliação da cobertura no sequenciamento do exoma	22

5. RESULTADOS.....	24
Workflow para análise de dados de sequenciamento de alto desempenho.....	24
Controle de qualidade pré-alinhamento propicia identificar comportamentos anômalos nas bibliotecas sequenciadas.....	26
Controle de qualidade pós-alinhamento propicia um diálogo entre o experimento biológico e os resultados “in silico”	29
Anotação das variantes encontradas adiciona informações que possibilitam filtragem posterior	34
Whole exome sequencing depth of coverage is susceptible to technical and population differences	38
7. DISCUSSÃO	53
8. CONCLUSÃO	58
REFERÊNCIAS	59
ANEXOS.....	68

DEDICATÓRIA

A todos os sonhadores,
que mesmo acordados,
fazem deste mundo
um lugar melhor.

Aos meus queridos
e todos os que ainda estou
por amar.

AGRADECIMENTOS

*Sede bendito, Senhor Deus de nossos pais,
digno de louvor e de eterna glória!*

*Que seja bendito o vosso santo nome glorioso,
digno do mais alto louvor e de eterna exaltação!*

*Sede bendito no templo de vossa glória santa,
digno do mais alto louvor e de eterna glória!*

*Sede bendito por penetrardes com o olhar os abismos,
e por estardes sentado sobre os querubins,
digno do mais alto louvor e de eterna exaltação!*

*Glorificai o Senhor porque ele é bom,
porque eterna é a sua misericórdia.*

Homens piedosos,

bendizei o Senhor,

*Deus dos deuses, louvai-o,
glorificai-o, porque é eterna a sua misericórdia!*

Daniel 3

Três jovens clamam a Deus em um hino de louvor. Em um contexto inusitado, sim, mas completamente centrados. Como gostaria de me unir a eles em um agradecimento tão sincero e inteiro: completo! Mas ainda me falta muito! Muito ainda devo caminhar, muitos ainda devo amar, muitas coisas ainda me pesam. Mas temos que começar! Dizem por aí que um homem só pode ser feliz se é agradecido pela vida que (pensa que) tem. De verdade, felicidade e gratidão só podem vir juntas, de mãos dadas (por que não?) no destino da vida... não faz sentido viver uma vida sem sabor, tão pouco em excesso.

Agradeço imensamente a Deus por me guiarem e preservarem até aqui. Agradeço por cada vez mais me mostrarem quem sou eu por inteiro.

Agradeço a meus pais que desde muito me quiseram para mostrar ao mundo um sinal concreto de seu amor. Agradeço por eles terem dado tudo de si para que eu me moldasse e me formasse como sou hoje. Espero poder também eu um dia fazer pelos meus filhos e pelos que amo o que fizeram por mim.

Agradeço a minha pequena irmã, uma fagulha que vi crescer até se tornar uma estrela por si só. Obrigado pela sua luz e seu calor! Agradeço ao meu grande irmão e sua rosa sempre tão doce, cheios de amor que transbordam e inundam toda a terra! Agradeço a minha família venial e a família que escolhi como minha, são vocês quem me lembram quem eu sou e o que almejo ser.

Agradeço a minha venerável futura esposa, tão esperada e amada! Faltam adjetivos superlativos para lhe oferecer! Te amo! E como posso a cada instante te

amar mais? Obrigado por derramar amor por onde passa e onde toca. Amor que cura o corpo e acalma a alma.

Agradeço aos meus formadores e instrutores por me ajudarem na difícil tarefa de moldar um homem. Quem me dera um dia me assemelhar um pouco a vocês.

Agradeço aos grandes e fiéis amigos que fiz, sem vocês tudo seria menos divertido!

Como não agradecer aos que intercederam por mim? Agradeço imensamente por depositarem neste pobre vassalo suas esperanças.

Agradeço pelas condições de permanecer na pesquisa, pela infraestrutura da Universidade Estadual de Campinas, da Faculdade de Ciências Médicas e pelo financiamento da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior.

Agradeço sobretudo pelo dia que está por vir depois do ocaso... Que simplesmente pensar nessa luz e nesse calor já me abraça o coração, me torne o que desejo ser: luz do mundo, quase uma estrela!

*Podemos percorrer
muitos caminhos,
e ficar sem futuro
cheio de metros
na planta dos pés.*

*Podemos dar
um passo,
e antecipar nele
o gozo da meta.*

*Podemos olhar
muitas paisagens,
e ficar vazios
cheios de imagens
na superfície da cor.*

*Podemos contemplar
um só horizonte
e ver aparecer nele
a plenitude do infinito.*

Benjamín González Buelta, SJ

1. INTRODUÇÃO

Sequenciamento de alto desempenho

O surgimento de métodos que propiciaram o sequenciamento de alto desempenho de moléculas de DNA [1, 2], associado ao seu desenvolvimento, consolidou o sequenciamento de nova geração, *NGS* (sigla em inglês para *Next Generation Sequencing*), como um método para identificação de todos os tipos de variações genéticas, já que a resolução alcançada é dos menores constituintes do genoma: as bases nitrogenadas. Além disso, o desenvolvimento destas plataformas de sequenciamento se encontra em franco aprimoramento, tanto nas já consolidadas tecnologias policromáticas, quanto nas novas tecnologias monocromáticas ou acromáticas. Deste modo, formas tradicionais de mapeamento genético, como por cariótipo [3], análise de ligação [4], mapeamento de homozigocidade [5] e análises de *CNV* (sigla em inglês para *Copy Number Variation*) [6], bem como o sequenciamento de um gene candidato por capilaridade ou estudos com *arrays*, gradativamente vêm cedendo lugar ao sequenciamento de partes ou da totalidade do genoma.

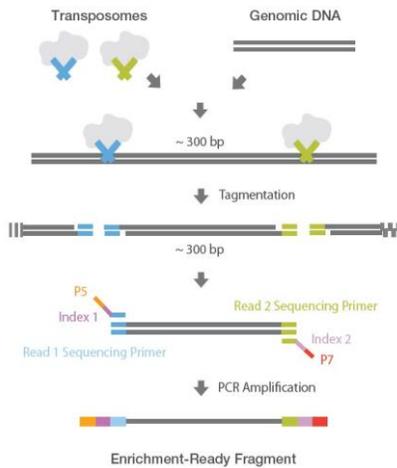
Captura do exoma

Neste contexto, surge uma nova abordagem de sequenciamento, não do genoma completo (*WGS*, do inglês *Whole Genome Sequencing*), mas de regiões codificantes ou de interesse, correspondente aos exons (*WES*, do inglês *Whole Exome Sequencing*). Se por um lado, o sequenciamento completo do genoma

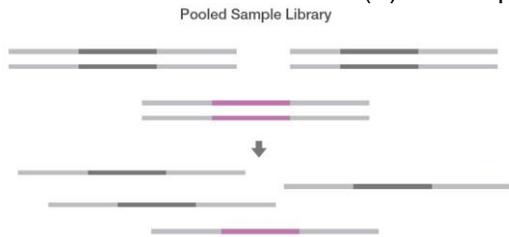
propicia a detecção de todas as variantes genéticas de um indivíduo em um único experimento, o sequenciamento do exoma, que corresponde a menos de 2% do genoma, é atualmente uma alternativa mais viável e custo efetiva. Em comparação com as abordagens tradicionais de mapeamento genético, o *WGS* e o *WES* apresentam-se como métodos diretos, já que seus resultados finais serão as bases resultantes dos fragmentos amplificados. Isto permite uma associação direta entre o fenótipo e a(s) variante(s) detectada(s) [7, 8, 9].

Quando a opção de sequenciamento de partes do genoma humano se apresenta (ex: painéis de genes candidatos ou *WES*), a primeira fase experimental corresponde a seleção das regiões específicas a serem sequenciadas. Nesse contexto, o método de hibridização por fase líquida destaca-se. De modo simplificado (Figura 1), após sua desnaturação, os fragmentos a serem sequenciados são flanqueados por adaptadores biotinizados, que recebem posteriormente a ligação de uma esfera de streptavidina, propiciando posterior captura por diferença de potencial, resultando assim em uma biblioteca contendo apenas os fragmentos de interesse para amplificação e sequenciamento [10].

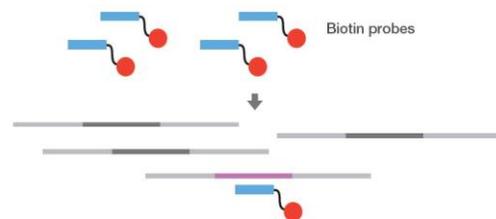
Espera-se da tecnologia de captura do exoma que sua extensão cubra a maior parte dos bancos de dados que contêm as regiões codificantes ou potencialmente codificantes do genoma. A Tabela 1 representa esta cobertura para o kit de captura do exoma utilizado [11]. A definição do exoma utilizada se estende por cerca de 62 Mb, capturando 20.794 genes e 201.121 exons com 340.427 sondas de 95 pb sem sobreposição, em uma biblioteca de tamanho recomendado entre 300 e 350 pb.



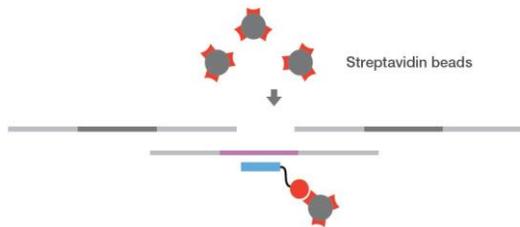
(A) Preparação da amostra



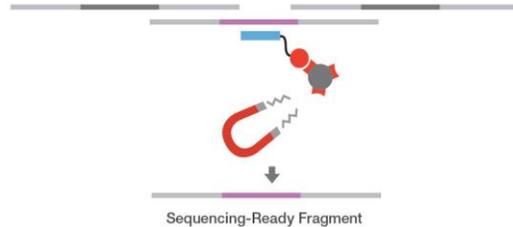
(B) Desnaturação da fita dupla de DNA (adaptadores e indexadores não mostrados)



(C) Hibridização de adaptadores biotinilados as regiões-alvo



(D) Etapa de enriquecimento utilizando *beads* de streptavidina



(E) Eluição dos *beads*

Figura 1. Etapas simplificadas para captura de partes específicas do genoma. Após sua desnaturação, os fragmentos são flanqueados por adaptadores biotinilados, que recebem posteriormente a ligação de uma esfera de streptavidina, propiciando posterior captura por diferença de potencial, resultando assim em uma biblioteca contendo apenas os fragmentos de interesse para amplificação e sequenciamento. Extraído de [11].

Tabela 1. Descrição da utilização de diversos bancos de dados de regiões codificantes do genoma pela tecnologia de captura empregada. Adaptado de [11].

Banco de dados	Tamanho	% utilizada	Descrição
CCDS	31,3 Mb	97,2%	Principal conjunto de regiões codificantes do genoma consistentemente anotadas e de alta qualidade
RefSeq	33,2 Mb	96,4%	Genes conhecidos como codificantes do <i>NCBI RNA Reference Collection</i>
RefSeq exons +	67,8 Mb	88,3%	Genes conhecidos como codificantes do <i>NCBI RNA Reference Collection</i> juntamente com DNA não codificante
Encode/Genecode	25,6 Mb	93,2%	Projeto que busca identificar todos os elementos funcionais do genoma humano
Alvos de microRNA	9,0 Mb	77,6%	Alvos preditos de microRNAs

É importante salientar que o *WES* possui limitações e desafios passíveis de aprimoramento. Podemos citar primeiramente a perda de informações relacionadas a sítios de regulação transcricional ou *splicing*. Dificuldades também surgem no âmbito da integração dos dados provenientes de grupos de pesquisa que utilizam kits de captura distintos, devido à falta de consenso na definição do exoma, pois muitos kits disponíveis no mercado contemplam não apenas as regiões codificantes do genoma [11]. Além disso, existem limitações técnicas que podem comprometer os resultados obtidos e dificultar a análise e interpretação dos dados. Exemplos de tais limitações são a falta de uniformidade na cobertura, regiões não-capturadas do exoma e diferenças na eficiência de captura a depender da existência de polimorfismos em regiões de hibridação das sondas.

Aplicações em medicina genômica

Aplicações do *WES* como ferramenta para identificar variantes patogênicas no âmbito da medicina genômica têm sido desenvolvidas recentemente [12, 13, 14]. Considerar um grupo ou apenas uma variante como causal para um determinado fenótipo de interesse não é uma tarefa simples. Para as doenças Mendelianas, caracterizadas por fenótipos causados por uma ou várias mutações em um gene e herdadas de modo recessivo, dominante ou de forma ligada ao sexo, temos como abordagens possíveis: considerar as alterações para os indivíduos afetados, a filtragem de variantes depositadas em bancos de dados públicos, ou ainda aquelas presentes em controles definidos no experimento. Essa abordagem tende a diminuir o número de candidatos, podendo igualmente serem filtrados pela presença em genes que melhor explicam o fenótipo apresentado, ou ainda por serem classificadas como promotores de alterações significativas na tradução. Esta estratégia tem sido bem sucedida, principalmente para doenças com padrão recessivo de segregação, como por exemplo, nos resultados apresentados por Choi *et al* [15].

Uma estratégia possível consiste no sequenciamento de trios pais-filho, principalmente para manifestações de fenótipos esporádicos, com o objetivo de identificar mutações *de novo* herdadas por esta geração. A efetividade deste método é compatível com um perfil de doença Mendeliana, apesar de também ter aplicabilidades para doenças com um perfil de herança mais complexo.

Fenótipos estudados

Nos últimos 20 anos, nosso grupo de pesquisa tem se dedicado ao estudo dos aspectos genéticos das epilepsias e dos distúrbios da formação cortical. Durante esse período, vários núcleos familiares segregando diversas formas de epilepsia e distúrbios corticais foram identificados e caracterizados cuidadosamente [17, 18, 19, 20, 21, 22].

Constituídas por uma classe de cerca de cinquenta doenças que afetam o sistema nervoso central, as epilepsias têm como manifestação a presença de uma atividade elétrica cerebral anormal, sendo estas generalizadas ou parciais [23]. Outra subclassificação as rotula como sintomáticas e idiopáticas. Os pacientes com uma causa conhecida ou suspeita de crises, associada a um déficit neurológico, são classificados como sintomáticos. Já os idiopáticos, aproximadamente 40% dos pacientes, não apresentam uma justificativa ou lesões cerebrais que expliquem esse quadro, levantando a hipótese de que suas causas podem estar vinculadas a variações genéticas [24]. Dentre as epilepsias idiopáticas, podemos citar as epilepsias mioclônicas juvenis, as epilepsias de lobo temporal e as epilepsias rolândicas benignas. Já entre os pacientes que apresentam epilepsia sintomática, as malformações corticais representam uma causa importante de crises. Também foram objeto de estudo no presente trabalho sequências obtidas de pacientes isolados ou parte de pequenas famílias de alguns projetos de colaboradores, tais como doenças oftalmológicas (projeto em colaboração com a Profa. Dra. Mônica Mello, UNICAMP) e doenças metabólicas (projeto em colaboração com o Prof. Dr. Roberto Giugliani, UFRGS).

Análises de bioinformática

Como produto resultante do sequenciamento, milhões de sequências (de comprimento entre 100-150 pb para Illumina®) precisam ser alinhadas a uma referência. Diversas implementações para esta tarefa são aplicáveis [26, 27, 28, 29, 30]. No entanto, aquelas baseadas na transformada de Burrows-Wheeler se destacam [31] devido a robustez computacional, compatibilidade com diferentes plataformas de NGS e velocidade [32]. Com relação ao sequenciamento em geral, divergências com o genoma de referência, denominadas variantes, devem ser avaliadas como possivelmente causais do fenótipo observado. São populares dois fluxos de trabalho para esse fim: um baseado no pacote de ferramentas SAMtools [33] e outro no *Genome Analysis Toolkit* [34, 35]. Para uma amostra humana, usualmente são extraídas aproximadamente 50.000 variantes das regiões codificantes do genoma, contra cerca de 4.000.000 em *WGS* [16]. Estes números ainda podem ser influenciados pelas diferenças na definição do exoma e por ancestralidade, como demonstrado nos resultados de [36]. Contudo, após processamento das sequências alinhadas, temos que o cenário mais provável para uma proporção considerável destas variantes será o de classificá-las como falso-positivos.

Desta forma, torna-se necessária a filtragem daquelas variantes possivelmente associadas ao fenótipo de interesse e aquelas que podem resultar em associações equivocadas. Estas se devem a variações introduzidas pela metodologia de sequenciamento empregada [37], por variantes frequentes na

população normal, variantes sem associação direta com a manifestação fenotípica de interesse [38] ou heterogeneidade de cobertura [39, 40].

Assim, fica evidente a necessidade de formas eficientes de anotação destas variantes. Deve-se, então, estabelecer critérios de priorização e métricas eficientes de discriminação de variantes inseridas por erros de alinhamento ou sequenciamento. Estratégias de classificação das variáveis baseadas no desenho experimental e em características dos indivíduos afetados e controles são igualmente cabíveis para efetuar sua filtragem. A predição teórica do comportamento deletério das variantes subsequentes e a determinação de seu impacto na transcrição podem ser estimados mediante a utilização de programas de predição, como PolyPhen2 [41] e SIFT [42].

Dadas as complexidades das etapas de alinhamento e descoberta de variantes, além da preocupação sempre presente em se garantir a qualidade das análises *in silico* realizadas [43, 44]. O desenvolvimento e aplicação de protocolos em bioinformática para análise dos dados são essenciais para lidar com o alto volume de dados gerados e realizar a ligação entre o experimento biológico e a interpretação dos dados gerados: representam assim uma excelente oportunidade para desenvolvimento. De fato, intervenções são aplicáveis em várias etapas a fim de se garantir a qualidade das variantes descobertas pela análise em *NGS*. Estas visam a verificação dos perfis de qualidade das sequências resultantes da reação de sequenciamento; alinhamento contra um genoma de referência, bem como a garantia de minimização de erros de alinhamento em regiões sabidamente ou potencialmente problemáticas para esta etapa; minimização da variância pelo uso

de covariáveis de contexto que podem interferir na interrogação por variantes, entre outras.

Diante da complexidade para execução das etapas supracitadas, observamos a necessidade da aplicação de um *workflow* robusto e baseado em ferramentas confiáveis, para contribuir com a automatização das análises, bem como identificar potenciais pontos de intervenção nos protocolos aplicados.

2. OBJETIVOS

GERAL

Aplicar e avaliar protocolos e métodos em bioinformática para análise de sequenciamento de exomas humanos.

ESPECÍFICOS

- Avaliar protocolos e aplicações disponíveis para serem utilizados na análise dos dados gerados pelo sequenciamento de exomas humanos.
- Avaliar o impacto de diferenças técnicas e populacionais na cobertura resultante da captura do exoma.

3. MATERIAIS E MÉTODOS

Perfil das amostras

As análises realizadas trataram principalmente com amostras de pacientes provenientes do biorrepositório para estudos de genética molecular em doenças neuropsiquiátricas da Faculdade de Ciências Médicas da Unicamp [45], cujo parecer de aceite se encontra no Anexo 1. Os pacientes selecionados para as análises neste projeto já tinham suas amostras de DNA coletadas e armazenadas. Foram selecionadas famílias com pelo menos 3 indivíduos afetados e um não afetado dentre as disponíveis no grupo de amostras do subprojeto “Epilepsias e Malformações do Desenvolvimento Cortical” [45]. Todos os pacientes incluídos no estudo foram cuidadosamente estudados do ponto de vista fenotípico, incluindo exames de eletrofisiologia e neuroimagem de alta resolução. Além disso, foram incluídas sequências obtidas de pacientes isolados ou parte de pequenas famílias de alguns projetos de colaboradores, devidamente aprovados pelos seus comitês de ética, tais como doenças oftalmológicas (projeto em colaboração com a Profa. Dra. Mônica Mello, UNICAMP) e doenças metabólicas (projeto em colaboração com o Prof. Dr. Roberto Giugliani, UFRGS).

Para a maior parte dos estudos que buscam associações genéticas que possam justificar o quadro clínico que afeta o paciente, é necessário ainda que se estabeleça um grupo de indivíduos controles: espera-se que não possuam a doença, que possivelmente tenham a mesma origem étnica e ou geográfica do paciente. Isto deve-se à difícil interpretação funcional da alteração no DNA de

pacientes e complexidade na inferência de alteração de função. Nesses casos, a pesquisa da mesma alteração em indivíduos sem a doença é crucial para auxiliar nas conclusões sobre patogenicidade da mutação encontrada. A associação entre as amostras e os projetos que as representam estão disponíveis no Anexo 3.

Para o “Proj1”, temos pacientes não relacionados. Todos são afetados e as amostras são pareadas, uma de sangue e outra de tecido cerebral displásico. Portanto estamos estudando uma malformação cortical chamada de Displasia Cortical Focal que é causa de crises epiléticas recorrentes. Para estas amostras estamos procurando mutações em mosaicismo nas vias mTOR e TAU.

Para o “Proj2”, temos indivíduos com um tipo de epilepsia temporal familiar. Ela é denominada Epilepsia Autossômica Dominante com Sintomas Auditivos, portanto é esperado um perfil monogênico de herança autossômica dominante. Temos dois indivíduos controles para essa família.

Para o “Proj3”, temos duas famílias em que se investiga epilepsia do lobo temporal mesial. As amostras de 1-4 constituem a família um com três indivíduos afetados e um indivíduo não afetado de uma família segregando Epilepsia do Lobo Temporal Mesial (ELTM). Os indivíduos 5-8 também são três indivíduos afetados e um não afetado de outra família segregando ELTM. Realizamos o sequenciamento do exoma a fim de identificar variantes que possam estar relacionadas com a etiologia da ELTM. Além disso, as duas famílias apresentam um padrão de herança autossômico dominante com penetrância incompleta.

Para o “Proj4”, temos 2 pacientes afetados, pai e filho, e dois não afetados, mãe e filha, com glaucoma primário de ângulo aberto, os afetados apresentam

aumento da pressão intraocular, escavação (que corresponde a morte das células ganglionares da retina) com correspondente perda de campo visual. Quanto ao padrão de herança ainda não temos certeza já que a doença geralmente apresenta padrão complexo de herança e fenótipo bastante variável, inclusive nesta família, buscou-se inicialmente apenas uma mutação em heterozigose no pai e filho e ausente na mãe e filha.

Para o “Proj5”, trata de diversos indivíduos não relacionados onde se estudam doenças metabólicas.

Delineamento de experimentos

A escolha dos indivíduos sequenciados, baseados em estudos de suas famílias e na forma em que a herança é segregada, representam parte crucial no delineamento experimental que foi proposto aos pesquisadores que obtiveram o material genético. Visto que a escolha dos indivíduos a serem sequenciados implica diretamente nas estratégias de sua análise posterior, o delineamento do experimento, bem como aspectos de tomada de decisão nas abordagens de sequenciamento e captura do exoma são considerados de extrema relevância.

Assim sendo, várias abordagens experimentais podem ser propostas, como elucidadas na Figura 2: análises de ligação são aplicadas a indivíduos múltiplos afetados em uma mesma família; análises de homozigocidade são efetivas para um indivíduo afetado proveniente de uma família consanguínea; estratégias de *double-hit* se aplicam ao caso de um indivíduo afetado com uma doença de herança

recessiva; as estratégias de *overlap* se aplicam a múltiplos afetados com uma doença dominante; busca de mutações *de novo* se aplicam a um afetado esporádico, onde toda ou parte da família é sequenciada; por fim, as estratégias baseadas em candidatos específicos têm um indivíduo afetado com uma doença dominante da família sequenciado sem outros membros da mesma família [16].

Como estratégia de sequenciamento, utilizamos o sequenciamento cíclico reversível por síntese, no equipamento Illumina® HiSeq 2500. As amostras foram sequenciadas no “Laboratório Central de Tecnologias de Alto Desempenho em Ciências da Vida”. O sequenciamento se baseia na repetição de três etapas cíclicas: a incorporação de um nucleotídeo; a aquisição de uma imagem da fluorescência emitida pela base nitrogenada incorporada e a clivagem do grupamento inibitório de ligação da base ligada. Para o passo de incorporação do nucleotídeo, uma molécula de DNA-polimerase se liga a amostra hibridizada nos *primers* dispersos pela superfície onde ocorre a reação de sequenciamento. Uma base complementar àquela presente na amostra é ligada. Uma próxima ligação é inibida devido a um grupamento inibitório de ligação presente nos nucleotídeos fornecidos para a reação. As bases não incorporadas são removidas e uma imagem é obtida da fluorescência da base ligada que é estimulada por laser. Após a obtenção da imagem, o grupamento inibitório é clivado, propiciando uma nova etapa de ligação de uma das bases fornecidas para a próxima etapa da reação.

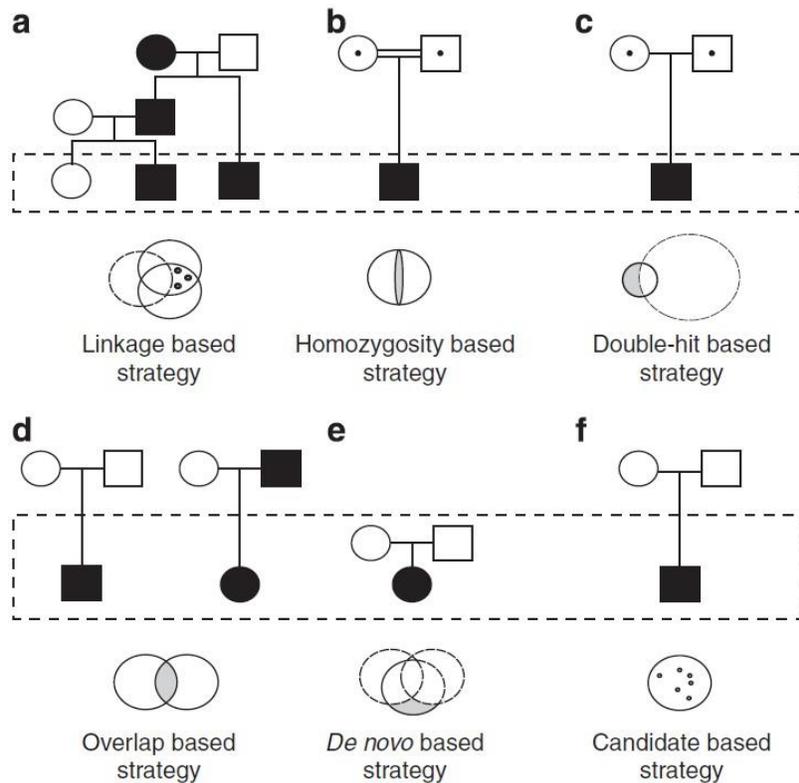


Figura 2. Estratégias de identificação de genes mediante ao sequenciamento completo do exoma. Os indivíduos sequenciados estão envolvidos por retângulos tracejados. Os diagramas de Venn representam as variações genéticas encontradas em cada exoma. Os círculos com contorno preenchido constituem as variantes em potencial que explicam o fenótipo apresentado. Análises de ligação são aplicadas a indivíduos múltiplos afetados em uma mesma família; análises de homozigocidade são efetivas para um indivíduo afetado proveniente de uma família consanguínea; estratégias de *double-hit* se aplicam ao caso de um indivíduo afetado com uma doença de herança recessiva; as estratégias de *overlap* se aplicam a múltiplos afetados com uma doença dominante; busca de mutações de novo se aplicam a um afetado esporádico, onde toda ou parte da família é sequenciada; por fim, as estratégias baseadas em candidatos específicos tem um indivíduo afetado com uma doença dominante da família sequenciado sem outros membros da mesma família. Extraído de [18].

Análises de bioinformática

Controle de qualidade pré alinhamento

Antes da realização do alinhamento foi realizado o controle de qualidade das sequências. Nesta etapa pudemos detectar possíveis problemas ou erros sistemáticos que potencialmente afetaram a reação de sequenciamento, e que podem introduzir vieses nas etapas de interrogação das bases pelo sequenciador, produção das bibliotecas, designação de valores de qualidade às bases interrogadas, demultiplexação das amostras (quando aplicável) e, finalmente, repercutindo em todas as etapas de processamento dos dados, culminando em erros na classificação das variantes e sua eventual interpretação errônea [46]. Foram utilizados programas como o FastQC (0.11.3) [44] e Rqc (1.2.0) [43], que fornecem informações sobre estatísticas básicas das sequências; qualidade; conteúdo de GC; distribuição do tamanho das sequências; níveis de duplicação das sequências; e sequências sobre-representadas.

Alinhamento das sequências

Como etapa central das análises de dados em sequenciamento genômico, o alinhamento das milhares de sequências a um genoma de referência é crítico, e diretamente dependente da metodologia de sequenciamento empregada para interrogação das bases nitrogenadas. O tempo para indexação e alinhamento de sequências a um genoma de referência obedecem a uma tendência linear a medida em que se aumenta o tamanho da região a ser alinhada e sua cobertura final para

várias ferramentas disponíveis para este fim, como demonstrado por [47]. São determinantes também a infraestrutura do servidor em que serão realizadas as análises, bem como o potencial de escalonamento de processos e disponibilidade de unidades de processamento para cada um dos processos iniciados.

Para a realização do alinhamento das sequências contra o genoma de referência GRCh38 utilizou-se o alinhador BWA (0.7.12) [31], ferramenta baseada na transformada de Burrows-Wheeler, e amplamente utilizada pela comunidade científica, incluso em grandes projetos de pesquisa, como o Projeto 1000 Genomas [48]. BWA é um software que implementa três algoritmos distintos: BWA-backtrack, BWA-SW e BWA-MEM. O primeiro se aplica a sequências oriundas de tecnologias policromáticas com comprimento maior que 100 pares de base. Os outros dois lidam com sequências de 70 a 1000 bases. No entanto, a implementação BWA-MEM, que é a mais recente, é recomendada para sequências de alta qualidade, sendo mais rápido e acurado [49].

Processamento pós-alinhamento

Após o alinhamento das sequências realizado, iniciamos o processamento das sequências alinhadas, de forma a remover vieses introduzidos pela dificuldade em alinhar estas sequências, seja pela qualidade ou similaridade múltipla das mesmas. Visto a complexidade destes arquivos, houve a necessidade de mantê-los sempre ordenados e indexados, para que acessos a partes específicas fossem otimizados para se obter maior desempenho nas análises. Para a etapa de

ordenação do arquivo com as sequências alinhadas em coordenadas genômicas, utilizamos a função “SortSam” do conjunto de ferramentas Picard (1.131) [50].

Ainda no contexto de minimização das influências técnicas nos resultados finais do alinhamento, faz-se necessária a marcação de sequências duplicadas. A incorporação de duplicatas pode introduzir falsos-positivos ou falsos-negativos, e ainda puderam apontar possíveis problemas na etapa de preparação da biblioteca. Para tal, novamente foi utilizado o suíte do Picard na implementação “MarkDuplicates”.

Não obstante, a forma de preparação da biblioteca, incluindo a forma de multiplexação e sequenciamento pode influenciar o resultado final da posterior chamada das variantes. Para tanto, inserimos informações como identificação da amostra; biblioteca utilizada para amplificação das amostras; centro sequenciador; tecnologia empregada para interrogação das bases; câmara de fluxo e canaleta onde ocorreu a reação de sequenciamento e adaptadores utilizados para multiplexação das amostras. Para tanto, a ferramenta utilizada foi Picard “AddOrReplaceReadGroups”. Esta etapa foi de extrema importância para as etapas posteriores de recalibração da qualidade e chamada de variantes.

Quando se é necessário fazer acessos rápidos e otimizados ao arquivo, se faz necessária a criação de um índice para o arquivo, que como o nome ilustra, faz com que o acesso a um arquivo binário e altamente compactado seja feito de modo otimizado. Para criação dos índices nas análises, sempre se utilizou a implementação “BuildBamIndex” do Picard. Pontos de checagem para definir se erros ocorreram durante a análise são primordiais para evidenciar falhas que podem

estar relacionadas a inconsistências nos arquivos. Para tal, frequentemente utilizamos a função “ValidateSamFile” do Picard, que além de verificar a consistência dos arquivos BAM também checa se os índices dos respectivos arquivos estão em concordância.

Realinhamento local

Com o intuito de minimizar as bases alinhadas de modo não inteiramente correspondente ao genoma de referência, uma etapa de realinhamento local foi introduzida nas análises. Em geral, regiões que possuem deleções ou inserções são mais suscetíveis a um realinhamento local, visto que há uma sucessão de bases que podem ser confundidas com variantes sucessivas, quando na verdade refletem uma organização diferente da do genoma de referência. Para esta distinção, a ferramenta utiliza todo o arredor da ocorrência para executar o realinhamento de modo mais sensível e específico que no alinhador utilizado anteriormente. Para tal, utilizamos o conjunto de ferramentas do GATK (3.3-0) [35], determinando os intervalos e regiões propícias a um realinhamento local com a implementação “RealignerTargetCreator” e o realinhamento propriamente dito com “IndelRealigner”. Utilizamos as inserções e deleções disponíveis no banco de dados do dbSNP e os *indels* já disponíveis na própria amostra. Foi utilizado um limite de significância de 40%, ou seja, se o melhoramento no alinhamento foi significativo o suficiente para ser maior que este valor, ele foi realizado. Note que este parâmetro pode ser ajustado para valores menores, caso tenha-se um experimento com baixa cobertura ou quando se buscam *indels* com baixa frequência alélica.

Recalibração da qualidade

Ao passo que um realinhamento local com *indels* reduz o número de falso-positivos, é necessário também distinguir entre as qualidades das variantes pontuais onde observamos quais são mais prováveis de serem representações de uma situação biologicamente verdadeira daquelas introduzidas por variações outras, como o preparo da biblioteca, reação de sequenciamento e alinhamento das sequências. Para tanto, utilizamos a ferramenta “BaseRecalibrator” do GATK, que assume todas estas variantes como errôneas e indicativas de baixa qualidade de alinhamento ou de baixa qualidade das bases interrogadas nas sequências consideradas. Para cada uma destas variantes foi calculada uma lista de covariáveis contextuais: o grupo ao qual pertencem os *reads*, escores de qualidade associados a variante e ao ciclo de sequenciamento que àquela base pertence e ao contexto de dinucleotídeos. Dadas estas informações, o GATK implementa um modelo estatístico que utiliza estas informações para a estimação da probabilidade de erro dada uma covariável em particular, obtida ao se calcular a razão de variantes pelo número total de observações. Assim, foi atribuído um valor de qualidade que variou de 0 a 127 na escala *phred* que será substituído no arquivo contendo as sequências alinhadas mediante o uso da função “PrintReads” também do GATK.

Controle de qualidade pós alinhamento

O alinhamento de milhares de sequências de DNA a um genoma de referência, seguido dos diversos passos descritos nas seções acima, ilustram o

quão complexo são as análises em sequenciamento de alto desempenho. Contudo, mesmo mediante o sucesso da execução dos passos acima, o controle de qualidade pós-alinhamento é essencial para confrontar as expectativas quanto ao experimento com os dados reais obtidos pelas análises.

Para tanto, utilizamos a ferramenta “CollectMultipleMetrics” do Picard [50], que estima métricas referentes a estatísticas básicas do alinhamento, distribuição do tamanho dos insertos após alinhamento, distribuição dos scores de alinhamento, bem como métricas ligadas a qualidade por ciclo da reação. De modo a analisar a cobertura na definição do exoma, utilizamos a ferramenta “CalculateHsMetrics” também do Picard. Com estes resultados em mãos, traçamos o perfil de vários níveis de cobertura para as amostras consideradas.

Descoberta de variantes

Para a etapa de descoberta das várias variações em comparação ao genoma de referência GRCh38, utilizamos o algoritmo “HaplotypeCaller” do GATK. Este algoritmo interroga SNPs e *indels* simultaneamente, dados os haplótipos de uma região ativa. As regiões ativas são determinadas a partir de evidências da presença destas variantes. Para cada uma, o programa constrói um grafo de De Bruijn para identificar os possíveis haplótipos representados. Os sítios variantes são posteriormente obtidos via o realinhamento utilizando o algoritmo de Smith-Waterman, obtendo a probabilidade de cada um dos haplótipos. Aquele mais provável é atribuído a amostra. O método de emissão das posições utilizado foi de blocos condensados, ou seja, a emissão não é somente das variantes pontuais,

mas também das regiões de consenso entre as sequências alinhadas e a referência. Esta estratégia foi utilizada para examinar os blocos de variantes de todas as amostras em separado. Para interrogar as variantes em grupos de amostras, utilizamos a ferramenta “GenotypeGVCFs” também do suíte do GATK.

Anotação das variantes

Como resultado final e mais interpretável do sequenciamento, uma lista resultante das variantes contém os trechos ou localizações que não constituem um consenso no alinhamento contra o genoma de referência. Contudo, posições genômicas e variações de base por si só não são informativas: faz-se necessária a introdução de informações a respeito daquela posição que auxiliarão no processo de filtragem das variantes de interesse. O *Variant Effect Predictor (VEP)* [51] é uma ferramenta que executa esta etapa e foi utilizada para inserir informações como: genes e transcritos afetados pela variante em questão, localização da variante no contexto genômico em que ela se insere, consequência na codificação de proteínas, potencial deletério e associações entre esta variante e condições conhecidas, entre outras. Com a lista de variantes em mãos, foi finalmente possível filtrá-las tendo em vista o conhecimento biológico prévio ao se realizar o experimento. Assim sendo, a lista de variantes anotadas foi o produto final do protocolo aqui proposto.

Avaliação da cobertura no sequenciamento do exoma

Utilizamos dados públicos do Projeto 1000 Genomas para investigar a variação na cobertura do exoma nestas amostras. Para avaliação de diferenças

metodológicas influenciando a cobertura, utilizamos 120 indivíduos de 4 populações. Estas amostras são provenientes de três fases do projeto que diferem na metodologia empregada para captura do exoma. Para avaliação das diferenças populacionais utilizamos 120 indivíduos da fase III, de 12 populações distintas. Foi utilizado *MDS (multidimensional scaling)* para lidar com a alta dimensionalidade e permitir uma comparação mais direta entre os grupos. Maior detalhamento na seção de resultados “Whole exome sequencing depth of coverage is susceptible to technical and population differences”.

5. RESULTADOS

Diversas etapas em bioinformática foram realizadas no intuito de descobrir variantes em nossas amostras. A Tabela 2 sumariza com uma breve descrição as ferramentas utilizadas neste trabalho.

Tabela 2. Sumarização dos softwares empregados para análise.

Software	Versão	Descrição
FastQC	0.11.3	Sumarização de controle de qualidade por amostra considerada.
Rqc	1.2.0	Visualização de gráficos de controle de qualidade em alta resolução.
BWA	0.7.12	Alinhador que implementa a transformada de Burrows-Wheeler.
Picard	1.131	Conjunto de ferramentas utilizado para tarefas como ordenação, marcação de duplicados, indexação e validação.
GATK	3.3-0	Conjunto de ferramentas utilizado para realinhamento local, recalibração da qualidade e descoberta de variantes.
VEP	78	Anotação das variantes.

Workflow para análise de dados de sequenciamento de alto desempenho

Como o resultado deste trabalho, temos a aplicação de um *workflow* robusto para análise de dados de sequenciamento de alto desempenho, cujas etapas estão exemplificadas na Figura 3 e disponíveis em <https://goo.gl/zZ88F3>.



Figura 3. Esquema simplificado do workflow desenvolvido.

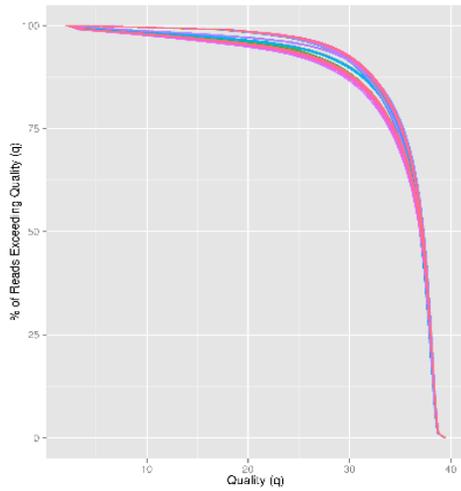
O controle de qualidade pré e pós-alinhamento ofereceram-nos pontos de tomada de decisão quanto a qualidade e robustez dos dados apresentados, nos quais pudemos identificar comportamentos atípicos, passíveis de investigação de suas causas, desde a etapa de preparação da biblioteca a interrogação das variantes. As etapas de pré e pós alinhamento (que incluíram o realinhamento ao redor de inserções e deleções e recalibração da qualidade das variantes) e chamada de variantes foram realizadas com um dos algoritmos mais utilizados pela comunidade científica para tratar com dados humanos [35]. Nossa intervenção nestas etapas contemplou ajustar entradas e saídas dos programas, sempre nos preocupando em rastrear os registros de execução dos mesmos, garantindo que cada um dos passos tenha sido executado com sucesso, resultando em variantes potenciais para cada um dos fenótipos estudados. Pontos que ainda carecem de intervenção e aprimoramento também foram identificados, como por exemplo, a utilização de um banco de variantes normais da população brasileira.

Controle de qualidade pré-alinhamento propicia identificar comportamentos anômalos nas bibliotecas sequenciadas

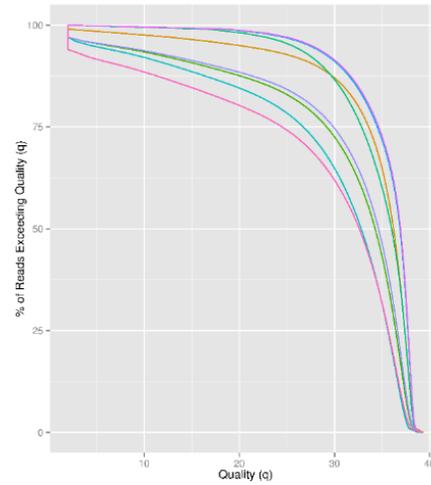
Métricas de qualidade dadas pelo programa FastQC [42] foram sumarizadas para cada um dos experimentos realizados, possibilitando um panorama geral quanto a qualidade das sequências para cada um dos arquivos das sequências utilizados como entrada nesta etapa da análise. De forma a explicitar melhor os resultados, com representações visuais para cada um dos ciclos da reação de sequenciamento, utilizamos o pacote Rqc [41]. Observamos que algumas amostras possuíram comportamentos fora daqueles esperados para cada um dos testes realizados.

A distribuição da qualidade das sequências é um bom indicativo do sucesso do sequenciamento de milhares de sequências. A Figura 4 apresenta um exemplo deste perfil de qualidade dada por uma curva de sobrevivência. Nesta representação, visualiza-se a proporção de fragmentos cuja qualidade média excede o limiar dado na abcissa. Como os gráficos são estratificados por fita e alocação física, é possível diferenciar padrões específicos destas variáveis.

Avaliamos a qualidade média dos fragmentos ao longo dos ciclos de sequenciamento. A Figura 5 apresenta quedas gradativas habitualmente observadas, sendo mais pronunciada ao final das sequências. Identificamos também quedas pontuais em alguns ciclos, possivelmente associadas a falhas no sistema de sequenciamento. Esta variabilidade nos perfis de qualidade foi minimizada pela recalibração destes índices.

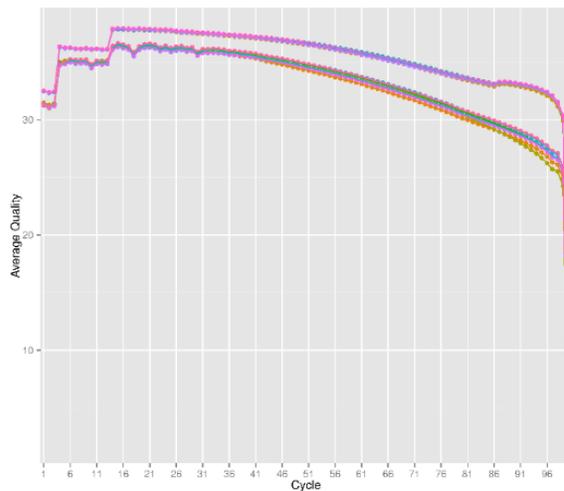


A

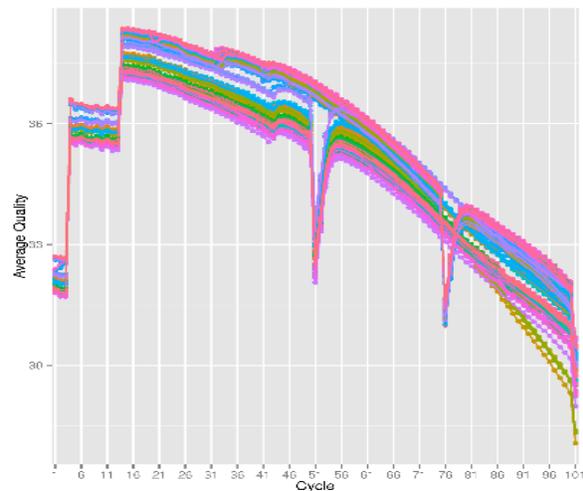


B

Figura 4. Comparação entre curvas de sobrevivência representando a distribuição das qualidades para dois experimentos. (A) Quase não se percebe a distinção entre as fitas direta e reversa, com mais de 80% das sequências com qualidade média superior a q30. Comportamento diferente de (B) onde se nota diferentes padrões de qualidade influenciados pelo fita e canaleta a que pertencem as sequências.



A



B

Figura 5. Comparação entre a qualidade média por ciclo para dois experimentos. (A) Temos uma leve queda na qualidade na fita reversa considerada normal pela metodologia de sequenciamento adotada. (B) Quedas abruptas da qualidade também podem ser observadas em ciclos específicos.

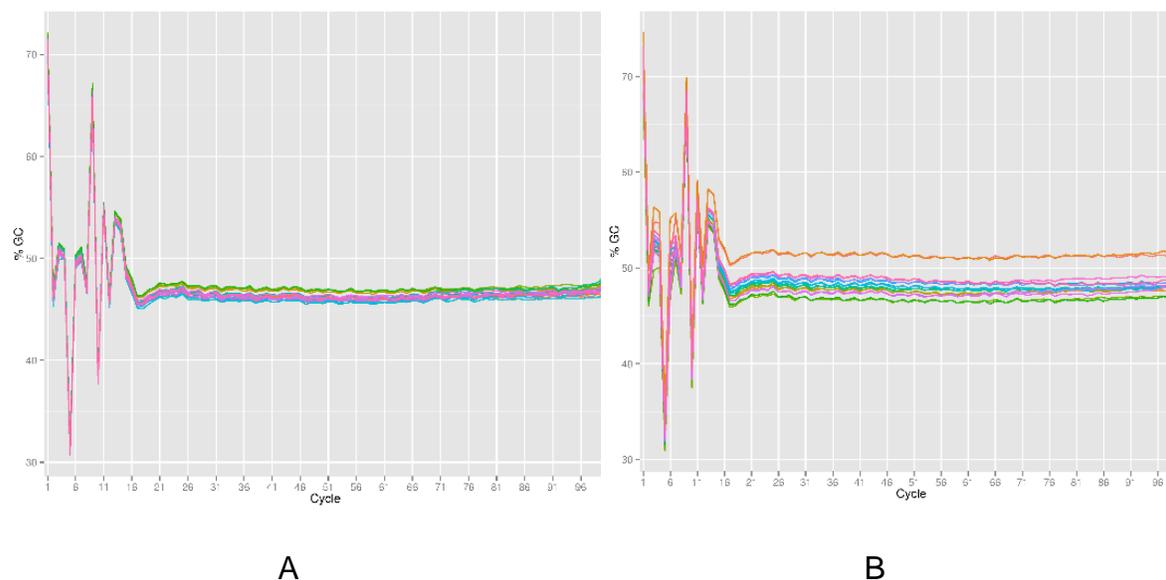


Figura 6. Comparação entre a distribuição de GC em dois experimentos. Em (A), um padrão esperado para metodologia empregada no preparo da biblioteca. O painel (B) evidencia um comportamento anômalo em uma amostra com alto percentual de GC.

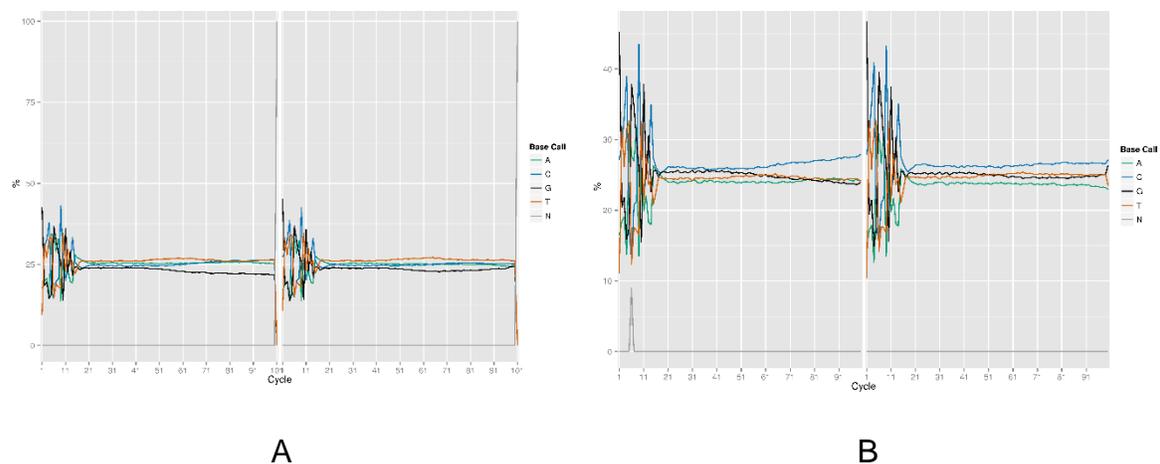


Figura 7. Exemplos de comportamentos anômalos nas percentagens de bases nitrogenadas. (A) apresenta um padrão mais homogêneo de distribuição das bases nitrogenadas, porém apresenta um pico de bases não identificadas ao final das fitas direta e reversa. Padrões que afetam uma base também podem ser identificados, como em (B), na qual a percentagem de C apresenta aumento significativo ao final da fita direta e há um aumento de bases não interrogadas “N” no início desta mesma fita.

Analisar a contribuição de cada uma das bases nitrogenadas obtidas na reação de sequenciamento é de extrema importância para se identificar o sobre-sequenciamento de um dado nucleotídeo. A percentagem de GC encontrada resultou num padrão variante esperado no início das sequências, devido a metodologia empregada na fragmentação das moléculas de DNA, como na Figura 6. Padrões espúrios em outras regiões das sequências, apesar de inesperados, foram encontrados, como a proeminência na representação de uma única base ou o aumento de bases não interrogadas pelo sequenciador (Figura 7).

Controle de qualidade pós-alinhamento propicia um diálogo entre o experimento biológico e os resultados “in silico”

O realinhamento local ocorreu em cerca de 10 milhões de regiões para cada uma das amostras analisadas, afetando em média 79% das sequências de cada um dos arquivos em uma região correspondente a mais de 15% do exoma.

A recalibração contextual das variantes encontradas ocorreu para todas as amostras. A Figura 8 ilustra o resultado desta calibração para a amostra P20 (Proj 1), para a qual observamos um perfil de qualidade bem mais homogêneo que o original reportado.

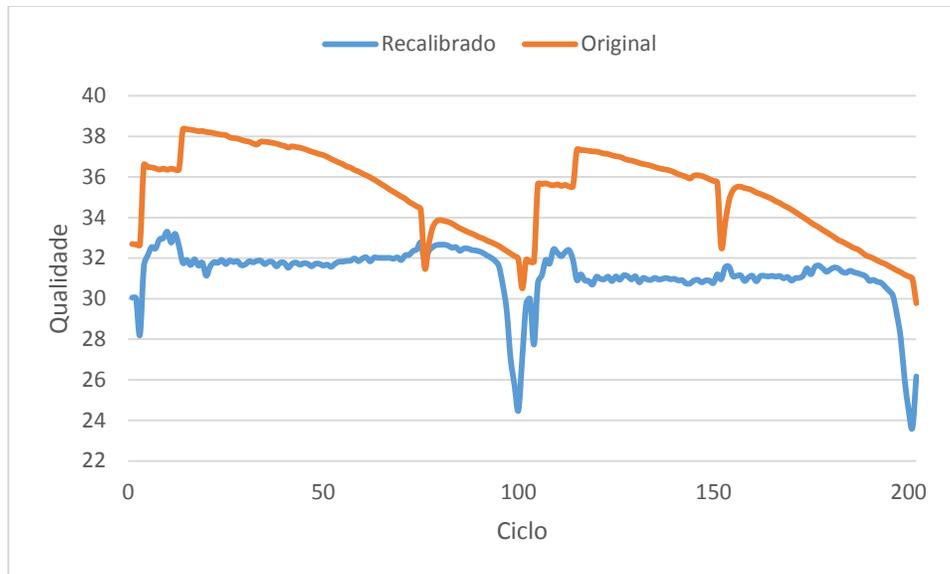


Figura 8. Perfil de qualidade para cada um dos ciclos de sequenciamento antes e após recalibração de qualidade.

Temos disponíveis na Tabela 3 um sumário dos dados de controle de qualidade pós alinhamento feito para cada uma das amostras utilizadas. Com os resultados em mãos, pudemos analisar o percentual de sequências alinhadas e daquelas propriamente pareadas (em que ambas as fitas foram alinhadas satisfatoriamente contra o genoma, sem sequências duplicadas), resultando em média uma taxa de alinhamento de 73%. Em média, 46% das sequências foram alinhadas fora da definição do exoma. A cobertura média alcançada varia para cada experimento, e em média é de 59X.

Tabela 3. Sumário do controle de qualidade pós-alinhamento.

Amostras	Total de sequências	Pareados	Únicos	Fora do alvo	Cobertura média	>30X	>50X
19313	3.746.31.790	73,2%	63,5%	55,3%	149,48	84,4%	78,6%
4411	304.306.104	54,3%	48,1%	60,8%	81,20	72,7%	61,5%
54112	342.193.426	65,3%	58,8%	54,2%	131,60	82,3%	75,6%
74212	351.752.804	67,1%	60,5%	49,0%	156,66	84,0%	78,5%
p13	402.834.332	72,0%	65,4%	57,7%	154,66	84,9%	79,5%
p16	416.400.890	71,8%	65,3%	56,9%	163,39	85,3%	80,4%
p19	511.190.768	65,7%	59,5%	53,4%	203,37	85,8%	81,9%
p20	448.049.716	68,7%	62,3%	54,1%	180,62	85,8%	81,4%
586	32.846.950	93,7%	85,2%	47,6%	22,20	25,5%	7,7%
587	35.732.650	93,6%	84,9%	46,9%	24,83	30,2%	11,3%
588	57.401.346	93,1%	84,5%	49,2%	37,02	47,7%	25,3%
920	37.460.606	92,7%	83,8%	50,7%	23,81	28,9%	11,3%
931	33.567.366	93,7%	84,9%	51,7%	20,97	23,7%	7,1%
938	36.561.778	93,1%	84,3%	53,4%	21,91	25,3%	8,3%
942	37.951.510	94,5%	85,8%	52,3%	23,14	26,7%	9,0%
943	15.422.152	95,3%	86,2%	49,5%	10,51	4,3%	0,4%
F1-1	59.844.796	82,4%	73,3%	38,6%	43,46	48,3%	29,3%
F1-2	42.439.336	84,0%	75,3%	42,3%	29,02	36,0%	15,5%
F1-3	39.275.014	86,3%	77,5%	45,0%	26,26	32,2%	12,6%
F1-4	58.830.740	83,0%	74,4%	43,7%	37,66	47,6%	24,2%
F2-5	61.215.924	80,5%	71,6%	39,5%	42,07	51,6%	30,4%
F2-6	56.880.030	81,7%	73,0%	38,3%	40,53	50,1%	27,8%
F2-7	82.673.450	80,2%	71,9%	48,3%	46,56	55,0%	34,9%
F2-8	69.037.872	80,7%	71,7%	39,7%	46,86	53,9%	33,9%
P4-1	50.591.714	88,9%	79,2%	38,8%	40,15	46,1%	26,4%
P4-2	48.087.722	88,7%	79,4%	40,3%	37,54	43,0%	24,0%
P4-3	53.810.560	88,7%	79,2%	41,7%	40,33	47,3%	27,8%
P4-4	48.705.924	89,0%	79,6%	41,0%	36,64	43,2%	23,8%
102214	104.266.352	75,4%	67,2%	39,5%	66,99	62,3%	50,0%
64814	115.238.336	77,5%	68,3%	36,1%	78,68	68,1%	56,0%
93814	81.409.760	80,3%	71,1%	37,1%	57,77	57,4%	44,2%
93914	106.591.324	69,4%	60,4%	38,4%	62,62	60,7%	48,4%

As distribuições do tamanho dos insertos e da qualidade do alinhamento das sequências foram obtidas com Picard “CollectMultipleMetrics” [46]. Temos que os picos de distribuição estiveram todos entre 100 e 200, que são uma métrica da

distância entre os finais das sequências do par, significando que os tamanhos dos fragmentos da biblioteca após alinhamento consistem entre 200 e 300 pb (Figura 9). Contudo, temos que segundo o protocolo empregado, o tamanho ideal dos insertos deveria ser de 300 a 350 pb.

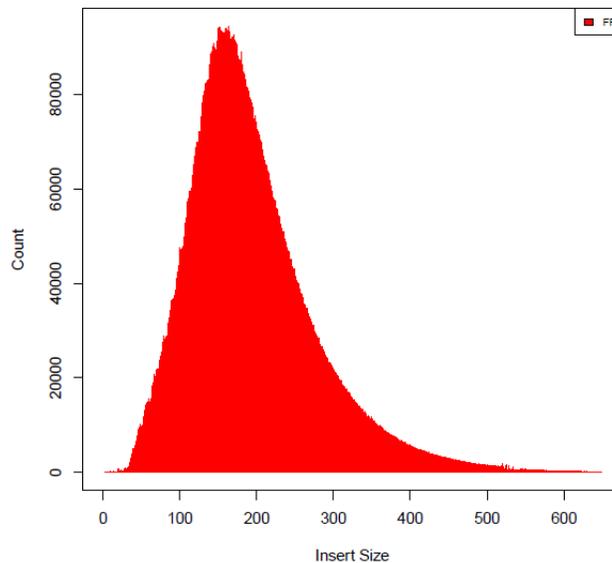


Figura 9. Distribuição do tamanho dos insertos após alinhamento.

A ferramenta “CalculateHsMetrics”, também do Picard, permitiu-nos traçar um perfil de vários níveis de cobertura para as amostras de nossos experimentos. Pelo padrão observado, observamos que em torno de 1×10^8 sequências foram necessárias para assegurar pelo menos uma cobertura maior que 50x em 50% da definição do exoma utilizando o protocolo empregado na preparação das bibliotecas (Figura 10).

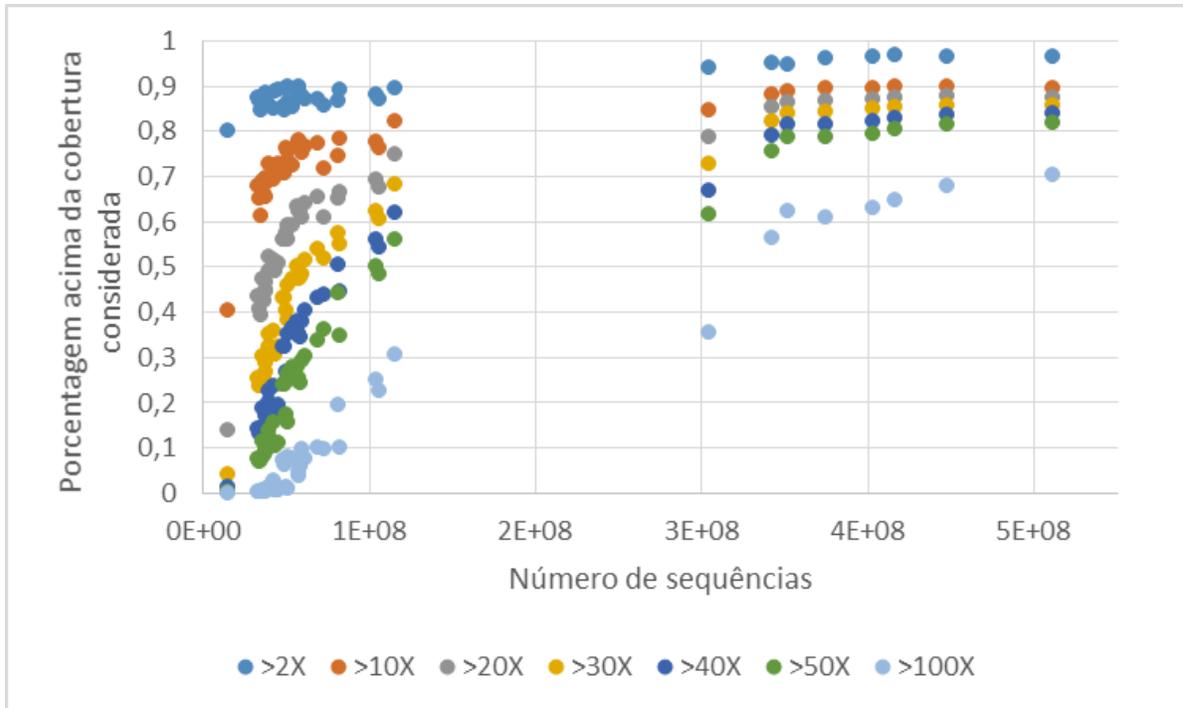


Figura 10. Gráfico da dispersão de cobertura pelo número de sequências consideradas.

Também comparamos a cobertura e o percentual de bases fora do alvo se considerarmos a intersecção entre a versão do kit de captura que somente contempla a versão restrita do exoma. Para esta região, temos que 58% das sequências seriam alinhadas fora do alvo, com uma cobertura maior que em sua versão estendida de 66X. Um perfil geral de cobertura está disponível na Figura 11. Note que a cobertura média apresenta um padrão linear para as duas regiões consideradas, mas se consideramos a distribuição das coberturas, esse padrão apresenta um limiar de saturação como apresentado na Figura 10.

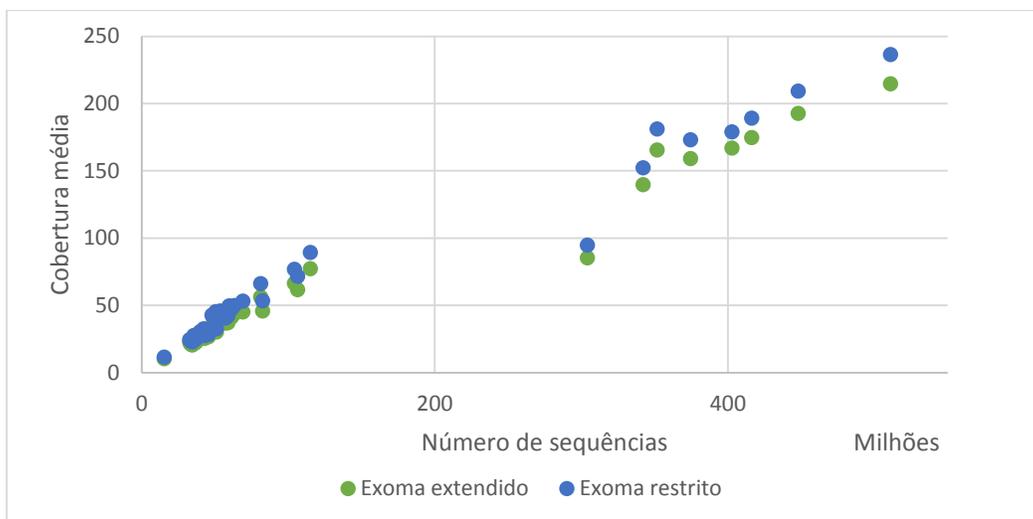


Figura 11. Gráfico da variação de cobertura média pelo número de sequências consideradas.

Anotação das variantes encontradas adiciona informações que possibilitam filtragem posterior

As regiões diferentes do consenso foram interrogadas de forma a fornecer uma lista de variantes para potencial associação com os fenótipos em questão. A Tabela 4 ilustra os resultados obtidos. O experimento “Proj 1” possui o maior número de variantes, visto que possuía o maior número de sequências. Em média 86% das variantes encontradas já foram descritas no banco de dados do dbSNP [52]. A maioria das variantes encontradas está em regiões intrônicas e intergênicas. Em média, 2,7% das variantes têm como consequência uma alteração sinonímia e 0,8% das variantes, em média, foram classificadas como potencialmente patogênicas ou deletérias pelos algoritmos de predição de impacto SIFT e Polyphen [41, 42]. No entanto, como evidenciado na Tabela 5, grande parte destas variantes possuem uma baixa cobertura, menor que 10x, podemos notar que em média 53% das variantes foram filtradas por este critério. Quanto a presença destas variantes em

bancos de dados, a proporção se manteve em 87%. A maior parte das variantes continua presente nos introns, contudo notamos uma queda considerável da proporção de variantes em regiões intergênicas e conseqüentemente a um aumento na proporção de impactos previstos na transcrição destas variantes.

Tabela 4. Perfil das variantes encontradas. O projeto Proj3 possuiu duas famílias que foram analisadas em separado.

Experimento	Total	Presentes no dbSNP	Intrônicas	Intergênicas	Sinonímias	Previstas como patológicas
Proj1	6.549.506	88,6%	49,6%	38,9%	0,4%	0,1%
Proj2	976.764	85,2%	50,5%	31,4%	2,2%	0,6%
Proj3 – F1	489.251	86,0%	52,7%	24,9%	3,5%	0,9%
Proj3 – F2	591.509	86,2%	52,1%	27,4%	2,9%	0,8%
Proj4	423.106	86,8%	51,4%	24,9%	3,9%	1,0%
Proj5	678.047	83,5%	51,7%	25,9%	3,5%	1,1%

Tabela 5. Perfil das variantes encontradas com cobertura maior ou igual a 10X em pelo menos uma das amostras para cada projeto.

Experimento	Percentual do total	Presentes no dbSNP	Intrônicas	Intergênicas	Sinonímias	Previstas como patológicas
Proj1	97,2%	88,7%	49,8%	38,7%	0,4%	0,1%
Proj2	29,3%	84,0%	52,3%	11,4%	7,4%	2,1%
Proj3 – F1	38,5%	88,7%	54,8%	5,2%	8,8%	2,4%
Proj3 – F2	34,2%	87,8%	55,3%	5,9%	8,3%	2,2%
Proj4	38,8%	88,6%	52,0%	6,3%	9,7%	2,5%
Proj5	39,8%	87,0%	53,9%	6,6%	8,7%	2,6%

Tendo aplicado todos os passos da análise, avaliamos o tamanho e tempo das análises, sempre em relação ao número total de sequências geradas pela reação de sequenciamento. O tamanho corresponde majoritariamente ao arquivo binário com as sequências alinhadas contra o genoma de referência e ao arquivo com as posições variantes no genoma, bem como dos arquivos de *logs* intermediários das análises (Figura 12). O tempo diz respeito as etapas de processamento deste arquivo, incluindo as etapas a partir da marcação de duplicados até a descoberta de variantes (Figura 13). Assim, a grosso modo, esperamos gerar aproximadamente 40GB de dados em 30 horas a partir de 100 milhões de sequências.

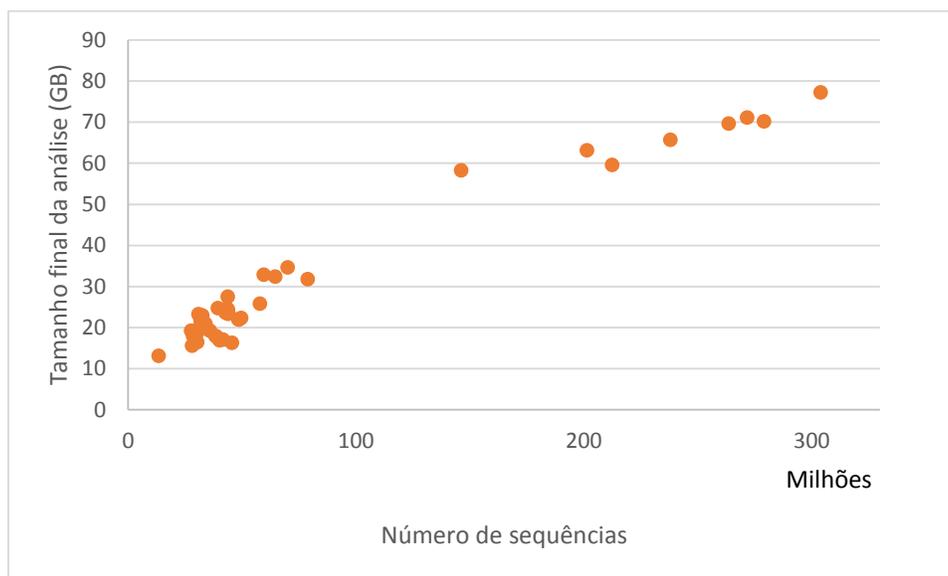


Figura 12. Variação do tamanho final dos arquivos de alinhamento pelo número de sequências.

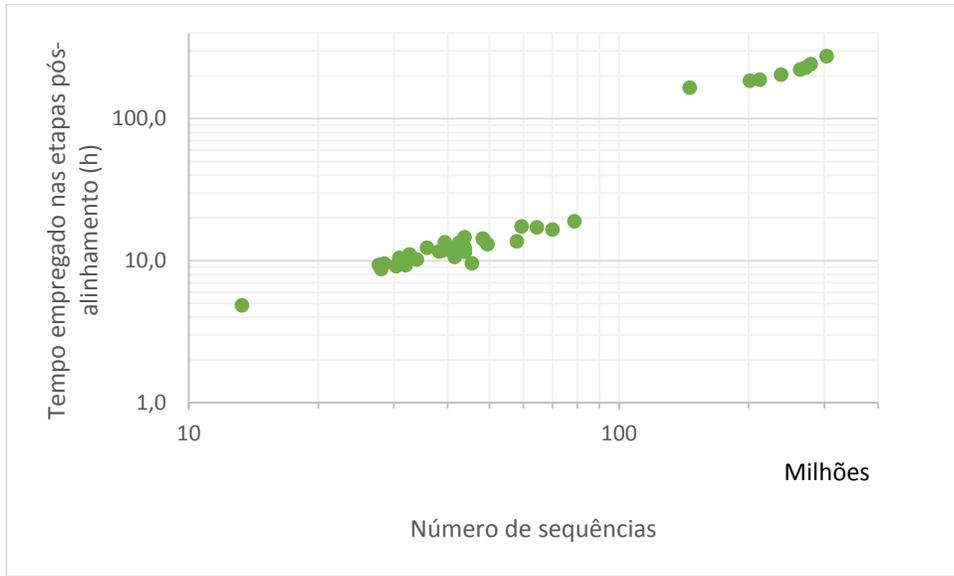


Figura 13. Variação do tempo para etapas de processamento do arquivo de alinhamento e chamada de variantes pelo número de sequências.

Whole exome sequencing depth of coverage is susceptible to technical and population differences

Murilo G Borges^{1,2}; Cristiane S Rocha^{1,2}; Benilton S Carvalho^{1,2} and Iscia Lopes-Cendes^{1,2*}

Author details

¹ University of Campinas, School of Medical Sciences, Department of Medical Genetics, R. Tessália Vieira de Camargo, 126, Cidade Universitária “Zeferino Vaz”, 13083-887 Campinas, Brazil.

² Brazilian Institute of Neuroscience and Neurotechnology (BRAINN), Rua Vital Brasil, 251, HC-UNICAMP, 2th floor, 13083-888 Campinas, Brazil.

Abstract

Background: The coding region of the human genome corresponds to less than 2% of its entirety. This portion, called exome, concentrates most of known pathologic variations. After alignment, each base position can be interrogated a certain number of times. This metric is referred as depth. It is important to determine whether technical and ethnic differences can affect this parameter. In the present work, we aim to investigate and understand the technical and ethnic patterns of base-specific depth on whole exome sequenced subjects from the 1000 Genomes Project.

Results: Comparative analysis with multidimensional scaling projections suggests that exome capture behaves differently across different capture methodologies and is susceptible to population differences. We believe that one reason for this is the fact that probes used for capture may require population-specific design, as they do not account for many population specificities. Genetic differences due to isolation and selection between the populations may explain these findings.

Conclusions: The success and reliability of exome sequencing strongly depends on the capture phase reaction. Lack of effectiveness in capturing sequences from different ethnicities may represent a concern when dealing with population studies. Technical integration is challenging as well, as different methodologies directly impact final exome coverage patterns.

Keywords: depth of coverage; whole exome sequencing; population genetics; ethnicity; bioinformatics

Background

Whole exome sequencing (WES) has emerged as a powerful tool in genomic medicine, as it provides the possibility of interrogating the most interpretable portion of the genome [1]. This strategy allowed the identification of causal variants with high success rate in several Mendelian disorders [2, 3]. But if the genotype tends to assume a more complex profile, exome resequencing can still add important information, but some issues in interpretation as well [5, 6]. The development of many different capture technologies adds complexity to data integration. This difficulty is due different exome capture efficiency [4]. Results obtained from next generation sequencing technologies may suffer biases due to experimental design, sequencing strategies and variant calling methods [7]. However, WES may also include another source of bias: the exons' capture reaction efficiency, which directly affects the sequences final depth uniformity, affecting final interpretation [8, 9].

In fact, reads uniformity, depth and quality depend directly on the enrichment phase and primordial to the WES reaction. The depth of coverage, depth or still the coverage, is the average number of sequenced and properly aligned bases or reads at a certain position or region. Its expected value is one of the first parameters considered in sequencing studies design [10]. Depth has big variations in WES studies. Even when the expected coverage is high, capturing some regions can still be problematic [11]. Some capture methodologies promise to cover certain regions more efficiently, bringing difficulties to researchers when opting for a capture platform [12]. In spite of its complexity, understanding depth distribution in a capture

experiment is essential to establish a relationship between the number of reads, costs and efforts required to answer the question of interest [13]. Those difficulties in capture affect experimental results, yielding regions with different depths and introducing depth differences between samples [14].

Whole genome sequencing may not be a feasible alternative when the interest is on population studies. Focusing on exons reduces the complexity, drops costs and simplifies interpretation. On the other hand, inferring copy number is not trivial. Depending on the methodology and the exome definition, different capture protocols may also add distinct bias patterns to the results [15, 16, 17].

Recent population growth and weak purification selection contributed to increase the number of previously unknown and population-specific variants [18]. In fact, there is an increasing need for the development of population-specific markers [19]. It is already in use in SNP-array designs, allowing for greater resolution in population-specific variants [20, 21]. One serious issue is the fact that using a single reference genome as model for probes design may not account for population-specific variations and genomic rearrangements [22, 23, 24, 25, 26].

In this context, we describe depth variations in the coding regions of a subset of whole exome sequenced samples from the 1000 Genomes Project [27], aiming to analyze the impact of technical and population differences on depth.

Materials and Methods

We used public data from the 1000 Genomes Project Consortium (Additional File 2) to investigate the variation of depth [27]. We selected 120 unrelated individuals from

four populations (GBR - British in England and Scotland; ACB - African Caribbean from Barbados; YRI - Yoruba in Nigeria and JPT - Japanese in Tokyo) and three technically and temporally different phases. The subset of targets used in the capture reactions in phase I and II are intersections of the different technologies used by the Consortium, combined with regions corresponding to the Consensus Coding DNA Sequences (CCDS) gene list [31, 32, 33]. The targets in phase III are from NimbleGen EZ exome (version 1) and Agilent Sure Select (version 2).

To investigate population specific changes in depth, we selected 120 unrelated individuals from phase III and from 12 different populations: GBR - British in England and Schotland; IBS - Iberian population in Spain; ACB African Caribbeans in Barbados; GWD - Gambian in Western Divisions in The Gambia; ESN - Esan in Nigeria; MSL - Mende in Sierra Leone; YRI - Yoruba in Ibadan, Nigeria; JPT - Japanese in Tokyo, Japan; LWK - Luhya in Webuye, Kenya and TSI - Toscani in Italy.

We developed our own comparison method by combining multidimensional scaling with permutation tests [34]. We estimated the distribution of the statistic of interest under the null hypothesis (i.e., no differences between the groups) by randomly shuffling the group memberships of the observations. This allowed us to perform hypothesis testing without distributional assumptions [35]. Our strategy uses MDS to project the data onto lower dimensions, while preserving the distance between the data points. We determined the statistic of interest using these projections obtaining the Mahalanobis distance between the groups. The method empirically estimates

the null distribution of the distances between the groups using permutation. The strategy uses this distribution to assess the evidences of differences between them. The method is available as an R function at <https://goo.gl/v1tzdW> [36]. We used SAMtools depth (version 1.0) [37, 38] to estimate base-by-base depth over all the CCDS regions from the 22 autosomal chromosomes. We used the R statistical environment (version 3.1.1) [39] to conduct our analysis.

Results

Technical differences strongly affect depth distribution in WES

By using the projections obtained by multidimensional scaling, we identified that samples cluster together according to the study phase they are part of. Depth patterns suggest strong separation between samples from phase III (p-value < 0.001, for both comparisons) and those from phases I and II (p-value = 0.004). This suggests that changes in capture methodologies affect coverage patterns (Figure 1A).

On Figure 1B, we observe differences when comparing accumulated depth distribution. Low depth positions are not relevant for variant calling and may introduce false-positive results. However, sectors with extremely high depths adds no new information, as sequencing efforts results on a redundant interrogation of those regions.

Different populations present different depth distributions

Figure 1A shows that samples from phase III have a smaller dispersion, suggesting that the most recent capture definition used by the 1000 Genomes Consortium is more consistent in comparison with the other phases.

High dimension data 2D projection over the whole exome definition depth for this phase suggests a separation into two groups: one mainly composed by Negroid populations (GWD, ESN and MSL) and other mixed group with the remaining populations (Figure 2A). We calculated the distances based on the centroid of each population cluster (Additional File 1) and Figure 2B illustrates these distances among populations.

Discussion

Technical and population differences affect the observed coverage in WES. Theoretical depth estimation for these studies is complex and depends on a number of technical factors like library preparation and capture, generating the variability between the planned and observed depth [13]. Despite this issue, establishing the expected average depth is one of the most important parameters when designing sequencing experiments. It influences the number of unique fragments or pairs aligned to a reference genome with acceptable quality scores [10].

Our results indicate that technical differences in the capture phase proved to play the most important role while separating different samples depth over their captured

coding regions. This presents a challenge for large and long-term exome sequencing projects that expect to aggregate methodological advancements over time.

Different capture profiles between samples from different or mixed ethnicities may represent another concern. Our hypothesis is that this stratification is a result of population isolation and selection over time. This corroborates the fact that probes used for capturing require population-specific designs, like what is already in use for genotyping microarrays. This illustrates how important it is to take into account that some information may have been lost or causing misunderstanding while using a single human reference genome that may not account for population-specific common variations or genetic rearrangements. The same may occur to the probes for any targeted region [23, 26]. Some effort is in progress in order to understand the genome complex organization: we can cite population-based reference graphs, common unaligned sequences databases and advances made in minimizing off-target reads, decreasing costs and increasing sensibility [24, 25, 28, 29, 30]. In fact, the new human genome assembly, GRCh38 contains 178 alternative locus sequences, and more than 150 genes not previously represented, what is a short but important advance in this direction [26].

Conclusions

Our study indicates that WES depth is liable to technical and population differences, given that the initial step for a WES experiment is the capture of the target regions to be subsequently enriched and sequenced. This step is dependent on probe

hybridization, whose construction is based on a single reference genome that might not account for population-specific genetic variability. It is fundamental to account for such specificities, as they directly influence the efficiency and effectiveness in the capture phase, directly influencing the depth distribution.

Abbreviations

CCDS - Consensus Coding DNA Sequences; MDS - Multidimensional Scaling Analysis; WES - Whole Exome Sequencing.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

Research developed with the support of CENAPAD-SP (High Performance National Center - São Paulo, Brazil), project proj595 - UNICAMP / FINEP - MCT. Study supported by CAPES (Coordenadoria de Aperfeiçoamento de Pessoal).

References

1. Coffey, A.J., Kokocinski, F., Calafato, M.S., Scott, C.E., Palta, P., Drury, E., Joyce, C.J., LeProust, E.M., Harrow, J., Hunt, S., et al.: The gencode exome: sequencing the complete human exome. *European Journal of Human Genetics* 19(7), 827–831 (2011)
2. Gilissen, C., Hoischen, A., Brunner, H.G., Veltman, J.A.: Unlocking mendelian disease using exome sequencing. *genome* 11, 64 (2011)
3. Gilissen, C., Hoischen, A., Brunner, H.G., Veltman, J.A.: Disease gene identification strategies for exome sequencing. *European Journal of Human Genetics* 20(5), 490–497 (2012)
4. Clark, M.J., Chen, R., Lam, H.Y., Karczewski, K.J., Chen, R., Euskirchen, G., Butte, A.J., Snyder, M.: Performance comparison of exome DNA sequencing technologies. *Nature biotechnology* 29(10), 908–914 (2011)
5. Kiezun, A., Garimella, K., Do, R., Stitzel, N.O., Neale, B.M., McLaren, P.J., Gupta, N., Sklar, P., Sullivan, P.F., Moran, J.L., et al.: Exome sequencing and the genetic basis of complex traits. *Nature genetics* 44(6), 623–630 (2012)
6. Karakoc, E., Alkan, C., O’Roak, B.J., Dennis, M.Y., Vives, L., Mark, K., Rieder, M.J., Nickerson, D.A., Eichler, E.E.: Detection of structural variants and indels within exome data. *Nature methods* 9(2), 176–178 (2012)
7. Goldstein, D.B., Allen, A., Keebler, J., Margulies, E.H., Petrou, S., Petrovski, S., Sunyaev, S.: Sequencing studies in human genetics: design and interpretation. *Nature Reviews Genetics* 14(7), 460–470 (2013)
8. Do, R., Kathiresan, S., Abecasis, G.R.: Exome sequencing and complex disease: practical aspects of rare variant association studies. *Human molecular genetics* 21(R1), 1–9 (2012)
9. Parla, J.S., Iossifov, I., Grabill, I., Spector, M.S., Kramer, M., McCombie, W.R.: A comparative analysis of exome capture. *Genome Biol* 12(9), 97 (2011)

10. Sims, D., Sudbery, I., Illott, N.E., Heger, A., Ponting, C.P.: Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics* 15(2), 121–132 (2014)
11. Sampson, J., Jacobs, K., Yeager, M., Chanock, S., Chatterjee, N.: Efficient study design for next generation sequencing. *Genetic epidemiology* 35(4), 269–277 (2011)
12. Chilamakuri, C.S.R., Lorenz, S., Madoui, M.-A., Vodák, D., Sun, J., Hovig, E., Myklebost, O., Meza-Zepeda, L.A.: Performance comparison of four exome capture systems for deep sequencing. *BMC genomics* 15(1), 449 (2014)
13. Daley, T., Smith, A.D.: Predicting the molecular complexity of sequencing libraries. *Nature methods* 10(4), 325–327 (2013)
14. Veal, C.D., Freeman, P.J., Jacobs, K., Lancaster, O., Jamain, S., Leboyer, M., Albanes, D., Vaghela, R.R., Gut, I., Chanock, S.J., et al.: A mechanistic basis for amplification differences between samples and between genome regions. *BMC genomics* 13(1), 455 (2012)
15. Arcos-Burgos, M., Muenke, M.: Genetics of population isolates. *Clinical genetics* 61(4), 233–247 (2002)
16. Tennessen, J.A., O'Connor, T.D., Bamshad, M.J., Akey, J.M.: The promise and limitations of population exomics for human evolution studies. *Genome Biol* 12(127), 10–1186 (2011)
17. Auer, P.L., Johnsen, J.M., Johnson, A.D., Logsdon, B.A., Lange, L.A., Nalls, M.A., Zhang, G., Franceschini, N., Fox, K., Lange, E.M., et al.: Imputation of exome sequence variants into population-based samples and blood-cell-trait-associated loci in african americans: Nhlbi go exome sequencing project. *The American Journal of Human Genetics* 91(5), 794–808 (2012)
18. Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., et al.: Evolution and functional

impact of rare coding variation from deep sequencing of human exomes. *Science* 337(6090), 64–69 (2012)

19. Price, A.L., Patterson, N., Yu, F., Cox, D.R., Waliszewska, A., McDonald, G.J., Tandon, A., Schirmer, C., Neubauer, J., Bedoya, G., et al.: A genome-wide admixture map for latino populations. *The American Journal of Human Genetics* 80(6), 1024–1036 (2007)

20. Bryc, K., Auton, A., Nelson, M.R., Oksenberg, J.R., Hauser, S.L., Williams, S., Froment, A., Bodo, J.-M., Wambebe, C., Tishkoff, S.A., et al.: Genome-wide patterns of population structure and admixture in west africans and african americans. *Proceedings of the National Academy of Sciences* 107(2), 786–791 (2010)

21. International HapMap 3 Consortium, et al.: Integrating common and rare genetic variation in diverse human populations. *Nature* 467(7311), 52–58 (2010)

22. Dewey, F.E., Chen, R., Cordero, S.P., Ormond, K.E., Caleshu, C., Karczewski, K.J., Whirl-Carrillo, M., Wheeler, M.T., Dudley, J.T., Byrnes, J.K., et al.: Phased whole-genome genetic risk in a family quartet using a major allele reference sequence. *PLoS genetics* 7(9), 1002280 (2011)

23. Burgess, D.J.: Genomics: Getting personal and regional. *Nature Reviews Genetics* 12(11), 744–744 (2011)

24. Genovese, G., Handsaker, R.E., Li, H., Altemose, N., Lindgren, A.M., Chambert, K., Pasaniuc, B., Price, A.L., Reich, D., Morton, C.C., et al.: Using population admixture to help complete maps of the human genome. *Nature genetics* 45(4), 406–414 (2013)

25. Paten, B., Novak, A., Haussler, D.: Mapping to a reference genome structure. *arXiv preprint arXiv:1404.5010* (2014)

26. Church, D.M., Schneider, V.A., Steinberg, K.M., Schatz, M.C., Quinlan, A.R., Chin, C.-S., Kitts, P.A., Aken, B., Marth, G.T., Hoffman, M.M., et al.: Extending reference assembly models. *Genome biology* 16(1), 13 (2015)

27. Via Garc'ia, M., 1000 Genomes Project Consortium, et al.: An integrated map of genetic variation from 1,092 human genomes. *Nature*, 2012, vol. 491, p. 56-65 (2012)
28. Dilthey, A., Cox, C.J., Iqbal, Z., Nelson, M.R., McVean, G.: Improved genome inference in the mhc using a population reference graph. *bioRxiv*, 006973 (2014)
29. Marcus, S., Lee, H., Schatz, M.C.: Splitmem: a graphical algorithm for pan-genome analysis with suffix skips. *Bioinformatics* 30(24), 3476–3483 (2014)
30. Wang, C., Zhan, X., Bragg-Gresham, J., Kang, H.M., Stambolian, D., Chew, E.Y., Branham, K.E., Heckenlively, J., Study, T.F., Fulton, R., et al.: Ancestry estimation and control of population stratification for sequence-based association studies. *Nature genetics* 46(4), 409–415 (2014)
31. Pruitt, K.D., Harrow, J., Harte, R.A., Wallin, C., Diekhans, M., Maglott, D.R., Searle, S., Farrell, C.M., Loveland, J.E., Ruef, B.J., et al.: The consensus coding sequence (ccds) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome research* 19(7), 1316–1323 (2009)
32. Harte, R.A., Farrell, C.M., Loveland, J.E., Suner, M.-M., Wilming, L., Aken, B., Barrell, D., Frankish, A., Wallin, C., Searle, S., et al.: Tracking and coordinating an international curation effort for the ccds project. *Database* 2012, 008 (2012)
33. Farrell, C.M., O'Leary, N.A., Harte, R.A., Loveland, J.E., Wilming, L.G., Wallin, C., Diekhans, M., Barrell, D., Searle, S.M., Aken, B., et al.: Current status and new features of the consensus coding sequence database. *Nucleic acids research* 42(D1), 865–872 (2014)
34. Kruskal, J.B., Wish, M.: *Multidimensional scaling*. Sage (1978)
35. Hotelling, H.: *The generalization of Student's ratio*. Springer (1992)
36. Mahalanobis, P.C.: On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)* 2, 49–55 (1936)

37. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., et al.: The sequence alignment/map format and samtools. *Bioinformatics* 25(16), 2078–2079 (2009)
38. Li, H.: A statistical framework for snp calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27(21), 2987–2993 (2011)
39. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2014). R Foundation for Statistical Computing. <http://www.R-project.org/>

Figures

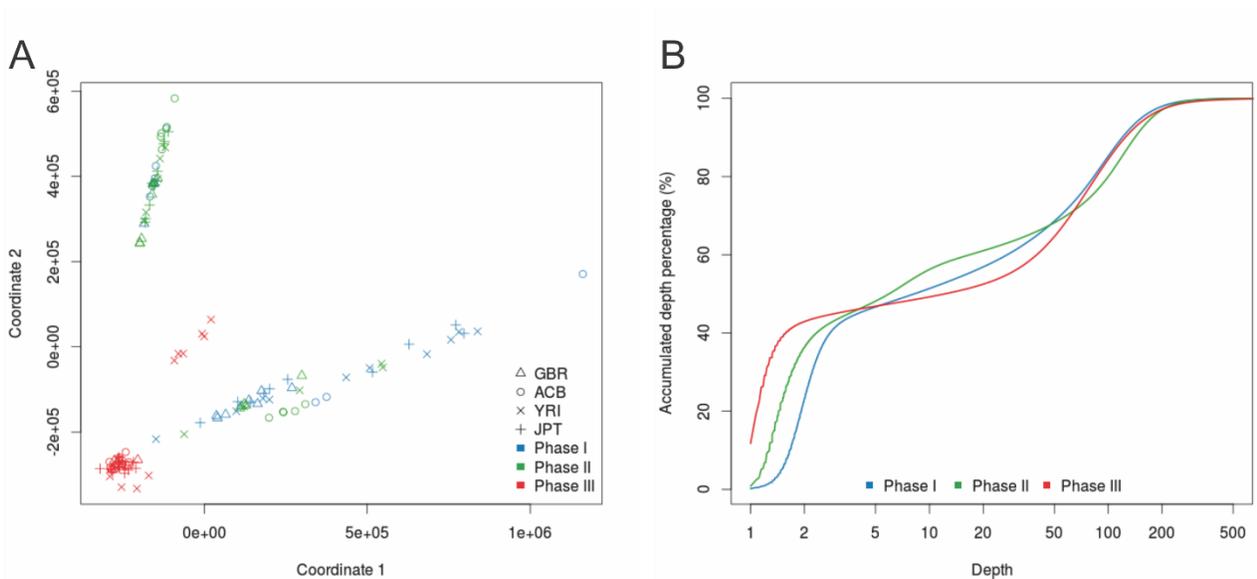


Figure 1. Technical differences affect WES depth: (A) Metric multidimensional scaling 2D projection for the whole exome depth of 120 samples: Notice that samples from phase III tend to cluster together in a different way of samples from phases I and II. This suggests that differences in the capture protocols deeply influence the final depth distribution over samples. (B) Accumulated depth distribution for the three 1000 Genomes exome sequencing phases: In average, the three sequencing phases have different depth distributions. Note that about 40% of the data have at most a 5x coverage. Low or extremely high depths do not add relevant information or may represent redundant information, affecting experiment costs.

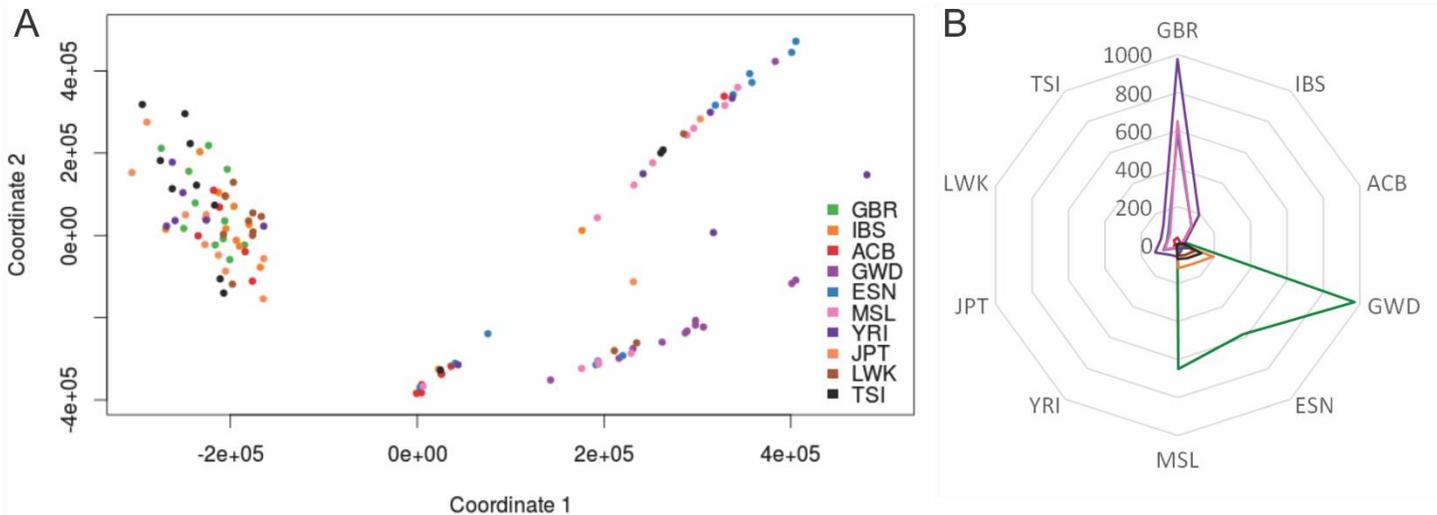


Figure 2. Metric multidimensional scaling and cluster's distances for the depth of 120 samples from phase III: (A) The depth MDS suggests a separation of our samples into two groups: one mainly composed by Negroid populations (GWD, ESN and MSL) and other mixed group with the remaining populations. (B) The radar plot presents these centroid distances for each of the populations. Each polygon represents the distances in (A) from one population's clusters to all the others. One possible explanation for those differences may be that older and conserved populations have aggregated much more variants to their genomes, which impacts in the capture reactions, and consequently in the depth obtained.

Additional Files (<https://goo.gl/tEftzz>)

Additional file 1 - Population distances based on the MDS from Figure 2B (xlsx) Spreadsheet containing the distances between population groups and p-values.

Additional file 2 - Samples information (xlsx) Spreadsheet containing information for the 1000 Genomes Project samples used in this study.

7. DISCUSSÃO

Dados os vários passos para análise de sequenciamento de alto desempenho apresentados até aqui, aplicamos e testamos uma forma eficaz de executar e analisar os resultados obtidos através do controle de qualidade pré e pós-alinhamento, bem como toda a aplicação do *workflow* desde o sequenciamento até a identificação de variantes consistentes e anotadas para filtragem posterior.

Nosso *pipeline* se baseou principalmente naquele proposto por [34, 35]. A escolha do alinhador BWA consistiu no fato de ser uma ferramenta amplamente utilizada pela comunidade científica e por apresentar uma alta sensibilidade e especificidade em comparação com uma ampla gama de alinhadores de código aberto [53]. O suíte de ferramentas do GATK possui várias funções que propiciam a interrogação consistente de regiões ou posições variantes no genoma. Ao se comparar o GATK com outros algoritmos [54, 55, 56], nota-se uma maior especificidade do GATK [35]. Este conjunto de ferramentas, escrito em Java, se destaca por disponibilizar vários módulos para análise, em uma arquitetura de processamento com *MapReduce* [57], onde pudemos nos focar nas ferramentas disponíveis para tratar com os desafios impostos aos dados de exomas humanos, propiciando aplicação nas etapas de chamada de variantes através da correção de erros sistemáticos no alinhamento de sequências e designação da qualidade das bases, sempre com o intuito de minimizar falso-positivos.

As associações entre variantes encontradas por sequenciamento e o fenótipo em estudo são complexas e possuem várias dificuldades como

apresentadas por [58, 59, 60]. Nosso *workflow* permitiu detectar comportamentos anômalos na reação de sequenciamento através do controle de qualidade pré alinhamento, evidenciando falhas pontuais em determinados ciclos, quedas de qualidade fora dos padrões esperados, bem como discrepâncias sistemáticas nas percentagens de GC e dos nucleotídeos em separado. Apesar de aparentes, estes achados não se apresentaram danosos ao ponto de inviabilizar as análises seguintes, e foram corrigidos mediante a etapas de realinhamento e recalibração de qualidade. Desafios parecidos foram encontrados por [61, 62, 63, 64, 65, 66].

Nosso experimento versa a respeito do alinhamento de sequências provenientes do exoma humano. A grande parte das sequências contudo não são alinhadas à região que corresponde ao alvo da definição do exoma utilizada pelo kit de captura utilizado, sugerindo possíveis falhas ou ineficiência na metodologia de captura destas sequências flanqueadas fora de sua definição. Apesar destes achados, nossa cobertura média foi de 59X, variando para cada experimento (Tabela 3), considerada satisfatória para os padrões esperados. Além disso, ao considerarmos apenas as regiões de intersecção entre o kit de captura estendido que utilizamos e o restrito, temos que para esta região a cobertura é de 66X, apresentando melhor cobertura que na versão estendida por si só (Figuras 10 e 11). Ao traçarmos um perfil das amostras sequenciadas, temos que para o protocolo de preparo das bibliotecas e sequenciamento, sequenciar em torno de cem milhões de sequências é plausível para se obter uma cobertura superior de 50 vezes em mais de 50% da definição do exoma, dentro de padrões de qualidade e cobertura em sequenciamento expostos por [67] em vários experimentos em sequenciamento.

Baseado no número inicial de sequências utilizadas para as análises, também obtivemos um perfil das demandas de armazenamento, e tempo de processamento para as análises.

A maior parte do total das variantes encontradas não estavam nos exons, evidenciando que as baixas coberturas (<10x) nos arredores das regiões propriamente capturadas têm uma influência muito maior na introdução de variantes que são possivelmente errôneas ou que não possuem uma cobertura adequada para classificá-las como causais no fenótipo (Tabelas 4 e 5). Ao considerar esse fato, poderemos modificar nossa abordagem de filtragem de variantes encontradas, podendo prender nossa atenção em variantes bem cobertas e com alta qualidade de alinhamento em regiões codificantes e possivelmente, a sua extensão mais próxima. De fato, a detecção de variantes que são na verdade falsos-positivos é de extrema importância em *WES* para uma associação genótipo-fenótipo [68].

Para as etapas intermediárias de alinhamento, identificamos a necessidade real de criação de um banco de variantes genômicas da população brasileira, empregada para as etapas de realinhamento ao redor de *indels* e recalibração de qualidade. Neste trabalho já o fizemos com bancos como dbSNP [52]. De fato, vários projetos visam a criação de grandes bancos de dados de variantes genômicas [48, 69, 70, 71]. Acreditamos que, ao utilizar um banco criado a partir de nossa população, poderemos ter um maior poder de identificação de regiões passíveis de intervenção de realinhamento e recalibração de qualidade, eventualmente refletindo nas variantes interrogadas e na associação de variantes

aos fenótipos, excluindo as que pela sua frequência na população normal, não são tomadas como patogênicas.

Diferenças técnicas e populacionais afetam a distribuição da cobertura em exomas do Projeto 1000 Genomas. Devido à complexidade das etapas de preparação das bibliotecas, comumente são observadas grandes variações entre os valores esperados e obtidos [72]. Diferenças técnicas inseridas na fase de captura são mais impactantes, elucidando o grande desafio em se integrar dados obtidos por diferentes metodologias. Contudo, para uma mesma fase, podemos observar uma segregação populacional em dois grupos: um majoritariamente constituído por uma população com background étnico africano conservado e outro com as demais etnias consideradas. Julgamos que esse achado reflita a diferença genômica entre as populações consideradas que não são necessariamente contempladas pelo uso de um único genoma de referência para criação de alvos para captura do exoma.

Como pontos positivos para este estudo, temos a aplicação de um *workflow* a dados de sequenciamento de alto desempenho de exomas humanos que leva em consideração correções no alinhamento local ao redor de inserções e deleções, bem como a recalibração da qualidade das bases interrogadas na reação de sequenciamento, culminando com uma lista anotada de variantes passíveis de filtragem pelos pesquisadores responsáveis. As ferramentas utilizadas são todas softwares livres, executadas em sistema operacional Linux, todas distribuídas por grupos de pesquisa renomados, amplamente utilizadas pela comunidade científica e em constante manutenção e atualização. Como fragilidades no dado estudo,

temos a falta de um banco de variantes da população brasileira, que poderia contribuir para eliminação de falsos-positivos e descoberta de variantes nos indivíduos da população. Fragilidades relacionadas a técnica de captura do exoma também podem influenciar os resultados finais das análises.

Temos assim que grande avanço foi realizado nas etapas de análises dos dados de exomas humanos, possibilitando aos pesquisadores responsáveis estabelecer uma real associação entre o genótipo encontrado e o fenótipo apresentado pelos indivíduos sequenciados.

8. CONCLUSÃO

Ao concluirmos nosso trabalho, temos um *workflow* consistente e robusto para interrogação de variantes, que levou em consideração a qualidade das sequências fornecidas pelo sequenciador, o alinhamento contra o genoma, realinhamento ao redor de regiões sabidamente conhecidas como portadoras de variações, recalibração da qualidade e anotação.

Além disso, fomos capazes de anotar as variantes encontradas com o intuito de facilitar a aplicação dos filtros biológicos e posterior análise, colocando em evidência, por exemplo, variantes em regiões possivelmente codificantes e aquelas que são tidas como patogênicas em potencial.

Temos evidências que a cobertura obtida pelo sequenciamento do exoma foi influenciada por diferenças técnicas e populacionais, refletindo que a complexidade do genoma pode interferir na reação de captura das sequências, influenciando a efetividade da técnica empregada.

REFERÊNCIAS

1. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, et al. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*. 2005;309:1728–1732.
2. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*. 2005;437:376–380.
3. Kurotaki N, Imaizumi K, Harada N. Haploinsufficiency of NSD1 causes Sotos syndrome. *Nat Genet*. 2002;30:365–366.
4. Kerem B, Rommens JM, Buchanan JA. Identification of the cystic fibrosis gene: genetic analysis. *Science*. 1989;245:1073–1080.
5. Lander ES, Botstein D. Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science*. 1987;236:1567–1570.
6. Vissers LE, Veltman JA, Kessel VAG, Brunner HG. Identification of disease genes by whole genome CGH arrays. *Hum Mol Genet*. 2005;14(Spec No. 2):215.
7. Kryukov GV, Pennacchio LA, Sunyaev SR. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am J Hum Genet*. 2007;80:727–739.

8. Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, et al. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet.* 2008;4(e1000083).
9. Fay JC, Wyckoff GJ, Wu CI. Positive and negative selection on the human genome. *Genetics.* 2001;158:1227–1234.
10. Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol.* 2009;27:182–189.
11. Illumina. Nextera Exome Enrichment Kit. Illumina, Inc.; 2012. 770-2012-028.
12. Need AC, Shashi V, Hitomi Y, Schoch K, Shianna KV, McDonald MT, et al. Clinical application of exome sequencing in undiagnosed genetic conditions. *J Med Genet.* 2012.
13. Do R, Kathiresan S, Abecasis GR. Exome sequencing and complex disease: practical aspects of rare variant association studies. *Human Molecular Genetics.* 2012.
14. Yang Y, Muzny DM, Reid JG, Bainbridge MN, Eng CM. Clinical WholeExome Sequencing for the Diagnosis of Mendelian Disorders. *The new england journal of medicine.* 2013.
15. Choi M, Scholl UI, Ji W, T Liu and Tikhonova I, Zumbo P, Nayir A, et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci U S A.* 2009;106:19096–19101.

16. Gilissen C, Hoischen A, Brunner HG, Veltman JA. Disease gene identification strategies for exome sequencing. *European Journal of Human Genetics*. 2012;20:490–497.
17. Torres FR, Santos NF, Secolin R, Gonsales MC, Kobayashi E, Sardinha LAC, et al. A locus for familial mesial temporal lobe epilepsy mapped on chromosome 18p. 56th Annual Meeting of The American Society of Human Genetics. 2006; p. 288–288.
18. Santos NF, Secolin R, Brandão-Almeida IL, Silva MS, Torres FR, Tsuneda SS, et al. A new candidate locus for bilateral perisylvian polymicrogyria mapped on chromosome Xq27. *American Journal of Medical Genetics Part A*. 2008;146A(9):1151–1157.
19. Torres FR, Montenegro MA, Marques-de Faria AP, Guerreiro MM, Cendes F, Lopes-Cendes I. Mutation screening in a cohort of patients with lissencephaly and subcortical band heterotopia. *Neurology*. 2004;62(5):799–802.
20. Tsuneda SS, Souza-Kols DA, Torres FR, Secolin R, Maurer-Morelli CV, Guerreiro MM, et al. Estudo genético de famílias com polimicrogria perisylviana bilateral congênita: screening de mutações no gene SRPX2. XXXII Congresso Brasileiro de Epilepsia / XVIII Jornada Brasileira de Neurofisiologia Clínica. 2008.
21. Lopes-Cendes I, Scheffer IE, Berkovic SF, Rousseau M, Andermann E, Rouleau GA. A New Locus for Generalized Epilepsy with Febrile Seizures Plus Maps to Chromosome 2. *The American Journal of Human Genetics*. 2000;66(2):698 – 701.

22. Maurer-Morelli CV, Secolin R, Morita ME, Domingues RR, Marchesini RB, Santos NF, et al. A locus identified on chromosome 18p11.31 is associated with hippocampal abnormalities in a family with mesial temporal lobe epilepsy. *Frontiers in Neurology*. 2012;3(124).
23. Baulac S, Baulac M. Advances on the genetics of Mendelian idiopathic epilepsies. *Clin Lab Med*. 2010;30:911–29.
24. Steinlein OK. Channelopathies can cause epilepsy in man. *Eur J Pain*. 2002;6:Suppl A:27–34.
25. Barkovich AJ, Guerrini R, Kuzniecky RI, Jackson GD, Dobyns WB. A developmental and genetic classification for malformations of cortical development: update 2012. *Brain*. 2012;135(5):1348–1369.
26. Ning Z, Cox AJ, Mullikin JC. SSAHA: a fast search method for large DNA databases. *Genome research*. 2001;11(10):1725–1729.
27. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome research*. 2008;18(11):1851–1858.
28. Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, et al. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*. 2009;25(15):1966–1967.
29. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*. 2009;25(14):1754–1760.
30. Chen Y, Souaiaia T, Chen T. PerM: efficient mapping of short sequencing reads with periodic full sensitive spaced seeds. *Bioinformatics*. 2009;25(19):2514–2521.

31. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics*. 2009;25:1754–1760.
32. Matthew R, Thomas L, Mehmet K. Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics*. 2011 August;27(20):2790–2796.
33. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*. 2009;25:2078–9.
34. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernysky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20:1297– 303.
35. DePristo M, Banks E, Poplin R, Garimella K, Maguire J, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*. 2011;43:491–498.
36. Durbin RM, Abecasis G, Altshuler D, Auton A, Brooks L, Gibbs R, et al. A map of human genome variation from population-scale sequencing. *Nature*. 2010;467:1061–1073.
37. Durtschi, Jacob, et al. VarBin, a novel method for classifying true and false positive variants in NGS data. *BMC bioinformatics* 14.Suppl 13 (2013): S2.
38. Fuentes Fajardo, Karin V., et al. Detecting false-positive signals in exome sequencing. *Human mutation* 33.4 (2012): 609-613.
39. Chan EY. Next-generation sequencing methods: impact of sequencing accuracy on SNP discovery. *Methods Mol Biol*. 2009;578:95–111.

40. Garner C. Confounded by sequencing depth in association studies of rare alleles. *Genet Epidemiol.* 2011.
41. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. *Nat Meth.* 2010;7:248–249.
42. Kumar P, Henikoff S, Pauline CNG. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protocols.* 2009.
43. Souza W, Carvalho B. Rqc: Quality Control Tool for High-Throughput Sequencing Data. R package version 1.2.0; 2014.
44. BABRAHAM BIOINFORMATICS. A quality control tool for high throughput sequence data; 2013.
<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.
45. Cendes IL. Biorrepositório: Estudos de Genética Molecular em Doenças Neuropsiquiátricas – FASE – I; 2013. Projeto de Pesquisa.
46. Bravo HC, Irizarry RA. Model-Based Quality Assessment and Base-Calling for Second-Generation Sequencing Data. *Biometrics.* 2010;66(3):665–674.
47. Ruffalo M, LaFramboise T, Koyuturk M. Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics.* 2011;27(20):2790–2796.
48. 1000 GENOMES PROJECT CONSORTIUM. A map of human genome variation from population-scale sequencing. *Nature.* 2010;467.

49. Li H, Durbin R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*. 2010;26(5):589–595.
50. Broad Institute. Picard; 2015. Available from: <http://broadinstitute.github.io/picard/>.
51. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics*. 2010;26(16):2069–2070.
52. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. *Nucleic acids research*. 2001;29(1):308–311.
53. Li, Heng. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:1303.3997 (2013).
54. Gurtowski J, Schatz MC, Langmead B. Genotyping in the cloud with crossbow. *Current Protocols in Bioinformatics*. 2012;p. 15–3.
55. O’Rawe, Jason, et al. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome med* 5.3 (2013): 28.
56. Liu, Xiangtao, et al. Variant callers for next-generation sequencing data: a comparison study. *PLoS One* 8.9 (2013): e75619.
57. Yang, Hung-chih, et al. Map-reduce-merge: simplified relational data processing on large clusters. *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*. ACM, 2007.

58. Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proceedings of the National Academy of Sciences*. 2009;106(45):19096–19101.
59. Montenegro G, Powell E, Huang J, Speziani F, Edwards YJ, Beecham G, et al. Exome sequencing allows for rapid gene identification in a Charcot-Marie-Tooth family. *Annals of neurology*. 2011;69(3):464–470.
60. O’Rawe J, Jiang T, Sun G, Wu Y, Wang W, Hu J, et al. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome med*. 2013;5(3):28.
61. Li M, Nordborg M, Li LM. Adjust quality scores from alignment and improve sequencing accuracy. *Nucleic acids research*. 2004;32(17):5183– 5191.
62. Brockman W, Alvarez P, Young S, Garber M, Giannoukos G, Lee WL, et al. Quality scores and SNP detection in sequencing-by-synthesis systems. *Genome research*. 2008;18(5):763–770.
63. Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K, et al. SNP detection for massively parallel whole-genome resequencing. *Genome research*. 2009;19(6):1124–1132.
64. Walsh T, Shahin H, Elkan-Miller T, Lee MK, Thornton AM, Roeb W, et al. Whole exome sequencing and homozygosity mapping identify mutation in the cell polarity protein GPM2 as the cause of nonsyndromic hearing loss DFNB82. *The American Journal of Human Genetics*. 2010;87(1):90–94.

65. Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ, et al. De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature*. 2012;485(7397):237–241.
66. Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*. 2013;493(7431):216–220.
67. Sims D, Sudbery I, Illott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics*. 2014;15(2):121–132.
68. Fuentes Fajardo KV, Adams D, Mason CE, Sincan M, Tiffit C, Toro C, et al. Detecting false-positive signals in exome sequencing. *Human Mutation*. 2012;33(4):609–613.
69. Gibbs RA, Belmont JW, Hardenbol P, Willis TD, Yu F, Yang H, et al. The international HapMap project. *Nature*. 2003;426(6968):789–796.
70. Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*. 2001;409(6822):928–933.
71. Hinds DA, Stuve LL, Nilsen GB, Halperin E, Eskin E, Ballinger DG, et al. Whole-genome patterns of common DNA variation in three human populations. *Science*. 2005;307(5712):1072–1079.
72. Daley, Timothy, and Andrew D. Smith. Predicting the molecular complexity of sequencing libraries. *Nature methods* 10.4 (2013): 325-327.

ANEXOS

Anexo 1. Parecer do comitê de ética em pesquisa

FACULDADE DE CIÊNCIAS
MÉDICAS - UNICAMP
(CAMPUS CAMPINAS)



PARECER CONSUBSTANCIADO DO CEP

DADOS DO PROJETO DE PESQUISA

Título da Pesquisa: BIORREPOSITÓRIO
ESTUDOS DE GENÉTICA MOLECULAR EM DOENÇAS NEUROPSIQUIÁTRICAS
FASE I

Pesquisador: Iscia Teresinha Lopes Cendes

Área Temática: Área 1. Genética Humana.
(Trata-se de pesquisa envolvendo genética humana não contemplada acima.);

Versão: 2

CAAE: 12112913.3.0000.5404

Instituição Proponente: Hospital de Clínicas da UNICAMP

Patrocinador Principal: Financiamento Próprio

DADOS DO PARECER

Número do Parecer: 257.020

Data da Relatoria: 12/04/2013

Apresentação do Projeto:

Trata-se de um projeto para implantação de biorepositório de doenças neuro-psiquiátricas e casos-controle. O estudo prevê recrutamento e coleta de 700 pacientes e 300 indivíduos controle.

O presente projeto pretende estudar os aspectos moleculares das seguintes doenças: epilepsias, malformações do desenvolvimento cortical, coreias, ataxias, paraparesias espásticas, distonias, transtorno afetivo bipolar, esquizofrenia, doenças musculares, doenças mitocondriais, doença de Parkinson, acidente vascular cerebral e demências. O projeto está dividido em sub-projetos, com a descrição detalhada das estratégias que serão utilizadas para cada uma dessas doenças.

Serão utilizadas diversas técnicas de biologia molecular para identificação de mutações conhecidas ou novas, como PCR, sequenciamento convencional e de terceira geração e análises de bioinformática. No Subprojeto 1 (Epilepsias e Malformações do Desenvolvimento Cortical), serão avaliadas mutações através da implantação da tecnologia de sequenciamento, baseada em equipamento de terceira geração e um novo sistema de detecção eletrônico, sistema Ion Torrent®. No subprojeto 2 será realizada a captura e o sequenciamento do exoma em amostras de DNA de

Endereço: Rua Tessália Vieira de Camargo, 126

Bairro: Barão Geraldo

CEP: 13.083-887

UF: SP

Município: CAMPINAS

Telefone: (19)3521-8936

Fax: (19)3521-7187

E-mail: cep@fcm.unicamp.br

FACULDADE DE CIÊNCIAS
MÉDICAS - UNICAMP
(CAMPUS CAMPINAS)



material biológico. Estão presentes todos os termos de apresentação obrigatória previstos pela Resolução 196/96 e complementares, assim como o "Regulamento do Biorepositório de Doenças Neuropsiquiátricas!".

Recomendações:

Nada a declarar.

Conclusões ou Pendências e Lista de Inadequações:

Foram acrescentadas ao projeto principal as informações sobre o recrutamento dos voluntários do grupo controle, com priorização inicial para membros da família de pacientes, porém não portadores das doenças. As amostras serão coletadas nos ambulatórios de Genética e Neurologia HC-Unicamp e Hemocentro-UNICAMP.

Anexo 2. Termos de consentimento e requisição de exames empregados



Universidade Estadual de Campinas
Departamento de Genética Médica

TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO PARA PESQUISA

Título do projeto: Estudos de genética molecular em doenças neuropsiquiátricas – Fase I

Investigador principal: Dra. Iscia Lopes Cendes (tel: 19 3521 8907)

OBJETIVO DA PESQUISA:

Eu entendo que fui convidado (a) a participar em um projeto de pesquisa envolvendo pacientes e famílias de indivíduos com

- Epilepsias
- Malformações corticais
- Coreias
- Ataxias
- Paraparesias
- Distonias
- Transtorno afetivo bipolar
- Esquizofrenia
- Doenças musculares
- Doenças mitocondriais
- Parkinson
- Acidente Vascular Cerebral
- Demências
- Grupo Controle

O objetivo do estudo é identificar a alteração genética que causa a doença. Isso poderá melhorar o diagnóstico da doença (se a alteração for encontrada) e poderá levar a um melhor tratamento no futuro. No entanto, sei que muito provavelmente o meu tratamento não será modificado com a participação nesse estudo.

PROCEDIMENTO:

Eu entendo que se concordar em participar desse estudo, os pesquisadores participantes farão perguntas a respeito dos meus antecedentes médicos e familiares. Eu serei submetido a um exame físico neurológico e ou psiquiátrico para estabelecer meu estado clínico. Além disso, poderei ser submetido a um eletroencefalograma (EEG), ou a uma eletromiografia (EMG) e talvez uma tomografia computadorizada ou uma ressonância magnética de crânio. Uma amostra de sangue venoso será colhida (20 a 30 ml, o equivalente a duas colheres de sopa). Hospitalização não será necessária. Os procedimentos mencionados acima, com exceção da coleta da amostra de sangue, fazem parte dos cuidados médicos de rotina para um paciente com doença neuropsiquiátrica.

Rubrica do pesquisador	Rubrica do sujeito de pesquisa ou seu representante
------------------------	---



RISCO E DESCONFORTO:

Uma coleta de 20 a 30 ml de sangue venoso será efetuada. Os riscos associados a esse procedimento são mínimos, podendo ocorrer dor e manchas roxas (equimoses) no local da coleta do sangue. O desconforto será mínimo, pois se trata de uma coleta de sangue geralmente da veia do braço que será realizada por profissional treinado e habilitado para realizar esse procedimento.

VANTAGENS:

Eu entendo que não obterei nenhuma vantagem direta com a minha participação nesse estudo e que o meu diagnóstico e o meu tratamento provavelmente não serão modificados. Contudo, os resultados desse estudo podem, a longo prazo, oferecer vantagens para os indivíduos com doenças neuropsiquiátricas e suas famílias, possibilitando um melhor diagnóstico e tratamento mais adequado. É importante notar que o diagnóstico pré-sintomático não faz parte dessa pesquisa, mas se eu desejar obter orientação genética, ela será oferecido no ambulatório de neurogenética do Hospital de Clínicas (HC) da Universidade Estadual de Campinas (UNICAMP), tel. (19) 35217754.

SIGILO:

Eu entendo que toda informação médica, mas não os resultados dos testes genéticos decorrentes desse projeto de pesquisa, farão parte do meu prontuário médico e serão submetidos aos regulamentos do HC- UNICAMP referentes ao sigilo da informação médica. Se os resultados ou informações fornecidas resultarem em publicação científica, nenhum nome será utilizado.

FORNECIMENTO DE INFORMAÇÃO ADICIONAL:

Eu entendo que posso requisitar informações adicionais relativas ao estudo a qualquer momento. A Dra. Iscia Lopes Cendes, tel (19) 35217754 estará disponível para responder minhas questões e preocupações. Em caso de recurso, dúvidas ou reclamações posso contatar a secretaria do Comitê de Ética em Pesquisa, tel. (19) 3521-8936.

RECUSA OU DESCONTINUAÇÃO DA PARTICIPAÇÃO:

Eu entendo que a minha participação é voluntária e que eu posso me recusar a participar ou retirar meu consentimento e interromper a minha participação no estudo a qualquer momento (incluindo a retirada da amostra de sangue) sem comprometer os cuidados médicos que recebo atualmente ou receberei no futuro no HC- UNICAMP. Eu reconheço também que a Dra. Iscia Lopes Cendes pode interromper a minha participação nesse estudo a qualquer momento que julgar apropriado.

Rubrica do pesquisador	Rubrica do sujeito de pesquisa ou seu representante
------------------------	---



Universidade Estadual de Campinas
Departamento de Genética Médica

Eu confirmo que o(a) Dr(a) _____
me explicou o objetivo do estudo, os procedimentos aos quais serei submetido, os riscos, os desconforto e as possíveis vantagens advindas desse projeto de pesquisa. Eu li e compreendi (ou me foi explicado) esse termo de consentimento e estou de pleno acordo em participar desse estudo. Além disso, informo que:

- Autorizo o armazenamento do material biológico e dispense a necessidade de novo consentimento em caso de seu uso em outras pesquisas.
- Autorizo o armazenamento do material biológico e desejo ser consultado para consentimento em caso de seu uso em outras pesquisas.
- NÃO** autorizo o armazenamento do material biológico, devendo o mesmo ser descartado após o encerramento de minha participação nessa pesquisa.

Nome do participante ou responsável

Assinatura do participante ou responsável

data

Nome da testemunha

Assinatura da testemunha

data

RESPONSABILIDADE DO PESQUISADOR:

Eu expliquei a _____
o objetivo do estudo, os procedimentos requeridos e os possíveis riscos e vantagens que poderão advir do estudo, usando o melhor do meu conhecimento. Eu me comprometo a fornecer uma cópia desse termo de consentimento ao participante ou responsável. Caso uma nova pesquisa seja realizada utilizando o material biológico coletado e armazenado por ocasião dessa pesquisa, comprometo-me a submeter e aguardar o parecer do sistema CEP/CONEP para sua utilização.

Nome do pesquisador ou associado

Assinatura do pesquisador ou associado

data

ANEXO 3. Lista das mostras tratadas neste trabalho.

Tabela 5. Associação entre amostras e projetos de pesquisa

Amostra	Experimento
19313_1	Proj1
4411_1	Proj1
54112_1	Proj1
74212_1	Proj1
p13_1	Proj1
p16_1	Proj1
p19_1	Proj1
p20_1	Proj1
19313	Proj1
4411	Proj1
54112	Proj1
74212	Proj1
p13	Proj1
p16	Proj1
p19	Proj1
p20	Proj1
586	Proj2
587	Proj2
588	Proj2
920	Proj2
931	Proj2
938	Proj2
942	Proj2
943	Proj2
F1-1	Proj3 – F1
F1-2	Proj3 – F1
F1-3	Proj3 – F1
F1-4	Proj3 – F1
F2-5	Proj3 – F2
F2-6	Proj3 – F2
F2-7	Proj3 – F2
F2-8	Proj3 – F2
P4-1	Proj4
P4-2	Proj4
P4-3	Proj4
P4-4	Proj4
102214	Proj5
64814	Proj5
93814	Proj5
93914	Proj5