Universidade do Rio Grande do Sul
Instituto de Biociências
Programa de Pós-Graduação em Genética e Biologia Molecular

SEQUENCIAMENTO DE NOVA GERAÇÃO: EXPLORANDO APLICAÇÕES CLÍNICAS DE DADOS DE *TARGETED GENE PANEL* E *WHOLE EXOME SEQUENCING*

Delva Pereira Leão

Dissertação submetida ao Programa de Pós-Graduação em Genética e Biologia Molecular da Universidade Federal do Rio Grande do Sul como requisito parcial para a obtenção do grau de Mestre em Genética e Biologia Molecular.

Orientadora: Profa. Dra. Ursula da Silveira Matte

Porto Alegre, abril de 2017

Este trabalho foi desenvolvido no Centro de Terapia Gênica do Centro de Pesquisa Experimental do Hospital de Clínicas de Porto Alegre em parceria com o Serviço de Genética Médica do Hospital de Clínicas de Porto Alegre. Esta pesquisa foi apoiada pelo Programa de Pós-Graduação em Genética e Biologia Molecular (PPGBM-UFRGS) e pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) através da concessão da bolsa de mestrado.

DEDICATÓRIA

"Existe um único ponto de vital importância: é a forte vontade de crescer e expandir custe o que custar. Esta atitude é fundamental. O pior pensamento a seu próprio respeito é pensar 'Não tenho capacidade'. Pense assim: 'Eu também sou um ser humano. Se aquela pessoa está fazendo, eu também serei capaz de fazer'. Aquele que nunca desiste, com uma forte determinação para realizar o seu trabalho – mesmo cometendo falhas ou sendo ridicularizado pelos outros – certamente crescerá bastante. Eu próprio trabalho com essa atitude. Entretanto, aquelas que desistem após o primeiro fracasso não servem, de fato, para o trabalho. Há um provérbio que diz: 'A resignação é importante'. Em algumas circunstâncias isso é verdadeiro mas, neste caso, a não resignação é fundamental. Em resumo, devemos desistir quando a causa não for boa, pelo contrário, ter força de vontade para as boas causas".

Atitude Mental. Meishu Sama – Alicerce do Paraíso, v.4

À todos aqueles que buscam tornar seus sonhos realidade.

AGRADECIMENTOS

Ao Supremo Deus e ao Messias Meishu Sama, por me permitir e ensinar a viver com gratidão. Por essa etapa de minha vida ter sido um momento de intenso aprendizado, recheado de amizades e de constantes desafios que me tornaram a Delva de hoje e que servem de molde para a Delva de amanhã.

Aos meus pais, pois "sem a existência de vocês, não existiriam, neste mundo, meu corpo e minha alma". Mãe, alicerce de minha existência, te amo e te dedico esta vitória. Pai, não há palavras para definir o que é ter você junto a mim neste momento. À vocês, meu eterno amor e gratidão.

Às minhas avós Doroty e Delva. Vó Beti: seu amor, cuidado e educação fizeram de mim o que sou hoje. Sei que você fez diferente comigo e sei que me ama como tua filha. Eu também te amo como minha mãe, muito obrigada! Vó Delva: sinto seu amor pulsando em meu sangue, carrego seu nome e sua força. O amor que nos une é mais forte que a distância que nos separa, te amo e obrigada por me proteger sempre.

À minha irmã Livia. Emociona-me saber que você não desiste dos teus sonhos e luta diariamente para torná-los realidade. Você é um exemplo pra mim: menina doce, mulher guerreira, riso fácil e carinhos não raros. Sou muito grata por ter em ti uma amiga, te amo.

Ao meu irmão Vitor. Por ter me feito sentir que sempre estivemos juntos. Creio que, de fato, estávamos. Obrigada por todo cuidado com o pai, por ser um homem correto e honesto. Tu me provou que os laços de família são eternos e removem quaisquer obstáculos, amo você.

À minha grande família! Meu ninho, meu lar, meu chão. Não há como descrever como é difícil estar longe de vocês todos estes anos. Minha eterna gratidão por vocês existirem, por nos amarmos e estarmos sempre juntos. Em especial, aos meus amados primos: Bia,

Mari, Neto, Isis, Julia, Rafa, Leo, Junior e Zé. Sem vocês nada disso seria igual! Nunca desistam dos seus sonhos.

Ao meu afilhado João Fernando, meus sonhos custaram deixar de te ver crescer. Saiba que valorizo cada momento que passamos juntos. Daqui a alguns anos lembre-se: a vida deve ser vivida e tudo vai valer a pena. Te amo meu pequeno!

À minha querida orientadora Ursula Matte. Obrigada por ter me aceitado como sua aluna e acreditado em mim quando nem eu mesma acreditei. Obrigada por ter me dado tanta responsabilidade, por ter sido compreensiva quando precisava e me impulsionado para a frente, sempre. Você foi essencial para meu amadurecimento profissional. Tenho em ti um espelho, tenho orgulho de poder ter convivido com alguém que buscou e realizou grande parte dos seus sonhos.

Aos meus amigos! Sem vocês essa jornada que foi viver em Porto Alegre não teria sido a mesma.
À Dani, minha nega, por tudo o que passamos juntas desde que chegamos nesta cidade. A vida ainda nos reserva muitas aventuras! Te amo preta <3
À Cle, minha surpresa de Poa. Teu modo de encarar a vida é uma inspiração à qualquer um. Te admiro muito, obrigada pelo presente que é a tua amizade.
À Marina, que pessoa forte! Muito obrigada por todos os cafés, aprendizados do Ion e conselhos. Mas acima de tudo, obrigada por sua amizade, por abrir as portas da tua família para mim e por me dar tanto carinho sincero.
À Patilu, por me aconselhar sempre, oferecer sua amizade, casa e carinho. Te admiro por ter batalhado tanto para conquistar o que tens e por ainda ter vontade de lutar por muito mais. Muito obrigada por tudo!
À Iáskara, meu exemplo de pessoa com forte vontade de crescer e expandir custe o que custar. Tua garra, tua força e teu amor me ajudaram tantas vezes...mesmo sem você saber. Sou muito grata por poder conviver contigo, muito obrigada!
Ao Bruno e Luana, querido casal. Obrigada por me darem tanto carinho e me escutarem quando precisei. Estarei sempre com vocês.

Ao Jussiê, querido amigo. É incrível como somos parecidos, talvez por isso sempre foi fácil conversar, desabafar e também te ouvir. Sei que estará sempre ao meu lado e eu estarei ao seu. Muito obrigada!
À Julia, muito obrigada por tua amizade sincera. Admiro a pureza de sentimento que tens e por me enxergar pelo ser humano que sou. Muito obrigada minha amiga amada.
Aos meus amigos e colegas do CTG, CPE e PPGBM: ia ficar difícil escrever sobre cada um. Obrigada pelos risos, cafés, conselhos e discussões científicas. Certamente foi um prazer contar com vocês nesta jornada. Em especial: Everaldo, Elmo, Aninha, Diana, Igor, Michael, Hugo, Jeferson, Pati K, Presuntinho, Gabi Pasqualim, Grazi, Mónica, Talita, Natan, Rose, Édina, Raíssa, Suelen, Leandro, Eduardo, Douglas, Santiago, Fernanda Souza, Mariana Mendoza, Zu, Perpétua, Luiza M e todos que por ventura eu tenha esquecido.

Aos meus amigos do Johrei! Jovens messiânicos, obrigada por me fazerem acreditar em futuro onde o ser humano será pleno de prosperidade, saúde e paz. Muito obrigada pela parceria nas dedicações, risadas e amizade.

Por fim, mas não menos importante, agradeço à vida. Que pulsa, que passa e que se sente. Vida que vivo intensamente. Vida que me leva para os caminhos que devo trilhar. À espera do meu próximo desafio, me lanço ao desconhecido com a certeza de que vai valer à pena, como tudo o que vivi valeu, como materializado por esta conquista.

SUMÁRIO

LISTA DE ABREVIATURAS

DNA – Ácido Desoxirribonucleico (***D****eoxyribo****n****ucleic* ***A****cid*)

ExAC – Exome **A**ggregation **C**onsortium

NGS – Sequenciamento de Nova Geração (***N****ext-****g****eneration* ***S****equencing*)

PCR – Reação em Cadeia da Polimerase (***P****olymerase* ***C****hain* ***R****eaction*)

RNA – Ácido Ribonucleico (***R****ibo****n****ucleic* ***A****cid*)

SNP – Polimorfismo de Nucleotídeo Único

sSNP - Polimorfismo de Nucleotídeo Único Sinônimo

TGP – Sequenciamento de painel de genes (***T****argeted* ***G****ene* ***P****anel)*

VCF – *Variant Call Format*

WGS – Sequenciamento de genoma completo (***W****hole* ***G****enome* ***S****equencing*)

WES – Sequenciamento de exoma completo (***W****hole* ***E****xome* ***S****equencing*)

RESUMO

A tecnologia de sequenciamento de nova geração (*next-generation sequencing* – NGS) e suas aplicações tem sido cada vez mais utilizada na prática médica para elucidar a base molecular de doenças Mendelianas. Embora seja uma poderosa ferramenta de pesquisa, ainda existe uma importante transição quanto à análise dos dados entre as tecnologias tradicionais de sequenciamento e o NGS. A primeira parte deste trabalho aborda aspectos analíticos envolvidos nesta mudança, com foco na plataforma Ion Torrent *Personal Genome Machine.* Esta é uma plataforma amplamente utilizada para sequenciar painéis de genes, já que esta aplicação requer menor rendimento de dados. Este trabalho demonstra indicadores adequados para avaliar a qualidade de corridas de sequenciamento e também uma estratégia baseada em valores de profundidade de cobertura para avaliar a performance de *amplicons* em diferentes cenários. Por outro lado, o NGS permitiu a realização de estudos populacionais em larga escala que estão mudando nossa compreensão sobre as variações genéticas humanas. Um desses exemplos são as mutações até então chamadas de silenciosas, que estão sendo implicadas como causadoras de doenças humanas. A segunda parte deste trabalho investiga a patogenicidade de polimorfismos de núcleotídeo único sinônimos (*synonymous single nucleotide polymorphisms* – sSNP) baseado em dados públicos obtidos do *Exome Aggregation Consortium* (ExAC) (exac.broadinstitute.org/) utilizando o *software Silent Variant Analysis* (SilVA) (compbio.cs.toronto.edu/silva/) e outros recursos para reunir informações adicionais sobre consequências funcionais visando fornecer um panorama dos efeitos patogênicos de sSNP em mais de 60.000 exomas humanos. Nós demonstramos que de 1,691,045 variantes sinônimas, um total de 26,034 foram classificadas como patogênicas pelo SilVA, com frequência alélica menor que 0,05. Análises funcionais *in silico* revelaram que as variantes sinônimas patogênicas estão envolvidas em processos biológicos importantes, como regulação celular, metabolismo e transporte. Ao expor um cenário de variações sinônimas patogênicas em exomas humanos, nós concluímos que filtrar sSNP em *workflows* de priorização é razoável, no entanto em situações específicas os sSNP podem ser considerados. Pesquisas futuras neste campo poderão fornecer uma imagem clara do papel de tais variações em doenças genéticas.

ABSTRACT

Next-generation sequencing (NGS) technologies and its applications are increasingly used in medicine to elucidate the molecular basis of Mendelian diseases. Although it is a powerful research tool, there is still an important transition regarding data analysis between traditional sequencing techniques and NGS. The first part of this work addresses analytical aspects involved on this switch-over, focusing on the Ion Torrent Personal Genome Machine platform. This is a widely used platform for sequencing gene panels, as this application demands lower throughput of data. We present indicators suitable to evaluate quality of sequencing runs and also a strategy based on depth of coverage values to evaluate amplicon performance on different scenarios. On the other hand, NGS enabled large-scale population studies that are changing our understanding about human genetic variations. One of these examples are the so-called silent mutations, that are being implied as causative of human diseases. The second part of this work investigates the pathogenicity of synonymous single nucleotide polymorphisms (sSNP) based on public data obtained from the Exome Aggregation Consortium (ExAC) (exac.broadinstitute.org/) using the software Silent Variant Analysis (SilVA) (compbio.cs.toronto.edu/silva/) and other sources to gather additional information about affected protein domains, mRNA folding and functional consequences aiming to provide a landscape of harmfulness of sSNP on more than 60,000 human exomes. We show that from 1,691,045 synonymous variants a total of 26,034 were classified as pathogenic and by SilVA, with allele frequency lower than 0.05. *In silico* functional analysis revealed that pathogenic synonymous variants found are involved in important biological process, such as cellular regulation, metabolism and transport. By exposing a scenario of pathogenic synonymous variants on human exomes we conclude that filtering out sSNP on prioritization workflows is reasonable, although in some specific cases sSNP should be considered. Future research on this field will provide a clear picture of such variations on genetic diseases.

CAPÍTULO I

INTRODUÇÃO

I.1. A Busca pelo Diagnóstico Molecular

A identificação de alterações genéticas responsáveis pelo surgimento de um fenótipo patológico específico é um dos principais objetivos da genética médica. A determinação de uma mutação causal permite o tratamento e aconselhamento genético adequados. A importância desta questão reflete-se na contínua incorporação de novas tecnologias que possibilitem explicar a base genética de fenótipos específicos de relevância médica (Manolio et al., 2009; Bamshad et al., 2011).

Diversas estratégias genômicas têm sido utilizadas de forma eficaz para a avaliação de pacientes com suspeita de desordens genéticas. Dentre elas, a tecnologia de sequenciamento de nova geração (*next-generation sequencing* ou NGS) revolucionou a genética médica. A primeira plataforma de NGS foi lançada no mercado em 2005 e permitia o sequenciamento de 25 milhões de bases em apenas quatro horas (Margulies et al., 2005). Desde então, o NGS se tornou a ferramenta de escolha para estudos de genética por permitir a análise simultânea de diferentes regiões genômicas (Goldstein et al., 2013; Metzker, 2010). Após uma década de uso, a tecnologia de NGS continua a evoluir (Goodwin et al., 2016) e passou a fazer parte da rotina da pesquisa biológica (Shen et al., 2015) e do diagnóstico genético (Rehm, 2017), possibilitando que pesquisadores entendam como as variações nas sequências do genoma humano delineiam o fenótipo e as doenças.

I.2. A Tecnologia de Sequenciamento de Nova Geração

O NGS pode ser caracterizado como um sequenciamento automatizado, paralelo e de alto rendimento. Os protocolos iniciam com o preparo de uma biblioteca que consiste, basicamente, na fragmentação do DNA inicial (de forma randômica ou sistemática) em fragmentos pequenos, seguido pela ligação de adaptadores (uma sequência curta de bases conhecidas) em ambas as extremidades de cada fragmento. Adicionalmente, diferentes amostras podem ser indexadas por sequências curtas chamadas *barcodes*, permitindo o sequenciamento de múltiplas amostras ao mesmo tempo. Os adaptadores são utilizados como primers universais e para distribuir espacialmente os fragmentos de DNA em uma superfície sólida, fixando-os através da complementariedade de bases entre as sequências do adaptador e da superfície sólida (Goodwin et al., 2016). Os fragmentos são, então,

amplificados por PCR gerando clusters de sequências idênticas. A presença dos clusters garante que a incorporação de cada base durante a reação de sequenciamento, produza um sinal suficientemente forte para ser detectado pelo sistema óptico ou eletroquímico do sequenciador. Após esta etapa de enriquecimento, o sequenciamento ocorre através de sucessivos ciclos de incorporação e a sequência de cada cluster é obtida através de softwares plataforma-específicos, que convertem em bases o sinal obtido em cada ciclo de incorporação, um processo denominado *base calling* (Pfeifer, 2017).

É importante destacar que o protocolo de preparo das bibliotecas pode ser *single-end*, onde cada fragmento é sequenciado na orientação *forward*, ou *paired-end*, onde cada fragmento é sequenciado em ambas as orientações, *forward* e *reverse*, gerando pares de leituras (*reads*). Bibliotecas *single-end* geram *reads* com boa qualidade, estão disponíveis para todas as plataformas e são amplamente utilizadas para detecção de Polimorfismos de Nucleotídeos Únicos (SNPs) e pequenas inserções e deleções (indels) (Morey et al., 2013). Os pares de *reads*, por sua vez, facilitam a detecção de rearranjos genômicos (Suzuki et al., 2014). Entretanto, os kits possuem um custo maior e não estão disponíveis para todas as plataformas.

Apesar da similaridade entre os protocolos de NGS, as plataformas de sequenciamento possuem características únicas quanto à quantidade de DNA inicial necessária, preparo da biblioteca, dependência e método de amplificação por PCR, rendimento (do inglês, *throughput*), tamanho médio dos *reads* e, principalmente, viéses associados ao preparo da biblioteca, amplificação e sequenciamento, assim como erros sistemáticos, que resultam em uma taxa de erro médio que difere de acordo com a plataforma de escolha (Reuter, et al., 2015; Ambardar *et al*., 2016). Um exemplo que vale a pena ser citado é a diferença existente entre as plataformas Illumina e Ion Torrent, amplamente distribuídas no mercado mundial. A plataforma Illumina possui uma taxa de erro de ~0,1% e sub-representação de regiões ricas em AT ou CG. Já a plataforma Ion Torrent possui uma taxa de erro de ~1% derivada, principalmente, da inconsistente representação de regiões de homopolímeros (Pfeifer, 2017; Reinert et al., 2015). Diversas revisões demonstram as características das principais plataformas de NGS disponíveis comercialmente, delineando informações como preparo da biblioteca, tamanho dos *reads*, rendimento por corrida, taxa de erro, limitações e vantagens, tempo de corrida, custo da plataforma e outros (Goodwin et al., 2016; Pfeifer, 2017).

I.3. As Etapas Básicas de um *Workflow* para Análise de Dados

Embora o NGS proporcione valiosa quantidade de informações biológicas e aplicações, os desafios no processamento e interpretação dos dados são complexos e tornaram a bioinformática uma área indispensável e de intenso desenvolvimento (Pabinger et al., 2013). O processo de *base calling*, realizado automaticamente pelas plataformas de sequenciamento, gera um arquivo .FASTQ, formato de dados mais popular para recordar leituras brutas de sequências curtas (Mielczarek & Szyda, 2016). Este arquivo é o ponto inicial dos *workflows* de análise e abriga todos os *reads* e o valor de qualidade de cada uma das bases geradas no sequenciamento. O valor de qualidade é representado pelo *Phred score* (Q), que expressa a probabilidade de que uma base determinada no *base calling* esteja errada (Ewing & Green, 1998). As informações do .FASTQ possibilitam realizar uma etapa importante de controle de qualidade, na qual *reads* de baixa qualidade são descartados, com o objetivo de assegurar um conjunto de dados de alta qualidade.

Após o processamento inicial dos dados é realizado o alinhamento dos *reads* ao genoma de referência existente ou, alternativamente, uma montagem *de novo* sem referência. Os genomas de referência estão disponíveis para muitas espécies, tornando esta estratégia de alinhamento muito popular. Em geral, o alinhamento com o genoma de referência inicia com uma indexação e então o processo de alinhamento é, de fato, realizado. A indexação aumenta a velocidade do alinhamento e a maioria dos *softwares* constrói índices (estruturas de dados auxiliares) para a sequência de referência, mas também é possível indexar os *reads* de cada amostra individualmente. A vantagem de indexar a sequência referência é o menor custo computacional requerido, uma vez que esta precisa ser conduzida apenas uma vez (Mielczarek & Szyda, 2016; Reinert *et al*., 2015). Os algoritmos principais de indexação incorporados na maioria dos *softwares* são baseados em tentativas sufixo/prefixo ou em tabelas *hash* (Li & Homer, 2010). Independentemente do método de indexação, o alinhamento *per se* é realizado utilizando o algoritmo de Smith-Waterman (Smith & Waterman, 1981) ou o algoritmo de Needle-Wunsch (Needleman & Wunsch, 1970) e, dependendo do *software* utilizado, o alinhamento resultante pode ou não conter *gaps*. Alinhamentos com *gaps* resultam de rearranjos genômicos de pequena escala, como inserções e deleções. Deste modo, a existência destas lacunas é uma característica

desejável e a maioria das ferramentas disponibiliza essa opção, bem como suporte ao alinhamento de *reads* de bibliotecas *paired-end* (Mielczarek & Szyda, 2016; Reinert *et al*., 2015).

Independentemente do *software* utilizado, o *output* do alinhamento é representado no formato *Sequence Alignment/Map* (SAM) ou na versão binária *Binary Alignment/Map* (BAM). O formato SAM é um formato de alinhamento genérico flexível, compacto no tamanho e eficiente no acesso aleatório por indexação, usado para armazenar alinhamentos de *reads* contra sequências de referência, suporta sequências curtas e longas produzidas por diferentes plataformas de sequenciamento (Li *et al*., 2009). Os desenvolvedores do formato SAM também implementaram um conjunto de utilitários chamado SAMtools, que fornece ferramentas universais para o pós-processamento do alinhamento (Li *et al*., 2009). O pós-processamento é uma prática frequentemente aplicada nos *workflows* de descoberta de variantes (Pfeifer, 2017), dois passos muito utilizados incluem realinhamento local em localizações conhecidas de indels e recalibração do escore de qualidade de bases (BQSR) (Mielczarek & Szyda, 2016). Entretanto, o impacto desta prática vem sendo discutido e um estudo recente demonstrou que a eficiência do pós-procesamento nem sempre aumenta a detecção de variantes, variando de acordo com os *softwares* utilizados no alinhamento e chamada de variantes, profundidade de cobertura e nível de divergência (Tian *et al*., 2016).

Com os *reads* alinhados é possível identificar as bases que diferem do genoma de referência, um processo chamado *variant call* ou chamada de variantes. Muitos algoritmos e *softwares* foram desenvolvidos para identificar SNPs em dados de NGS (Pabinger et al., 2014), entretanto para uma detecção confiável é necessária uma alta profundidade de cobertura, pois o número de *reads* alinhados a cada base é fundamental para diferenciar erros de sequenciamento de polimorfismos verdadeiros (Mielczarek & Szyda, 2016). Quando uma única amostra é analisada, a chamada de variantes e a definição do genótipo são similares, já que um genótipo homozigoto ou heterozigoto que difere da referência implica na presença de um SNP. Na análise simultânea de várias amostras, um SNP é identificado se pelo menos um indivíduo é homozigoto ou heterozigoto para um alelo que difere da referência. Os *softwares* utilizados para determinar a presença de SNPs e os genótipos são baseados em métodos heurísticos ou probabilísticos. Os métodos heurísticos demandam maior poder computacional e definem as variantes baseados em informações de múltiplas fontes associando a estrutura e qualidade dos dados. Os métodos probabilísticos

são mais populares e frequentemente baseados na inferência Bayesiana, fornecendo medidas de incerteza estatística para os genótipos chamados que permitem monitorar a acurácia do genótipo determinado (Mielczarek & Szyda, 2016; Pabinger et al., 2014; Pfeifer, 2017). A informação resultante da chamada de variantes é recordada no arquivo *variant call format* (VCF), um formato genérico que armazena dados de variações genéticas, tais como SNPs, inserções, deleções e variações estruturais. O arquivo VCF apresenta uma versão binária e pode ser indexado para rápida recuperação de variantes em um intervalo de posições de um genoma de referência, por exemplo. Possui, ainda, um conjunto de utilitários semelhante ao SAMtools chamado VCFtools, desenvolvido para realizar o processamento de arquivos VCF, incluindo validação, concatenação, comparação, conversão e outros (Danecek *et al*., 2011).

Uma métrica determinante para todos os passos do *workflow* de análise descrito anteriormente é a profundidade de cobertura, sendo oportuno tecer alguns comentários. Previamente ao sequenciamento existe a cobertura teórica ou esperada, que corresponde ao número médio de vezes esperado de que cada nucleotídeo seja sequenciado dado um certo número de *reads* de determinado comprimento e a concomitante suposição de que *reads* são distribuídos aleatoriamente através de um genoma idealizado. Após o sequenciamento, a cobertura empírica real por base (*per-base coverage*) representa o número exato de vezes que uma base na referência é coberta por um *read* de alta qualidade alinhado. A redundância de cobertura também é chamada de profundidade de cobertura ou *depth*, sendo frequentemente citada como profundidade média de *reads* brutos ou alinhados, que indica a cobertura esperada com base no número e no comprimento de *reads* de alta qualidade antes ou depois do alinhamento com a referência (Sims *et al*., 2014).

I.4. As Aplicações do Sequenciamento de Nova Geração

O NGS é uma tecnologia que oferece inúmeras possibilidades de aplicações e novos métodos têm sido desenvolvidos continuamente. De acordo com o objetivo experimental, o NGS pode ser usado para (1) construir um novo genoma de um organismo desconhecido; (2) avaliar a variação genética de um organismo contra um genoma de referência existente; (3) analisar globalmente a transcrição de um organismo ou célula a partir do DNA complementar; (4) estudar o epigenoma e mecanismos regulatórios de um organismo e (5)

investigar a diversidade microbiana de amostras ambientais a partir de amostras ambientais não cultiváveis (Park & Kim; 2016). Diversos estudos abordam a variedade de aplicações do NGS. Em artigo recente de revisão sobre a tecnologia, Reuter *et al* (2016) subdividem as aplicações em grandes áreas: expressão gênica, biologia do RNA, regulação do genoma, sequenciamento do genoma, organização do genoma, transcrição, replicação, tradução e outras (Reuter *et al*., 2016).

Um dos motivos que possibilitou esta grande variedade de aplicações é a diminuição dos custos de sequenciamento (van Nimwegen *et al*., 2016). Atualmente, plataformas de sequenciamento de NGS são comuns em universidades e laboratórios particulares. As aplicações tornaram-se cada vez mais robustas, permitindo a realização de projetos de larga escala que têm fornecido recursos valiosos à comunidade e abordado questões que seriam laboriosas para laboratórios individuais atingirem, como frequência populacional de variações genéticas (Reuter *et al*., 2016). Um destes projetos é o *Exome Aggregation Consortium* (ExAC) (Lek *et al*., 2016), uma coalizão de pesquisadores que procuram agregar e harmonizar dados de variação genética de regiões codificantes de proteínas a partir de 60.706 indivíduos sequenciados.

Uma das aplicações de NGS amplamente utilizada na prática clínica é a avaliação das variações genéticas individuais e a associação destas com desordens genéticas (Shen *et al*., 2015). É possível interrogar o genoma de um indivíduo na busca de SNPs e indels com consequências deletérias através do sequenciamento de genoma completo (WGS), sequenciamento de exoma completo (WES) ou pelo sequenciamento de um painel de genes (*targeted gene panel*, ou TGP). No WGS, todo o conteúdo genômico é avaliado, o DNA do invidívuo é extraído e submetido ao preparo da biblioteca e *workflow* de análise. Apesar de fornecer uma visão global da variação genética, o WGS ainda possui uma utilidade clínica limitada, já que a contribuição de regiões não codificantes para a etiologia de doenças genéticas ainda não é completamente elucidada. Ainda, o custo real do sequenciamento e custo computacional para análise e armazenamento dos dados contribuem para este cenário (Goodwin *et* al., 2016; Petersen *et al*., 2017).

Alternativamente ao WGS, o WES e o TGP utilizam estratégias de enriquecimento de regiões de interesse para realizar o sequenciamento de regiões codificantes. O enriquecimento é baseado, principalmente, em amplificação ou hibridização e oferecem alta especificidade de captura do alvo (Samorodnitsky *et al*., 2015; Sun *et al*., 2015). O

exoma refere-se às regiões codificantes do genoma (1-2%) e o WES permite a análise global das variações presentes em todas as regiões codificantes de proteínas de um indivíduo. A análise de um único indivíduo tem potencial para identificar mais de 30.000 variações genéticas (Smedley et al., 2014) e o desenho experimental de um estudo de WES abrange três tipos principais: WES baseado em família (Stitziel et al., 2013; Yu et al., 2013), WES caso-controle (Fu et al., 2013, Li et al., 2013; Yang et al., 2013) e WES de caso único (Worthey et al., 2011). O WES de caso único é muito utilizado na prática clínica, já que o custo do exame restringe a realização de testes adicionais em outros indivíduos além dos afetados. Sua utilização tem sido crescente para a análise de casos isolados onde o diagnóstico genético por metodologias tradicionais é inconclusivo (Wu et al., 2015). No TGP um conjunto customizado de genes ligados à uma condição específica de interesse é sequenciado. O TGP tem sido amplamente utilizado na prática do diagnóstico molecular (Dallol *et al*., 2016; Hegele *et al*., 2015; Johansen *et al*., 2014; Lim *et al*., 2015) pelo baixo custo quando comparado à outras aplicações de NGS (van Nimwegen *et al*., 2016), por necessitar de menor quantidade de DNA para o preparo da biblioteca, permitir a análise de diversas amostras em uma mesma corrida devido à menor quantidade de dados gerados, menor tempo para realizar a análise dos dados e, principalmente, por possibilitar que cada fragmento de DNA seja sequenciado diversas vezes independentemente (profundidade de cobertura) devido ao reduzido número de alvos (Morey *et al*., 2013).

I.5. Interpretação de variantes genéticas e a importância das mutações sinônimas

A popularização do WES e do TGP estimulou a criação de ferramentas e estratégias específicas para lidar com a priorização de variantes patogênicas (Li et al., 2012; Alemán et al., 2014; Jiang, 2015; Cooper & Shendure, 2012; Moreau & Tranchevent, 2012). Uma vez realizado o processamento dos dados, o arquivo VCF é o ponto de partida para a priorização. Porém a priorização é uma tarefa laboriosa já que cada indivíduo carrega milhares de alterações na região codificante (>30.000) (Smedley et al., 2014) e uma parcela significativa destas (5-10%) não está descrita em bancos de dados como dbSNP e não possui dados de frequência alélica relatada, que ainda podem apresentar variações populacionais (Koboldt et al., 2014).

A escolha da estratégia ideal para priorizar as variantes patogênicas entre as milhares anotadas após o processamento dos dados depende do tipo de estudo de WES, da disponibilidade de indivíduos e famílias com fenótipo bem caracterizado, do modo de herança, severidade da doença (Gilissen et al., 2012) e frequência populacional das alterações encontradas (Shearer et al., 2014). Ferramentas e *workflows* desenvolvidos para priorização, implementam uma combinação de filtros para excluir variantes comuns e eleger a variante causal da doença investigada (Robinson et al., 2014; Smedley et al., 2014; Koboldt et al., 2014). É pertinente ressaltar que filtros heurísticos comumente empregados baseiam-se em várias suposições sobre a variante patogênica a ser encontrada: altera a sequência codificante da proteína; é rara; possui penetrância completa, por exemplo (Stitziel et al., 2011).

A anotação funcional é um dos principais filtros utilizados e classifica as variantes em sinônimas e não-sinônimas (Ohanian et al., 2015). As mutações não-sinônimas tem consequências óbvias para função protéica e seu papel na patogênese de diversas doenças é bastante explorado. Considerando a quantidade de alterações reportadas após o processamento dos dados de WES, focar nas alterações não-sinônimas facilita muito o processo de priorização de variantes patogênicas. Assim, 50-75% das variantes identificadas (Stitziel et al., 2011) são simplesmente excluídas e as alterações não-sinônimas tornam-se foco das filtragens subsequentes para determinar a mutação responsável pela condição. Essa estratégia de priorização tem sido amplamente empregada e várias publicações relatam o sucesso deste método (Fu et al., 2013; Katsonis et al., 2014; Li et al., 2013). Porém, estas são menos efetivas do que se espera: em apenas 21% dos exomas de caso único é possível definir o diagnóstico (Farwell *et al*., 2015).

Mais recentemente, consequências funcionais de alterações sinônimas na expressão, conformação e função protéica passaram a ser exploradas e relatadas na literatura (Plotkin & Kudla, 2011). Tem-se conhecimento de que mais de 50 doenças estão associadas a mutações sinônimas e que variações sinônimas e não-sinônimas possuem similar probabilidade de associação à doenças (Chen *et* al., 2010; Hunt et al., 2014). Implicações de mutações sinônimas em mecanismos essenciais para homeostase celular como acurácia de *splicing*, fidelidade de tradução, estrutura de RNA mensageiro e dobramento protéico e a relação destas alterações com a patogênese de doenças tornaram-se foco de inúmeras pesquisas (Sauna & Kimchi-Sarfaty, 2011), assim como abordagens

computacionais para analisar a patogenicidade de alterações sinônimas (Buske et al., 2013), adicionando uma nova camada de complexidade na análise do significado biológico das variações genéticas do genoma humano.

CAPÍTULO II

JUSTIFICATIVA

Sem dúvidas o sequenciamento de nova geração inaugurou uma nova era na genômica e no diagnóstico de doenças genéticas, com consequências claras para a prática do diagnóstico molecular. Entretanto, a robustez própria da tecnologia de NGS tornou a análise dos dados e a interpretação de variantes genéticas uma tarefa pouco trivial.

Existe uma importante transição que deve ser feita por médicos geneticistas e pesquisadores no que concerne à interpretação de variantes patogênicas. Metodologias tradicionais de sequenciamento forneciam uma visão restrita da variação genética e com o NGS essa visão foi ampliada diversas vezes. Somado à isso, novas nomenclaturas, parâmetros e formatos de arquivos que compõem *workflows* de análise também devem ser apreendidos para que o profissional que realiza o diagnóstico molecular com aplicações de NGS possa assegurar a identificação da variante causal da condição investigada com acurácia. Considerando este cenário de transição, trabalhos que abordem aspectos analíticos relevantes servindo de guia para interpretação acurada de resultados de sequenciamento de painél de genes (*targeted gene sequencing*) são de extrema valia.

O NGS também possibilita avaliar globalmente as variações genéticas das regiões codificantes de proteínas do genoma humano através do WES. Além do *workflow* básico para processar os dados brutos, o WES agrega uma camada de complexidade na priorização e interpretação de variantes genéticas, já que um único exoma pode identificar ~30.000 SNP. A priorização de variantes baseia-se numa série de passos de filtragem para identificar um conjunto menor de variantes que possam estar ligadas à condição investigada. Um dos primeiros passos de filtragem utilizados é a remoção de variantes sinônimas. Entretanto, a relação deste tipo de variação com alteração da homeostase celular alterada e doenças humanas tem sido recentemente reportada na literatura. Mergulhando mais na complexidade da interpretação de variantes genéticas, explorar a patogenicidade de variantes sinônimas pode contribuir para o entendimento da relação destas variações e doenças genéticas.

CAPÍTULO III

OBJETIVOS

III.1. Objetivo geral

Analisar aspectos relevantes para aplicação clínica do sequenciamento de nova geração.

III.2. Objetivos específicos

1. Explorar abordagens para avaliar com precisão o desempenho de sequenciamento de painéis de genes em dados de *targeted gene sequencing*.

2. Avaliar a patogenicidade de variantes sinônimas raras obtidas dados públicos de sequenciamento completo de exoma.

CAPÍTULO IV

MANUSCRITO I

Next-generation sequencing using Ion Torrent PGM platform:

how to handle a gene panel results

(Formatado para submissão ao periódico PLOS ONE)

Next-generation sequencing using Ion Torrent PGM platform: how to handle a gene panel results

Delva Pereira Leão[1,4], Diana Elizabeth Rojas Málaga[1,4], Marina Siebert[2,4], Silvia Liliana Cossio[4], Ana Carolina Brusius-Facchin[1,5], Bárbara Alemar[1,4], Roberto Giugliani[1,3,5], Ursula Matte[1,3,4*].

[1]Post-Graduation Program on Genetics and Molecular Biology, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil;
[2]Post-Graduation Program on Sciences: Gastroenterology and Hepatology, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil;
[3]Department of Genetics, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil;
[4]Experimental Research Center, Hospital de Clínicas de Porto Alegre, Porto Alegre, RS, Brazil.
[5]Medical Genetics Service, Hospital de Clínicas de Porto Alegre, Porto Alegre, RS, Brazil.

*Corresponding author:
E-mail: umatte@hcpa.edu.br (UM)

Abstract

Next-generation sequencing (NGS) is a high-throughput technique that generates enormous amount of data and has several advantages over Sanger sequencing, such as being less time-consuming, having lower costs and scalability. Targeted sequencing of genes has shown to be an interesting approach to research groups that want to embody this technology with focus on a specific group of diseases. Ion Torrent Personal Genome Machine (PGM) is one of the most affordable and user-friendly NGS platforms available. However, in order to evaluate customized gene panels, a deeper understanding of data meaning is needed. Based on this, the aim of this study was to explore approaches to accurately assess performance of gene panel data, with a focus on customized Ion AmpliSeq® Panels. Five customized gene panels and a total of 120 unrelated human samples were included in this study. All raw data processing was performed at Torrent Suite™ software v5.0. Quality processed and not processed data were compared to demonstrate the reliability of data provided after accurate base calling, trimming and filtering processes using default parameters of Torrent Suite™ software. Coverage charts were generated to evaluate depth distribution of amplicons throughout a gene panel. Although panels were different in number of target genes and size, we established a serie of features that can be used to evaluate the overall performance of any Ion Ampliseq® Panel. The present study covered different aspects of data analysis, highlighting the importance of the evaluation of sequencing quality, coverage distribution of amplicons and adequate data processing using recommended parameters. As a result, reliable data will be generated for subsequent identification of mutations and clinical interpretation.

Keywords: Next-generation sequencing, gene panel, targeted sequencing, genetic disorders, Ion Torrent PGM

Introduction

Recently, advances in sequencing technologies have enabled the global analysis of genetic alterations. The advantages of these technologies include more scalable and lower-cost solutions with the possibility of obtaining real-time results (Rothberg et al., 2011). The Ion Torrent Personal Genome Machine (PGM) (Thermo Fisher Scientific) is a semiconductor-based high throughput sequencing platform. Its detection method is based on modification of surface potential of sensors triggered by pH alteration upon the incorporation of a nucleotide into a growing strand of DNA (Rothberg et al., 2011). Ion Torrent PGM is suitable to explore sets of genes, small genomes and targeted metagenomics (Morey et al., 2013). Currently, depending on the Ion chip throughput, it has a maximum output of two gigabytes of raw data in 3.0 to 7.0 hr, with a read length of 35 up to 400 bases and 200 bases average, generating up to 5.5 million reads per run, according to manufacturer's information. Ion PGM is especially useful for targeted sequencing of genes and the manufacturer offers an auxiliary platform for the design of customized primer pools to construct targeted libraries of genomic regions of interest.

Consistent with its ambition to be an entry platform for NGS suited for low experienced users, data analysis on PGM is especially useful for targeted sequencing of genes and capable of efficiently identify point mutations and small insertions/deletions. Data analysis on PGM is carried out by Torrent Suite™ software, a robust pipeline that provide its output through Torrent Browser, an user-friendly graphic interface designed for researchers with little proficiency in bioinformatics. Data processing, a common NGS bottleneck, was reshaped with the pipeline implemented on Torrent Suite™ software. However, although a simplified and automated system is available, users must be able to evaluate their customized panels performance. This is mainly represented by quality and coverage, to

access accurate identification of mutations and clinical interpretation. In this study, we aim to discuss relevant practices that help on the evaluation of performance of custom designed gene panels using the tools embedded in Ion PGM.

Material and Methods

Datasets and samples

Datasets for this study were obtained from five customized gene panels generated with the Ion AmpliSeq™ Designer tool (www.ampliseq.com; Thermo Fisher Scientific). All panels were designed for germline application with high specificity for standard DNA. Three different sequencing runs of each panel were selected for further analysis.

A total of 120 unrelated human samples were included in this study. These samples belong to patients with different genetic disorders that were previously diagnosed by Sanger sequencing at the Medical Genetics Service of Hospital de Clínicas de Porto Alegre, located in Brazil. Informed consent and ethical approval were obtained by Ethics Research Committee of Hospital de Clínicas de Porto Alegre for the original research projects in which these individuals were enrolled. Stored gDNA samples extracted from peripheral blood were quantified using NanoDrop 1000 Spectrophotometer (Thermo Fisher Scientific) to evaluate the $A_{260/280}$ ratio and obtain the initial DNA concentration. Then, samples were diluted to 8 ng/μL and a new quantification was performed on Qubit® 2.0 fluorometer (Thermo Fisher Scientific) using the Qubit® dsDNA HS kit (Thermo Fisher Scientific).

Next-generation sequencing assay

For library construction, twenty nanograms of gDNA were used as input for library construction using Ion AmpliSeq™ Library kit 2.0 (Thermo Fisher Scientific). Eight

samples barcoded with Ion Xpress™ Barcode Adapters kit (Thermo Fisher Scientific) were included in each set of library preparation. Unamplified libraries were purified with Agencourt AMPure XP kit (Beckman Coulter). Libraries were prepared in equimolar concentrations using the Ion Library Equalizer™ kit (Thermo Fisher Scientific) or quantified using the Qubit® dsDNA HS kit (Thermo Fisher Scientific), followed by dilution to the same concentration. Both methods are described in the Ion AmpliSeq™ DNA and RNA Library Preparation (Life Technologies, 2014) user guide.

For template preparation, the eight barcoded libraries were pooled in equimolar concentrations of 100 pM each and were subsequently submitted to emulsion PCR using the Ion PGM™ Template OT2 200 kit (Thermo Fisher Scientific) on the Ion OneTouch2™ Instrument (Thermo Fisher Scientific). The percent of positive Ion Sphere Particles (ISPs) was defined with flow cytometry performed on Attune® Acoustic Focusing Flow Cytometer (Thermo Fisher Scientific) according to the demonstrated protocol (Part. no. 4477181, Thermo Fisher Scientific). Positive ISPs were enriched using Ion OneTouch™ ES (Enrichment System; Thermo Fisher Scientific) and submitted to sequencing on the PGM equipment. The enriched template-positive ISPs were loaded onto Ion 314™ chip v2 (Thermo Fisher Scientific) and sequenced using the Ion PGM™ 200 Sequencing kit (Thermo Fisher Scientific), following the manufacturer's instructions.

Evaluation of custom designed panels

All raw data processing was performed at Torrent Suite™ software v5.0 (Thermo Fisher Scientific) using default parameters. For a sequencing run to be included in the dataset, the following thresholds on parameters available on the run report had to be met: 1) test fragments (TF) percent 50AQ17 > 60%; 2) key signal > 40; 3) read length distribution close to the expected for the library; 4) chip loading > 30%; 5) Mean depth of 150x per

sample. TF are known sequences spiked into the library before the loading into the chip to provide a positive control of sequencing and play an important role on reporting the quality of a run. The key signal is the strength of the signal associated with the key sequence, which is a short sequence that designates the reads as a library or a test fragment. The cutoff used for these parameters were defined based on the need and experience of our group, and should not be interpreted as established or limiting values. Nevertheless, these metrics provide a valuable overview of possible failures and should be observed each time a sequencing run is performed. After assessment of the sequencing run performance, we carried out further analysis.

Base quality evaluation

To demonstrate that the Torrent Suite™ software v5.0 performs quality processing of sequenced bases, we compared two conditions named as processed data (default output of software pipeline) and not processed data (disabled quality and filtering processing output). Processed data information was gathered based on run report provided by Torrent Browser. For the same run, to obtain the not processed condition, we reanalyzed the data changing default base calling and filtering parameters of quality. By default, the parameter "disable-all-filters" is off; when this parameter is turned on, it disables all filtering and trimming and overrides other filtering and trimming settings. The "keypass-filter" parameter when on, filters out reads that do not produce either the library key or the key test fragment signal match. To graphically demonstrate the difference between processed and not processed data and the importance of quality control, we selected metrics that explain the heterogeneity of reads analyzed in the two conditions, namely: the total number of bases, in megabase (Mb), and measures of central tendency, in base pair (bp).

Panel coverage performance assessment

The Coverage Analysis v5.0 plugin was executed to our dataset. For each run, the output file barcode/amplicon coverage matrix was downloaded. Depth of Coverage (DoC) information of each amplicon across sequenced samples was used to create two charts that provide an overview of uniformity of coverage depth and run reproducibility. The intra-run graph is used to evaluate the intra-run Relative Depth of Coverage (RDoC) for each amplicon of a single sequencing run. RDoC was computed using the formula below:

$$\text{RDoC}=\log\frac{DoC\ for\ each\ amplicon}{Median\ DoC\ of\ the\ run}$$

DoC: Depth of Coverage

RDoC values equal or close to 0 indicate that a given amplicon had a depth of coverage similar to the median coverage of the run. Values above or below 0 indicate that the amplicon is over- or underrepresented, respectively. The inter-run graph was helpful to analyze the amplicon inter-run reproducibility. The mean RDoC of the amplicon in every run was calculated and plotted, this was done for all the three runs analyzed per panel. Furthermore, to show possible different coverage distribution scenarios a specific chart was created modifying depth values of barcode/amplicon coverage matrix file before computing RDoC. The representation of a poorly covered amplicon across all samples was obtained by decreasing the amplicon DoC in each sample. On the other hand, to represent an entire sample under or overrepresented, we assigned a lower or higher DoC value to each amplicon, respectively. Finally, to represent a non amplified target, we assumed 1 as value of DoC for a particular amplicon in a sample.

Results

Gene panel specifications

Table 1 summarizes the main attributes of gene panels included in the present study. The number of target genes varied across the panels and influences the number of designed amplicons and target size. Differences on breath of coverage were due to heterogeneous nature of coding regions. Despite the different versions of Ion AmpliseqSeq™ Designer tool used, no inequality on sequencing performance was observed.

Table 1. Specifications of customized gene panels designed using Ion Ampliseq™ Designer tool.

Panel	Number of target genes	Target size (Kb)	Number of amplicons	Breadth of coverage (%)	Ampliseq version
A	14	55.72	293	98.44	4.0
B	4	13.42	83	95.91	2.0
C	4	7.81	66	95.92	2.0
D	2	8.61	49	90.15	2.0
E	4	13.10	72	97.74	3.4

Run performance features

Based on the run report, the runs that reached the established thresholds described above were included in our dataset. Three runs of each panel were selected for this study and mean performance of each panel is shown in Table 2. The values available in the Library ISPs correspond to reads that match the library sequencing key and should be close to the output expected for a given Ion chip. Final library values, also referred to as total reads at Torrent Browser interface, stand for the total of bases suitable to be used in subsequent analysis. This output is obtained from Library ISPs after subsequent steps of filtering out polyclonal, low quality and adapter dimer reads.

Table 2. Mean performance of five gene panels among different runs: final output.

		Filtered			
Panel	**Library ISPs***	**Polyclonal**	**Low quality**	**Adapter dimer**	**Final library** ***
A	924,589	251,642	59,685	19,405	593,857
B	1,058,371	383,792	76,095	10,499	587,986
C	998,962	294,25	126,826	13,848	564,037
D	971,215	290,017	109,725	14,679	556,794
E	1,047,190	293,591	98,306	25,538	629,755

*Library ISPs: live wells that have library template (loaded wells with live signal - test fragment wells). This parameter depends on loading performance.

** Final library: percentage of sequence available for analysis after filtering.

Base quality evaluation

The main goal of this analysis was to demonstrate that the quality checks performed by the Torrent Suite™ software v5.0 are enough to ensure subsequent analysis and may reveal important aspects of customized panels and individual runs. The flow processing of raw information performed by the software includes the assignment of base quality (represented by the Phred score) and the removal of poor quality bases by filtering and trimming. Fig 1 shows the difference on run patterns when such quality checks are disabled. The solid line represent reads processed with default quality parameters and the dotted line represent reads obtained by disabling the base quality processing.

As expected, the total number of not processed bases is greater than that of processed data and the same is observed for central tendency measures (mean, mode and median). The only exception to this is observed in run 1 of Panel C, in which a prominent decrease in the mode length of not processed data is observed. Therefore, each run, regardless of the specific panel, generated a set of reads with a different read length distribution. For most runs, these reads are of enough quality and the difference between the solid and dotted lines is mainly due to trimming. In this process, the 3’ ends of reads are checked for

matches to the adapter sequence and for regions of low quality, thereby ensuring the accuracy of data written out to the unmapped BAM file.

The approach used for removal of adapter sequence is based on the search and testing of candidate positions that matches the 3’ end to the known adapter sequence in flow-space. As to the removal of regions with low quality score, the trimming is performed using a per-base quality score combined with a fixed-length window (Fig 2) that slides through the read extension starting from 3’end, assessing at each position the mean quality score. When the mean quality score of the window falls below to the established threshold the read is trimmed, thus leading to different read lengths distributions.

By default, the two trimming strategies are applied separately and a given read length is taken as the shorter of the two. If the resultant read length is shorter than the minimum read length threshold, the read is filtered out entirely from the output file. Low-quality reads are also removed by filtering to withdraw library sequence reads with insufficient quality from final output. The goal of this processing is the removal of short reads, adapter dimers, reads without the key sequencing, with off-scale signal and polyclonals. Table 2 demonstrates filtered values with input values corresponding to Library ISPs and final values corresponding to the Final library. The two conditions demonstrated, especially for run 1 of Panel C, that quality processing is essential to generate a high quality dataset. Even though the distribution of not processed data may vary between different runs of the same panel, the processed data seems to have a panel-specific distribution. Despite of variations on the final output represented by Total Bases (Mb), as observed between the runs of Panel D, the quality checks ensure that resultant bases are reliable for subsequent analysis.

Panel coverage performance assessment

Amplicon coverage distribution was analyzed using the intra-run graph (Fig 3). In the evaluation of RDoC across a single sequencing run (Fig 3A), we observed that all amplicons present a RDoC value equal or close to 0, indicating coverage depth close to the median. Coverage distribution across several runs of a panel can be analyzed with Inter-run graph (Fig 3B). Mean distribution of coverage values equal, or close, in the different runs, as well as small standard deviation, demonstrate a high reproducibility. This was observed for all panels analyzed in this study. For the sake of comparison, we summarize in Fig 3C a hypothetic situation with uneven coverage distribution: amplicons poorly covered in all samples (amplicon 12); an entire sample under or overrepresented (sample 2 and 8, respectively) and targets not amplified (amplicons 9-11, sample 6).

Discussion

Sequencing coding regions of genes associated with genetic disorders using NGS technology became an interesting approach to interrogate genetic alterations, with many advantages over other methods, such as Sanger sequencing (Goldstein et al., 2013). There are key factors in the definition and evaluation of any NGS assay: 1) read length, 2) expected throughput, 3) read accuracy measured by Phred Score and 4) depth of coverage. Another important point to be considered is the management, processing and interpretation of the generated data (Morey et al., 2013; Bahassi and Stambrook, 2014). The Torrent Suite™ software encompasses solutions to deal with these factors, providing results through a friendly interface to a robust system. With this, the researcher can focus on interpreting issues that may influence the outcome, such as those presented here. One of the important factors to consider is the sequencing itself. Our previous experience of

carrying out several sequencing runs allowed us to conduct a detailed analysis on the run report and establish a set of indicators suitable to demonstrate the quality of runs included in the dataset presented on Table 1. Among these, TF is one of the most important. It reveals the sequencing performance and is important when a run did not perform as expected. In this case, if TF value is high, is possible to infer that the sequencing reaction itself occurred without any failures and explore through troubleshooting checkpoints possible problems that occurred in steps prior to sequencing. Taken together, these indicators provide a reliable way to evaluate the sequencing performance of a gene panel. Filtered values of Table 2 can be evaluated for a given run, but should not be used as performance parameters since they vary according to panel and run conditions.

The pipeline used to analyze our dataset performs default quality processing of sequenced bases, as shown in Fig 1. Interesting, the not processed data of Panel C on run 1 exhibited an excess of short length reads that masked the actual run quality. This condition resulted on reads with mode length of 50 bp in contrast to reads with 200 bp when quality checks were enabled. A possible scenario that explains this result is that the sequencing run yielded too many short reads and during the quality processing stage, reads examined with the sliding window approach and identified as low quality score regions were trimmed resulting in further shortening of read length. At the end of trimming, the resulting read lengths were lower than the default filtering threshold and these too short reads were filtered out from final output file, thus restoring the pattern observed for this panel with processed condition. After trimming and filtering those reads, the mode represents a high quality set: reads with expected length and established quality. The comparison between modes of both conditions demonstrates, notably in this panel, the combination between trimming and filtering processes used to obtain accurate reads. Despite the observed higher

output of not processed data, there is sufficient evidence to consider that the extra bases of this condition are not precise and accurate. The processed bases represented by solid line, even though with a smaller throughput, have higher quality and ensure that subsequent analyzes can be held with enough reliability. This demonstrates that the demand for higher throughput in terms of number of bases must not be dissociated from quality to avoid misleading analysis. On Panel D, even though the final output from run 1 is smaller than for runs 2 and 3, the reads are reliable for further analysis because the base calling filter settings are the same for all sequencing runs. The lower throughput could be due to poor sample integrity, biases in library preparation or due to the method of library quantification. Evaluating the mapped reads per sample, we noticed a marked lower throughput in libraries quantified by Ion Library Equalizer Kit when compared with Qubit dsDNA HS kit, which provided an uniform result across samples. We also performed several depth of coverage analysis to explore different sequencing landscapes of heterogenous gene panels (Fig 3). Ideally, all the amplicons should be equally represented across the entire region of interest, but some target regions could have depth values higher than expected while other regions could be poorly or not covered. Besides the natural bias introduced by the capture method based on amplification of target genes (Samorodnitsky et al., 2015; Tattini et al., 2015), other reasons can account for an unequal coverage distribution, as demonstrated by Fig 3C: a) the amount and quality of input DNA sample (sample 2 and 8) can result in an entire sample under- or over-represented; b) a region with an increased GC content is prone to be underrepresented (amplicon 12 in all samples) due to the difficulty of amplification; c) repeated elements or tandem repeats; d) biases in library preparation and sequencing. For these poorly or not covered regions, increase overall coverage per sequencing run or analyze these regions by Sanger sequencing provide a good solution to

fill these gaps. Consequently, it is important to consider the uniformity of sequencing depth as a representation of data quality, since this feature reduce the robustness of variant calling and identification of the genetic variants responsible for the investigated condition.

It is important to stress out that the graphs shown in Fig 3 are meant to show amplicon depth distribution throughout a gene panel. However, the situation represented by amplicons 9 to 11 of sample 6 on Fig 3C could suggest a copy number variation (CNV), that need to be confirmed by other methods, such as multiplex ligation-dependent probe amplification (MLPA). It is worth noticing that PCR-amplification steps required for the preparation of sequencing libraries and nonuniform read depth among amplified regions imply on low coverage homogeneity of amplicon sequencing, essential to CNV estimation through depth of coverage methods. In addition, amplicon sequencing data normalization can be less effective due to limited number of target regions (Duan et al., 2013; Tattini et al., 2015).

In summary we presented important aspects to be considered when evaluating performance of customized gene panels both by using real data and a simulation of possible situations that decrease the expected performance of particular samples or amplicons.

Acknowledgments

We would like to thank Universidade Federal do Rio Grande do Sul (UFRGS) and Hospital de Clínicas de Porto Alegre (HCPA) for their support during this work. Financial support for this study was provided by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, Brazil), Fundo de Incentivo à Pesquisa e Eventos do Hospital de Clínicas de Porto Alegre and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES, Brazil).

References

Bahassi EM, Stambrook PJ. Next-generation sequencing technologies: breaking the sound barrier of human genetics. Mutagenesis. 2014;29(5):303-310. doi: 10.1093/mutage/geu031

Duan J, Zhang J-G, Deng H-W, Wang Y-P. Comparative Studies of Copy Number Variation Detection Methods for Next-Generation Sequencing Technologies. PLoS ONE. 2013;8(3): e59128. doi:10.1371/journal.pone.0059128

Goldstein DB, Allen A, Keebler J, Margulies EH, Petrou S, Petrovski S, Sunyaev S. Sequencing studies in human genetics: design and interpretation. Nat Rev Genet. 2013;14(7):460-70. doi: 10.1038/nrg3455

Morey M, Fernández-Marmiesse A, Castiñeiras D, Fraga JM, Couce ML, Chocho JA. A glimpse into past, present, and future DNA sequencing. Molecular Genetics and Metabolism. 2013; 110(1-2): 3-24. doi:10.1016/j.ymgme.2013.04.024

Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, et al. An integrated semiconductor device enabling non-optical genome sequencing. Nature. 2011; 475(7356):348-352. doi:10.1038/nature10242

Samorodnitsky E, Datta J, Jewell BM, Hagopian R, Miya J, Wing MR, et al. Comparison of Custom Capture for Targeted Next-Generation DNA Sequencing. The Journal of Molecular Diagnostics. 2015;17(1):64-75. doi:10.1016/j.jmoldx.2014.09.009

Tattini L, D'Aurizio R, Magi A. Detection of genomic structural variants from next-generation sequencing data. Frontiers in Bioengineering and Biotechnology. 2015;3(92). doi:10.3389/fbioe.2015.00092

Supporting Information

S1 Figure. Comparision of two quality conditions altering specific parameters in the Torrent Suite™ software.

Fig 1. Comparision of two quality conditions altering specific parameters in the Torrent Suite™ software.

To each run of our dataset, two conditions named processed and not processed data was obtained by changing the parameters *disable-all-filters* and *keypass-filter* of base calling. Results for panels C and D are shown, solid line stands for reads processed using default parameters, while the dotted line represents reads obtained by disabling quality filters. The established metrics used to demonstrate the difference between the two conditions are: total bases (Mb) and Mean, Mode and Median (bp).

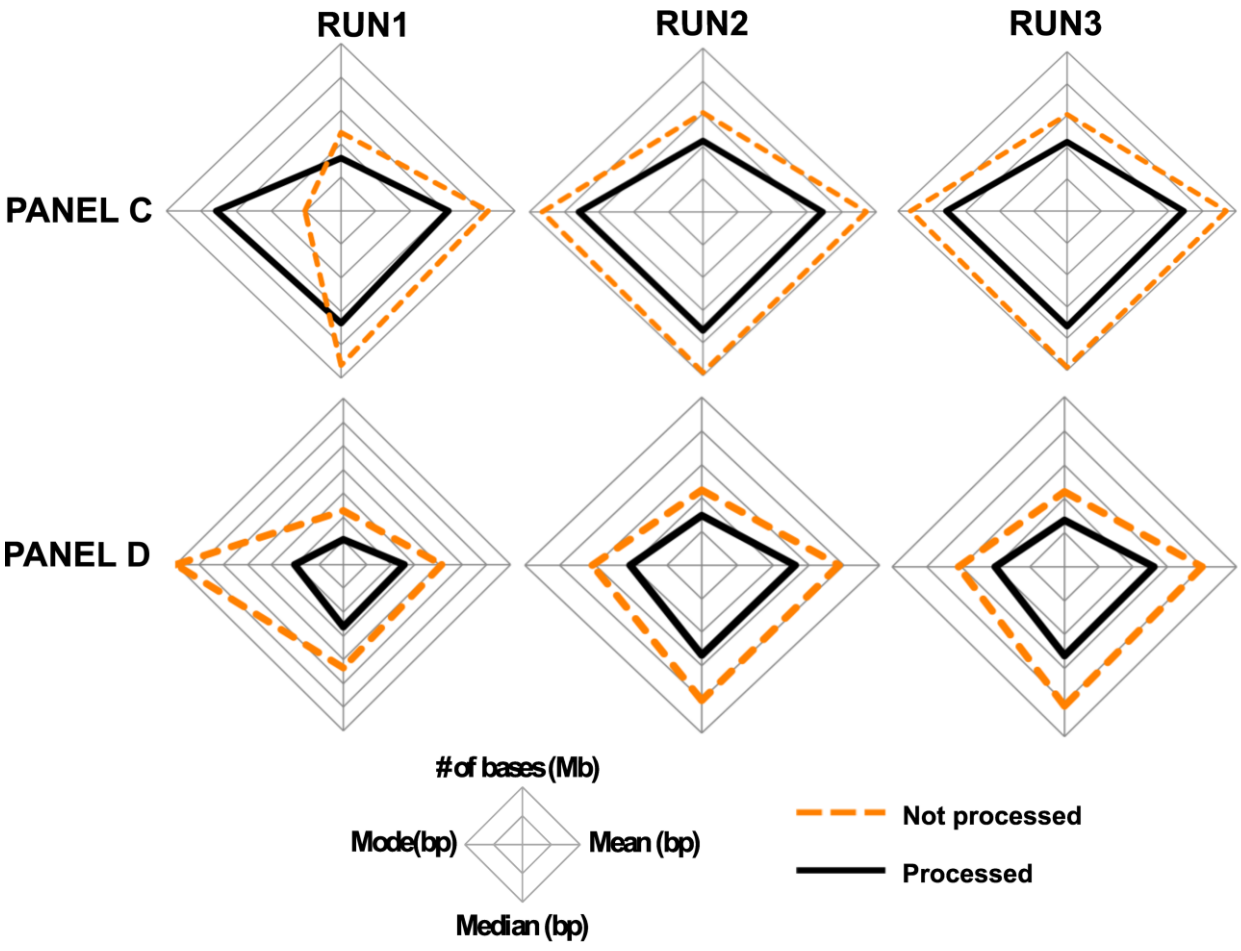


Fig 2. Demonstration of strategy used for removal of low quality reads.

(A) Adapter sequence (in yellow) is trimmed out by searching through the read for candidate matches to the known adapter sequence. (B) The sliding window: the window (size is set to 30 bases) slides trough the read (in blue) searching for low quality calls. In each movement (one base each time), a quality score average is calculated. (C) When the average of the per-base quality score drops below a fixed threshold (15, by default), the sequence is cut from this point (forward 5'). The trim point is just before the earliest (5' most) base. (D) The distribution of quality scores within Ion Torrent reads is such that the highest quality calls tend to occur at the start of the read where signal is strongest and phase errors are smallest in magnitude.

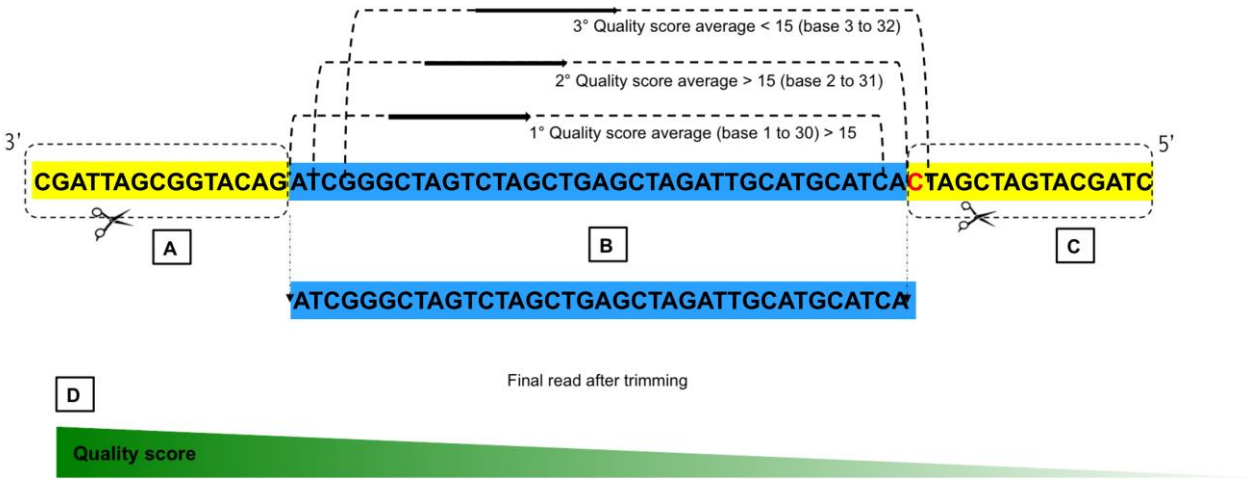
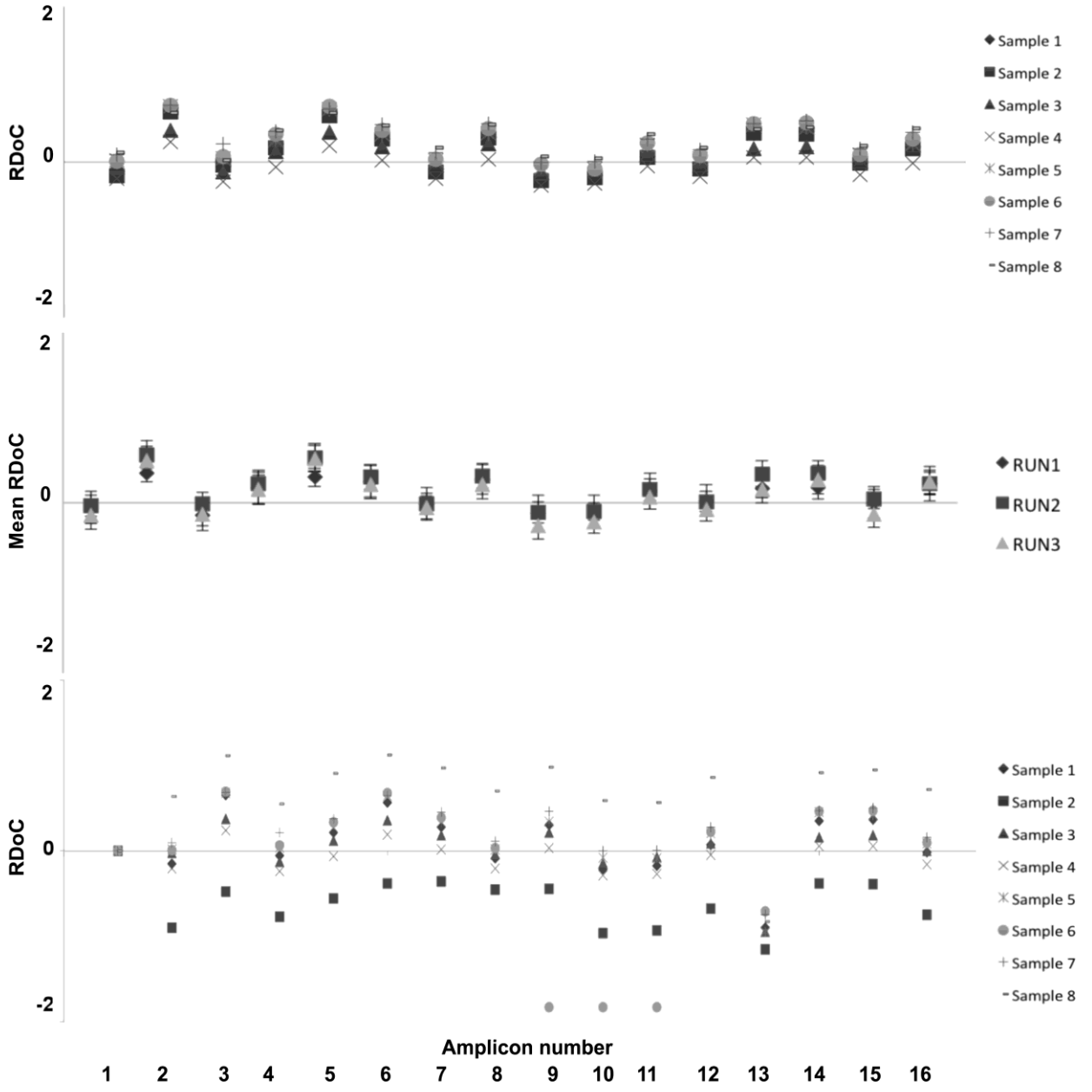


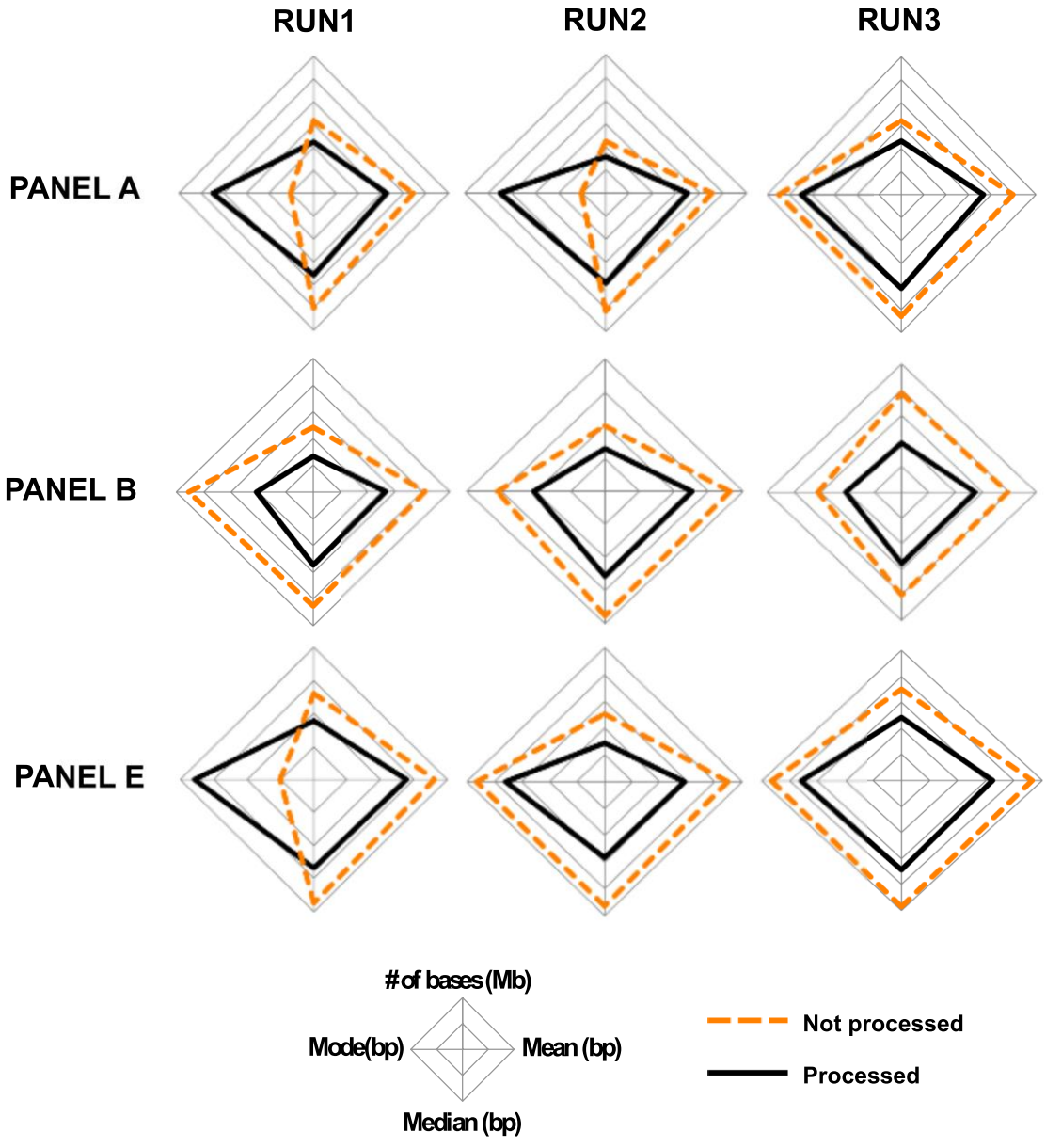
Fig 3. Coverage distribution of amplicons of a gene panel.

(A) Intra-run relative depth of coverage (RDoC) at 16 amplicons (panel E, run 2), corresponding to one gene in 8 samples sequenced in one run. (B) Inter-run reproducibility. This chart represents the mean RDoC at 16 amplicons of panel E in different samples sequenced in three different sequencing runs. Error bars represent standard deviation about the mean of the RDoC of each amplicon in 8 samples analyzed in one run. (C) Simulated coverage scenarios. Intra-run relative depth of coverage (RDoC) at 16 amplicons corresponding to 8 samples sequenced in one run.



S1 Figure. Comparision of two quality conditions altering specific parameters in the Torrent Suite™ software.

To each run of our dataset, two conditions named processed and not processed data was obtained by changing the parameters *disable-all-filters* and *keypass-filter* of base calling. The solid line stands for reads processed using default parameters, while the dotted line represents reads obtained by disabling quality filters. The established metrics used to demonstrate the difference between the two conditions are: total bases (Mb) and Mean, Mode and Median (bp).



CAPÍTULO V

MANUSCRITO II

A comprehensive evaluation of deleterious synonymous variants on human exomes

Artigo em preparação

(Formatado para submissão ao periódico GENE)

A comprehensive evaluation of deleterious synonymous variants on human exomes

Delva Pereira Leão[1,3], Ursula Matte[1,2,3*].

[1]Post-Graduation Program on Genetics and Molecular Biology, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil;
[2]Department of Genetics, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil;
[3]Experimental Research Center, Hospital de Clínicas de Porto Alegre, Porto Alegre, RS, Brazil.

*Corresponding author:
E-mail: umatte@hcpa.edu.br

Abstract

The coming of next-generation sequencing (NGS) technologies has drastically changed the research and diagnostic of genetic disorders. Exome sequencing is the main NGS application used in clinics as it enables global interrogation of coding gene regions. The great amount of information generated on a single exome lead to development of prioritization strategies that are mainly based on filtering of neutral variants. One common step to almost all prioritization workflows is the removal of synonymous variants. More recently, however, several works are demonstrating the functional role of synonymous codon changes on altered cellular homeostasis and its relation with several human diseases. Based on public data of >60,000 exomes obtained from Exome Aggregation Consortium (ExAC) and *in silico* pathogenicity prediction by Silent Variant Analyzer (SilVA), this work demonstrate the landscape of harmfulness of synonymous variants. We show that from 1,691,045 variants classified as synonymous by SilVA, 26,034 were predicted to be pathogenic with allele frequency lower than 0.05. *In silico* functional analysis revealed that pathogenic synonymous variants found are involved in important biological processes, such as cellular regulation, metabolism and transport. By exposing a scenario of pathogenic synonymous variants on human exomes we conclude that filtering out sSNP on prioritization workflows is reasonable, although in some specific cases sSNP should be considered. Future research on this field will provide a clear picture of such variations on genetic diseases.

Introduction

It is well known that the majority of known disease-causing mutations affect highly conserved protein residues. The role of non-synonymous single point mutations on the pathogenesis of genetic diseases has been extensively characterized (Hecht *et* al., 2013; Kumar *et al*, 2014; Nakken *et al*, 2007; Yates & Sternberg, 2013). The coming of next-generation sequencing (NGS) technologies has drastically changed the research and diagnostic of genetic disorders opening a new era in genomics (Goodwin *et* al., 2016; Koboldt et al., 2014; Shen *et al*., 2015). Whole-genome and whole-exome sequencing (WES) strategies allows for global interrogation of human genetic variations, with the latter one being the main NGS clinical application (Seaby *et al*., 2016; Petersen *et* al., 2017; Yang et al., 2013). The term exome refers to coding regions of human genome and it is estimated that ~85% of disease-causing mutations lies in the exome, what justify its recently widespread use on diagnostic practice (Bamshad *et al*., 2011; Botstein & Risch, 2003). For rare diseases of unknown etiology, WES proves to be successful with a diagnostic rate up to 25% (Farwell et al., 2015; Taylor et al., 2015; Yang et al., 2013). Moreover, an individual exome can identify ~30,000 variants when compared with the genomic reference sequence (Smedley *et al*., 2014) leading to a big challenge: prioritizing disease-causing mutations from neutral variants. Prioritization aims to exclude benign variants and also check whether the function of a mutated gene is actually relevant for the disease (Bamshad *et al*., 2011; Pabinger *et al*., 2014). For that, several filtering steps are made based on assumptions which lead to a smaller set of relevant variants (Koboldt et al., 2014; Pfeifer, 2017). An integral part of most filtering workflows is the removal of synonymous single nucleotide polymorphisms (sSNP) (Seaby *et al*., 2016; Stitziel, 2011). However, the functional role of synonymous codon changes on mechanisms essential for

cellular homeostasis, such as splicing and translation accuracy, codon usage bias, mRNA secondary structure, microRNA binding and protein folding became focus of several works (Bali & Bebok, 2014; Chaney & Clark, 2015; Hunt *et* al., 2014; Plotkin & Kudla, 2011; Sauna & Kimchi-Sarfaty, 2011; Shabalina *et al*., 2013). It is now clear that synonymous mutations can affect multiple levels of cellular biology, from DNA and RNA to protein-based features, leading to a wealth of diseases (Hunt *et* al., 2014; Gotea *et al*., 2015; Takata *et al*., 2016). In agreeing with it, efforts to predict harmfulness of synonymous variants became available (Buske *et al*., 2013). Therefore, this article aims to access the landcape of sSNP pathogenicity on public exome data providing new insights on this type of variation to human diseases.

Material and Methods

Dataset

Public data from 60,706 exomes was obtained from Exome Aggregation Consortium (ExAC) (Lek *et al*., 2016) release 0.3.1. FTP link available at ftp://ftp.broadinstitute.org/pub/ExAC_release provides variation data on VCF format. Integrity of 34GB downloaded data was checked with md5sum tag.

Pathogenicity prediction

Silent Variant Analyzer (SilVA) (Buske *et al*., 2013) software was used to perform *in silico* prediction of harmfulness. SilVA is freely available from http://compbio.cs.toronto.edu/silva. It takes a list of variants on VCF format, annotates each one with 26 features and then score variants based on a trained random forest machine learning model. Synonymous variants are ranked and classified as likely benign,

potentially pathogenic and likely pathogenic based on score thresholds of 0.27 and 0.485, respectively.

Functional analysis

To gain insights about the function of genes damaged by synonymous variants revealed by SilVA, we performed an overrepresentation test with PANTHER classification system (Mi *et al*., 2016) version 11.1 using PANTHER GO – Slim Biological Process annotation dataset. Binomial test was applied to compare our findings against a reference list of human genes ($p<0.05$) followed by Bonferroni correction for multiple testing.

Statistical computing

Tabular text file outputted from SilVA was analyzed on R to perform data manipulation, graphs generation and statistical testing using the following packages: dplyr, tidyr, ggplot2 and RcolorBrewer.

Results

Damaging variants

A total of 1,691,045 from 9,362,318 ExAC variants dataset were classified as synonymous by SilVA. Of that, 26,034 variants with allele frequency <0.05 were classified as likely pathogenic (5,862) and potentially pathogenic (20,172), as shown in Fig 1.

Overrepresented genes

Gene ontology (GO) analysis for biological process revealed a total of 46 enriched GO classes with $p<0.05$, shown in Fig 2. More than 25% of our gene list was implicated on the following GO classes: cellular process (GO:0009987), metabolic process (GO:0008152) and primary metabolic process (GO:0044238). Noteworthy, biological regulation

(GO:0065007), regulation of biological process (GO:0050789), transport (GO:0006810), protein transport (GO:0015031), intracellular protein transport (GO:0006886) and intracellular signal transduction (GO:0035556) classes were also represented with statistical significance.

Discussion

Prioritization of disease-causing variants from benign ones is a laborious task, being compared with finding needles in stacks of needles (Cooper *et al*., 2011). Filtering frameworks apply different approaches to reduce the number of variants to a manageable list of candidates (Pfeifer, 2017; Stitiziel *et al*., 2011). Considering this scenario, prioritization turns into a trade off between what is and isn't worthy looking at. Removal of synonymous variants is one of the key assumptions made on prioritization process (Seaby *et al*., 2016). Our work demonstrated a comprehensive evaluation of synonymous variants pathogenicity on human exomes, showing that a relatively small number of variants could be considered as disease-causing candidates. This might be due to limitations of SilVA performance, which is currently restraint by the small number of training examples (Buske *et al*., 2013). With novel pathogenic synonymous variants being experimentally confirmed and development of additional features to better predict harmfulness, SilVA performance could be improved as well as the number of disease-related variants. Despite of that, functional analysis revealed that pathogenic synonymous variants found are involved in important biological process, such as cellular regulation, metabolism and transport. Better characterization of 26,034 variants found is needed to fully understand the relation and potential contribution of each one to human diseases. Lastly, by exposing a scenario of pathogenic synonymous variants on human exomes we conclude that filtering out sSNP on prioritization workflows is reasonable, although in some specific cases sSNP should be

considered. Future research on this field will provide a clear picture of such variations on genetic diseases.

Acknowledgments

We would like to thank Universidade Federal do Rio Grande do Sul (UFRGS) and Hospital de Clínicas de Porto Alegre (HCPA) for their support during this work. Financial support for this study was provided by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, Brazil).

References

Bali V, Bebok Z. Decoding mechanisms by which silent codon changes influence protein biogenesis and function. Int J Biochem Cell Biol. 2015 Jul;64:58-74. doi: 10.1016/j.biocel.2015.03.011.

Bamshad MJ, Ng SB, Bigham AW, et al. Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet 2011;12(11):745–55. doi: 10.1038/nrg3031.

Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nature Genet 2003;33:228–37. doi:10.1038/ng1090.

Buske OJ, Manickaraj A, Mital S, Ray PN, Brudno M. Identification of deleterious synonymous variants in human genomes. Bioinformatics. 2013 Aug 1;29(15):1843-50. doi: 10.1093/bioinformatics/btt308.

Chaney JL, Clark PL. Roles for Synonymous Codon Usage in Protein Biogenesis. Annu Rev Biophys. 2015;44:143-66. doi: 10.1146/annurev-biophys-060414-034333.

Cooper GM, Shendure J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat Rev Genet. 2011 Aug 18;12(9):628-40. doi: 10.1038/nrg3046.

Farwell KD, Shahmirzadi L, El-Khechen D, et al. Enhanced utility of family-centered diagnostic exome sequencing with inheritance model-based analysis: results from 500 unselected families with undiagnosed genetic conditions. Genet Med. 2015 Jul;17(7):578-86. doi: 10.1038/gim.2014.154.

Goodwin S, McPherson JD, McCombie WR.Coming of age: ten years of next-generation sequencing technologies.Nat Rev Genet. 2016 May 17;17(6):333-51. doi:10.1038/nrg.2016.49.

Gotea V, Gartner JJ, Qutob N, et al. The functional relevance of somatic synonymous mutations in melanoma and other cancers. Pigment Cell Melanoma Res. 2015 Nov;28(6):673-84. doi: 10.1111/pcmr.12413.

Hecht M, Bromberg Y, Rost B. News from the protein mutability landscape. J Mol Biol. 2013 Nov 1;425(21):3937-48. doi: 10.1016/j.jmb.2013.07.028.

Hunt RC, Simhadri VL, Iandoli M, et al. Exposing synonymous mutations. Trends Genet. 2014 Jul;30(7):308-21. doi: 10.1016/j.tig.2014.04.006.

Koboldt, D. C. et al. Exome-Based Mapping and Variant Prioritization for Inherited Mendelian Disorders. The American Journal of Human Genetics 2014, 94, 373–384. doi: 10.1016/j.cell.2013.09.006.

Kumar A, Rajendran V, Sethumadhavan R, et al. Computational SNP analysis: current approaches and future prospects. Cell Biochem Biophys. 2014 Mar;68(2):233-9. doi: 10.1007/s12013-013-9705-6.

Lelieveld SH, Veltman JA, Gilissen C. Novel bioinformatic developments for exome sequencing. Hum Genet. 2016 Jun;135(6):603-14. doi: 10.1007/s00439-016-1658-6.

Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. Aug 2016; 536, 285–291. doi:10.1038/nature19057.

Mi H, Huang X, Muruganujan A, et al. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. Nucleic Acids Res. 2017 Jan 4; 45 (D1): D183-D189. doi: 10.1093/nar/gkw1138.

Nakken S, Alseth I, Rognes T. Computational prediction of the effects of non-synonymous single nucleotide polymorphisms in human DNA repair genes. Neuroscience. 2007 Apr 14;145(4):1273-9.

Pabinger S, Dander A, Fischer M, et al. A survey of tools for variant analysis of next-generation genome sequencing data. Brief Bioinform. 2014 Mar; 15(2): 256–278. doi: 10.1093/bib/bbs086.

Petersen BS, Fredrich B, Hoeppner MP, et al. Opportunities and challenges of whole-genome and -exome sequencing. BMC Genet. 2017 Feb 14;18(1):14. doi: 10.1186/s12863-017-0479-5.

Pfeifer SP. From next-generation resequencing reads to a high-quality variant data set. Heredity (Edinb). 2017 Feb;118(2):111-124. doi: 10.1038/hdy.2016.102.

Plotkin JB, Kudla G. Synonymous but not the same: the causes and consequences of codon bias. Nat Rev Genet. 2011 Jan;12(1):32-42. doi: 10.1038/nrg2899.

Sauna ZE, Kimchi-Sarfaty C. Understanding the contribution of synonymous mutations to human disease. Nat Rev Genet. 2011 Aug 31;12(10):683-91. doi: 10.1038/nrg3051.

Seaby EG, Pengelly RJ, Ennis S. Exome sequencing explained: a practical guide to its clinical application. Brief Funct Genomics. 2016 Sep;15(5):374-84. doi: 10.1093/bfgp/elv054.

Shabalina SA, Spiridonov NA, Kashina A. Sounds of silence: synonymous nucleotides as a key to biological regulation and complexity. Nucleic Acids Res. 2013 Feb; 41(4): 2073–2094. doi: 10.1093/nar/gks1205.

Shen T, Lee A, Shen C, Lin CJ. The long tail and rare disease research: the impact of next-generation sequencing for rare Mendelian disorders. Genet. Res., Camb. (2015), vol. 97, e15. doi:10.1017/S0016672315000166.

Smedley, D. et al. Walking the interactome for candidate prioritization in exome sequencing studies of Mendelian diseases. Bioinformatics. 2014,Vol. 30 no. 22,3215–3222. doi: 10.1093/bioinformatics/btu508.

Stitziel, N. O. et al. Computational and statistical approaches to analyzing variants identified by exome sequencing. Genome Biology 2011, 12:227. doi: 10.1186/gb-2011-12-9-227.

Takata A, Ionita-Laza I, Gogos JA, et al. De Novo Synonymous Mutations in Regulatory Elements Contribute to the Genetic Etiology of Autism and Schizophrenia. Neuron. 2016 Mar 2;89(5):940-7. doi: 10.1016/j.neuron.2016.02.024.

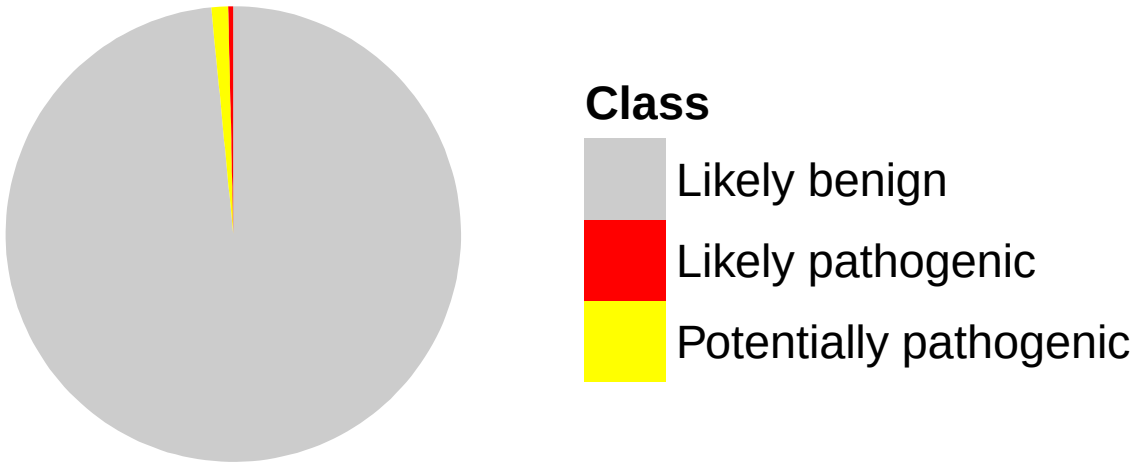
Taylor JC, Martin HC, Lise S, et al. Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. Nat Genet 2015;47:717–26. doi: 10.1038/ng.3304.

Yang Y, Muzny DM, Reid JG, et al. Clinical whole-exome sequencing for the diagnosis of mendelian disorders. N Engl J Med 2013;369(16):1502–11. 5. doi: 10.1056/NEJMoa1306555.

Yates CM, Sternberg MJ. The effects of non-synonymous single nucleotide polymorphisms (nsSNPs) on protein-protein interactions. J Mol Biol. 2013 Nov 1;425(21):3949-63. doi: 10.1016/j.jmb.2013.07.012.

Fig 1. Classification of synonymous variants according to SilVA.

Silent Variant Analyzer was used for processing ExAC dataset (release 0.3.1) revealing a total of 1,691,045 synonymous variants. From that, a total of 26,034 variants were classified as likely pathogenic (5,862) or potentially pathogenic (20,172) with allele frequency <0.05.



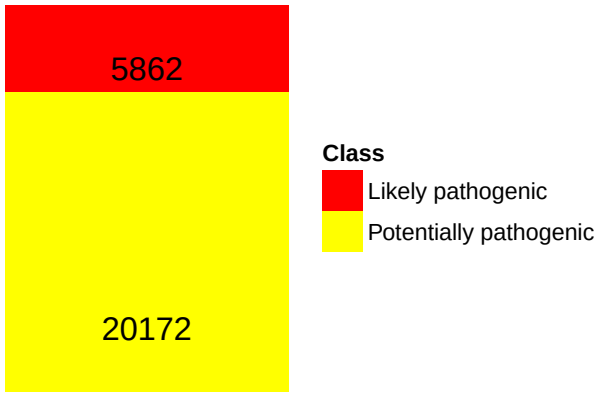
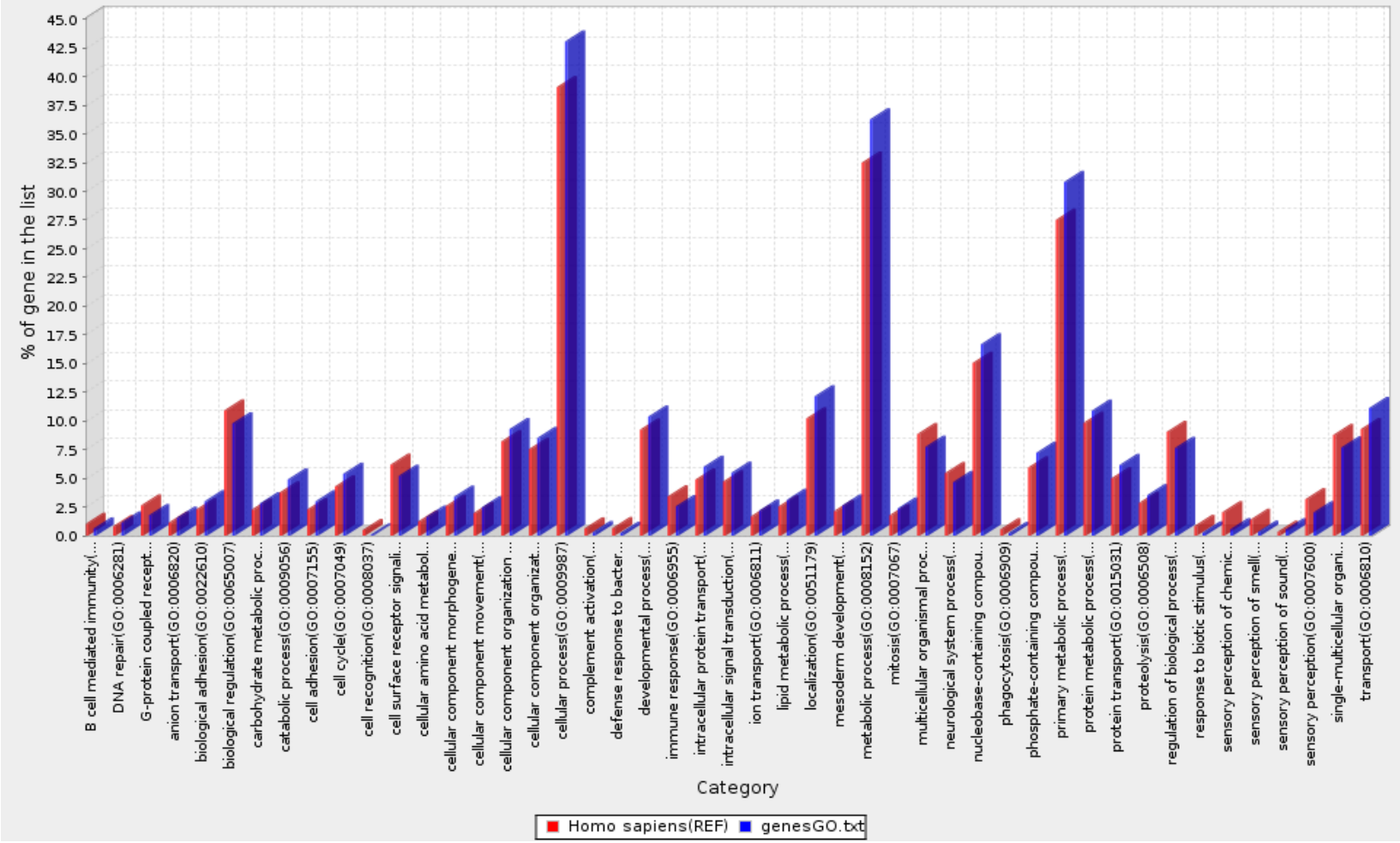


Fig 2. Overrepresentation test results according to PANTHER GO version 11.1.

Genes pointed by SilVA as carring a pathogenic synonymous variant were submited to *in silico* functional analysis using the tool PANTHER GO. Gene ontology (GO) analysis for biological process revealed a total of 46 enriched GO classes with $p<0.05$.



CAPÍTULO VI

DISCUSSÃO

As discussões específicas aos resultados obtidos no presente estudo encontram-se nos manuscritos apresentados no Capítulo IV e V. Neste capítulo serão mencionados aspectos mais gerais referentes ao tema, retomando questões não discutidas anteriormente.

Mais do que ser uma novidade, a tecnologia de sequenciamento de nova geração agora faz parte da rotina da pesquisa biológica e tornou possível a realização de pesquisas consideradas impossíveis até poucos anos atrás. Com o NGS integrado à prática clínica, iniciativas recentes como programas de medicina de precisão estão sendo desenvolvidos e aplicados em todo o mundo. Assim, sequenciar a um custo relativamente baixo e de forma veloz proporciona aos médicos e pesquisadores as ferramentas necessárias para traduzir informações genômicas em ações clínicas. Entretanto, essa revolução abriga um novo conjunto de desafios. Um deles é a transição de tecnologias tradicionais para o NGS, uma questão relevante abordada no primeiro manuscrito deste trabalho.

Outro ponto, porém, merece ser destacado: a interpretação de variantes genéticas. A complexidade advinda do NGS para classificação de novas variantes com significado incerto é surpreendente. Menos de duas décadas atrás, era comum que pesquisadores ao encontrarem uma variante nova, interrogando poucas regiões genômicas, classificassem-na como mutação causal a partir apenas da comprovação da sua ausência em um número restrito de indivíduos normais. Uma ação justificável, pois, no caso de uma doença Mendeliana, ao estudar o gene responsável pela condição, em um paciente com o fenótipo característico, os indícios apontavam para esta direção. É preciso considerar a escassez de recursos ou, quem sabe, a falta de necessidade, por parte da comunidade científica, de realizar testes adicionais, como ensaios de expressão, por exemplo, para confirmar a relação causa-efeito. As ferramentas disponíveis eram estas e inúmeros artigos foram publicados seguindo essa linha de raciocínio.

Com o NGS o paradigma mudou. Passamos a ter acesso à variação genética global e com isso a compreensão sobre a inferência de patogenicidade e significância clínica de uma variante genética sofreram mudanças drásticas. Hoje, uma nova variante é considerada como sendo de significado incerto, até que se prove o contrário. E para tal, inúmeros esforços têm sido realizados. O *American College of Human Genetics* lançou, em 2015, um *guideline* para a interpretação destas variantes, recomendando terminologias específicas e uma tabela descrevendo o processo para classificação de variantes de significado incerto (Richards *et al*., 2015). Esforços também tem sido empreendidos no

desenvolvimento de ferramentas e softwares que auxiliem no processo de priorização e predição de patogenicidade. Ainda assim, muitas das informações genéticas consideradas como menos importantes também têm demonstrado possuir um papel relevante em doenças humanas, como é o caso das variantes sinônimas, exploradas no segundo manuscrito deste trabalho.

Enquanto a prática do diagnóstico genético utilizando NGS continua a ser desafiada, outro cenário torna-se relevante: a quantidade de dados que as plataformas de sequenciamento têm gerado. Em 2013, foi reportado que o mundo pode gerar ~15 *petabytes* (10^{15} bytes) de dados de sequenciamento por ano (Schatz & Langmead, 2013) e o número de plataformas só cresceu desde então. Essa quantidade de dados têm desafiado a análise e infraestrutura computacional necessária e inovações para o armazenamento de dados e soluções de bioinformática serão de extrema valia em um futuro próximo. Assim, torna-se claro que a bioinformática é um ramo de pesquisa desafiador e promissor. Talvez o maior de todos os desafios seja traduzir a vasta quantidade de dados genéticos em um contexto biológico relevante. Certamente ainda há muito para avançarmos até a compreensão holística da complexidade biológica dos seres vivos. E, sem dúvidas, o NGS tem o potencial necessário para nos fazer chegar a este ponto.

CAPÍTULO VII

CONCLUSÕES

Os resultados obtidos neste trabalho cumpriram os objetivos que haviam sido propostos:

1. Avaliar com precisão o desempenho de sequenciamento e performance de painéis de genes em dados de *targeted gene sequencing*.

Evidenciamos indicadores adequados para avaliar a qualidade das corridas de sequenciamento da plataforma Ion Torrent, bem como parâmetros que asseguram a qualidade dos dados gerados. Também demonstramos, através de um método baseado em profundidade de cobertura, como avaliar o desempenho de cada região (*amplicon*) que compõe o painel de genes de interesse.

2. Avaliar a patogenicidade de variantes sinônimas raras obtidas dados públicos de sequenciamento de exoma completo.

Utilizamos dados públicos de mais de 60.000 exomas e demonstramos um panorama da patogênicidade de variantes sinônimas. Um total de 26.034 variantes raras (com frequência alélia menor que 0,05) foram classificadas como patogênicas. Análises funcionais revelaram que as variantes sinônimas patogênicas estão envolvidas em processos biológicos importantes, como regulação celular, metabolismo e transporte.

CAPÍTULO VIII

PERSPECTIVAS

Como perpectivas e sugestões adicionais de análise destacamos as seguintes, específicas do manuscrito apresentado no capítulo V:

- Caracterizar as variantes sinônimas patogênicas encontradas quanto à localização em domínios protéicos, *folding* de mRNA e outras evidências de consequências funcionais que nos permitam corroborar com a classificação fornecida pelo SilVA;

- Explorar a frequência alélica das variantes sinônimas patogênicas de forma específica a cada população que compõe o ExAC;

- Investigar se os genes que abrigam as variantes sinônimas patogênicas estão associados à doenças Mendelianas;

- Explorar a ocorrência de seleção purificadora nas regiões que abrigam as variantes sinônimas patogênicas.

REFERÊNCIAS BIBLIOGRÁFICAS

Alemán A, Garcia-Garcia F, Salavert F, et al (2014) A web-based interactive framework to assist in the prioritization of disease candidate genes in whole-exome sequencing studies. Nucleic Acids Research 42 (W1): W88-W93.

Ambardar S, Gupta R, Trakroo D, et al (2016) High Throughput Sequencing: An Overview of Sequencing Chemistry. Indian J Microbiol 56(4):394-404.

Bamchad, MJ. et al (2011) Exome sequencing as a tool for Mendelian disease gene discovery. Nature Reviews Genetics 12, 745-755.

Buske OJ, Manickaraj A, Mital S, Ray PN, Brudno M (2013) Identification of deleterious synonymous variants in human genomes. Bioinformatics 29(15):1843-50.

Chen, R. et al. (2010) Non-synonymous and synonymous coding SNPs show similar likelihood and effect size of human disease association. PLoS ONE 5(10):e13574

Cooper GM, Shendure J (2011) Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nature Reviews Genetics 12(9):628-40.

Dallol A, Daghistani K, Elaimi A, et al (2016) Utilization of amplicon-based targeted sequencing panel for the massively parallel sequencing of sporadic hearing impairment patients from Saudi Arabia. BMC Med Genet 17(Suppl 1):67

Danecek P, Auton A, Abecasis G, et al (2011) The variant call format and VCFtools. Bioinformatics 27(15): 2156–2158.

Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 8(3):186-94.

Farwell KD, Shahmirzadi L, El-Khechen D, et al (2015) Enhanced utility of family-centered diagnostic exome sequencing with inheritance model-based analysis: results from 500 unselected families with undiagnosed genetic conditions. Genet Med 17(7):578-86.

Fu W, O'Connor TD, Jun G, et al (2013) Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature 493, 216–220

Gilissen C, Hoischen A, Brunner HG, Veltman JA (2012) Disease gene identification strategies for exome sequencing. Eur J Hum Genet 20(5): 490–497.

Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: ten years of next-generation sequencing technologies.Nat Rev Genet 17(6):333-51.

Goldstein DB,Allen A, Keebler J, et al (2013) Sequencing studies in human genetics: design and interpretation. Nat Rev Genet 14(7):460-70.

Hegele RA, Ban MR, Cao H, et al (2015) Targeted next-generation sequencing in monogenic dyslipidemias. Curr Opin Lipidol 26(2):103-13.

Hunt RC, Simhadri VL, Iandoli M, et al (2014) Exposing synonymous mutations. Trends Genet 30(7):308-21.

Jiang, R (2015) Walking on multiple disease-gene networks to prioritize candidate genes. J Mol Cell Biol 7(3):214-30.

Johansen CT, Dubé JB, Loyzer MN, et al (2014) LipidSeq: a next-generation clinical resequencing panel for monogenic dyslipidemias. J Lipid Res 55(4):765-72.

Katsonis P, Koire A, Wilson SJ, et al (2014) Single nucleotide variations: Biological impact and theoretical interpretation. Protein Sci 23(12):1650-66.

Lek M, Karczewski KJ, Minikel EV, et al (2016) Analysis of protein-coding genetic variation in 60,706 humans. Nature 536, 285–291.

Li H, Homer N (2010) A survey of sequence alignment algorithms for next-generation sequencing. Brief Bioinform 11(5): 473–483.

Li H, Handsaker B, Wysoker A, et al (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25(16): 2078–2079.

Li M, Kwan JSH, Bao S, et al (2013) Predicting Mendelian Disease-Causing Non-Synonymous Single Nucleotide Variants in Exome Sequencing Studies. PLoS Genet 9(1): e1003143.

Li MX, Gui HS, Kwan JS, et al (2012) A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. Nucleic Acids Research 40(7):e53.

Lim EC, Brett M, Lai AH, et al (2015) Next-generation sequencing using a pre-designed gene panel for the molecular diagnosis of congenital disorders in pediatric patients. Hum Genomics 9:33.

Manolio, T. A. et al (2009) Finding the missing heritability of complex diseases. Nature 461, 747-753.

Margulies M, Egholm M, Altman WE, et al (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 15;437(7057):376-80.

Metzker, ML (2010) Sequencing technologies - the next generation. Nat Rev Genet 11(1):31-46.

Mielczarek M & Szyda J (2016) Review of alignment and SNP calling algorithms for next-generation sequencing data.J Appl Genetics 57:71–79

Moreau Y, Tranchevent LC (2012) Computational tools for prioritizing candidate genes: boosting disease gene discovery. Nat Rev Genet 13(8):523-36.

Morey M, Fernández-Marmiesse A, Castiñeiras D, et al (2013) A glimpse into past, present, and future DNA sequencing. Mol Genet Metab 110(1-2):3-24.

Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48(3):443–453

Ohanian M, Otway R, Fatkin D (2012) Heuristic Methods for Finding Pathogenic Variants in Gene Coding Sequences. J Am Heart Assoc 1(5): e002642.

Pabinger S, Dander A, Fischer M, et al (2014) A survey of tools for variant analysis of next-generation genome sequencing data. Brief Bioinform 15(2): 256–278.

Park ST, Kim J (2016) Trends in Next-Generation Sequencing and a New Era for Whole Genome Sequencing. Int Neurourol J 20(Suppl 2): S76-83.

Petersen B, Fredrich B, Hoeppner MP, et al (2017) Opportunities and challenges of whole-genome and -exome sequencing. BMC Genetics 18:14

Pfeifer SP (2017) From next-generation resequencing reads to a high-quality variant data set.Heredity (Edinb) 118(2):111-124.

Plotkin JB, Kudla G (2011) Synonymous but not the same: the causes and consequences of codon bias. Nat Rev Genet 12(1):32-42.

Rehm HL (2017) Evolving health care through personal genomics. Nat Rev Genet 18(4):259-267.

Reinert K, Langmead B, Weese D, Evers DJ (2015) Alignment of Next-Generation Sequencing Reads.Annu Rev Genomics Hum Genet 16:133-51.

Reuter JA, Spacek DV, Snyder MP (2015) High-Throughput Sequencing Technologies. Mol Cell 58(4): 586–597.

Richards S, Aziz N, Bale S, et al (2015) Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genetics in Medicine 17, 405–423

Robinson PN, Kohler S, Oellrich A, et al (2014) Improved exome prioritization of disease genes through cross-species phenotype comparison. Genome Research 24:340–348.

Samorodnitsky E, Datta J, Jewell BM, et al (2015) Comparison of Custom Capture for Targeted Next-Generation DNA Sequencing. J Mol Diagn 17(1): 64–75.

Sauna ZE, Kimchi-Sarfaty C (2011) Understanding the contribution of synonymous mutations to human disease. Nat Rev Genet 12(10):683-91.

Schatz, M. C. & Langmead, B (2013) The DNA data deluge: fast, efficient genome sequencing machines are spewing out more data than geneticists can analyze. IEEE Spectr. 50, 26–33.

Shearer AE, Eppsteiner RW, Booth KT, et al (2014) Utilizing Ethnic-Specific Differences in Minor Allele Frequency to Recategorize Reported Pathogenic Deafness Variants. Am J Hum Genet 95(4):445-53.

Shen T, Lee A, Shen C, Lin CJ (2015) The long tail and rare disease research: the impact of next-generation sequencing for rare Mendelian disorders. Genet Res 14;97:e15.

Sims D, Sudbery I, Ilott NE, et al (2014) Sequencing depth and coverage: key considerations in genomic analyses. Nature Reviews Genetics 15, 121–132

Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147:195–197

Smedley, D, Kohler S, Czeschik JC (2014) Walking the interactome for candidate prioritization in exome sequencing studies of Mendelian diseases. Bioinformatics 15; 30(22): 3215–3222.

Stitziel NO, Kiezun A, Sunyaev S (2011) Computational and statistical approaches to analyzing variants identified by exome sequencing. Genome Biol 14;12(9):227.

Sun Y, Ruivenkamp CAL, Hoffer MJV, et al (2015) Next-Generation Diagnostics: Gene Panel, Exome, or Whole Genome? Hum Mutat 36(6):648-55

Suzuki T, Tsurusaki Y, Nakashima M, et al (2014) Precise detection of chromosomal translocation or inversion breakpoints by whole-genome sequencing.J Hum Genet 59(12):649-54.

Tian S, Yan H, Kalmbach M, Siager SL (2016) Impact of post-alignment processing in variant discovery from whole exome data. BMC Bioinformatics 17:403.

van Ninwegen KJM, van Soest RA, Veltman JA, et al (2016) Is the $1000 Genome as Near as We Think? A Cost Analysis of Next-Generation Sequencing. Clinical Chemistry 62:11;1458–1464.

Worthey EA, Mayer AN, Syverson GD, et al (2011) Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. Genet. Med 13(3):255-62.

Wu L, Schaid DJ, Sicotte H, et al (2015) Case-only exome sequencing and complex disease susceptibility gene discovery: study design considerations. J Med Genet 52(1): 10–16.

Yang Y, Muzny DM, Reid JG, et al (2013) Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders. N Engl J Med 369(16):1502-11.

Yu JH, Jamal SM, Tabor HK, Bamshad MJ (2013) Self-guided management of exome and whole-genome sequencing results: changing the results return model. Genet Med 15(9):684-90.