Universidade de São Paulo
Programa de Pós-graduação em Bioinformática

**Busca de variantes em sequência de DNA proveniente de pacientes com
deficiência em processos de reparo do genoma**

Livia Maria Silva Moura

Orientação:

Dr. Carlos Frederico Martins Menck

Co-orientação:

Dr. Pedro A. F. Galante

São Paulo, 2015

Livia Maria Silva Moura

**Busca de variantes em sequência de DNA proveniente de pacientes com
deficiência em processos de reparo do genoma**

Dissertação apresentada ao Programa de pós-graduação em Bioinformática da Universidade de São Paulo, como parte dos requisitos para obtenção do título de Mestre em Bioinformática.

Área de concentração: Bioinformática

Orientação: Prof. Dr. Carlos Frederico Martins Menck

Co-orientação: Prof. Dr. Pedro A. F. Galante

São Paulo, 2015

AGRADECIMENTOS

A Deus por ter me dado forças para não desanimar e saúde para continuar lutando por um futuro melhor, apesar de todas as dificuldades que enfrentei.

Aos meus pais sempre presentes em minha vida, sempre apostando em minha capacidade, sempre desejando o meu melhor e me amando incondicionalmente.

Aos meus irmãos que sempre me apoiaram, sofreram e vibraram comigo em várias passagens nesse período.

A minha tia Fátima, que Deus a tenha, por ter incentivado os meus estudos desde pequena até o dia que partiu.

Aos meus orientadores Dr. Carlos Menck e Dr. Pedro Galante, pela grande contribuição na minha formação científica, pela paciência e incentivo, dando-me a oportunidade de desenvolver este trabalho, e me ensinando em todo o tempo. Meus sinceros agradecimentos pela confiança e conhecimentos repassados.

Ao meu grande amigo Fábio Higa, por estar presente em vários momentos dessa minha jornada e pela sua colaboração no desenvolvimento deste projeto.

A grande amiga que adquiri durante o mestrado, a Dr. Huma Asif, por todas as conversas, conselhos, amizade e ensinamentos.

Aos meus colegas de laboratório pelos momentos de convivência e amizade durante esses anos.

A todos os professores da Pós-Graduação com os quais tive a oportunidade e o prazer de aprender.

Aos colegas da pós-graduação pela convivência e amizade durante todo o curso.

As minhas amigas com quem tive o prazer de dividir o mesmo teto e compartilhar risadas e ótimos momentos. Muito obrigada pelo suporte quando mais precisei.

A todos aqueles que de maneira direta ou indireta participaram da minha formação profissional e permitiram que este estudo fosse realizado, meus sinceros agradecimentos.

A FAPESP e a CAPES, pelo apoio financeiro que me permitiram realizar esse trabalho.

SUMÁRIO

Lista de abreviaturas

Lista de Figuras

Lista de Tabelas

Lista de Anexos

Resumo

Abstract

1. INTRODUÇÃO	1
2. OBJETIVOS	11
2.1. OBJETIVOS GERAIS.....	11
2.2. OBJETIVOS ESPECÍFICOS	11
3. METODOLOGIA.....	11
3.1. AMOSTRAS E OBTENÇÃO DE SUAS SEQUÊNCIAS NUCLEOTÍDICAS	11
3.2. ANÁLISE DAS SEQUÊNCIAS E PARÂMETRO DE QUALIDADE	15
3.3. BUSCA DE SUBSTITUIÇÕES E <i>INDELS</i>	15
3.3.1. Primeira etapa:	16
3.3.1.1. SAMTOOLS	17
3.3.1.2. GATK.....	18
3.3.1.3. GATK+SAMTOOLS.....	21
3.3.2. Segunda etapa:	22
3.4. DESCOBRINDO O POTENCIAL PATOGÊNICO DAS VARIANTES	23
3.5. DESENVOLVIMENTO DA INTERFACE <i>WEB</i> DE DOMÍNIO PÚBLICO.....	24
3.5.1. O script da plataforma	25
3.6. INTERFACE <i>WEB</i> À PROVA: AMOSTRAS PROVENIENTES DA PLATAFORMA MISEQ (ILLUMINA).....	26
3.6.1. Processamento das amostras Illumina	27
4. RESULTADOS E DISCUSSÃO	28

4.1.	A QUALIDADE DO SEQUENCIAMENTO	28
4.2.	EM BUSCA DE VARIANTES	29
4.3.	A SAÍDA FINAL	35
4.4.	A ANÁLISE DAS VARIANTES POTENCIONALMENTE PATOGÊNICAS.....	38
4.5.	A PLATAFORMA DE DOMÍNIO PÚBLICO E SUA INTERFACE	42
4.6.	ANÁLISE DAS AMOSTRAS ILLUMINA: TESTANDO A EFETIVIDADE DA INTERFACE <i>WEB</i>	46
5.	CONCLUSÃO GERAL	49
6.	REFERÊNCIAS	50

Anexos

LISTA DE ABREVIATURAS

6-4 PP: fotoproduto pirimidina 6-4 pirimidona (do inglês 6-4 Photoproduct)

8-oxodG: 8-oxo-7,8-dihidro-20-desoxiguanosina

BAM: formato binário ordenado, indexado e extremamente compacto de nucleotídeos alinhados (do inglês *Binary Alignment/Mapping*)

bp: pares de base (do inglês *base pair*)

CPD: dímero de pirimidina ciclobutano (do inglês *cis-syn Cyclobutane Pyrimidine Dimers*)

CS: Síndrome de Cockayne (do inglês *Cockayne Syndrome*)

DNA: ácido desoxirribonucleico

GGR: reparo do genoma global (do inglês *Global Genome Repair*)

HC: do inglês *Haplotype Caller*

Indel: mutação gerada pela inserção ou deleção de um ou mais nucleotídeos

NER: reparo por excisão de nucleotídeos (do inglês *Nucleotide Excision Repair*)

NGS: plataforma de sequenciamento da nova geração (do inglês *Next Generation Sequencing*)

PCR: reação em cadeia da polimerase (do inglês *Polymerase Chain Reaction*)

RPA: Replication Protein A

SBS: sequenciamento por síntese (do inglês *Sequencing by synthesis*)

Sc: parâmetro de busca de variantes *-stand_call_conf*

Se: parâmetro de busca de variantes *-emit_call_conf*

TCR (ou TC-NER): reparo acoplado à transcrição (do inglês *Transcription-Coupled Repair*)

TFIIH: fator de transcrição (do inglês *Transcription Factor*) IIH

TLS: síntese translesão (do inglês *Translesion Synthesis*)

TTD: tricotiodistrofia

UC: do inglês *Unified Genotyper*

UV: luz ultravioleta

UVA: luz ultravioleta no comprimento de onda de 315 a 400 nm

UVB: luz ultravioleta no comprimento de onda de 280 a 315 nm

UVC: luz ultravioleta no comprimento de onda de 200 a 280 nm

VCF: arquivo que contém variantes chamadas (do inglês *variant call format*),

XP: Xeroderma Pigmentosum

LISTA DE FIGURAS

Figura 1 - Representação esquemática do sequenciamento pela plataforma SOLiD..	8
Figura 2 - Representação esquemática do <i>bridge</i> PCR das plataformas da Illumina.	9
Figura 3 - Genealogia dos pacientes de Araras-GO.	14
Figura 4 - Demonstração de um alinhamento antes e após se remover os reads. ...	17
Figura 5 - Demonstração das qualidades de <i>reads</i> antes e após a recalibração de bases.	21
Figura 6 - Fluxograma que resume a primeira etapa na busca por variantes.	22
Figura 7 - Heredograma de alguns pacientes que tiveram suas sequências nucleotídicas analisadas pela plataforma da Illumina.	27
Figura 8 - <i>Layout</i> principal da <i>Human Variants Finder Interface</i> desenvolvida e disponibilizada.	43
Figura 9 - <i>Layout</i> de conclusão de tarefas da <i>Human Variants Finder Interface</i>	43
Figura 10a - Representação dos resultados finais já anotados bem como as diversas possibilidades de filtros que podem ser realizados com o arquivo VCF final (informações iniciais que são obtidas).	44
Figura 10b - Representação dos resultados finais bem como as diversas possibilidades de filtros que podem ser realizados com o arquivo VCF final (informações finais que são obtidas).	45
Figura 11 - Demonstração de como estão sendo organizados os bancos de dados na página MySQL acessado localmente.	46

LISTA DE TABELAS

Tabela 1 - Relação das amostras sequenciadas segundo o tecido de extração do DNA.....	13
Tabela 2 - Genes abordados na análise.....	13
Tabela 3 - Resumo do relatório de enriquecimento da amostra pelo método de SureSelect.	31
Tabela 4 - Quantidades de variantes chamadas pelos 3 pipelines de busca distintos utilizados na primeira etapa.	32
Tabela 5 - As variantes restante foram divididos em filtrados e os novos.	32
Tabela 6 - Resultados ao se comparar a quantidade de <i>reads</i> filtrados entre os dois buscadores de variantes do programa GATK.	34
Tabela 7 - Comparação da quantidade de <i>indels</i> chamado pelos dois buscadores de variantes do programa GATK.	34
Tabela 8 - Comparação de diferentes valores de busca envolvendo os parâmetros - <i>stand_call_conf</i> e - <i>emit_call_conf</i>	37
Tabela 9 - Exemplo do tipo de saída e as algumas informações resultantes de cada análise.	38
Tabela 10 - Levantamento das possíveis variantes que possam justificar o diagnóstico clínico dos pacientes, muitas delas ainda não descritas ou reportadas em banco de dados (SOLiD).	40
Tabela 11 - Resultados gerados pelas plataformas PROVEAN e SIFT para as potenciais variantes candidatas ao perfil clínico dos pacientes (SOLiD)	42
Tabela 12 - Levantamento das possíveis variantes que possam justificar o diagnóstico clínico dos pacientes, muitas delas ainda não descritas ou reportadas em banco de dados (Illumina).	48
Tabela 13 - Resultados gerados pelas plataformas PROVEAN e SIFT para as potenciais variantes candidatas ao perfil clínico dos pacientes sequenciados pelo MiSeq Illumina.	49

LISTA DE ANEXOS

Anexo 1

Aqui estão as variantes filtradas de todas as amostras sequenciados por SOLiD que não estão no banco de dados do 1000genomes, removendo as variantes intrônicas, com exceção dos sítios de *splicing*.

Anexo 2

Análise das amostras sequenciadas pela plataforma MiSeq da Illumina.

Anexo 3

Parecer do Comitê de ética em Pesquisa do projeto raiz.

RESUMO

Apesar de altamente estável, o DNA sofre milhares de alterações em sua estrutura diariamente, sejam essas espontâneas ou pela exposição a agentes mutagênicos. A maior parte dessas alterações é prontamente removida por um conjunto de eventos de reparo de DNA. A via de reparo por excisão de nucleotídeos (NER) é a mais versátil e flexível lidando com uma variedade de lesões que podem gerar distorções das hélices do DNA. Esses danos resultam em alterações características que, caso não reparadas, podem gerar mutações ou morte celular e, conseqüentemente, câncer e envelhecimento. Algumas síndromes, nas quais os pacientes são sensíveis à luz solar, estão relacionadas à deficiência no processo de NER, como a Xeroderma Pigmentosum (XP), síndrome de Cockayne (CS) e Tricotiodistrofia (TTD).

Indivíduos brasileiros, incluindo pacientes com diagnóstico clínico de XP e membros das famílias, passaram por um processo *in silico* para a identificação variantes em genes relacionados aos processos de reparo do DNA após o sequenciamento do DNA por plataformas de nova geração (NGS: plataforma ABI 5500XL SOLiD e MiSeq Illumina) e análises de Bioinformática. Para cada paciente, foram selecionados os melhores valores de parâmetros para se realizar a busca por variantes considerando a qualidade de alinhamento e a taxa de cobertura das bases alvo. SNPs já depositados no banco de dados do projeto 1000genomes foram removidos de nossos dados. O restante das variantes foi analisado para encontrar potenciais candidatos que poderiam explicar o diagnóstico clínico do paciente. Em muitas amostras foi possível determinar pelo menos uma variante (mutação) com uma elevada possibilidade de ser responsável pelos sintomas XP. Para alguns pacientes, a má qualidade do sequenciamento ou eventos não esclarecidos durante este, dificultou a identificação de candidatos à mutação patogênica. Potenciais mutações não sinônimas foram analisadas com os programas SIFT e PROVEAN, que identificaram a potencial capacidade deletéria da alteração de aminoácido na proteína. Finalmente, foi desenvolvida uma interface de domínio público amigável, a *Human Variantes do Finder Interface* (<http://www.varfinderhg.com.br>), que visa facilitar a identificação de variantes em dados gerados por NGS.

Abstract

Although highly stable, DNA molecule undergoes thousands of damage in its structure every day, due to spontaneous lesions or exposure to various mutagens. Most of these lesions are readily removed by a number of cellular DNA repair processes. The process of nucleotide excision repair (NER) is the most versatile and flexible dealing with a variety of lesions that can lead to distortions of the DNA strands. Ultraviolet irradiation induced DNA damage are the main substrates for NER. These DNA damage, if not repaired, can generate mutations or cell death causing several diseases, including cancer and aging. Some syndromes, sensitive to sunlight, are related to deficiencies in the NER process, such as Xeroderma Pigmentosum (XP), Cockayne syndrome (CS) and Trichothiodystrophy (TTD).

Brazilian individuals, including patients with clinical diagnosis of XP and family members, went through *in silico* process for the identification of variants in genes related to DNA repair processes after DNA sequencing by next generation sequencing (NGS in the platforms ABI 5500XL SOLiD and MiSeq Illumina) and dedicated Bioinformatics pipelines. For each patient the best search pattern of variant calling was used considering the alignment quality and coverage rate of bases in target. SNPs already deposited at the 1000genomes project database were removed from the data. The remaining variants were analyzed to find potential candidates that could explain the clinical diagnosis. In many samples, it was possible to determine at least one variant (mutation) with a high possibility of being responsible for the clinical XP. For some patients, the poor quality of the sequencing or unclear events during sequencing hampered the identification of clear mutation candidates. Potential nonsynonymous mutations were analyzed with SIFT and PROVEAN softwares, which identified the potential deleterious capacity of the amino acid change in the protein. Finally, we developed a user-friendly public domain interface, the *Human Variants Finder Interface* (<http://www.varfinderhg.com.br>), which, we expect, will facilitate the identification of variants in data generated by NGS.

1. INTRODUÇÃO

O material genético dos seres vivos precisa ser altamente estável devido sua importância no armazenamento da informação genética, essencial à vida do organismo. Entretanto, o DNA sofre milhares de danos em sua sequência nucleotídica diariamente, sejam esses espontâneos, como a depurinação das bases adenina (A) e guanina (G), ou mesmo a desaminação da citosina (C) para uracila (U) ou geradas pela constante exposição do material genético a variados agentes mutagênicos, sejam esses físicos, químicos ou biológicos. Felizmente, a maior parte desses danos é prontamente removida por processos biológicos que visam manter a integridade e estabilidade do DNA através de um conjunto de eventos conhecido como reparo de DNA. Parte desses danos, no entanto, pode não ser removida podendo resultar em mutações (MORAES; CABRAL NETO; MENCK, 2012).

Entre as vias conhecidas de reparo de DNA, a via de reparo por excisão de nucleotídeos (NER) é a mais versátil e flexível, lidando com uma variedade de lesões que geram distorções das cadeias do DNA interferindo no emparelhamento das bases e, portanto, na replicação e na transcrição dessa molécula, comprometendo a integridade celular e contribuindo para que ocorram mutações. Os tipos de lesões que utilizam o NER como via de reparo geralmente são ocasionadas por radiação ultravioleta (UV) ou por vários tipos de compostos químicos (COSTA *et al.*, 2003; DE LIMA-BESSA *et al.*, 2008; SUGASAWA, 2008).

A via de NER é composta de pelo menos 30 proteínas que atuam de acordo com a necessidade e o tipo de lesão a ser removida do DNA. Essa via se baseia na mudança conformacional da estrutura do DNA causada principalmente por alterações químicas em sua cadeia, para ser ativada (NOUSPIKEL, 2009). Ela é subdividida em duas outras vias: o reparo do genoma global (GG-NER), que identifica e repara lesões pelo genoma, e o reparo acoplado à transcrição (TC-NER), que remove as lesões de genes transcritos. Apesar de atuarem em locais distintos, a diferença entre ambas as vias NER se limitam na forma de identificação da lesão, que ocorrem através de eventos diferenciados, sendo o restante do processo de reparo semelhante (COSTA *et al.*, 2003; GREDILLA; GARM; STEVNSNER, 2012; LEIBELING; LASPE; EMMERT, 2006). Na via GG-NER o reconhecimento ocorre principalmente pela ação do

complexo das proteínas *XPC*, *HR23B* e *Centrina 2*. No entanto, pode ocorrer também através do complexo proteico *XPE* (formado pelas subunidades *DDB1* e *DDB2*), que reconhece lesões particulares causadas por luz UV (GREDILLA; GARM; STEVNSNER, 2012; LIU; LEE; ZHOU, 2011). Em relação ao TC-NER, o reconhecimento ocorre pela parada da ação da RNA polimerase através do bloqueio do molde do DNA devido à lesão. O TC-NER então é iniciado através do recrutamento da proteína CSB, seguida de CSA, o que permite a chegada da maquinaria NER ao local de lesão (COSTA *et al.*, 2003; GREDILLA; GARM; STEVNSNER, 2012). Após a lesão ter sido identificada, o GG-NER e o TC-NER convergem para uma via comum de clivagem de fita lesada, com o recrutamento de fatores como o TFIIH, XPA, RPA e XPG. O fator TFIIH é um complexo proteico multifuncional TFIIH que contém 10 subunidades. Dentre essas, XPB e XPD (com atividade de helicases) desenrolam a dupla hélice localmente expondo a lesão. A estabilidade da cadeia simples é mantida por um conjunto de proteínas que compreendem XPA, XPG e a proteína de replicação A (RPA). As proteínas com atividade de endonuclease, XPG e ERCC1/XPF, clivam, respectivamente, a 3' e 5' da lesão, resultando num fragmento de 24 a 32 pares de bases (bp) contendo a lesão, que é removido. A lacuna resultante da remoção desse fragmento, posteriormente é preenchida pela maquinaria de replicação, utilizando-se da outra fita do DNA como molde (GREDILLA; GARM; STEVNSNER, 2012; SUGASAWA, 2011).

Dentre os agentes que atuam no DNA e que geram lesões que são substratos para NER, destaca-se a importância da radiação UV. Esta é a faixa do espectro magnético entre a luz visível e os raios-X. Ela é subdividida em três faixas espectrais de acordo com o tamanho do comprimento de onda: em UVC (220–280 nm) e UVB (280–315 nm) e UVA, subclassificada em UVA1 (340–400 nm) e UVA2 (315–340 nm) (GREINERT *et al.*, 2012; IKEHATA; ONO, 2011; SCHUCH *et al.*, 2013). As ondas curtas (UVC) e mais energéticas são totalmente filtradas pela camada de ozônio da estratosfera, chegando à superfície terrestre apenas a luz UVB, que corresponde em cerca de 5% do total de UV, e a UVA que corresponde a aproximadamente 95% (GRUBER *et al.*, 2007). Essa faixa de comprimento maior que chega a Terra é capaz de atravessar as camadas da pele (quanto maior seu comprimento de onda, mais profundo na pele ela penetra), causando danos nos tecidos e à molécula de DNA (GREINERT *et al.*, 2012; MCMILLAN *et al.*, 2008; WELSH *et al.*, 2011). Os tipos de

lesão geradas pela UV também variam dependendo do comprimento de onda. O mais bem compreendido é o UVB que, através de uma excitação direta à molécula de DNA, faz com que a mesma sofra reações de dimerização entre bases de pirimidinas adjacentes produzindo duas formas de lesões principais: a formação dos fotoprodutos de dímeros de pirimidina ciclobutano (CPDs) e de pirimidina (6-4) pirimidona (6-4PPs). Essas reações fazem com que as fitas de DNA sofram uma mudança conformacional interferindo na integridade das mesmas, como citado anteriormente. Lesões não reparadas podem originar mutações, sendo que são características de UV a substituição de base citosina (C) para timina (T) em locais dipirimídicos ou de CC para TT, quando as pirimidinas estão em *tandem* na sequência. Em uma célula humana, os fotoprodutos 6-4PP são reparados rapidamente (cerca de 6 horas), enquanto os CPDs demoram pelo menos o dobro do tempo (DIGIOVANNA; KENNETH; H. KRAEMER, 2012; GREINERT *et al.*, 2012; IKEHATA; ONO, 2011). Os CPDs são formados quando a UV induz a formação de ligações covalentes adicionais entre os carbonos C-5 e C-6 de uma pirimidina com os mesmo carbonos de outra adjacente, enquanto os 6-4PPs formam-se quando ocorre uma ligação covalente entre o carbono C-4 de um pirimidina com o carbono C-6 de outra, também adjacente (NOUSPIKEL, 2009). Já as lesões induzidas pelo UVA são menos compreendidas quando comparadas aos casos de UVB. Sabe-se que a luz UVA causa o envelhecimento e o aparecimento de manchas amarronzadas na pele, além de estimular a produção de espécies reativas de oxigênio e nitrogênio, que atacam o DNA podendo produzir oxidação das bases, sendo a guanina a mais susceptível, gerando lesões como o 8-oxo-7,8-dihidro-2'-desoxiguanosina (8-oxodG), ou mesmo pode ocasionar quebras na molécula de DNA, simples ou dupla fita. Além disso, a luz UVA também está envolvida na formação de alguns fotoprodutos CPD (principalmente TT-CPDs) e quando absorvida pela base do DNA que sofreu previamente alteração para um 6-4PP, induz a formação de um isômero conhecido como Dewar (IKEHATA; ONO, 2011; MCMILLAN *et al.*, 2008; SCHUCH; MENCK, 2010). As mutações geradas pela atuação da UVA ainda estão em estudo, porém, algumas delas já estão bem caracterizadas como é o caso da transversão de G para T, induzida pela 8-oxodG, e a substituição de C para T relacionada ao CPD (IKEHATA; ONO, 2011, RUNGER *et al.*, 2012, SAGE *et al.*, 2012).

Síndromes de herança autossômica recessiva, caracterizadas pela sensibilidade aumentada aos raios UV, estão relacionadas à deficiência no processo de reparo por excisão de nucleotídeo, como a Xeroderma Pigmentosum (XP), síndrome de Cockayne (CS) e Tricotiodistrofia (TTD) (MENCK; MUNFORD, 2014).

O XP (OMIM #278700) é uma doença hereditária geneticamente heterogênea que afeta todos os grupos raciais, e é caracterizada por sintomas cutâneos na área de pele exposta à ação da luz solar devido ao defeito no reparo de lesões induzidas pelos raios UV. Os pacientes XP são extremamente sensíveis à luz solar podendo apresentar uma resposta exagerada e prolongada em relação às queimaduras solares. Além disso, podem apresentar mudanças de pigmentação na pele, elevada propensão ao desenvolvimento de câncer de pele (incluindo melanoma), problemas neurodegenerativos e perda de audição, sendo os dois últimos menos frequentes. (DIGIOVANNA; KENNETH; H. KRAEMER, 2012; NOUSPIKEL, 2009; TOTONCHY *et al.*, 2013). Os pacientes podem apresentar durante suas vidas vários tumores de pele, sendo a idade média para surgir os primeiros em torno de 9 anos de idade, que é muito baixa quando comparada com a idade média de 67 anos em pessoas que não possuem a doença. Células de pacientes XP são sensíveis a irradiação com luz UV e apresentam alto nível de mutagenicidade induzida por luz UV (MENCK; MUNFORD, 2014). O fenótipo XP resulta de mutações que acarretam perda de função do produto de 8 genes, cada qual caracterizando um grupo distinto de XP (*XPA-XPG* e XP variante, *XPV*) que diferem pelos sintomas apresentados e a capacidade de NER dependendo da mutação e da proteína afetada: XPA, XPB/ERCC3, XPC, XPD/ERCC2, XPE/DDB2, XPF/ERCC4, XPG/ERCC5 e POLH/XPV (BUDDEN; BOWDEN, 2013; DIGIOVANNA; KENNETH; H. KRAEMER, 2012). No caso do grupo XPV, os pacientes não possuem defeitos na via de NER, e sim perda de função da DNA polimerase η (pol η), codificada pelo gene *POLH/XPV*, que atua na replicação do DNA lesado, ou via de síntese translesão (TLS). Essa alteração ocorre devido à demora de se remover lesões, como CPDs, pelo GG-NER, fazendo com as DNA polimerases replicativas, altamente especializadas e fidedignas, sejam bloqueadas durante a replicação. Nesse momento, as polimerases de translesão atuam ignorando a lesão e utilizando a lesão como molde, em geral adicionando as bases AA em locais com os dímeros TT. Nos pacientes XPV, devido à ausência ou perda de função da pol η , os CPDs são utilizados como moldes por outros tipos de polimerases que

realizam a translesão, porém são menos precisas e propensas a mutarem o DNA (IKEHATA; ONO, 2011; INUI *et al.*, 2008; SUGASAWA, KAORU, 2008). Apesar de ser uma doença rara, em uma comunidade (Araras) situada na cidade de Faina no estado de Goiás foram diagnosticados mais de vinte pacientes com XP em uma população de cerca de mil habitantes, com idades que variam de 8 a 78 anos, muitos deles nascidos de casamentos consanguíneos. Estudos realizados pelo nosso grupo de pesquisa apontam que esses pacientes têm mutações no gene *POLH*, sendo então diagnosticados como XP-V.

A síndrome de Cockayne (OMIM #216400) é uma doença autossômica recessiva caracterizada por nanismo, retinopatia, microcefalia, calcificação ganglionar, surdez, defeitos neurais, retardo de crescimento e desenvolvimento após o nascimento, e não está relacionada com aumento de incidência de câncer. Esses pacientes são fotossensíveis, porém não exibem alterações pigmentares nem uma maior incidência de câncer de pele como no XP. A gravidade dos sintomas, classificado em grupos I, II e III, pode ser muito variável, dependendo do grupo de complementação e a natureza da mutação. Esta síndrome resulta de mutações que acarretam perda de função em cinco genes: genes *CSA* e *CSB*, que são especificamente necessárias para o TC-NER, e mutações nos genes *XPB*, *XPD* e *XPG*. Os pacientes que apresentam mutações nos genes XP geralmente sofrem do tipo grave da doença (tipo II), podendo também apresentar alguns sintomas semelhantes a pacientes XP, entretanto a média de vida desses pacientes é muito baixa (12 anos em média) (BERQUIST; WILSON, 2012; CLEAVER, 2005; NOUSPIKEL, 2009).

Tricotiodistrofia (OMIM #601675) é uma síndrome também de transmissão autossômica recessiva caracterizada pela deficiência de proteínas ricas em enxofre. Os pacientes com TTD exibem cabelos ralos e quebradiços e outras características que incluem neuroectodermias, retardo mental e de crescimento, tendência a infecções, anormalidades das unhas, diminuição da fertilidade, características de envelhecimento precoce e fotossensibilidade. As células de indivíduos com TTD também exibem prejuízo na via de NER, cuja gravidade varia de paciente para paciente. Esta síndrome pode resultar por mutações que afetam três genes, que codificam subunidades do complexo TFIIH: TTD-A, XPB e XPD. Dado que TFIIH opera

na transcrição, bem como em NER, isso levou à sugestão de que TTD, como para CS, pode resultar de um defeito sutil na transcrição, e não apenas deficiência de NER (NOUSPIKEL, 2009; STEFANINI *et al.*, 2010).

Nosso grupo de pesquisa tem potencial interesse em estudar e entender as doenças correlacionadas a falhas do sistema de reparo do DNA, sobretudo NER e TLS. Nesse sentido, nossa expectativa foi desenvolver processos que permitam localizar e mapear em nível molecular as mutações ocorridas nos genes envolvidos nesses sistemas, possibilitando o diagnóstico molecular e proporcionando maior compreensão dos eventos moleculares afetados. Para que isso fosse possível, amostras de DNA de pacientes com diagnóstico clínico dessas síndromes foram sequenciadas por uma plataforma de sequenciamento de nova geração (NGS) e suas variações de nucleotídeos foram posteriormente identificadas através de diferentes ferramentas de bioinformática.

As novas tecnologias são desenvolvidas visando à elucidação das bases genéticas relacionadas a cada doença e, com o surgimento das plataformas de sequenciamento de nova geração, a identificação de variações genéticas está progredindo rapidamente, o que facilita o diagnóstico da doença, contribuindo para o tratamento e aconselhamento genético de familiares. Dentre as plataformas atuais e presentes em nosso país, pode-se citar o ABI SOLiD da *Life Technologies* e as plataformas MiSeq e HiSeq da Illumina. A primeira plataforma utiliza a tecnologia de sequenciamento baseado na ligação de oligonucleotídeos marcados com fluoróforo através das enzimas ligases. As bibliotecas de DNA, geradas através de fragmentação mecânica, são amplificadas através de PCR em emulsão (*Polymerase Chain Reaction*) em que cada fragmento único da biblioteca gerada, após desnaturação, se liga aos adaptadores que se conectam a uma microesfera metálica e, então, emulsionadas em uma mistura de água e óleo com reagentes para amplificação da fita simples obtendo múltiplas cópias desse fragmento. Em seguida, ocorre o sequenciamento através da utilização de um *primer* universal e de ligações de sondas de oito bases que contém: local de ligação (a primeira e a segunda base, que consistem em 1 das 16 possibilidades de combinação entre 2 bases: AG, AC, etc), local de clivagem (a quinta base), e quatro fluoróforos diferentes (ligada à última base). O sinal fluorescente será registrado nas bases complementares e, em seguida, o

fluoróforo será clivado na sua extremidade 3'. Nos próximos ciclos, utilizam-se *primers* de tamanho $(n-1)$ até $(n-4)$ para que toda a sequência do fragmento alvo seja determinada (Figura 1) (MARDIS, 2008; METZKER, 2010; VOELKERDING; DAMES; DURTSCHI, 2009). O comprimento de cada leitura (*read*) alcançado nas sequências geradas é de 50 a 75 bases pelo ABI SOLiD 5500XL. Como cada combinação de duas bases é determinada com uma cor fluorescente diferente e a adição do *primer* $(n-x)$ faz com que as bases sequenciadas sofram dupla constatação, a taxa de erro da base chamada é reduzida, tornando o SOLiD uma plataforma de elevada precisão, auxiliando na investigação de possíveis SNPs (do inglês, *single nucleotide polymorphisms*) e mutações correlacionadas à doenças (VOELKERDING; DAMES; DURTSCHI, 2009). Já as plataformas da Illumina utilizam o método de sequenciamento por síntese (SBS) (http://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf). Após a formação da biblioteca de DNA, os fragmentos formados passam pelo processo de PCR em fase sólida, conhecida também como amplificação em “ponte” (*bridge PCR*), uma característica dessa tecnologia. Esses fragmentos são ligados em adaptadores em suas extremidades que, após desnaturados, ligam-se às sondas complementares contidas na superfície de clonagem (*flow cell*) pela extremidade 5', deixando a 3' livre para o processo de amplificação e consequente sequenciamento (ANSORGE, 2009; SHENDURE; JI, 2008). Cada fita simples imobilizada forma uma ponte ao ligar o adaptador da extremidade livre à uma sonda complementar contida na superfície sólida, que age como um *primer* para que a fita seja amplificada após a adição de reagentes de PCR. Após vários ciclos de amplificação, em torno de mil cópias da fita simples são obtidas formando aglomerados de sequências iguais muito próximos chamados de *cluster* ou *colonies* (Figura 2). A mistura de reagentes para que ocorra o sequenciamento contém nucleotídeos terminadores marcados com diferentes fluoróforos para cada base, além de *primers* e uma DNA polimerase são fornecidos para as reações de sequenciamento que ocorrem em cada *cluster*. Após a incorporação do nucleotídeo terminador na cadeia de DNA durante o sequenciamento, a luz emitida, bem como a posição do *cluster*, são capturadas através de uma câmera. Em seguida, o grupo terminador e o fluoróforo são removidos e lavados, e outro ciclo se inicia até a cobertura total dos fragmentos. A alta densidade dos *clusters* de sequenciamento possibilita que o sinal de fluorescência gerado com a incorporação

de cada um dos nucleotídeos terminadores tenha uma intensidade suficiente para garantir sua detecção exata, diminuindo o erro de falsos positivos por fragmento (ANSORGE, 2009; SHENDURE; JI, 2008; ZHOU *et al.*, 2010). O comprimento de cada *read* alcançado nas sequências geradas é de 75 a 300 bases pelo Illumina MiSeq (http://www.illumina.com/documents/products/datasheets/datasheet_miseq.pdf).

Figura 1: Representação esquemática do sequenciamento pela plataforma SOLiD. Visualiza-se a sequência alvo ligada ao adaptador e o *primer* universal de tamanho n hibridizado, bem como a primeira sonda marcada sendo ligada pela enzima ligase (A), com a consecutiva clivagem e captura da fluorescência (B). Esse processo se repete até que toda a sequência alvo seja coberta (C e D). Ao final dessa etapa, é utilizado nas 4 demais etapas *primers* universais de tamanho $(n-1)$ até $(n-4)$, em que as bases sofrem um segundo questionamento e, conseqüentemente, uma segunda contestação de sua chamada, proporcionando uma baixa taxa de erro da base sequenciada (adaptado de VOELKERDING; DAMES; DURTSCHI, 2009).

Figura 2: Representação esquemática do *bridge* PCR das plataformas da Illumina. Visualiza-se dois diferentes fragmentos de DNA se hibridizando nas sondas complementares contidas na superfície sólida de amplificação. Em seguida, nota-se a amplificação em ponte dessas fitas gerando os *clusters* para cada fragmento que serão posteriormente sequenciados pela plataforma utilizada (adaptado de METZKER, 2010).

As mutações pontuais (substituições) são caracterizadas pela troca de um nucleotídeo por outro em uma mesma posição na sequência de DNA e podem surgir pela incorporação de bases errôneas durante a replicação, ou mesmo por modificação química nas bases. São classificadas em sinônimas ou não-sinônimas. No caso das sinônimas, o aminoácido codificado pelo códon que contém a variante é o mesmo que aquele codificado pelo códon original, enquanto que na não sinônima, o códon modificado codifica um aminoácido diferente daquele codificado pelo códon original (TENG; MICHONOVA-ALEXOVA; ALEXOV, 2008). Essas mudanças que levam a alteração do aminoácido, não necessariamente alteram a estrutura e função de uma proteína, mas caso isso ocorra, podem resultar na formação de peptídeos não funcionais. Pode também ocorrer a formação de um códon de parada (*stop codon*) prematuro, gerando peptídeos truncados e, muitas vezes, não funcionais. As mutações que envolvem deleção ou inserção de um ou mais nucleotídeos na sequência de DNA são conhecidos pelo termo molecular *indel*. Os *indels* podem englobar desde 1 até milhares de nucleotídeos (como aqueles provocados por transposons), sendo os menores mais frequentes no genoma humano e podem gerar um grande impacto nas

funções de genes (MILLS *et al.*, 2006). Os efeitos que a presença de uma mutação acarreta são variáveis. Podem estar relacionados com (a) modificações no enovelamento proteico, sua estabilidade e flexibilidade, essenciais para a realização de suas funções, (b) substituição de resíduos em sítios catalíticos de proteínas, ou próximos desses, afetando sua função ou cinética, (c) modificação das funções das proteínas, (d) ou mesmo a alterações em sítios de ligações que poderiam afetar a ligação entre moléculas que interagem com essa proteína (TENG; MICHONOVA-ALEXOVA; ALEXOV, 2008). Quando uma variação da sequência de nucleotídeos impede ou reduz a capacidade funcional da proteína codificada pelo gene, ela é considerada deletéria. Variantes que surgem na população em decorrência dos processos de mutação podem ser chamadas de SNPs. Por definição, uma variante é considerada um SNP quando ocorre variação numa única posição na sequência de DNA em pelo menos 1% da população (<http://www.nature.com/scitable/definition/single-nucleotide-polymorphism-snp-295>).

O principal objetivo desse projeto é a identificação de variantes presentes na sequência de nucleotídeos do DNA de células de pacientes com doenças relacionadas ao reparo de DNA (principalmente XP, TTD e CS), em amostras sequenciadas por plataformas NGS.

Este projeto, em uma primeira etapa, visou executar diferentes protocolos (*pipelines*) de identificação e análise de substituições e pequenos *indels* nas amostras sequenciadas. Dado que estas amostras são provenientes de pacientes com características clínicas que indicavam deficiência no sistema de reparo de DNA, concentramos o estudo nas variantes contidas em genes sabidamente envolvidos na manifestação das doenças relacionadas ao processo de reparo de DNA por NER. Além disso, após determinar o melhor protocolo para buscar variantes em sequências de DNA, foi desenvolvido uma interface *web* eficiente e amigável para que qualquer pesquisador/usuário (formado em áreas não necessariamente ligada a informática), que tenha interesse em realizar identificação, anotação e pré-análise de variantes a partir de sequências de DNA geradas por diferentes plataformas NGS, possa utilizá-lo.

2. OBJETIVOS

2.1. OBJETIVOS GERAIS

Desenvolver e aplicar um *pipeline* eficiente na identificação de variantes em sequências de DNA geradas através de NGS e desenvolver uma interface web de domínio público que visa facilitar a análise de mutações potencialmente responsáveis por doenças genéticas, tendo como alvo principalmente doenças genéticas relacionadas a reparo de DNA.

2.2. OBJETIVOS ESPECÍFICOS

- Desenvolver um sistema eficiente de busca e identificação de variantes em sequências geradas por NGS, específica para genes, incluindo exons e introns, relacionados ao reparo de DNA.
- Identificar variantes comparando-as com referência do genoma humano (Hg19) e com as mutações patogênicas já descritas como associadas às doenças em estudo.
- Caracterizar *in silico* as variantes candidatas quanto ao seu potencial deletério, sua localização (região codificadora, região não traduzida, introns e sítios de *splicing*), seu tipo (sinônimo ou não sinônimo) e possível mudança das características químicas da proteína correspondente ao gene.
- Identificar variantes potencialmente responsáveis às manifestações clínicas das síndromes.
- Desenvolver uma interface *web* de domínio público, eficiente e de fácil acesso para identificar variantes (principalmente substituições simples e pequenos *indels*) em amostras sequenciadas por NGS.

3. METODOLOGIA

3.1. AMOSTRAS E OBTENÇÃO DE SUAS SEQUÊNCIAS NUCLEOTÍDICAS

Neste projeto foram analisadas, em primeira instância, 16 amostras, sendo 14 amostras de pacientes com quadro clínico XP, e 2 amostras de pais (não afetados)

de um dos pacientes. Os pacientes apresentavam, em sua maioria, sintomas como fotossensibilidade e manchas na pele em diferentes níveis, alguns já haviam desenvolvido tumores cutâneos.

Das amostras utilizadas, nove foram obtidas em parceria com o Hospital AC Camargo, somando-se com as 7 amostras de pacientes que já são estudadas pelo nosso grupo de pesquisa. Nas amostras do Hospital AC Camargo, o DNA foi extraído a partir do sangue. Estes pacientes não possuíam qualquer parentesco entre eles. Nas amostras do nosso grupo de pesquisa, o DNA foi obtido a partir de culturas de fibroblastos de pele normal (não tumoral) dos pacientes, com exceção do paciente XP02AM, no qual a amostra obtida era de tecido tumoral (carcinoma basocelular) (Tabela 1). Dessas amostras, cinco delas são oriundas de pacientes da região de Araras e a relação de parentesco entre eles pode ser observada na Figura 3. As outras 2 amostras (XP02AM e XP02RJ) pertencem a indivíduos que não tem parentesco com nenhum outro paciente. As amostras foram processadas no Centro de Facilidades de Apoio à Pesquisa - USP (CEFAP-USP), onde se encontra a estrutura para a realização do sequenciamento. Para esta metodologia foram utilizados 5 µg de DNA de cada amostra.

As amostras de DNA foram submetidas ao enriquecimento por captura de sequências específicas para os genes relacionados à NER, e já descritos como responsáveis por síndromes como XP, ou relacionadas (CS e TTD). Também foram incorporados ao painel genes relacionados à anemia de Fanconi (KENNEDY; D'ANDREA, 2005) e genes já conhecidos por estarem envolvidos em vários tipos de cânceres e relacionados intimamente a processos de reparo de DNA, como o *BRCA1* e *TP53* (BROUSTAS; LIEBERMAN, 2014; XIE *et al.*, 2014). A captura customizada foi realizada com kit da Agilent (*SureSelect Custom DNA/RNA*- www.agilent.com) para esse fim. No desenho das sequências a serem capturadas foram incluídos exons e íntrons, além de 5.000 bases a montante e a jusante de cada gene, resultando um total de 1,218,069 bp. O sequenciamento das amostras foi realizado pela plataforma ABI SOLiD 5500XL (*Life Technologies*) gerando *reads* de 75 pb.

Tabela 1: Relação das amostras sequenciadas segundo o tecido de extração do DNA.

Amostra	XP02AM	XP02RJ	XPSPAC02F0	GO01	GO03	GO04	GO05	GO06
DNA extraído	cultura celular tumoral	cultura celular	sangue	cultura celular	cultura celular	cultura celular	cultura celular	cultura celular
	XPSPAC01F0	XPSPAC04F0	XPSPAC06F0	XPSPAC08F0	XPSPAC11F0	XPSPAC12F1	XPSPAC13F0	XPSPAC31F0
	sangue	sangue	sangue	sangue	sangue	sangue	sangue	sangue

Os genes capturados pelo sequenciamento enriquecido são listados abaixo (Tabela 2):

Tabela 2: Genes abordados na análise. Genes de reparo por NER: *XPA*, *ERCC3/XPB*, *XPC*, *ERCC2/XPB*, *DDB1/XPE*, *DDB2*, *ERCC4/XPF*, *ERCC1*, *ERCC5/XPG*, *ERCC8/CSA*, *ERCC6/CSB*, *POLH/XPV*, *GTF2H5/TTD*; genes relacionados a anemia de Fanconi: *FANCA*, *FANCC*, *FANCG*, *FANCD1/BRCA2*, *FANCD*; genes relacionados em vários tipos de cânceres e relacionados a processos de reparo de DNA: *BRCA1*, *TP53*.

Símbolo	Crom	Fita	#Exon	Tamanho RNAm	Tamanho genômico	Tamanho proteico	RefSeq ID	Descrição
<i>XPA</i>	chr9	-	6	1.491	22.501	293	NM_000380	Proteína de reparo de DNA complementar às células XP-A
<i>ERCC3 (XPB)</i>	chr2	-	15	2.751	36.887	782	NM_000122	Subunidade helicase XPB do complexo TFIIH
<i>XPC</i>	chr3	-	16	3.729	33.525	940	NM_004628	Proteína de reparo de DNA complementar às células XP-C
<i>ERCC2 (XPD)</i>	chr19	-	23	2.568	19.197	760	NM_000400	Subunidade helicase XPD do complexo TFIIH
<i>DDB1 (XPE)</i>	chr11	-	27	4.372	33.766	1.140	NM_001923	Reconhecimento de lesão
<i>DDB2</i>	chr11	+	10	1.870	24.277	429	NM_000107	Reconhecimento de lesão
<i>ERCC4 (XPF)</i>	chr16	+	11	6.765	32.192	916	NM_005236	Endonuclease de reparo de DNA
<i>ERCC1</i>	chr19	-	10	3.400	16.587	297	NM_001983	Proteína de reparo por excisão
<i>ERCC5 (XPG)</i>	chr13	+	15	4.091	30.161	1.186	NM_000123	Proteína de reparo de DNA complementar às células XP-G
<i>ERCC8 (CSA)</i>	chr5	-	12	2.044	71.247	396	NM_000082	Proteína de reparo por excisão
<i>ERCC6 (CSB)</i>	chr10	-	21	8.993	84.644	1.493	NM_000124	Proteína de reparo por excisão
<i>POLH</i>	chr6	+	11	8.412	44.383	713	NM_006502	DNA polimerase eta
<i>GTF2H5 (TTD-A)</i>	chr6	+	3	7.503	30.998	71	NM_207118	Fator de transcrição IIH subunidade 5
<i>FANCA</i>	chr16	-	43	5.460	79.107	1.455	NM_000135	Proteína grupo A da anemia Fanconi
<i>FANCC</i>	chr9	-	15	2.721	207.484	492	NM_000136	Proteína grupo C da anemia Fanconi
<i>FANCG</i>	chr9	-	14	2.649	6.179	622	NM_004629	Proteína grupo G da anemia Fanconi
<i>BRCA2 (FANCD1)</i>	chr13	+	27	11.386	84.193	3.418	NM_000059	Proteína de susceptibilidade do tipo 2 de câncer de mama
<i>FANCD2</i>	chr3	+	43	5.204	73.232	1.471	NM_033084	Proteína grupo D2 da anemia Fanconi
<i>BRCA1</i>	chr17	-	24	7.287	81.189	1.884	NM_007300	Proteína de susceptibilidade do tipo 1 de câncer de mama
<i>TP53</i>	chr17	-	11	2.591	19.149	354	NM_001276760	Antígeno celular tumoral

Figura 3: Genealogia dos cinco pacientes de Araras que tiveram suas seqüências nucleotídicas analisadas (imagem cedida pela doutoranda Ligia Pereira Castro do Laboratório de Reparo de DNA, USP-SP).

3.2. ANÁLISE DAS SEQUÊNCIAS E PARÂMETRO DE QUALIDADE

Os *reads* obtidos do sequenciamento eram do tipo *single-end* e foram inicialmente alinhados contra o genoma de referência (hg19/GRCh37.p13) pelo programa LifeScope utilizando os parâmetros padrões de alinhamento sugerido pela empresa através do manual do usuário (<http://www.appliedbiosystems.com/lifescopes>). No alinhamento, os *reads* foram processados pelo SAET (SOLiD Accuracy Enhancement Tool), programa integrado ao Lifescope, que visa melhorar a qualidade do alinhamento aumentando a precisão das bases sequenciadas e eliminando *reads* de baixíssima qualidade. Esses *reads* são alinhados gerando um arquivo de saída final no formato BAM (um formato binário ordenado, indexado e extremamente compacto), e estatísticas sobre a qualidade do alinhamento, que serviram de base para testes de parâmetros a serem utilizados para cada situação. Os dados de formato BAM foram visualizados pelo programa *Integrative Genomics Viewer* (IGV) (THORVALDSDÓTTIR; ROBINSON; MESIROV, 2013).

3.3. BUSCA DE SUBSTITUIÇÕES E *INDELS*

Em busca das variantes presentes nas amostras, a análise foi dividida em duas etapas: a primeira etapa visou analisar as saídas das duas principais ferramentas de bioinformática que buscam variantes, o GATK (*Genome Analysis Toolkit*, versão 2.7.2) (MCKENNA *et al.*, 2010) e o SAMTOOLS (versão 0.1.19) (LI *et al.*, 2009), para se determinar qual delas seria utilizada na interface de análise final; a segunda etapa, visou estabelecer os melhores parâmetros dependendo da qualidade inicial alinhamento dos *reads*. O genoma de referência humano utilizado foi o hg19/GRCh37.p13, o mesmo utilizado no alinhamento dos *reads* das amostras.

O SAMTOOLS é um pacote de programas capaz, entre outras funções, de manipular arquivos de formato SAM (um formato de texto delimitado por tabulação consistindo de cabeçalho, alinhamento e informações de qualidade) (<http://samtools.github.io/hts-specs/SAMv1.pdf>) ou BAM. Ele oferece diversas ferramentas para conversão de formato, controle de qualidade básica, indexação e ordenação de arquivo, remoção de *reads* duplicados e, principalmente, busca por variantes, além de outras funções (LI *et al.*, 2009). O GATK é um programa em JAVA que processa os arquivos alinhados para se aumentar a qualidade destes e, em

seguida, buscar por variantes com uma maior precisão. Esse software foi desenvolvido e disponibilizado pelo *Broad Institute*, um grande centro independente de pesquisa genômica e biomédica formado pela parceria entre as Universidade de Harvard e do Instituto de Tecnologia de Massachussetts (<https://www.broadinstitute.org/>). Para se aumentar a qualidade da amostra alinhada, o GATK é capaz de realizar o realinhamento de regiões ricas em *indels* (baseado em bancos de dados de referência) e recalibrar a qualidade das bases dos *reads* sequenciados devido a erros dos ciclos da máquina de sequenciamento (DEPRISTO *et al.*, 2011).

3.3.1. Primeira etapa:

Devido ao evento de PCR que ocorre previamente ao sequenciamento, em que os fragmentos de DNA são replicados várias vezes, é quase que inevitável a ocorrência de *reads* duplicados. Essas duplicações ocorrem quando duas cópias do mesmo fragmento de DNA resultam, através do sequenciamento, em 2 ou mais *reads* idênticos (ex: mesmo tamanho, mesma posição inicial e final quando alinhado no genoma). A leitura de bases cobertas por esses *reads* duplicados trazem problemas na especificidade na busca por variantes, pois podem capturar algum erro de sequenciamento, que foi lido mais de uma vez por estar duplicado, aumentando o número de variantes que são falsos positivos. Para evitar esse evento, foram removidos os *reads* duplicados presentes no arquivo alinhado em todos os três *pipelines* utilizados, com a ferramenta *rmDup* do pacote SAMTOOLS (Figura 4). Com a mesma finalidade, também foi utilizado o programa PICARD *MarkDuplicates* (versão 1.97, <http://picard.sourceforge.net>), para uma análise comparativa entre ambos, sendo que o segundo mostrou-se um pouco mais eficaz que o *samtools rmDup* na remoção de *reads* duplicados. Na utilização do PICARD foi necessário inserir a expressão regular específica para o nome dos *reads* provenientes de sequenciadores SOLiD. Com a remoção de *reads* duplicados, foi ainda necessário criar uma nova indexação para o novo arquivo BAM gerado e, para esse fim, utilizou-se a ferramenta PICARD *BuildBamIndex* (1).

(1)

- `samtools rmdup sample.bam`
- `java -jar picard/MarkDuplicates.jar I=sample_rmdup.bam
O=sample_realrmdup.bam VALIDATION_STRINGENCY=LENIENT
REMOVE_DUPLICATES=true ASSUME_SORTED=true READ_NAME_REGEX="([0-9]+)_([0-9]+)_([0-9]+)" 2>picard_rel.txt`
- `java -jar picard/BuildBamIndex.jar I=sample.bam O=sample.bam.bai`

Figura 4: Demonstração, através da amostra XPSPAC02F0, de regiões dos genes *DDB2* (1) e *XPA* (2) de um alinhamento antes de se remover os *reads* duplicados (gráfico superior) e após a remoção (gráfico inferior). Nota-se uma maior concentração de *reads* em ambas regiões antes do processo de remoção dos duplicados (visualizado pelo IGV).

Utilizamos três *pipelines* distintos e amplamente divulgado pela comunidade científica para buscar as variantes: SAMTOOLS, GATK e SAMTOOLS + GATK.

3.3.1.1. SAMTOOLS

Nesse procedimento foi utilizado o pacote de ferramentas SAMTOOLS na busca ou chamada de variantes.

Com o novo arquivo BAM gerado, sem os *reads* duplicados, foi utilizado a ferramenta *samtools view*. Esta tem como finalidade (em outras) capturar os *reads* que atingem um determinado valor de qualidade de mapeamento frente ao genoma humano de referência para posteriormente realizar a busca de variantes, e o parâmetro utilizado nesse filtro foi de 22 em escala *Phred*. Em seguida utilizamos outras duas ferramentas, o *samtools mpileup* e o *bcftools*. O primeiro calcula as

proporções genotípicas das bases e um dos parâmetros que utilizamos foi a de sobreposição de *reads* com valor de 7%, ou seja, para que os *reads* sejam contabilizados na cobertura de uma base referência, eles precisam que pelo menos 7% das bases entre os *reads* estejam se sobrepondo. Ao fim desse passo é gerado um arquivo no formato BCF (formatação binária das variantes). Esse arquivo então é lido pela segunda ferramenta que chama as potenciais variantes, agindo como um filtro da saída do *mpileup*. O arquivo no formato BCF então foi convertido para um arquivo no formato VCF (*variant call format*), que contém um cabeçalho seguido pelas variantes filtradas com suas devidas informações para a análise. Logo em seguida, foi utilizado um *script* em *Perl* interno do samtools, o *vcfutils*, que realiza uma filtragem nesse arquivo VCF com os parâmetros do interesse do pesquisador. Nessa filtragem foram chamados apenas variantes com um mínimo de cobertura de 5 *reads*, com pelo menos 1 alteração, sendo que, caso houvesse mais de 500 *reads* para aquela variante, a abordagem seria em *reads* aleatórios, ou seja, seriam escolhidos 500 *reads* aleatórios entre todos os possíveis para se realizar a contagem. Além disso, reforçamos que o mínimo de qualidade de mapeamento é de 22 (2) (<http://samtools.sourceforge.net/samtools.shtml>).

(2)

```
• samtools view -b -q 22 sample.bam | samtools mpileup -ug -F 0.07 -f
reference.fa - | bcftools view -bcvg - > sample.bcf

• bcftools view sample.bcf | vcfutils.pl varFilter -a 1 -d 5 -Q 22 -D 500 >
saída.vcf
```

3.3.1.2. GATK

No procedimento utilizando apenas o GATK, foram realizados os procedimentos descritos no livro escrito pelos próprios desenvolvedores, o “*The GATK Guide Book*” (http://www.broadinstitute.org/gatk/pdftdocs/GATK_GuideBook_2.7-2.pdf), bem como a utilização de informações presentes na comunidade científica do GATK, em que os desenvolvedores criam tutoriais, comentam *pipelines* e respondem à perguntas dos usuários (<http://gatkforums.broadinstitute.org/>).

Como os programas que realizam alinhamento são otimizados para alinhar milhares de *reads* rapidamente, pode acontecer desses *reads* terem sido alocados em uma região errônea, principalmente devido às deleções e inserções. Devido a este fato, realizamos o realinhamento dos *reads* baseado nas inserções e deleções contidas no banco de dado de variantes do *National Center for Biotechnology Information* (NCBI): o dbSNP (versão 138). Esse banco de dados contém, além de substituições, pequenas inserções e deleções, pequenas repetições, marcadores de microssatélite, entre outros (SHERRY; WARD; SIROTKIN, 1999). Primeiramente criamos uma lista das regiões que devem passar pelo realinhamento dos *reads* pela opção *RealignerTargetCreator* do GATK, que determina pequenos intervalos suspeitos que possam necessitar de realinhamento. Logo em seguida, realinhamos os *reads* contidos no arquivo BAM pela opção *IndelRealigner*. Após esse processo, podemos enfim remover os *reads* duplicados. Utilizamos então o programa PICARD *MarkDuplicates*, seguido pelo PICARD *BuildBamIndex* (3).

(3)

```
• java -jar GenomeAnalysisTK.jar -T RealignerTargetCreator -R reference.fa
-I sample.bam -o sample.bam.intervals -known dnSNP138.vcf

• java -jar GenomeAnalysisTK.jar -T IndelRealigner -R reference.fa -I
sample.bam -targetIntervals sample.bam.intervals -known dnSNP138.vcf
sample_reali.bam

• java -jar picard/MarkDuplicates.jar I=sample_rmdup.bam
O=sample_reali_rmdup.bam VALIDATION_STRINGENCY=LENIENT
REMOVE_DUPLICATES=true ASSUME_SORTED=true READ_NAME_REGEX="([0-9]+)_([0-
9]+)_([0-9]+)" 2>picard_rel.txt

• java -jar picard/BuildBamIndex.jar I=sample.bam O=sample.bam.bai
```

Recalibrou-se então os valores de qualidade dos *reads* restantes pela opção *BaseRecalibrator* do GATK. Isso é realizado pois as qualidades de base atribuídas pela máquina podem não refletir verdadeiramente a base, chamando erros e podendo levar a retenção de falsos positivos (Figura 5). Para realizar tal tarefa, ele considera a base dentro de seu contexto geral, ou seja, a qualidade lida, a posição que ela se encontra dentro do *read* e as bases adjacentes a ela (<http://gatkforums.broadinstitute.org/discussion/44/base-quality-score-recalibration->

[bqsr](#)). Nesse processo foi necessário utilizar parâmetros definindo a origem (SOLID) de sequenciamento daqueles *reads*, pois ele trabalha com a leitura de cores. Ao final, obtém-se o levantamento de todas as informações para a recalibração, que é utilizado pela ferramenta *PrintReads* que recaptura os *reads* atribuindo os novos valores (4).

(4)

```
• java -jar GenomeAnalysisTK.jar -T BaseRecalibrator -R reference.fa -I
sample_real1_rmdup.bam -knownSites dbsnp138.vcf -o sample.recal_data.grp --
covariate QualityScoreCovariate --covariate ReadGroupCovariate --covariate
ContextCovariate --covariate CycleCovariate --solid_nocall_strategy
PURGE_READ --solid_recal_mode SET_Q_ZERO_BASE_N

• java -jar GenomeAnalysisTK.jar -T PrintReads -R reference.fa -I
sample_real1_rmdup.bam -BQSR sample.recal_data.grp -o
sample_real1_rmdup_recali.bam
```

Com a recalibração completada, podemos chamar pelas variantes. O GATK possui dois programas distintos para realizar tal função: *Unified Genotyper (UC)* e *Haplotype Caller (HC)*. Segundo os desenvolvedores, o primeiro chama por substituições e *indels* por dois modelos estatísticos distintos, tornando assim mais rápido, porém é menos específico quando se trata em se chamar *indels*. Já o *Haplotype Caller*, chama as variantes dentro de um mesmo modelo estatístico e parte do processo constitui no realinhamento *de novo* dos *reads* em torno de cada potencial variante, gerando dados de alta acurácia. As amostras já processadas foram submetidas a busca de variantes por ambos os programas, utilizando os mesmos parâmetros (5).

(5)

```
• java -jar GenomeAnalysisTK.jar -R reference.fa -T UnifiedGenotyper -I
sample_real1_rmdup_recali.bam -o sample.vcf -stand_call_conf 22 -
stand_emit_conf 10 -glm BOTH -D dnSNP138.vcf

• java -jar GenomeAnalysisTK.jar -R reference.fa -T HaplotypeCaller -I
sample_real1_rmdup_recali.bam -o sample.vcf -stand_call_conf 22 -
stand_emit_conf 10 -D dbSNP.vcf
```

Figura 5: Demonstração, através de parte de sequenciamento da amostra XPSPAC02F0, das qualidades de *reads* antes (A) e após (B) a recalibração de bases. Nota-se na imagem que o mesmo *read* apresentado (Reads_name demarcado em vermelho) possui uma de suas bases T com qualidade phred 30 (A), o que seria muito elevado, entretanto, após a recalibração essa mesma base foi reduzida pra uma qualidade phred de 11 (B).

3.3.1.3. GATK+SAMTOOLS

Neste procedimento foi realizado o processamento das amostras pelas ferramentas do GATK, e a busca pelas variantes foi executada pelo pacote SAMTOOLS.

Após a realização desses três procedimentos distintos, delimitaram-se as regiões de interesse a serem analisadas por uma filtragem que visava englobar toda a extensão dos genes em questão bem como 1.500 bases a montante e a jusante a esses. A razão pela qual não foram utilizadas as 5.000 bases anteriores e posteriores aos genes que foram sequenciados é que visávamos reduzir a quantidade de variantes chamadas em regiões intergênicas e focar apenas na porção estimada da presença da região promotora (0-1.000 aproximadamente *upstream* na maioria dos genes). Essa filtragem realizada foi necessária pois, apesar de ter sido um

resultados obtidos. Neste passo da análise, utilizamos uma versão mais atualizada que foi lançada do pacote GATK (versão 3.1.1), além mais nova versão dbSNP (versão 142). Os parâmetros alterados na busca de variantes, através das ferramentas *UnifiedGenotyper* e *HaplotypeCaller*, foram o *emit_call_conf* (se) e o *stand_call_conf* (sc). O primeiro parâmetro define o limiar mínimo de confiança em escala *Phred* no qual o programa deve emitir possíveis variantes. Já o segundo parâmetro é o limiar de confiança mínimo no qual o programa deve chamar uma possível variável. Se a variável chamada possui uma pontuação inferior ao limiar imposto, o programa emitirá essa variável como “filtrada” e vai mostrar que é *LowQual*, separando as variáveis de alta confiança das baixas. Após a obtenção do arquivo VCF final contendo as variantes, analisou-se os resultados obtidos.

3.4. DESCOBRINDO O POTENCIAL PATOGÊNICO DAS VARIANTES

Foram utilizados os programas de predição PROVEAN, *Protein Variation Effect Analyzer* (CHOI *et al.*, 2012), e SIFT, *Sorting Intolerant From Tolerant* (KUMAR; HENIKOFF; NG, 2009), nas variantes não sinônimas filtradas como candidatas a mutações patogênicas. Esses analisam o potencial deletério da substituição em questão, ou seja, o potencial da mudança de aminoácido afetar a função do produto proteico.

SIFT prevê se uma substituição de aminoácido afeta a função da proteína com base no grau de conservação de resíduo de aminoácido ao se alinhar com sequências proximamente relacionadas. As pontuações de SIFT variam de 0 a 1, e pontuações ≤ 0.05 pontuações são previstas como substituições “danosas” pelo algoritmo, enquanto que contagens > 0.05 são consideradas “toleráveis” (KUMAR; HENIKOFF; NG, 2009). PROVEAN é uma ferramenta que prevê o impacto funcional para todas as classes de variações na sequência de proteína, não apenas substituições, mas também inserções, deleções e substituições múltiplas. Este algoritmo é baseado na pontuação de alinhamento que mede a mudança na similaridade de uma sequência proteica com sequências homólogas da proteína antes e após a introdução de uma variação de aminoácidos. Se a pontuação for $\leq -2,5$, a proteína variante é prevista com um efeito “deletério”, enquanto que, se a pontuação for $> -2,5$, a variante é “neutra” (CHOI *et al.*, 2012). Ambos os *softwares* estão disponíveis na *homepage* do

Instituto J. Craig Venter: a ferramenta SIFT está em <http://sift.jcvi.org>, e o PROVEAN em <http://provean.jcvi.org>.

3.5. DESENVOLVIMENTO DA INTERFACE *WEB* DE DOMÍNIO PÚBLICO

O *script* do *pipeline* final que é usado na interface *web* foi desenvolvido em linguagem Perl (<http://www.perl.org>) para trabalhar especificamente com o genoma humano. Neste *script* está concentrada toda a parte de processamento da amostra (extensões BAM ou VCF), desde a busca por variantes através de parâmetros chave providos pelo usuário (entrada: bam), até a obtenção de cada variante anotada baseada nos bancos de dados de grande referência científica como Refseq, 1000genomes, COSMIC, clinvar, ESP6500 e Exac (entrada: *bam* e *vcf*). Além disso, ele é integrado com um sistema de gerenciamento de banco de dados SQL, o MySQL (<http://www.mysql.com>), o que permite que o arquivo final resultante das análises executadas por essa interface aceite customizações de filtros, possibilitando que o usuário tenha um maior aproveitamento de seus dados.

A integração entre o *script* em Perl e o usuário, via *web*, foi sendo desenvolvida em PHP (<http://php.net>), gerando páginas em *html* que trabalham com HTML5, Javascript/JQuery e CSS3, o que resulta na interface *web*. Esta integração entre o *script* e a página *web* está sendo feita pelo em colaboração do programador independente Fábio T. Higa.

Através da interface é possível entrar com os arquivos a serem analisados, informações sobre a amostra (por exemplo: tipo de *reads*, plataforma de sequenciamento) e mudar alguns parâmetros colocados como “*default*” no projeto (ex: região a ser analisada, qualidade mínima de busca por variantes, remoção de *reads* duplicados). Os arquivos de entrada aceitos são do formato BAM ou VCF de até 1GB (gigabase), podendo estar compactado no formato ZIP. Caso o usuário opte em carregar o arquivo BAM, todo o processo de busca e anotação será executado pelo *script* baseado nos parâmetros padrões ou modificados pelo usuário. Caso a entrada seja do formato VCF, esse será submetido apenas à anotação e filtragem das variantes.

Essa interface web foi nomeada de *Human Variants Finder Interface*, de uso livre, e foi registrado sob o domínio <http://www.varfinderhq.com.br>, sendo necessário apenas realizar um cadastro para ter acesso total ao serviço prestado.

3.5.1. O script da plataforma

O *script* da plataforma de busca é iniciado após a conclusão de envio dos arquivos pelo usuário e irá verificar o tipo de solicitação escolhida por este (ex: BAM or VCF), escolhendo a subrotina que seja adequada. Após essa escolha, acessando uma página criada em PHP (*“exec.php”*) que possui as informações de acesso ao banco de dados, o *script* busca o caminho de acesso dos programas que são utilizados para a busca das variantes (GATK, PICARD, ANNOVAR, SAMTOOLS), além do caminho e identificação única dos arquivos enviados e os parâmetros escolhidos pelo usuário, caso tenha inserido. Essa página, além de ser responsável por essas informações básicas para que o *script* seja executado corretamente, é responsável por eliminar todos os dados enviados e gerados pelo *script* após 1 semana finalizada a análise da amostra. Além disso, essa página é responsável por informar via *e-mail* ao usuário quando o processo de análise está concluído.

O *script Perl* contém todo o *pipeline* de execução, processamento e anotação, direciona todos os dados gerados para posterior obtenção dos relatórios de processamento do arquivo inicial, remove arquivos temporários ao longo da execução, atua como um sinalizador de estado do processo (ex: início, término, erros de execução), salva pequenos relatórios no banco de dados e controla as funções contidas na página *“exec.php”* (ex: *e-mail*). O *link* de acesso aos resultados finais é enviado por *e-mail* ao usuário, porém o usuário também pode acessar pela sua página de perfil cadastrado na interface *web*. No perfil do usuário existe a opção de visualizar o estado completo do processamento, relatórios gerados, tempo restante antes da remoção dos arquivos gerados do sistema, customização de filtragem dos resultados e opções para baixar os dados finais (inclusive pós customização). Os resultados são mostrados em formato tabelar numa página HTML com informações claras para que os usuários, mesmo não habituados com arquivos de saída de texto, possam realizar análises posteriores.

A *Human Variants Finder Interface* foi desenvolvida para se ter sigilo absoluto e segurança em relação ao acesso e disponibilização dos dados. Todos os usuários serão previamente cadastrados e um termo de utilização e política de acesso será disponibilizado a cada um.

3.6. INTERFACE WEB À PROVA: AMOSTRAS PROVENIENTES DA PLATAFORMA MISEQ (ILLUMINA)

Com a finalidade de se constatar a efetividade de nossa interface *web* frente a diferentes tipos de sequenciamento de nova geração, foram analisadas outras 15 amostras. Dessas 15, 8 são de pacientes com quadro clínico XP e 2 são amostras de pais (não afetados). As demais já haviam sido analisadas através do sequenciamento pela plataforma SOLiD (XP02RJ, XPSPAC06F0, XPSPAC11F0, XPSPAC31F0 e XP03AM que, diferente da amostra XP02AM proveniente de tecido tumoral, é uma amostra de tecido normal do mesmo paciente – Tabela 1, página 13). Todas as amostras foram obtidas da cultura de fibroblasto de pele normal e os pacientes não possuem qualquer parentesco entre eles. Os dois parentes dos pacientes correspondem às amostras BA02, que é a mãe da paciente de amostra XP01BA, e PE02, que é mãe da paciente de amostra XP01PE. O histórico familiar de alguns desses pacientes pode ser visualizado na Figura 7 para auxiliar na interpretação dos resultados.

Assim como as amostras anteriores, essas também foram processadas no CEFAP-USP. A captura customizada foi realizada com kit da Agilent, porém selecionando apenas os exons dos genes abordados nesse projeto (Tabela 2, página 13), resultando um total de 293,8 Kbp. O sequenciamento das amostras foi realizado pela plataforma MiSeq da Illumina gerando *pair-reads* de 2x 150 bp.

Figura 7: Heredograma de alguns pacientes que tiveram suas sequências nucleotídicas analisadas pela plataforma da Illumina (imagem cedida pela doutoranda Ligia Pereira Castro do Laboratório de Reparo de DNA, USP-SP).

3.6.1. Processamento das amostras Illumina

Os *reads* obtidos do sequenciamento foram alinhados contra o genoma de referência (hg19/GRCh37.p13) pelo programa BWA (LI; DURBIN, 2009) utilizando os parâmetros padrões necessários para que o resultado final de alinhamento, o arquivo BAM, fosse compatível com os softwares de análise utilizados no nosso pipeline (Gatk e PICARD, já abordados anteriormente) (6). Primeiramente indexamos o genoma de referência (o *database*) utilizando o parâmetro ‘-a bwtsv’, o indicado para a indexação de genomas grandes, como o genoma humano. Na fase do alinhamento, utilizamos a ferramenta *bwa mem* que, segundo o manual (<http://bio-bwa.sourceforge.net/bwa.shtml>), é usado preferencialmente em amostras com os *reads* maiores do que 70 bp. O parâmetro ‘-R’ utilizado foi para que o arquivo BAM contivesse informações essenciais no cabeçalho (*header*) para que o programa GATK funcionasse, baseado na documentação disponível na comunidade (<https://www.broadinstitute.org/gatk/guide/article?id=1317>). Já o parâmetro ‘-M’ é

necessário para que o PICARD funcione corretamente, pois marca *reads* de um alinhamento quimérico como 'alinhamento secundário' em vez de 'alinhamento suplementar'. Para a verificação da qualidade de alinhamento das amostras, foram utilizadas ferramentas dos programas GATK e PICARD, onde várias informações e estatísticas foram geradas (7).

(6)

```
• bwa index -a bwtsv reference.fa

• bwa mem -M -R '@RG\tID:foo\tLB:Agilent\tSM:foo\tPL:ILLUMINA' reference.fa
sample_reads1.fastq sample_reads2.fastq | samtools view -bS - > sample.bam
```

(7)

```
• picard.jar CollectAlignmentSummaryMetrics R=reference.fa I=sample.bam
O=sample.txt AS=true

• java -jar GenomeAnalysisTK.jar -T DepthOfCoverage -R reference.fa -I
sample_sort.bam -o sample_gatk.txt
```

Após a obtenção do arquivo BAM, as amostras foram utilizadas como entrada na *Human Variants Finder Interface* alterando os parâmetros para que o *pipeline* trabalhe com a plataforma Illumina com *reads pair-end*, além da inclusão de um arquivo contexto as coordenadas dos genes que estamos abordando neste projeto. Os parâmetros de chamada utilizados foram de SE=15 e SC=30, com remoção de *reads* duplicados.

4. RESULTADOS E DISCUSSÃO

4.1. A QUALIDADE DO SEQUENCIAMENTO

Ao analisar os relatórios de qualidade do sequenciamento de nossas amostras (Tabela 3), podemos afirmar que, em sua maioria, as amostras possuem uma boa quantidade de *reads* cobrindo as regiões delimitadas (pelo menos 78% das bases de

interesses possuem uma cobertura igual ou superior a 10 *reads*) dos genes selecionados no painel de sequenciamento. Dentre as amostras, apenas três delas estavam insuficientes para a realização de uma análise mais acurada de variantes com menos de 60% das bases com cobertura igual ou superior a 10 *reads* (GO03, XPSPAC06F0 e XPSPAC11F0) e duas delas estariam um abaixo da média, em torno de 60% a 80% das bases com cobertura $\times 10$ (XP02RJ, GO04). Assim esperava-se um número reduzido de variantes nessas amostras.

A análise dos relatórios de sequenciamento e alinhamento dos dados não faz parte da interface desenvolvida neste projeto, sendo necessário que o usuário a realize e avalie a qualidade de sua própria amostra.

4.2. EM BUSCA DE VARIANTES

A partir dos resultados obtidos na comparação entre os três *pipelines* de busca de variantes (Tabela 4), foi possível determinar qual deles seria utilizado na interface. Analisando cada procedimento isoladamente, notou-se uma mudança drástica nas quantidades de variantes após passar pelo filtro baseado nas regiões a serem estudadas. Isso é explicado pelo fato de que muitos *reads* (mais de 65% deles, em média) estarem situados fora da região alvo como demonstrado previamente na Tabela 3. Para tornar a análise mais específica foram removidas essas variantes fora das regiões de interesse do nosso estudo. Além disso, os *reads* abaixo da qualidade de chamada estabelecida, possuindo a assinatura “*LowQual*” no arquivo VCF, foram removidos. Apesar da qualidade *Phred* de filtragem utilizada para este procedimento (QUAL = 22) ter sido abaixo do recomendado pelos desenvolvedores dos programas de busca (QUAL = 30) e, conseqüentemente, mais variantes seriam chamadas, salientamos que essa medida foi tomada visando contemplar todas as amostras, independentes do seu grau de cobertura.

Ao se limitar a região de interesse e remover as variantes sinalizadas como de baixa qualidade, restaram as variantes de maior confiabilidade para aqueles parâmetros utilizados. Pode-se notar que os valores quantitativos finais das variantes identificadas alteraram significativamente entre os procedimentos, sendo que o procedimento que utiliza apenas o pacote SAMTOOLS foi o qual identificou mais

variantes, enquanto o do GATK e GATK+SAMTOOLS apresentaram resultados próximos.

Para descobrir qual dos procedimentos identifica as variantes com uma maior confiabilidade, elas foram comparadas às contidas no dbSNP138 (Tabela 5). Isso foi utilizado como critério uma vez que muitas das variantes já foram descritas na literatura anteriormente e estão depositadas no NCBI (<http://www.ncbi.nlm.nih.gov/>), o que faz com que a probabilidade de se encontrar variantes novas em grande quantidade, ou seja, nunca descritas, seja baixa.

Nota-se que o procedimento utilizando o programa SAMTOOLS foi o do qual resultaram mais variantes, sendo que em torno de 25% destas não foram encontradas no dbSNP138. Isso provavelmente é devido à ausência da etapa de controle de qualidade do arquivo alinhado, como ocorre com o GATK ao se realizar o realinhamento e a recalibração das bases. Já o procedimento utilizando o GATK foi o segundo que retornou mais variantes, entretanto foi o qual chamou menos variantes novas entre os três. Apesar do procedimento GATK+SAMTOOLS ter gerado dados semelhantes ao do GATK sozinho, essa diferença é dada pelo fato do buscador de variantes ter sido o SAMTOOLS. Já foi demonstrado, em trabalhos que comparam os programas que buscam por variantes, que o GATK possui uma maior sensibilidade e especificidade dentre os programas abordados (LIU, X. *et al.*, 2013; YU; SUN, 2013), confirmado por nossos dados. Logo, decidiu-se que a busca por variantes em nossa interface *web* será realizada com o software GATK.

Deve-se salientar que a remoção das variantes com a assinatura “*LowQual*” acarreta perda de informações importantes na análise, principalmente nas amostras com baixa cobertura de suas bases, como podemos observar em algumas de nossas amostras (XPSPAC06F0, XPSPAC11F0, GO03, por exemplo). A análise final foi realizada considerando todas as variantes chamadas, mesmo com baixa qualidade, entretanto, na interface *web* existe a possibilidade de o usuário remover essas de acordo com o critério por ele utilizado.

Tabela 3: Resumo do relatório de enriquecimento da amostra pelo método de SureSelect. Conclui-se que três amostras estão pobremente sequenciadas (GO03, XPSPAC06F0, XPSPAC11F0), duas delas estão boas (XP02RJ, GO04), e o restante possuem uma quantidade de *reads* cobrindo as bases (80% das bases com $\geq 10x$ cobertura) acima da média.

Amostra	Reads no alvo	Reads fora do alvo	% de bases alvos não cobertas	Cobertura de bases $\geq 1X$	Cobertura de bases $\geq 5X$	Cobertura de bases $\geq 10X$	Cobertura de bases $\geq 20X$	Cobertura média das bases
XP02AM	1085585 (34.9579%)	2019826 (65.0421%)	2% (24498 bp)	97.99%	93.42%	88.94%	80.78%	65,43
XP02RJ	318871 (29.4291%)	764653 (70.5709%)	4.1% (49593 bp)	95.93%	84.27%	68.72%	36.05%	19,2
XPSPAC02F0	433576 (30.5978%)	983441 (69.4022%)	3.1% (37553 bp)	96.92%	88.79%	78.54%	52.52%	26,1
GO01	565474 (37.5395%)	940869 (62.4605%)	2.7% (32606 bp)	97.32%	90.42%	83.18%	66.50%	34,22
GO03	19366 (1.718%)	1107877 (98.282%)	45% (545040 bp)	55.25%	3.92%	0.19%	0.00%	1,15
GO04	285789 (32.1965%)	601850 (67.8035%)	4.4% (53702 bp)	95.59%	82.85%	64.92%	30.32%	17,26
GO05	2055402 (34.06%)	3979248 (65.94%)	1.4% (17234 bp)	98.59%	95.71%	92.99%	88.20%	123,91
GO06	1742542 (35.8671%)	3115787 (64.1329%)	1.5% (17808 bp)	98.54%	95.22%	91.81%	86.25%	105,07
XPSPAC01F0	826372 (35.4351%)	1505701 (64.5649%)	2% (24697 bp)	97.97%	92.87%	87.53%	76.69%	49,77
XPSPAC04F0	24541 (31.2939%)	53880 (68.7061%)	37% (450634 bp)	63.00%	6.56%	0.49%	0.00%	1,48
XPSPAC06F0	930483 (33.5518%)	1842793 (66.4482%)	1.9% (22910 bp)	98.12%	93.62%	88.81%	79.86%	56,17
XPSPAC08F0	667531 (37.4423%)	1115293 (62.5577%)	2.8% (34384 bp)	97.18%	90.44%	83.49%	68.17%	40,16
XPSPAC11F0	1221464 (35.4174%)	2227300 (64.5826%)	1.7% (21138 bp)	98.26%	94.13%	89.97%	82.79%	73,66
XPSPAC12F1	2079663 (36.7495%)	3579364 (63.2505%)	1% (12680 bp)	98.96%	96.80%	94.51%	90.55%	125,47
XPSPAC13F0	57934 (36.087%)	102606 (63.913%)	19% (237420 bp)	80.51%	26.52%	7.10%	1.08%	3,5
XPSPAC31F0	1051976 (35.7986%)	1886623 (64.2014%)	1.7% (21169 bp)	98.26%	94.18%	89.96%	82.06%	63,47

Tabela 4: Quantidades de variantes chamadas pelos 3 pipelines de busca distintos utilizados na primeira etapa. A região alvo corresponde às variantes presentes apenas nos genes estudados e a região alvo de alta qualidade corresponde às variantes que não possuem a assinatura de baixa qualidade (*LowQual*).

Amostra	Variantes capturadas								
	Gatk	Região alvo	Região alvo alta qualidade	Samtools	Região alvo	Região alvo alta qualidade	Gatk + Samtools	Região alvo	Região alvo alta qualidade
XP02AM	8515	829	776	4345	1473	1113	2039	766	665
XP02RJ	2664	703	608	1883	1216	896	976	628	444
XPSPAC02F0	2794	627	558	2018	1177	845	960	596	467
GO01	4242	771	727	2395	1303	980	1257	693	581
GO03	2553	63	42	543	19	11	366	4	1
GO04	2228	621	551	1695	1195	806	782	545	405
GO05	15970	683	641	7526	1289	929	3077	610	541
GO06	14760	844	788	6806	1465	1082	3034	762	664
XPSPAC01F0	178	118	85	42	33	25	21	16	10
XPSPAC04F0	7397	878	819	3967	1474	1118	1868	796	681
XPSPAC06F0	4540	912	858	2859	1510	1166	1468	827	652
XPSPAC08F0	6533	919	860	3295	1521	1149	1616	813	695
XPSPAC11F0	7948	587	563	4001	1128	787	1732	541	495
XPSPAC12F1	16585	816	779	8025	1353	1021	3465	727	664
XPSPAC13F0	346	183	140	207	143	100	112	76	49
XPSPAC31F0	7711	858	798	4132	1549	1188	1870	771	665

Tabela 5: As variantes restantes foram divididas em filtradas (aquelas que foram encontradas no dbSNP138) e as novas (sem entrada no banco de dados). A probabilidade de se encontrar uma variante nova é pequena, quando comparada às já depositadas, logo a frequência dessa distribuição realizada reflete a precisão do buscador.

Amostra	Comparação com dbSNP 138								
	Gatk			Samtools			Gatk+Samtools		
	Filtrado	Novo	% de combinação	Filtrado	Novo	% de combinação	Filtrado	Novo	% de combinação
XP02AM	748	29	96,27	773	340	69,45	624	41	93,83
XP02RJ	587	21	96,55	707	189	78,91	424	20	95,50
XPSPAC02F0	545	14	97,50	636	209	75,27	442	25	94,65
GO01	692	35	95,19	720	260	73,47	549	32	94,49
GO03	41	1	97,62	8	3	72,73	1	0	100,00
GO04	536	15	97,28	608	198	75,43	384	21	94,81
GO05	605	36	94,38	622	307	66,95	504	37	93,16
GO06	754	34	95,69	774	308	71,53	625	39	94,13
XPSPAC01F0	84	1	98,82	22	3	88,00	9	1	90,00
XPSPAC04F0	786	33	95,97	838	280	74,96	640	41	93,98
XPSPAC06F0	834	24	97,20	882	284	75,64	616	36	94,48
XPSPAC08F0	822	39	95,47	842	307	73,28	659	36	94,82
XPSPAC11F0	542	21	96,27	536	251	68,11	470	25	94,95
XPSPAC12F1	734	45	94,22	729	292	71,40	635	29	95,63
XPSPAC13F0	138	2	98,57	83	17	83,00	47	2	95,92
XPSPAC31F0	754	44	94,49	779	409	65,57	629	36	94,59

Para analisar as diferenças entre os dois buscadores de variantes oferecidos pelo GATK, o *Unified Genotyper* (UC) e *Haplotype Caller* (HC), foi feita a busca por variantes com os dois programas utilizando os mesmos parâmetros (sc 22, se 10). A variação no desempenho de ambos programas foi significativa (Tabela 6) e, ao se analisar o arquivo de saída, pôde-se observar que algumas variantes que estavam na saída de um dos programas, não estavam no outro, e vice-versa. Além disso, dentro do nosso critério de escolha, baseado na maior quantidade de variantes chamadas que estão contidas no dbSNP, o *Unified Genotyper* seria o escolhido para ser utilizado na interface *web*. Entretanto, apesar desses resultados, discussões e notas de atualizações oferecidas pela própria equipe de desenvolvimento do programa GATK, aconselham o uso do *Haplotype Caller* como o principal buscador de variantes pois a sua capacidade para chamar substituições é equivalente ao do UC, porém a sua capacidade para chamar *indels* é muito superior (<http://gatkforums.broadinstitute.org/discussion/3151/should-i-use-unifiedgenotyper-or-haplotypecaller-to-call-variants-on-my-data>; http://www.broadinstitute.org/gatk/pdfdocs/GATK_GuideBook_2.7-2.pdf). Analisamos então as quantidades de *indels* chamados por ambos os buscadores (Tabela 7) e podemos observar que o HC chegou a chamar em torno de 50% mais *indels* que o UC. Devido a isso, é possível entender melhor a variação apresentada na Tabela 6. Sabe-se que a identificação, anotação e o depósito de *indels* em banco de dados ainda é um desafio e não possui um formato estabelecido padrão. Por esse motivo, como a quantidade de *indels* encontrada no HC é muito superior, aumenta-se a probabilidade dessas variantes não estarem depositados no banco de dados, o que faz com que a porcentagem de compartilhamento com as variantes presentes no dbSNP142 tenha uma leve queda. Por esses motivos, o HC foi escolhido para ser utilizado na interface *web*.

Tabela 6: Resultados ao se comparar a quantidade de *reads* filtrados entre os dois buscadores de variantes do programa GATK. Nota-se que ambos chamam aproximadamente a mesma quantidade de variantes, porém o HC chama muito mais variantes como novas quando comparada com o dbSNP142.

Amostra	Unified Genotyper			Haplotype Caller		
	sc22se10		% de variantes presentes no DB	sc22se10		% de variantes presentes no DB
	Filtrado	Novo		Filtrado	Novo	
XP02AM	829	39	95,51	848	59	93,50
XP02RJ	703	20	97,23	641	15	97,71
XPSPAC02F0	627	14	97,82	594	32	94,89
GO01	771	45	94,49	801	55	93,57
GO03	63	1	98,44	62	3	95,38
GO04	621	19	97,03	619	33	94,94
GO05	683	45	93,82	702	79	89,88
GO06	844	47	94,73	892	83	91,49
XPSPAC01F0	919	40	95,83	904	42	95,56
XPSPAC04F0	816	59	93,26	838	85	90,79
XPSPAC06F0	183	2	98,92	171	3	98,28
XPSPAC08F0	858	50	94,49	895	94	90,50
XPSPAC11F0	118	5	95,93	114	6	95,00
XPSPAC12F1	878	45	95,12	869	43	95,29
XPSPAC13F0	912	21	97,75	954	48	95,21
XPSPAC31F0	587	50	92,15	620	67	90,25

Tabela 7: Comparação da quantidade de *indels* chamados pelos dois buscadores de variantes do programa GATK. Nota-se que o HC chama em torno de 50% mais variantes que o UC.

Amostra	Indels chamados	
	Unified Genotyper	Haplotype Caller
XP02AM	47	88
XP02RJ	23	50
XPSPAC02F0	25	51
GO01	33	83
GO03	0	3
GO04	21	48
GO05	35	99
GO06	40	105
XPSPAC01F0	47	78
XPSPAC04F0	35	99
XPSPAC06F0	4	8
XPSPAC08F0	37	103
XPSPAC11F0	0	7
XPSPAC12F1	45	75
XPSPAC13F0	39	85
XPSPAC31F0	37	88

Após determinar o buscador de variantes do GATK a ser utilizado em nossa interface final, todas as amostras foram submetidas a diferentes valores de dois parâmetros de busca com a finalidade de identificar quais seriam os melhores valores para cada amostra. Utilizamos para os parâmetros “-stand_call_conf” e “-emit_call_conf” os valores de, respectivamente (22,10), (20,15), (30,10), (30,15) (30,20) e (50,15). Podemos observar, baseando nesses valores atribuídos, que as variantes chamadas para o arquivo VCF final são dependentes do parâmetro de emissão, não importando o valor atribuído -stand_call_conf. Entretanto, quanto mais se aumenta esse último, mais variáveis são filtradas como “LowQual”. Além disso, as bases pouco cobertas por *reads*, que apresentam uma qualidade geral menor, são perdidas (Tabela 8). Isso ocorre devido ao baixo número de *reads* cobrindo determinada base, o que faz a qualidade de filtragem geral diminuir, porém, o limiar de emissão de variantes, que é baseada na qualidade de todas as bases dos *reads* que a cobrem a mesma posição, é superior e, então, essa variante torna-se mais confiável. Caso não atinja nenhum dos casos, a variante não aparece no arquivo de saída.

Para analisar as amostras desde projeto, visando a obtenção de um resultado mais confiável, utilizamos os valores sc30/se15 para as amostras com boa qualidade de sequenciamento, enquanto que para as 3 amostras com baixa qualidade utilizamos os valores sc30/se10. Além disso, não foram removidas as variantes marcadas com a assinatura “LowQual”, pois ao se realizar a busca por variantes com esses valores utilizados, garantimos uma maior qualidade de bases para serem chamadas (http://www.broadinstitute.org/gatk/pdfdocs/GATK_GuideBook_2.7-2.pdf). Apesar desses valores terem sido escolhidos para a análise das nossas amostras, o usuário do *Human Variants Finder Interface* poderá alterar esses valores para adequar sua análise as qualidades de cada amostra dele.

4.3. A SAÍDA FINAL

Após a chamada das variantes, a anotação foi feita e filtros foram aplicados na tentativa de encontrar mutações possivelmente relacionadas às doenças de reparo do DNA. O principal filtro utilizado nos resultados foi a de remoção de variantes que já tinham sido caracterizadas como um SNP (frequência alélica $\geq 1\%$) em várias

populações, como a asiática (sul e leste), americana, europeia e africana, pelo projeto *1000genomes* (<http://www.1000genomes.org/>). Para restringir ainda mais em busca do possível candidato à mutação patogênica, foram removidas as variantes presentes nos íntrons, caso não estivessem localizadas em regiões de sítios de *splicing*.

O resultado final, para cada paciente dessa primeira etapa de testes, teve a finalidade de reunir todas as informações relevantes geradas através da busca de variantes e anotação gênica. Além disso, visou apresentar esse resultado de forma clara e objetiva para aqueles que não são habituados com análises de arquivos de texto, gerando uma saída, em primeira instância, como a da Tabela 9 (paciente XP02AM). As das outras amostras poderão ser visualizadas no Anexo 1. A partir desses resultados, o usuário está apto a analisar seus próprios dados com uma maior facilidade e praticidade e tomar decisão de quais variantes podem ser candidatas como responsável pelo fenótipo do paciente. No caso apresentado, a variante em homozigose no gene *DDB2* (ou XPE) que resulta em um códon de parada aparece como mutação candidata mais provável como responsável ao fenótipo XP do paciente XP02AM.

emit_call_conf.

Amostra	Haplotype Caller											
	sc22se10		sc20se15		sc30se10		sc30se15		sc30se20		sc50se15	
	Filtrado	Variantes "LowQual"	Filtrado	Variantes "LowQual"	Filtrado	Variantes "LowQual"	Filtrado	Variantes "LowQual"	Filtrado	Variantes "LowQual"	Filtrado	Variantes "LowQual"
XP02AM	848	45	826	18	848	67	826	45	808	27	826	98
XP02RJ	641	81	600	27	641	137	600	96	573	69	600	191
XPSPAC02F0	594	58	565	22	594	86	565	57	543	35	565	126
GO01	801	40	779	14	801	74	779	52	765	38	779	109
GO03	62	19	54	3	62	34	54	26	51	23	54	42
GO04	619	56	589	15	619	94	589	64	574	49	589	151
GO05	702	25	693	11	702	42	693	33	682	22	693	82
GO06	892	28	878	11	892	55	878	41	867	30	878	90
XPSPAC01F0	904	48	883	18	904	79	883	58	865	40	883	131
XPSPAC04F0	838	26	825	8	838	41	825	28	817	20	825	59
XPSPAC06F0	171	39	155	14	171	66	155	50	141	36	155	87
XPSPAC08F0	895	49	880	22	895	74	880	59	858	37	880	120
XPSPAC11F0	114	28	102	14	114	59	102	47	88	33	102	74
XPSPAC12F1	869	36	852	14	869	66	852	49	838	35	852	118
XPSPAC13F0	954	66	929	28	954	106	929	81	901	53	929	169
XPSPAC31F0	620	26	608	10	620	38	608	26	598	16	608	74

Tabela 9: Exemplo do tipo de saída e as algumas informações resultantes de cada análise (paciente XP02AM). Nota-se a presença de uma variável não depositada no bando de dados dbSNP142 (RS_ID = “.”), em homozigose, na região exônica gerando um códon de parada no exon 7 do gene DDB2 (XPE) – em amarelo. Provavelmente essa mutação é uma forte candidata a ser patogênica.

CHROM	POSITION	RS_ID	REF	ALT	#REF	#ALT	GENE	REGION	GENOTYPE	VAR_TYPE	GENE_ID	EXON	CDNA_POS	PROT_POS	#QUAL
chr3	10088299	rs112887807	C	T	8	3	FANCD2	exonic	het	synonymous SNV	NM_001018115	exon15	c.C1170T	p.S390S	59
chr16	14045774	.	A	G	10	8	ERCC4	UTR3	het		NM_005236	-	c.*3570A>G		78
chr3	10088266	rs72492997	G	T	17	4	FANCD2	exonic	het	synonymous SNV	NM_001018115	exon15	c.G1137T	p.V379V	16
chr19	45911319	rs11314106	CA	C	2	7	ERCC1	UTR3	het		NM_001166049	-	c.*1613delT		107
chr3	10108898	rs77246387	A	G	16	7	FANCD2	exonic	het	synonymous SNV	NM_001018115	exon26	c.A2391G	p.V797V	67
chr11	47256943	.	C	T	0	15	DDB2	exonic	hom	stopgain	NM_000107	exon7	c.C1003T	p.Q335X	364
chr16	89804335	rs17233826	CT	C	1	13	FANCA	UTR3	hom		NM_000135	-	c.*673delA		414
chr16	14045297	rs183916977	T	C	10	14	ERCC4	UTR3	het		NM_005236	-	c.*3093T>C		228
chr3	10088308	rs72492998	T	C	6	3	FANCD2	exonic	het	synonymous SNV	NM_001018115	exon15	c.T1179C	p.T393T	89
chr16	14043490	rs185626419	A	G	6	5	ERCC4	UTR3	het		NM_005236	-	c.*1286A>G		49

4.4. A ANÁLISE DAS VARIANTES POTENCIONALMENTE PATOGÊNICAS

Essa análise visou avaliar os genes relacionados ao reparo de DNA com a finalidade de tentar esclarecer o diagnóstico molecular dos pacientes cujas amostras foram sequenciadas. Na Tabela 10, listamos as variantes que acreditamos que possam ser as mutações responsáveis pelo quadro clínico dos pacientes identificados com XP, com base nas variantes demonstradas no Anexo 1.

As amostras dos pacientes de Araras (GO01, GO03~6), assim como a amostra do paciente XP02RJ, foram utilizadas como controle, pois já eram conhecidas suas mutações. Como já demonstrado na Figura 3, os pacientes de Araras possuem grau de parentesco e alguns não são acometidos pela doença, sendo apenas portadores do alelo com a mutação (heterozigotos). Esse é o caso dos pacientes GO01 e GO05, que são os que possuem uma mutação na mesma posição do gene *POLH* em heterozigose, afetando um sítio de *splicing*, justificando o caso clínico de XP-V no seu filho, o paciente GO06, homozigoto para com mutação.

Alguns resultados ainda foram inconclusivos pelo fato das amostras estarem com qualidade abaixo do esperado, ou simplesmente por não apresentarem as variantes esperadas, como no caso de algumas amostras controles. Não encontramos uma das duas mutações já conhecidas da amostra controle GO04, apesar da boa qualidade de alinhamento deste sequenciamento. O mesmo pode se dizer da amostra GO03 que, no entanto, apresentava baixa cobertura no sequenciamento. Além desses

pacientes, na amostra controle XP02RJ também não foram encontradas as duas mutações já conhecidas. Neste paciente, haviam sido descritas duas mutações no gene *ERCC5* (SOLTYS *et al.*, 2013). Entretanto, nenhuma das duas foi identificada pelo sequenciamento por SOLiD, mesmo com uma média de cobertura relativamente boa (36% das bases com >20x de cobertura). Coincidentemente, as amostras GO04 e XP02RJ foram classificadas como de boa qualidade quanto à cobertura, quando comparadas com as demais amostras, o que torna difícil atribuir a ausência de detecção dessas variantes como um problema de regiões ricas em CG, comum viés entre todas as plataformas de sequenciamento, principalmente em sequenciador SOLiD (RIEBER *et al.*, 2013). De qualquer forma, algumas amostras foram reavaliadas em novo processo de sequenciamento NGS em outra plataforma (MiSeq, Illumina).

O paciente XP02AM, como indicado acima, apresentou uma mutação nova em homozigose, localizada no gene *DDB2* exon 7, que gera um códon de parada. O surgimento desse códon de parada geraria, durante a tradução, uma proteína truncada, provavelmente inativando-a e, conseqüentemente, poderia explicar o quadro clínico do paciente. Chamamos atenção que mutações no gene *DDB2/XPE* são muito raras, e resultam em um fenótipo mais leve nos pacientes. No caso, o fenótipo é muito agressivo, o que provavelmente se explica pela sua vida próximo ao Equador (Manaus, Amazônia). O paciente XPSPAC02F0 também apresentou uma substituição nova em homozigose, não sinônima, localizada no gene *POLH (XPV)* exon 5. Foi necessário submeter esta mutação aos programas que analisam o efeito sobre a proteína (Tabela 11). Nota-se que o PROVEAN relatou que esta mutação possui caráter deletério, enquanto o SIFT demonstrou que a mesma é tolerável. Entretanto, as amostras XPSPAC08F0 e XPSPAC11F0, também apresentam a mesma mutação, sendo ambos pacientes com sintomatologia XP. Esses três pacientes não possuem nenhuma relação de parentesco, segundo as pesquisadoras do Hospital AC Camargo, com isso podemos inferir que esta mutação provavelmente é de fato clinicamente significativa no diagnóstico desses pacientes. O paciente XPSPAC01F0 apresentou uma mutação nova em homozigose, não sinônima, localizada no gene *POLH (XPV)* exon 8. A análise pelo PROVEAN identifica essa mutação como deletéria, e SIFT como danosa, sendo possível inferir essa mutação como provável responsável pelo quadro clínico do paciente.

Tabela 10: Levantamento das possíveis variantes que possam justificar o diagnóstico clínico dos pacientes, muitas delas ainda não descritas ou reportadas em banco de dados. Cinza: Amostra controle que não obtivemos as variantes, apesar da boa qualidade. Vermelho – amostra de baixa qualidade. Laranja – variantes que supomos estar envolvidas com a doença, provavelmente pela possibilidade de ser uma mutação heterozigótica composta (XPSPAC04F0 e XPSPAC12F1).

Amostra	CHROM	POSITION	RS_ID	REF	ALT	#REF	#ALT	#DEPH	GENE	REGION	GENOTYPE	VAR_TYPE	GENE_ID	EXON	CDNA_POS	PROT_POS	#QUAL
XP02AM	chr11	4725943	.	C	T	0	15	15	DB2	exonic	hom	stopgain	NM_000107	exon7	c.C1003T	p.O335X	364
XP02RJ																	
XPSPAC02F0	chr6	43565513	.	A	C	0	9	9	POLH	exonic	hom	nonsynonymous S	NM_006502	exon5	c.A571C	p.T191P	176
G001	chr6	43568829	.	G	A	16	7	23	POLH	splicing	het		NM_006502	exon6	c.764+1G>A		50
G003	chr6	43568829	.	G	A	0	2	2	POLH	splicing	hom		NM_006502	exon6	c.764+1G>A		25
G004	chr6	43568829	.	G	A	7	12	19	POLH	splicing	het		NM_006502	exon6	c.764+1G>A		173
G005	chr6	43568829	.	G	A	17	9	26	POLH	splicing	het		NM_006502	exon6	c.764+1G>A		123
G006	chr6	43568829	.	G	A	3	33	36	POLH	splicing	hom		NM_006502	exon6	c.764+1G>A		737
XPSPAC01F0	chr6	43572429	.	G	T	0	31	31	POLH	exonic	hom	nonsynonymous S	NM_001291969	exon6	c.G590T	p.C197F	765
XPSPAC04F0	chr6	43578436	.	ATACT	A	28	12	43	POLH	exonic	het	frameshift dele	NM_006502	exon10	c.1221_1224del	p.N407fs	383
XPSPAC06F0																	
XPSPAC08F0	chr6	43565513	.	A	C	1	27	28	POLH	exonic	hom	nonsynonymous S	NM_001291969	exon3	c.A199C	p.T67P	558
XPSPAC11F0	chr6	43565513	.	A	C	0	2	2	POLH	exonic	hom	nonsynonymous	NM_006502	exon5	c.A571C	p.T191P	21,77
XPSPAC12F1	chr10	50723780	rs142477138	C	T	25	11	36	ERCC6-PGBD3,PGBD3	exonic	het	nonsynonymous S	NM_170753	exon2	c.G1381A	p.G461R	145
	chr10	5080422	rs145720191	C	T	13	7	20	ERCC6	exonic	het	nonsynonymous S	NM_000124	exon16	c.G2924A	p.R975Q	101
XPSPAC13F0	chr3	14190232	.	C	G	0	4	4	XPC	splicing	hom		NM_004628	exon14	c.2251+1G>C		93
XPSPAC31F0	chr6	43565576	.	T	C	2	18	20	POLH	exonic	hom	nonsynonymous S	NM_006502	exon5	c.T634C	p.C212R	165

Apesar da amostra do paciente XPSPAC04F0 apresentar uma boa cobertura, não foram encontradas variantes no sequenciamento que justificassem seu caso clínico, com exceção do fato dele apresentar uma deleção (que provoca uma *frameshift* e, portanto, deve resultar em uma proteína truncada) em heterozigose no exon 10 do gene *POLH* (*XPV*). Nesse caso, é provável que outra variante patogênica no outro alelo cromossômico desse gene que não foi detectada por esse sequenciamento. O paciente XPSPAC13F0 foi utilizado como controle pelo grupo do Hospital AC Camargo. Sua mutação (em sítio de *splicing* no gene *XPC*), além de ter sido encontrada pela análise das variantes através desse projeto, foi encontrada e confirmada por sequenciamento Sanger pelos pesquisadores do hospital. Por fim, o paciente XPSPAC31F0 apresentou uma mutação nova em provável homozigose, localizada no gene *POLH* (*XPV*) exon 5, sendo classificada como deletéria pelos programas de análise PROVEAN e danosa pelo SIFT. Portanto, essa deve ser a mutação responsável pelos sintomas XP do paciente.

Finalmente, no caso da amostra XPSPAC12F1 não encontramos mutações em nenhum dos genes XP analisados. Curiosamente, encontramos 2 variantes em heterozigose relacionadas ao gene *ERCC6* (*CSB*), sendo que uma delas localiza-se no transposon piggyback (*PGBD3*), que se encontra inserido no intron 5 do gene *ERCC6* (*CSB*). É relatado que, por *splicing* alternativo, ocorre a produção de uma proteína de fusão CSB-PGBD3, que foi relatada recentemente como participante do processo de reparo de DNA, atuando como um sinérgico do produto de CSB, aumentando 200~250% o reparo de lesões UV (WEINER; GRAY, 2014). Com essas informações, imagina-se que caso essa proteína de fusão e o CSB estejam reduzidos no indivíduo, possa não apresentar um desempenho satisfatório em suas funções do reparo da lesão. As duas variantes são identificadas como deletérias (PROVEAN) ou danosas (SIFT) para as proteínas, o que nos chama ainda mais a atenção. Este seria o segundo caso de paciente XP com mutações no gene *ERCC6* (*CSB*), sendo o primeiro caso relatado para um paciente com sintomas clínicos graves que incluem neurodegeneração (característica de pacientes CS) (COLELLA *et al.*, 2000). Entretanto, o paciente XPSPAC12F1 apresenta apenas fenótipo XP. Além disso, as duas variantes já foram encontradas no projeto *1000genomes*. Sendo assim, certamente a amostra desse paciente deverá ser melhor investigada através de um novo sequenciamento (que deverá incluir familiares afetados e/ou não afetados), além

de prevermos obter cultura de fibroblastos do paciente, através de biópsia de pele. Essas células poderão ajudar nosso grupo a entender o que ocorre nesse caso.

Tabela 11: Resultados gerados pelas plataformas PROVEAN e SIFT, respectivamente, para as potenciais variantes candidatas ao perfil clínico dos pacientes.

INPUT	LENGTH	CODON_CHANGE	POS	RESIDUE REF	RESIDUE ALT	TYPE	SCORE	PREDICTION (cutoff=-2.5)	SCORE2	PREDICTION (cutoff=0.05)	MEDIAN INFO
6,43565513,A,C,XPSPAC02F0 XPSPAC08F0 XPSPAC11F0	713	CTC [A/C]CC GTG	191	T	P	Single AA Change	-2.79	Deleterious	0.144	Tolerated	2.84
6,43572429,G,T,XPSPAC01F0	713	GGC T[G/T]T AGT	321	C	F	Single AA Change	-9.46	Deleterious	0.000	Damaging	2.84
6,43578436,ATACT,A,XPSPAC04F0	713					Frameshift	NA	NA	NA	NA	NA
10,50723780,C,T,XPSPAC12F1	1061	ATG [G/A]GA GGC	929	G	R	Single AA Change	-2.57	Deleterious	0.007	Damaging	3.97
10,50680422,C,T,XPSPAC12F1	1493	CAC C[G/A]A CAA	975	R	Q	Single AA Change	-3.84	Deleterious	0.000	Damaging	2.83
6,43565576,T,C,XPSPAC31F0	713	CAG [T/C]GT TCA	212	C	R	Single AA Change	-11.66	Deleterious	0.000	Damaging	2.84

4.5. A PLATAFORMA DE DOMÍNIO PÚBLICO E SUA INTERFACE

A *Human Variants Finder Interface* engloba todos os procedimentos descritos nesse trabalho, com exceção da análise de relatórios de qualidade do sequenciamento e da análise final das variantes filtradas, pois estes precisam de um julgamento crítico-humano. Sua interface visa facilitar a análise deste tipo de dado por pesquisadores que não conhecem o mundo da bioinformática, bem como suas ferramentas e ambiente Linux, e necessitam de um pré-processamento de qualidade com resultados confiáveis e mais objetivo possível através da realização de filtros chaves para a obtenção de melhores resultados.

Essa plataforma foi desenvolvida para abranger, além dos genes envolvidos na via de reparo NER, regiões específicas de interesse do usuário, como análise do exoma completo, genoma ou mesmo genes específicos. Nas Figuras 8 e 9, está sendo ilustrado o *layout* intuitivo de nossa plataforma. O usuário pode entrar com seu arquivo (BAM ou VCF), escolher a plataforma em que sua amostra foi sequenciada, as regiões alvo de interesse, os valores dos parâmetros utilizados na busca de variantes e se gostaria que os *reads* duplicados fossem removidos (para estudo genômico é altamente recomendável remover esses *reads*). Ao finalizar as análises, o usuário será notificado por e-mail sobre a conclusão do processo e poderá visualizar e aplicar filtros, em tempo real, para apurar a obtenção de suas variantes dependendo de seu interesse (Figuras 10a e 10b).

Toda interface *web* está acoplada a um *script* desenvolvido em *Perl* que realiza toda a análise e que está vinculado ao gerenciador de bancos MySQL (Figura 11). Atualmente a plataforma se encontra hospedada em um dos nossos servidores mantidos na Nuvem provida pela USP (<https://nuvem.uspdigital.usp.br>).

A *Human Variants Finder Interface* foi criada visando a identificação das variantes, facilitando e agilizando a análise genômica em poucos cliques, gerando resultados confiáveis de forma simplificada e podendo ser utilizada por usuários de diferentes níveis de conhecimento da área de bioinformática.

Figura 8: *Layout* principal da *Human Variants Finder Interface* desenvolvida e disponibilizada. Ela foi criada para ser limpo e intuitivo para que qualquer pessoa não tenha dificuldade em utilizá-lo.

Figura 9: *Layout* de conclusão de tarefas da *Human Variants Finder Interface*. O usuário poderá obter esses o relatório do processamento de suas amostras providos pelos programas que foram utilizados.

This table contains 89 results (Showing results 1 to 30)

CHROM	POSITION	RS_ID	REF	ALT	#QUAL	#REF	#ALT	#DEPH	GENE	REGION	GENOTYPE	VAR_TYPE	GENE_ID
chr9	97872712	rs1042961...	C	T	442	0	22	22	FANCC	Not Equal	hom		NM_0012437...
chr9	32888473	rs206115	C	T	545	14	17	31	ZAR1L<->BR...	Not Equal	het		-
chr13	103493180	rs2018836	A	G	163	9	10	19	BIVM	OR	het		NM_017693
chr19	45923653	rs11615	A	G	25	6	4	10	ERCC1	Contains	het	synonym...	NM_001983
chr13	32973012	rs15869	A	C	50	12	8	20	BRCA2	UTR3	het		NM_000059
chr13	103527849	rs9514066	G	C	333	0	22	22	BIVM-ERCC5...	exonic	hom	nonsynonym...	NM_000123
chr19	45873942	rs3810366	G	C	49	0	3	4	ERCC2	upstream	hom		-
chr13	103493487	rs1571061	T	C	43	18	11	29	BIVM	UTR3	het		NM_017693
chr16	14047418	rs9646271	C	T	347	10	19	29	ERCC4<->L...	Intergenic	het		-
chr19	45910672	rs1046282	A	G	44	16	10	26	ERCC1	UTR3	het		NM_0011660...
chr10	50740876	rs2228524	G	C	391	0	18	18	ERCC6-ERCC...	exonic	hom	synonym...	NM_001270...
chr13	103504517	rs1047768	T	C	41	15	10	25	BIVM-ERCC5...	exonic	het	synonym...	NM_000123
chr6	158615926	rs1019682	A	T	55	2	5	7	GTF2H5	UTR3	het		NM_207118
chr17	7579801	rs1042785	G	C	63	1	4	5	TP53	UTR5	het		NM_0011261...
chr10	50662182	rs3980751...	A	AT	585	8	25	39	ERCC6	downstream	het		-
chr11	47261043	rs326221	T	C	560	0	24	24	ACP2	UTR3	hom		NM_001610
chr10	50680471	rs1908638...	C	A	132	13	9	22	ERCC6	exonic	het	nonsynonym...	NM_000124
chr13	32889483	rs206116	G	A	538	15	18	33	ZAR1L<->BR...	Intergenic	het		-
chr19	45927610	rs3212930	A	G	129	8	10	18	ERCC1	upstream	het		-
chr17	7578645	rs2009430	C	T	499	0	26	26	TP53	UTR5	hom		NM_0011261...
chr10	50666808	rs4253231	A	G	181	9	12	22	ERCC6	UTR3	het		NM_000124
chr13	32974288	rs1148321	T	A	193	0	11	11	BRCA2.N4BP...	downstream	hom		-
chr16	14046149	rs3837752	AT	A	129	9	7	22	ERCC4	UTR3	het		NM_005236
chr16	14045399	rs4781563	G	A	142	17	12	29	ERCC4	UTR3	het		NM_005236
chr19	45912736	rs3212986	C	A	26	6	3	9	CD3EAP	exonic	het	nonsynonym...	NM_0012975...
chr6	158618588	rs6456077	C	T	101	14	9	23	GTF2H5	UTR3	het		NM_207118
chr16	14045236	rs4781562	G	T	160	21	12	33	ERCC4	UTR3	het		NM_005236
chr13	32906729	rs144848	A	C	77	16	8	24	BRCA2	exonic	het	nonsynonym...	NM_000059
chr19	45854919	rs13181	T	G	349	2	19	21	ERCC2	exonic	hom	synonym...	NM_000400
chr13	32913055	rs206075	A	G	308	0	19	19	BRCA2	exonic	hom		NM_000059

Figura 10a: Representação dos resultados finais já anotados bem como as diversas possibilidades de filtros que podem ser realizados com o arquivo VCF final (informações iniciais que são obtidas). O usuário poderá obter, além da versão final original, os dados processados após a customização dos filtros.

This table contains 89 results (Showing results 1 to 30)

EXON	CDNA_POS	PROT_POS	ISOFORMS	CLINVAR	COSMIC	#ESP6500	#EXAC	#IKGEN_AMR	#IKGEN_EUR	#IKGEN_SAS	#IKGEN_AFR	#IKGEN_EAS
-	c.*766G>A	-	-	-	-	0	0	1	0.9901	0.9901	1	1
-	c.*965A>G	-	-	-	-	0	0	0.4788	0.5596	0.5596	0.4788	0.4788
exon4	c.T354C	p.N118N	ERCC1:NM...	-	-	0	0	0.68	0.7147	0.7147	0.68	0.68
-	c.*105A>C	-	-	yes	-	0.5401	0.5085	0.9629	0.3777	0.3777	0.9629	0.9629
exon15	c.G3157C	p.G1053R	BIVM-ERCC5...	-	-	0	0	0.0113	0.2356	0.2413	0.0113	0.0113
GENE ID: NM_000123	-	-	-	-	-	0.9975	0.9994	0.9932	1	1	0.9932	0.9932
-	c.*1272T>C	-	-	-	-	0	0	0.9266	0.4473	0.4479	0.9266	0.9266
-	c.*2261T>C	-	-	-	-	0	0	0.9365	1	1	0.9365	0.9365
exon2	c.C135G	p.L45I	ERCC6-PGBD...	yes	-	0.6276	0.6946	0.3896	0.2565	0.3016	0.3896	0.3896
exon2	c.T138C	p.H46H	BIVM-ERCC5...	yes	COSM3753...	0.6118	0.5291	0.3336	0.7346	0.8231	0.3336	0.3336
-	c.*2737A>T	-	-	-	-	0	0	0.6778	0.5845	0.4663	0.6778	0.6778
-	c.*232C>G	-	-	yes	-	0.6669	0.6689	0.9576	0.6909	0.6544	0.9576	0.9576
-	-	-	-	-	-	0	0	0.4924	0.7117	0.5051	0.4924	0.4924
-	c.*624A>G	-	-	-	-	0	0	0.211	0.5249	0.4867	0.211	0.211
exon16	c.G2875T	p.V959L	-	-	-	0	0	0.9728	1	1	0.9728	0.9728
-	-	-	-	-	-	0	0.0017	0	0	0	0	0
-	-	-	-	-	-	0	0	0.3268	0.3926	0.362	0.3268	0.3268
-	c.*112G>A	-	-	-	-	0	0	0.0703	0.2147	0.183	0.0703	0.0703
-	c.*53T>C	-	-	-	-	0	0	0.7141	0.8579	0.8149	0.7141	0.7141
-	-	-	-	-	-	0	0	0.1059	0.0994	0.2229	0.1059	0.1059
-	-	-	-	-	-	0	0	0.8752	0.7545	0.7106	0.8752	0.8752
-	c.*3946delT	-	-	-	-	0	0	0.1626	0.2893	0.2331	0.1626	0.1626
-	c.*3195G>A	-	-	-	-	0	0	0.1679	0.2893	0.2331	0.1679	0.1679
exon3	c.C1516A	p.Q506K	CD3EAP:NM...	-	-	0.2612	0.3237	0.2912	0.2505	0.3006	0.2912	0.2912
-	c.*5399C>T	-	-	-	-	0	0	0.2625	0	0	0.2625	0.2625
-	c.*3032G>T	-	-	-	-	0	0	0.3245	0.2893	0.3221	0.3245	0.3245
exon10	c.A1114C	p.N372H	-	yes	COSM1476...	0.2332	0.2779	0.084	0.2893	0.3538	0.084	0.084
exon23	c.A2251C	p.K751Q	-	yes	COSM4132...	0.326	0.3308	0.1921	0.3638	0.3466	0.1921	0.1921
exon11	c.A4563G	p.L1521L	-	yes	-	0.9758	0.9931	0.9039	0.999	1	0.9039	0.9039

Figura 10b: Representação dos resultados finais bem como as diversas possibilidades de filtros que podem ser realizados com o arquivo VCF final (informações finais que são obtidas). O usuário poderá obter, além da versão final original, os dados processados após a customização dos filtros.

#	Nome	Tipo	Agrupamento (Collation)	Atributos	Nulo
1	id	int(10)		UNSIGNED	Não
2	user_id	int(10)		UNSIGNED	Sim
3	status	tinyint(4)			Sim
4	file_one	text	latin1_swedish_ci		Sim
5	ext_one	tinytext	latin1_swedish_ci		Sim
6	file_two	text	latin1_swedish_ci		Sim
7	ext_two	tinytext	latin1_swedish_ci		Sim
8	custom_regions	text	latin1_swedish_ci		Sim
9	job_type	text	latin1_swedish_ci		Sim
10	options	text	latin1_swedish_ci		Sim
11	gatk_log	text	latin1_swedish_ci		Sim
12	picard_log	text	latin1_swedish_ci		Sim
13	annovar_log	text	latin1_swedish_ci		Sim
14	generic_log	text	latin1_swedish_ci		Sim
15	original_filename_one	text	latin1_swedish_ci		Sim
16	timestamp	timestamp		on update CURRENT_TIMESTAMP	Não

Figura 11: Demonstração de como estão sendo organizados os bancos de dados na página MySQL acessado localmente.

4.6. ANÁLISE DAS AMOSTRAS ILLUMINA: TESTANDO A EFETIVIDADE DA INTERFACE WEB

Assim como na primeira análise, visamos estudar os genes relacionados ao reparo de DNA na tentativa esclarecer o diagnóstico dos pacientes cujas amostras foram sequenciadas pela plataforma MiSeq da Illumina. As amostras, em sua maioria, apresentaram uma grande quantidade de *reads* em um mapeamento frente ao genoma de referência em torno de 98% (dado não mostrado). Na Tabela 12 listamos as variantes que acreditamos que possam ser as mutações responsáveis pelo diagnóstico clínico dos pacientes como XP, baseadas nas variantes demonstradas no Anexo 2, além das amostras que não tiveram um bom sequenciamento (em vermelho).

Como já mencionado, algumas amostras foram sequenciadas novamente (XP02RJ, XPSPAC06F0, XPSPAC11F0, XPSPAC31F0 e XP03AM/XP02AM) utilizando a plataforma MiSeq da Illumina com a finalidade de se confirmar mutações e entender o motivo de amostras, como a XP02RJ, não terem apresentado nenhuma das variantes já previamente relatadas. Através desse sequenciamento conseguimos encontrar as duas mutações já publicadas que não tinham sido encontradas na amostra do paciente XP02RJ, c.83C>A (p.Ala28Asp) e c.2904G>C (p.Trp968Cys). Além disso, conseguimos confirmar a presença da mutação dos pacientes XPSPAC11F0 e XPSPAC31F0. Já a amostra XP03AM, que é uma amostra de tecido

normal (não tumoral), também confirmou a presença da mutação em sua sequência, demonstrando que a mutação encontrada em XP02AM não era intrínseca do tumor, ou seja, não era uma mutação somática. Outra amostra que apresentou variantes, na qual anteriormente não havíamos detectado, foi a XPSPAC06F0. Foram encontradas 2 variantes no gene *ERCC2/XPD* em heterozigose. Então submetemos as variantes aos preditores de potencial deletério, SIFT e PROVEAN (Tabela 13). Através do resultado obtido por esses programas, que indicaram que as duas mutações podem ser deletérias ou danosas, caso essas mutações estejam em diferentes alelos, podem explicar as características clínicas desse paciente, sendo o diagnóstico mais provável como sendo um paciente XPD.

Prosseguindo para as amostras dos pacientes que não passaram pelo ressequenciamento, TTD02SP e XP01TO, apesar de possuírem uma boa cobertura de bases e várias variantes chamadas, não apresentaram nenhuma variante que pudessem nos auxiliar a inferir um possível gene candidato como causador da doença. Os pacientes que tiveram um dos seus parentes também sequenciados, o XP01BA e a XP01PE, além de apresentarem uma mutação candidata ao fenótipo XP, conseguimos verificar a origem de um desses alelos herdados de suas mães, BA02 e PE02, respectivamente, portadoras de um dos alelos mutados. XP01BA possui uma mutação em homozigose, não sinônima, no exon 5 do gene *XPA* gerando um códon de parada, provavelmente truncando a proteína tornando-a inativa ou não suficiente. Já a paciente XP01PE, possui 2 mutações em heterozigose composto no gene *XPC*, uma deleção (de 1 base) no exon 15 herdado de sua mãe e uma no exon 10 que gera um códon de parada, provavelmente herdada de seu pai. Os pacientes XP01RN e XP136GO, apesar de não terem nenhum parentesco, eles possuem a mesma mutação em homozigose formando um códon de parada no exon 3 do gene *POLH*, provavelmente sendo pacientes XPV. Por fim, na análise do sequenciamento da paciente XP03BA não foram encontradas variantes que justificassem seu caso clínico, com exceção do fato dela apresentar uma inserção (que provoca uma *frameshift* e, portanto, deve resultar em uma proteína truncada) em heterozigose no exon 6 do gene *POLH*. Nesse caso, possivelmente existe outra variante no outro alelo desse gene que não foi detectada por esse sequenciamento.

Tabela 12: Levantamento das possíveis variantes que possam justificar o diagnóstico clínico dos pacientes, muitas delas ainda não descritas ou descritas em banco de dados. Cinza: Amostra em que não obtivemos variantes candidatas. Vermelho – amostra de baixa qualidade. Laranja – variantes que supomos estar envolvidas com a doença, provavelmente pela possibilidade de ser uma mutação heterozigótica composta (XP03BA).

Amostra	CHROM	POSITION	RS_ID	REF	ALT	#REF	#ALT	#DEPH	GENE	REGION	GENOTYPE	VAR_TYPE	GENE_ID	EXON	CDNA POS	PROT_POS	#QUAL
XP01BA	chr9	100447232	.	G	A	3	133	137	XPA	exonic	hom	stopgain	NM_000380	exon5	c.C646T	p.Q216X	3486
XP136GO	chr6	43565580	.	C	G	0	341	342	POLH	exonic	hom	stopgain	NM_001291969	exon3	c.C266G	p.S89X	9399
TTD02SP																	
XP01RN	chr6	43565580	.	C	G	0	431	432	POLH	exonic	hom	stopgain	NM_001291969	exon3	c.C266G	p.S89X	11731
BA02	chr9	100447232	.	G	A	59	61	120	XPA	exonic	het	stopgain	NM_000380	exon5	c.C646T	p.Q216X	1266
XP01TO																	
XP01PE	chr3	14197899	.	C	A	187	190	377	XPC	exonic	het	stopgain	NM_004628	exon10	c.G1969T	p.E657X	3852
	chr3	14188829	.	CA	C	220	164	405	XPC	exonic	het	frameshift deletion	NM_004628	exon15	c.2564delT	p.L855fs	5334
PE02	chr3	14188829	.	CA	C	204	189	418	XPC	exonic	het	frameshift deletion	NM_004628	exon15	c.2564delT	p.L855fs	6287
XP03BA	chr6	43568736	.	A	AT	131	139	270	POLH	exonic	het	frameshift insertion	NM_001291970	exon6	c.672_673insT	p.K224fs	5454
	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?
XP02RJ	chr13	103525633	rs267607280	G	C	184	146	330	BIM-ERC5,ERC5	exonic	het	nonsynonymous	NM_00123	exon14	c.G2904C	p.W968C	2925
	chr13	103498699	rs267607281	C	A	204	218	422	ERC5	exonic	het	nonsynonymous	NM_00123	exon1	c.C83A	p.A28D	4449
XP03AM	chr11	47256943	.	C	T	0	343	343	DB2	exonic	hom	stopgain	NM_00107	exon7	c.C1003T	p.C335X	10118
XP05AM																	
XPSPA06F0	chr19	45855610	rs41556519	G	A	12	14	26	ERC2	exonic	het	nonsynonymous	NM_000400	exon22	c.C2047T	p.R883W	261
	chr19	45855507	rs144564120	G	C	8	9	17	ERC2	exonic	het	nonsynonymous	NM_000400	exon22	c.C2150G	p.A717G	164
XPSPAC1F0	chr6	43565513	.	A	C	0	26	27	POLH	exonic	hom	nonsynonymous	NM_001291970	exon5	c.A571C	p.T191P	696
XPSPAC31F0	chr6	43565576	.	T	C	0	37	37	POLH	exonic	hom	nonsynonymous	NM_001291970	exon5	c.T634C	p.C212R	950

Tabela 13: Resultados gerados pelas plataformas PROVEAN e SIFT para as potenciais variantes candidatas ao perfil clínico dos pacientes sequenciados pelo MiSeq Illumina.

INPUT	LENGTH	CODON_CHANGE	POS	RESIDUE REF	RESIDUE ALT	TYPE	SCORE	PREDICTION (cutoff=-2.5)	SCORE2	PREDICTION (cutoff=0.05)	MEDIAN INFO
9,100447232,G,A,XP01BA BA02	273	AAA [C/T]AG AAG	216	Q	*	Nonsense	NA	NA	NA	NA	NA
6,43565580,C,G,XP136GO	713	TGT T[C/G]A GCT	213	S	*	Nonsense	NA	NA	NA	NA	NA
3,14197899,C,A,XP01PE	940	TAT [G/T]AG GCC	657	E	*	Nonsense	NA	NA	NA	NA	NA
3,14188829,CA,C,XP01PE PE02	940					Frameshift	NA	NA	NA	NA	NA
6,43568736,A,AT,XP03BA	713					Frameshift	NA	NA	NA	NA	NA
11,47256943,C,T,XP03AM	427	TTC [C/T]AG CAC	335	Q	*	Nonsense	NA	NA	NA	NA	NA
19,45855610,G,A,XPSPAC06F0	760	AAG [C/T]GG TTT	683	R	W	Single AA Change	-7.31	Deleterious	0.000	Damaging	2.83
19,45855507,G,C,XPSPAC06F0	760	GTG G[C/G]C AAG	717	A	G	Single AA Change	-2.49	Neutral	0.007	Damaging	2.82
6,43565513,A,C,XPSPAC11F0	713	CTC [A/C]CC GTG	191	T	P	Single AA Change	-2.79	Deleterious	0.144	Tolerated	2.84
6,43565576,T,C,XPSPAC31F0	713	CAG [T/C]GTTCA	212	C	R	Single AA Change	-11.66	Deleterious	0.000	Damaging	2.84

5. CONCLUSÃO GERAL

Como demonstrado, os objetivos deste projeto foram alcançados. Conseguimos identificar as principais mutações candidatas da grande maioria das amostras e, devido ao sequenciamento posterior pela plataforma MiSeq Illumina, pudemos confirmar algumas dessas mutações e encontrar outras mutações não relatadas anteriormente.

Acreditamos que a realização deste trabalho está sendo um primeiro passo na análise bioinformática de busca de mutações em amostras de DNA de pacientes com as síndromes XP, CS e TTD através de um serviço de grande importância social para esses pacientes e familiares. Além disso, acreditamos que o desenvolvimento dessa interface *web* de busca de variantes poderá ser empregado na análise de mutações induzidas no DNA genômico em geral, seja por um tratamento específico em laboratório (como irradiação com luz UVA ou mesmo luz solar ambiental), seja na análise de um tecido tumoral de pacientes XP, por exemplo.

Podemos constatar que a interface *web* desenvolvida, a *Human Variants Finder Interface*, funciona como o esperado na análise de amostras provenientes de diferentes tipos de plataformas. Além disso, a nossa ferramenta *web* visa mimetizar o caráter poderoso do Galaxy, porém através de uma interface prática e de fácil uso. Essa ferramenta oferece um *pipeline* completo e conciso para análise de dados humanos baseado na experiência compartilhada de vários pesquisadores que trabalham com dados NGS, juntamente com um forte poder de manipulação dos

dados finais, através de filtragem específicas permitindo uma análise direcionada, resultando em uma apresentação clara e intuitiva. É nossa expectativa que o pesquisador que tenha relativamente pouca instrução em bioinformática possa ser beneficiado, identificando alterações que podem ser responsáveis por síndromes genéticas, sobretudo aquelas de vias específicas, como as relacionadas a mecanismos de reparo de DNA.

6. REFERÊNCIAS

- ANSORGE, W. J. Next-generation DNA sequencing techniques. *New biotechnology*, v. 25, n. 4, p. 195–203, abr. 2009.
- BERQUIST, B. R.; WILSON, D. M. Pathways for repairing and tolerating the spectrum of oxidative DNA lesions. *Cancer letters*, v. 327, n. 1-2, p. 61–72, 31 dez. 2012.
- BROUSTAS, C. G.; LIEBERMAN, H. B. DNA damage response genes and the development of cancer metastasis. *Radiation research*, v. 181, n. 2, p. 111–30, fev. 2014.
- BUDDEN, T.; BOWDEN, N. A. The Role of Altered Nucleotide Excision Repair and UVB-Induced DNA Damage in Melanomagenesis. *International journal of molecular sciences*, v. 14, n. 1, p. 1132–51, jan. 2013.
- CHOI, Y. *et al.* Predicting the functional effect of amino acid substitutions and indels. *PloS one*, v. 7, n. 10, p. e46688, jan. 2012.
- CLEAVER, J. E. Cancer in xeroderma pigmentosum and related disorders of DNA repair. *Nature reviews. Cancer*, v. 5, n. 7, p. 564–73, jul. 2005.
- COLELLA, S. *et al.* Identical mutations in the CSB gene associated with either Cockayne syndrome or the DeSanctis-cacchione variant of xeroderma pigmentosum. *Human molecular genetics*, v. 9, n. 8, p. 1171–5, 1 maio 2000.
- COSTA, R. M. A. *et al.* The eukaryotic nucleotide excision repair pathway. *Biochimie*, v. 85, n. 11, p. 1083–1099, nov. 2003.
- DE LIMA-BESSA, K. M. *et al.* CPDs and 6-4PPs play different roles in UV-induced cell death in normal and NER-deficient human cells. *DNA repair*, v. 7, n. 2, p. 303–12, 1 fev. 2008.
- DEPRISTO, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, v. 43, n. 5, p. 491–498, 2011.

- DIGIOVANNA, J. J.; KENNETH, M. D. AND; H. KRAEMER, M. D. Shining a light on Xeroderma Pigmentosum. *J Invest Dermatol.*, v. 132, n. 3, p. 785–796, 2012.
- GREDILLA, R.; GARM, C.; STEVNSNER, T. Nuclear and mitochondrial DNA repair in selected eukaryotic aging model systems. *Oxidative medicine and cellular longevity*, v. 2012, p. 282438, jan. 2012.
- GREINERT, R. *et al.* UVA-induced DNA double-strand breaks result from the repair of clustered oxidative DNA damages. *Nucleic acids research*, v. 40, n. 20, p. 10263–73, 1 nov. 2012.
- GRUBER, F. *et al.* Photocarcinogenesis-Molecular mechanisms. *Collegium antropologicum*, v. 31 Suppl 1, p. 101–6, jan. 2007.
- IKEHATA, H.; ONO, T. The Mechanisms of UV Mutagenesis. *Journal of Radiation Research*, v. 52, n. 2, p. 115–125, 2011.
- INUI, H. *et al.* Xeroderma Pigmentosum-Variant Patients from America, Europe, and Asia. *J Invest Dermatol*, v. 128, n. 8, p. 2055–2068, 2008.
- KENNEDY, R. D.; D'ANDREA, A. D. The Fanconi Anemia/BRCA pathway: new faces in the crowd. *Genes & development*, v. 19, n. 24, p. 2925–40, 15 dez. 2005.
- KUMAR, P.; HENIKOFF, S.; NG, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature protocols*, v. 4, n. 7, p. 1073–81, jan. 2009.
- LEIBELING, D.; LASPE, P.; EMMERT, S. Nucleotide excision repair and cancer. *Journal of molecular histology*, v. 37, n. 5-7, p. 225–38, set. 2006.
- LI, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, v. 25, n. 16, p. 2078–9, 15 ago. 2009.
- LI, H.; DURBIN, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, v. 25, n. 14, p. 1754–1760, 2009.
- LIU, L.; LEE, J.; ZHOU, P. Navigating the Nucleotide Excision Repair Threshold. *J Cell Physiol*, v. 224, n. 3, p. 585–589, 2011.
- LIU, X. *et al.* Variant callers for next-generation sequencing data: a comparison study. *PloS one*, v. 8, n. 9, p. e75619, jan. 2013.
- MARDIS, E. R. Next-generation DNA sequencing methods. *Annual review of genomics and human genetics*, v. 9, p. 387–402, jan. 2008.
- MCKENNA, A. *et al.* The Genome Analysis Toolkit : A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, v. 20, p. 1297–1303, 2010.

MCMILLAN, T. J. *et al.* Cellular effects of long wavelength UV light (UVA) in mammalian cells. *The Journal of pharmacy and pharmacology*, v. 60, n. 8, p. 969–76, ago. 2008.

MENCK, C. F. M.; MUNFORD, V. DNA repair diseases: What do they tell us about cancer and aging? *Genetics and Molecular Biology*, v. 37, p. 220–233, 2014.

METZKER, M. L. Sequencing technologies - the next generation. *Nature reviews. Genetics*, v. 11, n. 1, p. 31–46, jan. 2010.

MILLS, R. E. *et al.* An initial map of insertion and deletion (INDEL) variation in the human genome An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome research*, v. 16, p. 1182–1190, 2006.

MORAES, M. C. S.; CABRAL NETO, J. B.; MENCK, C. F. M. DNA repair mechanisms protect our genome from carcinogenesis. *Frontiers in Bioscience*, v. 17, n. 1, p. 1362, 2012.

NOUSPIKEL, T. DNA repair in mammalian cells : Nucleotide excision repair: variations on versatility. *Cellular and molecular life sciences : CMLS*, v. 66, n. 6, p. 994–1009, mar. 2009.

RIEBER, N. *et al.* Coverage bias and sensitivity of variant calling for four whole-genome sequencing technologies. *PloS one*, v. 8, n. 6, p. e66621, jan. 2013.

SCHUCH, A. P. *et al.* DNA damage as a biological sensor for environmental sunlight. *Photochemical & photobiological sciences : Official journal of the European Photochemistry Association and the European Society for Photobiology*, v. 12, n. 8, p. 1259–1272, 2013.

SCHUCH, A. P.; MENCK, C. F. M. The genotoxic effects of DNA lesions induced by artificial UV-radiation and sunlight. *Journal of photochemistry and photobiology. B, Biology*, v. 99, n. 3, p. 111–6, 1 jun. 2010.

SHENDURE, J.; JI, H. Next-generation DNA sequencing. *Nature biotechnology*, v. 26, n. 10, p. 1135–1145, 2008.

SHERRY, S. T.; WARD, M.; SIROTKIN, K. dbSNP -- Database for Single Nucleotide Polymorphisms and Other Classes of Minor Genetic Variation dbSNP — Database for Single Nucleotide Polymorphisms and Other Classes of Minor Genetic Variation. *Genome research*, v. 9, p. 677–679, 1999.

SOLTYS, D. T. *et al.* Novel XPG (ERCC5) Mutations Affect DNA Repair and Cell Survival after Ultraviolet but not Oxidative Stress. *Human mutation*, v. 34, n. 3, p. 481–9, mar. 2013.

STEFANINI, M. *et al.* Trichothiodystrophy: from basic mechanisms to clinical implications. *DNA repair*, v. 9, n. 1, p. 2–10, 2 jan. 2010.

SUGASAWA, K. Multiple DNA damage recognition factors involved in mammalian nucleotide excision repair. *Biochemistry (Moscow)*, v. 76, n. 1, p. 16–23, 16 fev. 2011.

SUGASAWA, K. Xeroderma pigmentosum genes: functions inside and outside DNA repair. *Carcinogenesis*, v. 29, n. 3, p. 455–65, mar. 2008.

TENG, S.; MICHONNOVA-ALEXOVA, E.; ALEXOV, E. Approaches and resources for prediction of the effects of non-synonymous single nucleotide polymorphism on protein function and interactions. *Current pharmaceutical biotechnology*, v. 9, n. 2, p. 123–33, abr. 2008.

THORVALDSDÓTTIR, H.; ROBINSON, J. T.; MESIROV, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics*, v. 14, n. 2, p. 178–92, mar. 2013.

TOTONCHY, M. B. *et al.* Auditory analysis of xeroderma pigmentosum 1971-2012: hearing function, sun sensitivity and DNA repair predict neurological degeneration. *Brain : a journal of neurology*, v. 136, n. Pt 1, p. 194–208, jan. 2013.

VOELKERDING, K. V; DAMES, S. A; DURTSCHI, J. D. Next-generation sequencing: from basic research to diagnostics. *Clinical chemistry*, v. 55, n. 4, p. 641–58, abr. 2009.

WANG, K.; LI, M.; HAKONARSON, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research*, v. 38, n. 16, p. e164, set. 2010.

WEINER, A. M.; GRAY, L. T. What role (if any) does the highly conserved CSB-PGBD3 fusion protein play in Cockayne syndrome? *Mech Ageing Dev.*, v. 134, n. 0, p. 225–233, 2014.

WELSH, M. M. *et al.* Genetic determinants of UV-susceptibility in non-melanoma skin cancer. *PloS one*, v. 6, n. 7, p. e20019, jan. 2011.

XIE, M. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature*, v. 502, n. 7471, p. 333–339, 2014.

YU, X.; SUN, S. Comparing a few SNP calling algorithms using low-coverage sequencing data. *BMC bioinformatics*, v. 14, n. 1, p. 274, jan. 2013.

ZHOU, X. *et al.* The next-generation sequencing technology and application. *Protein & cell*, v. 1, n. 6, p. 520–36, jun. 2010.

Anexos

Anexo 1

Aqui estão apresentadas as variantes filtradas de todas as amostras sequenciados por SOLiD que não estão no banco de dados do 1000genomes, após remoção das variantes intrônicas, com exceção dos sítios de *splicing*. A seleção em amarelo mostra a variante mais provável de estar relacionada à patologia ou a condição de portador da mutação, enquanto a não seleção indica que as variantes não são candidatas a explicar o quadro do paciente. Já as seleções em laranja indicam variantes que supomos estar relacionada com a doença, mas estão presentes em heterozigose (não acompanhadas de segunda mutação do mesmo gene, como esperado para doença recessiva).

a- Amostra XP02RJ

CHROM	POSITION	RS_ID	REF	ALT	#REF	#ALT	GENE	REGION	GENOTYPE	VAR_TYPE	GENE_ID	EXON	CDNA_POS	PROT_POS	#QUAL
chr16	89804335	rs17233826	CT	C	8	3	FANCA	UTR3	het		NM_000135	-	c.*673delA		44
chr3	10089723	rs12330369	G	A	17	9	FANCD2	exonic	het	synonymous SNV	NM_033084	exon16	c.G1401A	p.T467T	94
chr3	10108898	rs77246387	A	G	5	7	FANCD2	exonic	het	synonymous SNV	NM_033084	exon26	c.A2391G	p.V797V	110
chr3	10108913	rs80258959	G	T	4	4	FANCD2	exonic	het	nonsynonymous	NM_033084	exon26	c.G2406T	p.Q802H	49

b- Amostra XPSPAC02F0

CHROM	POSITION	RS_ID	REF	ALT	#REF	#ALT	#DEPH	GENE	REGION	GENOTYPE	VAR_TYPE	GENE_ID	EXON	CDNA_POS	PROT_POS	#QUAL
chr3	10089723	rs12330369	G	A	17	8	25	FANCD2	exonic	het	synonymous SNV	NM_033084	exon16	c.G1401A	p.T467T	34
chr16	89804053	.	TTAA	T	10	3	14	FANCA	UTR3	het		NM_000135	-	c.*955_*953delT		59
chr16	89804335	rs17233826	CT	C	10	3	13	FANCA	UTR3	het		NM_000135	-	c.*673delA		42
chr9	100459731	rs528384632	C	A	1	3	4	XPA	upstream	het		-	-	-		38
chr3	10108898	rs77246387	A	G	5	3	8	FANCD2	exonic	het	synonymous SNV	NM_033084	exon26	c.A2391G	p.V797V	18
chr3	10088308	rs72492998	T	C	3	2	5	FANCD2	exonic	het	synonymous SNV	NM_033084	exon15	c.T1179C	p.T393T	62
chr3	10088299	rs112887807	C	T	5	4	9	FANCD2	exonic	het	synonymous SNV	NM_033084	exon15	c.C1170T	p.S390S	81
chr3	10088343	rs73126218	A	G	0	2	2	FANCD2	exonic	hom	nonsynonymous S	NM_033084	exon15	c.A1214G	p.N405S	23
chr6	43565513	.	A	C	0	9	9	POLH	exonic	hom	nonsynonymous S	NM_006502	exon5	c.A571C	p.T191P	176

c- Amostra GO01

CHROM	POSITION	RS_ID	REF	ALT	#REF	#ALT	#DEPH	GENE	REGION	GENOTYPE	VAR_TYPE	GENE_ID	EXON	CDNA_POS	PROT_POS	#QUAL
chr17	7577539	rs121912651	G	A	16	11	27	TP53	exonic	het	nonsynonymous S	NM_001276699	exon3	c.C265T	p.R89W	165
chr9	35076975	rs372234656	C	T	11	7	18	FANCG	exonic	het	nonsynonymous S	NM_004629	exon6	c.G770A	p.R257H	101
chr6	43568829	.	G	A	16	7	23	POLH	splicing	het		NM_006502	exon6	c.764+1G>A		50
chr3	10108898	rs77246387	A	G	12	6	18	FANCD2	exonic	het	synonymous SNV	NM_033084	exon26	c.A2391G	p.V797V	47
chr3	10089723	rs12330369	G	A	31	12	43	FANCD2	exonic	het	synonymous SNV	NM_033084	exon16	c.G1401A	p.T467T	169

d- Amostra GO03

CHROM	POSITION	RS_ID	REF	ALT	#REF	#ALT	#DEPH	GENE	REGION	GENOTYPE	VAR_TYPE	GENE_ID	EXON	CDNA_POS	PROT_POS	#QUAL
chr6	43568829	.	G	A	0	2	2	POLH	splicing	hom		NM_006502	exon6	c.764+1G>A		25

e- Amostra GO04

CHROM	POSITION	RS_ID	REF	ALT	#REF	#ALT	#DEPH	GENE	REGION	GENOTYPE	VAR_TYPE	GENE_ID	EXON	CDNA_POS	PROT_POS	#QUAL
chr19	45853665	.	A	C	10	10	20	KLC3	exonic	het	nonsynonymous S	NM_177417	exon9	c.A1210C	p.K404Q	135
chr3	10108913	rs80258959	G	T	6	4	10	FANCD2	exonic	het	nonsynonymous S	NM_001018115	exon26	c.G2406T	p.Q802H	50
chr3	10088299	rs112887807	C	T	4	4	8	FANCD2	exonic	het	synonymous SNV	NM_001018115	exon15	c.C1170T	p.S390S	114
chr3	10088266	rs72492997	G	T	7	4	11	FANCD2	exonic	het	synonymous SNV	NM_001018115	exon15	c.G1137T	p.V379V	43
chr3	10108898	rs77246387	A	G	10	7	17	FANCD2	exonic	het	synonymous SNV	NM_001018115	exon26	c.A2391G	p.V797V	82
chr3	10088308	rs72492998	T	C	2	3	5	FANCD2	exonic	het	synonymous SNV	NM_001018115	exon15	c.T1179C	p.T393T	117
chr6	43568829	.	G	A	7	12	19	POLH	splicing	het		NM_006502	exon6	c.764+1G>A		173
chr3	10089723	rs12330369	G	A	19	15	36	FANCD2	exonic	het	synonymous SNV	NM_001018115	exon16	c.G1401A	p.T467T	226

f- Amostra GO05

CHROM	POSITION	RS_ID	REF	ALT	#REF	#ALT	#DEPH	GENE	REGION	GENOTYPE	VAR_TYPE	GENE_ID	EXON	CDNA_POS	PROT_POS	#QUAL
chr6	43568829	.	G	A	17	9	26	POLH	splicing	het		NM_006502	exon6	c.764+1G>A		123
chr3	10088343	rs73126218	A	G	0	2	2	FANCD2	exonic	hom	nonsynonymous S	NM_001018115	exon15	c.A1214G	p.N405S	21
chr13	32910426	.	G	GGGGA	15	4	20	BRCA2	exonic	het	frameshift inse	NM_000059	exon11	c.1934_1935insG	p.R645fs	87
chr3	10108898	rs77246387	A	G	36	16	52	FANCD2	exonic	het	synonymous SNV	NM_001018115	exon26	c.A2391G	p.V797V	159
chr3	10089723	rs12330369	G	A	33	7	41	FANCD2	exonic	het	synonymous SNV	NM_001018115	exon16	c.G1401A	p.T467T	39
chr16	14026041	.	C	T	16	15	31	ERCC4	exonic	het	nonsynonymous S	NM_005236	exon6	c.C1001T	p.S334L	238
chr13	32950873	.	A	C	22	5	27	BRCA2	exonic	het	nonsynonymous S	NM_000059	exon21	c.A8699C	p.D2900A	91
chr3	10088308	rs72492998	T	C	13	3	16	FANCD2	exonic	het	synonymous SNV	NM_001018115	exon15	c.T1179C	p.T393T	79
chr13	32910421	.	GAAAA	G	15	4	20	BRCA2	exonic	het	frameshift dele	NM_000059	exon11	c.1930_1933del	p.K644fs	87
chr16	14013450	.	GAAA	G	8	3	11	ERCC4	upstream	het		-	-	-		65
chr3	10108913	rs80258959	G	T	28	8	36	FANCD2	exonic	het	nonsynonymous S	NM_001018115	exon26	c.G2406T	p.Q802H	72
chr3	10088299	rs112887807	C	T	16	4	20	FANCD2	exonic	het	synonymous SNV	NM_001018115	exon15	c.C1170T	p.S390S	70
chr16	14013455	.	A	AG	6	3	9	ERCC4	upstream	het		-	-	-		16
chr13	32950870	.	AAG	A	24	5	31	BRCA2	exonic	het	frameshift dele	NM_000059	exon21	c.8697_8698del	p.Q2899fs	88
chr13	32950875	.	G	GT	25	5	30	BRCA2	exonic	het	frameshift inse	NM_000059	exon21	c.8701_8702insT	p.G2901fs	88
chr16	14013453	.	A	AGG	8	3	11	ERCC4	upstream	het		-	-	-		16
chr13	32950874	.	T	TC	24	5	29	BRCA2	exonic	het	frameshift inse	NM_000059	exon21	c.8700_8701insC	p.D2900fs	85

g- Amostra G006

CHROM	POSITION	RS_ID	REF	ALT	#REF	#ALT	#DEPH	GENE	REGION	GENOTYPE	VAR_TYPE	GENE_ID	EXON	CDNA_POS	PROT_POS	#QUAL
chr16	14026041	.	C	T	21	15	36	ERCC4	exonic	het	nonsynonymous S	NM_005236	exon6	c.C1001T	p.S334L	231
chr3	10088299	rs112887807	C	T	17	14	31	FANCD2	exonic	het	synonymous SNV	NM_001018115	exon15	c.C1170T	p.S390S	409
chr3	10088266	rs72492997	G	T	19	14	33	FANCD2	exonic	het	synonymous SNV	NM_001018115	exon15	c.G1137T	p.V379V	244
chr3	10089723	rs12330369	G	A	41	19	61	FANCD2	exonic	het	synonymous SNV	NM_001018115	exon16	c.G1401A	p.T467T	266
chr3	10108913	rs80258959	G	T	18	10	29	FANCD2	exonic	het	nonsynonymous S	NM_001018115	exon26	c.G2406T	p.Q802H	139
chr16	89804335	rs17233826	CT	C	7	6	17	FANCA	UTR3	het		NM_000135	-	c.*673delA		155
chr6	43568829	.	G	A	3	33	36	POLH	splicing	hom		NM_006502	exon6	c.764+1G>A		737
chr3	10066649	.	TGGA	T	5	3	9	CIDECP	ncRNA_intronic	het		-	-	-		74
chr3	10066653	.	C	CA	5	3	9	CIDECP	ncRNA_intronic	het		-	-	-		74
chr3	10108898	rs77246387	A	G	22	17	39	FANCD2	exonic	het	synonymous SNV	NM_001018115	exon26	c.A2391G	p.V797V	261
chr3	10066648	.	C	CGT	5	3	9	CIDECP	ncRNA_intronic	het		-	-	-		74
chr3	10088308	rs72492998	T	C	12	9	21	FANCD2	exonic	het	synonymous SNV	NM_001018115	exon15	c.T1179C	p.T393T	345
chr9	35076975	rs372234656	C	T	27	13	40	FANCG	exonic	het	nonsynonymous S	NM_004629	exon6	c.G770A	p.R257H	152
chr10	50724711	.	A	C	28	7	36	ERCC6-PGBD3,PGBD3	exonic	het	synonymous SNV	NM_001277058	exon6	c.T1854G	p.L618L	39
chr10	50724710	.	C	A	28	7	36	ERCC6-PGBD3,PGBD3	exonic	het	nonsynonymous S	NM_001277058	exon6	c.G1855T	p.D619Y	39

h- Amostra XPSPAC01F0

CHROM	POSITION	RS_ID	REF	ALT	#REF	#ALT	#DEPH	GENE	REGION	GENOTYPE	VAR_TYPE	GENE_ID	EXON	CDNA_POS	PROT_POS	#QUAL
chr3	10088308	rs72492998	T	C	9	3	12	FANCD2	exonic	het	synonymous SNV	NM_001018115	exon15	c.T1179C	p.T393T	55
chr9	97862127	.	G	A	11	9	20	FANCC	UTR3	het		NM_001243743	-	c.*1862C>T		135
chr19	45911319	rs11314106	CA	C	1	6	9	ERCC1	UTR3	het		NM_001166049	-	c.*1613delT		116
chr3	10088299	rs112887807	C	T	11	3	14	FANCD2	exonic	het	synonymous SNV	NM_001018115	exon15	c.C1170T	p.S390S	52
chr3	14207005	rs370445216	C	T	11	12	23	XPC	exonic	het	synonymous SNV	NM_004628	exon6	c.G702A	p.L234L	210
chr6	43572429	.	G	T	0	31	31	POLH	exonic	hom	nonsynonymous S	NM_001291969	exon6	c.G590T	p.C197F	765
chr3	10089723	rs12330369	G	A	31	8	39	FANCD2	exonic	het	synonymous SNV	NM_001018115	exon16	c.G1401A	p.T467T	45
chr3	10108898	rs77246387	A	G	18	11	29	FANCD2	exonic	het	synonymous SNV	NM_001018115	exon26	c.A2391G	p.V797V	107
chr16	89804335	rs17233826	CT	C	16	7	27	FANCA	UTR3	het		NM_000135	-	c.*673delA		146
chr3	10108913	rs80258959	G	T	13	6	19	FANCD2	exonic	het	nonsynonymous S	NM_001018115	exon26	c.G2406T	p.Q802H	73

i- Amostra XPSPAC04F0

CHROM	POSITION	RS_ID	REF	ALT	#REF	#ALT	#DEPH	GENE	REGION	GENOTYPE	VAR_TYPE	GENE_ID	EXON	CDNA_POS	PROT_POS	#QUAL
chr6	43578436	.	ATACT	A	28	12	43	POLH	exonic	het	frameshift dele	NM_006502	exon10	c.1221_1224del	p.N407fs	383
chr3	10088343	rs73126218	A	G	0	4	4	FANCD2	exonic	hom	nonsynonymous S	NM_001018115	exon15	c.A1214G	p.N405S	60
chr16	14043455	rs532485638	T	C	26	22	49	ERCC4	UTR3	het		NM_005236	-	c.*1251T>C		177
chr9	100459849	.	TA	T	38	6	52	XPA	upstream	het		-	-	-		98
chr13	32910426	.	G	GGGGA	11	4	17	BRCA2	exonic	het	frameshift inse	NM_000059	exon11	c.1934_1935insG	p.R645fs	81
chr17	41196753	.	GAACAC	G	32	13	47	BRCA1	UTR3	het		NM_007300	-	c.*941_937delG		392
chr9	100459857	.	G	GTCC	39	6	46	XPA	upstream	het		-	-	-		95
chr3	10108913	rs80258959	G	T	25	13	38	FANCD2	exonic	het	nonsynonymous S	NM_001018115	exon26	c.G2406T	p.Q802H	177
chr3	10088308	rs72492998	T	C	11	5	16	FANCD2	exonic	het	synonymous SNV	NM_001018115	exon15	c.T1179C	p.T393T	100
chr17	41196758	.	C	CGGTGT	31	13	46	BRCA1	UTR3	het		NM_007300	-	c.*936_937insA		392
chr3	10088299	rs112887807	C	T	18	3	21	FANCD2	exonic	het	synonymous SNV	NM_001018115	exon15	c.C1170T	p.S390S	61
chr3	10108898	rs77246387	A	G	35	17	52	FANCD2	exonic	het	synonymous SNV	NM_001018115	exon26	c.A2391G	p.V797V	126
chr13	32910421	.	GAAAA	G	12	4	20	BRCA2	exonic	het	frameshift dele	NM_000059	exon11	c.1930_1933del	p.K644fs	81
chr17	41243948	rs56214134	C	A	16	15	31	BRCA1	exonic	het	nonsynonymous S	NM_007300	exon10	c.G3600T	p.Q1200H	203
chr9	100459855	.	AGG	A	40	6	47	XPA	upstream	het		-	-	-		95

j- Amostra XPSPAC06F0 – nenhuma variante encontrada após os filtros.

k- Amostra XPSPAC08F0

CHROM	POSITION	RS_ID	REF	ALT	#REF	#ALT	#DEPH	GENE	REGION	GENOTYPE	VAR_TYPE	GENE_ID	EXON	CDNA_POS	PROT_POS	#QUAL
chr13	103528224	.	C	T	18	17	36	BIVM-ERCC5,ERCC5	exonic	het	nonsynonymous S	NM_000123	exon15	c.C3532T	p.R1178C	247
chr13	32918767	.	AGGCTT	A	16	5	22	BRCA2	exonic	het	frameshift dele	NM_000059	exon12	c.6915_6919del	p.K2305fs	76
chr9	100459849	.	TA	T	24	6	32	XPA	upstream	het		-	-	-		133
chr6	43565513	.	A	C	1	27	28	POLH	exonic	hom	nonsynonymous S	NM_001291969	exon3	c.A199C	p.T67P	558
chr3	10088299	rs112887807	C	T	15	6	21	FANCD2	exonic	het	synonymous SNV	NM_001018115	exon15	c.C1170T	p.S390S	170
chr3	10088308	rs72492998	T	C	12	5	17	FANCD2	exonic	het	synonymous SNV	NM_001018115	exon15	c.T1179C	p.T393T	126
chr13	32918772	.	T	TAATCC	15	5	21	BRCA2	exonic	het	stopgain	NM_000059	exon12	c.6919_6920insA	p.S2307_K2308de	79
chr10	50661816	rs397827318	GT	G,GT	0	9	29	ERCC6	downstream	het		-	-	-		309
chr3	10108898	rs77246387	A	G	14	7	21	FANCD2	exonic	het	synonymous SNV	NM_001018115	exon26	c.A2391G	p.V797V	22
chr9	100459857	.	G	GTCC	27	6	34	XPA	upstream	het		-	-	-		108
chr3	10108913	rs80258959	G	T	9	5	14	FANCD2	exonic	het	nonsynonymous S	NM_001018115	exon26	c.G2406T	p.Q802H	53
chr9	100459855	.	AGG	A	27	6	35	XPA	upstream	het		-	-	-		127
chr3	10088343	rs73126218	A	G	0	2	2	FANCD2	exonic	hom	nonsynonymous S	NM_001018115	exon15	c.A1214G	p.N405S	23
chr19	45912489	rs35729377	CAAG	C	14	11	30	CD3EAP	exonic	het	nonframeshift d	NM_012099	exon3	c.1264_1266del	p.422_422del	383

l- Amostra XPSPAC11F0

CHROM	POSITION	RS_ID	REF	ALT	#REF	#ALT	#DEPH	GENE	REGION	GENOTYPE	VAR_TYPE	GENE_ID	EXON	CDNA_POS	PROT_POS	#QUAL
chr6	43565513	.	A	C	0	2	2	POLH	exonic	hom	nsynonymd	NM_006502	exon5	c.A571C	p.T191P	21,77

m- Amostra XPSPAC12F1

CHROM	POSITION	RS_ID	REF	ALT	#REF	#ALT	#DEPH	GENE	REGION	GENOTYPE	VAR_TYPE	GENE_ID	EXON	CDNA_POS	PROT_POS	#QUAL
chr16	89877386	rs139160837	G	A	17	16	33	FANCA	exonic	het	nonsynonymous S	NM_001018112	exon4	c.C377T	p.T126M	267
chr3	10108913	rs80258959	G	T	12	4	16	FANCD2	exonic	het	nonsynonymous S	NM_001018115	exon26	c.G2406T	p.Q802H	29
chr3	10088308	rs72492998	T	C	8	5	13	FANCD2	exonic	het	synonymous SNV	NM_001018115	exon15	c.T1179C	p.T393T	125
chr16	89804335	rs17233826	CT	C	19	6	25	FANCA	UTR3	het		NM_000135	-	c.*673delA		116
chr3	10108898	rs77246387	A	G	16	8	24	FANCD2	exonic	het	synonymous SNV	NM_001018115	exon26	c.A2391G	p.V797V	46
chr3	10088299	rs112887807	C	T	13	5	18	FANCD2	exonic	het	synonymous SNV	NM_001018115	exon15	c.C1170T	p.S390S	97
chr13	103492647	.	A	AT	12	6	19	BIVM	UTR3	het		NM_017693	-	c.*432_*433insT		98
chr3	10088343	rs73126218	A	G	0	4	4	FANCD2	exonic	hom	nonsynonymous S	NM_001018115	exon15	c.A1214G	p.N405S	61
chr3	10089723	rs12330369	G	A	32	17	49	FANCD2	exonic	het	synonymous SNV	NM_001018115	exon16	c.G1401A	p.T467T	233

m2-

CHROM	POSITION	RS_ID	REF	ALT	#REF	#ALT	#DEPH	GENE	REGION	GENOTYPE	VAR_TYPE	GENE_ID	EXON	CDNA_POS	PROT_POS	#QUAL
chr10	50723780	rs142477138	C	T	25	11	36	ERCC6-PGBD3,PGBD3	exonic	het	nonsynonymous S	NM_170753	exon2	c.G1381A	p.G461R	145
chr16	89877386	rs139160837	G	A	17	16	33	FANCA	exonic	het	nonsynonymous S	NM_001018112	exon4	c.C377T	p.T126M	267
chr16	89849480	rs2239359	C	T	22	22	44	FANCA	exonic	het	nonsynonymous S	NM_001286167	exon16	c.G1501A	p.G501S	365
chr17	41223094	rs1799966	T	C	11	16	27	BRCA1	exonic	het	nonsynonymous S	NM_007297	exon14	c.A4696G	p.S1566G	225
chr16	89815152	rs17233497	G	A	9	4	13	FANCA	exonic	het	nonsynonymous S	NM_001286167	exon33	c.C3263T	p.S1088F	44
chr3	10108913	rs80258959	G	T	12	4	16	FANCD2	exonic	het	nonsynonymous S	NM_001018115	exon26	c.G2406T	p.Q802H	29
chr10	50680422	rs145720191	C	T	13	7	20	ERCC6	exonic	het	nonsynonymous S	NM_000124	exon16	c.G2924A	p.R975Q	101
chr9	35079502	rs35984312	G	A	8	5	13	FANCG	exonic	het	nonsynonymous S	NM_004629	exon1	c.C20T	p.S7F	64
chr16	89836323	rs7195066	C	T	5	9	14	FANCA	exonic	het	nonsynonymous S	NM_001286167	exon26	c.G2426A	p.G809D	148
chr16	89839766	rs17232910	G	C	17	9	27	FANCA	exonic	het	nonsynonymous S	NM_001286167	exon22	c.C1927G	p.P643A	115
chr10	50724016	rs4253072	C	T	0	19	19	ERCC6-PGBD3,PGBD3	exonic	hom	nonsynonymous S	NM_170753	exon2	c.G1145A	p.R382K	457
chr17	41245471	rs4986850	C	T	32	26	58	BRCA1	exonic	het	nonsynonymous S	NM_007297	exon9	c.G1936A	p.D646N	748
chr13	32929387	rs169547	T	C	0	21	21	BRCA2	exonic	hom	nonsynonymous S	NM_000059	exon14	c.T7397C	p.V2466A	413
chr17	7579472	rs1042522	G	C	0	5	6	TP53	exonic	hom	nonsynonymous S	NM_001126114	exon4	c.C215G	p.P72R	60
chr17	41244000	rs16942	T	C	18	16	34	BRCA1	exonic	het	nonsynonymous S	NM_007297	exon9	c.A3407G	p.K1136R	212
chr16	89866043	rs7190823	T	C	0	13	13	FANCA	exonic	hom	nonsynonymous S	NM_001018112	exon9	c.A796G	p.T266A	167
chr17	41244936	rs799917	G	A	17	19	36	BRCA1	exonic	het	nonsynonymous S	NM_007297	exon9	c.C2471T	p.P824L	265
chr17	41244435	rs16941	T	C	23	14	37	BRCA1	exonic	het	nonsynonymous S	NM_007297	exon9	c.A2972G	p.E991G	143
chr3	10088343	rs73126218	A	G	0	4	4	FANCD2	exonic	hom	nonsynonymous S	NM_001018115	exon15	c.A1214G	p.N405S	61

n- Amostra XPSPAC13F0

CHROM	POSITION	RS_ID	REF	ALT	#REF	#ALT	#DEPH	GENE	REGION	GENOTYPE	VAR_TYPE	GENE_ID	EXON	CDNA_POS	PROT_POS	#QUAL
chr5	60169506	.	G	A	9	3	12	ERCC8	downstream	het		-	-	-		45
chr3	14190232	.	C	G	0	4	4	XPC	splicing	hom		NM_004628	exon14	c.2251-1G>C		93
chr3	10088299	rs112887807	C	T	4	4	8	FANCD2	exonic	het	synonymous SNV	NM_001018115	exon15	c.C1170T	p.S390S	70
chr16	89804335	rs17233826	CT	C	18	5	24	FANCA	UTR3	het		NM_000135	-	c.*673delA		83
chr5	60169505	.	G	A	9	3	12	ERCC8	downstream	het		-	-	-		45
chr3	10088308	rs72492998	T	C	3	3	6	FANCD2	exonic	het	synonymous SNV	NM_001018115	exon15	c.T1179C	p.T393T	74
chr3	10089723	rs12330369	G	A	18	11	30	FANCD2	exonic	het	synonymous SNV	NM_001018115	exon16	c.G1401A	p.T467T	180
chr3	10088266	rs72492997	G	T	8	6	14	FANCD2	exonic	het	synonymous SNV	NM_001018115	exon15	c.G1137T	p.V379V	53
chr9	100451855	.	A	ATGCGG	21	3	24	XPA	exonic	het	frameshift inse	NM_000380	exon3	c.349_350insCCG	p.L117fs	16
chr3	10088343	rs73126218	A	G	0	2	2	FANCD2	exonic	hom	nonsynonymous S	NM_001018115	exon15	c.A1214G	p.N405S	24

o- Amostra XPSPAC31F0

CHROM	POSITION	RS_ID	REF	ALT	#REF	#ALT	#DEPH	GENE	REGION	GENOTYPE	VAR_TYPE	GENE_ID	EXON	CDNA_POS	PROT_POS	#QUAL
chr10	50680471	rs190863815	C	A	13	9	22	ERCC6	exonic	het	nonsynonymous S	NM_000124	exon16	c.G2875T	p.V959L	132
chr19	45912489	rs35729377	CAAG	C	15	5	25	CD3EAP	exonic	het	nonframeshift d	NM_001297590	exon3	c.1270_1272del	p.424_424del	128
chr19	45911319	rs11314106	CA	C	8	6	14	ERCC1	UTR3	het		NM_001166049	-	c.*1613delT		79
chr6	43565576	.	T	C	2	18	20	POLH	exonic	hom	nonsynonymous S	NM_006502	exon5	c.T634C	p.C212R	165

Anexo 2

Análise das amostras sequenciadas pela plataforma MiSeq da Illumina com os mesmos critérios utilizados na análise anterior. A seleção em amarelo mostra a variante mais provável de estar relacionado à patologia ou condição de portador da mutação do paciente, enquanto a não seleção demonstra não ter nenhuma variável candidata filtrada.

a- Amostra XP01BA

CHROM	POSITION	RS_ID	REF	ALT	#REF	#ALT	#DEPH	GENE	REGION	GENOTYPE	VAR_TYPE	GENE_ID	EXON	CDNA_POS	PROT_POS	#QUAL
chr3	10085536	rs34046352	A	G	55	119	174	FANCD2	exonic	het	synonymous	NM_001018115	exon14	c.A1122G	p.V374V	2823
chr3	10088308	rs72492998	T	C	176	31	208	FANCD2	exonic	het	synonymous	NM_001018115	exon15	c.T1179C	p.T393T	659
chr3	10089723	rs12330369	G	A	214	105	320	FANCD2	exonic	het	synonymous	NM_001018115	exon16	c.G1401A	p.T467T	3343
chr3	10106532	rs3864017	C	T	172	64	236	FANCD2	exonic	het	nonsynonymous	NM_001018115	exon23	c.C2141T	p.P714L	1223
chr3	10088404	.	C	T	103	23	126	FANCD2	exonic	het	synonymous	NM_001018115	exon15	c.C1275T	p.Y425Y	595
chr3	10143061	.	CT	C	2	18	37	FANCD2	UTR3	hom	-	NM_001018115	-	c.*116delT	-	251
chr17	7578711	.	CTTTT	C,CT	0	17	91	TP53	UTR5	het	-	NM_001126115	-	c.-180_-182delA	-	2399
chr3	10088299	rs112887807	C	T	179	27	207	FANCD2	exonic	het	synonymous	NM_001018115	exon15	c.C1170T	p.S390S	616
chr3	10088407	.	AG	A	93	19	113	FANCD2	splicing	het	-	NM_033084	exon15	c.1278+1G>	-	151
chr3	10105516	rs142354499	A	C	221	86	307	FANCD2	exonic	het	nonsynonymous	NM_001018115	exon21	c.A1868C	p.Q623P	1472
chr19	45912489	rs35729377	CAAG	C	117	130	279	CD3EAP	exonic	het	nonframeshift deleti	NM_001297590	exon3	c.1270_1272del	p.A24_424del	4924
chr3	14186830	rs2470352	A	G	109	97	206	XPC	UTR3	het	-	NM_004628	-	c.*611T>C	-	2185
chr9	100447232	.	G	A	3	133	137	XPA	exonic	hom	stopgain	NM_000380	exon5	c.C646T	p.Q216X	3486
chr17	7572154	.	GAA	G,GA	2	22	49	TP53	UTR3	het	-	NM_001276761	-	c.*771delT	-	834
chr3	10088343	rs73126218	A	G	169	44	213	FANCD2	exonic	het	nonsynonymous	NM_001018115	exon15	c.A1214G	p.N405S	606

b- Amostra XP136GO

CHROM	POSITION	RS_ID	REF	ALT	#REF	#ALT	#DEPH	GENE	REGION	GENOTYPE	VAR_TYPE	GENE_ID	EXON	CDNA_POS	PROT_POS	#QUAL
chr6	43587101	.	CAAA	C	0	7	11	POLH	UTR3	hom	-	NM_006502	-	c.*4808_*4810de	-	190
chr6	43565580	.	C	G	0	341	342	POLH	exonic	hom	stopgain	NM_001291969	exon3	c.C266G	p.S89X	9399
chr17	7578711	.	CTTTT	C,CT	1	41	130	TP53	UTR5	het	-	NM_001126115	-	c.-180_-182delA	-	3853
chr17	7572154	rs200757381	GA	G	9	33	46	TP53	UTR3	het	-	NM_001276761	-	c.*772delT	-	488
chr3	10088404	.	C	T	103	52	155	FANCD2	exonic	het	synonymous	NM_033084	exon15	c.C1275T	p.Y425Y	1721
chr6	158613937	.	CAAAAA	C	8	4	15	GTF2H5	UTR3	het	-	NM_207118	-	c.*749_*753delA	-	67
chr16	89804335	rs17233826	CT	C	142	134	277	FANCA	UTR3	het	-	NM_000135	-	c.*673delA	-	4118
chr3	10089723	rs12330369	G	A	237	173	410	FANCD2	exonic	het	synonymous	NM_033084	exon16	c.G1401A	p.T467T	5875
chr3	10088266	rs72492997	G	T	194	39	233	FANCD2	exonic	het	synonymous	NM_033084	exon15	c.G1137T	p.V379V	705
chr3	10143061	.	CT	C	6	13	43	FANCD2	UTR3	het	-	NM_001018115	-	c.*116delT	-	112
chr3	10088308	rs72492998	T	C	164	63	227	FANCD2	exonic	het	synonymous	NM_033084	exon15	c.T1179C	p.T393T	1905
chr3	10106532	rs3864017	C	T	118	93	211	FANCD2	exonic	het	nonsynonymous	NM_033084	exon23	c.C2141T	p.P714L	2105
chr3	10085536	rs34046352	A	G	139	50	189	FANCD2	exonic	het	synonymous	NM_033084	exon14	c.A1122G	p.V374V	892
chr6	158616173	rs72413565	C	CT	72	26	102	GTF2H5	UTR3	het	-	NM_207118	-	c.*2984_*2985in	-	252
chr9	35076027	.	T	C	174	170	344	FANCG	splicing	het	-	NM_004629	exon10	c.1077-2A>G	-	3600
chr3	10088407	.	AG	A	92	46	138	FANCD2	splicing	het	-	NM_033084	exon15	c.1278+1G>	-	1727
chr3	10088343	rs73126218	A	G	156	82	238	FANCD2	exonic	het	nonsynonymous	NM_033084	exon15	c.A1214G	p.N405S	1537
chr3	10088299	rs112887807	C	T	174	52	226	FANCD2	exonic	het	synonymous	NM_033084	exon15	c.C1170T	p.S390S	1723

c- Amostra TTD02SP

CHROM	POSITION	RS_ID	REF	ALT	#REF	#ALT	#DEPH	GENE	REGION	GENOTYPE	VAR_TYPE	GENE_ID	EXON	CDNA_POS	PROT_POS	#QUAL
chr6	158618449	.	G	GA	29	24	55	GTF2H5	UTR3	het	-	NM_207118	-	c.*5260_*5261in	-	459
chr3	10088308	rs72492998	T	C	193	50	243	FANCD2	exonic	het	synonymous	NM_033084	exon15	c.T1179C	p.T393T	1311
chr19	45911319	rs11314106	CA	C	74	112	190	ERCC1	UTR3	het	-	NM_001166049	-	c.*1613delIT	-	2227
chr6	158616173	rs72413565	C	CT	71	33	115	GTF2H5	UTR3	het	-	NM_207118	-	c.*2984_*2985in	-	431
chr3	10088343	rs73126218	A	G	200	66	266	FANCD2	exonic	het	nonsynonymous	NM_033084	exon15	c.A1214G	p.N405S	973
chr17	7578711	.	CTTTT	C,CT	1	29	139	TP53	UTR5	het	-	NM_001126115	-	c.-180_-182delA	-	3859
chr3	10088404	.	C	T	132	42	174	FANCD2	exonic	het	synonymous	NM_033084	exon15	c.C1275T	p.Y425Y	1315
chr16	89804335	rs17233826	CT	C	2	313	333	FANCA	UTR3	hom	-	NM_000135	-	c.*673delA	-	9393
chr3	10089723	rs12330369	G	A	258	89	348	FANCD2	exonic	het	synonymous	NM_033084	exon16	c.G1401A	p.T467T	2575
chr3	10088407	.	AG	A	112	41	153	FANCD2	splicing	het	-	NM_033084	exon15	c.1278+1G>	-	1302
chr3	10088266	rs72492997	G	T	262	34	296	FANCD2	exonic	het	synonymous	NM_033084	exon15	c.G1137T	p.V379V	62
chr17	41196821	.	CTTT	C	5	4	10	BRCA1	UTR3	het	-	NM_007300	-	c.*873_*871delA	-	114
chr3	10088299	rs112887807	C	T	212	42	254	FANCD2	exonic	het	synonymous	NM_033084	exon15	c.C1170T	p.S390S	1139
chr6	158613937	.	CAAA	C	6	5	13	GTF2H5	UTR3	het	-	NM_207118	-	c.*749_*751delA	-	33
chr19	45856516	.	A	G	113	84	198	ERCC2	exonic	het	nonsynonymous	NM_000400	exon18	c.T1742C	p.L581P	1799

d- Amostra XP01RN

CHROM	POSITION	RS_ID	REF	ALT	#REF	#ALT	#DEPH	GENE	REGION	GENOTYPE	VAR_TYPE	GENE_ID	EXON	CDNA_POS	PROT_POS	#QUAL
chr6	43587101	.	CAA	C	2	12	17	POLH	UTR3	het	-	NM_006502	-	c.*4808_*4809de	-	288
chr3	10088308	rs72492998	T	C	200	33	233	FANCD2	exonic	het	synonymous	NM_033084	exon15	c.T1179C	p.T393T	546
chr10	50666628	.	CA	C	105	31	151	ERCC6	UTR3	het	-	NM_000124	-	c.*232delIT	-	216
chr17	41196821	.	CTTT	C	7	25	32	BRCA1	UTR3	het	-	NM_007300	-	c.*873_*871delA	-	632
chr3	10088299	rs112887807	C	T	215	26	241	FANCD2	exonic	het	synonymous	NM_033084	exon15	c.C1170T	p.S390S	454
chr6	158616173	.	CT	C	66	99	169	GTF2H5	UTR3	het	-	NM_207118	-	c.*2985delIT	-	1480
chr3	10089723	rs12330369	G	A	290	102	392	FANCD2	exonic	het	synonymous	NM_033084	exon16	c.G1401A	p.T467T	1718
chr17	7572154	.	GAA	G,GA	11	36	82	TP53	UTR3	het	-	NM_001276761	-	c.*771delIT	-	1209
chr3	10143061	.	CTT	C	12	14	56	FANCD2	UTR3	het	-	NM_001018115	-	c.*116_*117delIT	-	264
chr19	45911320	.	A	AAAAAATCAAAAAACCT	108	57	174	ERCC1	UTR3	het	-	NM_001166049	-	c.*1612_*1613in	-	1062
chr3	10088407	.	AG	A	108	27	135	FANCD2	splicing	het	-	NM_033084	exon15	c.1278+1G>	-	739
chr3	10088404	.	C	T	120	31	151	FANCD2	exonic	het	synonymous	NM_033084	exon15	c.C1275T	p.Y425Y	849
chr17	7578711	.	CTTTT	C,CT	0	35	145	TP53	UTR5	het	-	NM_001126115	-	c.-180_-182delA	-	4788
chr6	43565580	.	C	G	0	431	432	POLH	exonic	hom	stopgain	NM_001291969	exon3	c.C266G	p.S89X	11731
chr13	32973923	rs35930474	C	CT	23	16	53	BRCA2,N4BP2L1	downstream	het	-	-	-	-	-	155
chr3	10088343	rs73126218	A	G	203	44	247	FANCD2	exonic	het	nonsynonymous	NM_033084	exon15	c.A1214G	p.N405S	460

e- Amostra BA02

CHROM	POSITION	RS_ID	REF	ALT	#REF	#ALT	#DEPH	GENE	REGION	GENOTYPE	VAR_TYPE	GENE_ID	EXON	CDNA_POS	PROT_POS	#QUAL
chr3	10088266	rs72492997	G	T	196	38	234	FANCD2	exonic	het	synonymous	NM_001018115	exon15	c.G1137T	p.V379V	653
chr3	10089723	rs12330369	G	A	203	119	323	FANCD2	exonic	het	synonymous	NM_001018115	exon16	c.G1401A	p.T467T	3820
chr3	10105516	rs142354499	A	C	195	88	285	FANCD2	exonic	het	nonsynonymous	NM_001018115	exon21	c.A1868C	p.Q623P	1593
chr6	158613937	.	CAAAAA	C	8	6	20	GTF2H5	UTR3	het	-	NM_207118	-	c.*749_*753delA	-	134
chr3	14186830	rs2470352	A	G	102	104	206	XPC	UTR3	het	-	NM_004628	-	c.*611T>C	-	2358
chr3	10106532	rs3864017	C	T	146	90	236	FANCD2	exonic	het	nonsynonymous	NM_001018115	exon23	c.C2141T	p.P714L	1993
chr17	41196821	rs397857225	CTT	C	0	14	19	BRCA1	UTR3	hom	-	NM_007300	-	c.*873_*872delA	-	350
chr17	7578711	rs141204613	CTTT	C	0	96	109	TP53	UTR5	hom	-	NM_001126115	-	c.-179_-181delA	-	3340
chr3	10085536	rs34046352	A	G	46	83	129	FANCD2	exonic	het	synonymous	NM_001018115	exon14	c.A1122G	p.V374V	1964
chr19	45911319	rs11314106	CA	C	54	53	110	ERCC1	UTR3	het	-	NM_001166049	-	c.*1613delT	-	856
chr17	7572154	.	GAA	G,GA	6	25	62	TP53	UTR3	het	-	NM_001276761	-	c.*771delT	-	959
chr6	43584323	.	C	CAA	29	28	58	POLH	UTR3	het	-	NM_006502	-	c.*2029_*2030in	-	517
chr6	158616173	.	CT	C	64	29	100	GTF2H5	UTR3	het	-	NM_207118	-	c.*2985delT	-	274
chr9	100447232	.	G	A	59	61	120	XPA	exonic	het	stopgain	NM_000380	exon5	c.C646T	p.Q216X	1266
chr3	10088407	.	AG	A	104	47	151	FANCD2	splicing	het	-	NM_033084	exon15	c.1278+1G>-	-	1657
chr3	10088299	rs112887807	C	T	173	46	219	FANCD2	exonic	het	synonymous	NM_001018115	exon15	c.C1170T	p.S390S	1540
chr3	10088343	rs73126218	A	G	162	79	241	FANCD2	exonic	het	nonsynonymous	NM_001018115	exon15	c.A1214G	p.N405S	1406
chr3	10143061	.	CT	C	3	15	33	FANCD2	UTR3	het	-	NM_001018115	-	c.*116delT	-	226
chr3	10088308	rs72492998	T	C	166	61	227	FANCD2	exonic	het	synonymous	NM_001018115	exon15	c.T1179C	p.T393T	1728
chr3	10088404	.	C	T	114	50	164	FANCD2	exonic	het	synonymous	NM_001018115	exon15	c.C1275T	p.Y425Y	1634
chr11	47236322	.	TA	T	0	4	4	DDB2	upstream	hom	-	-	-	-	-	36

f- Amostra XP01TO

CHROM	POSITION	RS_ID	REF	ALT	#REF	#ALT	#DEPH	GENE	REGION	GENOTYPE	VAR_TYPE	GENE_ID	EXON	CDNA_POS	PROT_POS	#QUAL
chr3	10088308	rs72492998	T	C	147	52	200	FANCD2	exonic	het	synonymous	NM_033084	exon15	c.T1179C	p.T393T	1515
chr17	41196821	rs397857225	CTT	C	0	9	14	BRCA1	UTR3	hom	-	NM_007300	-	c.*873_*872delA	-	195
chr13	32973923	.	CT	C	14	7	28	BRCA2,N4BP2L1	downstream	het	-	-	-	-	-	54
chr3	10088404	.	C	T	77	51	128	FANCD2	exonic	het	synonymous	NM_033084	exon15	c.C1275T	p.Y425Y	1800
chr3	10088299	rs112887807	C	T	164	46	211	FANCD2	exonic	het	synonymous	NM_033084	exon15	c.C1170T	p.S390S	1410
chr13	32888981	.	C	T	0	2	2	BRCA2	upstream	hom	-	-	-	-	-	22
chr3	10088407	.	AG	A	65	47	112	FANCD2	splicing	het	-	NM_033084	exon15	c.1278+1G>-	-	1702
chr11	47236322	.	TA	T	2	6	8	DDB2	upstream	het	-	-	-	-	-	66
chr3	10089723	rs12330369	G	A	258	146	404	FANCD2	exonic	het	synonymous	NM_033084	exon16	c.G1401A	p.T467T	4774
chr17	7572154	.	GAA	G,GA	9	29	69	TP53	UTR3	het	-	NM_001276761	-	c.*771delT	-	990
chr19	45911319	rs11314106	CA	C	94	100	194	ERCC1	UTR3	het	-	NM_001166049	-	c.*1613delT	-	1774
chr6	158613937	.	CA	C	1	13	19	GTF2H5	UTR3	hom	-	NM_207118	-	c.*749delA	-	216
chr6	43587101	.	CAAA	C	5	4	10	POLH	UTR3	het	-	NM_006502	-	c.*4808_*4810de	-	60
chr3	10088343	rs73126218	A	G	140	68	209	FANCD2	exonic	het	nonsynonymous	NM_033084	exon15	c.A1214G	p.N405S	1164
chr3	10088266	rs72492997	G	T	199	37	238	FANCD2	exonic	het	synonymous	NM_033084	exon15	c.G1137T	p.V379V	708
chr10	50666628	.	CA	C	84	26	120	ERCC6	UTR3	het	-	NM_000124	-	c.*232delT	-	124
chr16	89804335	rs17233826	CT	C	0	264	264	FANCA	UTR3	hom	-	NM_000135	-	c.*673delA	-	9890
chr17	7578711	.	CTTTT	C,CT	0	28	117	TP53	UTR5	het	-	NM_001126115	-	c.-180_-182delA	-	3519

g- Amostra XP01PE

CHROM	POSITION	RS_ID	REF	ALT	#REF	#ALT	#DEPH	GENE	REGION	GENOTYPE	VAR_TYPE	GENE_ID	EXON	CDNA_POS	PROT_POS	#QUAL
chr17	7572154	.	GAA	G,GA	3	28	70	TP53	UTR3	het	-	NM_001276761	-	c.*771delT	-	1242
chr3	10088343	rs73126218	A	G	182	29	211	FANCD2	exonic	het	nonsynonymous	NM_033084	exon15	c.A1214G	p.N405S	101
chr3	10106532	rs3864017	C	T	178	86	264	FANCD2	exonic	het	nonsynonymous	NM_033084	exon23	c.C2141T	p.P714L	1621
chr3	10088404	.	C	T	119	22	141	FANCD2	exonic	het	synonymous	NM_033084	exon15	c.C1275T	p.Y425Y	484
chr3	14197899	.	C	A	187	190	377	XPC	exonic	het	stopgain	NM_004628	exon10	c.G1969T	p.E657X	3852
chr16	89804335	rs17233826	CT	C	2	280	295	FANCA	UTR3	hom	-	NM_000135	-	c.*673delA	-	8426
chr19	45913152	rs368190454	TA	T	0	4	4	CD3EAP	UTR3	hom	-	NM_001297590	-	c.*394delA	-	90
chr3	10143086	.	T	A	13	24	40	FANCD2	UTR3	het	-	NM_001018115	-	c.*140T>A	-	340
chr3	10089723	rs12330369	G	A	188	86	274	FANCD2	exonic	het	synonymous	NM_033084	exon16	c.G1401A	p.T467T	1461
chr3	10085536	rs34046352	A	G	145	48	193	FANCD2	exonic	het	synonymous	NM_033084	exon14	c.A1122G	p.V374V	684
chr3	14188829	.	CA	C	220	164	405	XPC	exonic	het	frameshift deletion	NM_004628	exon15	c.2564delT	p.L855fs	5334
chr13	32973923	.	CT	C	34	15	65	BRCA2,N4BP2L1	downstream	het	-	-	-	-	-	113
chr3	10088407	.	AG	A	109	19	128	FANCD2	splicing	het	-	NM_033084	exon15	c.1278+1G>-	-	465
chr17	7578711	.	CTTTT	C,CT	0	23	124	TP53	UTR5	het	-	NM_001126115	-	c.-180_-182delA	-	3663
chr3	10143061	.	CT	C	19	18	50	FANCD2	UTR3	het	-	NM_001018115	-	c.*116delT	-	148
chr6	158616173	.	CT	C	51	47	107	GTF2H5	UTR3	het	-	NM_207118	-	c.*2985delT	-	587
chr16	89807044	.	CAA	C	8	4	12	ZNF276	UTR3	het	-	NM_152287	-	c.*2391_*2392de	-	50

h- Amostra PE02

CHROM	POSITION	RS_ID	REF	ALT	#REF	#ALT	#DEPH	GENE	REGION	GENOTYPE	VAR_TYPE	GENE_ID	EXON	CDNA_POS	PROT_POS	#QUAL
chr6	43587101	.	CAAAA	C	3	3	9	POLH	UTR3	het	-	NM_006502	-	c.*4808_*4811de	-	44
chr6	158616173	.	CT	C	74	28	107	GTF2H5	UTR3	het	-	NM_207118	-	c.*2985delT	-	219
chr3	10088266	rs72492997	G	T	204	19	225	FANCD2	exonic	het	synonymous	NM_001018115	exon15	c.G1137T	p.V379V	69
chr6	158615020	.	CAA	C	0	4	5	GTF2H5	UTR3	hom	-	NM_207118	-	c.*1832_*1833de	-	72
chr16	89804335	rs17233826	CT	C	2	230	248	FANCA	UTR3	hom	-	NM_000135	-	c.*673delA	-	6851
chr3	10089723	rs12330369	G	A	243	172	415	FANCD2	exonic	het	synonymous	NM_001018115	exon16	c.G1401A	p.T467T	3829
chr3	10106532	rs3864017	C	T	222	106	328	FANCD2	exonic	het	nonsynonymous	NM_001018115	exon23	c.C2141T	p.P714L	2045
chr3	10088404	.	C	T	89	44	133	FANCD2	exonic	het	synonymous	NM_001018115	exon15	c.C1275T	p.Y425Y	1502
chr3	10088299	rs112887807	C	T	179	29	208	FANCD2	exonic	het	synonymous	NM_001018115	exon15	c.C1170T	p.S390S	618
chr17	7572154	.	GAA	G,GA	6	28	66	TP53	UTR3	het	-	NM_001276761	-	c.*771delT	-	1120
chr6	43584323	.	CA	C	68	20	89	POLH	UTR3	het	-	NM_006502	-	c.*2030delA	-	140
chr3	10088308	rs72492998	T	C	165	40	205	FANCD2	exonic	het	synonymous	NM_001018115	exon15	c.T1179C	p.T393T	874
chr3	10085536	rs34046352	A	G	150	34	184	FANCD2	exonic	het	synonymous	NM_001018115	exon14	c.A1122G	p.V374V	362
chr3	14188829	.	CA	C	204	189	418	XPC	exonic	het	frameshift deletion	NM_004628	exon15	c.2564delT	p.L855fs	6287
chr3	10088343	rs73126218	A	G	160	51	213	FANCD2	exonic	het	nonsynonymous	NM_001018115	exon15	c.A1214G	p.N405S	708
chr17	7578711	.	CTTTT	C,CT	1	27	108	TP53	UTR5	het	-	NM_001126115	-	c.-180_-182delA	-	3680
chr3	10088407	.	AG	A	78	42	120	FANCD2	splicing	het	-	NM_033084	exon15	c.1278+1G>-	-	1465

i- Amostra XP03BA

CHROM	POSITION	RS_ID	REF	ALT	#REF	#ALT	#DEPH	GENE	REGION	GENOTYPE	VAR_TYPE	GENE_ID	EXON	CDNA_POS	PROT_POS	#QUAL
chr6	43584323	.	C	CAA	28	19	47	POLH	UTR3	het	-	NM_006502	-	c.*2029_*2030in	-	298
chr13	32973923	.	CT	C	18	16	42	BRCA2,N4BP2L1	downstream	het	-	-	-	-	-	204
chr3	10088299	rs112887807	C	T	175	24	199	FANCD2	exonic	het	synonymous	NM_033084	exon15	c.C1170T	p.S390S	456
chr17	7572154	rs200757381	GA	G	9	30	41	TP53	UTR3	het	-	NM_001276761	-	c.*772delT	-	398
chr16	89804335	rs17233826	CT	C	1	304	324	FANCA	UTR3	hom	-	NM_000135	-	c.*673delA	-	9114
chr3	10088407	.	AG	A	81	30	111	FANCD2	splicing	het	-	NM_033084	exon15	c.1278+1G>	-	988
chr3	10088308	rs72492998	T	C	162	28	190	FANCD2	exonic	het	synonymous	NM_033084	exon15	c.T1179C	p.T393T	542
chr3	10088266	rs72492997	G	T	207	20	230	FANCD2	exonic	het	synonymous	NM_033084	exon15	c.G1137T	p.V379V	48
chr3	10089723	rs12330369	G	A	231	76	308	FANCD2	exonic	het	synonymous	NM_033084	exon16	c.G1401A	p.T467T	2272
chr19	45911319	rs11314106	CA	C	73	78	151	ERCC1	UTR3	het	-	NM_001166049	-	c.*1613delT	-	1371
chr3	10143607	rs532834749	C	T	61	55	116	FANCD2	UTR3	het	-	NM_001018115	-	c.*661C>T	-	1151
chr17	7578711	.	CTTTT	C,CTT	0	72	110	TP53	UTR5	het	-	NM_001126115	-	c.-181_-183delA	-	3306
chr3	10143061	.	CT	C	6	20	41	FANCD2	UTR3	hom	-	NM_001018115	-	c.*116delT	-	279
chr3	10088343	rs73126218	A	G	149	45	194	FANCD2	exonic	het	nonsynonymous	NM_033084	exon15	c.A1214G	p.N405S	615
chr6	43568737	.	C	T	131	142	273	POLH	exonic	het	synonymous	NM_001291970	exon6	c.C673T	p.L225L	5463
chr3	10088404	.	C	T	93	32	125	FANCD2	exonic	het	synonymous	NM_033084	exon15	c.C1275T	p.Y425Y	995
chr6	43568736	.	A	AT	131	139	270	POLH	exonic	het	frameshift insertion	NM_001291970	exon6	c.672_673insT	p.K224fs	5454
chr2	128051995	rs142337518	TGA	T	4	2	8	ERCC3	upstream	het	-	-	-	-	-	20
chr6	158616173	rs72413565	C	CT	54	32	91	GTF2H5	UTR3	het	-	NM_207118	-	c.*2984_*2985in	-	423

j- Amostra XP02RJ

CHROM	POSITION	RS_ID	REF	ALT	#REF	#ALT	#DEPH	GENE	REGION	GENOTYPE	VAR_TYPE	GENE_ID	EXON	CDNA_POS	PROT_POS	#QUAL
chr3	10088343	rs73126218	A	G	151	93	244	FANCD2	exonic	het	nonsynonymous S	NM_033084	exon15	c.A1214G	p.N405S	1793
chr13	103498699	rs267607281	C	A	134	116	250	ERCC5	exonic	het	nonsynonymous S	NM_000123	exon1	c.C83A	p.A28D	2638
chr3	10088299	rs112887807	C	T	186	59	245	FANCD2	exonic	het	synonymous SNV	NM_033084	exon15	c.C1170T	p.S390S	938
chr3	10088406	rs369823368	TAGTA	T	86	53	157	FANCD2	exonic	het	frameshift dele	NM_033084	exon15	c.1278_1278del	p.L426fs	2947
chr3	10089723	rs12330369	G	A	167	83	250	FANCD2	exonic	het	synonymous SNV	NM_033084	exon16	c.G1401A	p.T467T	1884
chr19	45912489	rs35729377	CAAG	C	114	125	249	CD3EAP	exonic	het	nonframeshift d	NM_012099	exon3	c.1264_1266del	p.422_422del	5485
chr19	45911319	rs11314106	CA	C	0	190	190	ERCC1	UTR3	hom	-	NM_001166049	-	c.*1613delT	-	5052
chr17	7578711	.	CTTTT	C,CT	5	48	206	TP53	UTR5	het	-	NM_001126115	-	c.-180_-182delA	-	6723
chr6	158616173	rs74377352	CT	C	106	33	143	GTF2H5	UTR3	het	-	NM_207118	-	c.*2985delT	-	368
chr3	10088308	rs72492998	T	C	179	70	249	FANCD2	exonic	het	synonymous SNV	NM_033084	exon15	c.T1179C	p.T393T	1120
chr3	10088404	.	C	T	111	57	168	FANCD2	exonic	het	synonymous SNV	NM_033084	exon15	c.C1275T	p.Y425Y	1094
chr3	10088266	rs72492997	G	T	206	44	250	FANCD2	exonic	het	synonymous SNV	NM_033084	exon15	c.G1137T	p.V379V	653
chr3	10143061	.	CT	C	28	40	90	FANCD2	UTR3	het	-	NM_001018115	-	c.*116delT	-	436
chr13	103525633	rs267607280	G	C	137	112	250	BIVM-ERCC5,ERCC5	exonic	het	nonsynonymous S	NM_000123	exon14	c.G2904C	p.W968C	2683

k- Amostra XP03AM

CHROM	POSITION	RS_ID	REF	ALT	#REF	#ALT	#DEPH	GENE	REGION	GENOTYPE	VAR_TYPE	GENE_ID	EXON	CDNA_POS	PROT_POS	#QUAL
chr16	14043490	rs185626419	A	G	52	58	110	ERCC4	UTR3	het	-	NM_005236	-	c.*1286A>G	-	1500
chr16	14045297	rs183916977	T	C	72	71	144	ERCC4	UTR3	het	-	NM_005236	-	c.*3093T>C	-	1657
chr17	7578711	.	CTTTT	C,CT	1	33	146	TP53	UTR5	het	-	NM_001126115	-	c.-180_-182delA	-	4648
chr10	50666628	.	CA	C	85	110	213	ERCC6	UTR3	het	-	NM_000124	-	c.*232delT	-	1727
chr6	43584323	.	C	CAAA	51	27	80	POLH	UTR3	het	-	NM_006502	-	c.*2029_*2030in	-	620
chr3	10088299	rs112887807	C	T	245	28	273	FANCD2	exonic	het	synonymous	NM_033084	exon15	c.C1170T	p.S390S	417
chr11	47236322	.	TA	T	6	8	14	DDB2	upstream	het	-	-	-	-	-	75
chr6	43587101	.	CAAAAA	C	3	6	10	POLH	UTR3	het	-	NM_006502	-	c.*4808_*4812de	-	156
chr6	158616173	rs72413565	C	CT	47	52	106	GTF2H5	UTR3	het	-	NM_207118	-	c.*2984_*2985in	-	897
chr3	10143061	.	CTT	C	19	16	63	FANCD2	UTR3	het	-	NM_001018115	-	c.*116_*117delT	-	279
chr3	10088407	.	AG	A	127	34	161	FANCD2	splicing	het	-	NM_033084	exon15	c.1278+1G>-	-	957
chr11	47256943	.	C	T	0	343	343	DDB2	exonic	hom	stopgain	NM_000107	exon7	c.C1003T	p.Q335X	10118
chr13	32973923	.	CT	C	34	12	58	BRCA2,N4BP2L1	downstream	het	-	-	-	-	-	72
chr3	10089723	rs12330369	G	A	338	109	449	FANCD2	exonic	het	synonymous	NM_033084	exon16	c.G1401A	p.T467T	3132
chr17	7572154	.	GAA	G,GA	10	29	81	TP53	UTR3	het	-	NM_001276761	-	c.*771delT	-	1165
chr3	10088343	rs73126218	A	G	242	46	288	FANCD2	exonic	het	nonsynonymous	NM_033084	exon15	c.A1214G	p.N405S	402
chr3	10088404	.	C	T	144	35	179	FANCD2	exonic	het	synonymous	NM_033084	exon15	c.C1275T	p.Y425Y	988
chr3	10088308	rs72492998	T	C	246	31	277	FANCD2	exonic	het	synonymous	NM_033084	exon15	c.T1179C	p.T393T	445
chr16	14045774	.	A	G	128	75	203	ERCC4	UTR3	het	-	NM_005236	-	c.*3570A>G	-	1687
chr16	89804335	rs17233826	CT	C	0	298	300	FANCA	UTR3	hom	-	NM_000135	-	c.*673delA	-	11220
chr19	45911319	rs11314106	CA	C	0	247	249	ERCC1	UTR3	hom	-	NM_001166049	-	c.*1613delT	-	5850

l- Amostra XP05AM

CHROM	POSITION	RS_ID	REF	ALT	#REF	#ALT	#DEPH	GENE	REGION	GENOTYPE	VAR_TYPE	GENE_ID	EXON	CDNA_POS	PROT_POS	#QUAL
chr6	43582427	.	AG	A	3	2	5	POLH	UTR3	het	-	NM_006502	-	c.*134delG	-	21
chr9	97860726	.	TC	T	0	2	2	FANCC	downstream	hom	-	-	-	-	-	38
chr9	97873168	.	TC	T	1	2	3	FANCC	UTR3	het	-	NM_001243744	-	c.*309delG	-	25
chr6	158616045	.	AT	A	5	2	7	GTF2H5	UTR3	het	-	NM_207118	-	c.*2857delT	-	22
chr16	89865058	.	GTA	G	8	2	10	FANCA	UTR3	het	-	NM_001018112	-	c.*428_*427delT	-	23
chr16	14045272	.	G	A	0	2	2	ERCC4	UTR3	hom	-	NM_005236	-	c.*3068G>A	-	22
chr6	43587871	.	TCTA	T	3	2	5	POLH	UTR3	het	-	NM_006502	-	c.*5578_*5580de	-	38
chr16	89804335	rs17233826	CT	C	5	2	7	FANCA	UTR3	het	-	NM_000135	-	c.*673delA	-	19
chr9	97863937	.	C	T	4	2	6	FANCC	UTR3	het	-	NM_001243743	-	c.*52G>A	-	44
chr13	103529006	.	G	T	0	2	2	BIVM-ERCC5,ERCC5	downstream	hom	-	-	-	-	-	22

m- Amostra XPSPAC06F0

CHROM	POSITION	RS_ID	REF	ALT	#REF	#ALT	#DEPH	GENE	REGION	GENOTYPE	VAR_TYPE	GENE_ID	EXON	CDNA_POS	PROT_POS	#QUAL
chr19	45855610	rs14556519	G	A	12	14	26	ERCC2	exonic	het	nonsynonymous	NM_000400	exon22	c.C2047T	p.R683W	261
chr2	128014819	.	C	T	0	2	2	ERCC3	downstream	hom	-	-	-	-	-	22
chr17	7579803	.	G	A	0	2	2	TP53	UTR5	hom	-	NM_001126118	-	c.-234C>T	-	63
chr10	50724064	.	G	A	9	2	11	ERCC6-PGBD3,PGBD3	exonic	het	nonsynonymous	NM_001277059	exon6	c.C2501T	p.S834L	29
chr17	7578711	rs141204613	CTTT	C	0	4	4	TP53	UTR5	hom	-	NM_001126115	-	c.-179_-181delA	-	111
chr19	45855507	rs144564120	G	C	8	9	17	ERCC2	exonic	het	nonsynonymous	NM_000400	exon22	c.C2150G	p.A717G	164
chr17	41244865	rs397508996	GT	G	2	2	4	BRCA1	exonic	het	frameshift deletion	NM_007294	exon10	c.2682delA	p.K894fs	20
chr10	50724069	.	G	A	9	2	11	ERCC6-PGBD3,PGBD3	exonic	het	synonymous	NM_001277059	exon6	c.C2496T	p.L832L	29

n- Amostra XPSPAC11F0

CHROM	POSITION	RS_ID	REF	ALT	#REF	#ALT	#DEPH	GENE	REGION	GENOTYPE	VAR_TYPE	GENE_ID	EXON	CDNA_POS	PROT_POS	#QUAL
chr3	10088404	.	C	T	1	3	4	FANCD2	exonic	het	synonymous	NM_001018115	exon15	c.C1275T	p.Y425Y	95
chr13	103528232	.	G	A	10	8	18	BIVM-ERCC5,ERCC5	exonic	het	synonymous	NM_000123	exon15	c.G3540A	p.R1180R	116
chr16	89804335	rs17233826	CT	C	0	23	23	FANCA	UTR3	hom	-	NM_000135	-	c.*673delA	-	831
chr6	43565513	.	A	C	0	26	27	POLH	exonic	hom	nonsynonymous	NM_001291970	exon5	c.A571C	p.T191P	696
chr3	10089723	rs12330369	G	A	16	5	21	FANCD2	exonic	het	synonymous	NM_001018115	exon16	c.G1401A	p.T467T	42
chr17	7572154	.	GAA	G	1	3	4	TP53	UTR3	het	-	NM_001276761	-	c.*772_*771delT	-	39
chr2	128014912	.	AC	A	4	2	6	ERCC3	UTR3	het	-	NM_000122	-	c.*259delG	-	25
chr3	10106532	rs3864017	C	T	16	8	24	FANCD2	exonic	het	nonsynonymous	NM_001018115	exon23	c.C2141T	p.P714L	121
chr3	10088299	rs112887807	C	T	12	3	15	FANCD2	exonic	het	synonymous	NM_001018115	exon15	c.C1170T	p.S390S	62
chr3	10088407	.	AG	A	1	3	4	FANCD2	splicing	het	-	NM_033084	exon15	c.1278+1G>	-	86
chr3	10088308	rs72492998	T	C	10	3	13	FANCD2	exonic	het	synonymous	NM_001018115	exon15	c.T1179C	p.T393T	68
chr3	10088343	rs73126218	A	G	6	4	10	FANCD2	exonic	het	nonsynonymous	NM_001018115	exon15	c.A1214G	p.N405S	63
chr3	10085536	rs34046352	A	G	6	3	9	FANCD2	exonic	het	synonymous	NM_001018115	exon14	c.A1122G	p.V374V	30

o- Amostra XPSPAC31F0

CHROM	POSITION	RS_ID	REF	ALT	#REF	#ALT	#DEPH	GENE	REGION	GENOTYPE	VAR_TYPE	GENE_ID	EXON	CDNA_POS	PROT_POS	#QUAL
chr10	50680471	rs190863815	C	A	16	8	24	ERCC6	exonic	het	nonsynonymous	NM_000124	exon16	c.G2875T	p.V959L	115
chr19	45912489	rs35729377	CAAG	C	10	25	36	CD3EAP	exonic	het	nonframeshift deleti	NM_001297590	exon3	c.1270_1272del	p.424_424del	983
chr17	7578711	rs141204613	CTTT	C	0	7	10	TP53	UTR5	hom	-	NM_001126115	-	c.-179_-181delA	-	181
chr6	43565576	.	T	C	0	37	37	POLH	exonic	hom	nonsynonymous	NM_001291970	exon5	c.T634C	p.C212R	950

Anexo 3

UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Biomédicas

Av. Prof. Lineu Prestes, 1374
Cidade Universitária
05508 São Paulo
Fax: 3091-7354
Fone: 3091-7499

São Paulo, 17 de agosto de 2015.

A quem possa interessar,

O projeto "Busca de variantes em sequência de DNA provenientes de pacientes com deficiência em processos de reparo do genoma", desenvolvido durante o projeto de Mestrado da aluna Livia Maria Silva Moura. Este projeto caracterizou-se como um estudo informático de sequências de DNA de amostras vinculadas ao projeto de pesquisa institucional intitulado "Determinação de mutação genética de isolado genético de xeroderma pigmentosum em Goiás" aprovado pelo Comitê de Ética em Pesquisa do ICB (PARECER 1007/CEP) no dia 29 de junho de 2011 e com foco ampliado (Of.CEPSH.045.13) no dia 08 de maio de 2013 com os pareceres em anexo.

Me coloco a seu inteiro dispor para qualquer informação que se fizer necessária.

Atenciosamente,

Dr. Carlos Frederico Martins Menck
Depto. de Microbiologia, ICB, USP
Av. Prof. Lineu Prestes, 1374.
Ed. Biomédicas 2, São Paulo, 05508-900, SP,
tel. 55.11.3091.7499; fax# 55.11.3091.7354

UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Biomédicas

Av. Prof. Lineu Prestes, 1374
Cidade Universitária
05508 São Paulo
Fax: 3091-7354

UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE CIÊNCIAS BIOMÉDICAS

Cidade Universitária "Armando de Salles Oliveira"
Av. Prof. Lineu Prestes, 2415 - cep. 05508-000 São Paulo, SP - Brasil
Telefone : (55) (11) 3091.7733 telefax : (55) (11) 3091-8405
e-mail: cep@icb.usp.br

São Paulo, 11 de agosto de 2011.

PARECER 1007/CEP

A Comissão de Ética em Pesquisas com Seres Humanos do ICB, na sessão realizada no dia 29.06.2011, **APROVOU** o projeto intitulado: " *Determinação de mutação genética de isolado genético de xeroderma pigmentosum em Goiás* " sob responsabilidade de execução do autor Prof. Dr. CARLOS FREDERICO MARTINS MENCK.

Cabe ao pesquisador executante elaborar e apresentar a este Comitê, relatórios anuais (parciais ou final), de acordo com a resolução 196/06 do Conselho Nacional da Saúde, item IX. 2 letra c.

O primeiro relatório deverá ser encaminhado à Secretaria deste CEP em **29.06.2012.**

Atenciosamente,

Prof. Dr. PAOLO M.A. ZANOTTO
Coordenador da Comissão de Ética em
Pesquisas com Seres Humanos - ICB/USP

UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE CIÊNCIAS BIOMÉDICAS

Cidade Universitária "Armando de Salles Oliveira"
Av. Prof. Lineu Prestes, 2415 - cep. 05508-000 São Paulo, SP - Brasil
Telefone: (55) (011) 3091.7733 - e-mail: cep@icb.usp.br

Of.CEPSH.045.13

Ilmo. Sr.
Prof.Dr. Prof. Dr. Carlos Frederico M. Menck
Chefe do Departamento de Microbiologia -ICB/USP

São Paulo, 14 de maio de 2013.

Prezado Professor Menck,

A Comissão de Ética em Pesquisa em Seres Humanos, em reunião de 08.05.2013, analisou sua solicitação datada de 02.04.13, decidiu-se por aprovar a proposta para ampliar o foco do projeto: "*Determinação de mutação genética de isolado genético de xeroderma pigmentosum em Goiás* " aprovado por esta Comissão em 29.06.2011.

Sugerimos ao pesquisador que faça todas as adequações do projeto de acordo com o solicitado e aprovado *antes* da migração do projeto para Plataforma Brasil, o que está sendo solicitado por colaboradores.

Atenciosamente,

Paulo Zanotto, D.Phil.
Coordenador- CEP SH-ICB/USP