

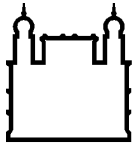
MINISTÉRIO DA SAÚDE  
FUNDAÇÃO OSWALDO CRUZ  
INSTITUTO OSWALDO CRUZ

Doutorado em Programa de Pós-Graduação em Biologia Computacional e Sistemas

O IMPACTO DE PEQUENAS DELEÇÕES GENÔMICAS EM DOMÍNIOS  
PROTÉICOS IDENTIFICADAS A PARTIR DE DADOS DE RNA-SEQ DE  
AMOSTRAS DE PACIENTES DE ADENOCARCINOMA DE PULMÃO

**GABRIEL WAJNBERG**

Rio de Janeiro  
Setembro de 2017



Ministério da Saúde

**FIOCRUZ**  
**Fundação Oswaldo Cruz**

## **INSTITUTO OSWALDO CRUZ**

**Programa de Pós-Graduação em Biologia Computacional e Sistemas**

***Gabriel Wajnberg***

O impacto de pequenas deleções genômicas em domínios protéicos identificadas a partir de dados de rna-seq de amostras de pacientes de adenocarcinoma de pulmão

Tese apresentada ao Instituto Oswaldo Cruz  
como parte dos requisitos para obtenção do título  
de Doutor em Ciências

**Orientador :** Dr. Fabio Passetti

**RIO DE JANEIRO**

Setembro de 2017

Wajnberg, Gabriel .

O Impacto de pequenas deleções genômicas em domínios proteicos identificadas a partir de dados de RNA-Seq de amostras de pacientes de adenocarcinoma de pulmão. / Gabriel Wajnberg. - Rio de Janeiro, 2017.

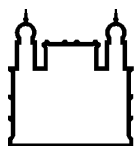
140 f.

Tese (Doutorado) - Instituto Oswaldo Cruz, Pós-Graduação em Biologia Computacional e Sistemas, 2017.

Orientador: Fabio Passetti.

Bibliografia: f. 113-127

1. Bioinformática. 2. Transcriptômica. 3. Genômica. 4. Deleções. 5. Câncer. I. Título.



Ministério da Saúde

FIOCRUZ

Fundação Oswaldo Cruz

## **INSTITUTO OSWALDO CRUZ**

**Programa de Pós-Graduação em Biologia Computacional e Sistemas**

**AUTOR: GABRIEL WAJNBERG**

**O impacto de pequenas deleções genômicas em domínios protéicos  
identificadas a partir de dados de rna-seq de amostras de pacientes de  
adenocarcinoma de pulmão**

**ORIENTADOR : Dr. Fabio Passetti**

**Aprovada em: 06/09/2017**

### **EXAMINADORES:**

**Prof. Dr. Thiago Estevam Parente Martins - Presidente** IOC/Fiocruz  
**Prof. Dr. Francisco Pereira Lobo** UFMG-MG  
**Prof. Dr. Diogo Antônio Tschoeke** UFRJ  
**Prof. Dr. Antônio Basílio de Miranda** IOC/FIOCRUZ  
**Prof. Dr. Gonzalo Bello Betancor** IOC/FIOCRUZ

Rio de Janeiro, 06 de Setembro de 2017



Ministério da Saúde

Fundação Oswaldo Cruz  
Instituto Oswaldo Cruz

Ata da defesa de tese de doutorado em Biologia Computacional e Sistemas de **Gabriel Wajnberg**, sob orientação do Dr. Fabio Passetti. Ao sexto dia do mês de setembro de dois mil e dezessete, realizou-se às treze horas, no Auditório Emmanuel Dias/FIOCRUZ, o exame da tese de doutorado intitulada: **“O impacto de pequenas deleções genômicas identificados a partir de dados de RNA-Seq de amostras de pacientes de adenocarcinoma de pulmão”** No programa de Pós-graduação em Biologia Computacional e Sistemas do Instituto Oswaldo Cruz, como parte dos requisitos para obtenção do título de Doutor em Ciências - área de concentração: Genômica Funcional, Evolução e Filogenômica, na linha de pesquisa: Genoma, transcriptoma, proteoma e metaboloma. A banca examinadora foi constituída pelos Professores: Dr. Thiago Estevam Parente Martins - IOC/FIOCRUZ (Presidente), Dr. Francisco Pereira Lobo - UFMG/MG, Dr. Diogo Antonio Tschoeke - UFRJ/RJ e como suplentes: Dr. Antonio Basilio de Miranda – IOC/FIOCRUZ e Dr. Gonzalo José Bello Bentancor – IOC/FIOCRUZ. Após arguir o candidato e considerando que o mesmo demonstrou capacidade no trato do tema escolhido e sistematização da apresentação dos dados, a banca examinadora pronunciou-se pela APROVAÇÃO da defesa da tese de doutorado. De acordo com o regulamento do Curso de Pós-Graduação em Biologia Computacional e Sistemas do Instituto Oswaldo Cruz, a outorga do título de Doutor em Ciências está condicionada à emissão de documento comprobatório de conclusão do curso. Uma vez encerrado o exame, o Coordenador do Programa, Dr. Ernesto Raúl Caffarena, assinou a presente ata tomando ciência da decisão dos membros da banca examinadora. Rio de Janeiro, 06 de setembro de 2017.

Dr. Thiago Estevam Parente Martins (Presidente da Banca):

Dr. Francisco Pereira Lobo (Membro da Banca):

Dr. Diogo Antonio Tschoeke (Membro da Banca):

Dr. Ernesto Raúl Caffarena (Coordenador do Programa):

**Dedico a todos que me ajudaram a chegar até este momento, em especial minha família: minha esposa Dayane, minha filha Hannah, meus pais Sérgio e Liliana e meu irmão Beni.**

## **AGRADECIMENTOS**

Gostaria de agradecer primeiramente a minha família que me apoiou em todos momentos me dando força para superar esses quatro anos de doutorado. Principalmente ao meu núcleo familiar formado ainda durante no mestrado com a minha esposa Dayane e com a adição neste ano de minha filha Hannah. Também não posso deixar de agradecer aos meus pais por todo apoio que me dão desde a minha decisão de me tornar biólogo até este momento.

Outra pessoa essencial já desde 2009 é meu orientador Dr. Fabio Passeti que me apresentou à bioinformática. Me formei mestre com a sua supervisão e saio deste doutorado mais maduro e independente para a minha próxima etapa como pesquisador.

Nosso laboratório mudou de lugar durante o doutorado e apenas algumas pessoas se mudaram comigo e meu orientador para a FIOCRUZ. Não posso deixar de agradecer a Dra. Nicole de Miranda Scherer que me ajudou bastante desde que assumiu seu cargo no INCA com todo tipo de problema até a problemas específicos em minha tese. Ela participou ativamente em ajudar a escrever os programas de construção das matrizes ternárias.

Preciso dedicar esse espaço a dois colegas que me acompanharam desde 2009 nessa jornada INCA/FIOCRUZ. O Dr. Raphael Tavares da Silva e a futura (até o momento da escrita deste texto) Dra. Natasha Andressa de Nogueira Jorge foram grandes amigos (além de colegas de trabalho) e também pudemos dividir experiências profissionais e técnicas. Também agradeço a Fernanda Cristina, minha ex-estagiária que me ajudou com alguns resultados e a me dar a oportunidade de orienta-la em sua monografia.

Meus novos colegas do LAGFB de trabalho não podem ficar de fora a este agradecimento principalmente da parte de bioinformática: Dra. Ana Carolina Guimarães e Dr. Marcos Catanho; Rafael Pierjorge, Vanessa, Phellippe, Edson, Alexander, Márcio e aos recém-chegados Mayla, Thais, Lucas e Aline.

Últimas pessoas a agradecer que foram muito importantes para mim desde meu início de jornada na bioinformática e que mantenho contato ainda, como: Dr. Gabriel Renaud e Dra. Mariana Brait que me ajudaram a entrar no meio da pesquisa

e me deram forças para procurar um pós-doutorado no exterior e ao meu amigo João Lucas Benaducci, que me ajudou a usar o Matlab para uma disciplina da pós-graduação.

Por último gostaria de agradecer ao Dr. Carlos Gil Ferreira por todo apoio a nossa pesquisa realizada, a FIOCRUZ, o INCA, CAPES, EACR e FAPERJ.



**“Life isn’t about finding yourself. Life is about creating yourself.”**

**George Bernard Shaw**

**O IMPACTO DE PEQUENAS DELEÇÕES GENÔMICAS EM DOMÍNIOS PROTÉICOS IDENTIFICADAS A PARTIR DE DADOS DE RNA-SEQ DE AMOSTRAS DE PACIENTES DE ADENOCARCINOMA DE PULMÃO**

**RESUMO**

**TESE DE DOUTORADO EM BIOLOGIA COMPUTACIONAL E SISTEMAS, IOC/FIOCRUZ**

**Gabriel Wajnberg**

Inserções e deleções (INDELS) são exemplos de alterações na sequência de DNA. As alterações causadas por deleções genômicas podem modificar nucleotídeos da região codificadora de proteínas com potencial para alterar os aminoácidos codificados e mudar o quadro de leitura da tradução. Estas deleções podem causar modificações substanciais em proteínas envolvidas com a biologia do câncer. Nós inovamos no uso de um método já existente de matrizes ternárias ao utilizá-lo para identificar deleções genômicas com o uso de dados de transcriptoma de alta vazão (RNA-Seq). Foram usados dados de RNA-Seq para identificar pequenas deleções de até 100 nucleotídeos de comprimento dentro de éxons humanos. Apresentamos dados a partir da análise do genoma de referência humano GRCh37/hg19 e de 66 genomas do projeto 1000G. Foram selecionados aleatoriamente três genomas representando cada uma das 22 populações disponíveis para o mapeamento de RNA-Seq de amostras da linhagem H1975, de seis pacientes não fumantes e de 14 pacientes fumantes de câncer de pulmão. Identificamos deleções de até 100 nucleotídeos que também foram identificadas utilizando o programa Varscan com o genoma de referência GRCh37/hg19. Entre elas, podemos citar uma deleção de 15 nucleotídeos no domínio tirosina kinase localizado no éxon 19 do gene *EGFR* em amostra tumoral de um paciente não fumante. Esta deleção também está previamente anotada na base de dados dbSNP. Em outro paciente não fumante foi encontrada uma deleção de quatro nucleotídeos que alterou o quadro de leitura do gene *CTSA* em amostras normais e tumorais. Encontramos a via de interferon  $\gamma$  com alta probabilidade de estar alterada ao analisar os dados de pacientes fumantes, entre os genes dessa via em que encontramos alterações estão o *INFR1* e o *INFR2*. Encontramos deleções a partir do mapeamento dos transcritos com 66 genomas do projeto 1000G que não encontramos com o mapeamento contra o genoma de referência GRCh37/hg19. Dentre elas, encontramos uma pequena deleção de 13 nucleotídeos no gene *EPDR1* que altera o domínio ependimina da proteína codificada. Além desta deleção, uma perda de sete nucleotídeos foi encontrada ao utilizar genomas de ancestralidade europeias como referência. No gene supressor tumoral *CDH1*, encontramos uma pequena deleção de dois nucleotídeos em pacientes não fumantes usando genomas de ancestralidade europeias e americanas. O oncogene *AKT1* apresentou uma deleção de cinco nucleotídeos em três pacientes fumantes. Estes achados poderão ser utilizados para o melhor entendimento da biologia do câncer de pulmão para novos métodos de diagnósticos e tratamentos.

THE IMPACT OF GENOMIC SMALL DELETIONS TO PROTEIN DOMAINS IDENTIFIED USING  
RNA-SEQ DATA FROM LUNG ADENOCARCINOMA PATIENTS

ABSTRACT

PHD THESIS IN COMPUTATIONAL AND SYSTEMS BIOLOGY, IOC/FIOCRUZ

Gabriel Wajnberg

Insertions and deletions (INDELs) are examples of genomic changes in the DNA sequence. One main effect of those deletions is the change of the coding region of proteins with the potential to alter the encoded amino acids and cause frameshift mutations. These deletions can cause substantial modifications in proteins involved with the biology of cancer. We innovated in the use of an already existing method, which uses ternary matrices to identify genomic deletions with the use of high-throughput transcriptome data (RNA-Seq). RNA-Seq data were used to identify small deletions up to 100 nucleotides in length within human exons. Here, we present data from the analysis of the human reference genome GRCh37/hg19 and 66 genomes from the 1000G project. We randomly selected three genomes representing each of the 22 available populations for mapping RNA-Seq data of the H1975 cell line, and samples from six nonsmoker patients and from 14 smokers patients with lung cancer. We identified deletions of up to 100 nucleotides that were also identified using the Varscan software when mapping to the reference genome GRCh37/hg19. We identified a small deletion of 15 nucleotides in tyrosine kinase domain in exon 19 in the *EGFR* gene in a tumor sample from a non-smoker patient. This deletion has been already present in the dbSNP database. In another non-smoker patient, a four-nucleotide deletion was found that caused frameshift mutation in the *CTSA* gene simultaneously in normal and tumor samples. Meanwhile, we found the interferon- $\gamma$  pathway with a high probability of being altered when analyzing the data of smoking patients, among the genes in which we found changes are *INFG1* and *INFG2*. We identified deletions when we mapped the transcripts with 66 genomes of the 1000G project, which we did not find while mapping with the reference genome GRCh37/hg19. We identified a small deletion of 13 nucleotides in the *EPDR1* gene that alters the ependymal domain of the encoded protein. In addition to this deletion, a loss of seven nucleotides was identified when using European ancestral genomes as a reference sequence. We also identified a small deletion in the *CDH1* tumor suppressor of two nucleotides in nonsmoking patients using European and American ancestry genomes and a deletion of five nucleotides in three smoking patients on the *AKT1* oncogene. In conclusion, we could use transcriptome data mapped on different genomes of different ancestry to identify small deletions of up to 100 nucleotides. These findings may be used to improve the lung cancer biology knowledge to develop new diagnosis and treatment methods.

## SUMÁRIO

<b>RESUMO</b>	<b>IX</b>
<b>ABSTRACT</b>	<b>X</b>
<b>1 INTRODUÇÃO</b>	<b>26</b>
1.1 O projeto genoma humano e as mutações no DNA .....	26
1.2 Mutações ligadas a doenças genéticas.....	28
1.3 O Câncer .....	29
1.4 Câncer de pulmão .....	30
1.5 Associação de deleções genômicas com o câncer .....	31
1.6 Bases de dados públicos de INDELS e de corridas de HTS .....	32
1.7 Justificativa.....	34
<b>2 OBJETIVOS</b>	<b>36</b>
2.1 Objetivo Geral.....	36
2.2 Objetivos Específicos .....	36
<b>3 MATERIAL E MÉTODOS</b>	<b>37</b>
3.1 Obtenção dos genomas de referência e transcritos do RefSeq para a construção das matrizes .....	37
3.2 Obtenção dos dados de RNA-Seq .....	38
3.3 Filtro dos “reads” e mapeamento .....	43
3.4 Montagem usando o Trinity.....	43
3.5 Identificação de pequenas deleções .....	44
3.6 Avaliação do impacto na sequência de proteínas.....	47
<b>4 RESULTADOS E DISCUSSÃO</b>	<b>48</b>
4.1 Identificação de pequenas deleções no genoma GRCh37/hg19 .....	48
4.1.1 Prova de conceito utilizando a linhagem celular H1975.....	48
4.1.2 Pequenas deleções identificadas em amostras de tecidos normais adjacentes ao tumor e tumorais de seis pacientes não fumantes de câncer de pulmão.....	55
4.1.3 Pequenas deleções identificadas em amostras normais e tumorais de 14 pacientes fumantes de câncer de pulmão.....	70

4.1.4	Análise de deleções encontradas exclusivamente em amostras normais, tumorais ou presente em ambas .....	83
<b>4.2</b>	<b>Identificação de pequenas deleções utilizando 66 genomas do 1000G.....</b>	<b>84</b>
4.2.1	Em amostras da linhagem celular H1975 .....	85
4.2.2	Pequenas deleções identificadas em amostras normais e tumorais de seis pacientes não fumantes de câncer de pulmão .....	90
4.2.3	Pequenas deleções identificadas em amostras normais e tumorais de 14 pacientes fumantes de câncer de pulmão.....	99
4.2.4	Análise de deleções encontradas exclusivamente em amostras normais, tumorais ou presente em ambas .....	106
<b>4.3</b>	<b>Comparação dos dados de pacientes fumantes e pacientes não fumantes .....</b>	<b>108</b>
4.3.1	Pequenas deleções identificadas a partir do genoma de referência GRCh37/hg19.....	108
4.3.2	Pequenas deleções identificadas a partir de 66 genomas do 1000G.....	110
<b>5</b>	<b>CONCLUSÕES</b>	<b>111</b>
<b>6</b>	<b>REFERÊNCIAS BIBLIOGRÁFICAS</b>	<b>113</b>
<b>7</b>	<b>ANEXOS</b>	<b>128</b>
7.1	Using high-throughput sequencing transcriptome data for INDEL detection: challenges for cancer drug discovery .....	128

Índice de Figuras	
Figura 1.1 Comparação entre DNA-Seq e RNA-Seq para a identificação de INDELS em amostras de câncer usando dados públicos disponíveis do TCGA, SRA e EGA (Adaptado de Wajnberg et al.,2016).....	34
Figura 3.1 Representação de transcritos do Refseq e gerado a partir da montagem do Trinity dispostos na matriz ternária. A) Representação do alinhamento dos dois transcritos, sendo o transcrito do Trinity contendo uma deleção no éxon 2, contra um gene do genoma de referência humano GRCh37/hg19. B) Deleção em vermelho representada com um zero nas matrizes ternárias.....	45
Figura 3.2 Esquema representando a identificação de pequenas deleções nas matrizes ternárias.....	46
Figura 4.1 Pequenas deleções identificadas por nossa metodologia utilizando as matrizes ternárias (verde) e utilizando o Varscan na amostra de RNA-Seq de H1975 (SRR1706863 e SRR1706864). ....	49
Figura 4.2 Pequena deleção de um nucleotídeo no gene MGLL visualizada no programa IGV.....	51
Figura 4.3 Pequena deleção de um nucleotídeo no gene MORF4L2 visualizada no programa IGV. ....	51
Figura 4.4 Pequena perda de um nucleotídeo no gene NUP50 visualizada no programa IGV.....	52
Figura 4.5 Pequena perda de um nucleotídeo no gene NAP1L4 visualizada no programa IGV.....	52
Figura 4.6 Pequena deleção de 21 nucleotídeos no gene CDC42EP1 visualizada no programa IGV. ....	53
Figura 4.7 Representação do impacto da pequena deleção de 21 nucleotídeos no gene CDC42EP1 na sequência de aminoácidos. A proteína normal (A) codificada pelo gene possui 391 aminoácidos, com dois domínios: CRIB (Domínio de ligação PAK, acesso cl00113) e BORG_CEP (Domínio de ligação de Rho GTPases. A proteína afetada (B) pela deleção mostra um encurtamento de sete aminoácidos.....	53
Figura 4.8 Pequena deleção de 17 nucleotídeos no gene SETD7 visualizada no programa IGV.....	54
Figura 4.9 Pequenas deleções identificadas utilizando as matrizes ternárias (verde) e utilizando o Varscan (azul) em amostras de tecido normal de RNA-	

Seq de seis pacientes de câncer de pulmão (Kim et al., 2013a). Entre parênteses a quantidade de deleções identificadas em mais de um paciente..56

Figura 4.10 Pequenas deleções identificadas utilizando as matrizes ternárias (verde) e utilizando o Varscan (azul) em amostras de tecido tumoral de RNA-Seq de seis pacientes de câncer de pulmão (Kim et al., 2013a). Entre parênteses a quantidade de deleções identificadas em mais de um paciente..56

Figura 4.11 Pequena deleção de 12 nucleotídeos visualizada na amostra normal do paciente 1 (P1N) no exon 19 do gene INF2. ....59

Figura 4.12. Representação do impacto da pequena deleção de 12 nucleotídeos no gene INF2 na sequência de aminoácidos. A proteína normal (A) codificada pelo gene possui 1249 aminoácidos, com quatro domínios: Drf\_GBD (Domínio de ligação de GTPase translúcido, acesso cl05720), Drf\_FH3 (Domínio FH3 translúcido, acesso pfam06367), Superfam\_FH2 (Domínio FH2, acesso cl19758) e Superfam\_ICP4\_C (Região C-terminal ICP4-like, acesso cl28033). A proteína afetada (B) pela deleção mostra um encurtamento de quatro aminoácidos. ....59

Figura 4.13 Pequena deleção de 15 nucleotídeos visualizada na amostra tumoral do paciente 1 (P1T) no exon 19 do gene EGFR. ....60

Figura 4.14 Representação do impacto da pequena deleção de 15 nucleotídeos no gene EGFR na sequência de aminoácidos. A proteína normal (A) codificada pelo gene possui 1210 aminoácidos, com seis domínios: dois domínios Dominio\_Rec\_L (domínios de receptores L, acesso pfam01030), Furin-like (Região rica em cisteína furina-like, acesso pfam00757), GF\_recep\_IV (Domínio fator de crescimento IV, acesso pfam14843), TM\_ErbB1 (Domínio Transmembrana ErbB1, acesso cd12093) e PTKc\_EGFR (Domínio catalítico tirosina quinase, acesso cd05108). A proteína alterada (B) pela deleção mostra o domínio tirosina quinase com menos cinco aminoácidos em sua estrutura. ...61

Figura 4.15 Pequena deleção de 10 nucleotídeos na amostra tumoral do paciente 4 (P4T) do gene TP53BP2.....62

Figura 4.16 Representação do impacto da pequena deleção de 10 nucleotídeos no gene TP53BP2 na sequência de aminoácidos. A proteína normal (A) codificada pelo gene possui 1134 aminoácidos, que inclui quatro domínios: SMC\_N (Domínio N-terminal SMC, acesso cl25732), Atrofina\_1 (Família atrofina-1, acesso cl26464), Rep\_ANK (Repetições anquirina, acesso cd00204), SH3 (Domínio de apoptose e de ligação de p53, acesso cd11953).A proteína

alterada (B) pela deleção possui menos aminoácidos por causa da mudança do quadro de leitura levando a perder diversos domínios da proteína.....	62
Figura 4.17 Pequena deleção de um nucleotídeo nas amostras tumorais do paciente 1 (P1T) e paciente 3 (P3T) do gene IFNGR1.....	64
Figura 4.18 Pequena deleção de um nucleotídeo nas amostras tumorais do paciente 1 (P1T) e paciente 3 (P3T) do gene BAMBI.....	65
Figura 4.19 Pequena deleção de um nucleotídeo nas amostras tumorais do paciente 1 (P1T) do gene PTEN.....	66
Figura 4.20 Pequenas deleções identificadas utilizando as matrizes ternárias (verde) e utilizando o VarScan (azul) encontradas, ao mesmo tempo, em amostras de tecido normal e tumoral de RNA-Seq de seis pacientes de câncer de pulmão (Kim et al., 2013a). .....	67
Figura 4.21 Pequena deleção de 4 nucleotídeo visualizada em amostras normais e tumorais do paciente 8 no gene CTSA. ....	69
Figura 4.22 Representação do impacto da pequena deleção de quatro nucleotídeos no gene CTSA na sequência de aminoácidos. A proteína normal (A) codificada pelo gene possui um domínio de Peptidase_S10 (carboxipeptidase, acesso pfam00450) enquanto a proteína afetada (B) pela deleção mostra este domínio perdido. ....	69
Figura 4.23. Pequena deleção de 1 nucleotídeo visualizada em amostras normais e tumorais de todos pacientes no gene EIF3A. ....	70
Figura 4.24 Pequenas deleções identificadas utilizando as matrizes ternárias (verde) e utilizando o VarScan (azul) encontradas em amostras de tecido normal adjacente ao tumor de RNA-Seq de 14 pacientes de câncer de pulmão do TCGA. Entre parênteses o valor de deleções identificadas em mais de um paciente.....	71
Figura 4.25 Pequenas deleções identificadas utilizando as matrizes ternárias (verde) e utilizando o VarScan (azul) encontradas em amostras de tecido tumoral de RNA-Seq de 14 pacientes de câncer de pulmão do TCGA. Entre parênteses o valor de deleções identificadas em mais de um paciente. ....	71
Figura 4.26. Pequena deleção de um nucleotídeo visualizada em amostra normal do paciente 6145 no gene VCP.....	75
Figura 4.27 Representação do impacto da pequena deleção de um nucleotídeo no gene VCP na sequência de aminoácidos. A proteína normal (A) possui 806 aminoácidos e um domínio TIP49 (Domínio C-Terminal TIP49, acesso cl27568).	



A proteína afetada (B) pela deleção perdeu 129 aminoácidos inclusive uma parte do domínio TIP49.....	75
Figura 4.28 Pequena deleção de um nucleotídeo visualizada em amostra tumoral do paciente 2655 no gene ZFP28.....	76
Figura 4.29 Representação do impacto da pequena deleção de um nucleotídeo no gene ZFP28 na sequência de aminoácidos que possui três domínios: dois domínios KRAB (Domínio associado "krueppel box", acesso smart00349) e FOG_dedo_Zn (Dedo de zinco FOG, acesso COG5048). A proteína normal (A) codificada possui sítios de ligação de dedo de zinco, enquanto a proteína alterada (B) pela deleção não possui. ....	77
Figura 4.30 Pequena deleção de um nucleotídeo visualizada em amostra tumoral do paciente 2662 e 6148 no gene IFNGR1.....	79
Figura 4.31 Pequena deleção de um nucleotídeo visualizada em amostra tumoral do paciente 2662 e 3398 no gene IFNGR2.....	79
Figura 4.32 Pequenas deleções identificadas por nossa metodologia utilizando as matrizes ternárias (verde) e utilizando o VarScan (azul) encontradas ao mesmo tempo em amostras de tecido normal e tumoral de RNA-Seq de 14 pacientes de câncer de pulmão de (Collisson et al., 2014).....	80
Figura 4.33 Pequena deleção de três nucleotídeos visualizada em amostras tumorais e normais de todos pacientes no gene VEGFC.....	82
Figura 4.34 Representação do impacto da pequena deleção de três nucleotídeos no gene VEGFC na sequência de aminoácidos que possui o domínio PDGF (Domínio de ligação ao receptor PDGFR, acesso smart00141). A proteína normal (A) codificada possui um aminoácido a mais do que a proteína alterada (B) pela deleção. ....	82
Figura 4.35 Comparação dos dois métodos de identificação de pequenas deleções no genoma de referência GRCh37/hg19 e nos 66 genomas do projeto 1000G sendo eles: utilização do Novoalign junto com o VarScan (A) e a utilização do Novoalign e Trinity para a montagem dos transcritos que depois são mapeados utilizando o BLAT e guardados nas matrizes ternárias (B). ....	85
Figura 4.36 Gráfico mostrando a proporção de pequenas deleções identificadas em linhagens celulares H1975 em regiões não codificadoras (azul), que não alteram o quadro de leitura (laranja) e que alteram o quadro de leitura (cinza). ....	86

Figura 4.37. Deleção de nove nucleotídeos no gene DENND4B em um transcrito (em amarelo) montado a partir de corridas de linhagem celular H1975 ao mapear contra um genoma asiático (HG02402) demonstrando a perda de três glutaminas (demarcado vermelho).....	87
Figura 4.38. Deleção de 13 nucleotídeos em um transcrito (em amarelo) montado a partir de dados da linhagem celular H1975 ao mapear contra um genoma europeu (HG00151) no gene EPDR1 (demarcado em vermelho) causando uma mutação que altera o quadro de leitura da tradução da proteína. ....	88
Figura 4.39 Representação do impacto da pequena deleção de 13 nucleotídeos no gene EPDR1 na sequência de aminoácidos. A proteína normal (A) codificada pelo gene possui um domínio Ependimina (Domínio ependimina, acesso pfam00811) enquanto a proteína afetada (B) pela deleção mostra este domínio perdido. ....	88
Figura 4.40 Deleção de sete nucleotídeos em um transcrito (em amarelo) montado a partir de dados de linhagem celular H1975 ao mapear contra um genoma europeu (HG00114) no gene PTK2 (demarcado em vermelho) causando uma mutação que altera o quadro de leitura da tradução. ....	89
Figura 4.41 Representação do impacto da pequena deleção de sete nucleotídeos no gene PTK2 na sequência de aminoácidos. A proteína normal (A) codificada pelo gene possui quatro domínios: B41 (Domínio ERM ou homólogos banda 4.1, acesso smart00295), Ferm_C_FAK (Domínio kinase FERM de adesão focal, acesso cd13190), PTKc_FAK (Domínio catalítico tirosina kinase de adesão focal, acesso cd05056) e Focal_AT (Região alvo de adesão focal, pfam03623). A proteína afetada (B) pela deleção perde uma porção do domínio final.....	90
Figura 4.42 Pequenas deleções identificadas com o uso das matrizes ternárias em tecido normal (azul) e em tecido tumoral (verde) de seis pacientes de câncer de pulmão (Kim et al., 2013a). Entre parênteses a quantidade de deleções identificadas em mais de um paciente.....	91
Figura 4.43 Deleção de 75 nucleotídeos em um transcrito (em amarelo) montado a partir de dados de amostras normais do paciente 5 ao mapear contra um genoma espanhol (HG01501) no gene MCM7 (demarcado em vermelho) causando uma mutação que adianta o códon de parada da tradução. ....	92

Figura 4.44 Representação do impacto da pequena deleção de 75 nucleotídeos no gene MCM7 na sequência de aminoácidos. A proteína normal (A) codificada pelo gene possui dois domínios: MCM_N (Domínio N-Terminal MCM, acesso pfam14551) e MCM (Proteína de manutenção minicromossomo, acesso smart00350) enquanto a proteína afetada (B) pela deleção foi encurtada. ....	93
Figura 4.45 Deleção de dois nucleotídeos em um transcrito (em amarelo) montado a partir de dados de amostras tumorais do paciente 4 ao mapear contra um genoma espanhol (HG01602) no gene CDH1 (demarcado em vermelho) causando uma mutação que altera o quadro de leitura da tradução. ....	94
Figura 4.46 Representação do impacto da pequena deleção de dois nucleotídeos no gene CDH1 na sequência de aminoácidos. A proteína normal (A) codificada pelo gene possui seis domínios: Pro_caderina (Prodomínio caderina-like, acesso pfam08758), CA_like (Domínio de repetição caderina-like, acesso cd00031), três domínios Rep_caderina (Domínio de repetição de caderina em tandem, acesso cd11304) e Domínio_caderina_C (Região citoplasmática caderina, acesso pfam01049) enquanto a proteína afetada (B) pela deleção um encurtamento com esses domínios perdidos.....	95
Figura 4.47 Deleção de cinco nucleotídeos em um transcrito (em amarelo) montado a partir de dados de amostra tumoral do paciente 3 ao mapear contra um genoma de ancestralidade europeia (HG01501) no gene ENGASE (demarcado em vermelho) causando uma mutação que altera o quadro de leitura da tradução. ....	96
Figura 4.48 Representação do impacto da pequena deleção de cinco nucleotídeos no gene ENGASE na sequência de aminoácidos. A proteína normal (A) codificada pelo gene possui um único domínio com sítios ativos GH85_Engase (Domínio Engase, acesso cd06547) enquanto a proteína afetada (B) pela deleção possui um encurtamento com esse domínio afetado, incluindo os sítios ativos .....	96
Figura 4.49 Deleção de cinco nucleotídeos em um transcrito (em amarelo) montado a partir de dados de amostras normais e tumorais dos pacientes 3 e 4 ao mapear contra um genoma de ancestralidade africana (HG02282) no gene SIRPB1 (demarcado em vermelho) causando uma mutação que altera o quadro de leitura da tradução. ....	97

Figura 4.50 Representação do impacto da pequena deleção de cinco nucleotídeos no gene SIRPB1 na sequência de aminoácidos. A proteína normal (A) codificada pelo gene possui três domínios de imunoglobulina: IgV_SIRPT (Domínio Imunoglobulina-like SIRP, acesso cd16097), Dominio_2_IgC_SIRP (Domínio imunoglobulina-like SIRP 2, acesso cd05772) e Dominio_3_IgC_SIRP (Domínio imunoglobulina-like SIRP 3, acesso cd16085). A proteína afetada (B) pela deleção tem um encurtamento. ....	98
Figura 4.51 Pequenas deleções identificadas ao utilizar as matrizes ternárias em tecido normal (azul) e em tecido tumoral (verde) de seis pacientes de câncer de pulmão (Collisson et al., 2014). Entre parênteses a quantidade de deleções identificadas em mais de um paciente.....	99
Figura 4.52 Deleção de 12 nucleotídeos em um transcrito (em amarelo) montado a partir de dados de amostra normal do paciente 3398 ao mapear contra um genoma de ancestralidade asiática (HG01869) no gene HCLS1 (demarcado em vermelho).....	100
Figura 4.53 Representação do impacto da pequena deleção de 12 nucleotídeos no gene HCLS1 na sequência de aminoácidos. A proteína normal (A) possui 486 aminoácidos e cinco domínios: três domínios Rep_HS1 (Repetição HS1, acesso pfam02218), Superfam_GGN (Gametogenetina, acesso cl25800) e SH3_HS1 (Domínio SH3, acesso cd12073) e a proteína afetada (B) pela deleção possui quatro nucleotídeos a menos.....	101
Figura 4.54 Deleção de 75 nucleotídeos em um transcrito (em amarelo) montado a partir de dados de amostra tumoral do paciente 3398 ao mapear contra um genoma de ancestralidade africana (HG01879) no gene GPX8 (demarcado em vermelho) causando uma mutação que altera o quadro de leitura da tradução. ....	102
Figura 4.55 Representação do impacto da pequena deleção de 75 nucleotídeos no gene GPX8 na sequência de aminoácidos. A proteína normal (A) possui 248 aminoácidos e o domínio Tioredoxina_like (Peroxidase GPX7 da superfamília tioredoxina-like, acesso TIGR02540) e a proteína afetada (B) pela deleção causou a perda de seu principal domínio proteico.....	102
Figura 4.56 Deleção de cinco nucleotídeos em um transcrito (em amarelo) montado a partir de dados de amostra tumoral dos pacientes 6776, 6777 e 6778 ao mapear contra um genoma de ancestralidade asiática (HG00956) no gene	

AKT1 (demarcado em vermelho) causando uma mutação que altera o quadro de leitura da tradução. ....	103
Figura 4.57 Representação do impacto da pequena deleção de cinco nucleotídeos no gene AKT1 na sequência de aminoácidos. A proteína normal (A) possui dois grandes domínios: Superfam_PH_like (Domínio Akt, acesso cd01241) e Superfam_PKc_like (Domínio catalítico Serina/Treonina kinase, acesso cd05594). A proteína afetada (B) pela deleção possui um encurtamento que afeta o domínio proteína kinase. ....	104
Figura 4.58 Deleção de um nucleotídeo em um transcrito (em amarelo) montado a partir de dados de amostras normais e tumorais dos pacientes 5645, 6778 e 6148 ao mapear contra um genoma de ancestralidade europeia (HG00361) no gene COL3A1 (demarcado em vermelho) causando uma mutação que altera o quadro de leitura da tradução. ....	105
Figura 4.59 Representação do impacto da pequena deleção de um nucleotídeo no gene COL3A1 na sequência de aminoácidos. A proteína normal (A) possui 1466 aminoácidos e seis domínios: VWC (Domínio fator von Willebrand tipo C, acesso pfam00093), quatro domínios Colag (Repetição tripla hélice de colágeno, acesso pfam01391) e COLFI (Domínio C-terminal de colágeno fibrilar, acesso pfam01410). A proteína afetada (B) pela deleção possui um encurtamento de 19 aminoácidos na sua porção C-terminal. ....	105
Figura 4.60 Exemplo de transcritos reconstruídos pelo programa Trinity para o gene MCM7 para a amostra do tecido normal do paciente 5 (P5N) e para o tecido tumoral deste mesmo pacientes (P5T). ....	107
Figura 4.61 Exemplo de transcritos reconstruídos pelo programa Trinity para o gene AKT1 para a amostra do tecido normal do paciente 6776 (6776N) e para o tecido tumoral deste mesmo pacientes (6776T). ....	107
Figura 4.62 Exemplo de transcritos reconstruídos pelo programa Trinity para o gene GPX8 para a amostra do tecido normal do paciente 3398 (3398N) e para o tecido tumoral deste mesmo pacientes (3398T). ....	108
Figura 4.63. Pequenas deleções identificadas pelas matrizes ternárias em amostras normais de Kim e colaboradores (2013a) (azul) e utilizando amostras normais de Collisson e colaboradores (2014) (verde). ....	109
Figura 4.64 Pequenas deleções identificadas pelas matrizes ternárias em amostras tumorais de Kim e colaboradores (2013a) (azul) e utilizando amostras tumorais de Collisson e colaboradores (2014) (verde). ....	109

**Figura 4.65 Pequena deleção de um nucleotídeo no gene IFNGR1 visualizada no programa IGV em amostra tumoral de paciente fumante (2662) e em paciente não fumante (p8).....110**

## LISTA DE TABELAS

Tabela 3.1 Genomas do 1000G utilizados no mapeamento dos dados de RNA-Seq.....	38
Tabela 3.2 Amostras utilizadas de Kim e colaboradores (2013a) do banco SRA com a quantidade de “reads” de cada corrida e estatísticas da montagem do transcriptoma produzida pelo programa Trinity.....	40
Tabela 3.3 Lista de amostras utilizadas do banco de dados TCGA sequenciados do “Chritiana Healthcare Center” com a quantidade de “reads” de cada corrida e estatísticas da montagem do transcriptoma produzida pelo programa Trinity.....	41
Tabela 4.1 Deleções identificadas exclusivamente pela abordagem das matrizes ternárias em dados de RNA-Seq da linhagem H1975 e a cobertura das leituras. ....	50
Tabela 4.2. Tabela de quantas deleções estão presentes na base de dados dbSNP e COSMIC em amostras normais e tumorais. ....	57
Tabela 4.3 Deleções identificadas em dados de RNA-Seq nas amostras de Kim e colaboradores (2013a) e a cobertura das leituras. ....	58
Tabela 4.4 Enriquecimento de genes que sofreram pequenas deleções em amostras tumorais identificadas pela nossa metodologia de matrizes ternárias. Abaixo encontramos as vias do Gene Ontology (GO) com probabilidade posterior maior que 0,5 de estarem superrepresentadas. ....	63
Tabela 4.5. Deleções identificadas em dados de RNA-Seq ao mesmo tempo em amostras normais e tumorais de Kim e colaboradores (2013a) e a cobertura das leituras. ....	68
Tabela 4.6 Representação de quantas deleções estão anotadas na base de dados COSMIC em amostras normais e tumorais. ....	72
Tabela 4.7 Deleções identificadas em dados de RNA-Seq nas amostras de Collisson e colaboradores (2014) e a cobertura das leituras. ....	73
Tabela 4.8 Enriquecimento de genes que sofreram pequenas deleções em amostras tumorais identificadas pela nossa metodologia de matrizes ternárias. Abaixo encontramos as vias do Gene Ontology (GO) com probabilidade posterior maior do que 0,5 de estarem superrepresentadas . ....	78
Tabela 4.9 Deleções identificadas em dados de RNA-Seq nas amostras de Collisson e colaboradores (2014) e a cobertura das leituras. ....	81

<b>Tabela 4.10 Tabela representando as deleções identificadas e seu impacto em regiões codificadoras e presentes no dbSNP e COSMIC. ....</b>	<b>85</b>
<b>Tabela 4.11 Tabela contendo as deleções identificadas e seu impacto em regiões codificadoras e presentes no dbSNP e COSMIC em amostras normais e tumorais de pacientes não fumantes de câncer de pulmão. ....</b>	<b>91</b>
<b>Tabela 4.12 Tabela representando as deleções identificadas e seu impacto em regiões codificadoras e presentes no dbSNP e COSMIC em amostras normais e tumorais de pacientes fumantes de câncer de pulmão. ....</b>	<b>100</b>
<b>Tabela 4.13 Enriquecimento de genes que sofreram pequenas deleções em amostras tumorais identificadas pela nossa metodologia de matrizes ternárias. Abaixo as vias do Gene Ontology (GO) com probabilidade com probabilidade posterior maior do que 0,5 de estarem superrepresentadas. ....</b>	<b>106</b>



## LISTA DE SIGLAS E ABREVIATURAS

1000G	do inglês, “1000 Genomes Project”
CRISPR	do inglês, “clustered regularly interspaced short palindromic repeats”
EMC	do inglês, “enzyme mismatch cleavage”
DNA	do inglês, “deoxyribonucleic acid”
DNA-Seq	do inglês, “DNA sequencing”
EBI	do inglês, “European Bioinformatics Institute”
EGA	do inglês, “European Genome-phenome Archive”
GATK	do inglês, “Genome Analysis Toolkit”
GRC	do inglês, “Genome Reference Consortium”
HGMD	do inglês, “Human Gene Mutation Database”
HTS	do inglês, “Highthroughput Sequencing”
INCA	Instituto Nacional de Câncer
INDEL	do inglês, “insertions and deletions”
LINEs	do inglês, “long interspersed repeated sequences”
LTR	do inglês, “long terminal repeats”
MSI	do inglês, “microsatellite instability”
NSCLC	do inglês, “non-small lung cancer”
OMS	Organização Mundial da Saúde
PCR	do inglês, “polymerase chain reaction”
RFLP	do inglês, “restriction fragment length polymorphisms”
RNA	do inglês, “ribonucleic acid”

RNA-Seq	do inglês, “high-throughput transcriptome sequencing”
SINEs	do inglês, “short interspersed repeated sequences”
SCLC	do inglês, “small cell lung cancer”
SCC	do inglês, “squamous carcinoma cell”
SMRT	do inglês, “single molecule real time”
SNV	do inglês, “single nucleotide variation”
SNP	do inglês, “single nucleotide polymorphism”
SRA	do inglês, “Sequence Read Archive”
TCGA	do inglês, “The Cancer Genome Atlas”
UTR	do inglês, “untranslated region”
WGS	do inglês, “whole genome shotgun”

# 1 INTRODUÇÃO

## 1.1 O projeto genoma humano e as mutações no DNA

A publicação do primeiro rascunho do genoma humano em 2001 propiciou novas perspectivas para a análise de genes humanos, incluindo a identificação de mutações em larga escala no genoma (Lander et al., 2001; Venter et al., 2001). Naquela época, as tecnologias disponíveis possibilitavam apenas o sequenciamento de regiões pontuais do genoma produzindo poucos dados e com alto custo por base sequenciada. O consórcio do genoma de referência (GRC) liberou a uma montagem do genoma humano de referência de 13 indivíduos, a qual a última versão é GRCh38 (Church et al., 2011; 2010). Esta montagem do genoma humano é usada para realizar comparações de sequências entre a sequência de referência e um determinado alvo para detectar mutações no DNA.

Diversos tipos de mutações podem ocorrer na sequência de DNA. O primeiro tipo é definido como substituição e ocorre quando um certo nucleotídeo de uma posição cromossômica é substituída por uma das outras três bases nucleotídicas disponíveis. Estas substituições ainda podem ser classificadas por transição, quando uma purina (A ou G) é substituída por outra purina, ou uma pirimidina (C ou T) é substituída por outra pirimidina; ou então transversão, em que uma purina substitui uma pirimidina. Esta substituição de um nucleotídeo também é chamada de variante de nucleotídeo único (SNV) e quando ocorre em mais de 1% da população é classificado como um polimorfismo ou polimorfismo de nucleotídeo único (SNP) (Vignal et al., 2002).

Outro exemplo de tipo de mutações que alteraram o tamanho da sequência de DNA, são as deleções e inserções (INDELs). A deleção ocorre quando um bloco de um ou mais nucleotídeos é perdido, enquanto a inserção ocorre quando um ou mais nucleotídeos são ganhos em uma dada região da sequência do DNA (Mills et al., 2006). As INDELs podem ser ainda categorizados de acordo com o seu tamanho: INDELs pequenas, para aqueles que variam de 1 a 543 nucleotídeos de comprimento (Bhangale et al., 2005), microINDELs quando atingem um máximo de 50 nucleotídeos (Scaringe et al., 2008); e variantes estruturais, que são frequentemente identificadas em tumores e normalmente são maiores do que 10.000 nucleotídeos de comprimento (Mullaney et al., 2010). Estas INDELs maiores podem

ser até observadas nos cromossomos em um exame de cariótipo. Eventos genômicos podem ocasionar INDELS, como por exemplo, o deslizamento da enzima DNA polimerase durante a replicação genômica que ocorre na meiose (Montgomery et al., 2013), gerando uma região de nucleotídeos repetidos denominados “microssatélites” que possuem de 5 a 50 repetições de sequências de nucleotídeos no total (Ellegren, 2004). Outro tipo de evento de INDEL ocorre por ação de elementos transponíveis ou transposons. Os transposons são sequências de DNA que podem se mover de um local para o outro do genoma (Slotkin e Martienssen, 2007). Estas sequências podem ser classificadas em retrotransposons, que são transcritos em moléculas de RNA antes de serem novamente inseridas no DNA através da transcriptase reversa, ou classificadas como transposons de DNA, em que uma sequência é cortada de uma região e inserida em outra através de uma enzima chamada transposase (Wicker et al., 2007). Os retrotransposons ainda podem ser denominados como retrotransposons de longas regiões terminais (ou LTR). Aqueles que não possuem LTR podem ser classificados como SINEs (do inglês, “short interspersed repeated sequences”) que podem possuir até 500 nucleotídeos ou LINEs (do inglês, “long interspersed repeated sequences”) que podem possuir até 5.000 nucleotídeos (Weiner, 2002). Os elementos transponíveis com maior prevalência no genoma humano pertencem à família de elementos Alu, pertencentes aos retrotransposons do tipo SINE. Esta família de elementos Alu, possui mais de um milhão de cópias no genoma humano, contribuindo a 11% do código total (Deininger, 2011). Por ser a maior família de elementos transponíveis, os elementos Alu servem como núcleos de recombinação homóloga que podem causar diversas alterações genéticas, entre elas perdas genômicas (Batzer e Deininger, 2002). Sen e colaboradores (2006) identificaram perdas genômicas associadas à recombinação de elementos Alu ao comparar regiões adjacentes destes retrotransposons em genomas de chimpanzé e humano.

Podemos também enfatizar dois tipos de mutações mais complexas que afetam grandes porções da estrutura do DNA, como as inversões, que compreendem as rotações de 180° de um segmento da molécula de DNA, e as translocações, em que partes de dois cromossomos não homólogos trocam de lugar.

Estas mutações ao mudar um gene ou o cromossomo podem causar a produção incorreta de proteínas e causar diferenças no fenótipo, como algumas doenças. No entanto mutações afetando a função de genes são raras, por exemplo,

a maioria das INDELS em regiões codificadoras de genes humanos não afetam domínios conhecidos de proteínas funcionais (de la Chaux et al., 2007).

Deleções que alteram a fase de leitura do gene podem levar ao mecanismo de decaimento do mRNA mediado por mutações sem sentido (NMD), principalmente por adiantar o códon de parada por mais de 55 nucleotídeos antes da região de junção éxon-éxon (Behm-Ansmant et al., 2007).

## 1.2 Mutações ligadas a doenças genéticas

O banco de dados “Human Gene Mutation” (HGMD) relaciona mutações com desordens genéticas. Na sua última versão disponibilizada em 2016, este banco de dados possuía 107.554 (64%) SNPs não sinônimos, 42.950 (26%) deleções, 15.454 (9%) de inserções e 1.766 (1%) de rearranjos (Stenson et al., 2016).

Apesar dos SNPs serem mais abundantes, espera-se que as INDELS possuam impacto maior na estrutura e função das proteínas por terem o potencial de alterar o quadro de leitura das proteínas (Iengar, 2012). Muitos SNPs, mesmo que causem a troca de aminoácidos na proteína codificada, podem não afetar consideravelmente a estrutura da proteína, diferente das INDELS que podem ocasionar uma mudança na matriz de leitura da tradução ocasionando o truncamento prematuro da proteína traduzida ou então a tradução de aminoácidos adjacentes não presentes na proteína original. Portanto, a probabilidade de modificar sítios ativos é maior nestes casos (Ipe et al., 2017). Quando só olhamos para as INDELS, vemos uma predominância de deleções estarem associadas com doenças. Em populações com alta prevalência de doenças genéticas, como as que tiveram altas taxas de endogamia ao longo da história, são observadas INDELS (sendo que quatro são deleções) frequentes nestas populações ou nos indivíduos que descendem delas. Dentre elas, podemos citar a deleção de dois nucleotídeos no gene *BRCA1*, que é ligado a vários tipos de câncer, inclusive de mama e deleções em outros genes ligados a outras doenças, como: *GJB2*, *CCR5* e *BLM* (Ostrer, 2001). Além disso, 24% das doenças genéticas mendelianas<sup>1</sup> são exclusivamente associadas a INDELS (Stenson et al., 2012).

---

<sup>1</sup> Doenças relacionadas a mutações em apenas um gene, também chamadas de monogênicas.

Outro exemplo é a associação de uma deleção com a fibrose cística. Esta doença está associada com a perda de três nucleotídeos na região codificadora do gene CFTR com a conseqüentemente perda da função da proteína traduzida (Collins et al., 1987). Em algumas doenças mendelianas, proteínas de ligação ao RNA (RBP, do inglês “RNA binding proteins”) não podem se ligar ao seu sítio de ligação na molécula de RNA devido a deleções que ocorreram em sua sequência. Estas deleções estão descritas no banco de dados HGMD e estão localizadas em regiões próximas de sítios de “splicing”, possivelmente alterando o mecanismo de “splicing”, ou em regiões dentro de éxons em que há evidências de eventos de “splicing” alternativo (Zhang et al., 2014).

No entanto, algumas deleções podem ser localizadas em regiões não codificadoras, como as regiões não traduzidas (UTR) 3' e 5'. Há na literatura exemplos de deleções em regiões 3' UTR que alteram o sítio de ligação ao miRNA. Podemos citar duas deleções descritas na região 3' UTR do gene *IKK1* que alteram o sítio de ligação do miR-223 (Bhattacharya et al., 2012). Em contraste, os INDELS que ocorrem na 5' UTR são preditos para alterar os motivos de iniciação de tradução, “upstream” do códon de início (uAUGs) e “upstream” da fase de leitura aberta (uORFs) (Chen et al., 2011). Por exemplo, duas deleções na região 5' UTR no gene DJ-1 foram identificadas num grande número de pacientes com Doença de Parkinson (Glanzmann et al., 2014).

Os genes afetados estão associados ao crescimento tumoral e criação de mecanismos de sobrevivência celular destes tumores (Greenman et al., 2007; Stratton et al., 2009). A seguir será revisada a associação de INDELS e câncer.

### **1.3 O Câncer**

O câncer, palavra usualmente utilizada para designar tumores malignos, não é uma terminologia que se refere a uma única patologia, mas sim a uma diversidade de patologias cujo processo carcinogênico, os fatores de risco associados, bem como o grau de malignidade podem ser completamente distintos (Weinberg, 2006). Um câncer é caracterizado pelo crescimento anormal de células (neoplasia), que formam uma massa de células (o neoplasma), comumente chamado de tumor (Kumar et al., 2010). Da mesma forma que não existe apenas um tipo da doença, também não existe uma causa única. As causas para o surgimento do câncer podem ser divididas entre fatores internos e externos. Dentre os fatores internos

temos: mutações herdadas, hormônios e condições imunes; e dentre os fatores externos: fumo, dieta, radiação, e organismos infecciosos (Anand et al., 2008).

A nomenclatura dos tumores é proveniente da origem tecidual das células de origem epitelial, por exemplo, originam os tumores humanos mais comuns existentes, os carcinomas. Há ainda alguns carcinomas específicos relacionados ao tipo de tecido epitelial como, por exemplo, aqueles formados por células epiteliais que selam cavidades para proteger células adjacentes. Tal tipo de tumor é denominado carcinoma de células escamosas. Como também existem células epiteliais que secretam substâncias nas cavidades que as revestem, tumores originados por estas células são denominados adenocarcinomas (Kumar et al., 2010).

Há ainda tumores que são originados a partir de tecidos não epiteliais, dentre eles temos: sarcomas (a partir de tecidos conjuntivos); leucemias (derivados de diversas linhagens de células hematopoiéticas); linfomas (surgem a partir de linhagens celulares linfoides, como linfócitos B e T) e tumores originados por componentes do sistema nervoso central (como gliomas, glioblastomas, neuroblastomas, schwannomas e meduloblastomas) (Weinberg, 2006).

## **1.4 Câncer de pulmão**

Na população brasileira, o câncer de pulmão foi estimado como o segundo tipo de malignidade mais incidente entre os homens (8,1% dos casos câncer) e o quarto mais incidente entre as mulheres (5,1% dos casos de câncer) no biênio de 2016-2017 (Instituto Nacional de Cancer José Alencar Gomes da Silva, 2016). Ademais de acordo com o relatório de 2014 da Organização Mundial de Saúde (OMS), o câncer de pulmão é aquele que possui maior incidência entre homens no mundo (16,7% dos casos de câncer) e o terceiro com maior incidência entre as mulheres (8,7% dos casos de câncer) (Stewart e Wild, 2014).

Entre 90 e 95% dos casos de tumores em pulmão consistem em carcinomas. Os carcinomas de pulmão possuem uma grande influência do tabagismo. Cerca de 87% dos pacientes com este tipo de câncer são fumantes ativos ou aqueles que pararam recentemente (Kumar et al., 2010). Além do tabagismo, outros fatores de risco desse tipo de câncer são: riscos industriais (trabalhadores envolvidos com

material radioativo ou asbestos<sup>2</sup>), poluição do ar e genética molecular (mutações em genes como *EGFR*, *KRAS*, *MYC* e *KIT*).

Os carcinomas podem ser classificados em dois tipos: o câncer de pulmão de células pequenas (SCLC) e o câncer de pulmão de células não pequenas (NSCLC). Há três tipos de tumores do tipo NSCLC: o adenocarcinoma, carcinoma de células escamosas (SCC) e o carcinoma de células grandes. Os tipos mais comuns são o adenocarcinoma e o SCC entre homens e mulheres (Travis et al., 2004).

## 1.5 Associação de deleções genômicas com o câncer

O gene *TP53* é frequentemente alvo de mutações em câncer, incluindo INDELS. A proteína p53 é codificada por este gene e desempenha função de reparo do DNA. Uma inserção de 16 nucleotídeos no íntron 3 do gene *TP53* aumenta o risco de câncer colorretal (Gemignani et al., 2004). Em outro gene, *GAS5*, há uma deleção de cinco nucleotídeos que está associada ao aumento de risco de carcinoma hepatocelular na população chinesa (Tao et al., 2015). A instabilidade genômica é outra característica comum detectada em tumores. Por exemplo, alguns subtipos de tumores colorretais são causados por um fenômeno conhecido como instabilidade de microssatélites (MSI) capaz de ocasionar mutações, em que as células perdem a capacidade de reparo de pareamentos errados do DNA (Boland e Goel, 2010; Williams et al., 2010). Por exemplo, a ocorrência de MSI já foi associada à identificação de uma mutação na região 3' UTR do gene *EGFR* no câncer colorretal (Yuan et al., 2009). MicroINDELS também foram identificados no gene *TP53* com frequências semelhantes às de outros tumores humanos como na mama, bexiga, colorretal, ovário, boca, pulmão e estômago (Scaringe et al., 2008).

O número e a variedade de mutações em tumores é enorme e neste contexto foi criado o Catálogo de Mutações Somáticas em Câncer (base de dados COSMIC) (Forbes et al., 2008). A base de dados COSMIC armazena todas as mutações somáticas identificadas nos genes e as associa a diferentes tipos de informação, tais como a droga alvo e o tumor associado. Um destes genes é o *EGFR*, que compreende um grande número de mutações associados a tumores. Devido à forte associação entre as alterações do *EGFR* e a ocorrência de NSCLC, foi criada uma

---

<sup>2</sup> Material de silicato utilizado em construções de residências e comércios.



base de dados específica para este gene: a base de dados de mutações do gene *EGFR* (Gu et al., 2007).

## **1.6 Bases de dados públicos de INDELS e de corridas de HTS**

Até o amplo uso da tecnologia de sequenciamento de alta vazão (HTS), utilizava-se bases de dados como o dbSNP (do inglês “Single Nucleotide Polymorphism Database”) (Wheeler et al., 2007) para comparar SNPs e INDELS identificados em algum experimento com aqueles já descritos na literatura. Esta base de dados de polimorfismos em sequências de DNA foi criada em 1998 como um suplemento do Genbank do NCBI.

Em 2008, foi lançado o projeto “1000 Genomes” (1000G) com a missão de lançar o mais detalhado catálogo de variações em genomas humanos (The 1000 Genomes Consortium, 2010) ao sequenciar o genoma de seres humanos de diferentes populações mundiais.

Até o mês de Outubro de 2015, o referido projeto sequenciou 2.500 genomas de humanos de diferentes populações na sua primeira fase (The 1000 Genomes Consortium, 2012), o que possibilitou a busca de alterações genômicas (como INDELS e SNPs) não apenas relacionadas à sequência do genoma de referência para espécie humana.

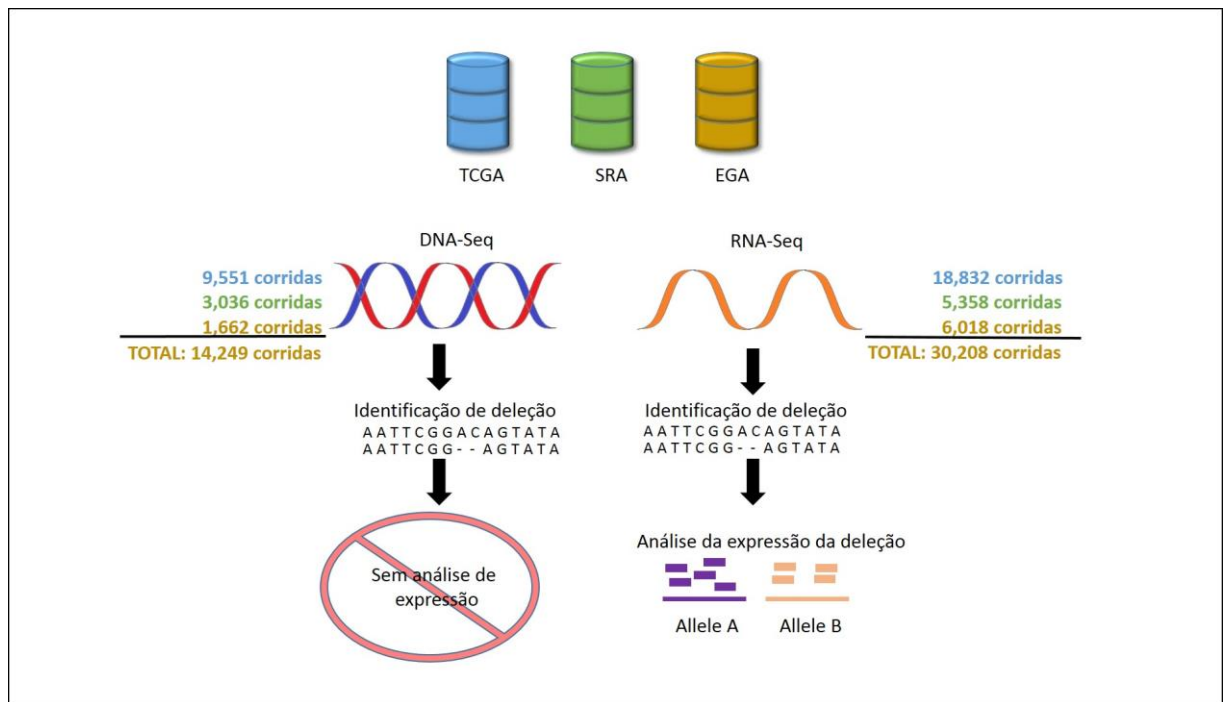
A imensa geração de dados de HTS permitiu a criação de diferentes bancos de dados públicos para que estes dados possam ser armazenados e acessados. Este tipo de iniciativa leva à possibilidade de novas descobertas em dados já analisados. A primeira base de dados deste tipo foi o SRA (do inglês, “Sequence Read Archive”) disponível pelo site do NCBI e é uma colaboração entre esta instituição, o EBI (do inglês “European Bioinformatics Institute”) e do DDBJ (do inglês “DNA data bank of Japan”) (Wheeler et al., 2008).

No entanto o EBI resolveu de certa forma criar uma base de dados que o permitisse controlar os acessos a eles. Por isso criou o EGA (do inglês “European Genome-phenome Archive”), e para acessar estes dados os usuários precisam se registrar e pedir permissão detalhando suas intenções (Lappalainen et al., 2015).

O SRA e o EGA possuem dados de HTS de diversas estratégias e organismos. Porém, uma nova iniciativa surgiu em 2013 com a intenção de guardar dados apenas de tumores em seres humanos, conhecido como “The Cancer Genome

Atlas” (TCGA) (Cancer Genome Atlas Research Network and Cancer Genome Atlas Research Network, 2013). Análises adicionais de dados de exoma e WGS de amostras de tumores de câncer colorretal e endometrial disponibilizados deste banco de dados contribuiu para a identificação de novos INDELS (Kim et al., 2013b).

Ao longo do tempo, começaram a usar mais o RNA-Seq em detrimento de dados genômicos para identificar INDELS (Kim e Park, 2014; Nielsen et al., 2011). A vantagem do uso de dados RNA-Seq se dá pelo fato que os dados de transcriptoma são mais abundantes do que os dados sobre as sequências genômicas disponíveis em bancos de dados públicos. Outro aspecto importante do uso de dados de transcriptoma é que, além de ser possível identificar alterações genômicas, a expressão destas variações pode ser analisada (Wajnberg e Passetti, 2016). Estimativas de 2015 mostram que a base de dados TCGA possuía 18.832 corridas RNA-Seq e 9.551 de DNA-Seq, o banco de dados SRA possuía 5.358 corridas de RNA-Seq e 3.036 de DNA-Seq, enquanto o banco de dados de EGA possuía 6.018 corridas de RNA-Seq e 1.662 de DNA-Seq (1.1). Portanto, há mais do que o dobro de dados de RNA-Seq em relação aos dados de DNA-Seq nas bases de dados acima mencionadas. Uma comparação entre uma mesma amostra analisada com WGS e RNA-Seq demonstrou que o uso de RNA-Seq pode identificar 81% das mutações identificadas utilizando WGS (Cirulli et al., 2010). Pelo exposto, cada vez mais esta tecnologia tem sido usada para buscar INDELS.



**Figura 1.1** Comparação entre DNA-Seq e RNA-Seq para a identificação de INDELS em amostras de câncer usando dados públicos disponíveis do TCGA, SRA e EGA (Adaptado de Wajnberg et al., 2016).

## 1.7 Justificativa

Os íntrons são geralmente sequências longas que são removidas durante o evento de “splicing” após a transcrição de um gene (Lander et al., 2001). Wang e Yu (2011) analisaram o genoma sequenciado de populações do 1000G e determinaram que o tamanho mínimo de um íntron é de 87 pares de bases. Esses íntrons mínimos foram identificados nas populações Yorubá, européia, chinesa Han e japonesa. O nosso grupo construiu um banco de dados de variantes de “splicing” (Tavares et al., 2014) a partir do banco de dados de sequências de dados de sequenciamento de transcriptomas. As sequências expressas foram organizadas usando a metodologia desenvolvida pelo nosso grupo denominada matrizes ternárias e agrupadas em transcritos hipotéticos. Nestes dados, foram identificadas possíveis perdas de até 100 nucleotídeos nas variantes de transcritos em porções centrais de éxons conhecidos, podendo ser classificados também como potenciais íntrons. Neste banco de dados, identificamos aproximadamente 56 mil eventos potenciais de perdas pequenas, sendo 63% delas com potencial para alterar a fase de leitura das proteínas. Na literatura, há ainda uma discussão se esses tipos de eventos seriam microéxons ou deleções (LaFlamme, 2015). Simultaneamente, em resultados

preliminares, identificamos 51 casos de perdas pequenas que ocorrem dentro de éxons humanos usando a base de dados do dbSNP (Wheeler et al., 2007). Pequenas deleções em regiões codificadoras de genes humanos já são conhecidas em genomas de diversas populações (Mills et al., 2006) e em alguns tipos de câncer, como por exemplo: ovário (Jones et al., 2010), glioblastoma (Parsons et al., 2008) e meduloblastoma (Parsons et al., 2011).

Não obstante, perdas pequenas em regiões nos éxons podem alterar domínios funcionais de proteínas ocasionando a alteração de sua função. Realizamos uma revisão bibliográfica e não foram encontrados dados que relacionam deleções genômicas a partir da análise de dados de transcriptoma nas diversas populações de seres humanos, apesar de existirem mais dados de transcriptoma disponíveis e publicados do que de genoma. Além disso, não há nenhum estudo que avalie pequenas alterações de sequências a partir deste tipo de mutação. Assim, pretendemos avaliar o impacto destas pequenas deleções na sequência das proteínas codificadas, no intuito de investigar domínios proteicos interrompidos.

Metodologicamente, há o desafio de se trabalhar com deleções encontradas em dados de RNA-Seq utilizando o genoma humano de referência e a identificação de novas deleções usando dados de anotação de polimorfismos de outros genomas oriundos do projeto “1000 Genomes” (1000G). O 1000G sequenciou o genoma e fez a anotação de polimorfismos de pessoas saudáveis de diversas populações humanas. Atualmente, os programas desenvolvidos para a identificação de deleções em dados de RNA-Seq não conseguem fazer essa identificação nos dados do 1000G. Assim, como desafio em termos de bioinformática, empregaremos a abordagem das matrizes ternárias para identificação de deleções em dados de RNA-Seq usando o genoma humano de referência, bem como a identificação usando dados de outros genomas humanos oriundos do 1000G. A metodologia das matrizes ternárias já foi empregada com sucesso para a identificação de variantes por “splicing” alternativo no transcriptoma e proteoma humano (Tavares et al., 2017; Tavares et al., 2014). Porém, o seu uso para detecção de deleções é ainda inédito.

## **2 OBJETIVOS**

### **2.1 Objetivo Geral**

Avaliar o possível impacto de pequenas deleções no genoma humano sobre as sequências de proteínas codificadas usando dados de RNA-Seq de câncer de pulmão.

### **2.2 Objetivos Específicos**

- Identificar deleções usando dados de RNA-Seq de câncer de pulmão, tentando as classificar como somáticas ou potenciais candidatas a germinativas;
- Comparar a metodologia de matrizes ternárias com o programa VarScan;
- Reconstruir três genomas a partir das deleções catalogadas de cada população do projeto “1000 Genomes” (1000G);
- Traduzir as sequências dos genes com deleções identificadas e avaliar o impacto destas alterações na sequência das proteínas; e
- Associar os polimorfismos identificados com as vias de sinalização possivelmente alteradas.

### **3 MATERIAL E MÉTODOS**

Todos os programas e dados foram executados, processados e armazenados no Laboratório de Bioinformática e Biologia Computacional no Instituto Nacional de Câncer (INCA), no Laboratório de Genômica Funcional e Bioinformática do Instituto Oswaldo Cruz da Fundação Oswaldo Cruz (IOC/FIOCRUZ) e na Plataforma de Bioinformática da Fiocruz RPT04A/RJ da Fiocruz.

#### **3.1 Obtenção dos genomas de referência e transcritos do RefSeq para a construção das matrizes**

Arquivos de extensão VCF que contém os polimorfismos de três genomas de cada uma das 22 populações humanas, disponíveis no 1000G (<ftp://ftp.1000genomes.ebi.ac.uk>) correspondente à fase 3 do projeto, versão 5a foram utilizados (Tabela 3.1). Estes arquivos possuem a localização cromossômica e tipo de cada polimorfismo. Todos mapeamentos realizados pelo 1000G foram realizados no genoma de referência humano versão GRCh37/hg19. A reconstrução dos três genomas de cada população humana do 1000G foi realizada utilizando programas escritos na linguagem de programação Perl para localizar a posição de cada polimorfismo apenas nas regiões gênicas do genoma de referência humano versão GRCh37/hg19.

Os dados referentes aos transcritos e proteínas foram obtidos através do projeto RefSeq (Pruitt et al., 2009) onde é possível obter dados de qualidade que foram utilizados para constituir a base de dados de humano.

É importante ressaltar que o nosso grupo já possuía os dados de mapeamento de todo dbEST no genoma de referência humano GRCh37/hg19, esta metodologia já está estabelecida no nosso grupo e este banco de dados vem sendo utilizado em outros projetos em andamento no laboratório. Cada transcrito depositado na base de dados é analisado utilizando a metodologia de matriz ternária (Tavares et al., 2014) já existente no nosso grupo para a identificação de perdas pequenas em sequências.

**Tabela 3.1** Genomas do 1000G utilizados no mapeamento dos dados de RNA-Seq.

Genoma	População	Ancestralidade
HG01455, HG01148, HG01113	Colombiana (CLM)	Americana
HG01948, HG01961, HG02299	Peruana (PEL)	Americana
HG00641, HG01055, HG01197	Porto riquenha (PUR)	Americana
HG01879, HG02549, HG02282	Afro-caribenha (ACB)	Africana
HG03105, HG03112, HG03169	Esan da Nigéria (ESN)	Africana
HG02852, HG02798, HG02896	Gambiana (GWD)	Africana
NA19041, NA19042, NA19020	Luthya do Quênia (LWK)	Africana
HG03066, HG03449, HG03433	Mende de Serra Leoa (MSL)	Africana
NA18878, NA18486, NA18517	Yoruba da Nigéria (YRI)	Africana
HG03910, HG03937, HG03916	Bengali de Bangladesh (BEB)	Sul asiática
HG03772, HG03714, HG04056	Indiano do Reino Unido (ITU)	Sul asiática
HG02731, HG02736, HG02654	Paquistanesa (PJL)	Sul asiática
HG03896, HG03756, HG03745	Tamil do Reino Unido (STU)	Sul asiática
HG00956, HG02156, HG02402	Chinesa Dai (CDX)	Leste asiática
NA18643, NA18641, NA18648	Chinesa Han de Pequim (CHB)	Leste asiática
HG00556, HG00674, HG00534	Chinesa Han do sul (CHS)	Leste asiática
NA18946, NA18975, NA18939	Japonesa (JPT)	Leste asiática
HG02020, HG01869, HG02061	Vietnamita (KHV)	Leste asiática
HG00151, HG00114, HG00122	Britânica (GBR)	Europeia
HG00341, HG00361, HG00367	Finlandesa (FIN)	Europeia
HG01537, HG01501, HG01602	Espanhola (IBS)	Europeia
NA12761, NA12762, NA12413	Americana caucasiana (CEU)	Europeia

### 3.2 Obtenção dos dados de RNA-Seq

Utilizamos dados de sequenciamento de transcriptoma de tumores de pacientes para os quais também estejam disponíveis dados de sequenciamento de tecido normal adjacente ao tumor para possível identificação de perdas que ocorrem apenas em tecido transformado, com exceção da amostra com réplica biológica proveniente da linhagem celular H1975 (acessos SRR1706863 e SRR1706864 do banco SRA utilizando o sequenciador “Illumina HiSeq 2500” e “reads single-end”), em que o doador não era fumante. A análise de pequenas deleções foi realizada

utilizando dados de RNA-Seq de adenocarcinoma de pulmão disponíveis no banco de dados SRA do NCBI (<ftp://ftp.ncbi.nih.gov/sra/>) e no banco de dados do TCGA (<https://tcga-data.nci.nih.gov/tcga/>).

Utilizamos o estudo SRP012656 da base de dados SRA, no total 12 corridas utilizando a plataforma “Illumina Genome Analyzer II” com leituras “paired-end” de 50 nucleotídeos correspondentes a amostras de adenocarcinoma de pulmão e tecido normal adjacente ao tumor de 6 pacientes coreanas que nunca fumaram (Tabela 3.2) (Kim et al., 2013a). Vale ressaltar que o trabalho supracitado não menciona nada sobre dois pacientes (P2 e P7) em suas análises. Além deste estudo, também utilizamos um grupo de amostras de pacientes de adenocarcinoma de pulmão do banco de dados TCGA sequenciados pelo “Christiana Healthcare Center” utilizando o sequenciador “Illumina HiSeq 2000” gerando leituras “paired-end” de 50 nucleotídeos que totalizam 28 amostras pareadas de 14 pacientes fumantes (Tabela 3.3) (Collisson et al., 2014). O acesso a estes dados foi autorizado pelo comitê de acesso a dados do dbGAP (projeto de número 9636).



**Tabela 3.2** Amostras utilizadas de Kim e colaboradores (2013a) do banco SRA com a quantidade de “reads” de cada corrida e estatísticas da montagem do transcriptoma produzida pelo programa Trinity.

ID amostra	Acesso SRA	Tipo de amostra	leituras brutas	leituras após filtro de qualidade	leituras mapeadas	Contigs montados	N50
P1N	SRR493938	Normal	76.902.448	74.301.521	67.168.110	682.755	546
P1T	SRR493940	Tumoral	68.084.332	65.825.445	58.505.818	806.563	463
P3N	SRR493942	Normal	68.426.110	66.322.532	61.616.904	519.749	715
P3T	SRR493944	Tumoral	83.275.440	80.114.213	72.150.855	700.773	407
P4N	SRR493946	Normal	71.277.616	68.824.114	63.568.988	205.709	486
P4T	SRR493948	Tumoral	69.054.280	66.528.755	60.605.403	707.098	463
P5N	SRR493950	Normal	64.767.290	62.225.438	56.730.775	585.151	377
P5T	SRR493952	Tumoral	58.892.784	56.652.120	50.907.099	386.764	470
P6N	SRR493954	Normal	53.335.026	52.105.854	47.931.290	592.680	424
P6T	SRR493956	Tumoral	59.276.726	56.327.257	52.384.986	696.496	342
P8N	SRR493958	Normal	61.971.570	60.221.102	55.352.282	750.229	315
P8T	SRR493960	Tumoral	63.737.424	60.429.012	55.840.651	792.318	288

**Tabela 3.3** Lista de amostras utilizadas do banco de dados TCGA sequenciados do “Chritiana Healthcare Center” com a quantidade de “reads” de cada corrida e estatísticas da montagem do transcriptoma produzida pelo programa Trinity.

(continua)

ID amostra	TCGA Barcode	Tipo de amostra	leituras brutas	leituras após filtro de qualidade	leituras mapeados	Contigs montados	N50
2655_N	TCGA-44-2655-11	Normal	75.312.544	73.102.057	68.699.901	531.330	175
2655_T	TCGA-44-2655-01	Tumoral	74.898.411	73.502.233	68.683.965	830.960	95
2657_N	TCGA-44-2657-11	Normal	76.088.415	71.619.475	67.636.142	375.494	163
2657_T	TCGA-44-2657-01	Tumoral	75.300.578	71.874.171	64.573.144	391.288	146
2662_N	TCGA-44-2662-11	Normal	76.356.439	70.444.379	68.360.589	379.091	138
2662_T	TCGA-44-2662-01	Tumoral	75.826.248	70.971.323	66.821.175	652.862	123
2668_N	TCGA-44-2668-11	Normal	74.918.887	72.608.032	64.460.604	354.720	178
2668_T	TCGA-44-2668-01	Tumoral	74.660.362	72.517.768	68.010.771	1.691.885	97
3396_N	TCGA-44-3396-11	Normal	76.859.393	70.508.486	67.071.951	438.565	170
3396_T	TCGA-44-3396-01	Tumoral	76.685.297	70.699.805	63.521.447	659.999	129
3398_N	TCGA-44-3398-11	Normal	75.483.539	73.398.168	68.536.549	565.421	166
3398_T	TCGA-44-3398-01	Tumoral	74.408.644	70.653.374	69.252.155	659.819	124
5645_N	TCGA-44-5645-11	Normal	75.912.285	73.806.304	67.673.020	425.579	185
5645_T	TCGA-44-5645-01	Tumoral	75.486.898	71.947.452	67.318.294	460.562	159
6145_N	TCGA-44-6145-11	Normal	74.461.400	70.010.825	66.112.520	370.744	162
6145_T	TCGA-44-6145-01	Tumoral	76.629.817	71.060.927	66.130.021	329.022	194

**Tabela 3.3** Lista de amostras utilizadas do banco de dados TCGA sequenciados do Chritiana Healthcare Center com a quantidade de “reads” de cada corrida e estatísticas da montagem do transcriptoma produzida pelo programa Trinity.

(conclusão)

ID amostra	TCGA Barcode	Tipo de amostra	leituras brutas	leituras após filtro de qualidade	leituras mapeadas	Contigs montados	N50
6146_N	TCGA-44-6146-11	Normal	75.282.358	73.933.243	67.957.539	320.554	193
6146_T	TCGA-44-6146-01	Tumoral	75.082.597	70.779.186	68.036.431	323.506	174
6147_N	TCGA-44-6147-11	Normal	75.672.495	73.121.933	67.556.729	493.120	178
6147_T	TCGA-44-6147-01	Tumoral	75.328.638	73.697.347	67.110.433	1.507.745	78
6148_N	TCGA-44-6148-11	Normal	76.102.142	72.288.143	65.901.463	543.506	148
6148_T	TCGA-44-6148-01	Tumoral	74.717.538	72.311.169	64.114.604	440.751	173
6776_N	TCGA-44-6776-11	Normal	74.565.090	70.792.115	63.607.371	320.244	227
6776_T	TCGA-44-6776-01	Tumoral	76.371.643	70.136.497	68.868.050	244.706	256
6777_N	TCGA-44-6777-11	Normal	75.032.096	71.593.840	64.761.382	348.551	242
6777_T	TCGA-44-6777-01	Tumoral	74.518.447	71.460.520	67.822.312	384.869	245
6778_N	TCGA-44-6778-11	Normal	74.371.546	71.653.077	63.154.381	672.992	153
6778_T	TCGA-44-6778-01	Tumoral	74.232.676	71.539.354	66.800.118	673.375	154

### 3.3 Filtro das leituras e mapeamento

Os dados de sequenciamento foram alinhados com o programa Novoalign versão 3.03 (parâmetros “-r None -i 300,50”) (Hercus, 2009) após a retirada de adaptadores (retirando a sequência ‘AGATCGGAAGAGC’ de 13 nucleotídeos correspondente do Illumina) e filtro de qualidade (retirando as leituras com “phred score” inferior a 30) usando o programa Trim Galore versão 0.4.0 ([www.bioinformatics.babraham.ac.uk/projects/trim\\_galore/](http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/)). O parâmetro “-r None” do Novoalign corresponde ao comando do programa para selecionar as leituras que mapeam uma única vez no genoma evitando utilizar aqueles que poderiam estar mapeando em regiões repetitivas ou em famílias gênicas. Já o parâmetro “-i 300,50” corresponde a uma instrução para o programa considerar leituras P1 e P2 respeitem o tamanho do fragmento sequenciado de 300 nucleotídeos com desvio padrão de 50, presente nos artigos de Kim e colaboradores (2013a) e Collisson e colaboradores (2014) e que são geralmente usados no sequenciamento “paired-end” da Illumina. O programa samtools (versão 1.5) (Li et al., 2009) foi utilizado para converter os arquivos de extensão SAM em arquivos binários comprimidos de extensão BAM (usando os parâmetros “view -Sb”). O primeiro parâmetro “-S” para auto detectar o formato do arquivo de entrada e o parâmetro “-b” para usar o padrão de saída BAM. As tabelas 3.2 e 3.3 mostram a quantidade leituras nos arquivos brutos, após os filtros de qualidade, aqueles utilizados no mapeamento por amostra e informações da montagem dos transcriptomas utilizando o programa Trinity.

### 3.4 Montagem usando o Trinity

As leituras alinhadas ao genoma de referência GRCh37/hg19 foram montados pelo programa de montagem Trinity versão 2.1.1 (Grabherr et al., 2011) utilizando o arquivo de extensão bam (parâmetros “--genome\_guided\_bam --genome\_guided\_max\_intron 10000 --CPU 2 --min\_contig\_length 30 --max\_memory 50G”). Para cada gene, utilizamos o programa Trinity para o conjunto de leituras mapeadas na região. Assim, evitamos que haja contaminação de leituras mapeados em outros genes na montagem dos transcritos usando somente aqueles mapeados em cada gene. Escolhemos o tamanho mínimo de 30 para cada contig para

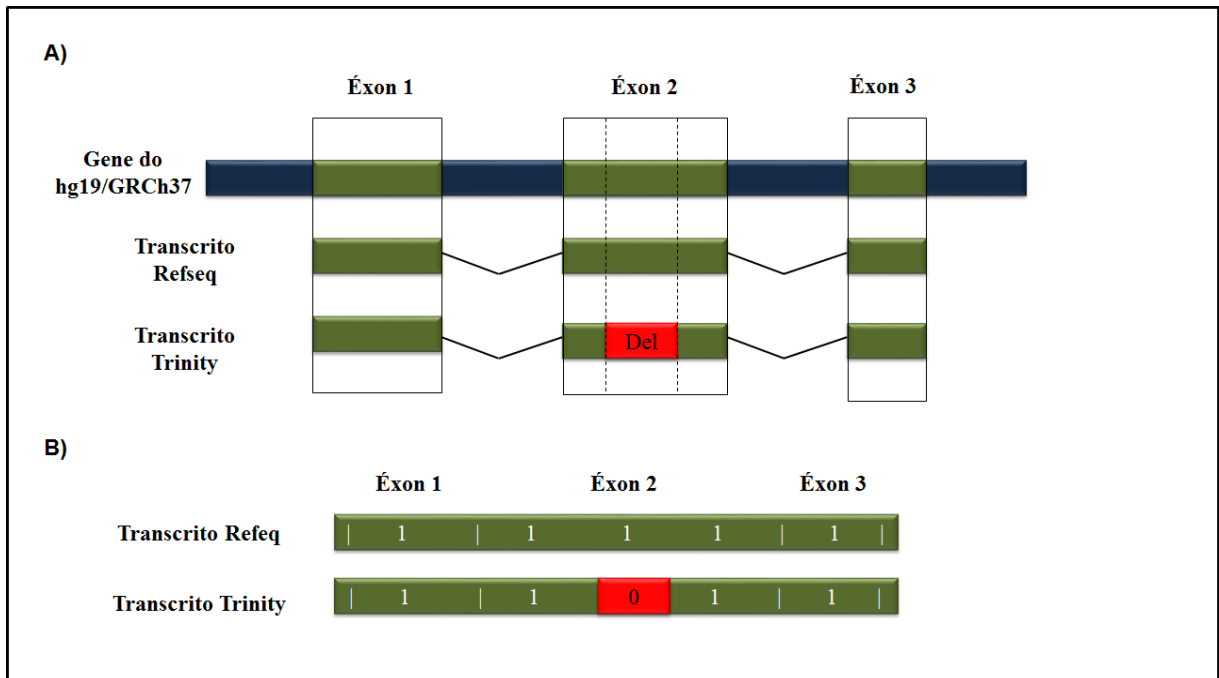
possibilitar que o programa Trinity monte utilizando o máximo de leituras possíveis para não perder respectivas leituras que tenham deleções. Outro parâmetro escolhido foi o tamanho máximo de íntron para 10.000 para que a montagem possa ocorrer mesmo em regiões distantes. As tabelas 3.2 e 3.3 mostram a quantidade de transcritos montados por amostra e o valor N50, ou seja, tamanho dos contigs em que, junto com os maiores contigs, representa metade da cobertura total da corrida (Pontius et al., 2007).

### **3.5 Identificação de pequenas deleções**

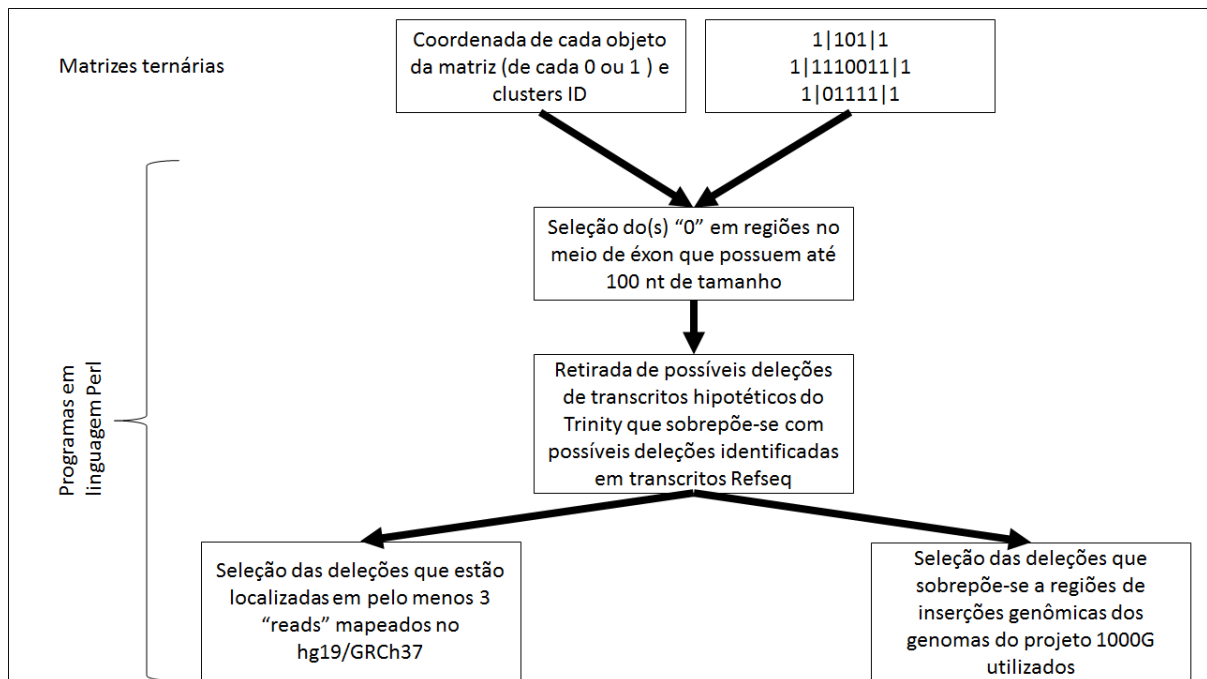
Como exposto, usamos uma abordagem de alinhar e montar os transcritos para cada gene para evitar a contaminação de transcritos de outros genes quando forem processados as leituras de um determinado gene. O produto desta montagem são os transcritos hipotéticos para cada gene que foram alinhados contra os seus respectivos genes presentes nos 66 genomas do 1000G e o genoma de referência humano GRCh37/hg19 utilizando o programa BLAT para cada gene (Kent, 2002) (parâmetros `-minIdentity=97 -out=sim4`). O parâmetro de mínima identidade foi utilizado para manter apenas os alinhamentos que possuíssem um mínimo de 97 de identidade (máximo de 100) para obter os alinhamentos mais confiáveis e o segundo parâmetro correspondente a saída na forma em que cada linha possua um bloco de início e fim de localizações cromossômicas de cada bloco de alinhamento entre aberturas do mapeamento (seja íntrons ou deleções). Novas deleções pequenas de até 100 nucleotídeos foram identificadas a partir desses transcritos quando dispostas nas matrizes ternárias e presentes internamente em éxons (Figura 3.1) (Tavares et al., 2014). Identificamos apenas deleções pelo motivo de que a abordagem das matrizes ternárias não permite a identificação de inserções. Após a identificação das pequenas deleções, filtramos os resultados em que as deleções podiam ser encontradas em pelo menos três leituras do alinhamento original realizado pelo Novoalign (nas deleções identificadas a partir do genoma de referência hg19/GRCh37) ou filtradas ao verificarmos inserções nos arquivos de extensão VCF nos genomas do 1000G (Figura 3.2).

Além disso, utilizamos o programa samtools para mostrar quais leituras representavam as regiões afetadas por pequenas deleções (utilizando o parâmetro

“view chrN:x-y”, onde “N” é o número do cromossomo, “x” é o início e “y” é o fim da localização da deleção. Com a lista de leituras que cobrem uma certa região podemos também identificar quais delas possuíam as deleções ao buscar a letra “D” no código CIGAR do mapeamento. Com esse fim, utilizamos o programa “awk” no terminal do Linux com o seguinte comando: “awk '{if \$6 ~ /D/} {print \$0}’”. Com esse comando apenas é mostrado as linhas que possuem as leituras em que a 6ª posição “\$6” possui um “D” de deleção.



**Figura 3.1** Representação de transcritos do Refseq e gerado a partir da montagem do Trinity dispostos na matriz ternária. **A)** Representação do alinhamento dos dois transcritos, sendo o transcrito do Trinity contendo uma deleção no éxon 2, contra um gene do genoma de referência humano GRCh37/hg19. **B)** Deleção em vermelho representada com um zero nas matrizes ternárias.



**Figura 3.2** Esquema representando a identificação de pequenas deleções nas matrizes ternárias.

Realizamos a validação dos resultados utilizando o programa de identificação de SNPs e INDELS chamado Varscan, versão 2.4.0 (Koboldt et al., 2009). Utilizamos este programa com filtro de p-valor de 0,05 para identificar apenas as pequenas perdas de até 100 nucleotídeos nos dados de RNA-Seq mapeados ao genoma de referência humano GRCh37/hg19. Para visualizar as deleções utilizamos o programa IGV (do inglês, "Integrative Genomics Viewer") (Robinson et al., 2011) para as leituras mapeadas no genoma de referência GRCh37/hg19 e o programa JBrowse (Skinner et al., 2009) para os transcritos montados mapeados nos 66 genomas do 1000G.

Com o intuito de verificar as vias de sinalização possivelmente alterada, utilizamos o pacote do Bioconductor MGSA (versão 3.5) (Bauer et al., 2011). Este programa calcula a probabilidade posterior, através de uma inferência bayesiana, de uma via de sinalização de um dado banco de dados estar super-representada. No nosso caso, utilizamos a base de dados "Gene Ontology" (Ashburne et al., 2000). Apenas consideramos em nossos resultados aquelas com probabilidade posterior acima de 0,5 em possuírem maior chance de estarem alteradas.

### 3.6 Avaliação do impacto na sequência de proteínas

A anotação destas pequenas deleções foi realizada utilizando o programa Segtor (Renaud et al., 2011) (parâmetros -s hg19 -d refseq -m 5). O parâmetro “s” corresponde a versão do genoma usado para a anotação, o parâmetro “d” é utilizado para escolha das informações contidas na base de dados RefSeq (O’Leary et al., 2016) e o parâmetro “m 5” que corresponde a opção de anotação de INDELS. Também foram produzidos programas próprios em linguagem de programação Perl para esta análise para processar os arquivos de resultado. O programa Segtor ao anotar as pequenas deleções, mostra a sequência de aminoácidos hipoteticamente produzida a partir da tradução da sequência genômica. Com a sequência de aminoácidos desta proteína hipotéticas, podemos verificar os domínios que foram alterados utilizando a base de dados CDD (do inglês, “Coding Domain Database”) (Marchler-Bauer et al., 2015) através da utilização do programa RPS-BLAST (Marchler-Bauer et al., 2002) disponível no pacote BLAST+ (v2.2.3) (<ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/>) com o limiar de  $1 \times 10^{-5}$  para o e-value (usando o parâmetro -evalue). Além desse parâmetro utilizamos também o “-db” para utilizar a base de dados construída com o “makeblastdb”. O programa RPS-BLAST (do inglês, “Reverse Position-Specific BLAST”) realiza a comparação entre uma matriz de “scores” posição específica (PSSM), gerada a partir do alinhamento de domínios proteicos conservados contidos nas seguintes bases de dados: NCBI (CDD) disponível em ([www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi](http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi)) (NCBI Resource Coordinators, 2013), SMART (Letunic et al., 2004), COG disponível em (<http://www.ncbi.nlm.nih.gov/COG/old/xognitor.html>) (Tatusov et al., 2000), KOG Protein clusters disponível em ([www.ncbi.nlm.nih.gov/proteinclusters](http://www.ncbi.nlm.nih.gov/proteinclusters)) (Tatusov et al., 2003) e TIGRFAM disponível em ([blast.jcvi.org/web-hmm/](http://blast.jcvi.org/web-hmm/)) (Haft et al., 2003).



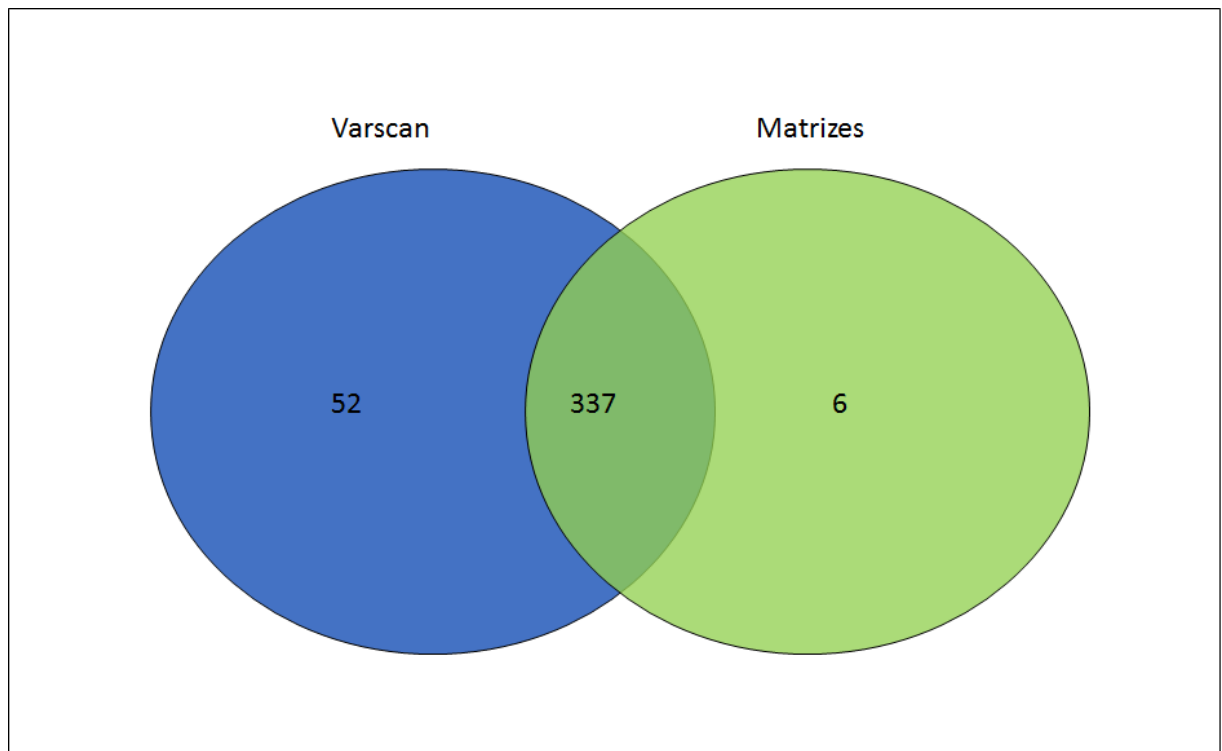
## 4 RESULTADOS E DISCUSSÃO

Para facilitar o entendimento dos resultados e seus desdobramentos no desenvolvimento deste projeto, apresentaremos os resultados seguido da discussão em cada uma das seções.

### 4.1 Identificação de pequenas deleções no genoma GRCh37/hg19

#### 4.1.1 *Prova de conceito utilizando a linhagem celular H1975*

Como prova de conceito, escolhemos o programa Varscan que foi adotado pela comunidade científica para a identificação de SNPs e INDELS em dados de DNA-Seq (Koboldt et al., 2009). Assim, utilizamos o programa Varscan para a identificação de 385 pequenas deleções de até 100 nucleotídeos em dados de RNA-Seq da linhagem celular humana H1975 (códigos de acesso no SRA SRR1706863 e SRR1706864). Utilizando estes dados brutos e empregando a metodologia das matrizes ternárias (Tavares et al., 2014), identificamos 343 pequenas deleções de até 100 nucleotídeos, sendo 87% delas identificadas pelos dois programas (Figura 4.1). Para a linhagem de células H1975, as deleções identificadas por nossa metodologia estavam presentes em regiões cobertas em média por 32 leituras, sendo cada deleção representada pela média de seis leituras (desvio padrão de 2). O valor máximo de leituras representando uma deleção foi de 15 e o valor mínimo para seleção de uma deleção foi de três leituras.



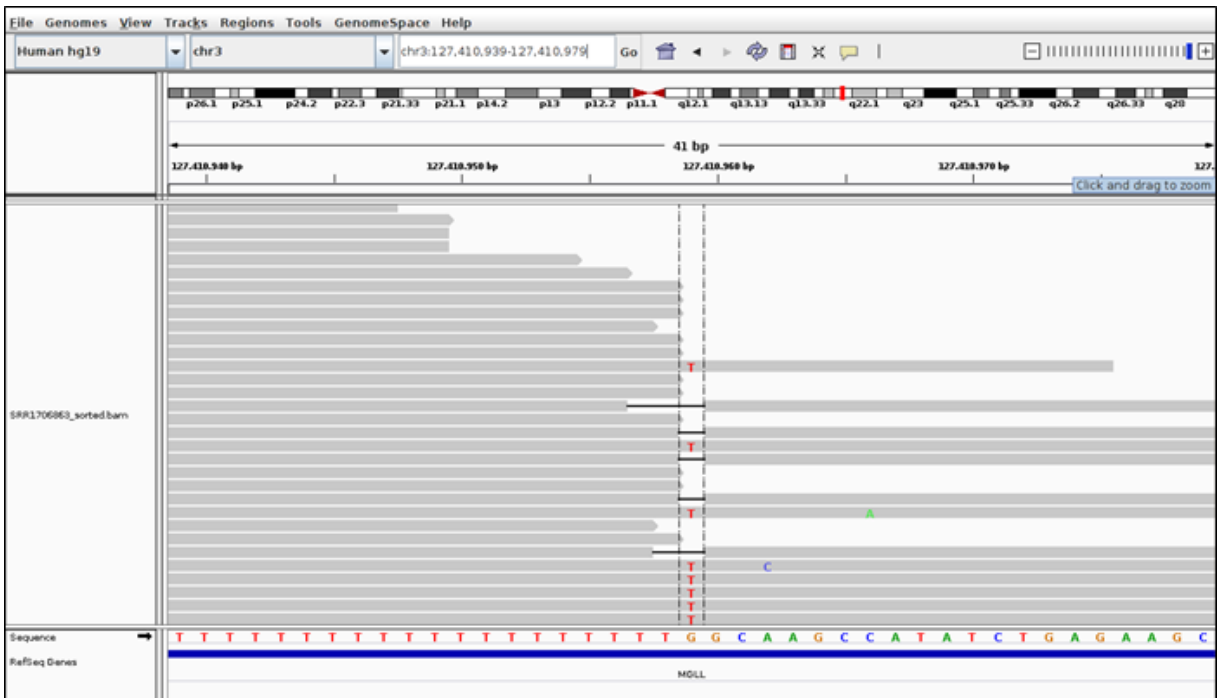
**Figura 4.1** Pequenas deleções identificadas por nossa metodologia utilizando as matrizes ternárias (verde) e utilizando o Varscan na amostra de RNA-Seq de H1975 (SRR1706863 e SRR1706864).

Ao compararmos os nossos achados com aqueles produzidos pelo programa Varscan, verificamos que 52 deleções pequenas de até 100 nucleotídeos foram detectadas exclusivamente pelo referido programa. Isto ocorreu devido ao fato de usarmos o programa Trinity para remontagem dos transcritos a partir de dados de RNA-Seq. Inspecionamos manualmente os 52 casos e apesar de alterarmos os parâmetros do programa Trinity, os transcritos hipotéticos sempre resultaram em alinhamentos sem aberturas (“gaps”). Este fato se repetiu nos resultados das análises nos conjuntos de dados de Kim e colaboradores (2013a) e Collisson e colaboradores (2014). Seis pequenas deleções foram detectadas exclusivamente pela abordagem das matrizes ternárias (Figura 4.1)

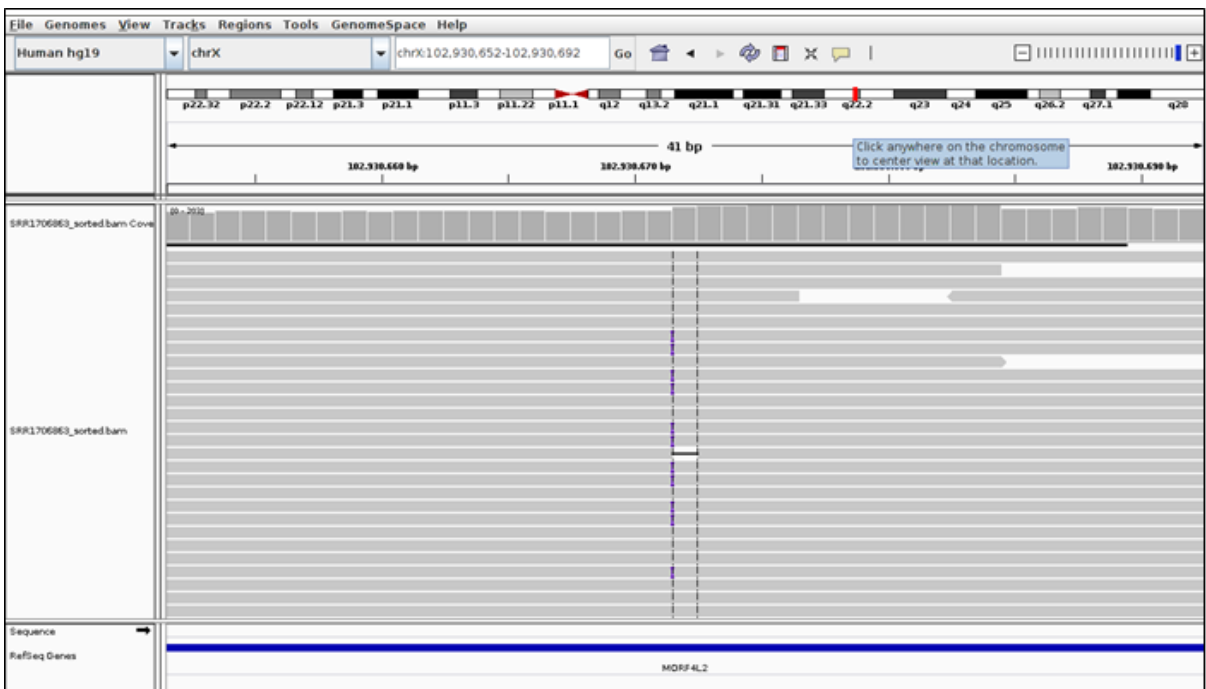
**Tabela 4.1** Deleções identificadas exclusivamente pela abordagem das matrizes ternárias em dados de RNA-Seq da linhagem H1975 e a cobertura das leituras.

Gene	Localização	Total de leituras na região com deleção (leituras com deleções)
<i>MGLL</i>	chr3:127410959-127410959	14 (5)
<i>MORF4L2</i>	chrX:102930671-102930671	164 (11)
<i>NUP50</i>	chr22:45583065-45583065	15 (6)
<i>NAP1L4</i>	chr11:2966276-2966276	45 (5)
<i>CDC42EP1</i>	chr22:37964412-37964432	28 (5)
<i>SETD7</i>	chr4:140431393-140431408	9 (7)

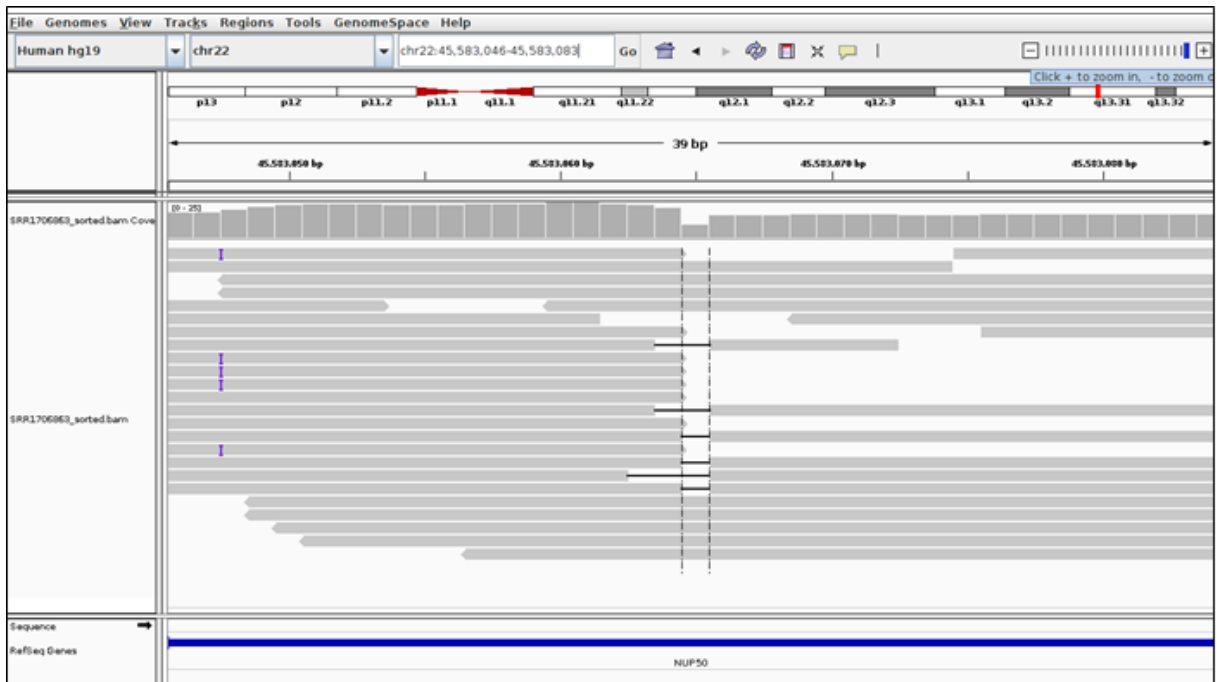
Como exposto na Tabela 4.1, 6 deleções não foram detectadas pelo programa Varscan, dentre elas: pequenas deleções de um nucleotídeo nos genes *MGLL* (Figura 4.2), *MORF4L2* (Figura 4.3), *NUP50* (Figura 4.4) e *NAP1L4* (Figura 4.5). Encontramos também deleções maiores que um nucleotídeo, sendo elas: uma de 21 nucleotídeos no gene *CDC42EP1* (Figura 4.6 e Figura 4.7) e outra de 17 nucleotídeos no gene *SETD7* (Figura 4.8). Dentre as deleções detectadas pela nossa abordagem, todas foram encontradas na porção 3' UTR dos genes, com exceção da deleção no gene *CDC42EP1* (Figura 4.6 e Figura 4.7). Esta deleção causa a mudança no quadro de leitura no último éxon deste gene. Outra deleção que também causa mudança no quadro de leitura na região codificadora do gene *CDC42EP1*, mas que não foi detectada pela nossa abordagem, já foi descrita em adenocarcinoma de pulmão (Imielinski et al., 2012). Podemos destacar a deleção na região 3' UTR do gene *MGLL* (Figura 4.2) que afeta a ligação do miRNA hsa-miR-182 (Arribas et al., 2013). Segundo nossas análises, nenhuma destas 5 deleções que afetam regiões 3' UTR de genes afetam sítios de ligação de miRNA conhecidos.



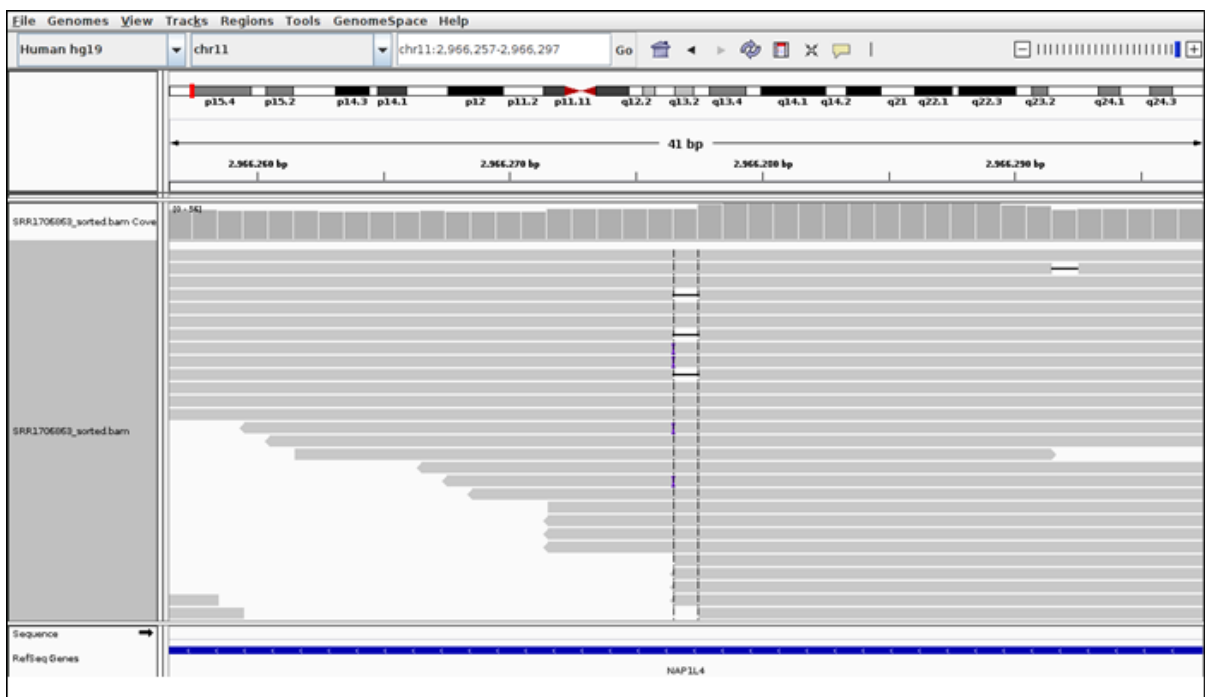
**Figura 4.2** Pequena deleção de um nucleotídeo no gene *MGLL* visualizada no programa IGV.



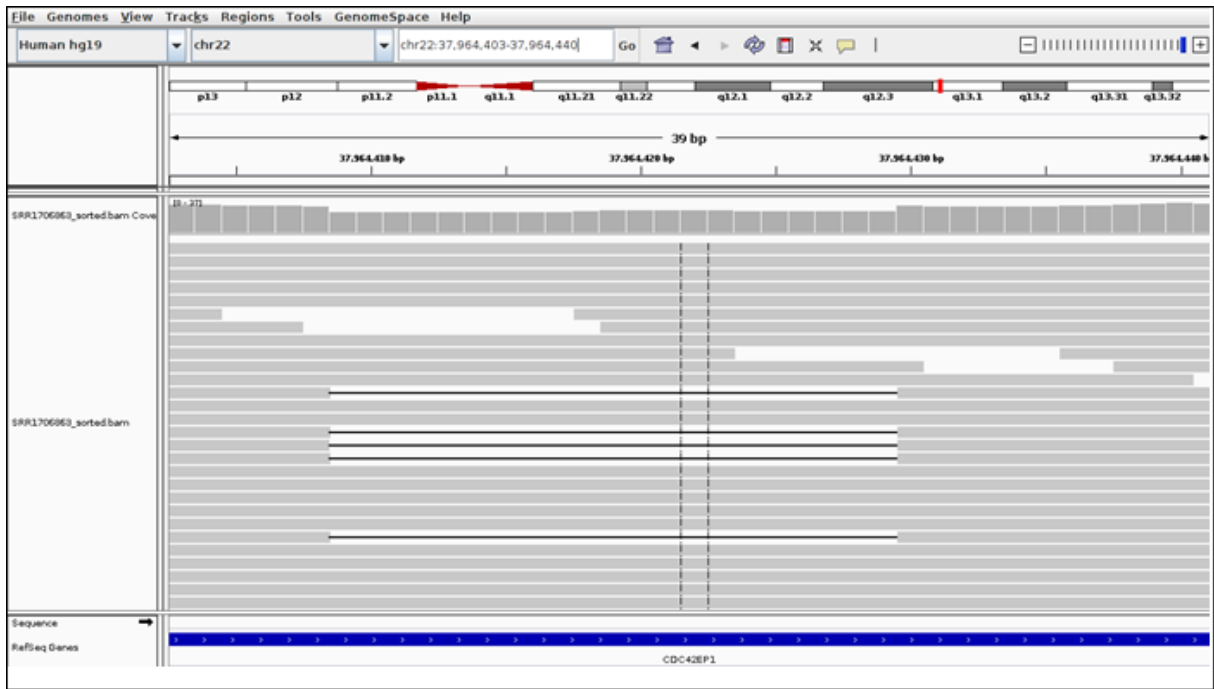
**Figura 4.3** Pequena deleção de um nucleotídeo no gene *MORF4L2* visualizada no programa IGV.



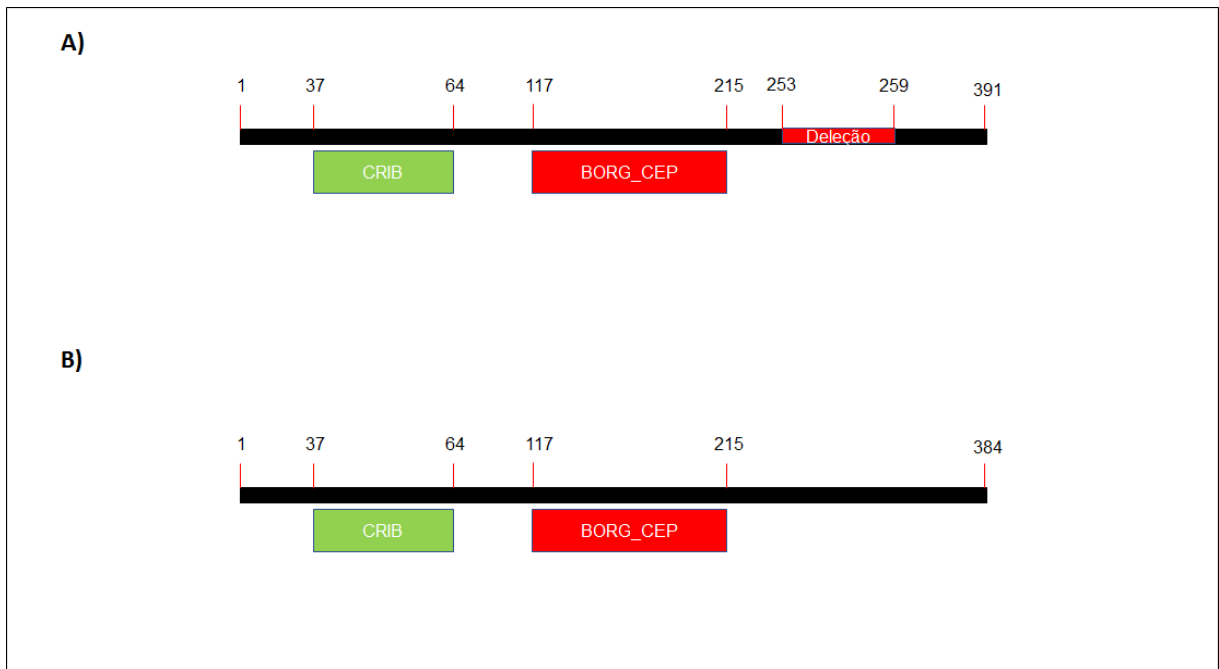
**Figura 4.4** Pequena perda de um nucleotídeo no gene *NUP50* visualizada no programa IGV.



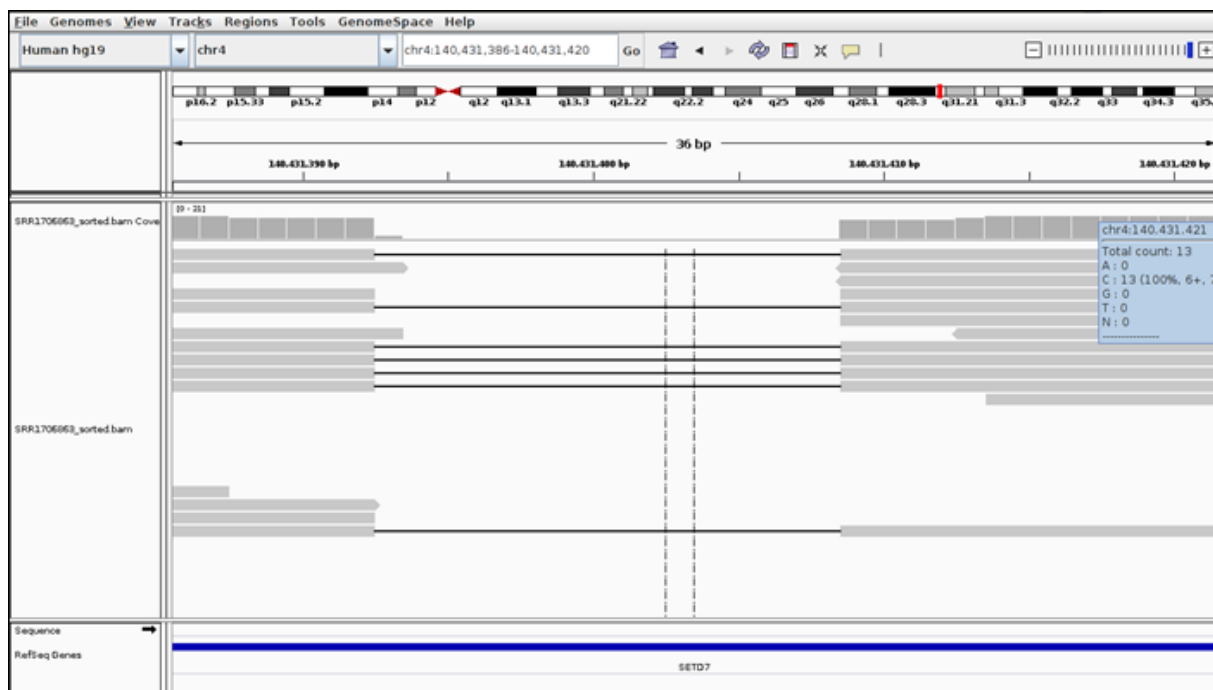
**Figura 4.5** Pequena perda de um nucleotídeo no gene *NAP1L4* visualizada no programa IGV.



**Figura 4.6** Pequena deleção de 21 nucleotídeos no gene *CDC42EP1* visualizada no programa IGV.



**Figura 4.7** Representação do impacto da pequena deleção de 21 nucleotídeos no gene *CDC42EP1* na sequência de aminoácidos. A proteína normal **(A)** codificada pelo gene possui 391 aminoácidos, com dois domínios: CRIB (Domínio de ligação PAK, acesso cl00113) e BORG\_CEP (Domínio de ligação de Rho GTPases). A proteína afetada **(B)** pela deleção mostra um encurtamento de sete aminoácidos.



**Figura 4.8** Pequena deleção de 17 nucleotídeos no gene *SETD7* visualizada no programa IGV.

Como esforço para avaliar os resultados obtidos quando empregamos a nossa abordagem e o programa Varscan, buscamos por deleções anotadas previamente pelo banco de dados de todas linhagens celulares ATCC (do inglês “American Type Culture Collection”) ([www.atcc.org](http://www.atcc.org)). No site do ATCC, só são listadas alterações de troca de um único nucleotídeo nesta linhagem e nenhuma deleção. Também buscamos no banco de dados COSMIC (Forbes et al., 2008), no qual encontramos 10 deleções nesta linhagem celular (9 em *EGFR* e 1 em *PIK3CA*), sendo que nenhuma delas foi identificada pelo Varscan ou por nossa metodologia, apesar de ser identificada a expressão desses dois genes nas amostras de RNA-Seq analisadas. Estabelecemos colaboração com a Dra. Mariana Caldas Waghbi do Laboratório de Genômica Funcional e Bioinformática do Instituto Oswaldo Cruz da Fundação Oswaldo Cruz para que ela faça a validação dos candidatos a deleção no genoma da linhagem celular H1975. Entretanto, a linhagem celular ainda está em fase de crescimento para a posterior extração do DNA.

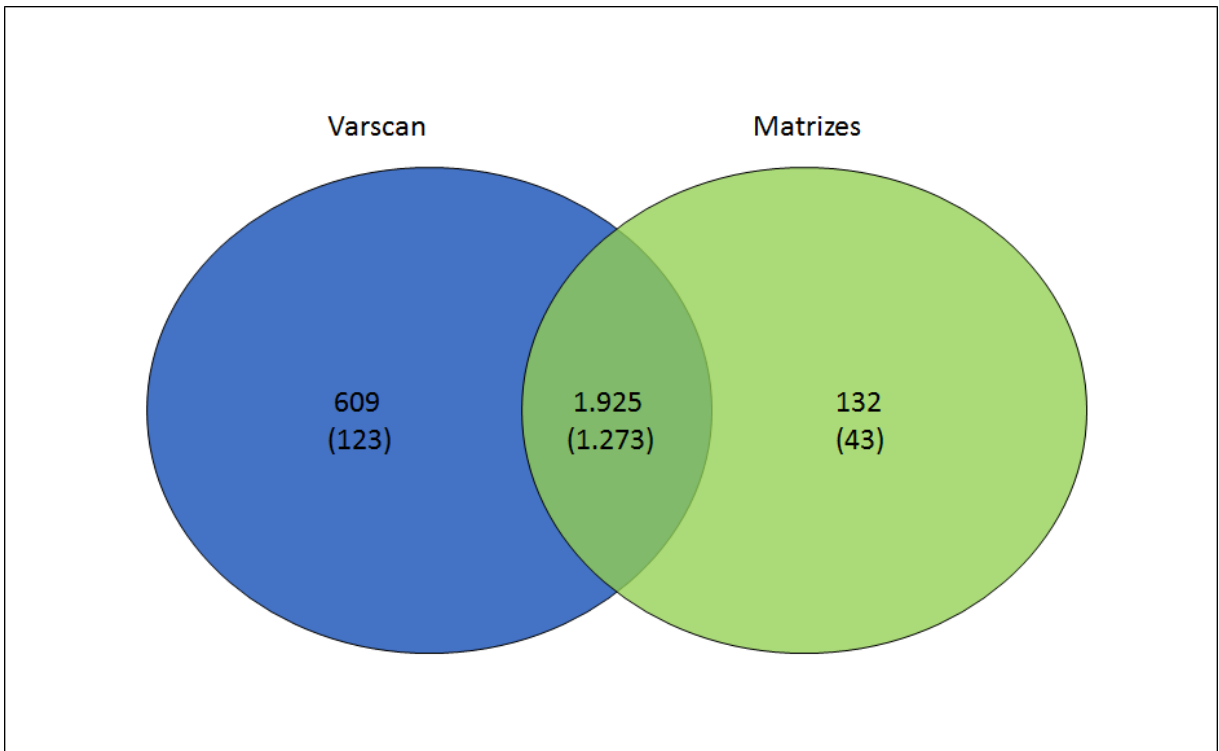
Uma vez que 98,25% (337/343) das deleções que encontramos na linhagem de células H1975 com a abordagem das matrizes ternárias também foram encontradas pelo programa Varscan e 86,63% (337/389) das deleções encontradas pelo referido programa terem sido encontradas pela nossa abordagem, acreditamos

que a nossa abordagem consiga detectar boa parte das deleções identificadas pelo programa Varscan. Como exposto na seção Justificativa, o programa Varscan não foi desenvolvido para a identificação de deleções no contexto de dados de RNA-Seq e os arquivos de anotação de deleções do 1000G. Assim, mostraremos a seguir, as deleções identificadas pelas matrizes ternárias e o programa Varscan em dados de RNA-Seq de pacientes com câncer de pulmão usando o genoma humano de referência, bem como a identificação de outras utilizando a metodologia das matrizes ternárias em dados de deleções genômicas disponibilizados pelo 1000G.

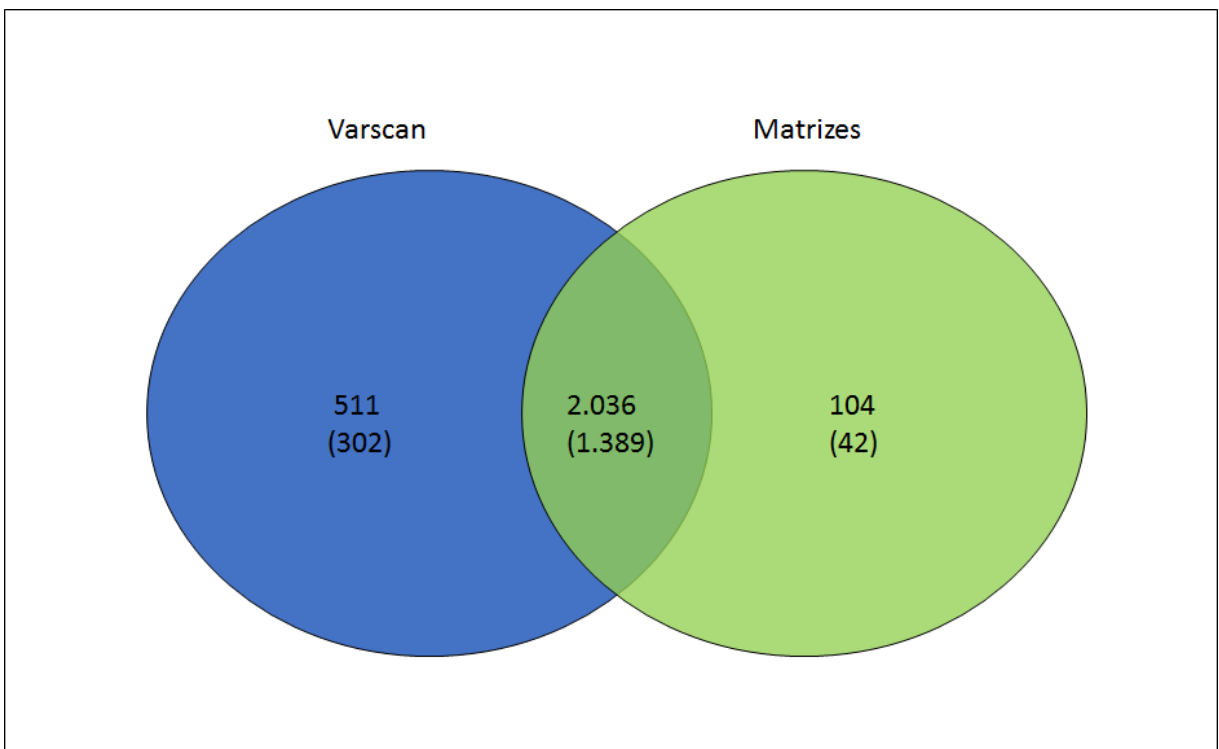
#### ***4.1.2 Pequenas deleções identificadas em amostras de tecidos normais adjacentes ao tumor e tumorais de seis pacientes não fumantes de câncer de pulmão***

Analizamos dados de corridas de RNA-Seq de amostras normais e tumorais pareadas de seis pacientes com adenocarcinoma de pulmão não fumantes publicados por Kim e colaboradores (2013a) para buscar por deleções de até 100 nucleotídeos. Utilizando a metodologia das matrizes ternárias, conseguimos identificar 2.057 pequenas deleções de até 100 nucleotídeos nos dados de amostras normais (1.316 delas em mais de uma amostra, 150 em todas) e 2.140 pequenas deleções apenas em amostras tumorais destes pacientes (1.431 delas ocorrem em mais de uma amostra, 231 em todas). Ao comparar os resultados obtidos pela nossa abordagem com aqueles obtidos com o programa Varscan, 76% das deleções oriundas de amostras normais e 80% das deleções originárias de amostras tumorais são encontradas pelas duas estratégias (Figura 4.9 e Figura 4.10).





**Figura 4.9** Pequenas deleções identificadas utilizando as matrizes ternárias (verde) e utilizando o Varscan (azul) em amostras de tecido normal de RNA-Seq de seis pacientes de câncer de pulmão (Kim et al., 2013a). Entre parênteses a quantidade de deleções identificadas em mais de um paciente.



**Figura 4.10** Pequenas deleções identificadas utilizando as matrizes ternárias (verde) e utilizando o Varscan (azul) em amostras de tecido tumoral de RNA-Seq de seis pacientes de câncer de pulmão (Kim et al., 2013a). Entre parênteses a quantidade de deleções identificadas em mais de um paciente.

Utilizando as matrizes ternárias, identificamos 140 pequenas deleções em amostras normais (Figura 4.9) e 122 pequenas deleções presentes em amostras tumorais (Figura 4.10) de pacientes que também foram identificadas por outros autores e constam no banco de dados dbSNP (Sherry et al., 2001) (Tabela 4.2). Além disso, 455 deleções identificadas em amostras normais e 457 em amostras tumorais estão presentes na base de dados COSMIC (Tabela 4.2).

**Tabela 4.2.** Tabela de quantas deleções estão presentes na base de dados dbSNP e COSMIC em amostras normais e tumorais.

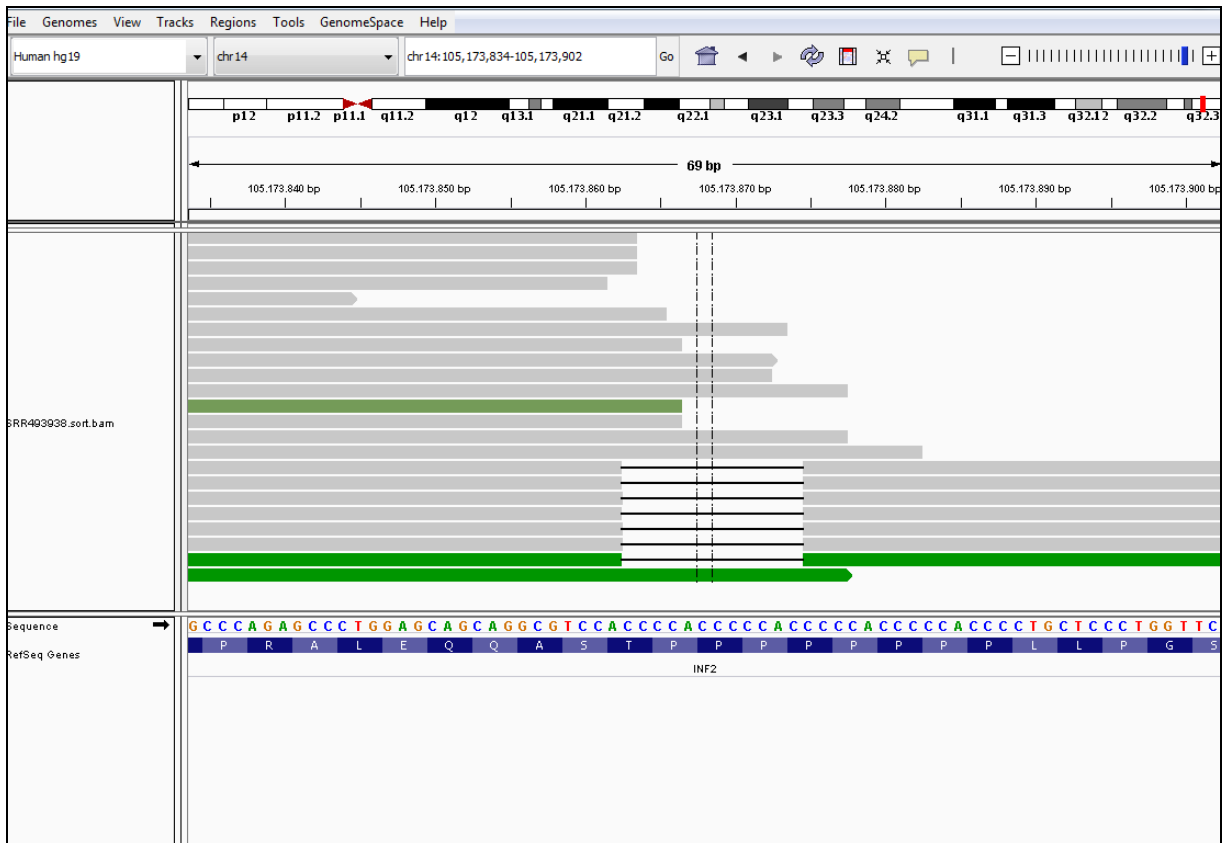
Banco	Tipo de amostra	Em regiões não codificadoras	Em regiões codificadoras
dbSNP	Normal	100	40
	Tumoral	72	
COSMIC	Normal	371	84
	Tumoral	378	79

As deleções identificadas por nossa metodologia estavam em regiões cobertas em média por 53 leituras, sendo que cada deleção é representada pela média de 17 leituras (com desvio padrão de 5). O valor máximo de leituras representando uma deleção foi de 42 e o mínimo de três. Identificamos oito pequenas deleções com relevância em nossos dados (Tabela 4.3).

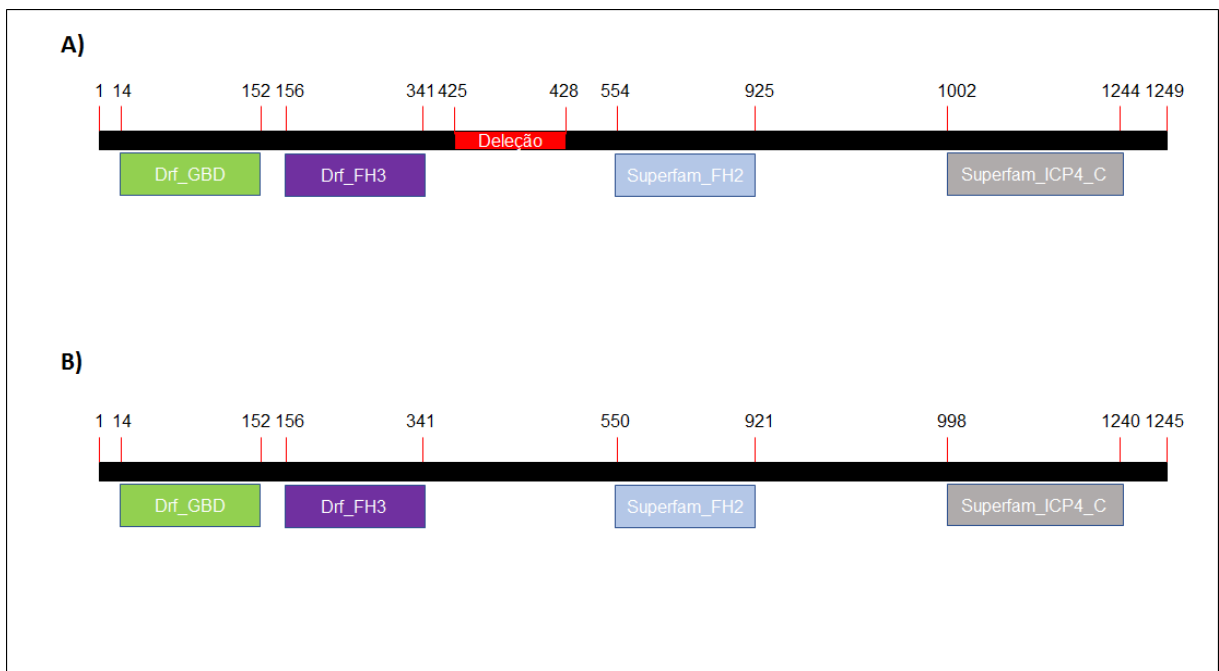
**Tabela 4.3** Deleções identificadas em dados de RNA-Seq nas amostras de Kim e colaboradores (2013a) e a cobertura das leituras.

Gene	Localização	Amostra	Total de leituras na região com deleção (leituras com deleções)
<i>INF2</i>	chr14:105173863-105173874	P1N	22 (8)
<i>INF2</i>	chr14:105173863-105173874	P3N	15 (7)
<i>EGFR</i>	chr7:55242465-55242479	P1T	78 (35)
<i>EGFR</i>	chr7:55242465-55242479	P5T	62 (23)
<i>TP53BP2</i>	chr1:223986236-223986245	P4T	57 (7)
<i>IFNGR1</i>	chr6:137518967-137518967	P1T	163 (25)
<i>IFNGR1</i>	chr6:137518967-137518967	P3T	180 (22)
<i>IFNGR1</i>	chr6:137518967-137518967	P4T	132 (8)
<i>IFNGR1</i>	chr6:137518967-137518967	P5T	98 (18)
<i>IFNGR1</i>	chr6:137518967-137518967	P6T	170 (22)
<i>IFNGR1</i>	chr6:137518967-137518967	P8T	160 (18)
<i>BAMBI</i>	chr10:28971526-28971526	P1T	7 (7)
<i>BAMBI</i>	chr10:28971526-28971526	P3T	10 (8)
<i>PTEN</i>	chr10:89725404-89725404	P1T	31 (16)
<i>CTSA</i>	chr20:44520258-44520261	P8T	269 (46)
<i>EIF3A</i>	chr10:120795211-120795211	P1T	47 (34)
<i>EIF3A</i>	chr10:120795211-120795211	P3T	60 (28)
<i>EIF3A</i>	chr10:120795211-120795211	P4T	42 (10)
<i>EIF3A</i>	chr10:120795211-120795211	P5T	30 (22)
<i>EIF3A</i>	chr10:120795211-120795211	P6T	45 (10)
<i>EIF3A</i>	chr10:120795211-120795211	P8T	25 (9)

Identificamos exclusivamente usando as matrizes ternárias uma pequena deleção em amostra normal (paciente 1 e paciente 3) de 12 nucleotídeos que não altera o quadro de leitura da tradução do gene *INF2* (Figura 4.11). Esta deleção levou a perda de quatro aminoácidos na proteína afetada (Figura 4.12)

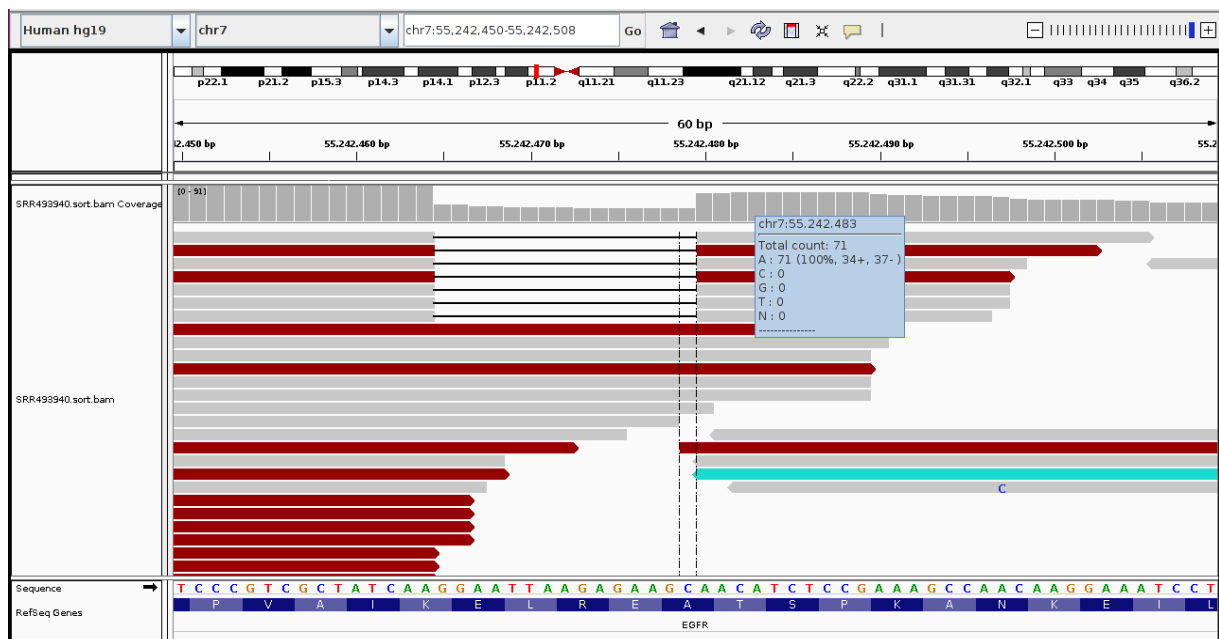


**Figura 4.11** Pequena deleção de 12 nucleotídeos visualizada na amostra normal do paciente 1 (P1N) no exon 19 do gene *INF2*.

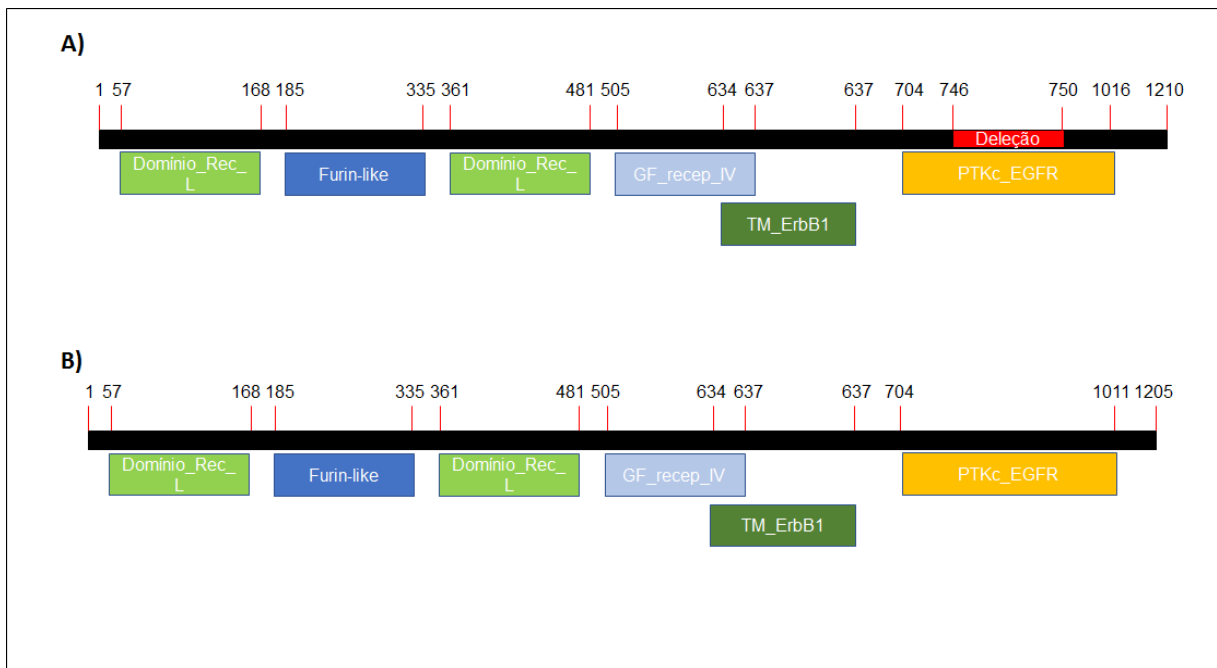


**Figura 4.12.** Representação do impacto da pequena deleção de 12 nucleotídeos no gene *INF2* na sequência de aminoácidos. A proteína normal **(A)** codificada pelo gene possui 1249 aminoácidos, com quatro domínios: Drf\_GBD (Domínio de ligação de GTPase translúcido, acesso cl05720), Drf\_FH3 (Domínio FH3 translúcido, acesso pfam06367), Superfam\_FH2 (Domínio FH2, acesso cl19758) e Superfam\_ICP4\_C (Região C-terminal ICP4-like, acesso cl28033). A proteína afetada **(B)** pela deleção mostra um encurtamento de quatro aminoácidos.

Identificamos uma deleção de 15 de nucleotídeos no éxon 19 do gene *EGFR* em amostras tumorais (paciente 1 e paciente 5) (Figura 4.13). Esta pequena deleção não foi encontrada pelo programa Varscan, mas possui anotação no dbSNP sob identificador “rs121913421”. Dos 78 leituras nessa região, aproximadamente metade deles (35) possuem deleção no paciente 1 e das 62 leituras nesta região, 23 possuem esta deleção no paciente 5. Esta deleção não altera o quadro de leitura da tradução da proteína e afeta o domínio tirosina quinase, que perde assim cinco aminoácidos (Figura 4.14). Este domínio tem como função se ligar ao ATP e ceder um fosfato (fosforilação) para uma proteína alvo para ativá-la. Neste caso a mutação pode reduzir a afinidade deste domínio pelo ATP (Lynch et al., 2004). Estudos demonstraram que este tipo de mutação ocorre em 48% dos tumores que possuem alteração neste gene (Mitsudomi e Yatabe, 2010). Esta mesma deleção já foi descrita sendo importante em conferir resistência à quimioterapia quando o composto gefitinib é usado no tratamento de pacientes com câncer de pulmão (Lynch et al., 2004).

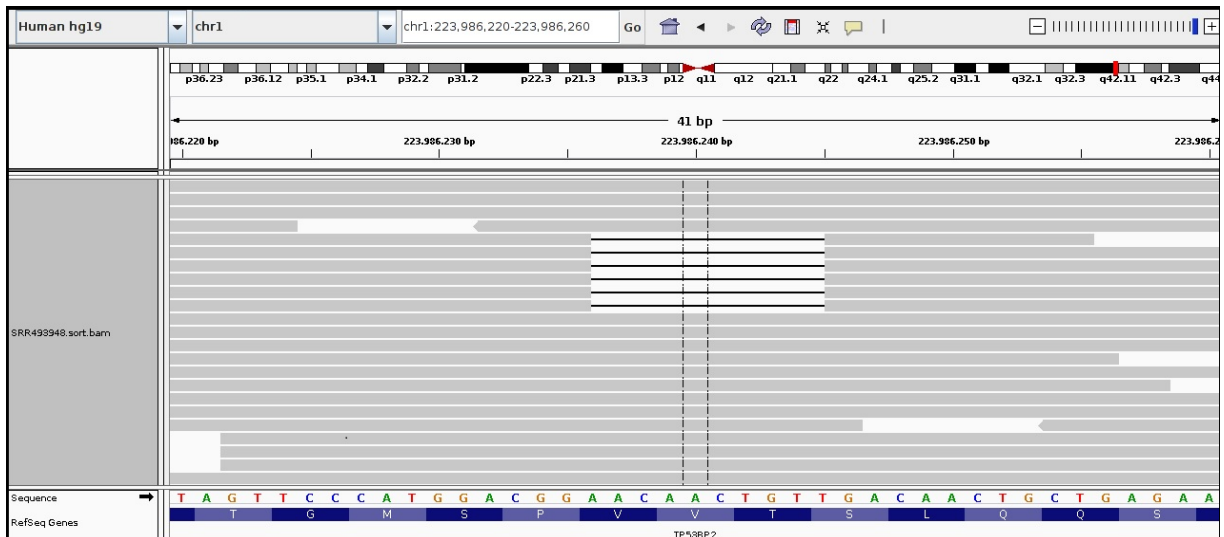


**Figura 4.13** Pequena deleção de 15 nucleotídeos visualizada na amostra tumoral do paciente 1 (P1T) no éxon 19 do gene *EGFR*.

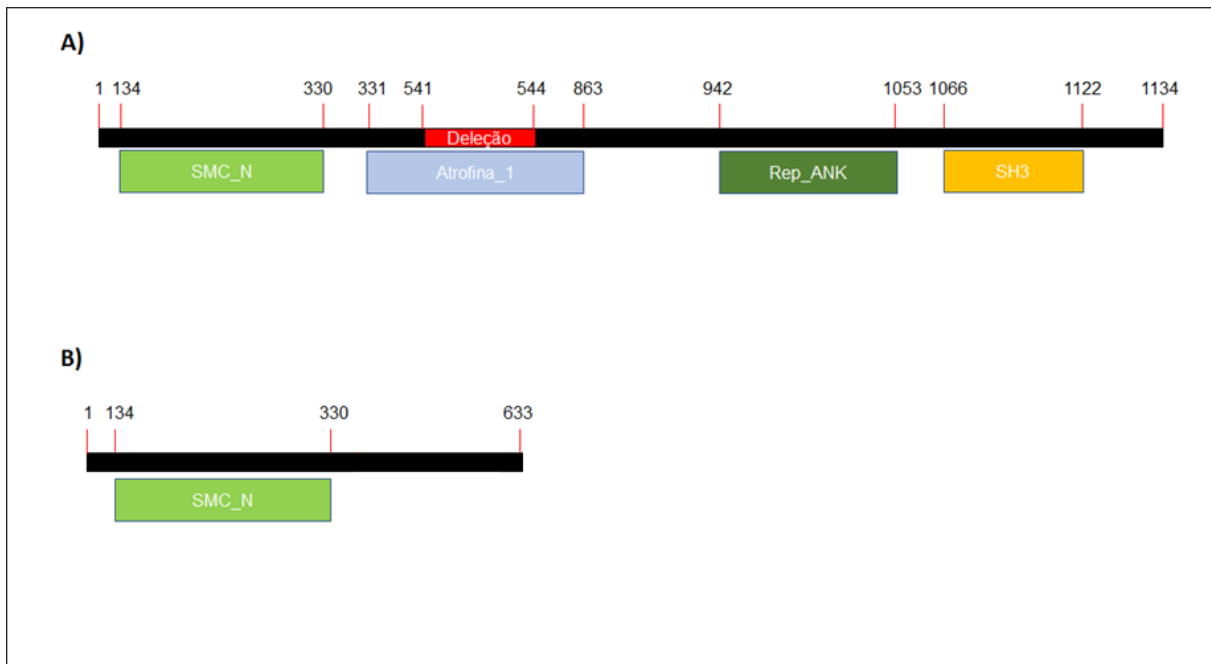


**Figura 4.14** Representação do impacto da pequena deleção de 15 nucleotídeos no gene *EGFR* na sequência de aminoácidos. A proteína normal **(A)** codificada pelo gene possui 1210 aminoácidos, com seis domínios: dois domínios Dominio\_Rec\_L (domínios de receptores L, acesso pfam01030), Furin-like (Região rica em cisteína furina-like, acesso pfam00757), GF\_recep\_IV (Domínio fator de crescimento IV, acesso pfam14843), TM\_ErbB1 (Domínio Transmembrana ErbB1, acesso cd12093) e PTKc\_EGFR (Domínio catalítico tirosina quinase, acesso cd05108). A proteína alterada **(B)** pela deleção mostra o domínio tirosina quinase com menos cinco aminoácidos em sua estrutura.

Encontramos também uma deleção de dez nucleotídeos alterando o quadro de leitura da tradução no gene *TP53BP2* na amostra tumoral do paciente 4 (P4T) (Figura 4.15). Esta deleção afetou o quadro de leitura da tradução da proteína removendo resíduos responsáveis pelo sítio de ligação com a proteína p53 (Figura 4.16). Consideremos importante salientar que somente 7 das 52 leituras na região com a deleção apresentam a mutação. A ligação da proteína codificada pelo *TP53BP2* com a p53 pode levar a proteína a apoptose e regular o crescimento celular (Naumovski e Cleary, 1996). Alguns estudos já mostraram esse gene com deleções que alteram o quadro de leitura da tradução adiando o códon de parada em adenocarcinoma de pulmão (Cancer Genome Atlas Network, 2012; Wang et al., 2014).



**Figura 4.15** Pequena deleção de 10 nucleotídeos na amostra tumoral do paciente 4 (P4T) do gene *TP53BP2*.



**Figura 4.16** Representação do impacto da pequena deleção de 10 nucleotídeos no gene *TP53BP2* na sequência de aminoácidos. A proteína normal **(A)** codificada pelo gene possui 1134 aminoácidos, que inclui quatro domínios: SMC\_N (Domínio N-terminal SMC, acesso cl25732), Atrofina\_1 (Família atrofina-1, acesso cl26464), Rep\_ANK (Repetições anquirina, acesso cd00204), SH3 (Domínio de apoptose e de ligação de p53, acesso cd11953). A proteína alterada **(B)** pela deleção possui menos aminoácidos por causa da mudança do quadro de leitura levando a perder diversos domínios da proteína.

A análise de enriquecimento de vias dos genes contendo as pequenas deleções identificadas pelas matrizes ternárias em amostras tumorais mostrou vias

com alta probabilidade de estarem alteradas como, por exemplo, via de sinalização mediada por interferon- $\gamma$  (probabilidade posterior de 1,00), expressão gênica (probabilidade posterior de 1,00), resposta ao estresse do Retículo Endoplasmático (probabilidade posterior de 0,96), regulação negativa da via de TGF- $\beta$  (probabilidade posterior de 0,95) e via de sinalização *PDGFR* (probabilidade posterior de 0,94) (Tabela 4.4).

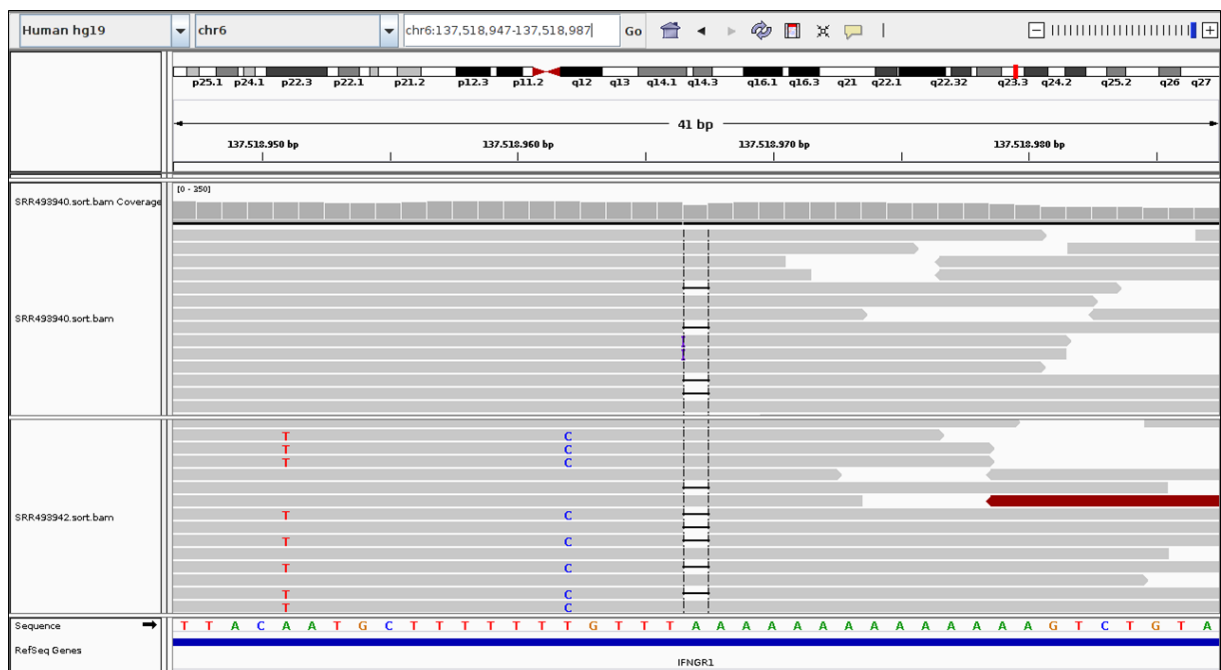
**Tabela 4.4** Enriquecimento de genes que sofreram pequenas deleções em amostras tumorais identificadas pela nossa metodologia de matrizes ternárias. Abaixo encontramos as vias do Gene Ontology (GO) com probabilidade posterior maior que 0,5 de estarem suprerrepresentadas.

GO id	Descrição	Probabilidade posterior	Quantidade de Genes
GO:0060333	Via de sinalização mediada pelo interferon- $\gamma$	1,00	22
GO:0010467	Expressão Gênica	1,00	14
GO:0034976	Resposta ao estresse do Retículo Endoplasmático	0,96	20
GO:0030512	Regulação negativa da via de TGF- $\beta$	0,95	16
GO:0048008	Via de sinalização PDGFR	0,94	11
GO:0001568	Regulação negativa da diferenciação de eritrócitos	0,84	11
GO:0045727	Regulação positiva da tradução	0,66	14
GO:0000188	Inativação da atividade de MAPK	0,59	9
GO:0032456	Reciclagem endocítica (membrana plasmática)	0,57	8
GO:0045647	Desenvolvimento de vasos sanguíneos	0,54	6

Um dos genes que pertence à via de sinalização mediada por interferon- $\gamma$  e que está afetado por uma deleção é *IFNGR1*, que é responsável por codificar um dos receptores do interferon- $\gamma$ . Essa deleção de um nucleotídeo foi encontrada na amostra tumoral de todos os seis pacientes (presente em 25 das 163 leituras no paciente 1, em 22 das 180 leituras no paciente 3, em 8 das 132 leituras do paciente



4, em 18 das 98 “reads” do paciente 5, em 22 das 170 leituras do paciente 6 e em 18 das 160 leituras do paciente 8) e afeta a porção 3’ UTR e possui anotação no dbSNP sob identificador “rs75851921” (Figura 4.17). Funcionando em perfeito estado, esta via pode desencadear alguns eventos que não são favoráveis para o crescimento e manutenção dos tumores como, por exemplo, a apoptose e se transformar em alvo do sistema imunológico (Parker et al., 2016). A via desencadeada através deste receptor estimula a transcrição de genes importantes para desencadear uma resposta imunológica que dificultará o crescimento do tumor (Yu et al., 2003)



**Figura 4.17** Pequena deleção de um nucleotídeo nas amostras tumorais do paciente 1 (P1T) e paciente 3 (P3T) do gene *IFNGR1*.

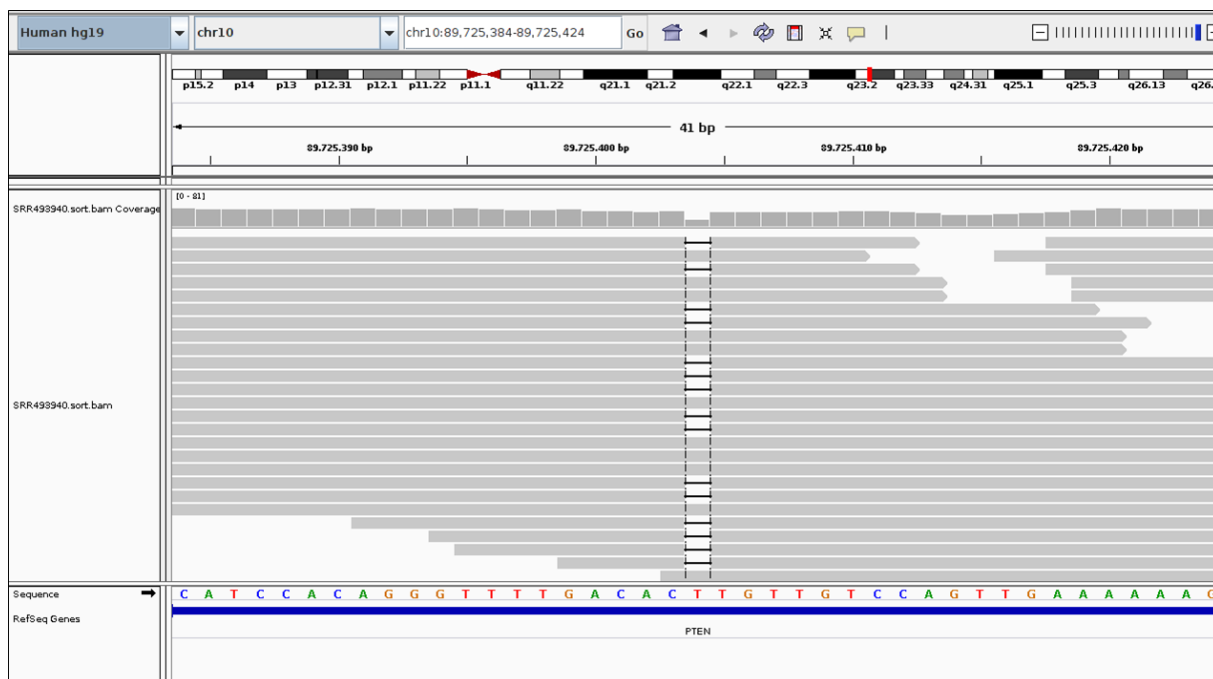
Outra via com alta probabilidade de estar alterada pelas perdas é a regulação negativa da via de TGF- $\beta$ . Esta via pode tanto funcionar como supressor tumoral como também estimular a transição epitélio-mesenquimal (EMT) associada com a metástase do tumor (Toonkel et al., 2010). Em nossos achados encontramos uma deleção de um nucleotídeo presente em todas as leituras localizadas na região da deleção (7/7) no gene *BAMBI* no paciente 1 e em 8 das 10 leituras do paciente 3, que está associada à esta via. A deleção neste gene se localiza na região 3’ UTR e também está presente no dbSNP sob o identificador “rs5784080”. Este gene codifica uma proteína receptora que pode funcionar como alternativa de ligação ao TGF- $\beta$  que desencadeia a reação de supressor tumoral. No entanto, a baixa expressão

desse gene já foi associada com a maior sinalização da via pró-tumor de TGF- $\beta$  aumentando o potencial de invasão em células de diferentes tumores de NSCLC, incluindo adenocarcinoma de pulmão (Marwitz et al., 2016). Como exposto, não encontramos em nossas análises que esta região com possui a deleção seja alvo de um miRNA conhecido.

Na via de *PDGFR* encontramos uma deleção confirmada por 16 das 31 “leitura localizadas na região da mutação no gene *PTEN*. Esta deleção também está presente na região 3’UTR do gene (Figura 4.19). Este gene é conhecido por suas funções supressoras tumorais para inibir o crescimento descontrolado de células transformadas (Leslie e Downes, 2004). Mutações somáticas nesse gene são encontradas em até 8% dos pacientes com NSCLC (Jin et al., 2010).



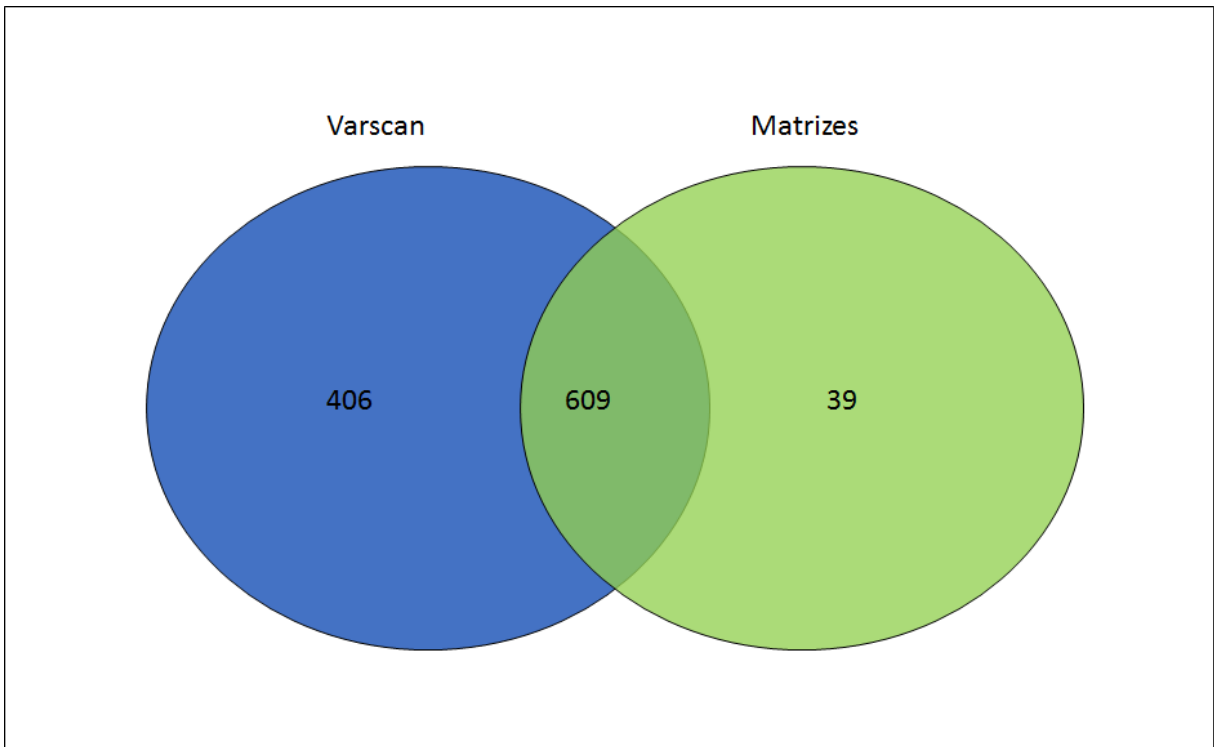
**Figura 4.18** Pequena deleção de um nucleotídeo nas amostras tumorais do paciente 1 (P1T) e paciente 3 (P3T) do gene *BAMBI*.



**Figura 4.19** Pequena deleção de um nucleotídeo nas amostras tumorais do paciente 1 (P1T) do gene PTEN.

Estas deleções estão identificadas em regiões da 3' UTR que não são normalmente alvos de miRNAs. No entanto, as deleções podem causar possíveis mudanças na sequência, tornando-as novos alvos para estas pequenas moléculas de RNA não codificantes (Dusl et al., 2015).

Uma análise adicional que fizemos foi a busca por potenciais candidatos a mutações germinativas ao serem detectadas tanto em dados de tecido normal e nos dados de tecido tumoral. Mutações germinativas já foram descritas como associadas ao câncer de pulmão e ao consumo de tabaco, como deleções no gene *TP53* (Gibbons et al., 2014) e diferentes mutações no gene *KRAS* (Ahrendt et al., 2001). Enquanto isso, para pacientes não fumantes deste tipo de câncer de pulmão se destacam deleções no gene *EGFR* e fusões nos genes *ROS1* e *ALK* (Govindan et al., 2012). Assim, identificamos um total de 648 pequenas deleções que ocorriam em nos tecidos normal e tumoral da mesma paciente. Destas 648 deleções, 609 também foram encontradas pelo Varscan (60% do total encontrado por este programa) (Figura 4.20).



**Figura 4.20** Pequenas deleções identificadas utilizando as matrizes ternárias (verde) e utilizando o Varscan (azul) encontradas, ao mesmo tempo, em amostras de tecido normal e tumoral de RNA-Seq de seis pacientes de câncer de pulmão (Kim et al., 2013a).

Entre os genes afetados em tecidos normais e tumorais dos mesmos pacientes, podemos destacar o *CTSA* e o *EIF3A*. A deleção de quatro nucleotídeos que ocorre no gene *CTSA* foi encontrada em média em 44 das 254,5 leituras (Tabela 4.5) localizados na região com a mutação no paciente 8 (Figura 4.21). Esta deleção altera o quadro de leitura da tradução da proteína podendo gerar uma proteína truncada com o domínio peptidase sendo afetado (Figura 4.22). Este gene codifica uma proteína com importante função protetiva para enzimas lisossomais como a  $\beta$ -galactosidase e neuramidase (van der Spoel et al., 1998). Deleções que alteram o quadro de leitura nesse gene já foram encontradas em outros adenocarcinomas como de cólon (Mouradov et al., 2014) e próstata (Kumar et al., 2016).

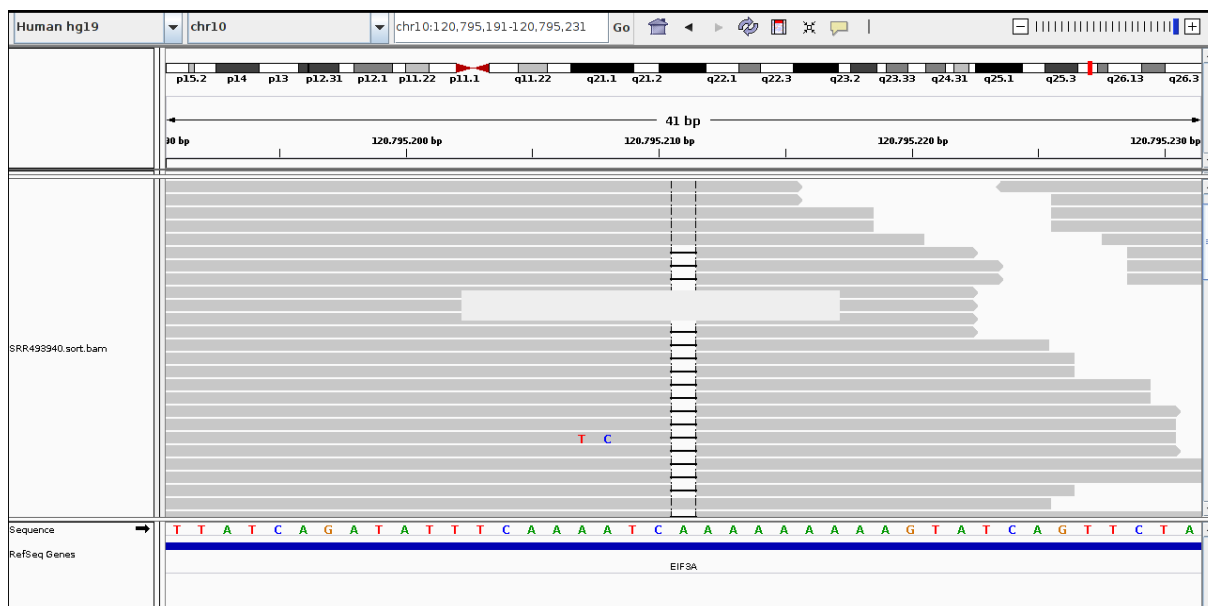
A deleção de um nucleotídeo encontrada no gene *EIF3A* foi identificada em todos pacientes (em média de em 32 das 38,5 leituras do paciente 1, em 29 das 66 leituras do paciente 3, em 10 das 39 leituras do paciente 4, em 18,5 das 29,5 leituras do paciente 5, em 9,5 das 42,5 leituras do paciente 6 e em 11 das 28 leituras do paciente 8) (Tabela 4.5), também foi identificada pelo programa Varscan e está localizada na região 3' UTR (Figura 4.23). A alta expressão deste gene está

correlacionada ao melhor resposta de tratamentos baseados em compostos de platina em câncer de pulmão (Yin et al., 2011). Esta deleção, assim como a deleção para o gene *CTSA* pode ser consideradas potenciais candidatas a mutações germinativas para pacientes não fumantes.

**Tabela 4.5.** Deleções identificadas em dados de RNA-Seq ao mesmo tempo em amostras normais e tumorais de Kim e colaboradores (2013a) e a cobertura das leituras.

Gene	Localização	Amostra	Total de leituras na região com deleção (leituras com deleções)
<i>CTSA</i>	chr20:44520258-44520261	P8N	240 (42)
		P8T	269 (46)
<i>EIF3A</i>	chr10:120795211-120795211	P1N	30 (22)
		P1T	47 (34)
<i>EIF3A</i>	chr10:120795211-120795211	P3N	72 (30)
		P3T	60 (28)
<i>EIF3A</i>	chr10:120795211-120795211	P4N	36(10)
		P4T	42 (10)
<i>EIF3A</i>	chr10:120795211-120795211	P5N	29 (15)
		P5T	30 (22)
<i>EIF3A</i>	chr10:120795211-120795211	P6N	40 (9)
		P6T	45 (10)
<i>EIF3A</i>	chr10:120795211-120795211	P8N	31 (13)
		P8T	25 (9)





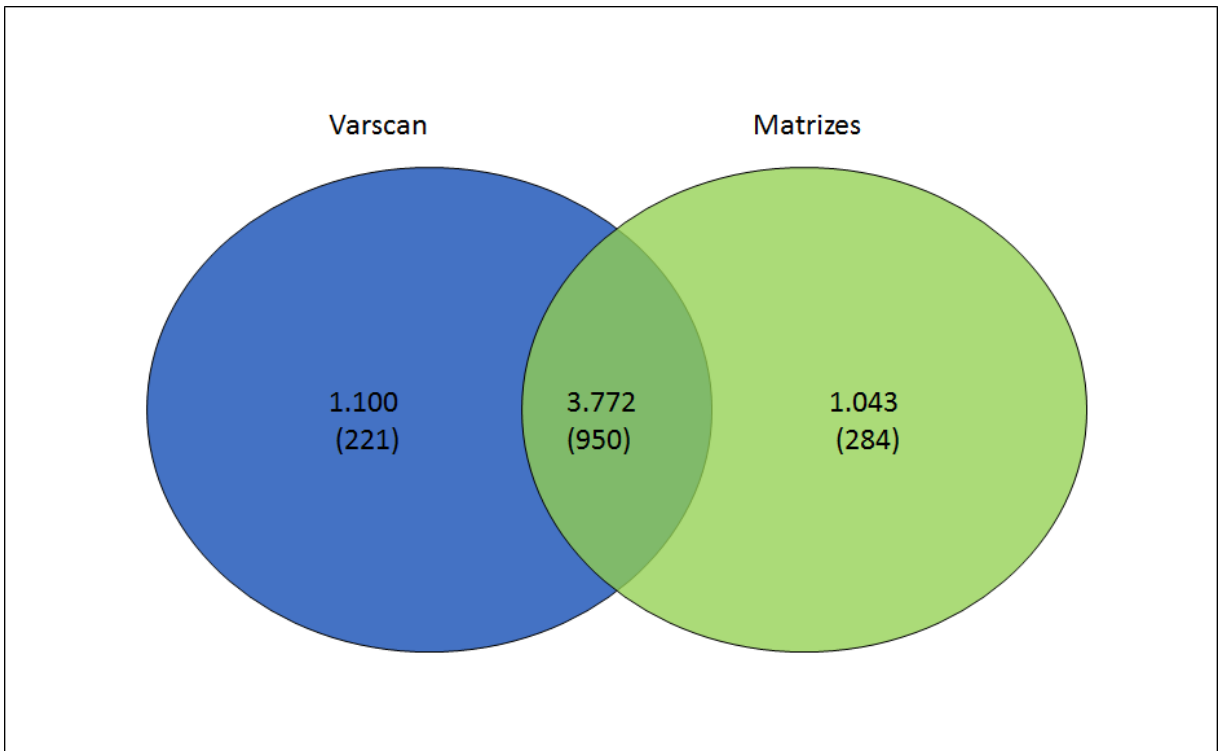
**Figura 4.23.** Pequena deleção de 1 nucleotídeo visualizada em amostras normais e tumorais de todos pacientes no gene *EIF3A*.

Vale ressaltar que não encontramos deleções com a mesma coordenada cromossômica nos dados de Kim e colaboradores (2013a) (pacientes não fumantes) e os dados da linhagem celular H1975, oriunda de um paciente não fumante.

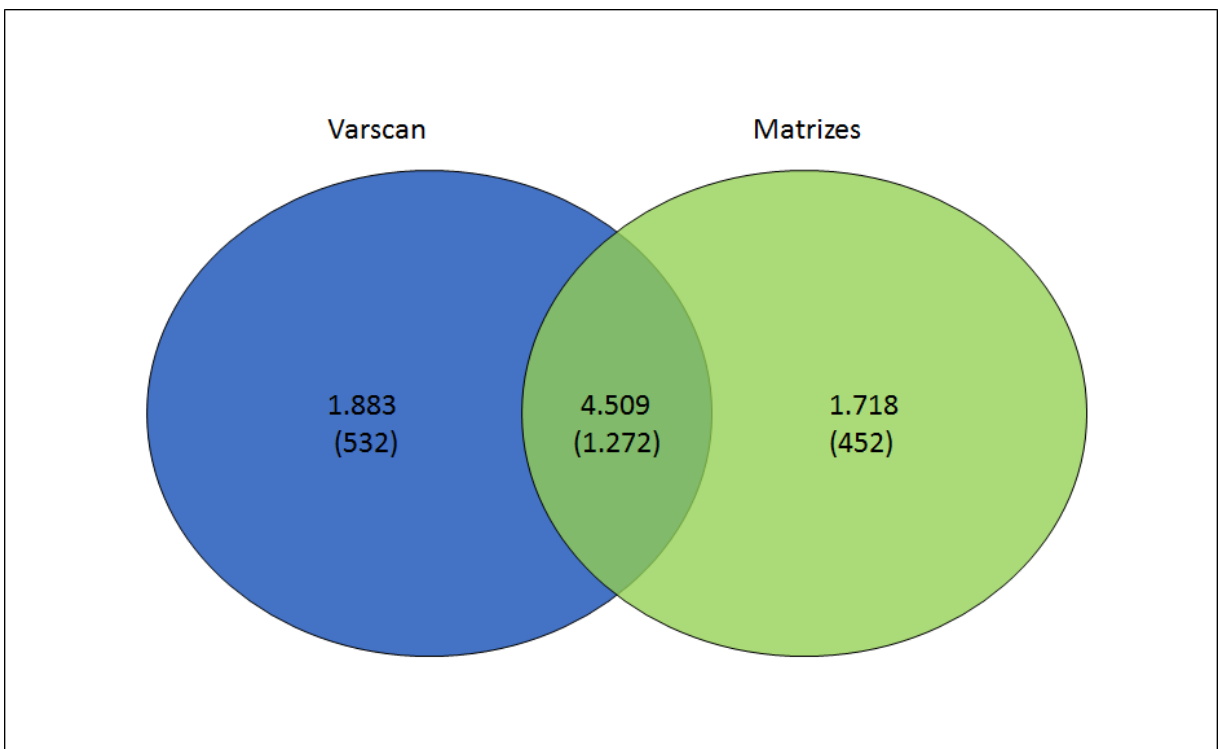
Não pudemos comparar os nossos resultados com os achados de Kim e colaboradores (2013a), porque estes não apresentam sem seus resultados a identificação de deleções. Os autores relatam apenas que mostram a 47 mutações somáticas de troca de um nucleotídeo que foram alvo de validação experimental (Kim et al., 2013a).

#### **4.1.3 Pequenas deleções identificadas em amostras normais e tumorais de 14 pacientes fumantes de câncer de pulmão**

Realizamos a busca de pequenas deleções em outro conjunto de dados, desta vez em pacientes fumantes com adenocarcinoma de pulmão cujos dados de sequenciamento estão depositados no banco de dados TCGA e publicados por Collisson e colaboradores (2014). Identificamos 4.815 pequenas deleções em amostras normais (1.234 em mais de um paciente, 87 em todos pacientes) (Figura 4.24) e outras 6.227 em amostras tumorais (1.724 em mais de um paciente, 85 em todos pacientes) (Figura 4.25).



**Figura 4.24** Pequenas deleções identificadas utilizando as matrizes ternárias (verde) e utilizando o Varscan (azul) encontradas em amostras de tecido normal adjacente ao tumor de RNA-Seq de 14 pacientes de câncer de pulmão do TCGA. Entre parênteses o valor de deleções identificadas em mais de um paciente.



**Figura 4.25** Pequenas deleções identificadas utilizando as matrizes ternárias (verde) e utilizando o Varscan (azul) encontradas em amostras de tecido tumoral de RNA-Seq de 14 pacientes de câncer de pulmão do TCGA. Entre parênteses o valor de deleções identificadas em mais de um paciente.



Comparamos as pequenas deleções identificadas com o banco de dados dbSNP e identificamos 270 pequenas deleções anotadas no dbSNP (6%) em amostras normais e 247 em amostras tumorais (4%) (Tabela 4.6). Encontramos 221 pequenas deleções em amostras normais e 188 em amostras tumorais que estavam anotadas na base de dados COSMIC (Tabela 4.6).

**Tabela 4.6** Representação de quantas deleções estão anotadas na base de dados COSMIC em amostras normais e tumorais.

Tipo de amostra	Tipo de amostra	Em regiões não codificadoras	Em regiões codificadoras
dbSNP	Normal	198	72
	Tumoral	191	56
COSMIC	Normal	193	28
	Tumoral	173	15

As deleções identificadas por nossa metodologia estavam em regiões cobertas em média por 115 leituras, sendo que cada deleção é representada pela média de 18 leituras (com desvio padrão de 6). O valor máximo de leituras representando uma deleção foi de 50 e o mínimo de 3. Identificamos 8 pequenas deleções com relevância em nossos dados (Tabela 4.7)

**Tabela 4.7** Deleções identificadas em dados de RNA-Seq nas amostras de Collisson e colaboradores (2014) e a cobertura das leituras.

(continua)

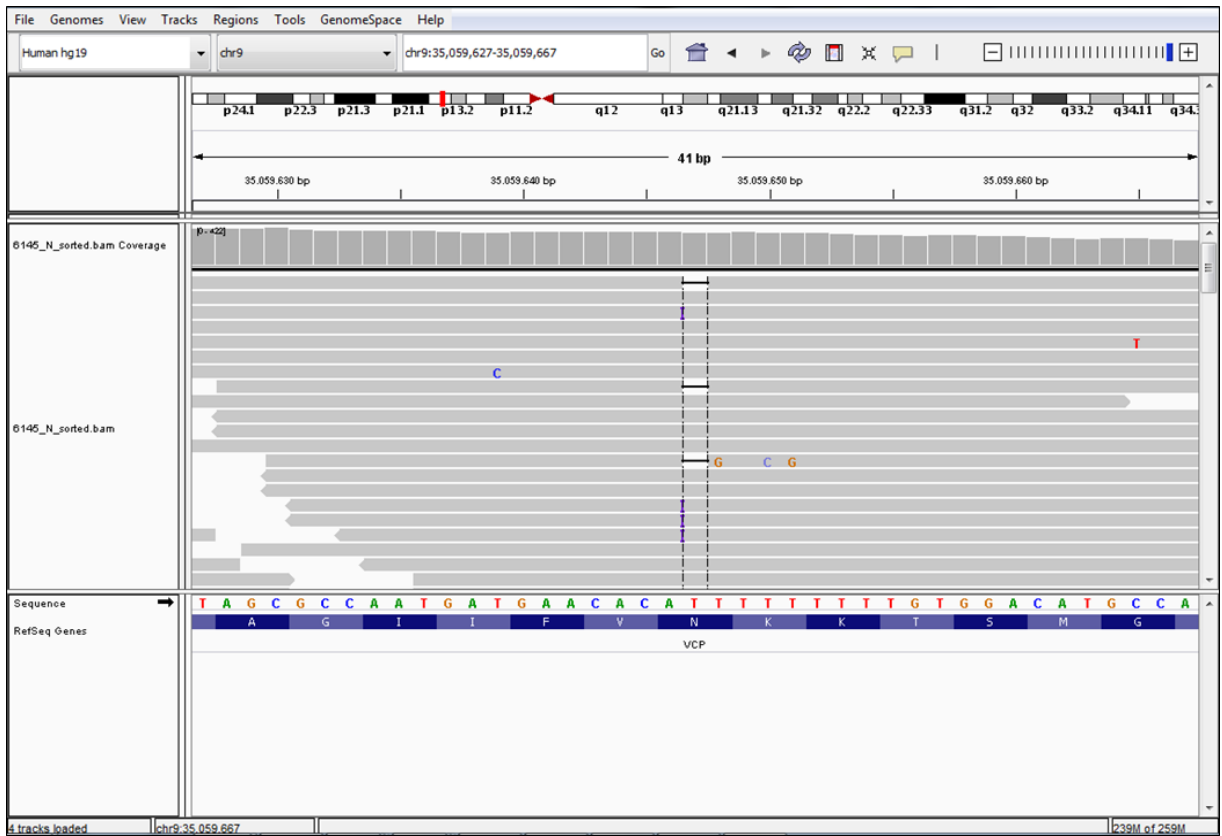
Gene	Localização	Amostra	Total de leituras na região com deleção (leituras com deleções)
<i>VCP</i>	chr9:35059647-35059647	2655N	200 (12)
<i>VCP</i>	chr9:35059647-35059647	2657N	139 (15)
<i>VCP</i>	chr9:35059647-35059647	2668N	240 (10)
<i>VCP</i>	chr9:35059647-35059647	3398N	97 (8)
<i>VCP</i>	chr9:35059647-35059647	5645N	353 (12)
<i>VCP</i>	chr9:35059647-35059647	6145N	328 (9)
<i>VCP</i>	chr9:35059647-35059647	6146N	231 (18)
<i>VCP</i>	chr9:35059647-35059647	6147N	299 (9)
<i>VCP</i>	chr9:35059647-35059647	6777N	302 (10)
<i>ZFP28</i>	chr19:57065508-57065508	2655T	9 (3)
<i>IFNGR1</i>	chr6:137518967-137518967	2655T	98 (10)
<i>IFNGR1</i>	chr6:137518967-137518967	2662T	312 (12)
<i>IFNGR1</i>	chr6:137518967-137518967	3396T	82 (10)
<i>IFNGR1</i>	chr6:137518967-137518967	3398T	212 (15)
<i>IFNGR1</i>	chr6:137518967-137518967	6146T	120 (16)
<i>IFNGR1</i>	chr6:137518967-137518967	6148T	132 (15)
<i>IFNGR1</i>	chr6:137518967-137518967	6776T	230 (18)
<i>IFNGR1</i>	chr6:137518967-137518967	6777T	112 (7)
<i>IFNGR2</i>	chr21:34809434-34809434	2655T	310 (20)
<i>IFNGR2</i>	chr21:34809434-34809434	2657T	152 (13)
<i>IFNGR2</i>	chr21:34809434-34809434	2662T	536 (22)
<i>IFNGR2</i>	chr21:34809434-34809434	2668T	402 (18)
<i>IFNGR2</i>	chr21:34809434-34809434	3396T	430 (32)
<i>IFNGR2</i>	chr21:34809434-34809434	3398T	220 (15)
<i>IFNGR2</i>	chr21:34809434-34809434	6145T	502 (13)
<i>IFNGR2</i>	chr21:34809434-34809434	6776T	421 (18)
<i>IFNGR2</i>	chr21:34809434-34809434	6777T	228 (13)
<i>VEGFC</i>	chr10:28971526-28971526	2655T	25 (19)
<i>VEGFC</i>	chr10:28971526-28971526	2662T	32 (16)

**Tabela 4.7** Deleções identificadas em dados de RNA-Seq nas amostras de Collisson e colaboradores (2014) e a cobertura das leituras.

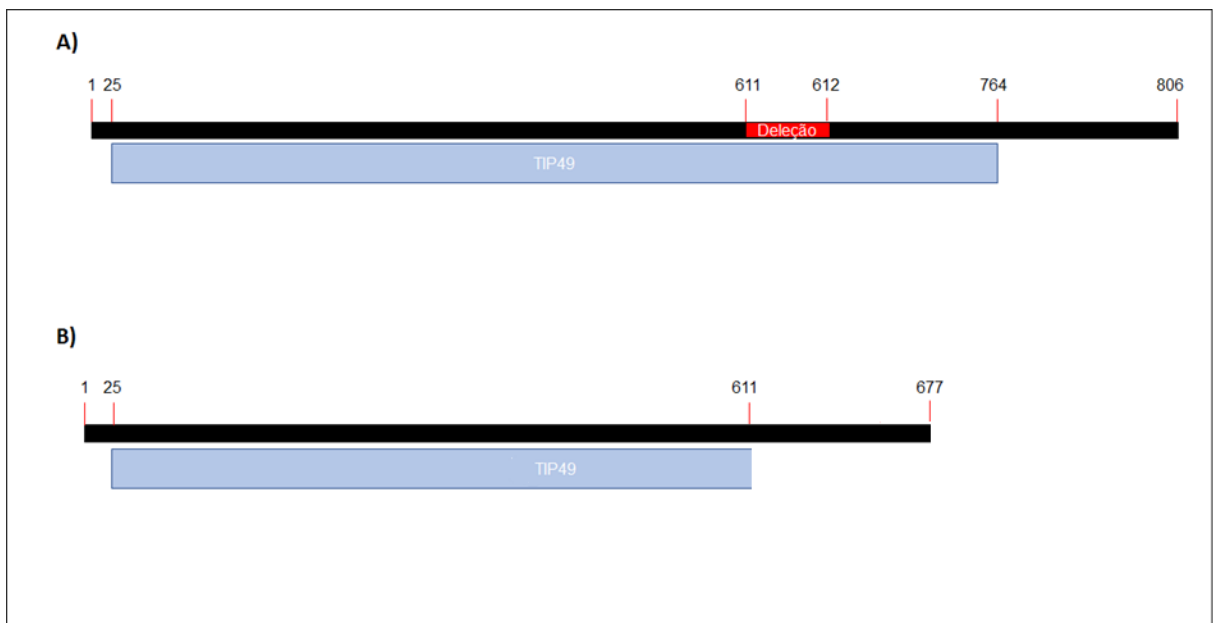
(conclusão)

Gene	Localização	Paciente	Total de leituras na região com deleção (leituras com deleções)
<i>VEGFC</i>	chr10:28971526-28971526	3396T	29 (12)
<i>VEGFC</i>	chr10:28971526-28971526	3398T	16 (8)
<i>VEGFC</i>	chr10:28971526-28971526	5645T	36 (20)
<i>VEGFC</i>	chr10:28971526-28971526	6145T	17 (16)
<i>VEGFC</i>	chr10:28971526-28971526	6148T	20 (12)
<i>VEGFC</i>	chr10:28971526-28971526	6777T	12 (7)
<i>VEGFC</i>	chr10:28971526-28971526	6778T	18 (15)

Como exemplo, temos a deleção em amostras normais de 10 pacientes ((em 12 leituras das 200 no paciente 2655, em 15 leituras das 139 no paciente 2657, em 10 leituras dos 240 no paciente 2668, em 8 leituras das 97 no paciente 3398, em 12 leituras das 353 no paciente 5645, em 9 leituras das 328 no paciente 6145, em 18 leituras das 231 no paciente 6146, em 9 leituras das 299 no paciente 6147, em 10 leituras das 302 no paciente 6148, em 10 leituras das 302 no paciente 6777)) de um único nucleotídeo no gene *VPC* (Figura 4.26). Esta deleção altera o quadro de leitura da tradução e isso afeta a proteína traduzida que perde 129 aminoácidos em seu domínio principal (Figura 4.27). Mutações nesse gene podem levar à ativação inapropriada do fator de transcrição NFκB (Sahab et al., 2010). Apesar de que mutações no gene *VCP* não foram associadas ao câncer, o NFκB é importante para desencadear vias da carcinogênese em câncer de pulmão de indivíduos fumantes (Chen et al., 2011b).



**Figura 4.26.** Pequena deleção de um nucleotídeo visualizada em amostra normal do paciente 6145 no gene *VCP*.

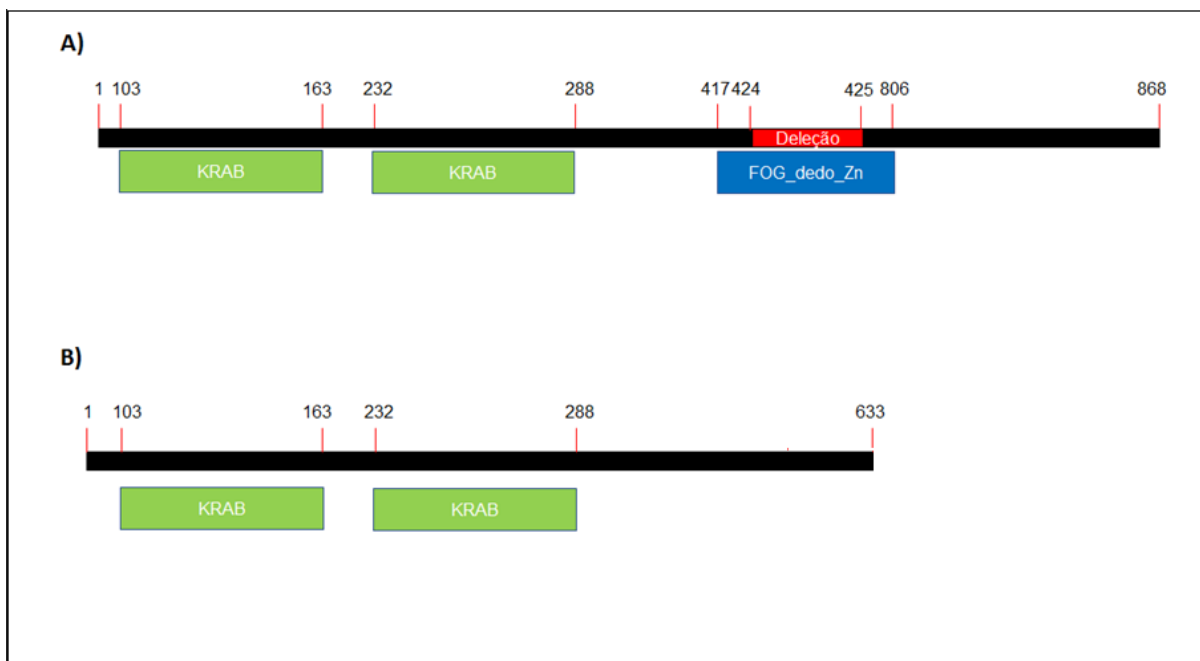


**Figura 4.27** Representação do impacto da pequena deleção de um nucleotídeo no gene *VCP* na sequência de aminoácidos. A proteína normal **(A)** possui 806 aminoácidos e um domínio TIP49 (Domínio C-Terminal TIP49, acesso cl27568). A proteína afetada **(B)** pela deleção perdeu 129 aminoácidos inclusive uma parte do domínio TIP49.

Encontramos em tecido tumoral, uma deleção de um único nucleotídeo ao mesmo tempo utilizando o VarScan e as matrizes ternárias no paciente 2655 (Figura 4.28). Esta deleção altera o quadro de leitura da tradução da proteína. Como resultado fica truncada com ausência do seu domínio de dedo de zinco (Figura 4.29). Os domínios dedo de zinco são responsáveis na ligação da proteína com o DNA, portanto mutações que perdem este domínio pode levar a não ligação da proteína com o DNA (Brayer and Segal, 2008). Uma deleção parecida foi identificada em adenocarcinoma de próstata que causa a perda desse domínio na proteína (Kumar et al., 2016). A proteína codificada por este gene regula a transcrição de diversos genes, inclusive o supressor tumoral *WT1* (Morrison et al., 2008).



**Figura 4.28** Pequena deleção de um nucleotídeo visualizada em amostra tumoral do paciente 2655 no gene *ZFP28*.



**Figura 4.29** Representação do impacto da pequena deleção de um nucleotídeo no gene *ZFP28* na sequência de aminoácidos que possui três domínios: dois domínios KRAB (Domínio associado "krueppel box", acesso smart00349) e FOG\_dedo\_Zn (Dedo de zinco FOG, acesso COG5048). A proteína normal **(A)** codificada possui sítios de ligação de dedo de zinco, enquanto a proteína alterada **(B)** pela deleção não possui.

Realizamos a análise por enriquecimento de genes que apresentaram pequenas deleções apenas nas amostras tumorais e encontramos vias com alta probabilidade de estarem alteradas como, por exemplo, adesão célula-célula (com probabilidade posterior de 1,00), via de sinalização mediada pelo interferon- $\gamma$  (com probabilidade posterior de 0,92), cicatrização (com probabilidade posterior de 0,78) e processamento de O-glicanos (com probabilidade posterior de 0,65) (Tabela 4.8).

**Tabela 4.8** Enriquecimento de genes que sofreram pequenas deleções em amostras tumorais identificadas pela nossa metodologia de matrizes ternárias. Abaixo encontramos as vias do Gene Ontology (GO) com probabilidade posterior maior do que 0,5 de estarem suprerrepresentadas .

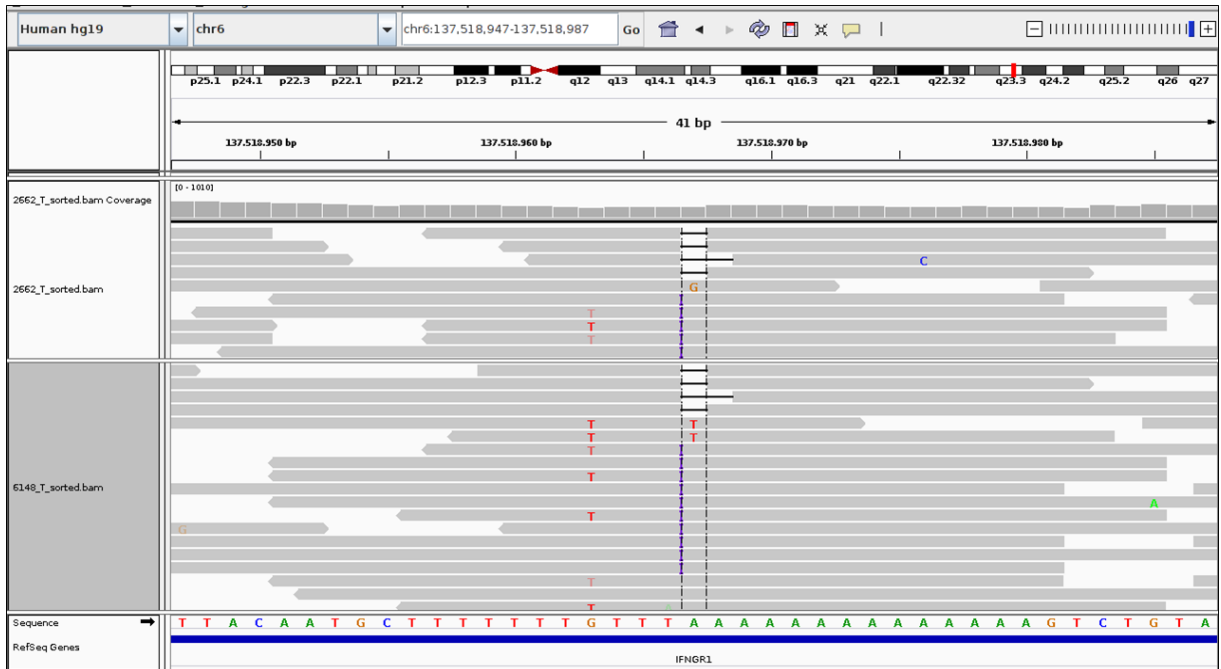
GO id	Descrição	Probabilidade posterior	Quantidade de Genes
GO:0098609	Adesão célula-célula	1,0	85
GO:0060333	Via de sinalização mediada pelo interferon- $\gamma$	0,92	27
GO:0042060	Cicatrização	0,78	32
GO:0016266	Processamento de O-glicanos	0,65	22

Analisamos também pequenas deleções em regiões não codificadoras. Encontramos duas deleções em regiões 3' UTR em dois genes associados com a via de sinalização mediada pelo interferon- $\gamma$ . Uma deleção de um nucleotídeo em 8 pacientes no gene *IFNGR1* (em 40 leituras das 98 no paciente 2655, em 12 leituras das 312 no paciente 2662, em 10 leituras das 82 no paciente 3396, em 15 leituras das 212 no paciente 3398, em 16 leituras das 120 no paciente 6146, em 15 leituras das 132 no paciente 6148, em 18 leituras das 230 no paciente 6776 e em 7 leituras das 112 no paciente 6777) (Figura 4.30).

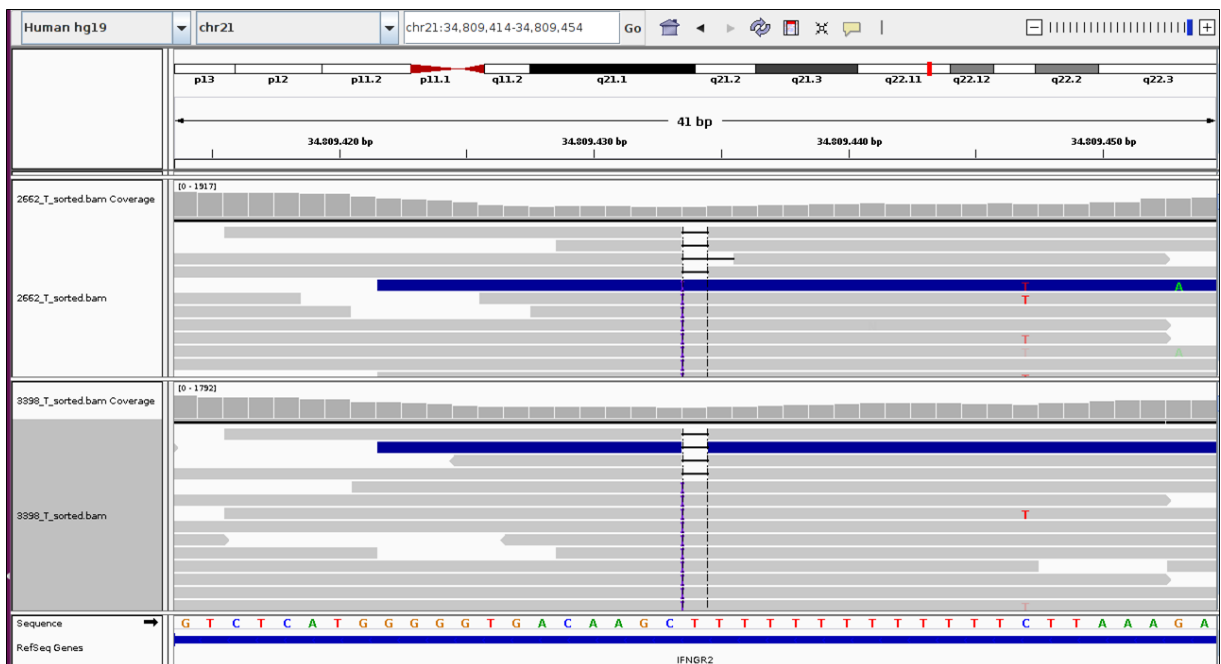
Encontramos também uma deleção de 1 nucleotídeo em 9 pacientes no gene *IFNGR2* (em 20 leituras das 310 no paciente 2655, em 13 leituras das 152 no paciente 2657, em 22 leituras das 536 no paciente 2662, em 18 leituras das 402 no paciente 2668, em 32 leituras das 430 no paciente 3396, em 18 leituras das 402 no paciente 3398, em 13 leituras das 502 no paciente 6145, em 18 leituras das 421 no paciente 6776, em 13 leituras das 228 no paciente 6777) (Figura 4.31).

A deleção encontrada em *IFNGR1* (Figura 4.31) é na mesma posição encontrada nos dados de Kim e colaboradores (2013a) (Figura 4.17). A deleção encontrada em *IFNGR2* também está presente no dbSNP sob o identificador "rs756768721" (Figura 4.32). Dentre as amostras tumorais de 14 pacientes do TCGA, encontramos 6 deles possuindo a deleção em ambos os genes desta família de receptores, 5 deles possuindo a deleção em um dos genes e apenas 3 deles não

possuíam deleção em nenhum dos genes (pacientes 6147, 5645 e 6778). Apesar de não termos identificado nenhum miRNA que tem como alvo as regiões em que ocorrem estas deleções, a sequência de nucleotídeos pode ter sido afetada de uma forma de que pudessem ser neste momento novos alvos para miRNA.



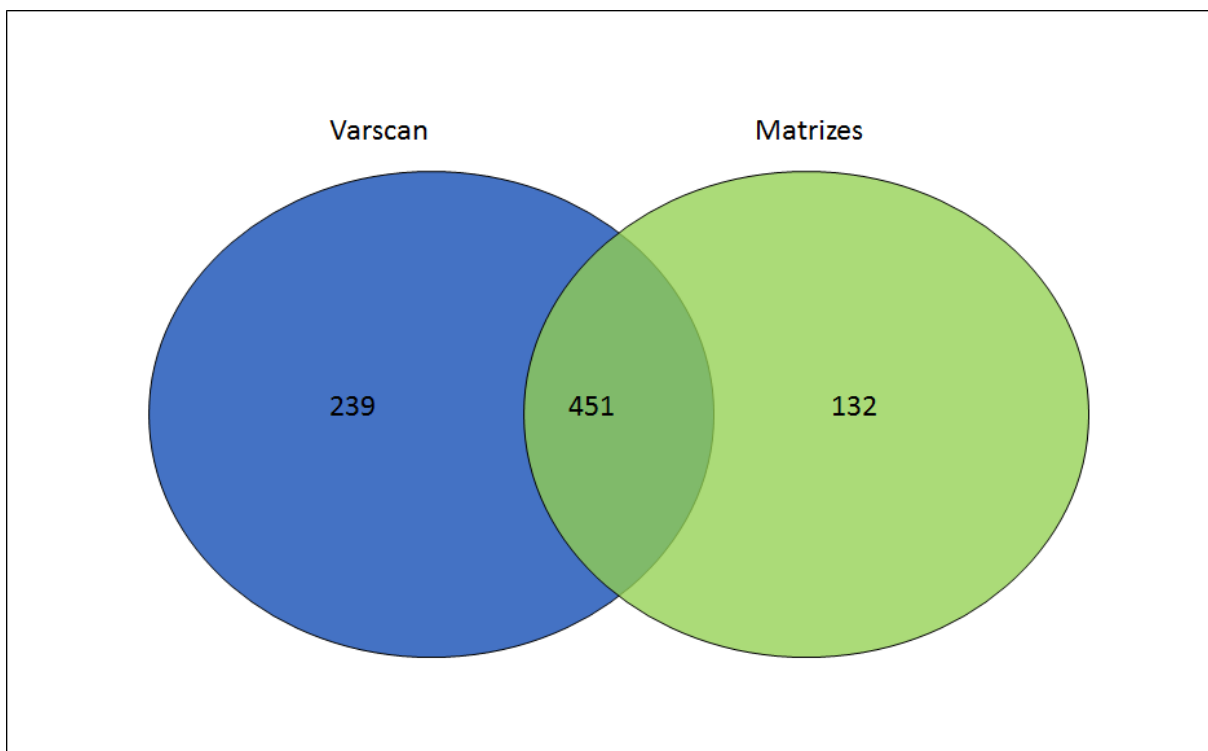
**Figura 4.30** Pequena deleção de um nucleotídeo visualizada em amostra tumoral do paciente 2662 e 6148 no gene *IFNGR1*.



**Figura 4.31** Pequena deleção de um nucleotídeo visualizada em amostra tumoral do paciente 2662 e 3398 no gene *IFNGR2*.



Buscamos também potenciais candidatos para mutações germinativas nestas amostras. Um total de 583 pequenas deleções foram encontradas ao mesmo tempo em amostras normais e tumorais, sendo 451 delas também foram identificadas pelo programa Varscan (65% do total identificado por este programa) (Figura 4.32).



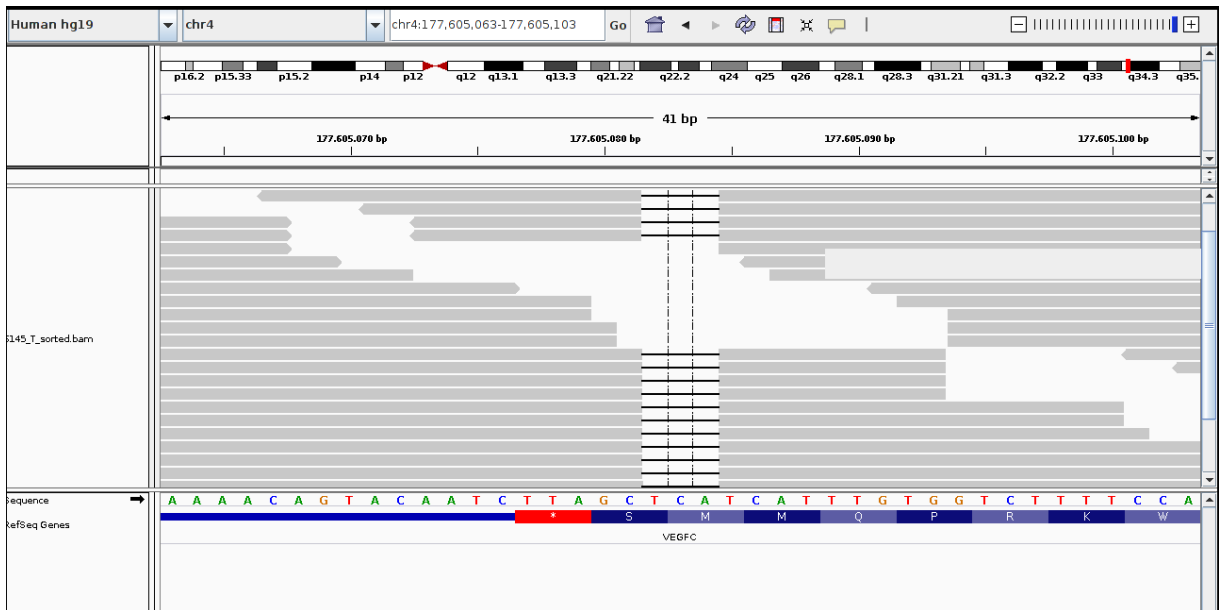
**Figura 4.32** Pequenas deleções identificadas por nossa metodologia utilizando as matrizes ternárias (verde) e utilizando o Varscan (azul) encontradas ao mesmo tempo em amostras de tecido normal e tumoral de RNA-Seq de 14 pacientes de câncer de pulmão de (Collisson et al., 2014).

Uma destas deleções afetou o final da região codificadora do gene *VEGFC* em nove pacientes (em média: 19,5 leituras das 27 no paciente 2655, em 12 leituras das 26,5 no paciente 2662, em 11 leituras das 25 no paciente 3396, em 19 leituras das 9 no paciente 3398, em 19 leituras das 31,5 no paciente 5645, em 12 leituras das 16 no paciente 6145, em 15,5 leituras das 20 no paciente 6148, em 6 leituras das 15,5 no paciente 6777 e em 41 leituras das 15,5 no paciente 6778) (Tabela 4.9) (Figura 4.33). Este gene codifica um ligante responsável em ligar ao seu receptor VEGFR, estimulando o crescimento de novos vasos sanguíneos para irrigar o tumor. Esta deleção diminuiu em um aminoácido o tamanho da proteína codificada (Figura 4.34). Os genes da família VEGF são muito utilizados como alvos terapêuticos e a deleção identificada pode servir para aumentar a resistência a essas drogas (Piperdi et al., 2014). No entanto a deleção observada afeta apenas o final da proteína e 1

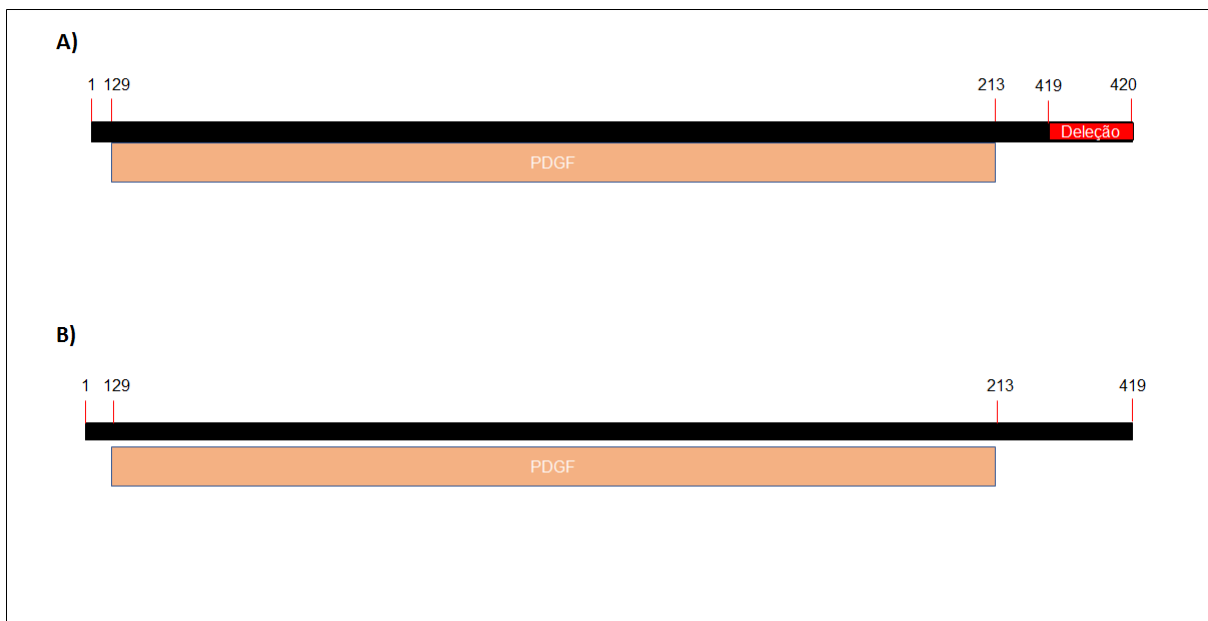
aminoácido é perdido. Os domínios dessa proteína estão próximos da extremidade N-terminal da proteína e por isso não são afetados (Figura 4.34). De forma análoga que encontramos as deleções dos genes CTSA e EI3FA como germinativas para pacientes fumantes e não fumantes, encontramos que a deleção no gene *VEGFC* pode ser considerada germinativa e exclusiva para pacientes fumantes.

**Tabela 4.9** Deleções identificadas em dados de RNA-Seq nas amostras de Collisson e colaboradores (2014) e a cobertura das leituras.

Gene	Localização	Amostra	Total de leituras na região com deleção (leituras com deleções)
<i>VEGFC</i>	chr10:28971526-28971526	2655N	29 (20)
		2655T	25 (19)
<i>VEGFC</i>	chr10:28971526-28971526	2662N	21 (8)
		2662T	32 (16)
<i>VEGFC</i>	chr10:28971526-28971526	3396N	25 (11)
		3396T	29 (12)
<i>VEGFC</i>	chr10:28971526-28971526	3398N	22 (10)
		3398T	16 (8)
<i>VEGFC</i>	chr10:28971526-28971526	5645N	27 (18)
		5645T	36 (20)
<i>VEGFC</i>	chr10:28971526-28971526	6145N	15 (8)
		6145T	17 (16)
<i>VEGFC</i>	chr10:28971526-28971526	6148N	20 (10)
		6148T	20 (12)
<i>VEGFC</i>	chr10:28971526-28971526	6777N	19 (5)
		6777T	12 (7)
<i>VEGFC</i>	chr10:28971526-28971526	6778N	23 (16)
		6778T	18 (15)



**Figura 4.33** Pequena deleção de três nucleotídeos visualizada em amostras tumorais e normais de todos os pacientes no gene *VEGFC*.



**Figura 4.34** Representação do impacto da pequena deleção de três nucleotídeos no gene *VEGFC* na sequência de aminoácidos que possui o domínio PDGF (Domínio de ligação ao receptor PDGFR, acesso smart00141). A proteína normal **(A)** codificada possui um aminoácido a mais do que a proteína alterada **(B)** pela deleção.

O banco de dados TCGA contém dados processados além de conter dados brutos de dados de HTS e dentre estes dados, encontramos também dados de polimorfismos identificados a partir apenas de corridas de DNA-Seq dos mesmos pacientes em que analisamos os dados de RNA-Seq (Collisson et al., 2014). Ainda assim, não pudemos utilizar estes dados para comparar com nossos resultados da

identificação de pequenas deleções a partir de dados de RNA-Seq, pois nestes resultados apenas continham variações de substituição de um único nucleotídeo.

#### **4.1.4 Análise de deleções encontradas exclusivamente em amostras normais, tumorais ou presente em ambas**

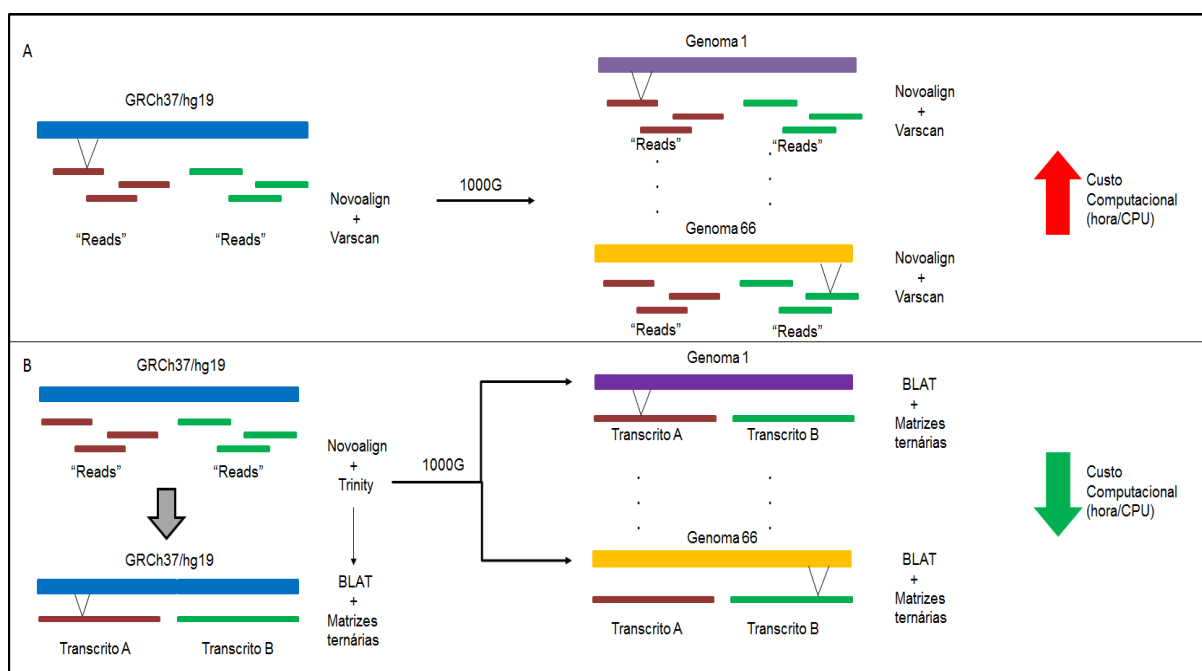
Encontramos um total de 1.409 (Figura 4.9) deleções exclusivas em amostras normais nos dados de Kim e colaboradores (2013a) e 4.232 (Figura 4.24) nos dados de Collisson e colaboradores (2014). Como exemplos de deleções exclusivas de tecidos normais como a deleção de 12 nucleotídeos no *INF2* identificadas em pacientes não fumantes (Figura 4.11) e a deleção de um nucleotídeo no gene *VCP* (Figura 4.26) em pacientes fumantes. Os genes *INF2* e *VCP* não apresentaram expressão nas respectivas amostras tumorais dos pacientes.

Deleções foram também encontradas exclusivamente em tecido tumoral, um total de 1.492 (Figura 4.10) nos dados de Kim e colaboradores (2013a), como a deleção de 15 nucleotídeos no gene *EGFR* (Figura 4.13), outra de dez nucleotídeos no gene *TP53BP2* (Figura 4.15), além de também afetar os genes *IFNGR1* (Figura 4.17), *BAMBI* (Figura 4.18) e *PTEN* (Figura 4.19). Nos dados de Collisson e colaboradores (2014), encontramos 5.644 (Figura 4.25) deleções também exclusivas de tecidos tumorais como a deleção nos genes *ZFP28* (Figura 4.28), *IFNGR1* (Figura 4.30) e *IFNGR2* (Figura 4.31). Em todos os casos descritos acima, encontramos a expressão desses genes sem a deleção nas respectivas amostras pareadas de tecidos normais desses pacientes.

Deleções com grande potencial de serem germinativas foram encontradas em tecidos normais e tumorais de mesmos pacientes nos genes *CTSA* (Figura 4.21) e *EIF3A* (Figura 4.23) nos dados de Kim e colaboradores (2013a) como amostras de um total de 648 deleções (Figura 4.20). Em pacientes fumantes encontramos uma deleção presente em tecido tumoral e normal em todos pacientes no gene *VEGFC* (Figura 4.33) dentre as 583 identificadas (Figura 4.32).

## **4.2 Identificação de pequenas deleções utilizando 66 genomas do 1000G**

A identificação de deleções em amostras de RNA-Seq utilizando como referência 66 genomas humanos diferentes pode ter um elevado gasto computacional caso tenha que ser feito o alinhamento do mesmo conjunto de RNA-Seq para cada um dos genomas para a posterior análise pelo programa Varscan para identificação de deleções. Para cada corrida teríamos então 66 arquivos de dados mapeados contra cada genoma, gerando alto custo computacional e cada mapeamento levaria considerável tempo para ser realizado (Figura 4.35A). Por exemplo, o alinhamento de uma única corrida de RNA-Seq demorou mais de 62 horas utilizando o programa Novoalign para mapeá-lo no genoma humano de referência. Assim, levando em consideração que teríamos 66 sequências de genomas para alinhar, o gasto computacional de tempo de processamento seria elevado. Com a metodologia das matrizes ternárias, o mapeamento do Novoalign é realizado uma única vez no genoma humano de referência para auxiliar na montagem dos transcritos que são reaproveitados (Figura 4.35B). Por estes motivos não utilizamos o Novoalign e o Varscan para identificar pequenas deleções nesta nova etapa. Assim, inovamos ao usar os transcritos recriados pelo programa Trinity, o seu mapeamento usando o programa BLAT em 66 genomas do 1000G e as matrizes ternárias para a identificação de deleções tomando como base dados de mapeamento de RNA-Seq no genoma humano de referência.



**Figura 4.35** Comparação dos dois métodos de identificação de pequenas deleções no genoma de referência GRCh37/hg19 e nos 66 genomas do projeto 1000G sendo eles: utilização do Novoalign junto com o Varscan **(A)** e a utilização do Novoalign e Trinity para a montagem dos transcritos que depois são mapeados utilizando o BLAT e guardados nas matrizes ternárias **(B)**.

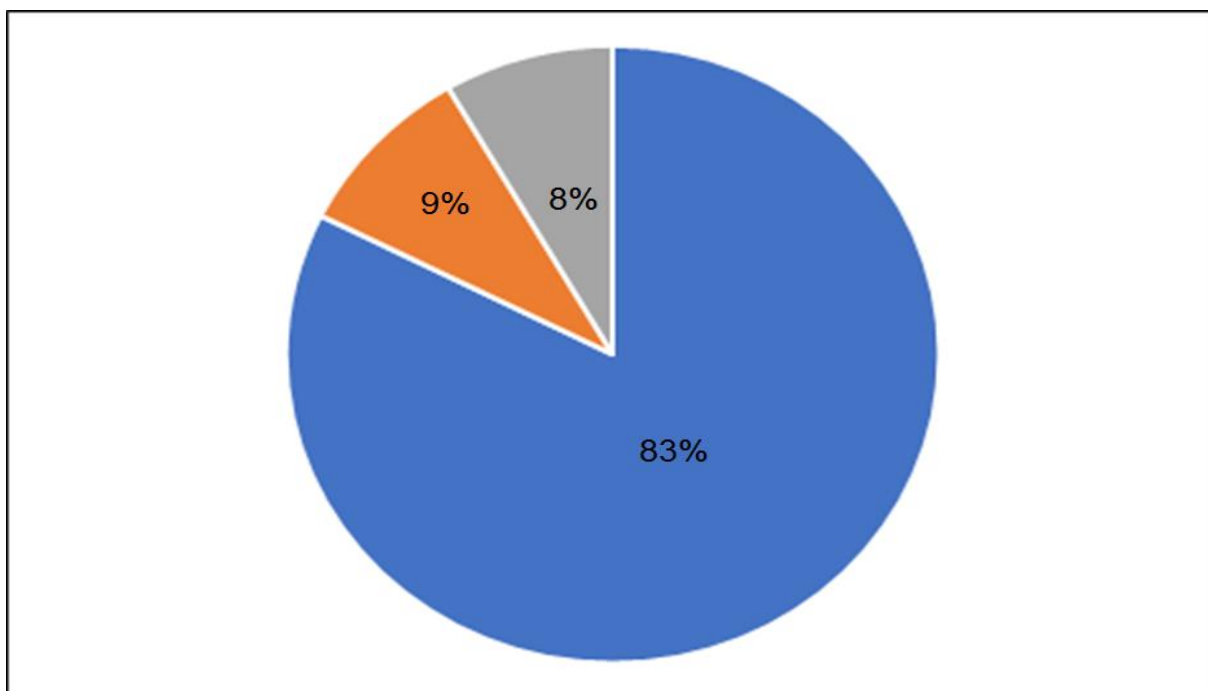
#### 4.2.1 Em amostras da linhagem celular H1975

Tomando como referência 66 genomas do 1000G, encontramos 1.332 pequenas deleções de até 100 nucleotídeos utilizando dados de RNA-Seq da linhagem celular H1975, sendo apenas 52 delas também foram identificadas no genoma de referência GRCh37/hg19. Apenas 834 ocorrem em mais de metade dos 66 genomas do 1000G e não são identificáveis utilizando o genoma de referência GRCh37/hg19. Encontramos 37 deleções com anotação nos bancos de dados dbSNP e COSMIC (Tabela 4.10).

**Tabela 4.10** Tabela representando as deleções identificadas e seu impacto em regiões codificadoras e presentes no dbSNP e COSMIC.

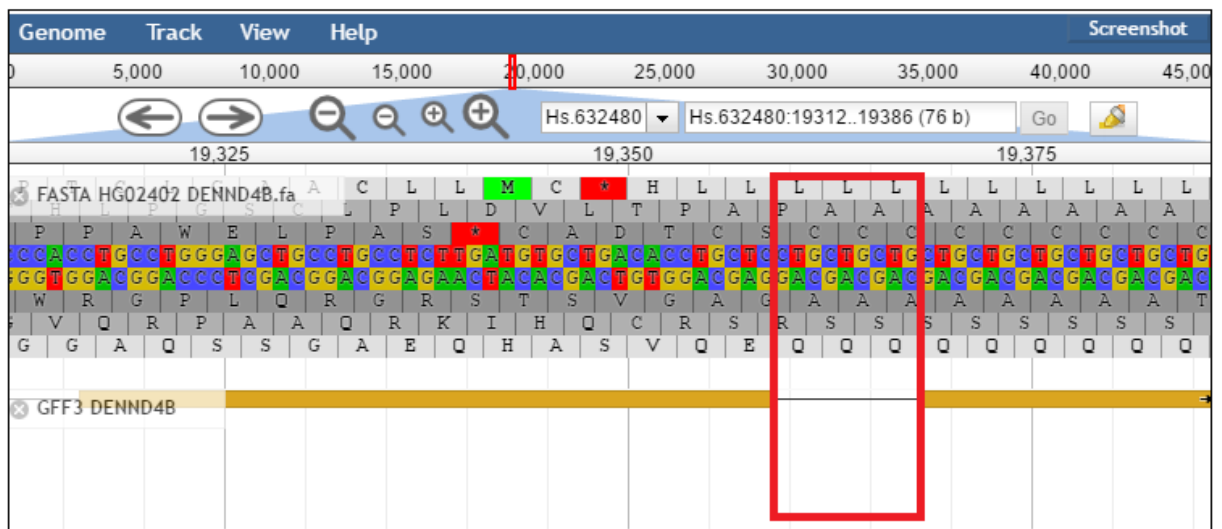
	Matrizes ternárias	dbSNP	COSMIC
Regiões não codificadoras	656	24(3,6%)	12 (1,8%)
Alteram o quadro de leitura	85	0	0
Não alteram o quadro de leitura	93	0	1(1%)

Apenas 178 deleções ocorrem em regiões codificadoras, sendo 85 delas com potencial para alterar o quadro de leitura da tradução e 93 não devem alterar o quadro de leitura (Figura 4.36).



**Figura 4.36** Gráfico mostrando a proporção de pequenas deleções identificadas em linhagens celulares H1975 em regiões não codificadoras (azul), que não alteram o quadro de leitura (laranja) e que alteram o quadro de leitura (cinza).

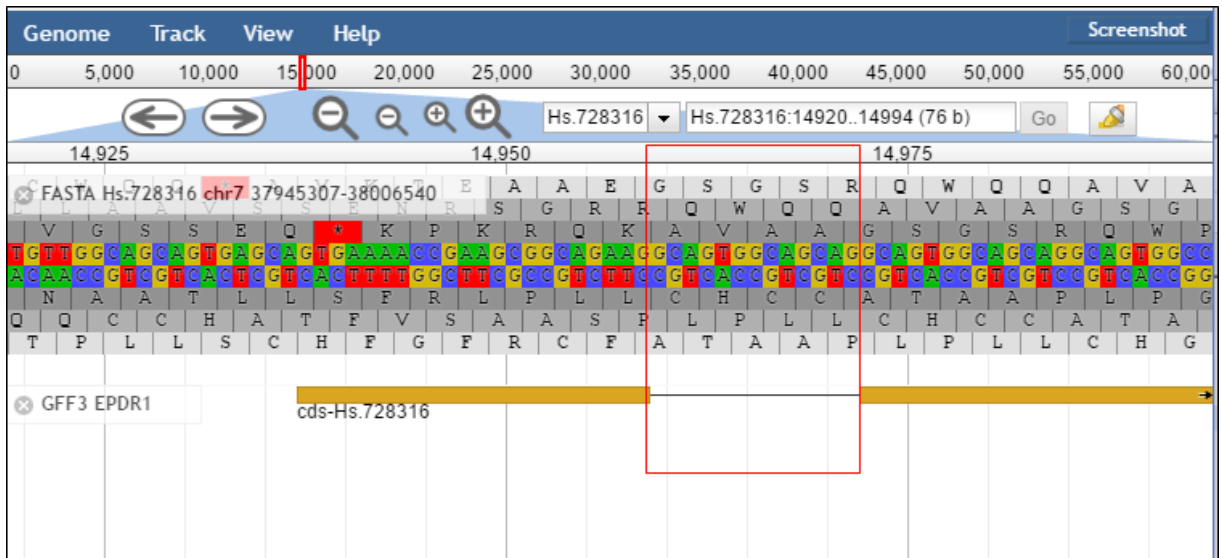
Encontramos uma deleção que possui anotação no COSMIC e que não muda o quadro de leitura da tradução da proteína correspondente ao gene *DENND4B*. Esta deleção corresponde à perda de 9 nucleotídeos e afeta uma região rica em poli glutamina (3 glutaminas seguidas são perdidas). Encontramos essa deleção utilizando 14 genomas de ancestralidade asiática (Figura 4.37). Corroborando o nosso achado, esta deleção já tinha sido identificada em adenocarcinoma de pulmão (Yin et al., 2014). Esta deleção causa a perda de três glutaminas de um domínio poli glutamina da proteína traduzida. Este domínio tem a função de modular a ligação desta proteína com seus alvos reguladores como as Rab GTPases responsáveis pelo transporte de vesículas pela membrana plasmática (Marat et al., 2011).



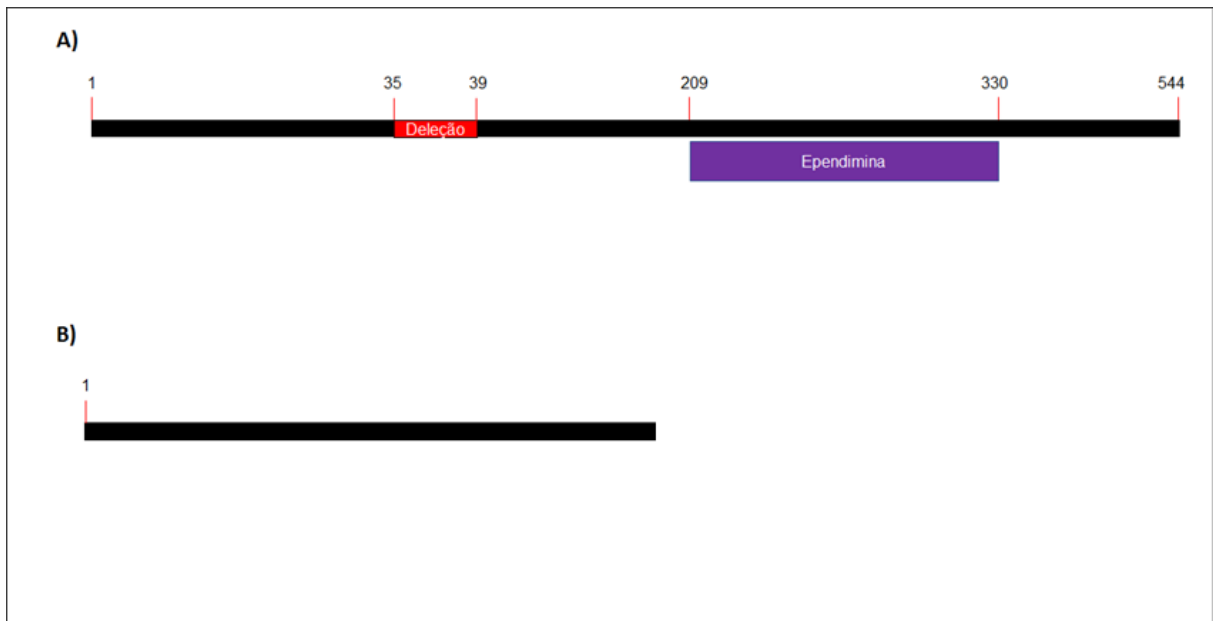
**Figura 4.37.** Deleção de nove nucleotídeos no gene *DENND4B* em um transcrito (em amarelo) montado a partir de corridas de linhagem celular H1975 ao mapear contra um genoma asiático (HG02402) demonstrando a perda de três glutaminas (demarcado vermelho).

A deleção que foi encontrada utilizando a maior quantidade de genomas foi identificada no gene *EPDR1*, totalizando 50 genomas do 1000G (12/12 europeias, 18/27 asiáticas, 7/9 americanas e 12/18 africanas). Este gene codifica uma proteína transmembrana associada com adesão celular. Esta deleção ocasionou a perda de 13 nucleotídeos alterando o quadro de leitura da tradução da proteína (Figura 4.38). Uma deleção com o mesmo tamanho foi identificada na mesma região do gene em pacientes de câncer de pulmão entre as regiões 37.960.263 e 37.960.275 do cromossomo 8 (Lee et al., 2010). No entanto, a deleção identificada por nossa metodologia encontrou esta deleção entre as coordenadas 37.960.265 e 37.960.277 (Figura 4.38). Esta deleção adiantou o códon de parada da tradução em 144 aminoácidos levando a alterar o domínio de endimina da proteína (Figura 4.39). A proteína com a perda deste domínio pode ter sua capacidade de adesão afetada por conta disso. Em pacientes com mieloma, um tipo de câncer que afeta os plasmócitos, este gene foi encontrado com baixa expressão que está associada a progressão tumoral (Wu et al., 2015).



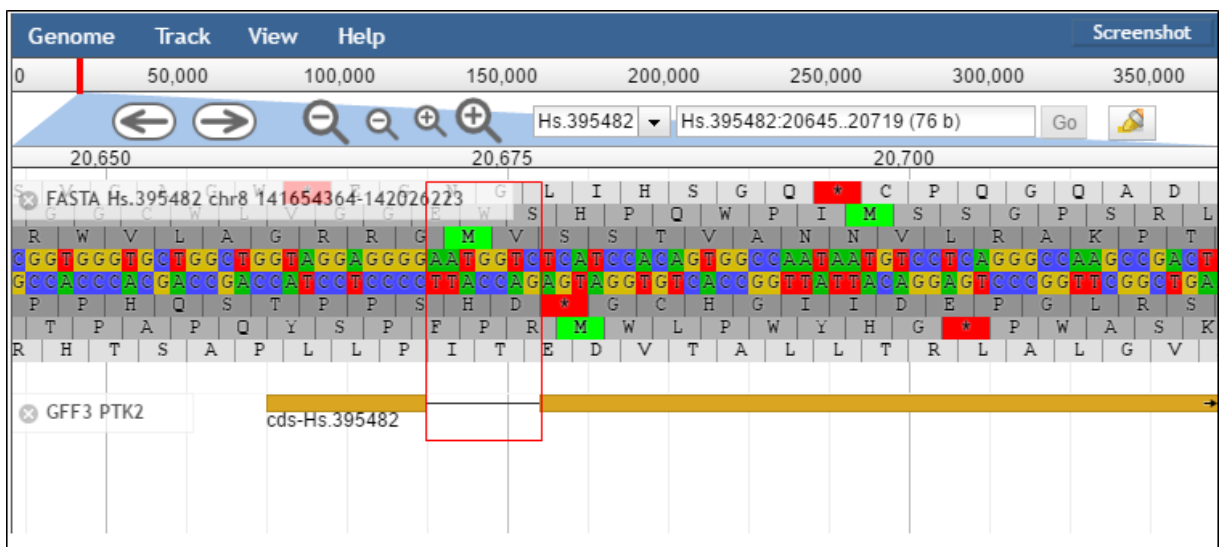


**Figura 4.38.** Deleção de 13 nucleotídeos em um transcrito (em amarelo) montado a partir de dados da linhagem celular H1975 ao mapear contra um genoma europeu (HG00151) no gene *EPDR1* (demarcado em vermelho) causando uma mutação que altera o quadro de leitura da tradução da proteína.

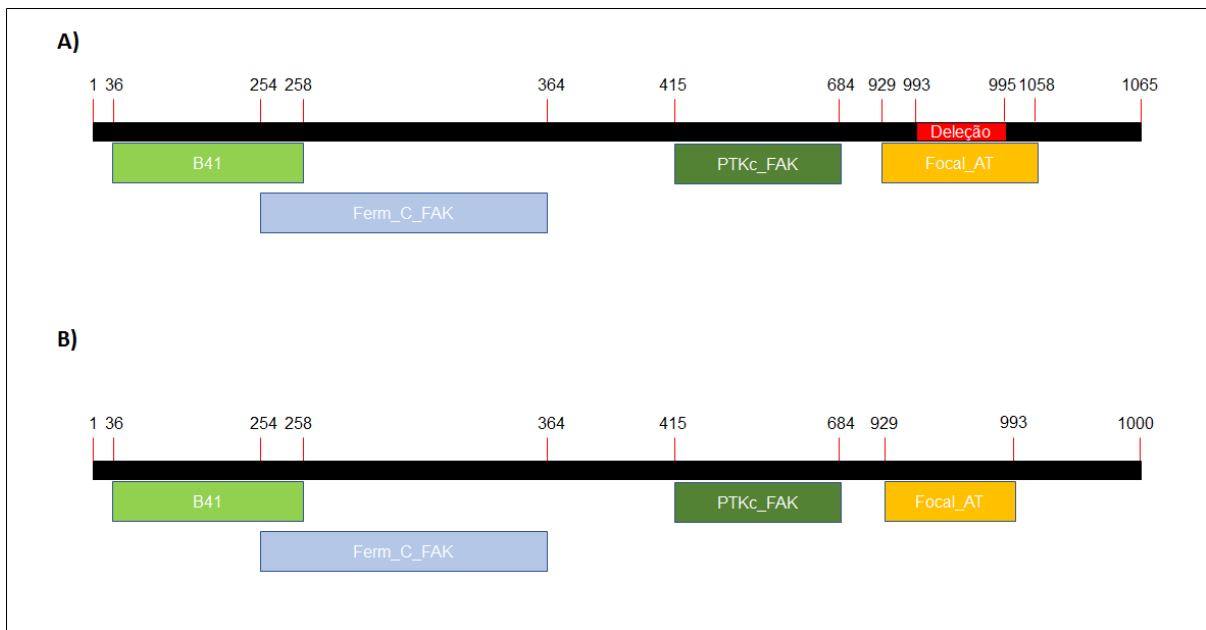


**Figura 4.39** Representação do impacto da pequena deleção de 13 nucleotídeos no gene *EPDR1* na sequência de aminoácidos. A proteína normal (A) codificada pelo gene possui um domínio Ependimina (Domínio ependimina, acesso pfam00811) enquanto a proteína afetada (B) pela deleção mostra este domínio perdido.

Outra deleção encontrada utilizando a maior parte dos genomas de ancestralidade europeia foi no gene *PTK2*. O tamanho desta deleção foi de 7 nucleotídeos (Figura 4.40). Esta deleção foi encontrada utilizando 34 genomas (12/12 europeus, 5/9 americanas, 10/27 asiáticas e 7/18 africana) e adianta o códon de parada da tradução em 65 aminoácidos. Não encontramos na literatura deleções que causem mudança no quadro de leitura em câncer de pulmão na literatura, porém uma deleção do mesmo tamanho foi encontrada em pacientes com melanoma (Sanborn et al., 2015). O domínio de adesão focal kinase foi o principal afetado por este adiantamento do códon de parada (Figura 4.41). Este domínio tem associação com a motilidade celular e sua sinalização é importante para a metástase (Prutzman et al., 2004). Com isso, podemos deduzir que há chance de que a motilidade dessas células possam estar prejudicadas por esta deleção.



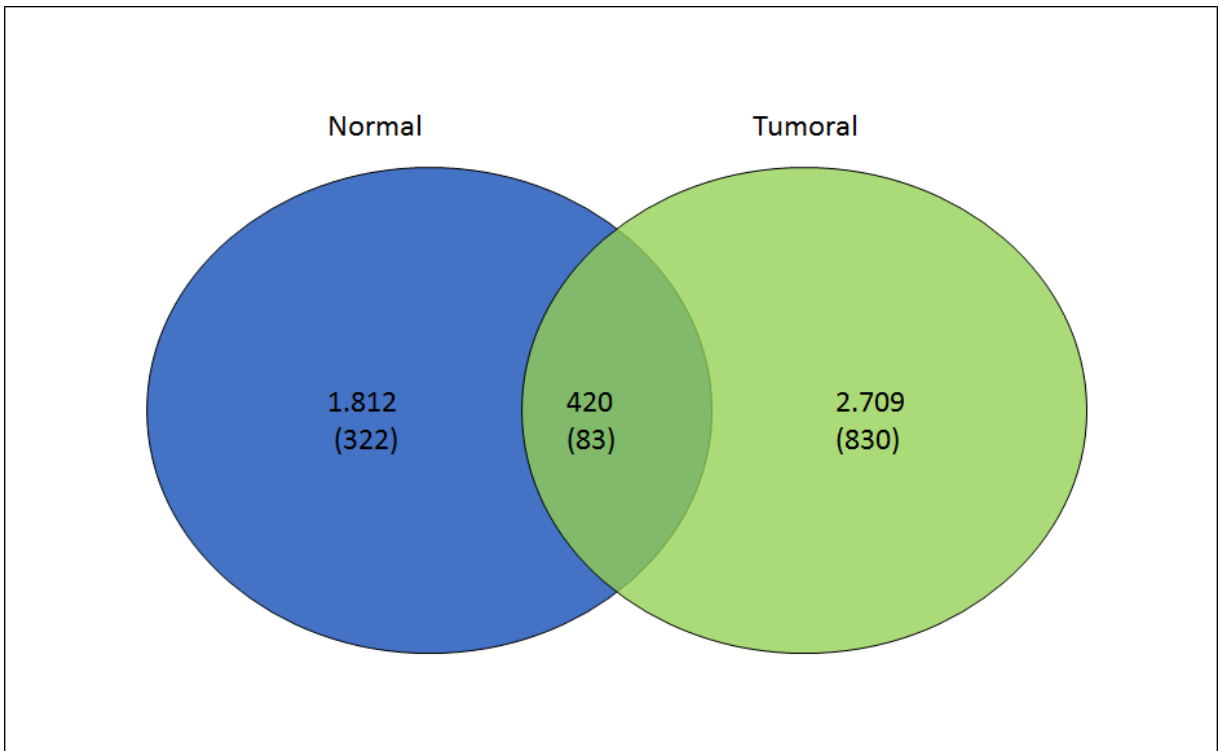
**Figura 4.40** Deleção de sete nucleotídeos em um transcrito (em amarelo) montado a partir de dados de linhagem celular H1975 ao mapear contra um genoma europeu (HG00114) no gene *PTK2* (demarcado em vermelho) causando uma mutação que altera o quadro de leitura da tradução.



**Figura 4.41** Representação do impacto da pequena deleção de sete nucleotídeos no gene *PTK2* na sequência de aminoácidos. A proteína normal **(A)** codificada pelo gene possui quatro domínios: B41 (Domínio ERM ou homólogos banda 4.1, acesso smart00295), Ferm\_C\_FAK (Domínio kinase FERM de adesão focal, acesso cd13190), PTKc\_FAK (Domínio catalítico tirosina kinase de adesão focal, acesso cd05056) e Focal\_AT (Região alvo de adesão focal, pfam03623). A proteína afetada **(B)** pela deleção perde uma porção do domínio final.

#### **4.2.2 Pequenas deleções identificadas em amostras normais e tumorais de seis pacientes não fumantes de câncer de pulmão**

Encontramos 2.332 pequenas deleções em amostras normais e 3.129 em amostras tumorais de pacientes com câncer de pulmão do estudo de Kim e colaboradores (2013a) utilizando como parâmetro a presença da deleção em pelo menos mais de metade dos genomas do 1000G e que estavam ausentes ao utilizar o genoma GRCh37/hg19. Foram encontradas 250 pequenas deleções em amostras normais e 282 em amostras tumorais que também foram identificadas no genoma de referência GRCh37/hg19. Dentre estes achados, apenas 420 estavam simultaneamente presentes em amostras normais e tumorais (Figura 4.42).



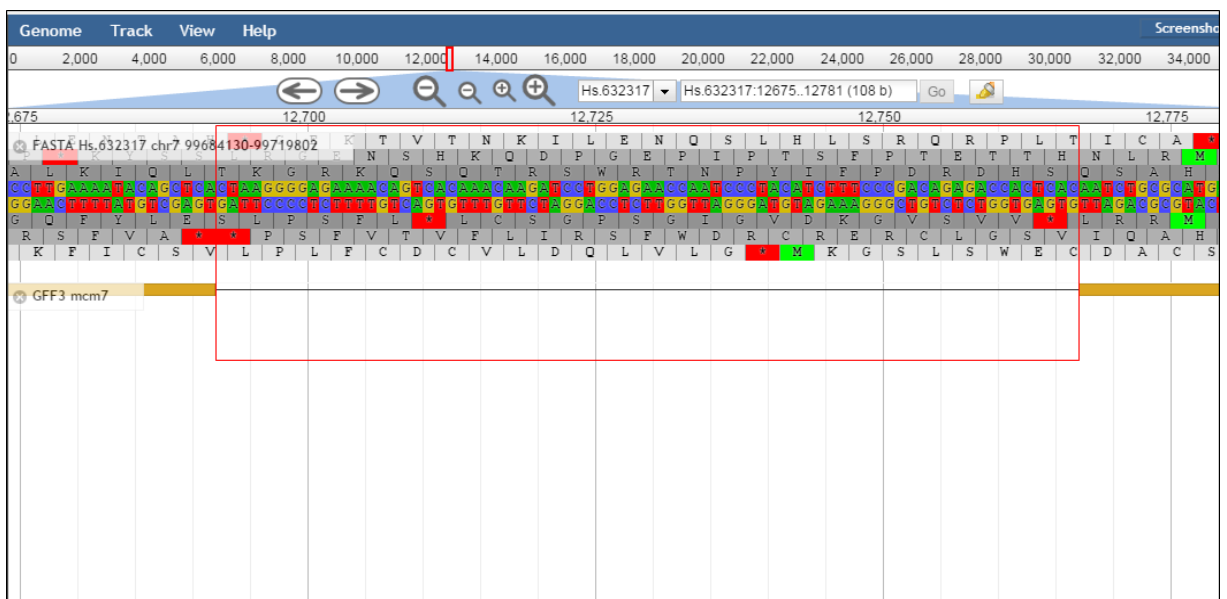
**Figura 4.42** Pequenas deleções identificadas com o uso das matrizes ternárias em tecido normal (azul) e em tecido tumoral (verde) de seis pacientes de câncer de pulmão (Kim et al., 2013a). Entre parênteses a quantidade de deleções identificadas em mais de um paciente.

Encontramos 99 pequenas deleções que ocorriam em regiões não codificadoras, 55 que não causaram mudança no quadro de leitura e 23 que causaram a mudança no quadro de leitura que possuem anotação no dbSNP ou COSMIC (Tabela 4.11).

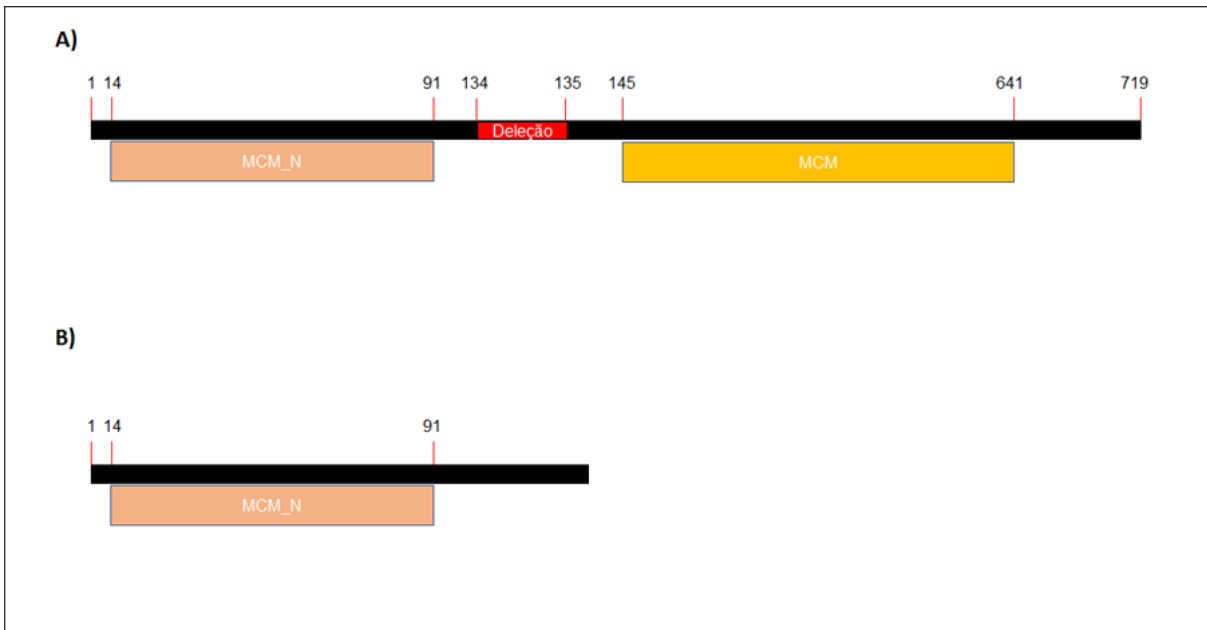
**Tabela 4.11** Tabela contendo as deleções identificadas e seu impacto em regiões codificadoras e presentes no dbSNP e COSMIC em amostras normais e tumorais de pacientes não fumantes de câncer de pulmão.

	Amostras normais			Amostras tumorais		
	Matrizes ternárias	dbSNP	Cosmic	Matrizes ternárias	dbSNP	Cosmic
Regiões não codificadoras	2.181	7	33	2.867	20	39
Alteram o quadro de leitura	21	2	6	50	5	10
Não alteram o quadro de leitura	130	3	15	212	12	25

Encontramos a deleção de 75 nucleotídeos no gene *MCM7* exclusivamente em amostra normal do paciente 5 (Figura 4.43) com o uso do mapeamento de 34 genomas do 1000G (10/12 europeus, 9/9 americanos, 13/27 asiáticos e 2/18 africanos). Esta deleção adianta o códon de parada da tradução fazendo a proteína ficar truncada e perder um domínio inteiro (Figura 4.44). Este gene está associado a regulação do ciclo celular, mais precisamente ativando o “checkpoint” ligado a proteína p53 que leva a parada do ciclo celular (Wei et al., 2013). Apesar de ter sido descrito esta função que poderia agir como supressor tumoral, a alta expressão deste gene já foi associada a baixa taxa de sobrevida em pacientes de NSCLC (Toyokawa et al., 2011). Em amostras tumorais esta deleção está ausente apesar de haver expressão desse gene (Figura 4.60). Acreditamos nesse caso que o alelo contendo a deleção tenha sido silenciado no tecido tumoral. Assim, demonstramos que o uso de RNA-Seq é eficaz tanto para a identificação de deleções como para a avaliação da expressão delas.

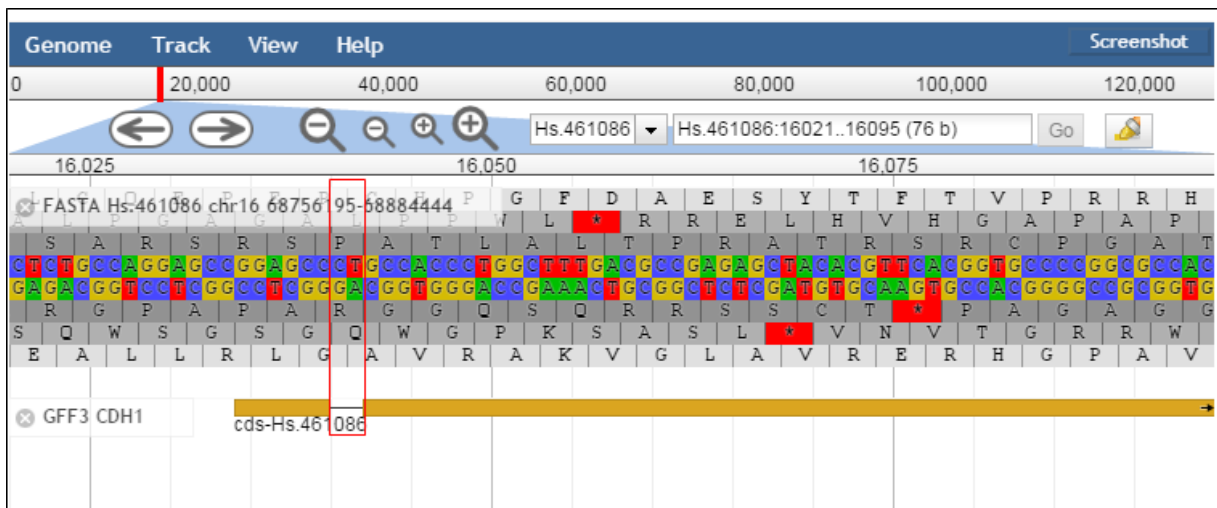


**Figura 4.43** Deleção de 75 nucleotídeos em um transcrito (em amarelo) montado a partir de dados de amostras normais do paciente 5 ao mapear contra um genoma espanhol (HG01501) no gene *MCM7* (demarcado em vermelho) causando uma mutação que adianta o códon de parada da tradução.



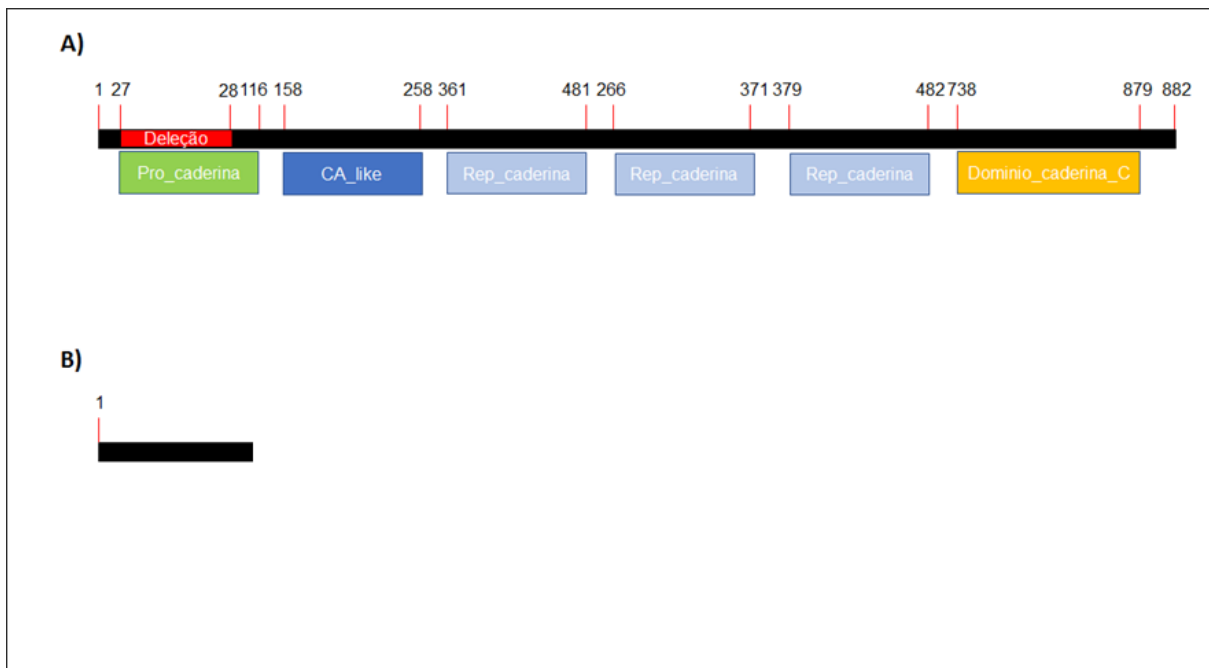
**Figura 4.44** Representação do impacto da pequena deleção de 75 nucleotídeos no gene *MCM7* na sequência de aminoácidos. A proteína normal (A) codificada pelo gene possui dois domínios: MCM\_N (Domínio N-Terminal MCM, acesso pfam14551) e MCM (Proteína de manutenção minicromossomo, acesso smart00350) enquanto a proteína afetada (B) pela deleção foi encurtada.

A deleção de dois nucleotídeos no gene *CDH1* foi identificada nos dados de RNA-Seq apenas na amostra tumoral do paciente 4, mas foi identificada usando metade dos genomas do 1000G utilizados como referência (6/12 europeus, 9/9 americanos, 10/27 asiáticos e 5/18 africanos) (Figura 4.45). Este gene codifica a proteína E-caderina, que é descrita por desempenhar ação de supressora tumoral, é responsável pela adesão célula-célula (Semb and Christofori, 1998)



**Figura 4.45** Deleção de dois nucleotídeos em um transcrito (em amarelo) montado a partir de dados de amostras tumorais do paciente 4 ao mapear contra um genoma espanhol (HG01602) no gene *CDH1* (demarcado em vermelho) causando uma mutação que altera o quadro de leitura da tradução.

Esta deleção afetou o quadro de leitura da proteína codificada pelo gene *CDH1* e todos domínios conhecidos da proteína foram alterados pela produção de uma proteína truncada (Figura 4.46). A deleção de dois nucleotídeos ocorre no início da região codificadora do gene e por isso adianta o códon de parada produzindo uma possível proteína de 56 aminoácidos ao invés de uma normal com os 882 aminoácidos (Figura 4.46). Por conta disso, as células com essa mutação podem ter sua capacidade de adesão ao tecido do pulmão reduzida podendo levar a uma possível metástase. Além disso, após este códon de parada são 61 nucleotídeos até o final do éxon, portanto é provável que este mRNA sofra o mecanismo de NMD. Em adenocarcinoma de pulmão já foi relatada a baixa expressão do gene *CDH1* com o mau prognóstico de pacientes em que o tumor se torna invasivo (Kase et al., 2000). Encontramos principalmente trabalhos que identificam SNVs no gene *CDH1* em adenocarcinoma de pulmão. Um deles encontra cinco diferentes localizações de SNVs afetando a região codificadora deste gene (Imielinski et al., 2012). No entanto, deleções nesse gene já foram identificados em outros adenocarcinomas como de mama (Stephens et al., 2012) e cólon (Seshagiri et al., 2012).

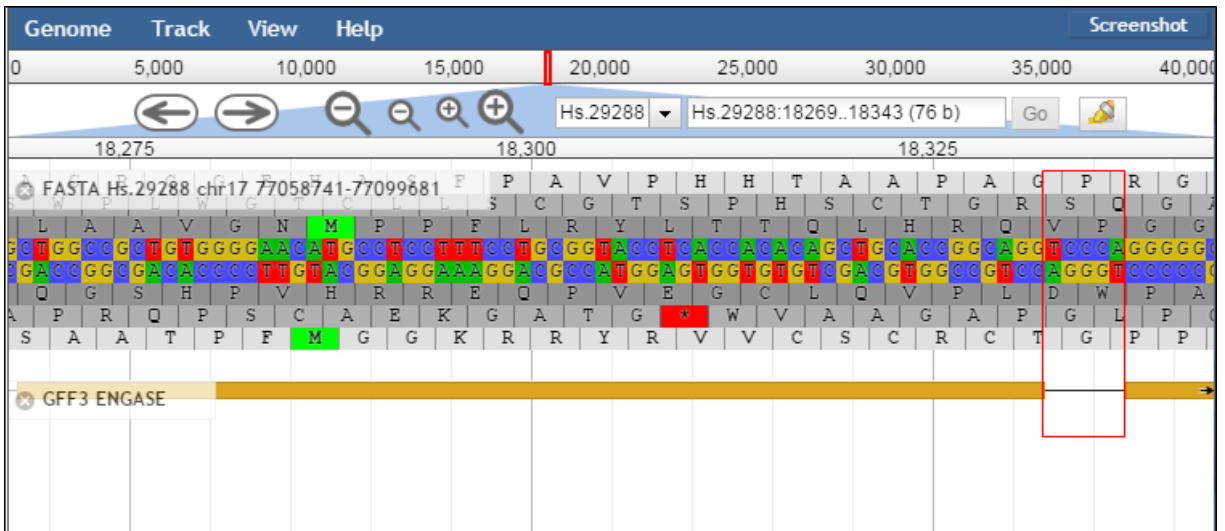


**Figura 4.46** Representação do impacto da pequena deleção de dois nucleotídeos no gene *CDH1* na sequência de aminoácidos. A proteína normal **(A)** codificada pelo gene possui seis domínios: Pro\_caderina (Prodomínio caderina-like, acesso pfam08758), CA\_like (Domínio de repetição caderina-like, acesso cd00031), três domínios Rep\_caderina (Domínio de repetição de caderina em *tandem*, acesso cd11304) e Domínio\_caderina\_C (Região citoplasmática caderina, acesso pfam01049) enquanto a proteína afetada **(B)** pela deleção um encurtamento com esses domínios perdidos.

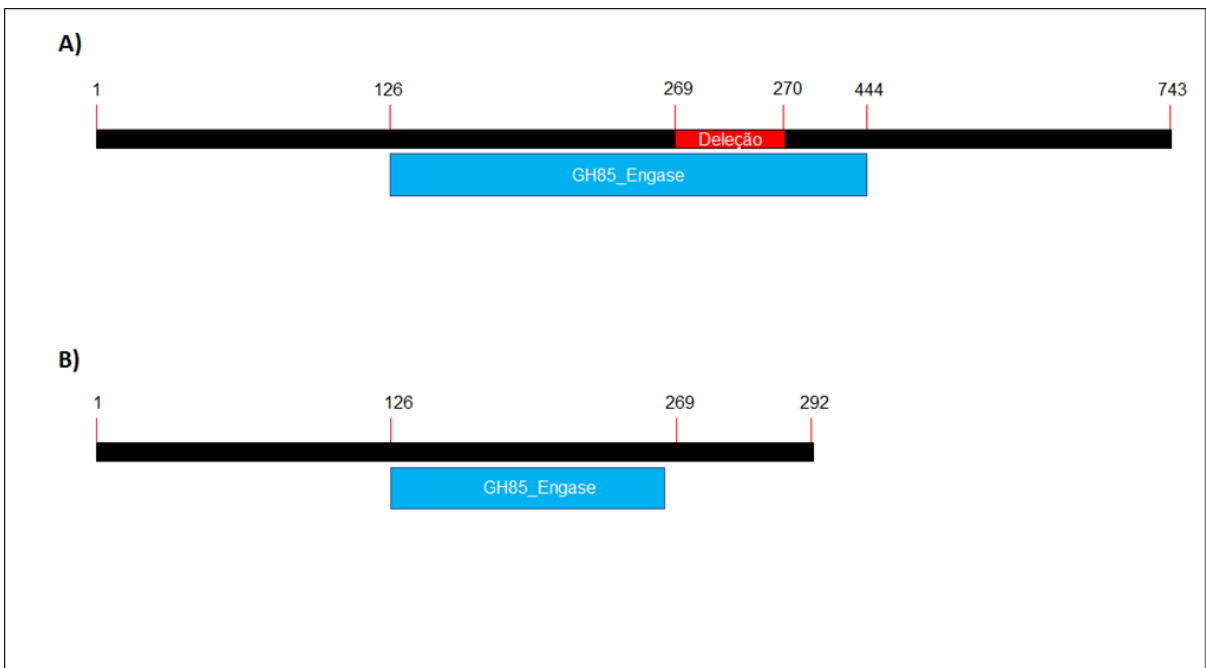
Outro exemplo que podemos citar é a deleção de cinco nucleotídeos no gene *ENGASE* encontrada a partir de dados de RNA-Seq em amostra tumoral do paciente 3. Encontramos esta deleção com o uso de 35 genomas do 1000G como referência (12/12 europeus, 9/9 americanos, 12/27 asiáticos e 2/18 africanos) (Figura 4.47). Este gene codifica uma enzima que participa do metabolismo de proteínas (Murakami et al., 2013).

Esta deleção alterou o quadro de leitura da tradução da proteína e sua sequência de aminoácidos ficou mais curta com os sítios ativos alterados (Figura 4.48). Como este domínio é para catalisar reações de hidrólise e de transglicosilação (Choragudi et al., 2014), esta mutação pode ter reduzido a afinidade dessa proteína em suas moléculas alvo. Foram encontradas deleções que também alteram o quadro de leitura da tradução deste gene em melanoma (Van Allen et al., 2014) e adenocarcinoma de cólon (Mouradov et al., 2014).





**Figura 4.47** Deleção de cinco nucleotídeos em um transcrito (em amarelo) montado a partir de dados de amostra tumoral do paciente 3 ao mapear contra um genoma de ancestralidade europeia (HG01501) no gene *ENGASE* (demarcado em vermelho) causando uma mutação que altera o quadro de leitura da tradução.

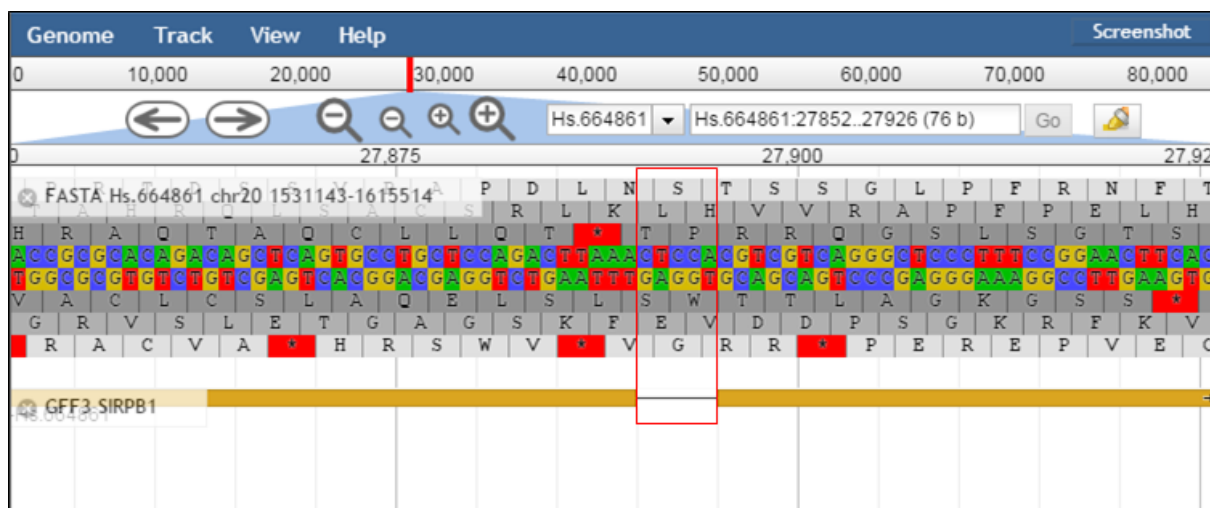


**Figura 4.48** Representação do impacto da pequena deleção de cinco nucleotídeos no gene *ENGASE* na sequência de aminoácidos. A proteína normal **(A)** codificada pelo gene possui um único domínio com sítios ativos GH85\_Engase (Domínio Engase, acesso cd06547) enquanto a proteína afetada **(B)** pela deleção possui um encurtamento com esse domínio afetado, incluindo os sítios ativos

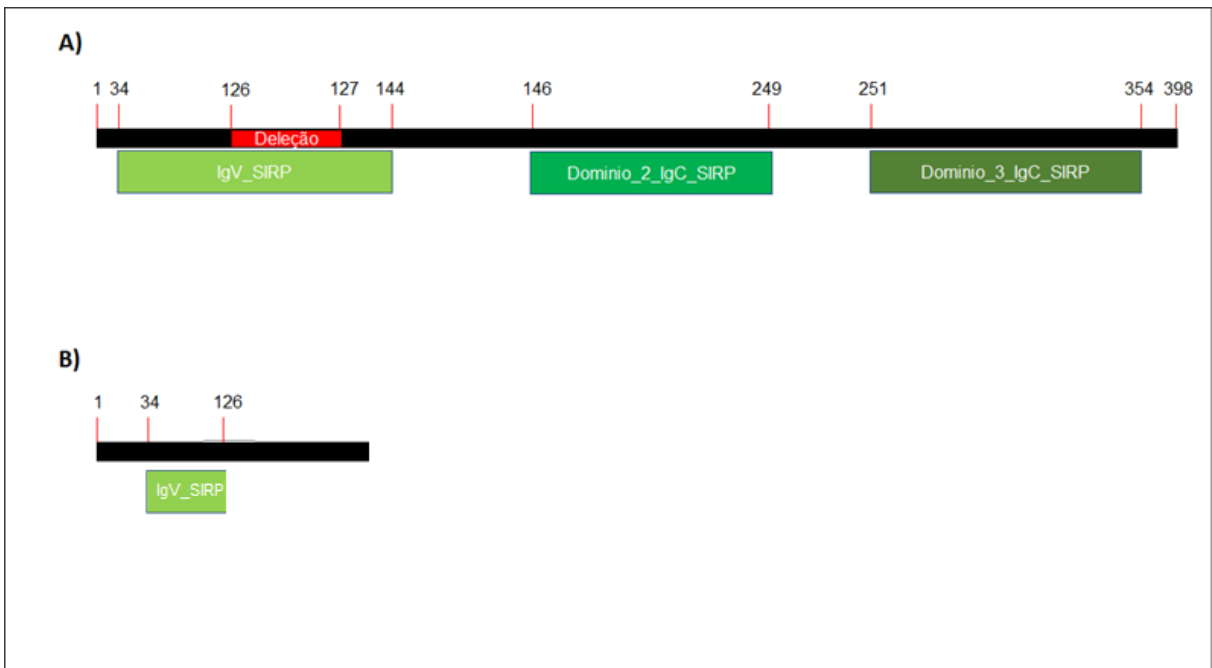
Analizamos os dados de RNA-Seq em busca de potenciais candidatos a mutações germinativas. Neste sentido, pudemos identificar presente em amostras normais e tumorais uma deleção de 5 nucleotídeos no gene *SIRPB1*. Esta deleção foi encontrada em amostras normais e tumorais de dois pacientes (pacientes 3 e 4)

(Figura 4.49). Este gene traduz uma glicoproteína transmembrana da família das imunoglobulinas que por sua vez, regula negativamente a sinalização de tirosina kinases (Kharitonenkov et al., 1997).

Esta pequena deleção foi encontrada a partir de 34 genomas do 1000G como referência (1/12 europeus, 9/9 americanos, 8/27 asiáticos e 16/18 africanos). A deleção alterou o quadro de leitura da tradução reduzindo o tamanho da proteína codificada de 398 para 131 aminoácidos (Figura 4.50). A proteína canônica possui três domínios de imunoglobulinas (Figura 4.50A), no entanto a deleção pode ter afetado todo o seu funcionamento pois seu tamanho reduziu para 131 aminoácidos e todos esses domínios foram afetados (Figura 4.50B). Alterações para controlar a expressão desse gene já foram observadas em adenocarcinoma de pulmão, como a metilação deste gene (Mullapudi et al., 2015) e a observação de uma deleção que causava a mudança do quadro de leitura na tradução da proteína (Imielinski et al., 2012). Segundo nossas análises, esta mutação poderia ser considerada potencial candidata a mutação germinativa e exclusiva de pacientes não fumantes.



**Figura 4.49** Deleção de cinco nucleotídeos em um transcrito (em amarelo) montado a partir de dados de amostras normais e tumorais dos pacientes 3 e 4 ao mapear contra um genoma de ancestralidade africana (HG02282) no gene *SIRPB1* (demarcado em vermelho) causando uma mutação que altera o quadro de leitura da tradução.



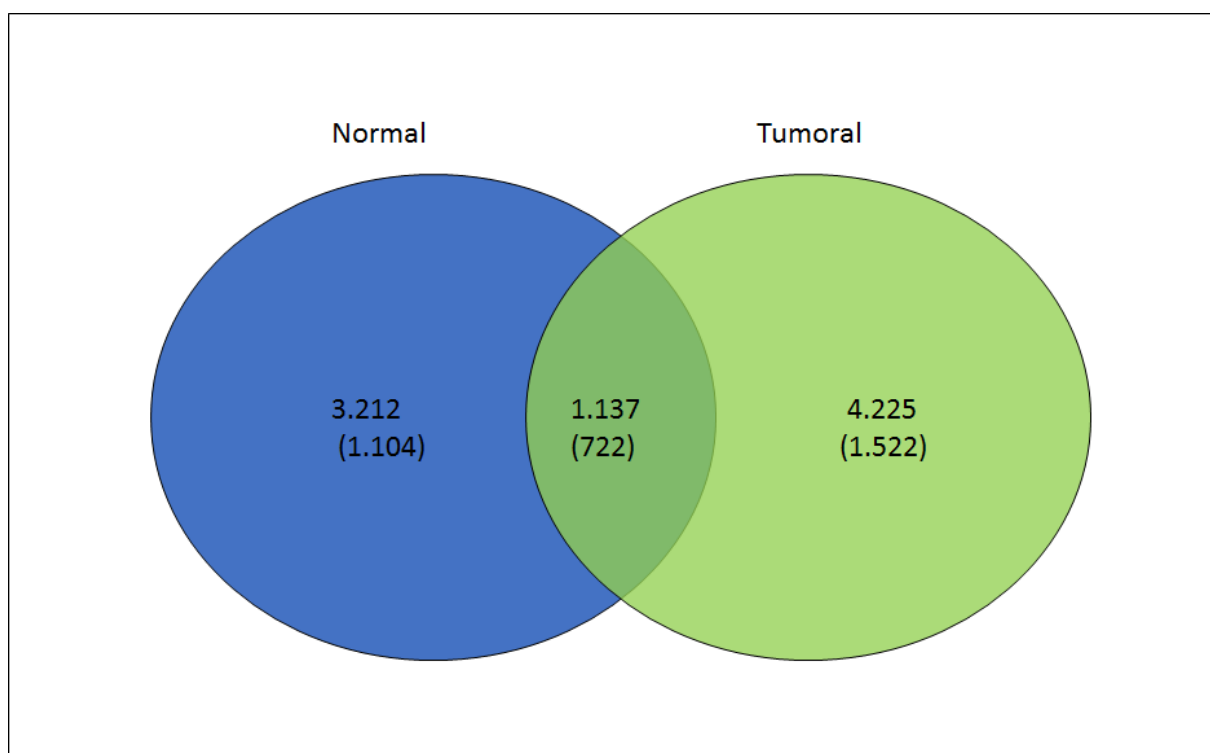
**Figura 4.50** Representação do impacto da pequena deleção de cinco nucleotídeos no gene SIRPB1 na sequência de aminoácidos. A proteína normal **(A)** codificada pelo gene possui três domínios de imunoglobulina: IgV\_SIRPT (Domínio Imunoglobulina-like SIRP, acesso cd16097), Domínio\_2\_IgC\_SIRP (Domínio imunoglobulina-like SIRP 2, acesso cd05772) e Domínio\_3\_IgC\_SIRP (Domínio imunoglobulina-like SIRP 3, acesso cd16085). A proteína afetada **(B)** pela deleção tem um encurtamento.

Realizamos uma análise de enriquecimentos de vias e encontramos apenas uma via com possibilidade de estar alterada pelas deleções identificadas nas amostras tumorais. A via encontrada com super-representação foi a de regulação do formato celular (GO:0008360) com probabilidade superior de 0,58. Identificamos pequenas deleções em 19 genes que participam dessa via.

Não encontramos nenhuma deleção que tivesse ocorrido ao mesmo tempo na linhagem celular H1975 que também é proveniente de tumores de pacientes não fumantes. Isto pode ter acontecido pelo fato das pacientes analisadas por Kim e colaboradores (2013a) serem de origem asiática e a linhagem ser de origem de um indivíduo caucasiano. Esperamos começar a elucidar a taxa de falsos positivos da nossa abordagem assim que tivermos os resultados da validação experimental dos resultados para a linhagem de células H1975 que está em andamento.

### 4.2.3 Pequenas deleções identificadas em amostras normais e tumorais de 14 pacientes fumantes de câncer de pulmão

Encontramos 4.349 pequenas deleções em amostras normais adjacentes ao tumor e 5.362 em amostras tumorais de pacientes com câncer de pulmão do estudo de Collisson e colaboradores (2014) utilizando como parâmetro a presença da deleção em pelo menos mais de metade dos genomas do 1000G. Utilizamos como critério de exclusão aquelas deleções presentes no genoma GRCh37/hg19. Dentre estes achados, 1.137 estavam presentes tanto na amostra normal como na tumoral (Figura 4.51). Encontramos 121 pequenas deleções que ocorriam em regiões não codificadoras, 40 que não causaram mudança no quadro de leitura e 22 que causaram a mudança no quadro de leitura que possuem anotação no dbSNP ou COSMIC (Tabela 4.12).

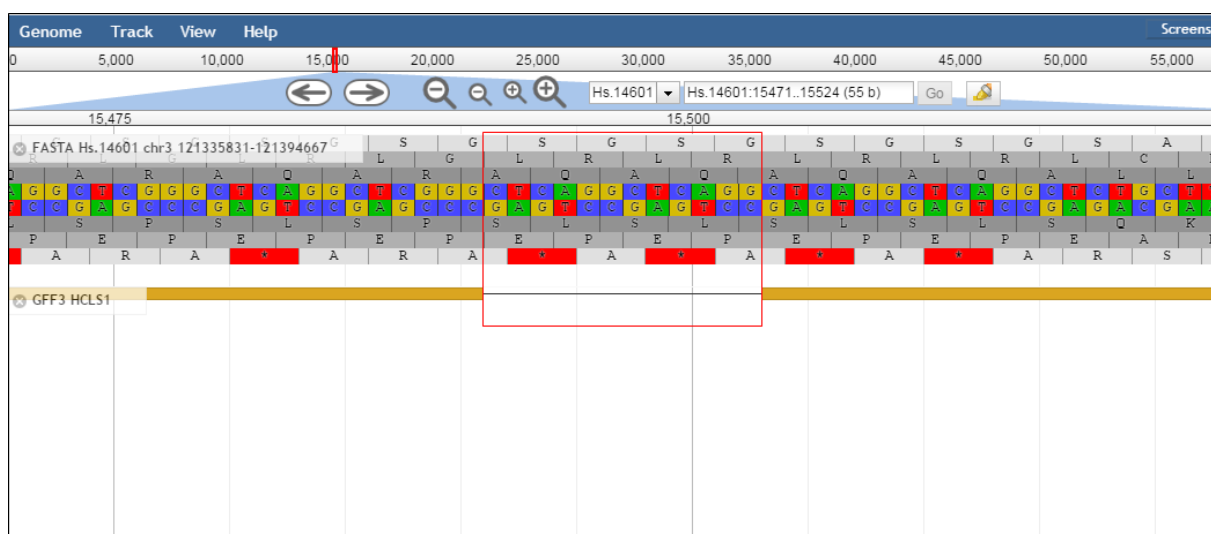


**Figura 4.51** Pequenas deleções identificadas ao utilizar as matrizes ternárias em tecido normal (azul) e em tecido tumoral (verde) de seis pacientes de câncer de pulmão (Collisson et al., 2014). Entre parênteses a quantidade de deleções identificadas em mais de um paciente.

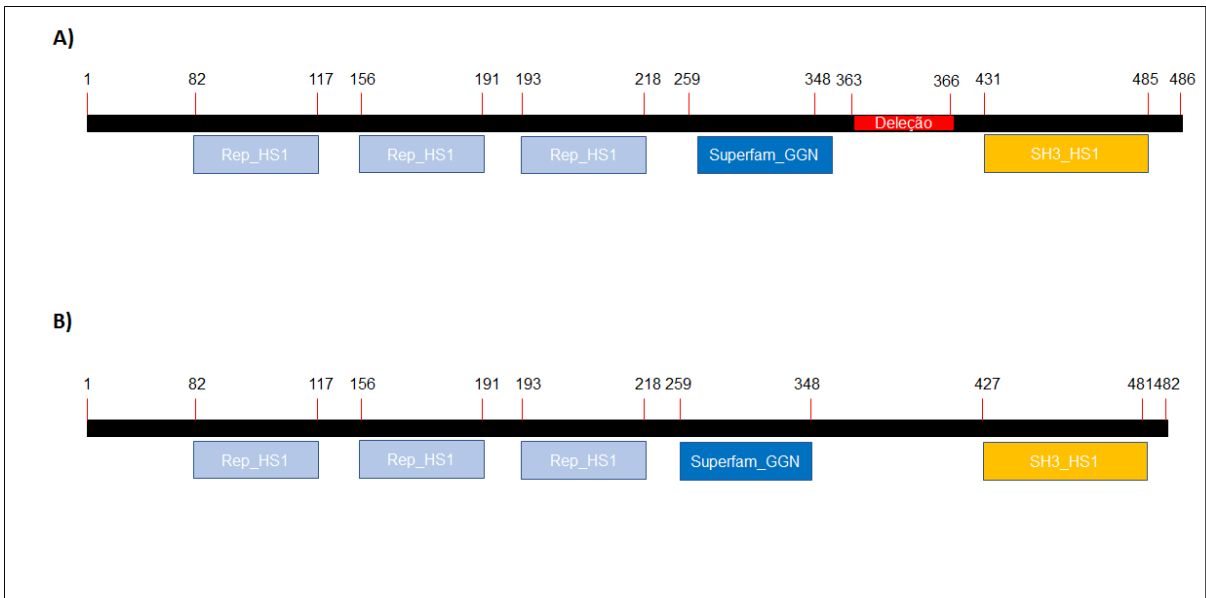
**Tabela 4.12** Tabela representando as deleções identificadas e seu impacto em regiões codificadoras e presentes no dbSNP e COSMIC em amostras normais e tumorais de pacientes fumantes de câncer de pulmão.

	Amostras normais			Amostras tumorais		
	Matrizes ternárias	dbSNP	Cosmic	Matrizes ternárias	dbSNP	Cosmic
Regiões não codificadoras	6.123	11	48	6.665	13	49
Alteram o quadro de leitura	264	3	6	271	6	7
Não alteram o quadro de leitura	271	5	10	386	11	14

Dentre as deleções que ocorreram apenas em amostras normais, identificamos uma pequena deleção de 12 nucleotídeos que afeta o gene *HCLS1* em 3 pacientes (2657, 3396 e 3398). Esta deleção foi encontrada a partir do uso de 33 genomas do 1000G como referência (Figura 4.52). Esta deleção basicamente fez a proteína perder quatro aminoácidos afetando o domínio de gametogenetina (Figura 4.53).



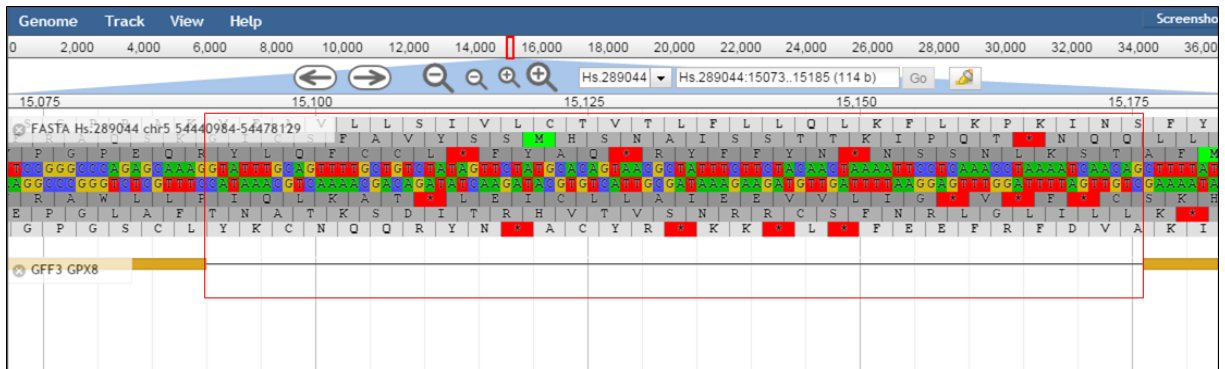
**Figura 4.52** Deleção de 12 nucleotídeos em um transcrito (em amarelo) montado a partir de dados de amostra normal do paciente 3398 ao mapear contra um genoma de ancestralidade asiática (HG01869) no gene *HCLS1* (demarcado em vermelho).



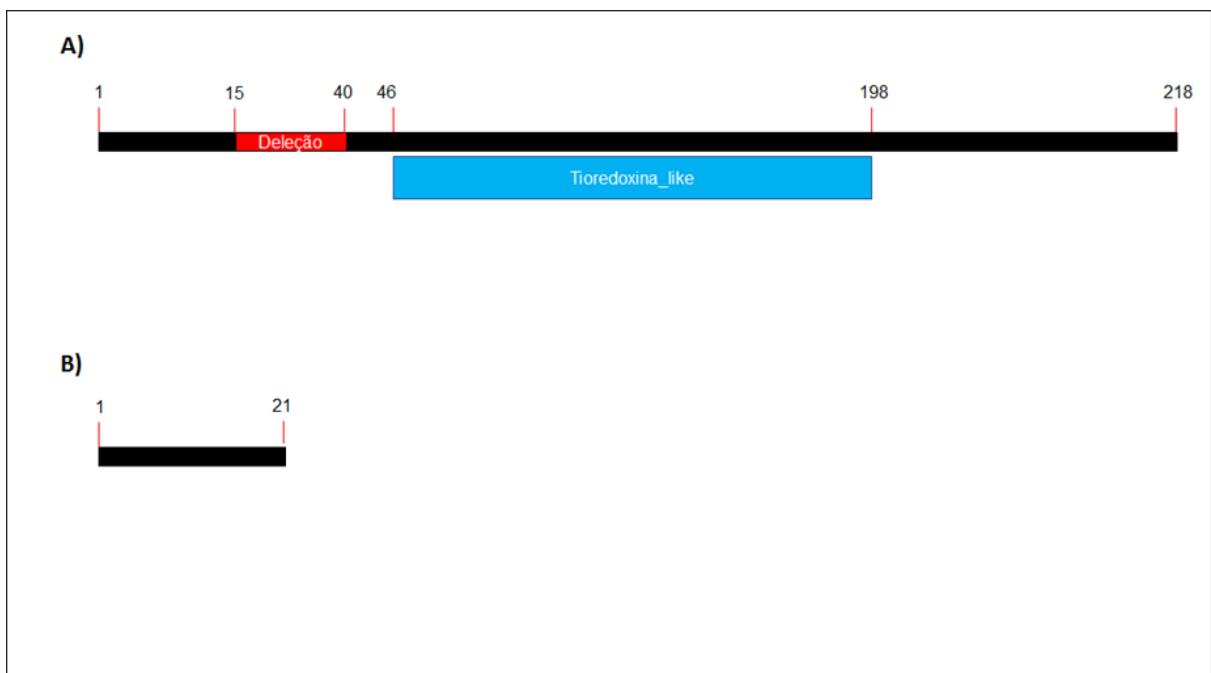
**Figura 4.53** Representação do impacto da pequena deleção de 12 nucleotídeos no gene *HCLS1* na sequência de aminoácidos. A proteína normal (**A**) possui 486 aminoácidos e cinco domínios: três domínios Rep\_HS1 (Repetição HS1, acesso pfam02218), Superfam\_GGN (Gametogenetina, acesso cl25800) e SH3\_HS1 (Domínio SH3, acesso cd12073) e a proteína afetada (**B**) pela deleção possui quatro nucleotídeos a menos.

Utilizando 60 genomas do 1000G como referência, encontramos uma pequena deleção de 75 nucleotídeos no gene *GPX8* na amostra tumoral de RNA-Seq do paciente 3398 (12/12 europeus, 9/9 americanos, 21/27 asiáticos e 18/18 africanos) (Figura 4.54). Este gene é responsável por produzir uma proteína que é expressa em resposta ao “stress” oxidativo em células epiteliais (Cortes et al., 2011).

Esta deleção adiantou o códon de parada da tradução da proteína em que a proteína normal possui foi reduzida de 208 para 21 aminoácidos. Este adiantamento do códon de parada afetou seu principal domínio em que possui resíduos catalíticos (Figura 4.55). Além disso, o códon de parada ocorre 66 nucleotídeos antes do final do éxon, portanto este mRNA pode ser degradado pelo mecanismo de NMD. Não encontramos eventos de INDELS relacionados a este gene em câncer. No entanto, uma substituição de um nucleotídeo foi observada adiantando o códon de parada desse gene em câncer de pulmão (Imielinski et al., 2012).



**Figura 4.54** Deleção de 75 nucleotídeos em um transcrito (em amarelo) montado a partir de dados de amostra tumoral do paciente 3398 ao mapear contra um genoma de ancestralidade africana (HG01879) no gene *GPX8* (demarcado em vermelho) causando uma mutação que altera o quadro de leitura da tradução.

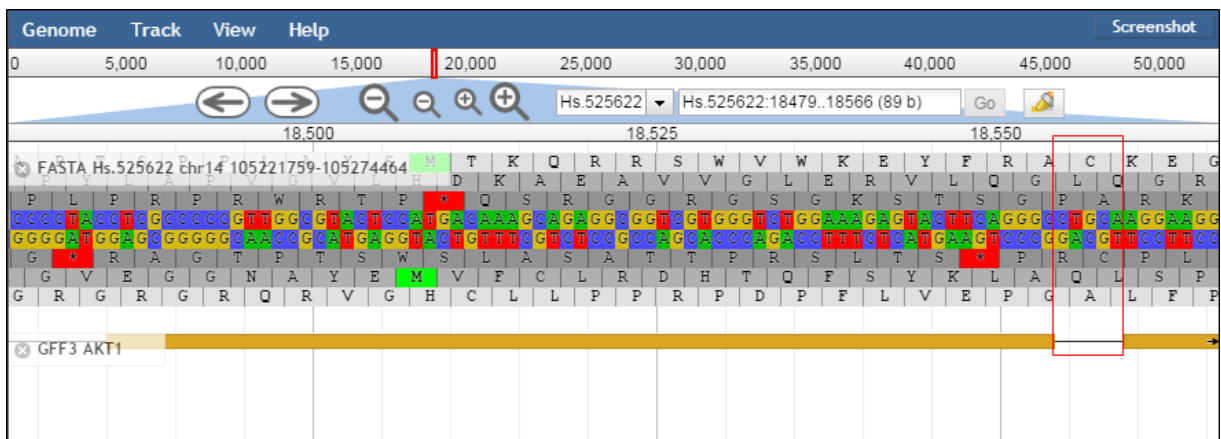


**Figura 4.55** Representação do impacto da pequena deleção de 75 nucleotídeos no gene *GPX8* na sequência de aminoácidos. A proteína normal **(A)** possui 218 aminoácidos e o domínio Tioredoxina\_like (Peroxidase GPX7 da superfamília tioredoxina-like, acesso TIGR02540) e a proteína afetada **(B)** pela deleção causou a perda de seu principal domínio proteico.

Outra deleção encontrada apenas em tecidos normais foi a deleção de cinco nucleotídeos no gene *AKT1* em três pacientes (6776, 6777 e 6778) (Figura 4.56). Esta deleção esteve presente em ao utilizar 33 genomas como referência (6/12 europeus, 2/9 americanos, 25/27 asiáticos e 0/18 africanos). As mutações do gene *AKT1* são conhecidas em câncer de pulmão, porém na maioria das vezes, essas mutações não afetam a proteína codificada. Isto se deve a sua importância para o

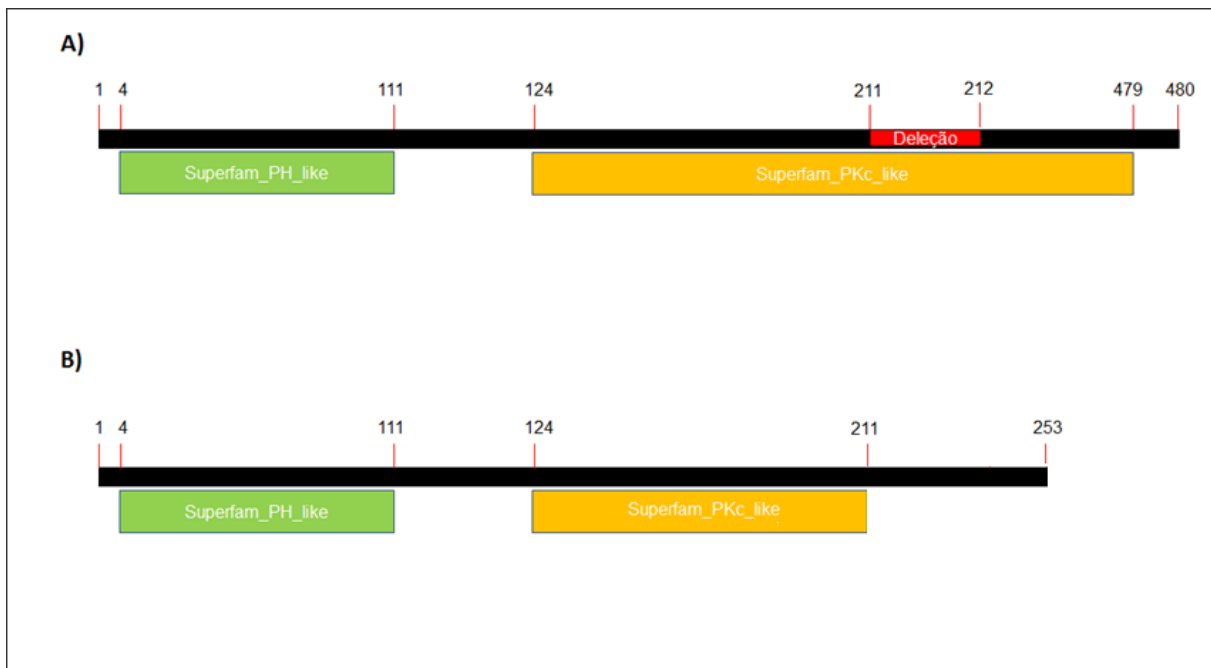
crescimento e proliferação do tumor mediado pela proteína mutada K-ras (Lee et al., 2011).

A deleção afetou o domínio proteína kinase adiantando o códon de parada da tradução desta (Figura 4.57). Um outro estudo relacionou que alguns pacientes de câncer de pulmão possuíam deleções em *AKT1* e por isso a tumorigenese não ocorria utilizando a via de K-ras com mutação (Hollander et al., 2011).



**Figura 4.56** Deleção de cinco nucleotídeos em um transcrito (em amarelo) montado a partir de dados de amostra tumoral dos pacientes 6776, 6777 e 6778 ao mapear contra um genoma de ancestralidade asiática (HG00956) no gene *AKT1* (demarcado em vermelho) causando uma mutação que altera o quadro de leitura da tradução.

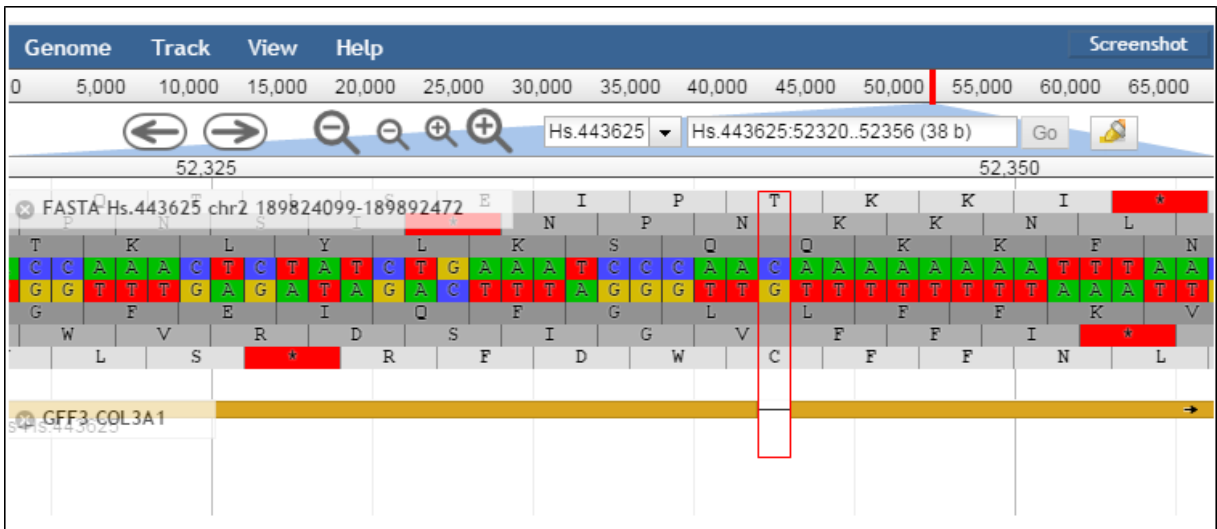




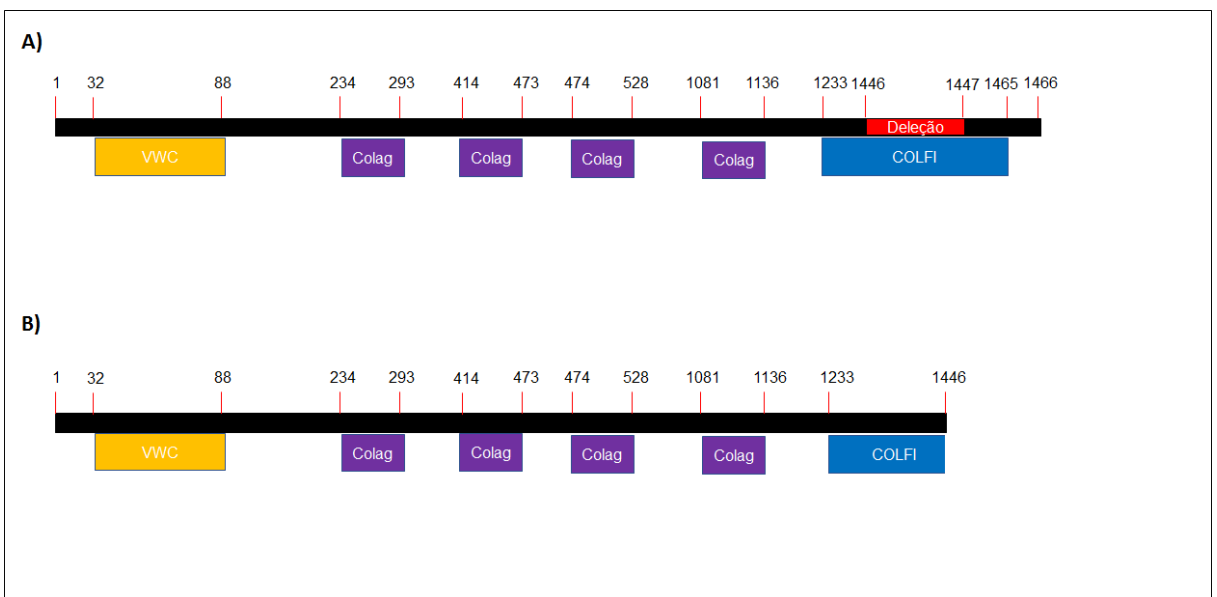
**Figura 4.57** Representação do impacto da pequena deleção de cinco nucleotídeos no gene *AKT1* na sequência de aminoácidos. A proteína normal **(A)** possui dois grandes domínios: Superfam\_PH\_like (Domínio Akt, acesso cd01241) e Superfam\_PKc\_like (Domínio catalítico Serina/Treonina kinase, acesso cd05594). A proteína afetada **(B)** pela deleção possui um encurtamento que afeta o domínio proteína kinase.

A deleção de um nucleotídeo no gene *COL3A1* foi encontrada a partir dos dados de RNA-Seq de amostras normais e tumorais de três pacientes diferentes (5645, 6778 e 6148) (Figura 4.58). Este gene codifica uma proteína colágeno importante para a matriz extracelular de vários órgãos, inclusive pulmão (Kuivaniemi et al., 1997).

Esta deleção foi identificada a partir de 39 genomas como referência (12/12 europeus, 9/9 americanos, 18/27 asiáticos e 0/18 africanos). A deleção de um único nucleotídeo ocorreu na região que corresponde o domínio C-terminal da proteína que leva a perder 19 aminoácidos deste domínio (Figura 4.59). Este domínio tem como função a ligação de vários substratos a esta proteína (Stembridge et al., 2015). Duas deleções neste gene já foram relatadas em adiantar o códon de parada da tradução da proteína codificada afetando o domínio C-terminal em pacientes com câncer de pulmão (Imielinski et al., 2012). Segundo nossas análises, esta mutação pode ser considerada como potencial candidata a mutação germinativa e exclusiva de pacientes fumantes de forma antagônica ao que encontramos a deleção do gene *SIRPB1* como potencial candidata a mutação germinativa e exclusiva de pacientes não fumantes.



**Figura 4.58** Deleção de um nucleotídeo em um transcrito (em amarelo) montado a partir de dados de amostras normais e tumorais dos pacientes 5645, 6778 e 6148 ao mapear contra um genoma de ancestralidade europeia (HG00361) no gene *COL3A1* (demarcado em vermelho) causando uma mutação que altera o quadro de leitura da tradução.



**Figura 4.59** Representação do impacto da pequena deleção de um nucleotídeo no gene *COL3A1* na sequência de aminoácidos. A proteína normal **(A)** possui 1466 aminoácidos e seis domínios: VWC (Domínio fator von Willebrand tipo C, acesso pfam00093), quatro domínios Colag (Repetição tripla hélice de colágeno, acesso pfam01391) e COLFI (Domínio C-terminal de colágeno fibrilar, acesso pfam01410). A proteína afetada **(B)** pela deleção possui um encurtamento de 19 aminoácidos na sua porção C-terminal.

O enriquecimento de vias com os genes em que foram encontradas deleções usando as matrizes ternárias em amostras tumorais utilizando os 66 genomas do 1000G resultou no aparecimento de vias com elevada probabilidade de serem

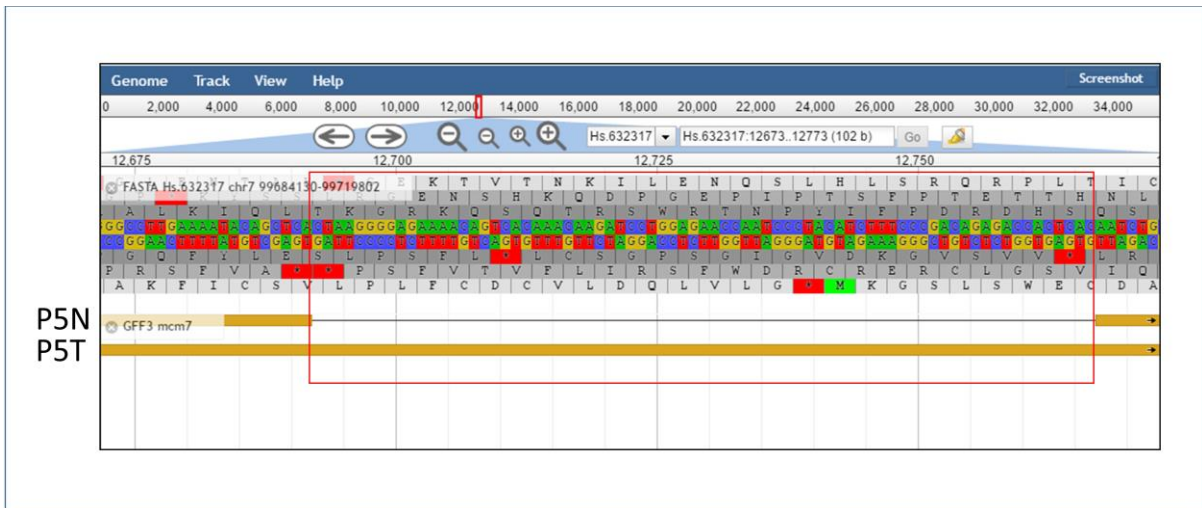
alteradas como: adesão célula-célula (probabilidade posterior de 1), adesão homofílica via moléculas de membrana plasmática (probabilidade posterior de 1) e regulação positiva da macroautofagia (probabilidade posterior de 0,88) (Tabela 4.13).

**Tabela 4.13** Enriquecimento de genes que sofreram pequenas deleções em amostras tumorais identificadas pela nossa metodologia de matrizes ternárias. Abaixo as vias do Gene Ontology (GO) com probabilidade com probabilidade posterior maior do que 0,5 de estarem superrepresentadas.

GO id	Descrição	Probabilidade posterior	Quantidade de Genes
GO:0098609	Adesão célula-célula	1,00	65
GO:0007156	Adesão homofílica via moléculas da membrana plasmática	1,00	33
GO:0016239	Regulação positiva da macroautofagia	0,88	11
GO:0050727	Regulação da resposta inflamatória	0,67	14
GO:0030033	Montagem de microvilosidade	0,66	9
GO:0071230	Resposta celular ao estímulo de aminoácido	0,64	16
GO:0048008	Via de sinalização de <i>PDGFR</i>	0,54	12

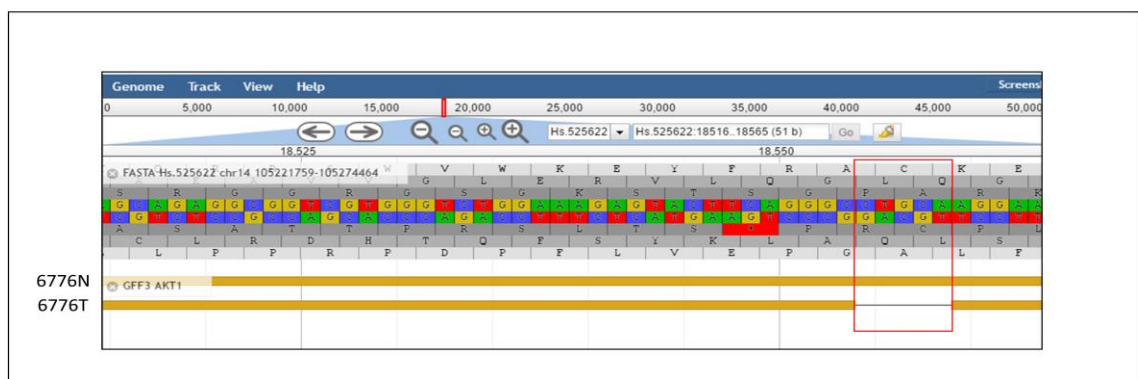
#### **4.2.4 Análise de deleções encontradas exclusivamente em amostras normais, tumorais ou presente em ambas**

Encontramos 1.812 deleções exclusivas de amostras normais no conjunto de dados de Kim e colaboradores (2013a) (Figura 4.42) e 3.212 no conjunto de amostras normais de Collisson e colaboradores (2014) (Figura 4.51). Dentre elas, encontramos uma deleção de 75 nucleotídeos no gene *MCM7* nos pacientes não fumantes (Figura 4.43) e outra de 12 nucleotídeos no gene *HCLS1* nos pacientes fumantes (Figura 4.52). Essas deleções não foram encontradas nas amostras tumorais desses pacientes, apesar de encontrarmos expressão destes genes no tecido tumoral (Figura 4.60).

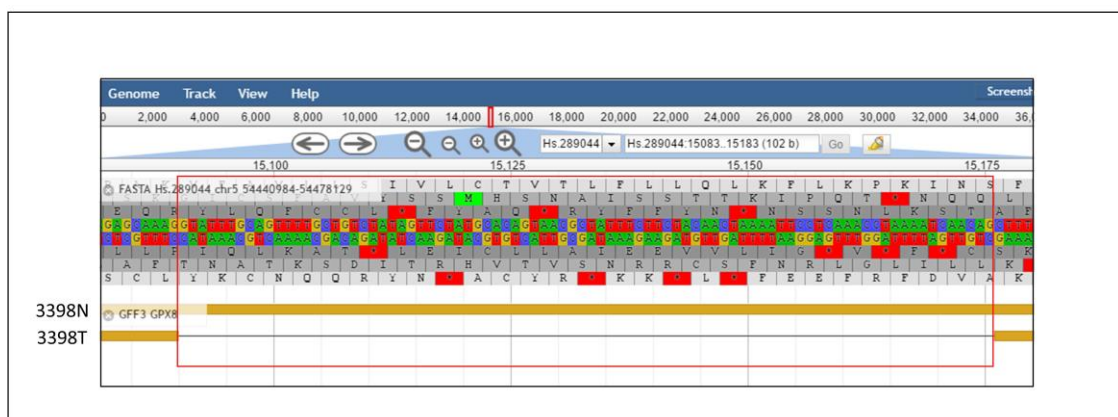


**Figura 4.60** Exemplo de transcritos reconstruídos pelo programa Trinity para o gene *MCM7* para a amostra do tecido normal do paciente 5 (P5N) e para o tecido tumoral deste mesmo pacientes (P5T).

Dentre as 2.709 deleções encontradas exclusivamente em tecido tumoral nos dados de Kim e colaboradores (2013a) (Figura 4.42) podemos citar como exemplos uma pequena deleção no gene *CDH1* (Figura 4.45) e outra no gene *ENGASE* (Figura 4.47). Nos dados de Collisson e colaboradores (2014), encontramos 4.225 deleções, sendo uma delas uma deleção no gene *GPX8* (Figura 4.54) e outra no gene *AKT* (Figura 4.56). Essas deleções não foram encontradas nas amostras normais desses pacientes, apesar de encontrarmos expressão destes genes no tecido normal (Figura 4.61 e Figura 4.62).



**Figura 4.61** Exemplo de transcritos reconstruídos pelo programa Trinity para o gene *AKT1* para a amostra do tecido normal do paciente 6776 (6776N) e para o tecido tumoral deste mesmo pacientes (6776T).



**Figura 4.62** Exemplo de transcritos reconstruídos pelo programa Trinity para o gene *GPX8* para a amostra do tecido normal do paciente 3398 (3398N) e para o tecido tumoral deste mesmo pacientes (3398T).

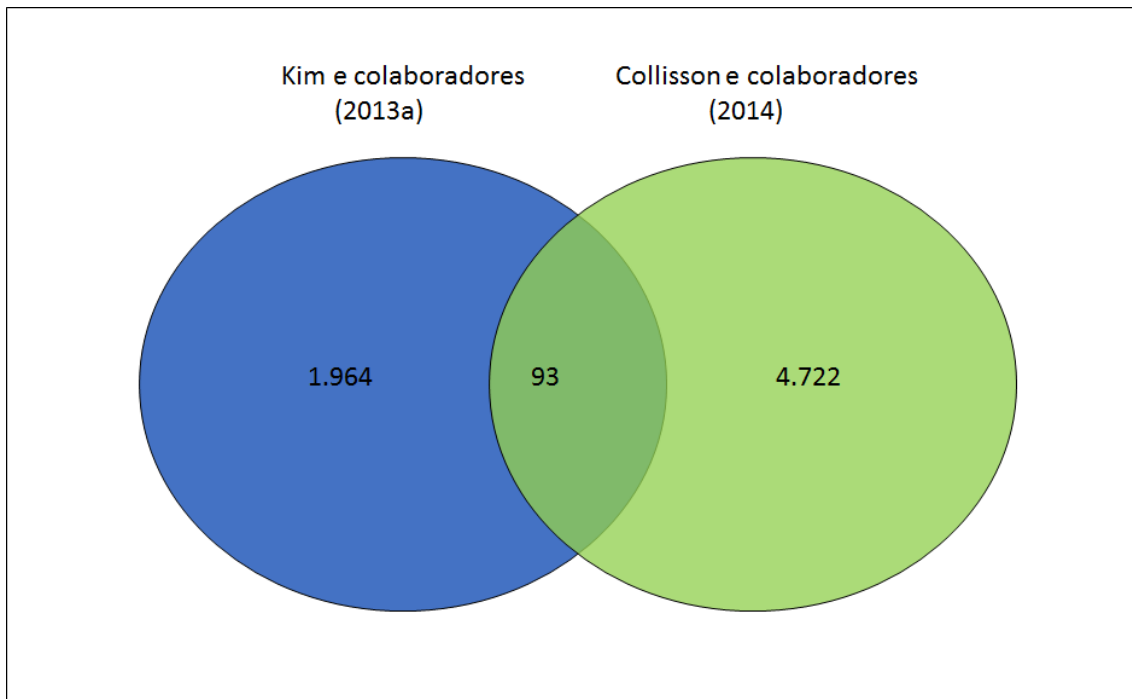
Também encontramos deleções ocorrendo ao mesmo tempo em amostras normais e tumorais. Nos dados de Kim e colaboradores (2013a), encontramos 420 deleções, sendo que uma delas ocorrendo no gene *SIRPB1* (Figura 4.49). Nos dados de Collisson e colaboradores (2014), encontramos 1.137 deleções e dentre elas uma que ocorre no gene *COL3A1* (Figura 4.58).

### 4.3 Comparação dos dados de pacientes fumantes e pacientes não fumantes

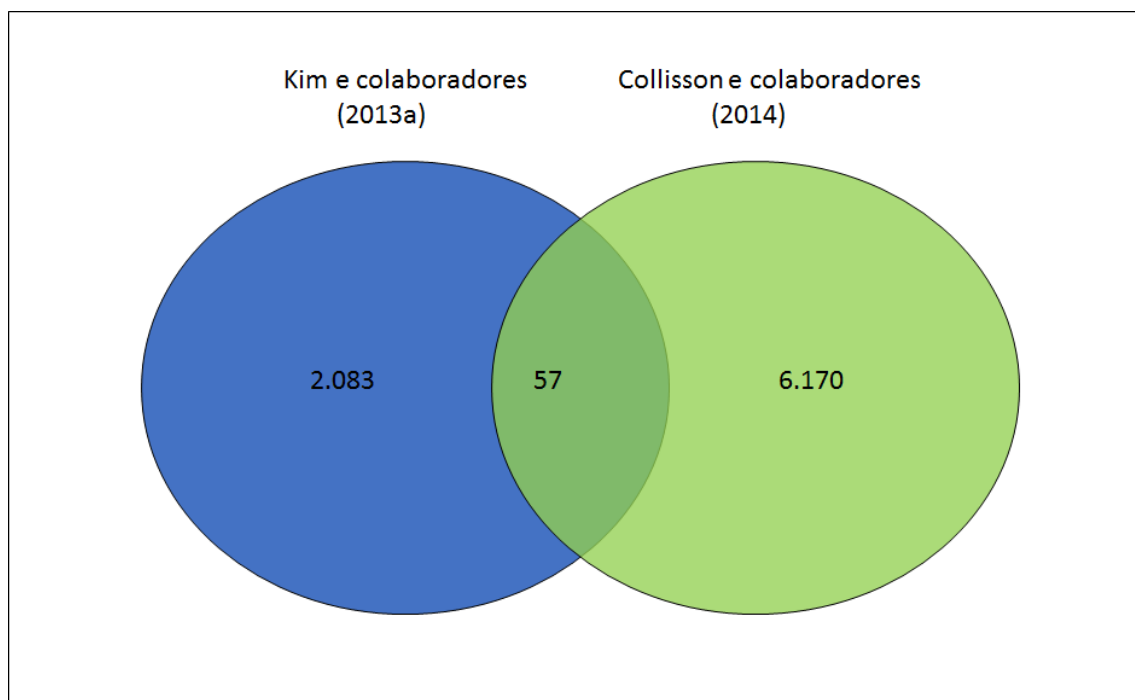
#### 4.3.1 Pequenas deleções identificadas a partir do genoma de referência GRCh37/hg19

Encontramos 93 deleções presentes tanto em amostras normais de pacientes não fumantes (Kim et al., 2013a) como em pacientes fumantes (Collisson et al., 2014) utilizando a metodologia das matrizes ternárias (Figura 4.63). Comparamos também as deleções encontradas em amostras tumorais nesses dois grupos de estudo e encontramos 57 deleções compartilhadas (Figura 4.64). Em todos os casos, as deleções foram encontradas em regiões não codificadoras e nenhuma delas é compartilhada entre os tecidos normais e tumorais. Entre as deleções compartilhadas por amostras tumorais, podemos citar uma deleção de um

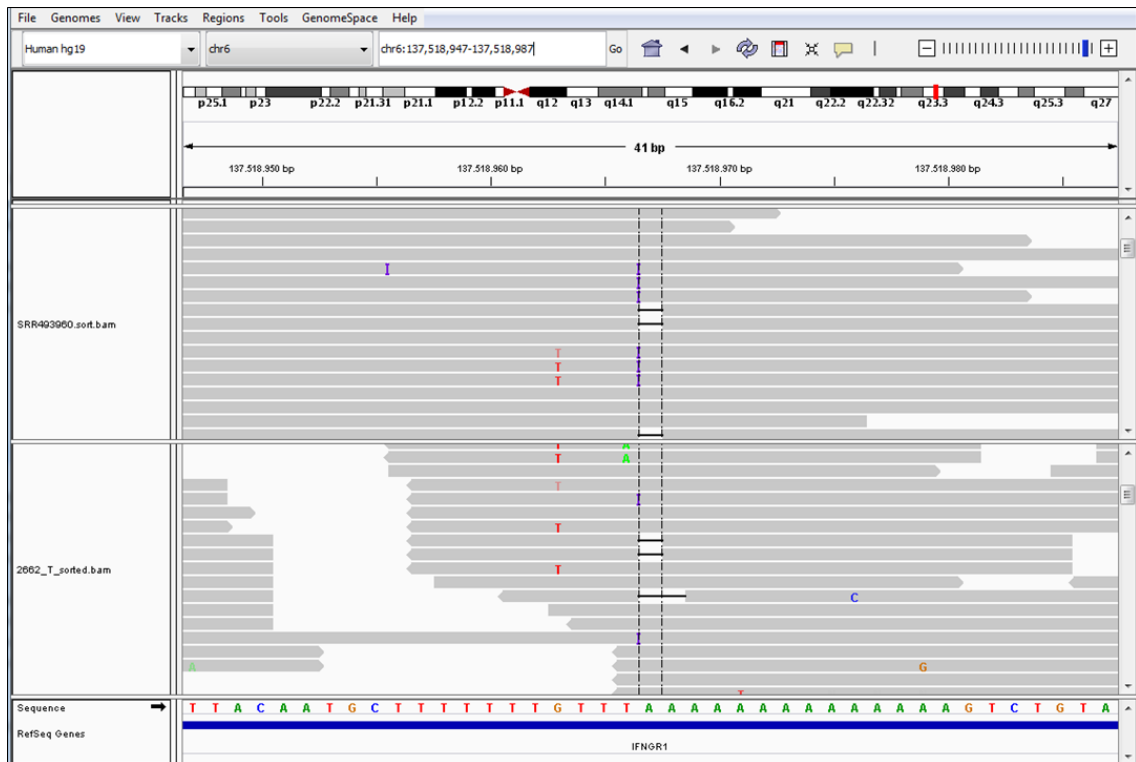
nucleotídeo na região 3' UTR do gene *IFNGR1* presente em 10 pacientes fumantes e em todos os pacientes fumantes (Figura 4.65).



**Figura 4.63.** Pequenas deleções identificadas pelas matrizes ternárias em amostras normais de Kim e colaboradores (2013a) (azul) e utilizando amostras normais de Collisson e colaboradores (2014) (verde).



**Figura 4.64** Pequenas deleções identificadas pelas matrizes ternárias em amostras tumorais de Kim e colaboradores (2013a) (azul) e utilizando amostras tumorais de Collisson e colaboradores (2014) (verde).



**Figura 4.65** Pequena deleção de um nucleotídeo no gene *IFNGR1* visualizada no programa IGV em amostra tumoral de paciente fumante (2662) e em paciente não fumante (p8).

#### 4.3.2 Pequenas deleções identificadas a partir de 66 genomas do 1000G

Não encontramos nenhuma deleção compartilhada em tecidos normais ou tumorais em pacientes não fumantes (Kim et al., 2013a) e em pacientes fumantes (Collisson et al., 2014) quando comparadas às deleções identificadas em 66 genomas do 1000G a partir dos critérios de exclusão empregados neste estudo.

## 5 CONCLUSÕES

- A metodologia de matrizes ternárias que foi desenvolvida primeiramente para armazenar dados de variantes de “splicing”, pode ser utilizada para identificar pequenas deleções de até 100 nucleotídeos em dados de RNA-Seq. Esta metodologia usou o genoma de referência GRCh37/hg19 e três conjuntos de dados distintos de adenocarcinoma de pulmão: uma linhagem celular, pacientes não fumantes e pacientes fumantes. Pudemos identificar deleções na linhagem H1975, como por exemplo, as deleções em *EPDR1* e *PTK2*. Estes achados começarão a ser validados experimentalmente em breve, uma vez que o nosso laboratório possui essa linhagem celular em cultura.

- Em média 80% das deleções identificadas pela metodologia das matrizes foram encontradas tanto com o uso de matrizes ternárias como pelo programa VarScan. Dentre as deleções encontradas nesta etapa, podemos citar uma deleção no gene *EGFR* com anotação no dbSNP e identificada em um paciente. Esta deleção é um exemplo de deleção já conhecida no câncer de pulmão. No entanto, pudemos identificar também deleções não associadas com o câncer de pulmão como é o caso da deleção no gene *CTSA*.

- Dados de RNA-Seq foram mapeados a 66 genomas do 1000G, sendo que três de cada população disponível no projeto. Esta etapa nos mostrou que podemos identificar pequenas deleções diferentes daquelas ao utilizar o genoma de referência GRCh37/hg19. Com essa estratégia diminuímos o viés da variação genética da população humana para identificar novos polimorfismos.

- Deleções em genes envolvidos de alguma forma com a biologia do câncer foram encontradas, como é o caso do supressor tumoral *CDH1* e do oncogene *AKT1* em dados de RNA-Seq de amostras de adenocarcinoma de pulmão.

- O uso de dados de transcriptoma ainda é pouco utilizado com o intuito de identificar polimorfismos de sequência de tamanho pequeno. Esses dados são mais abundantes em bancos de dados públicos e podemos ainda estudar a expressão através deles. Por isso, nosso trabalho é inovador por ter como foco o uso deste tipo de dado para a identificação de pequenas deleções. Neste sentido, encontramos exemplos de deleções como potenciais candidatas a serem germinativas exclusivas de pacientes fumantes ou não fumantes, abrindo oportunidade para a indicação de



novos candidatos a biomarcador que separem pacientes destes dois grupos em câncer de pulmão.

- Em nossas análises, não encontramos nenhuma deleção com sobreposição de coordenada entre as deleções identificadas a partir dos dados de Kim e colaboradores (2013a) e Collisson e colaboradores (2014) não mostrando nenhuma evidência de deleções associadas a fumantes e não fumantes.

- Os nossos achados podem contribuir para o estudo mais a fundo destas pequenas deleções de até 100 nucleotídeos e que podem em um futuro serem utilizadas como biomarcadores para o diagnóstico, estudo do prognóstico ou até para o desenvolvimento de novas drogas para o câncer de pulmão.

## 6 REFERÊNCIAS BIBLIOGRÁFICAS

Ahrendt SA, Decker PA, Alawi EA, Zhu Yr YR, Sanchez-Cespedes M, Yang SC, et al. Cigarette smoking is strongly associated with mutation of the K-ras gene in patients with primary adenocarcinoma of the lung. *Cancer*. 2001 Sep;92(6):1525–30.

Van Allen EM, Wagle N, Sucker A, Treacy DJ, Johannessen CM, Goetz EM, et al. The genetic landscape of clinical resistance to RAF inhibition in metastatic melanoma. *Cancer Discov*. 2014 Jan;4(1):94–109.

Anand P, Kunnumakkara AB, Kunnumakara AB, Sundaram C, Harikumar KB, Tharakan ST, et al. Cancer is a preventable disease that requires major lifestyle changes. *Pharm. Res.* Springer; 2008 Sep;25(9):2097–116.

Arribas AJ, Gó Mez-Abad C, Sánchez-Beato M, Martinez N, Dilisio L, Casado F, et al. Splenic marginal zone lymphoma: comprehensive analysis of gene expression and miRNA profiling. *Mod. Pathol*. 2013;26220(10):889–901.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry M, et al. Gene Ontology: tool for the unification of biology. *Nat Genet* . 2000 May ; 25(1): 25–29.

Batzer MA, Deininger PL. Alu repeats and human genomic diversity. *Nat. Rev. Genet*. 2002 May;3(5):370–9.

Bauer S, Robinson PN, Gagneur, J. Model-based gene set analysis for Bioconductor. *Bioinformatics*. 2011 Jul 1; 27(13): 1882–1883. Behm-Ansmant I, Kashima I, Rehwinkel J, Saulière J, Wittkopp N, Izaurralde E. mRNA quality control: An ancient machinery recognizes and degrades mRNAs with nonsense codons. *FEBS Lett*. 2007 Jun 19;581(15):2845–53.

Bhangale TR, Rieder MJ, Livingston RJ, Nickerson DA. Comprehensive identification and characterization of diallelic insertion-deletion polymorphisms in 330 human candidate genes. *Hum. Mol. Genet*. 2005 Jan 1;14(1):59–69.

Bhattacharya A, Ziebarth JD, Cui Y, Huntzinger E, Izaurralde E, Li L, et al. Systematic Analysis of microRNA Targeting Impacted by Small Insertions and Deletions in Human Genome. Castresana JS, editor. PLoS One; 2012 Sep 25;7(9):e46176.

Boland CR, Goel A. Microsatellite Instability in Colorectal Cancer. Gastroenterology. 2010 May;138(6):2073–2087.e3.

Brayer KJ, Segal DJ. Keep your fingers off my DNA: protein-protein interactions mediated by C2H2 zinc finger domains. Cell Biochem. Biophys. 2008;50(3):111–31.

Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012 Jul 18;487(7407):330–7.

Cancer Genome Atlas Research Network T, Cancer Genome Atlas Research Network. The Cancer Genome Atlas Pan-Cancer analysis project. Nat. Genet. Nature Research; 2013 Sep 26;45(10):1113–20.

Chen C-H, Liao B-Y, Chen F-C. Exploring the selective constraint on the sizes of insertions and deletions in 5' untranslated regions in mammals. BMC Evol. Biol. 2011a Dec 5;11(1):192.

Chen W, Li Z, Bai L, Lin Y. NF-kappaB in lung cancer, a carcinogenesis mediator and a prevention and therapy target. Front. Biosci. (Landmark Ed. NIH Public Access; 2011b Jan 1;16:1172–85.

Choragudi SF, Veeramachaneni G, Raman B, JS B. Molecular modeling and analysis of human and plant endo- $\beta$ -N-acetyl- glucosaminidases for mutations effects on function. Bioinformation. 2014 Aug 30;10(8):507–11.

Church DM, Schneider VA, Graves T, Auger K, Cunningham F, Bouk N, et al. Modernizing Reference Genome Assemblies. PLoS Biol. 2011 Jul 5;9(7):e1001091.

Cirulli ET, Singh A, Shianna K V, Ge D, Smith JP, Maia JM, et al. Screening the human exome: a comparison of whole genome and whole transcriptome

sequencing. *Genome Biol.* 2010;11(5):R57.

Collins FS, Drumm ML, Cole JL, Lockwood WK, Vande Woude GF, Iannuzzi MC. Construction of a general human chromosome jumping library, with application to cystic fibrosis. *Science.* 1987 Feb 27;235(4792):1046–9.

Collisson EA, Campbell JD, Brooks AN, Berger AH, Lee W, Chmielecki J, et al. Comprehensive molecular profiling of lung adenocarcinoma. *Nature.* 2014 Jul 9;511(7511):543–50.

Cortes DF, Sha W, Hower V, Blekherman G, Laubenbacher R, Akman S, et al. Differential gene expression in normal and transformed human mammary epithelial cells in response to oxidative stress. *Free Radic. Biol. Med.*. NIH Public Access; 2011 Jun 1;50(11):1565–74.

Deininger P. Alu elements: know the SINEs. *Genome Biol.* BioMed Central; 2011 Jan 28;12(12):236.

de la Chaux N, Messer PW, Arndt PF. DNA indels in coding regions reveal selective constraints on protein evolution in the human lineage. *BMC Evol Biol.* 2007 Oct 12;7:191

Dusl M, Senderek J, Muller JS, Vogel JG, Pertl A, Stucka R, et al. A 3'-UTR mutation creates a microRNA target site in the GFPT1 gene of patients with congenital myasthenic syndrome. *Hum. Mol. Genet.* 2015 Jun 15;24(12):3418–26.

Ellegren H. Microsatellites: simple sequences with complex evolution. *Nat. Rev. Genet.* 2004 Jun;5(6):435–45.

E pluribus unum. *Nat. Methods.* Nature Publishing Group; 2010 May;7(5):331–331.

Forbes SA, Bhamra G, Bamford S, Dawson E, Kok C, Clements J, et al. The Catalogue of Somatic Mutations in Cancer (COSMIC). *Curr. Protoc. Hum. Genet.*. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2008. p. Unit 10.11.

Gemignani F, Moreno V, Landi S, Moullan N, Chabrier A, Gutiérrez-Enríquez S, et al. A TP53 polymorphism is associated with increased risk of colorectal cancer and with reduced levels of TP53 mRNA. *Oncogene*. 2004 Mar 11;23(10):1954–6.

Gibbons DL, Byers LA, Kurie JM. Smoking, p53 mutation, and lung cancer. *Mol. Cancer Res. NIH Public Access*; 2014 Jan;12(1):3–13.

Glanzmann B, Lombard D, Carr J, Bardien S. Screening of two indel polymorphisms in the 5'UTR of the DJ-1 gene in South African Parkinson's disease patients. *J. Neural Transm*. 2014 Feb 20;121(2):135–8.

Govindan R, Ding L, Griffith M, Subramanian J, Dees ND, Kanchi KL, et al. Genomic Landscape of Non-Small Cell Lung Cancer in Smokers and Never-Smokers. *Cell. NIH Public Access*; 2012 Sep 14;150(6):1121–34.

Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, et al. Patterns of somatic mutation in human cancer genomes. *Nature*. 2007 Mar 8;446(7132):153–8.

Gu D, Scaringe WA, Li K, Saldivar J-S, Hill KA, Chen Z, et al. Database of somatic mutations in EGFR with analyses revealing indel hotspots but no smoking-associated signature. *Hum. Mutat*. 2007 Aug;28(8):760–70.

Haft DH, Selengut JD, White O. The TIGRFAMs database of protein families. *Nucleic Acids Res*. 2003 Jan;31(1):371–3.

Hollander MC, Maier CR, Hobbs EA, Ashmore AR, Linnoila RI, Dennis PA. Akt1 deletion prevents lung tumorigenesis by mutant K-ras. *Oncogene. NIH Public Access*; 2011 Apr 14;30(15):1812–21.

Iengar P. An analysis of substitution, deletion and insertion mutations in cancer genes. *Nucleic Acids Res.. Oxford University Press*; 2012 Aug;40(14):6401–13.

Imielinski M, Berger AH, Hammerman PS, Hernandez B, Pugh TJ, Hodis E, et al. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell*. 2012 Sep 14;150(6):1107–20.

Instituto Nacional de Cancer José Alencar Gomes da Silva. INCA - Instituto Nacional de Câncer - Estimativa 2016. Ministério da Saúde Inst. Nac. Cancer José Alencar Gomes da Silva. Rio de Janeiro; 2016.

Ipe J, Swart M, Burgess KS, Skaar TC. High-Throughput Assays to Assess the Functional Impact of Genetic Variants: A Road Towards Genomic-Driven Medicine. *Clin. Transl. Sci. Wiley-Blackwell*; 2017 Mar;10(2):67–77.

Jin G, Kim MJ, Jeon H-S, Choi JE, Kim DS, Lee EB, et al. PTEN mutations and relationship to EGFR, ERBB2, KRAS, and TP53 mutations in non-small cell lung cancers. *Lung Cancer*. 2010 Sep;69(3):279–83.

Jones S, Wang T-L, Shih I-M, Mao T-L, Nakayama K, Roden R, et al. Frequent Mutations of Chromatin Remodeling Gene ARID1A in Ovarian Clear Cell Carcinoma. *Science (80-. )*. 2010 Oct 8;330(6001):228–31.

Kase S, Sugio K, Yamazaki K, Okamoto T, Yano T, Sugimachi K. Expression of E-cadherin and beta-catenin in human non-small cell lung cancer and the clinical significance. *Clin. Cancer Res.*. 2000 Dec;6(12):4789–96.

Kent WJ. BLAT--the BLAST-like alignment tool. *Genome Res*. 2002 Apr;12(4):656–64.

Kharitonov A, Chen Z, Sures I, Wang H, Schilling J, Ullrich A. A family of proteins that inhibit signalling through tyrosine kinase receptors. *Nature*. 1997 Mar 13;386(6621):181–6.

Kim SC, Jung Y, Park J, Cho S, Seo C, Kim J, et al. A High-Dimensional, Deep-Sequencing Study of Lung Adenocarcinoma in Female Never-Smokers. *PLoS One*. 2013a;8(2).

Kim T-M, Laird PW, Park PJ. The Landscape of Microsatellite Instability in Colorectal and Endometrial Cancer Genomes. *Cell*. 2013b Nov 7;155(4):858–68.

Kim T-M, Park PJ. A Genome-wide View of Microsatellite Instability: Old Stories of Cancer Mutations Revisited with New Sequencing Technologies. *Cancer Res.* 2014 Nov 15;74(22):6377–82.

Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, et al. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*. 2009 Sep 1;25(17):2283–5.

Kuivaniemi H, Tromp G, Prockop DJ. Mutations in fibrillar collagens (types I, II, III, and XI), fibril-associated collagen (type IX), and network-forming collagen (type X) cause a spectrum of diseases of bone, cartilage, and blood vessels. *Hum. Mutat.* 1997;9(4):300–15.

Kumar A, Coleman I, Morrissey C, Zhang X, True LD, Gulati R, et al. Substantial interindividual and limited intraindividual genomic diversity among tumors from men with metastatic prostate cancer. *Nat. Med.* 2016 Apr;22(4):369–78.

Kumar V, Abbas AK, Fausto NF, Aster JC. Robbins e Cotran Bases Patológicas das Doenças. 8a ed. Elsevier Ltd; 2010.

de la Chaux N, Messer PW, Arndt PF. DNA indels in coding regions reveal selective constraints on protein evolution in the human lineage. *BMC Evol. Biol. BioMed Central*; 2007 Oct 12;7:191.

LaFlamme B. Microexons on the brain. *Nat. Genet.* 2015 Jan 28;47(2):105–105.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. *Nature*. 2001 Feb 15;409(6822):860–921.

Lappalainen I, Almeida-King J, Kumanduri V, Senf A, Spalding JD, ur-Rehman S, et al. The European Genome-phenome Archive of human data consented for biomedical research. *Nat. Genet.. Nature Research*; 2015 Jun 26;47(7):692–5.

Lee MW, Kim DS, Lee JH, Lee BS, Lee SH, Jung HL, et al. Roles of AKT1 and AKT2 in non-small cell lung cancer cell survival, growth, and migration. *Cancer Sci.* 2011 Oct;102(10):1822–8.

Lee W, Jiang Z, Liu J, Haverty PM, Guan Y, Stinson J, et al. The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature.* 2010 May 27;465(7297):473–7.

Leslie NR, Downes CP. PTEN function: how normal cells control it and tumour cells lose it. *Biochem. J.. Portland Press Ltd*; 2004 Aug 15;382(Pt 1):1–11.

Letunic I, Copley RR, Schmidt S, Ciccarelli FD, Doerks T, Schultz J, et al. SMART 4.0: towards genomic data integration. *Nucleic Acids Res.* 2004 Jan;32(Database issue):D142-4.

Li H, Handsaker B, Wysoker A, Fennel T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* 2009 Aug 15;25(16):2078-9.

Lynch TJ, Bell DW, Sordella R, Gurubhagavatula S, Okimoto RA, Brannigan BW, et al. Activating Mutations in the Epidermal Growth Factor Receptor Underlying Responsiveness of Non–Small-Cell Lung Cancer to Gefitinib. *N. Engl. J. Med.* 2004 May 20;350(21):2129–39.

Marat AL, Dokainish H, McPherson PS. DENN domain proteins: regulators of Rab GTPases. *J. Biol. Chem.* 2011 Apr 22;286(16):13791–800.

Marchler-Bauer A, Panchenko AR, Shoemaker BA, Thiessen PA, Geer LY, Bryant SH. CDD: a database of conserved domain alignments with links to domain



three-dimensional structure. *Nucleic Acids Res.* 2002 Jan 1;30(1):281–3.

Marwitz S, Depner S, Dvornikov D, Merkle R, Szczygie M, Müller-Decker K, et al. Downregulation of the TGF Pseudoreceptor BAMBI in Non-Small Cell Lung Cancer Enhances TGF Signaling and Invasion. *Cancer Res.* 2016 Jul 1;76(13):3785–801.

Mills RE, Luttig CT, Larkins CE, Beauchamp A, Tsui C, Pittard WS, et al. An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res.* 2006 Sep;16(9):1182–90.

Mitsudomi T, Yatabe Y. Epidermal growth factor receptor in relation to tumor development: EGFR gene and cancer. *FEBS J.* 2010 Jan;277(2):301–8.

Montgomery SB, Goode DL, Kvikstad E, Albers CA, Zhang ZD, Mu XJ, et al. The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. *Genome Res.* 2013 May;23(5):749–61.

Morrison AA, Viney RL, Ladomery MR. The post-transcriptional roles of WT1, a multifunctional zinc-finger protein. *Biochim. Biophys. Acta.* 2008 Jan;1785(1):55–62.

Mouradov D, Sloggett C, Jorissen RN, Love CG, Li S, Burgess AW, et al. Colorectal cancer cell lines are representative models of the main molecular subtypes of primary cancer. *Cancer Res.* 2014 Jun 15;74(12):3238–47.

Mullaney JM, Mills RE, Pittard WS, Devine SE. Small insertions and deletions (INDELs) in human genomes. *Hum. Mol. Genet.* 2010 Oct 15;19(R2):R131-6.

Mullapudi N, Ye B, Suzuki M, Fazzari M, Han W, Shi MK, et al. Genome Wide Methylation Alterations in Lung Cancer. *PLoS One.* 2015;10(12):e0143826.

Murakami S, Takaoka Y, Ashida H, Yamamoto K, Narimatsu H, Chiba Y. Identification and characterization of endo- $\beta$ -N-acetylglucosaminidase from methylotrophic yeast *Ogataea minuta*. *Glycobiology.* 2013 Jun;23(6):736–44.

Naumovski L, Cleary ML. The p53-binding protein 53BP2 also interacts with Bc12 and impedes cell cycle progression at G2/M. *Mol. Cell. Biol.* 1996 Jul;16(7):3884–92.

NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 2013 Jan;41(Database issue):D8–20.

Nielsen R, Paul JS, Albrechtsen A, Song YS. Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet.*. Nature Publishing Group; 2011 Jun];12(6):443–51.

O’Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 2016 Jan;44(D1):D733–45.

Ostrer H. A genetic profile of contemporary Jewish populations. *Nat. Rev. Genet.* 2001 Nov 1;2(11):891–8.

Parker BS, Rautela J, Hertzog PJ. Antitumour actions of interferons: implications for cancer therapy. *Nat. Rev. Cancer.* Nature Publishing Group; 2016;16(3):131–44.

Parsons DW, Jones S, Zhang X, Lin JC-H, Leary RJ, Angenendt P, et al. An Integrated Genomic Analysis of Human Glioblastoma Multiforme. *Science (80-. ).].* 2008 Sep 26;321(5897):1807–12.

Parsons DW, Li M, Zhang X, Jones S, Leary RJ, Lin JC-H, et al. The Genetic Landscape of the Childhood Cancer Medulloblastoma. *Science (80-. ).* 2011 Jan 28;331(6016):435–9.

Piperdi B, Merla A, Perez-Soler R. Targeting Angiogenesis in Squamous Non-Small Cell Lung Cancer. *Drugs.* 2014 Mar 28;74(4):403–13.

Pontius JU, Mullikin JC, Smith DR, Lindblad-Toh K, Gnerre S, Clamp M, et al. Initial sequence and comparative analysis of the cat genome. *Genome Res.* 2007

Nov; 17(11): 1675–1689.

Pruitt KD, Tatusova T, Klimke W, Maglott DR. NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res.* Oxford University Press; 2009 Jan;37(Database):D32–6.

Prutzman KC, Gao G, King ML, Iyer V V, Mueller GA, Schaller MD, et al. The focal adhesion targeting domain of focal adhesion kinase contains a hinge region that modulates tyrosine 926 phosphorylation. *Structure.* 2004 May;12(5):881–91.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. *Nat. Biotechnol. Nature Research;* 2011 Jan;29(1):24–6.

Sahab ZJ, Hall MD, Zhang L, Cheema AK, Byers SW. Tumor Suppressor RARRES1 Regulates DLG2, PP2A, VCP, EB1, and Ankrd26. *J. Cancer.* Ivyspring International Publisher; 2010 Jun 2;1:14–22.

Sanborn JZ, Chung J, Purdom E, Wang NJ, Kakavand H, Wilmott JS, et al. Phylogenetic analyses of melanoma reveal complex patterns of metastatic dissemination. *Proc. Natl. Acad. Sci. U. S. A.* 2015 Sep 1;112(35):10995–1000.

Scaringe WA, Li K, Gu D, Gonzalez KD, Chen Z, Hill KA, et al. Somatic microindels in human cancer: the insertions are highly error-prone and derive from nearby but not adjacent sense and antisense templates. *Hum. Mol. Genet.* 2008 Sep 15;17(18):2910–8.

Semb H, Christofori G. The tumor-suppressor function of E-cadherin. *Am. J. Hum. Genet.* 1998 Dec;63(6):1588–93.

Seshagiri S, Stawiski EW, Durinck S, Modrusan Z, Storm EE, Conboy CB, et al. Recurrent R-spondin fusions in colon cancer. *Nature.* 2012 Aug 30;488(7413):660–4.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al.

dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 2001 Jan 1;29(1):308–11.

Skinner ME, Uzilov A V., Stein LD, Mungall CJ, Holmes IH. JBrowse: A next-generation genome browser. *Genome Res.* 2009 Sep 1;19(9):1630–8.

Slotkin RK, Martienssen R. Transposable elements and the epigenetic regulation of the genome. *Nat. Rev. Genet.* 2007 Apr;8(4):272–85.

van der Spoel A, Bonten E, D’Azzo A. Transport of human lysosomal neuraminidase to mature lysosomes requires protective protein/cathepsin A. *EMBO J.* 1998 Mar 16;17(6):1588–97.

Stembridge NS, Vandersteen AM, Ghali N, Sawle P, Nesbitt M, Pollitt RC, et al. Clinical, structural, biochemical and X-ray crystallographic correlates of pathogenicity for variants in the C-propeptide region of the COL3A1 gene. *Am. J. Med. Genet. A.* 2015 Aug;167A(8):1763–72.

Stenson PD, Ball E V., Mort M, Phillips AD, Shaw K, Cooper DN. The Human Gene Mutation Database (HGMD) and Its Exploitation in the Fields of Personalized Genomics and Molecular Evolution. *Curr. Protoc. Bioinforma.* Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2012. p. Unit1.13.

Stenson PD, Ball E V., Mort M, Phillips AD, Shaw K, Cooper DN. The Human Gene Mutation Database. 2016. Available from: <http://www.hgmd.cf.ac.uk/>

Stephens PJ, Tarpey PS, Davies H, Van Loo P, Greenman C, Wedge DC, et al. The landscape of cancer genes and mutational processes in breast cancer. *Nature.* 2012 May 16;486(7403):400–4.

Stewart BW, Wild CP. World cancer report 2014. *World Heal. Organ.* 2014;1–2.

Stratton MR, Campbell PJ, Futreal PA. The cancer genome. *Nature.* Macmillan Publishers Limited. All rights reserved; 2009 Apr 9;458(7239):719–24.

Tao R, Hu S, Wang S, Zhou X, Zhang Q, Wang C, et al. Association between indel polymorphism in the promoter region of lncRNA GAS5 and the risk of hepatocellular carcinoma. *Carcinogenesis*. 2015 Oct;36(10):1136–43.

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin E V, et al. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*. BioMed Central; 2003 Sep;4:41.

Tatusov RL, Galperin MY, Natale DA, Koonin E V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res*. 2000 Jan;28(1):33–6.

Tavares R, de Miranda Scherer N, Pauletti BA, Araújo E, Folador EL, Espindola G, et al. SpliceProt: a protein sequence repository of predicted human splice variants. *Proteomics*. 2014 Feb;14(2–3):181–5.

Tavares R, Pauletti BA, Franco A, Leme P, Martins-de-souza D. Unveiling alternative splice diversity from human oligodendrocyte proteome data. *J Proteomics*. 2017 Jan 16;151:293-301.

The 1000 Genomes Consortium T 1000 GP. A map of human genome variation from population-scale sequencing. *Nature*. 2010 Oct 28;467(7319):1061–73.

The 1000 Genomes Consortium T 1000 GP. An integrated map of genetic variation from 1,092 human genomes. *Nature*. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2012 Nov 1;491(7422):56–65.

Toonkel RL, Borczuk AC, Powell CA. Tgf-beta signaling pathway in lung adenocarcinoma invasion. *J. Thorac. Oncol. NIH Public Access*; 2010 Feb;5(2):153–7.

Toyokawa G, Masuda K, Daigo Y, Cho H-S, Yoshimatsu M, Takawa M, et al. Minichromosome Maintenance Protein 7 is a potential therapeutic target in human cancer and a novel prognostic marker of non-small cell lung cancer. *Mol. Cancer*.

BioMed Central; 2011 May 28;10:65.

Travis WD, Brambilla E, Müller-Hermelink HK, Harris CC. Pathology and genetics of tumours of the lung. *Bull. World Health Organ.* 2004;50(1–2):9–19.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. *Science.* 2001 Feb 16 ;291(5507):1304–51.

Vignal A, Milan D, SanCristobal M, Eggen A. A review on SNP and other types of molecular markers and their use in animal genetics. *Genet. Sel. Evol. BioMed Central;* 2002;34(3):275.

Wajnberg G, Passeti F. Using high-throughput sequencing transcriptome data for INDEL detection: challenges for cancer drug discovery. *Expert Opin. Drug Discov..* 2016 Jan 20];

Wang K, Yuen ST, Xu J, Lee SP, Yan HHN, Shi ST, et al. Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer. *Nat. Genet.* 2014 Jun;46(6):573–82.

Wei Q, Li J, Liu T, Tong X, Ye X. Phosphorylation of minichromosome maintenance protein 7 (MCM7) by cyclin/cyclin-dependent kinase affects its function in cell cycle regulation. *J. Biol. Chem.. American Society for Biochemistry and Molecular Biology;* 2013 Jul 5;288(27):19715–25.

Weinberg RA. *The biology of cancer.* Garland Sciences. 2<sup>nd</sup> edition. 2006. 876 pages.

Weiner AM. SINEs and LINEs: the art of biting the hand that feeds you. *Curr. Opin. Cell Biol..* 2002 Jun;14(3):343–50.

Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res. Oxford University Press;* 2007 Jan;35(Database issue):D5-12.

Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* Oxford University Press; 2008 Jan;36(Database issue):D13-21.

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.*. 2007 Dec;8(12):973–82.

Williams DS, Bird MJ, Jorissen RN, Yu YL, Walker F, Zhang HH, et al. Nonsense Mediated Decay Resistant Mutations Are a Source of Expressed Mutant Proteins in Colon Cancer Cell Lines with Microsatellite Instability. Ko BCB, editor. *PLoS One*. 2010 Dec 31;5(12):e16012.

Wu P, Walker BA, Broyl A, Kaiser M, Johnson DC, Kuiper R, et al. A gene expression based predictor for high risk myeloma treated with intensive therapy and autologous stem cell rescue. *Leuk. Lymphoma*. 2015 Mar;56(3):594–601.

Yin J-Y, Shen J, Dong Z-Z, Huang Q, Zhong M-Z, Feng D-Y, et al. Effect of eIF3a on Response of Lung Cancer Patients to Platinum-Based Chemotherapy by Regulating DNA Repair. *Clin. Cancer Res.* 2011 Jul 1;17(13):4600–9.

Yin S, Yang J, Lin B, Deng W, Zhang Y, Yi X, et al. Exome sequencing identifies frequent mutation of MLL2 in non-small cell lung carcinoma from Chinese patients. *Sci. Rep.* 2014 Aug 12;4:6036.

Yu C-R, Mahdi RM, Ebong S, Vistica BP, Gery I, Egwuagu CE, et al. Suppressor of Cytokine Signaling 3 Regulates Proliferation and Activation of T-helper Cells. *J. Biol. Chem.*. BioMed Central; 2003 Aug 8;278(32):29752–9.

Yuan Z, Shin J, Wilson A, Goel S, Ling YH, Ahmed N, et al. An A13 repeat within the 3'-untranslated region of epidermal growth factor receptor (EGFR) is frequently mutated in microsatellite instability colon cancers and is associated with increased EGFR expression. *Cancer Res.* 2009;69(19):7811–8.

Zhang X, Lin H, Zhao H, Hao Y, Mort M, Cooper DN, et al. Impact of human

pathogenic micro-insertions and micro-deletions on post-transcriptional regulation.  
Hum. Mol. Genet.. 2014 Jun 1;23(11):3024–34.




## **7 ANEXO**

### **7.1 Using high-throughput sequencing transcriptome data for INDEL detection: challenges for cancer drug discovery**

REVIEW

## Using high-throughput sequencing transcriptome data for INDEL detection: challenges for cancer drug discovery

Gabriel Wajnberg  and Fabio Passetti 

Laboratory of Functional Genomics and Bioinformatics, Oswaldo Cruz Institute, Fundação Oswaldo Cruz (FIOCRUZ), Rio de Janeiro, RJ, Brazil

### ABSTRACT

**Introduction:** A cancer cell is a mosaic of genomic and epigenomic alterations. Distinct cancer molecular signatures can be observed depending on tumor type or patient genetic background. One type of genomic alteration is the insertion and/or deletion (INDEL) of nucleotides in the DNA sequence, which may vary in length, and may change the encoded protein or modify protein domains. INDELS are associated to a large number of diseases and their detection is done based on low-throughput techniques. However, high-throughput sequencing has also started to be used for detection of novel disease-causing INDELS. This search may identify novel drug targets.

**Areas Covered:** This review presents examples of using high-throughput sequencing (DNA-Seq and RNA-Seq) to investigate the incidence of INDELS in coding regions of human genes. Some of these examples successfully utilized RNA-Seq to identify INDELS associated to diseases. In addition, other studies have described small INDELS related to chemo-resistance or poor outcome of patients, while structural variants were associated with a better clinical outcome.

**Expert opinion:** On average, there is twice as much RNA-Seq data available at the most used repositories for such data compared to DNA-Seq. Therefore, using RNA-Seq data is a promising strategy for studying cancer samples with unknown mechanisms of drug resistance, aiming at the discovery of proteins with potential as novel drug targets.

### ARTICLE HISTORY

Received 14 September 2015  
Accepted 15 January 2016  
Published online 5 February 2016

### KEYWORDS

Bioinformatics; INDEL; transcriptomics; RNA-Seq; drug


### 1. Introduction

Genomic insertion and deletion (INDEL) may affect one to thousands of nucleotides in a DNA sequence.[1] INDELS can be classified according to their size: ‘small INDELS’ ranging from 1 to 543 nucleotides in length [2]; ‘microindels’, up to 50 nucleotides [3]; and ‘structural variant’, which is frequently detected in tumors and usually is more than 10,000 nucleotides long.[4] Other INDELS occur due to replication slippage during DNA copy through meiosis and may vary in length.[5] The region of repeated nucleotides is termed ‘microsatellite’ and consists of 5–50 tandem repetitions of short-nucleotide sequences.[6]

INDELS can potentially change the coding sequence by altering splice sites or the supposed encoded amino acid. One to multiple amino acid deletions or insertions can modify the protein sequence and may create a premature stop codon or a frameshift. Most INDELS in human coding regions do not affect known functional protein domains and have been detected less than expected in  $\alpha$ -helices.[7]

INDELS within coding regions have been associated to diseases. According to Stenson and colleagues

(2014) [8] 24% of all Mendelian diseases in the Human Gene Mutation Database (HGMD) are caused by INDELS. For example, cystic fibrosis is frequently caused by the loss of three nucleotides in the coding region of the *CFTR* gene, leading to the loss of function of the translated protein,[9] whereas the fragile X syndrome is associated with the insertion of a three-nucleotide repeat.[10] In addition, in some Mendelian diseases, RNA-binding proteins cannot bind to their RNA-binding site due to a sequence disruption caused by INDELS. Those cases are annotated in the HGMD and most are located closer to splice sites, possibly altering the splicing machinery, or within exons in which there is some evidence of alternative splicing.[11] However, some INDELS can be located in non-coding regions, primary in 3' and 5' untranslated regions (UTR). Besides the biological complexity behind such alterations and the challenge to apply this knowledge to drug discovery, considering that they do not affect the protein sequence, there are examples of INDEL detection on 3' UTRs that change the miRNA binding site. For example, two deletions were described in the 3' UTR region of the *IKK1* gene that alter the binding site of the miR-223, according to genome-wide association studies

CONTACT Fabio Passetti  [passetti@fiocruz.br](mailto:passetti@fiocruz.br)

© 2016 Taylor & Francis

data.[12] In contrast, INDELs occurring in the 5' UTR are predicted to alter translation initiation motifs, upstream start codons (uAUGs), and upstream open reading frames.[13] For example, two deletions in the 5' UTR in the *DJ-1* gene were identified in a large number of Parkinson's disease patients.[14]

INDELs and other polymorphisms have been described to be associated to tumor growth and cell survival mechanisms.[15,16] For example, a 16-nucleotide insertion at the *TP53* in intron 3 increases the risk of colorectal cancer,[17] and there is a 5-nucleotide INDEL in the *GAS5* gene that increases hepatocellular carcinoma risk in the Chinese population.[18] Genomic instability is another common feature detected in tumors. For example, some subtypes of colorectal cancers are caused by a hypermutable phenotype known as microsatellite instability (MSI), in which cells lose the DNA mismatch repair mechanism.[19,20] The occurrence of MSI has been identified in the 3' UTR region of the *EGFR* gene in colorectal cancer.[21] Microindels have also been identified in the *TP53* gene with similar frequencies to those in other human cancers such as breast, bladder, colorectal, ovary, mouth, lung, and stomach.[3]

The number and variety of sequence polymorphisms in tumors is enormous, leading to the creation of a database called the Catalog of Somatic Mutations in Cancer (COSMIC database).[22] The COSMIC database stores all somatic mutations identified in genes and associates them with different types of information, such as drug target. One of these genes is *EGFR*, which comprises a large number of mutations. Due to the strong association between *EGFR* changes and occurrence of non-small-cell lung cancer (NSCLC), a database specific for this gene was created: the *EGFR* mutation database.[23]

However, there are a variety of INDELs that are not associated to diseases. Gene polymorphism has been described as an important mechanism for evolution and a molecular feature of distinct wild populations in the tree of life. For instance, INDELs have been proposed as genetic markers for wild-wolf populations [24] and have also analyzed to determine sequence divergence in primate evolution. Due to their greater abundance between human and chimpanzee genomes, which is even higher than the occurrence of single nucleotide polymorphisms (SNP), INDELs are considered to have been important to the establishment of the human species.[25]

The publication of the first draft of the human genome in 2001 generated new perspectives for the analysis of human genes, particularly for the identification of polymorphisms on a large scale, including

INDELs.[26,27] Until the publication of the first draft of the human genome, no genome assembly based on linear sequences of chromosomes was available.

The Genome Reference Consortium (GRC) released the human reference genome assembly from 13 individuals, of which the latest version is GRCh38.[28,29] This assembly of the human genome is used to perform sequence comparisons among the reference sequence and a given target to detect SNP, INDEL, or structural variants, including copy number variation. For example, array-based Comparative Genomic Hybridization is a technique that identifies structural variants in chromosome regions [30,31] while a more specific array, called SNP-array, targets millions of SNPs and identifies small INDELs and structural variants.[32] Small deletions in genes such as *RUNX1*, *CEBPA*, *ETV6*, and *CDK2A* were identified in acute myeloid leukemia (AML) [33] and structural variants were also identified in rare tumors such as mesothelioma (lost copies of *NF2* gene) and ependymoma (lost copies of *SCHIP-1* gene) [34] using SNP-array. In 2005, the restriction of this technology to only detecting known sequence variations was surpassed with the launching of the first high-throughput sequencing (HTS) platform.[35] The HTS platform permits faster and higher throughput and the identification of new polymorphisms. Craig Venter was the first individual to have his complete genome sequenced; around 825,000 INDELs were identified.[36] The second complete human genome was published in 2008 using the DNA of James Watson,[37] and this study identified 222,718 INDELs, among which 113,539 were previously known.

To identify INDELs in HTS data, computer algorithms were created to support the analysis of new data formats and throughput. The raw data from an ordinary HTS run needs to be properly processed and analyzed. To this end, each step of the process demands different types of bioinformatics tools. These steps can be divided into: (1) Adapter removal and quality filtering; (2) Mapping; (3) Data manipulation and quality control; (4) Data visualization; and (5) INDEL calling (Figure 1). A variety of bioinformatics tools are used for variant calling, including VarScan [38] and GATK [39] for small INDELs; and PINDEL,[40] BreakSeek,[41] and IndelMINER [42] for structural variant breakpoint. Taking advantage of the HTS technology and some of these algorithms, 2500 individuals from different types of population had their polymorphism mapped and made available as part of the 1000 Genomes Project (1000G).[43,44] Because the current available human genome reference sequence is an assembly of only 13 different North American individuals, this project may increase the knowledge of genome diversity not



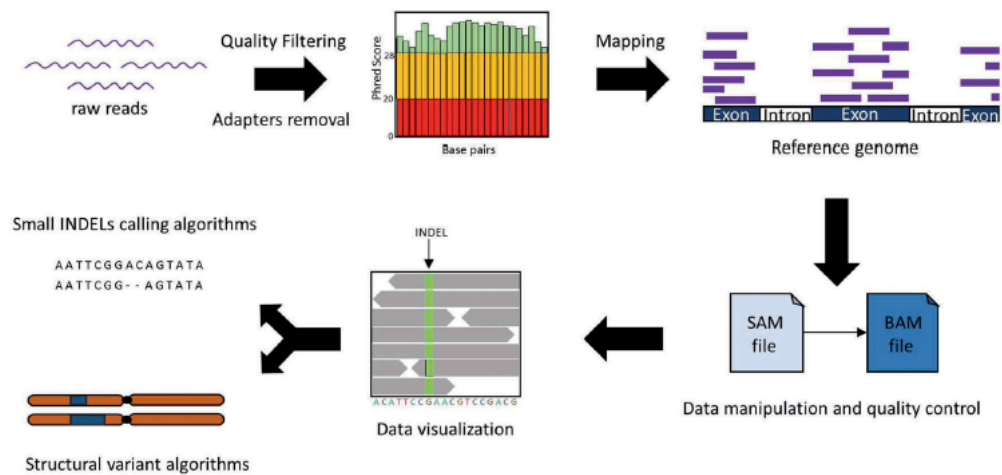


Figure 1. Workflow to call sequence variants from an HTS run.

associated to diseases and may improve the annotation of sequence polymorphisms.

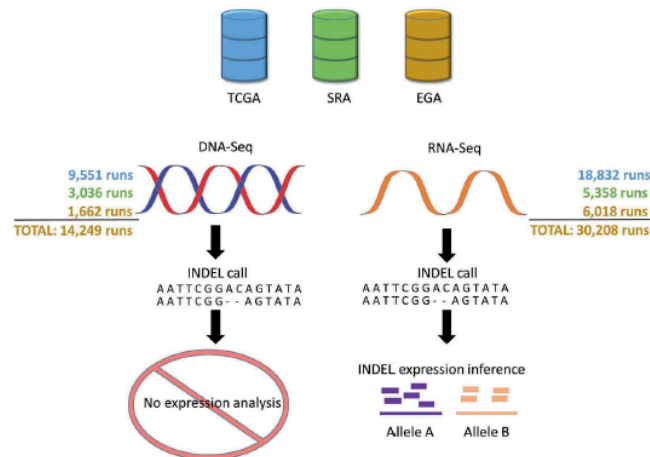
Different strategies emerged with the improvement of HTS technology, and exome sequencing can be cited as one of the most frequently used. In exome sequencing, only the exons are sequenced after probe capture, instead of using whole genome data.[45] MSIs were identified in coding regions of different types of cell lines using this technology.[46] HTS technology has also been used to identify INDELS in cancer genomes. Whole genome shotgun (WGS) and exome sequencing are the most commonly used approaches to identify INDELS. Initial efforts in cancer studies identified INDELS using WGS in breast cancer,[47] lung cancer,[48] and melanoma.[49] Using this approach, novel INDELS were identified in genes not previously associated with AML, such as *WAC*, *SMC3*, *DIS3*, *DDX41*, and *DAXX*.<sup>[50]</sup> Another study revealed a specific deletion of *CDKN2A* gene in only one sample of many in the same bladder tumor, revealing heterogeneity between cancer cells.<sup>[51]</sup> A pan-cancer cohort using 12 different types of tumors identified genes with high INDEL incidence, including the *TP53* gene in 42% of cancer samples and also in 35% of serous ovarian samples.<sup>[52]</sup>

Aiming to organize the many efforts in this field and the data generated, a database of cancer high-throughput raw data was created: The Cancer Genome Atlas (TCGA).<sup>[53]</sup> The reanalysis of these published data contributed to the identification of novel INDELS, with additional MSIs being identified in colorectal and endometrial cancers in TCGA exome and WGS data.<sup>[54]</sup> However, some studies started to use transcriptome instead of genomic data to identify INDELS.<sup>[55,56]</sup> The advantage of using RNA-Seq data is that transcriptome

data are more abundant than data on genomic sequences available in public databases. Another important aspect of using transcriptome data is that, besides identifying coding polymorphisms, the expression of such polymorphisms can be estimated. As depicted in Figure 2, TCGA database has 18,832 RNA-Seq runs and 9551 DNA-Seq runs; the Sequence Read Archive database has 5358 RNA-Seq runs and 3036 DNA-Seq runs; and the European Genome-Phenome Archive database has 6018 RNA-Seq runs and 1662 DNA-Seq runs. Therefore, there is more than twice as much RNA-Seq data as DNA-Seq on the aforementioned databases, and these valuable data have been rarely used for INDEL detection. A comparison between a sample analyzed with WGS and RNA-Seq shows that RNA-Seq can identify 81% of the coding polymorphisms identified using WGS.<sup>[57]</sup> Therefore, an emerging strategy is being employed by many research groups – using RNA-Seq data to search for INDELS.

## 2. INDELS identified in cancer high-throughput sequencing transcriptome data

The search for genomic sequence alterations is a challenge in cancer research, considering that DNA-Seq data from tumor samples are rare in public repositories. Xu and colleagues <sup>[58]</sup> innovatively used publicly available RNA-Seq data to search for somatic mutations in five prostate tumors. Using this approach, they were able to identify 116 somatic mutations in coding regions in 92 genes. Among these mutations, only 61 were associated with disruptive changes, such as frameshift mutations. For



**Figure 2.** Comparison between DNA-Seq and RNA-Seq for the identification of INDELs in cancer samples using available public data from The Cancer Genome Atlas (TCGA), NCBI's Sequence Read Archive (SRA), and the European Genome-phenome Archive (EGA).

example, an insertion of one nucleotide was detected in the *TNFSF10* gene (also known as TNF-related apoptosis-inducing ligand or *TRAIL*), which disrupts the open read frame of the protein's sequence responsible for inducing apoptosis.[58] In another study, INDELs were identified in another tumor suppressor gene (*ARID1A*) using RNA-Seq. In that study, Wiegand and colleagues [59] identified a deletion of a single guanine in an ovarian clear-cell carcinoma, which has also been identified using exome sequencing. In addition, these authors identified an insertion of a single cytosine in this same gene in an ovarian malignant adenocarcinoma cell line, a finding that was validated by exome sequencing.[59] RNA-Seq data were used to investigate 17 breast cancer patients, and were able to identify approximately 59 INDELs per sample, but only 38 unique INDELs in 10 breast-cancer-associated genes (*ATM*, *BRCA1*, *BRCA2*, *BRIP1*, *CASP8*, *CDH1*, *CHEK1*, *PTEN*, *STK11*, or *TP53*). In their findings, the authors identified a deletion of a single guanine within the *BRCA1* gene that shifted the stop codon downstream in a metastatic triple-negative breast cancer patient[60] and was nonresponsive to commonly used breast cancer drugs such as Trastuzumab, Pertuzumab, and Lapatinib.[61]

In order to evaluate the role of polymorphism in 26 patients with thyroid tumor and survivors of the nuclear accident in Chernobyl, Ricarte-Filho and colleagues [62] identified 17 INDELs predicted to cause frameshift. Among them, two were identified in the following cancer-associated genes: *PCNT*, in which overexpression is associated with uncontrolled cell

cycle in breast cancer,[63] and *PRDX5*, which is related to drug resistance in melanoma tumors.[64]

The use of exome sequencing simultaneously with RNA-Seq is also an innovative strategy. This approach has been used to identify 1661 INDELs in 48 gastric cancer patients, including a small INDEL in *MAP2K4* gene causing frameshift, and 139 INDELs affecting splice donor or acceptor sites.[65] The *MAP2K4* gene has already been described as a tumor suppressor in lung adenocarcinoma tumors inhibiting tumor cell invasion.[66] Another study used the same innovative strategy to identify polymorphisms in osteosarcoma patients; and three out of the five small INDELs found can cause modifications in the open read frame.[67] One of these INDELs occurs in the *ALK* gene, a frequently mutated gene in this type of cancer.[68,69]

Atak and colleagues [70] used RNA-Seq to dissect the mutational landscape of T-cell acute lymphoblastic leukemia and identified INDELs in this subtype of leukemia. The authors described two INDELs in the *PHF6* gene and an additional INDEL in the *NOTCH1* gene. In all cases, INDELs have also been confirmed by exome sequencing data obtained from the same patient.

The usage of the RNA-Seq platform to identify structural variants is another challenge in bioinformatics. For example, Patel and colleagues [71] analyzed glioblastoma tumors and were able to identify lost regions of chromosome 10 and gained regions in chromosome 7. In addition, glioblastoma samples with deletions in *EGFR* gene have been identified, including the EGFRV8 variant in a tumor sample,[71] a frequent INDEL in glioblastoma.[72]

A large prospective search using 675 cancer cell lines described 1437 novel INDELS in an effort using RNA-Seq and SNP-arrays. Most of these INDELS were detected in colorectal and uterus cancer cell lines and located in the following cancer-associated genes: *TP53*, *KRAS*, *CDKN2A*, *BRAF*, and *PTEN*. In addition, MAPK, cell cycle, and FGFR pathways were the ones most altered by these INDELS.[73] In conclusion, INDELS can be identified using RNA-Seq data and their expression can be estimated. This strategy can provide information on how the INDEL affects the protein's sequence and thus it might be used as a cancer drug target.

### 3. Cancer drugs targeting proteins with genomic INDELS

Genes affected by INDELS can potentially produce proteins with altered functional domains, which may alter known drug targets (Table 1). For example, the epidermal growth factor receptor, encoded by *EGFR* gene, has an important role in the initiation of different types of tumors. This gene encodes a cell-surface receptor of the ErbB family of tyrosine kinase that regulates cell proliferation and survival as well as differentiation of tumors. Many ligands are responsible to activate EGFR, including the epidermal growth factor and the transforming growth factor alpha. This protein has an extracellular domain, a hydrophobic transmembrane domain, an intracellular catalytic tyrosine kinase domain, and several intracellular tyrosine residues. The activation of EGFR leads to cellular growth, differentiation, and migration. Many tumor types are associated with dysregulation of EGFR, such as head and neck, breast, bladder, ovarian, renal, colon, and NSCLC.[74] Drugs such as Erlotinib and Gefitinib were developed to inhibit the EGFR protein, more specifically the tyrosine kinase domain.[75] However, a small deletion in exon 19 of the *EGFR* gene can lead to the deletion of four amino

acids, modifying this tyrosine kinase domain. NSCLC patients with this mutation showed Erlotinib [76] and Gefitinib [77] treatment resistance (Figure 3A), nevertheless patients with microindels have a better response to these drugs than patients with other types of mutations.[23] The alternative treatment for lung cancer patients with this *EGFR* mutation is platinum-based doublet with Pemetrexed, which inhibits three enzymes responsible for the synthesis of nucleic acids [78] and has shown a higher objective response rate in adenocarcinoma lung patients.[79] Another chemical compound that targets the encoded variant EGFRvIII with 801 base-pair deletion is called Temozolomide, and is usually used in the treatment of recurrent glioblastoma patients diagnosed with this *EGFR* mutation. The treatment consists in the use of this chemical compound in combination with radiotherapy.[72]

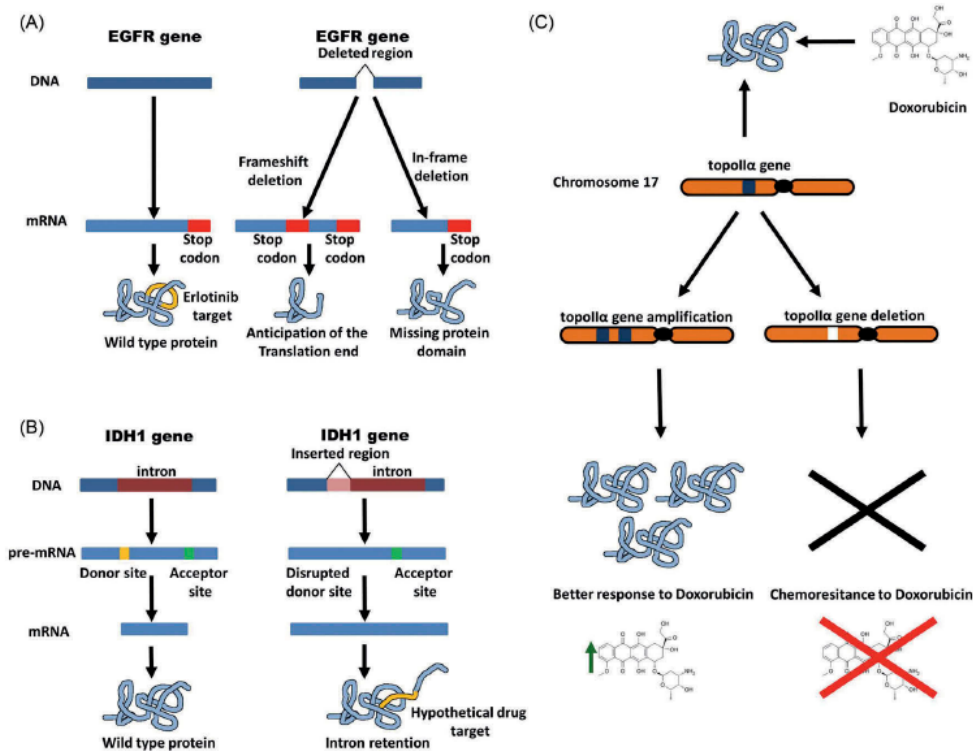
The proto-oncogene c-KIT is a cytokine receptor and drug target that plays an important role in cell proliferation and survival [80] and is inhibited by Imatinib. In an Imatinib trial with 11 melanoma patients with gained copies of the *KIT* gene, only one patient responded to the drug.[81] However, gain and loss of copies of the same gene can have opposite reactions to a cancer treatment. Doxorubicin is another well-known breast cancer drug, which targets the topoisomerase II. As consequence, the loss of copies of Topoisomerase IIa gene generates resistance to this chemical agent. However, the gain of copies of this gene, which usually occurs simultaneously with the gain of a known adjacent oncogene *HER-2*, increases the sensitivity of breast cancer tumor to this drug, leading to a better clinical response (Figure 3C).[82]

The loss of copies of the *CDKN2A* gene is frequent in some other cancers. This gene is targeted by Imatinib, unless structural variants occur. In a clinical trial with patients with dermatofibrosarcoma, a rare skin cancer,

**Table 1.** INDELS and respective genes occurring in tumors and the status of drug response.

Gene	INDEL	Cancer type	Drug response	References
<i>EGFR</i>	Deletion of exons 19 and 20	Head and neck, breast, bladder, ovarian, renal, colon, and NSCLC	Gefitinib resistance	[76]
<i>EGFR</i>	Deletion of exons 19 and 20	Head and neck, breast, bladder, ovarian, renal, colon, and NSCLC	Erlotinib resistance	[77]
<i>EGFR</i>	Deletion of 801 base pairs	Glioblastoma	Temozolomide sensitivity	[72]
<i>KIT</i>	Gained copies of the gene	Melanoma	Imatinib increased sensitivity	[81]
<i>CDKN2A</i>	Lost copies of the gene	Dermatofibrosarcoma and melanoma	Imatinib resistance	[82,83]
<i>TSC1</i>	c.1907_1908del	Bladder	Everolimus increased sensitivity	[86]
<i>PTEN</i>	Lost copies of the gene	Glioblastoma, uterine, cervical, lung squamous carcinoma, prostate, melanoma and stomach	Everolimus and Temsirolimus increased sensitivity	[88]
<i>PI3KCA</i>	Gained copies of the gene	Breast	Lapatinib and BYL719 increased sensitivity	[89]
<i>AKT2</i>	Gained copies of the gene	Breast	AKT inhibitor CCT128930 increased sensitivity	[90]
<i>HER2</i>	Gained copies of the gene	Breast	Trastuzumab increased sensitivity	[91]
<i>FGFR2</i>	Gained copies of the gene	Breast	Brivanib increased sensitivity	[92]





**Figure 3.** Different types of INDELs and respective possible drug response. (A) Possible *EGFR* gene INDELs causing frameshift or in-frame deletions, which can delete the Erlotinib targeting region and cause resistance. (B) Insertion in the *IDH1* gene causing the disruption of a splice site and forming a transcript with an intron retention, that when translated can produce a potential novel drug target. (C) Structural variants of the *topolla* gene can cause different Doxorubicin responses: gene amplification leads to a better clinical response to treatment whereas gene deletion leads to drug resistance.

patients with loss of copies of *CDKN2A* showed resistance to Imatinib treatment, but when treated with a novel compound (PD-0332991) that inhibits *CDK4* gene, the tumors shrunk.[83] This same compound was used in 47 melanoma cell lines with loss of copies of *CDKN2A* and showed similar results in reducing cell viability.[84]

INDELs can occur in genes encoding proteins that interact with known drug targets, activating or inhibiting them. For example, the deletion of two nucleotides in the *TSC1* gene in bladder cancer patients encodes a truncated protein not capable of regulating mTORC1, which is an Everolimus target. The *TSC1* gene usually encodes a protein that forms a complex with the protein encoded by the *TSC2* gene acting as a tumor suppressor inhibiting mTORC1. The former protein prevents the ubiquitination and degradation of the latter.[85] A cohort of 13 bladder patients, of whom 9 had this deletion in the *TSC1* gene, showed that sensitivity to this treatment is increased in patients with this mutation.[86]

An important pathway frequently affected by genomic alterations is phosphatidylinositol 3-kinase (PI3 K) signaling. This pathway plays an important role in cancer progression, including the regulation of cancer metabolism, survival, motility, and growth. The most common INDELs occur in the *PTEN* gene where protein truncation accumulates PIP<sub>3</sub>.<sup>[87]</sup> Some mTORC1 targeting drugs, such as Temsirolimus and Everolimus, have been used in mice with tumors with deletions in *PTEN*, with a good response.[88] Gained copies of the *PI3KCA* and *AKT2* genes, which also belong to this pathway, have been found in distinct tumors.[87] The amplification of these genes raises their sensibility to inhibitors of their encoded proteins. The combination of Lapatinib and a p110 $\alpha$ -selective PI3K inhibitor (BYL719) reduced the population of breast cancer cell line population BT474 LapR with amplifications of *PI3KCA* gene.[89] An adenosine triphosphate competitive AKT inhibitor is still in preclinical tests with tumor xenografts but already shows inhibition of the *AKT2* gene.[90]

In breast cancer tumors, it is common to find gained copies of genes related to cell growth signaling. *HER2* plays an important role in cell differentiation and survival in most breast tumors, being a stimulator of the PI3K/AKT anti-apoptosis pathway,[91] while the *FGFR2* gene is a tyrosine kinase receptor usually found with gained copies in breast cancer patients and is associated with tumorigenesis.[92] Some previously published studies have shown that the amplification of these genes may increase drug sensitivity. When amplification of the *HER2* gene is detected in breast cancer patients, the use of Trastuzumab normally improves survival. However, no more than 20% of invasive breast cancers present this amplification.[93] As an alternative, the *FGFR2* protein can be a target for patients with no *HER2* amplification. There is not a specific drug targeting *FGFR2* protein, but there is evidence that VEGFR inhibitors, such as Brivanib, can also inhibit its activity.[94] As presented here, many studies have associated INDEL detection to drug response. However, none of them used the RNA-Seq strategy to investigate those INDELs with potential to change protein sequence and clarify whether the allele is in fact expressed. Therefore, besides INDEL detection, RNA-Seq may assist to identify messenger RNA (mRNA) molecules that are expressed and provide valuable information for the development of new therapeutic targets.

#### 4. Challenges for drug discovery targeting proteins affected by INDELs

The discovery of new drugs targeting proteins affected by INDELs is a motivating field to be explored. Although INDEL detection is being used as a therapeutic target in other diseases, such as cystic fibrosis [95] and genetic skeletal diseases,[96] it has not been used to search for novel cancer therapeutic targets. The use of HTS to identify INDELs is an opportunity to find novel drug targets in cases where protein domains have been modified. RNA-Seq provides a further advantage, the possibility to identify INDELs that are significantly and differentially expressed between clinical conditions. However, to obtain tumor samples is more complicated than obtaining samples from patients with other diseases. While in such diseases it is possible to obtain a large number of samples using noninvasive methods (e.g. blood), in cancer research there is always the limitation of access to patient tumor samples. For this reason, the identification of INDELs based in cell lines is widely used to obtain the genetic portrait. For example, 675 cancer cell lines were studied to this end.[73] Nonetheless, the use of patient samples is crucial to gain an accurate picture

of this research field. There are many studies using HTS that identify INDELs from patients' samples, including those working with RNA-Seq. As an example, a cohort of 5 prostate patients was used to identify 116 mutations, including frameshift INDELs in 92 genes already associated with cancer.[58] In addition, more patient cohorts were used to identify INDELs using RNA-seq, such as cohorts of patients with ovarian carcinoma,[59] breast cancer,[60] gastric cancer,[65] osteosarcoma,[67] acute lymphoblastic leukemia,[70] and glioblastoma.[71]

Deletions in the *TP53* gene also modify its DNA binding domain, generating a truncated protein that lacks repairing ability, which reduces its detection in many cancer cells.[97] The commonly used chemotherapy, radiation, and other treatments that focus on DNA damage do not work properly in tumors with mutated p53 protein, conferring a poor chemo-sensitivity profile.[98] Some efforts to reactivate p53 using molecules to help reestablish its DNA binding ability have been done to confer stronger chemo-sensitivity. These efforts included the use of a synthetic peptide to restore a malformed domain [99] or the use of two compounds from the thiosemicarbazone family, which reactivate p53 function.[100]

Oncogenes affected by INDELs can escape the effects of chemotherapy. Thus, the development of molecules that help cancer cells carrying mutated oncogenes to become more sensitive to other drugs is also a good strategy to fight cancer. Drug development targeting the *KRAS* oncogene, for example, is still challenging and only a few efforts using different types of molecules have been used, such as miRNA [101] and iRNA.[102] Additional examples of intron retention events associated with INDELs in splice sites identified by RNA-Seq were found by Dvinge and Bradley [103] in the *IDH1* gene in AML samples. These intron retention mRNAs can produce novel amino acids, which can be used as drug targets (Figure 3B).

The strategy of RNA-Seq does sequencing of transcripts that are expected to have their introns spliced out. Therefore, some of the missing nucleotides might have resulted from an alternative splice event instead of a genomic loss. One example is exons with less than 51 nucleotides (micro exons) that are skipped in mature transcripts [104] and therefore are hard to correctly differentiate from INDELs. Another study detected regions captured by exome sequencing that presented low coverage in RNA-Seq; these regions were coined as 'exintron'. [105] These examples reveal the need to discuss and establish new parameters to establish whether a region represents genomic change or an alternative splicing event.



Although HTS sequencing platforms generate a large amount of data, there are many errors and not all potential INDELS are real.[106,107] A number of algorithms have been designed to reduce false positives by enabling gapped alignment while doing genome mapping.[108] Another challenge is the fact that many INDELS also occur in repeated regions of the genome, where they are difficult to be detected by bioinformatics tools.[109]

## 5. Conclusion

Several studies have already associated different INDEL types to tumors and indicated differential responses to drug treatment. HTS has proved to be a powerful tool for the identification of mutations and genomic alterations in tumor samples. The standard approach to INDEL identification is to use DNA as the source of data, but cancer genomic sequencing data are still rare. Some studies have already showed that the use of transcriptome data to identify INDEL candidates is possible, supporting the wide use of RNA-Seq. This strategy can improve our understanding of carcinogenesis and, therefore, the development of new drugs targeting mutated proteins or targeting proteins that interact with them.

In this review, we show that several studies have already used RNA-Seq as the source of data for INDEL identification in cancer, and that there are many cases of INDEL being applied to clinical practice. Therefore, INDEL is an emerging and promising methodological approach to be used in drug discovery.

## 6. Expert opinion

During the last 20 years, several technologies in genomics have emerged. In the first years, several groups used Sanger sequencing to search for cancer mutations. After that, microarrays were widely used to investigate for a list of known sequence polymorphisms. More recently, HTS was introduced as a technology to generate large amounts of data in an unprecedented volume and time. Several research groups moved their focus to use this technology, aiming to discover differentially expressed genes in cancer. However, apart from the identification of SNPs, INDELS have not been deeply investigated using transcriptome data. We present in this review article a selected list of research articles that: (1) show that RNA-Seq data can be used as a reliable source of data to search for INDEL; (2) INDELS are frequent in tumor samples; and (3) there are clinical cases in which it is possible to associate distinct types of INDELS to patient outcome. For example, INDELS

identified in a significant mutated gene (*TP53*) in colorectal cancer were associated with patient survival.[110] Therefore, a massive investigation of publicly available cancer RNA-Seq datasets is urgent to search for INDELS, which can be rapidly tested in clinical practice and used to improve personalized medicine approaches. These identified INDELS can cause important alterations in the encoded proteins such as inserting or deleting amino acids, truncating the encoded protein, and creating or disrupting splice sites. This clearly indicates that INDELS need more attention when developing new drug targets focused on rare tumor types. The RNA-Seq platform is a suitable methodology to generate large amounts of data to serve this purpose. There is twice as much RNA sequencing data as DNA-Seq. One challenge is to obtain enough RNA sequencing coverage in order to call a Sequence Nucleotide Variation (SNV). To be conservative, there is an expectation of having 30× nucleotide coverage for strong statistical confidence in calling an SNV. However, gene expression studies need to have enough sequence coverage to call differentially expressed genes. For most low-expressed genes, sequencing coverage is low and around 5×. One approach to solve this issue is to merge sequencing reads from a given tissue or pathological state to increase the statistical power of calling INDELS.

Although all genomes of the human species are very similar, there are still variations that may go unnoticed when the reference human sequence from the GRC is used for polymorphism detection. This type of approach can limit the identification of INDELS to differentiate healthy from disease states. To improve the knowledge of variants in the human genome, the 1000G was deployed. The use of 1000G data can also assist to establish whether specific INDELS can be considered as a normal genetic variant or a cancer-unique genetic variant. Yet, this field needs more studies with different cancer types to enable the identification of novel INDELS and molecular markers. A challenge of using genomes from projects like the 1000G is in having sufficient computational power to compare the obtained sequencing data against more than 10,000 distinct complete human genome sequences. To reduce time to compute such data, the 1000G provides a variant call format file with the annotation of sequence variations of each genome sequence when compared to the latest release of the human reference genome assembly from the GRC.


However, to determine INDEL candidates with higher confidence after RNA-Seq analysis, the acquisition of paired DNA and RNA samples from the same patient is important, besides the development of novel approaches in bioinformatics. Clinicians,

pathologists, molecular biologists, and bioinformaticians may work together to obtain samples with high-quality and develop software with higher accuracy. Moreover, other strategies, such as proteogenomics,<sup>[111,112]</sup> may be used to transform INDEL identification from RNA-Seq data to predicted translations to search for high-throughput proteomics data, such as the work done by Tavares and colleagues <sup>[113]</sup> for splice variants. This strategy may include an additional step of confidence after INDEL candidates are subjected to clinical validation and moved forward to drug discovery pipelines.

### Financial & Competing Interests Disclosure

Both authors are supported by the Fundação Oswaldo Cruz (FIOCRUZ). F Passetti is supported by the Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ) and the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) while G Wajnberg is supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES). The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

### ORCID

Gabriel Wajnberg  <http://orcid.org/0000-0002-4479-4063>  
Fabio Passetti  <http://orcid.org/0000-0001-5672-7848>

### References

- Mills RE, Luttig CT, Larkins CE, et al. An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res.* 2006 Sep;16(9):1182–1190.
- One of the first efforts to identify only IDELS using the human genome DNA sequencing.
- Bhangale TR, Rieder MJ, Livingston RJ, et al. Comprehensive identification and characterization of diallelic insertion-deletion polymorphisms in 330 human candidate genes. *Hum Mol Genet.* 2005 Jan 1;14(1):59–69.
- Scaringe WA, Li K, Gu D, et al. Somatic microindels in human cancer: the insertions are highly error-prone and derive from nearby but not adjacent sense and antisense templates. *Hum Mol Genet.* 2008 Sep 15;17(18):2910–2918.
- Mullaney JM, Mills RE, Pittard WS, et al. Small insertions and deletions (INDELs) in human genomes. *Hum Mol Genet.* 2010 Oct 15;19(R2):R131–136.
- Review about INDELs found in the human genome and their importance in genomics.
- Montgomery SB, Goode DL, Kvikstad E, et al. The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. *Genome Res.* 2013 May;23(5):749–761.
- Ellegren H. Microsatellites: simple sequences with complex evolution. *Nat Reviews Genet.* 2004 Jun;5(6):435–445.
- de La Chaux N, Messer PW, Arndt PF. DNA indels in coding regions reveal selective constraints on protein evolution in the human lineage. *BMC Evol Biol.* 2007;7:191.
- Stenson PD, Mort M, Ball EV, et al. The human gene mutation database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum Genet.* 2014 Jan;133(1):1–9.
- Collins FS, Drumm ML, Cole JL, et al. Construction of a general human chromosome jumping library, with application to cystic fibrosis. *Science.* 1987 Feb 27;235(4792):1046–1049.
- Warren ST, Zhang F, Licameli GR, et al. The fragile X site in somatic cell hybrids: an approach for molecular cloning of fragile sites. *Science.* 1987 Jul 24;237(4813):420–423.
- Zhang X, Lin H, Zhao H, et al. Impact of human pathogenic micro-insertions and micro-deletions on post-transcriptional regulation. *Hum Mol Genet.* 2014 Jun 1;23(11):3024–3034.
- Bhattacharya A, Ziebarth JD, Cui Y. Systematic analysis of microRNA targeting impacted by small insertions and deletions in human genome. *Plos One.* 2012;7(9):e46176.
- Chen CH, Liao BY, Chen FC. Exploring the selective constraint on the sizes of insertions and deletions in 5' untranslated regions in mammals. *BMC Evol Biol.* 2011;11:192.
- Review of how mutations, inclusive INDELs, are distributed in different types o tumors.
- Glanzmann B, Lombard D, Carr J, et al. Screening of two indel polymorphisms in the 5'UTR of the DJ-1 gene in South African Parkinson's disease patients. *J Neural Transm.* 2014 Feb;121(2):135–138.
- Greenman C, Stephens P, Smith R, et al. Patterns of somatic mutation in human cancer genomes. *Nature.* 2007 Mar 8;446:153–158.
- Stratton MRS, Peter JC, Futreal PA. The cancer genome. *Nature.* 2009;458:719–724.
- Gemignani F, Moreno V, Landi S, et al. A TP53 polymorphism is associated with increased risk of colorectal cancer and with reduced levels of TP53 mRNA. *Oncogene.* 2004 Mar 11;23(10):1954–1956.
- Tao R, Hu S, Wang S, et al. Association between indel polymorphism in the promoter region of lncRNA GASS and the risk of hepatocellular carcinoma. *Carcinogenesis.* 2015 Oct;36(10):1136–1143.
- Boland CR, Goel A. Microsatellite instability in colorectal cancer. *Gastroenterology.* 2010 Jun;138(6):2073–2087.
- Williams DS, Bird MJ, Jorissen RN, et al. Nonsense mediated decay resistant mutations are a source of expressed mutant proteins in colon cancer cell lines with microsatellite instability. *Plos One.* 2010;5(12):e16012.
- Yuan Z, Shin J, Wilson A, et al. An A13 repeat within the 3'-untranslated region of epidermal growth factor receptor (EGFR) is frequently mutated in microsatellite instability colon cancers and is associated with increased EGFR expression. *Cancer Res.* 2009 Oct 1;69(19):7811–7818.



22. Forbes SA, Bhamra G, Bamford S, et al. The catalogue of somatic mutations in cancer (COSMIC). 04/23th ed. In: Haines JL, et al., editors. *Current protocols in human genetics/editorial board*. Hoboken: John Wiley & Sons 2008 Apr;Chapter 10:Unit 10.11.
23. Gu D, Scaringe WA, Li K, et al. Database of somatic mutations in EGFR with analyses revealing indel hotspots but no smoking-associated signature. *Hum Mutat.* 2007 Aug;28(8):760–770.
24. Vali U, Brandstrom M, Johansson M, et al. Insertion-deletion polymorphisms (indels) as genetic markers in natural populations. *BMC Genetics.* 2008;9:8.
25. Chen FC, Chen CJ, Li WH, et al. Human-specific insertions and deletions inferred from mammalian genome sequences. *Genome Res.* 2007 Jan;17(1):16–22.
26. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature.* 2001 Feb 15;409:860–921.
27. Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. *Science.* 2001 Feb 16;291:1304–1351.
28. Church DM, Schneider VA, Graves T, et al. Modernizing reference genome assemblies. *PLoS Biol.* 2011 Jul;9(7):e1001091.
29. E pluribus unum. *Nat Methods [Internet].* 2010 May;7(5):331 [cited 2015 Dec 18]. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/20440876>
30. Pinkel D, Seagraves R, Sudar D, et al. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat Genet.* 1998;20:207–211.
31. Solinas-Toldo S, Lampel S, Stilgenbauer S, et al. Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosomes Cancer.* 1997;20:399–407.
32. Rauch A, Ruschendorf F, Huang J, et al. Molecular karyotyping using an SNP array for genomewide genotyping. *J Med Genet.* 2004;41:916–922.
33. Radtke I, Mullighan CG, Ishii M, et al. Genomic analysis reveals few genetic alterations in pediatric acute myeloid leukemia. *Proc Natl Acad Sci U S A.* 2009;106:12944–12949.
34. Wajnberg G, Carvalho BS, Ferreira CG, et al. Combined analysis of SNP array data identifies novel CNV candidates and pathways in ependymoma and mesothelioma. *Biomed Res Int.* 2015;2015:1–11.
  - **Identifies new CNVs in published data of Ependymoma and Mesothelioma of SNP-array data.**
35. Pareek CS, Smoczynski R, Tretyn A. Sequencing technologies and genome sequencing. *J Appl Genet.* 2011 Nov;52(4):413–435.
36. Levy S, Sutton G, Ng PC, et al. The diploid genome sequence of an individual human. *PLoS Biol.* 2007 Sep 4;5:e254.
37. Wheeler DA, Srinivasan M, Egholm M, et al. The complete genome of an individual by massively parallel DNA sequencing. *Nature.* 2008 Apr 17;452:872–876.
38. Koboldt DC, Chen K, Wylie T, et al. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics.* 2009 Sep 1;25(17):2283–2285.
39. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet.* 2011 May;43(5):491–498.
40. Ye K, Schulz MH, Long Q, et al. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics.* 2009 Nov 1;25(21):2865–2871.
41. Zhao H, Zhao F. BreakSeek: a breakpoint-based algorithm for full spectral range INDEL detection. *Nucleic Acids Res.* 2015;43(14):6701–6713.
42. Ratan A, Olson TL, Loughran TP Jr, et al. Identification of indels in next-generation sequencing data. *BMC Bioinformatics.* 2015;16:42.
43. Consortium TGP. An integrated map of genetic variation from 1,092 human genomes. *Nature.* 2012;491:56–65.
44. Challis D, Antunes L, Garrison E, et al. The distribution and mutagenesis of short coding INDELs from 1,128 whole exomes. *BMC Genomics.* 2015;16:143.
45. Turner EH, Ng SB, Nickerson DA, et al. Methods for genomic partitioning. *Annu Rev Genomics Hum Genet.* 2009;10:263–284.
46. Vaksman Z, Fonville NC, Tae H, et al. Exome-wide somatic microsatellite variation is altered in cells with DNA repair deficiencies. *Plos One.* 2014;9(11):e110263.
47. Shah SP, Kobel M, Senz J, et al. Mutation of FOXL2 in granulosa-cell tumors of the ovary. *N Engl J Med.* 2009 Jun 25;360:2719–2729.
48. Pleasance ED, Stephens PJ, O'Meara S, et al. A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature.* 2009 Jan 14;463:184–190.
49. Pleasance ED, Cheetham RK, Stephens PJ, et al. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature.* 2009 Jan 14;463:191–196.
50. Ding L, Ley TJ, Larson DE, et al. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. *Nature.* 2012 Jan 26;481(7382):506–510.
51. Morrison CD, Liu P, Woloszynska-Read A, et al. Whole-genome sequencing identifies genomic heterogeneity at a nucleotide and chromosomal level in bladder cancer. *Proc Natl Acad Sci U S A.* 2014 Feb 11;111(6):E672–E681.
52. Kandoth C, McLellan MD, Vandin F, et al. Mutational landscape and significance across 12 major cancer types. *Nature.* 2013 Oct 17;502(7471):333–339.
53. Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, et al. The cancer genome atlas pan-cancer analysis project. *Nat Genet.* 2013 Oct;45(10):1113–1120.
  - **Construction of The Cancer Genome Atlas with high-throughput sequencing raw and processed data.**
54. Kim TM, Laird PW, Park PJ. The landscape of microsatellite instability in colorectal and endometrial cancer genomes. *Cell.* 2013 Nov 7;155(4):858–868.
55. Nielsen R, Joshua SP, Anders A, et al. Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet.* 2011;12:443–451.
  - **Explain how to identify mutations in RNA-seq platform.**
56. Kim TM, Park PJ. A genome-wide view of microsatellite instability: old stories of cancer mutations revisited with

- new sequencing technologies. *Cancer Res.* 2014 Nov 15;74(22):6377–6382.
57. Cirulli ET, Singh A, Shianna KV, et al. Screening the human exome: a comparison of whole genome and whole transcriptome sequencing. *Genome Biol.* 2010;11(5):R57.
  58. Xu X, Zhu K, Liu F, et al. Identification of somatic mutations in human prostate cancer by RNA-seq. *Gene* 2013 May 1;519(2):343–347.
    - **Revealed important INDELS in prostate cancer using RNA-seq data.**
  59. Wiegand KC, Shah SP, Al-Agha OM, et al. ARID1A mutations in endometriosis-associated ovarian carcinomas. *N Engl J Med.* 2010 Oct 14;363:1532–1543.
    - **Used exome sequencing and RNA-seq together to identify INDELS and gene fusions in ovarian carcinomas.**
  60. Horvath A, Pakala SB, Mudvari P, et al. Novel insights into breast cancer genetic variance through RNA sequencing. *Sci Rep.* 2013;3:2256.
  61. Hudis CA, Gianni L. Triple-negative breast cancer: an unmet medical need. *Oncologist.* 2011;16(Suppl 1):1–11.
  62. Ricarte-Filho JC, Li S, Garcia-Rendueles ME, et al. Identification of kinase fusion oncogenes in post-chemotherapy radiation-induced thyroid cancers. *J Clin Invest.* 2013;123:4935–4944.
  63. Schneeweiss A, Sinn HP, Ehemann V, et al. Centrosomal aberrations in primary invasive breast cancer are associated with nodal status and hormone receptor expression. *Int J Cancer.* 2003 Nov 10;107:346–352.
  64. Schallreuter KU, Wood JM. New aspects in the pathophysiology of cutaneous melanoma: a review of the role of thioproteins and the effect of nitrosoureas. *Melanoma Res.* 1991 Aug–Sep;1:159–167.
  65. Liu J, McClelland M, Stawiski EW, et al. Integrated exome and transcriptome sequencing reveals ZAK isoform usage in gastric cancer. *Nat Commun.* 2014;5:3830.
  66. Ahn YH, Yang Y, Gibbons DL, et al. Map2k4 functions as a tumor suppressor in lung adenocarcinoma and inhibits tumor cell invasion by decreasing peroxisome proliferator-activated receptor  $\gamma$ 2 expression. *Mol Cell Biol.* 2011;31:4270–4285.
  67. Reimann E, Koks S, Ho XD, et al. Whole exome sequencing of a single osteosarcoma case—integrative analysis with whole transcriptome RNA-seq data. *Hum Genomics.* 2014;8:20.
  68. Choy E, Hornicek F, MacConaill L, et al. High-throughput genotyping in osteosarcoma identifies multiple mutations in phosphoinositide-3-kinase and other oncogenes. *Cancer.* 2011 Jun 1;118:2905–2914.
  69. Pant V, Jambhekar NA, Madur B, et al. Anaplastic large cell lymphoma (ALCL) presenting as primary bone and soft tissue sarcoma—a study of 12 cases. *Indian J Pathol Microbiol.* 2007;50:303–307.
  70. Atak ZK, Gianfelici V, Hulseimans G, et al. Comprehensive analysis of transcriptome variation uncovers known and novel driver events in T-cell acute lymphoblastic leukemia. *PLoS Genet.* 2013;9(12):e1003997.
  71. Patel AP, Tirosch I, Trombetta JJ, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science.* 2014 Jun 20;344:1396–1401.
    - **Describes an innovative methodology to identify CNVs using RNA-seq data from Glioblastoma.**
  72. Gan HK, Cvriljevic AN, Johns TG. The epidermal growth factor receptor variant III (EGFRvIII): where wild things are altered. *Febs J.* 2013 Nov;280(21):5350–5370.
    - **Review dissecting the importance of the mutant EGFRvIII to the carcinogenesis process.**
  73. Klijn C, Durinck S, Stawiski EW, et al. A comprehensive transcriptional portrait of human cancer cell lines. *Nature Biotechnology.* 2015;33(3):306–312.
    - **A pan cancer effort using RNA-seq and hundreds of cancer cell lines to identify different types of mutations.**
  74. Yewale C, Baradia D, Vhora I, et al. Epidermal growth factor receptor targeting in cancer: a review of trends and strategies. *Biomaterials.* 2013;34(34):8690–8707.
    - **Reviews the use of EGFR protein as a drug target.**
  75. Raymond E, Faivre S, Armand JP. Epidermal growth factor receptor tyrosine kinase as a target for anticancer therapy. *Drugs.* 2000;60(Suppl 1):15–23; discussion 41–2.
  76. Hirai F, Takenoyama M, Taguchi K, et al. Experience with erlotinib in lung adenocarcinoma harboring a coexisting KIF5B-RET fusion gene and EGFR mutation: report of a rare case. *J Thorac Oncol.* 2014;9:e37–39.
  77. Lynch TJ, Bell DW, Sordella R, et al. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *N Engl J Med.* 2004 May 20;350:2129–2139.
  78. McLeod HL, Cassidy J, Powrie RH, et al. Pharmacokinetic and pharmacodynamic evaluation of the glycinamide ribonucleotide formyltransferase inhibitor AG2034. *Clin Cancer Res.* 2000;6:2677–2684.
  79. Fang S, Wang Z, Guo J, et al. Correlation between EGFR mutation status and response to first-line platinum-based chemotherapy in patients with advanced non-small cell lung cancer. *Onco Targets Ther.* 2014;7:1185–1193.
  80. Edling CE, Hallberg B. c-kit—a hematopoietic cell essential receptor tyrosine kinase. *Int J Biochem Cell Biol.* 2007;39(11):1995–1998.
  81. Hodi FS, Corless CL, Giobbie-Hurder A, et al. Imatinib for melanomas harboring mutationally activated or amplified KIT arising on mucosal, acral, and chronically sun-damaged skin. *J Clin Oncol.* 2013 Sep 10;31:3182–3190.
  82. Jarvinen TA, Tanner M, Rantanen V, et al. Amplification and deletion of topoisomerase II $\alpha$  associate with ErbB-2 amplification and affect sensitivity to topoisomerase II inhibitor doxorubicin in breast cancer. *Am J Pathol.* 2000 Mar;156(3):839–847.
  83. Eilers G, Czaplinski JT, Mayeda M, et al. CDKN2A/p16 loss implicates CDK4 as a therapeutic target in imatinib-resistant dermatofibrosarcoma protuberans. *Mol Cancer Ther.* 2015 Jun;14(6):1346–1353.
  84. Young RJ, Waldeck K, Martin C, et al. Loss of CDKN2A expression is a frequent event in primary invasive melanoma and correlates with sensitivity to the CDK4/6 inhibitor PD0332991 in melanoma cell lines. *Pigment Cell Melanoma Res.* 2014 Jul;27(4):590–600.



85. Huang J, Manning BD. The TSC1-TSC2 complex: a molecular switchboard controlling cell growth. *Biochem J*. 2008 Jun 1;412(2):179–190.
86. Iyer G, Hanrahan AJ, Milowsky MI, et al. Genome sequencing identifies a basis for everolimus sensitivity. *Science*. 2012 Oct 12;338:221.
87. Courtney KD, Corcoran RB, Engelman JA. The PI3K pathway as drug target in human cancer. *J Clin Oncol*. 2010 Feb 20;28(6):1075–1083.
88. Dillon LM, Miller TW. Therapeutic targeting of cancers with loss of PTEN function. *Current Drug Targets*. 2014 Jan;15(1):65–79.
89. Zhang S, Huang WC, Li P, et al. Combating trastuzumab resistance by targeting SRC, a common node downstream of multiple resistance pathways. *Nat Med*. 2011 Apr;17(4):461–469.
90. Yap TA, Walton MI, Hunter LJ, et al. Preclinical pharmacology, antitumor activity, and development of pharmacodynamic markers for the novel, potent AKT inhibitor CCT128930. *Mol Cancer Ther*. 2011 Feb;10(2):360–371.
91. Gutierrez C, Schiff R. HER2: biology, detection, and clinical implications. *Arch Pathol Lab Med*. 2011 Jan;135(1):55–62.
92. Raskin L, Pinchev M, Arad C, et al. FGFR2 is a breast cancer susceptibility gene in Jewish and Arab Israeli populations. *Cancer Epidemiol Biomarkers Prev*. 2008 May;17(5):1060–1065.
93. Gajria D, Chandralapaty S. HER2-amplified breast cancer: mechanisms of trastuzumab resistance and novel targeted therapies. *Expert Rev Anticancer Ther*. 2011 Feb;11(2):263–275.
94. Jain VK, Turner NC. Challenges and opportunities in the targeting of fibroblast growth factor receptors in breast cancer. *Breast Cancer Res*. 2012;14:208.
95. Dorfman R. Modifier gene studies to identify new therapeutic targets in cystic fibrosis. *Curr Pharm Des*. 2012;18(5):674–682.
96. Briggs MD, Bell PA, Wright MJ, et al. New therapeutic targets in rare genetic skeletal diseases. *Expert Opin Orphan Drugs*. 2015;3(10):1137–1154.
97. Pavletich NP, Chambers KA, Pabo CO. The DNA-binding domain of p53 contains the four conserved regions and the major mutation hot spots. *Genes Dev*. 1993 Dec;7:2556–2564.
98. Koike M, Fujita F, Komori K, et al. Dependence of chemotherapy response on p53 mutation status in a panel of human cancer lines maintained in nude mice. *Cancer Sci*. 2004;95:541–546.
99. Selivanova G, Wiman KG. Reactivation of mutant p53: molecular mechanisms and therapeutic potential. *Oncogene*. 2007 Apr 2;26:2243–2254.
100. Yu X, Vazquez A, Levine AJ, et al. Allele-specific p53 mutant reactivation. *Cancer Cell*. 2012 May 15;21:614–625.
101. Dai X, Jiang Y, Tan C, et al. Let-7 sensitizes KRAS mutant tumor cells to chemotherapy. *Plos One*. 2015;10:e0126653.
102. Gysin S, Salt M, Young A, et al. therapeutic strategies for targeting ras proteins. *Genes Cancer*. 2011;2(3):359–372.
103. Dvinge H, Bradley RK. Widespread intron retention diversifies most cancer transcriptomes. *Genome Med*. 2015;7(1):45.
104. Li Yi, Sanchez-Pulido L, Haerty W, et al. RBFOX and PTBP1 proteins regulate the alternative splicing of micro-exons in human brain transcripts. *Genome Res*. 2015 Jan;25(1):1–13.
  - **Explains what types of problems someone can face using variant calling algorithms.**
105. Marquez Y, Höpfler M, Ayatollahi Z, et al. Unmasking alternative splicing inside protein-coding exons defines exons and their role in proteome plasticity. *Genome Res*. 2015 Jul;25(7):995–1007.
106. Le H-S, Schulz MH, McCauley BM, et al. Probabilistic error correction for RNA sequencing. *Nucleic Acids Res*. 2013;41:e109.
107. Li H. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*. 2014 Oct 15;30(20):2843–2851.
108. Krawitz P, Rodelsperger C, Jager M, et al. Microindel detection in short-read sequence data. *Bioinformatics*. 2010 Mar 15;26(6):722–729.
109. Jiang Y, Turinsky AL, Brudno M. The missing indels: an estimate of indel variation in a human genome and analysis of factors that impede detection. *Nucleic Acids Res*. 2015;43(15):7217–7228.
110. Yu J, Wu WK, Li X, et al. Novel recurrently mutated genes and a prognostic mutation signature in colorectal cancer. *Gut*. 2015 Apr;64(4):636–645.
111. Nesvizhskii AI. Proteogenomics: concepts, applications and computational strategies. *Nat Methods*. 2014 Nov;11(11):1114–1125.
112. Tavares R, Scherer NM, Ferreira CG, et al. Splice variants in the proteome: a promising and challenging field to targeted drug discovery. *Drug Discov Today*. 2015 Mar;20(3):353–360.
113. Tavares R, De Miranda Scherer N, Pauletti BA, et al. SpliceProt: a protein sequence repository of predicted human splice variants. *Proteomics*. 2014 Feb;14(2–3):181–185.