

**Universidade de São Paulo**  
**Faculdade de Medicina de Ribeirão Preto**  
**Programa de Pós Graduação em Genética**

**Érico Torrieri**

Abordagem computacional para avaliar o impacto da metilação do DNA na instabilidade  
Tumoral

**Ribeirão Preto - SP**

**2020**

**Universidade de São Paulo**  
**Faculdade de Medicina de Ribeirão Preto**  
**Programa de Pós Graduação em Genética**

**ÉRICO TORRIERI**

Abordagem computacional para avaliar o impacto da metilação do DNA na instabilidade tumoral

**Versão corrigida. A versão original encontra-se disponível tanto na Biblioteca da Unidade que aloja o Programa, quanto na Biblioteca Digital de Teses e Dissertações da USP (BDTD)”**

**Orientador: Prof. Dr. Wilson Araujo Silva Junior.**

**Ribeirão Preto - SP**

**2020**

**Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.**

**Torrieri, Érico**

**Abordagem computacional para avaliar o impacto da metilação do DNA na instabilidade tumoral. Ribeirão Preto, 2020.**

**124 p. : il. ; 30 cm**

**Tese de Doutorado, apresentada à Faculdade de Medicina de Ribeirão Preto/USP. Área de concentração: Genética. Orientador: Araujo Junior, Wilson.**

1. Desaminação 2. Metilação 3. Mutação 4. Instabilidade Tumoral 5. Citosinas

**FOLHA DE APROVAÇÃO**

**Torrieri, E. Abordagem computacional para avaliar o impacto da metilação do DNA na instabilidade tumoral. Tese apresentada à Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo, para obtenção do título de Doutor em Ciências - Área de concentração: Genética**

**Aprovada em:**

**Banca examinadora**

**Prof. Dr. \_\_\_\_\_ Instituição: \_\_\_\_\_**

**Julgamento: \_\_\_\_\_ Assinatura: \_\_\_\_\_**

**Prof. Dr. \_\_\_\_\_ Instituição: \_\_\_\_\_**

**Julgamento: \_\_\_\_\_ Assinatura: \_\_\_\_\_**

**Prof. Dr. \_\_\_\_\_ Instituição: \_\_\_\_\_**

**Julgamento: \_\_\_\_\_ Assinatura: \_\_\_\_\_**

**Prof. Dr. \_\_\_\_\_ Instituição: \_\_\_\_\_**

**Julgamento: \_\_\_\_\_ Assinatura: \_\_\_\_\_**

## **APOIO E SUPORTE FINANCEIRO**

Trabalho desenvolvido no laboratório de Genética Molecular e Bioinformática, do Hemocentro de Ribeirão Preto, da Universidade de São Paulo (USP). Essa tese recebeu o apoio financeiro das seguintes agências de fomento e instituições:

- Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP)
- Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)
- Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES)
- Programa de Pós-Graduação em Genética
- Faculdade de Medicina de Ribeirão Preto - USP.

## **Dedicatória**

*Dedico esta Tese à minha esposa Priscilla Aparecida Tartari Pereira e ao meu filho Artur Tartari Torrieri.*

*“Sempre se utilizam das menores ensanchas para inculir as idéias felizes e o bem operante. Vigilantes, são o apoio no desfalecimento, a coragem no receio e a força no momento do desânimo.”*

*Joanna de Angelis (Momento de Coragem – Divaldo Franco)*

## **Agradecimentos**

Agradeço primeiramente a Deus por tudo.

Agradeço ao meu orientador Wilson Araujo Silva Junior pelo aprendizado.

Agradeço aos meus colegas de laboratório, pelo companheirismo e aprendizado que me proporcionaram.

Agradeço aos meus familiares pelo apoio.

Agradeço a minha esposa Priscilla Aparecida Tartari Pereira pelo apoio e amor incondicional.

Agradeço ao meu filho Artur Tartari Torrieri por me fazer sentir um amor incondicional e alegrias todos os meus dias.

A secretária do programa de pós-graduação em genética Susie Nalon que sempre esteve pronta a me ajudar, aconselhar e acalmar. Meu muito obrigado!

## RESUMO

Torrieri, E. **Abordagem computacional para avaliar o impacto da metilação do DNA na instabilidade tumoral.** 124f. Tese (Doutorado). Faculdade de Medicina de Ribeirão Preto - Universidade de São Paulo, Ribeirão Preto, 2020.

**Palavras-Chaves:** Desaminação, metilação, mutação, instabilidade tumoral, citosinas.

A metilação do DNA é um mecanismo epigenético importante de regulação de expressão gênica, podendo desencadear diversos tipos de doenças complexas, como o câncer, diabetes, obesidade. A metilação do DNA ocorre em maior abundância sobre as bases de citosina, sendo sua maioria em dinucleotídeos CG (sítios CpG). As estimativas sugerem que 35% das mutações pontuais causadoras de distúrbios genéticos humanos ocorram nos dinucleotídeos CpG, e mais de 90% destes, são transições de C > T ou correspondentes transições G > A. A metilação de CpGs em tecidos normais pode aumentar a probabilidade de mutações nesses locais devido à capacidade da 5-metilcitosina sofrer desaminação, resultando na troca de citosina por timina. Em tumores esse evento, pode ser melhor observado já que regiões de ilhas CpG costumam estar abundantemente metiladas. Este trabalho tem como objetivo encontrar trocas C > T em sítios CpG metilados, inferindo que essas trocas ocorreram devido ao mecanismo de desaminação, para tal, foram usados dados de mutação e



metilação do projeto TCGA. Para dados de exoma, foi encontrado uma taxa de 37,67% de trocas C > T, enquanto que a segunda troca com maior probabilidade é, C > A com 24,18%. Quando olhamos somente para os sítios CpG essa porcentagem sobe chegando a 75,23% de trocas C > T e quando olhamos somente para trocas em sítios CpG metilados observamos que 76,98% são de trocas C > T. Dessas trocas 66,14% levam a mutações missense e nonsense e dessas 65,98% são mutações patogênicas. Esses resultados sugerem que mutações em citosinas metiladas contribuem para progressão da tumorigênese.

## ABSTRACT

Torrieri, E. **Computational approach to assess the impact of DNA methylation on tumor instability.** 124f. Thesis (Doctor degree). Faculty of Medicine of Ribeirão Preto - University of São Paulo, Ribeirão Preto, 2020.

**Key-words:** Deamination, methylation, mutation, tumor instability, cytosine.

DNA methylation is an important epigenetic mechanism for regulating gene expression, which can trigger several types of complex diseases, such as cancer, diabetes, obesity, etc. DNA methylation occurs in greater abundance over cytosine bases, most of them in CG dinucleotides (CpG sites). Estimates suggest that 35% of point mutations causing human genetic disorders occur in CpG dinucleotides, and more than 90% of these are transitions from C > T or corresponding G > A transitions. Methylation of CpGs in normal tissues may increase the likelihood of mutations at these sites due to the ability of 5-methylcytosine to deaminate, resulting in the exchange of cytosine for thymine. In tumors, this event can be better observed since regions of CpG islands are usually abundantly methylated. This work aims to find C > T exchanges at methylated CpG sites, inferring that these exchanges occurred due to the deamination mechanism, for this purpose, TCGA project mutation and methylation data were used. For exome data, a rate of 37.67% of C > T exchanges was found, while the second most likely exchange is C to A with 24.18%. When we look only at CpG sites this percentage goes up to 75.23% of C > T exchanges and when we look only at exchanges at methylated CpG

sites we see that 76.98% are C > T exchanges. and of those 65.98% are pathogenic mutations. These results suggest that mutations in methylated cytosines contribute to the progression of tumorigenesis.

## Lista de Figuras

<b>Figura 1.</b> Possibilidades de mutações, transições e transversões.....	23
<b>Figura 2.</b> Conversão das bases nitrogenadas ao sofrerem o mecanismo de desaminação.....	24
<b>Figura 3.</b> Mecanismo de desaminação em citocinas metiladas.....	25
<b>Figura 4.</b> Demonstração do aumento de metilação em sítios CpG de acordo com a progressão tumoral.....	27
<b>Figura 5.</b> Fluxograma do projeto.....	30
<b>Figura 6.</b> Barcodes do TCGA.....	35
<b>Figura 7.</b> Associação das regiões de mutação e metilação.....	39
<b>Figura 8.</b> Mutações C>T ocorrendo em sítios CpG metilados em regiões codificadoras tanto na fita + quanto na fita -.....	40
<b>Figura 9.</b> Mutação em sítio CpG metilado.....	40
<b>Figura 10.</b> Informações referentes a fita +.....	41
<b>Figura 11.</b> Mecanismo de busca do algoritmo.....	42
<b>Figura 12.</b> Proporções de todas as trocas possíveis para cada tumor.....	47
<b>Figura 13.</b> Proporções de todas as trocas possíveis em sítios CpG para cada tumor.....	48
<b>Figura 14.</b> Sítios investigados pelo array 450k.....	50
<b>Figura 15.</b> Metilação entre as amostras normais e tumorais versus ao tipo de mutação correspondente a troca .....	51
<b>Figura 16.</b> Oncoprint da quantidade de mutações por gene.....	56

**Figura 17.** Vias enriquecidas segundo os genes que sofreram mutações em sítios CpG causadas por metilação separadas por tipo tumoral.....58

## Lista de Tabelas

<b>Tabela 1.</b> Tabela contendo os tipos tumorais que contém amostras pareadas para os dados do array 450.....	33
<b>Tabela 2.</b> Tabela contendo as quantidades de amostras e variantes para cada um dos tumores quando pareados com os dados de metilação.....	45
<b>Tabela 3.</b> Tabela contendo os valores absolutos da quantidade de cada tipo de troca para cada tumor nos no arquivo .maf.....	69
<b>Tabela 4.</b> Tabela contendo a soma das trocas complementares para cada um dos tumores.....	71
<b>Tabela 5.</b> Tabela contendo a contagem total de trocas em sítios CpG para cada tumor.....	72
<b>Tabela 6.</b> Tabela contendo a soma das trocas complementares em sítios CpG para cada um dos tumores.....	73
<b>Tabela 7.</b> Tabela contendo as trocas em sítios CpG metilados para cada um dos tecidos em amostras normais.....	74
<b>Tabela 8.</b> Tabela contendo a soma das trocas complementares em sítios CpG metilados para cada um dos tecidos em amostras normais.....	75
<b>Tabela 9.</b> Tabela contendo trocas em sítios CpG não metilados para cada um dos tecidos em amostras normais.....	76
<b>Tabela 10.</b> Tabela contendo a soma das trocas complementares em sítios CpG não metilados para cada um dos tecidos em amostras normais.....	78

<b>Tabela 11.</b> Tabela contendo as trocas em sítios CpG metilados para cada um dos tecidos em amostras tumorais.....	79
<b>Tabela 12.</b> Tabela contendo a soma das trocas complementares em sítios CpG metilados para cada um dos tecidos em amostras tumorais.....	80
<b>Tabela 13.</b> Tabela contendo as troca em sítios CpG não metilados para cada um dos tecidos em amostras tumorais.....	81
<b>Tabela 14.</b> Tabela contendo a soma das trocas complementares em sítios CpG não metilados para cada um dos tecidos em amostras tumorais.....	82
<b>Tabela 15.</b> Tabela que relaciona os dados metilação entre amostras normais e tumorais.....	83
<b>Tabela 16.</b> Tabela contendo a contagem dos tipos de mutações.....	84
<b>Tabela 17.</b> Tabela contendo contagem de trocas patogênicas e não patogênicas.....	85

## **Lista de Abreviaturas**

*DNA - Desoxiribonucleic Acid*

*GDC - Genomic Data Commons*

*TCGA - The Cancer Genome Atlas*

*BLCA - Bladder Urothelial Carcinoma*

*BRCA - Breast invasive carcinoma*

*CESC - Cervical squamous cell carcinoma and endocervical adenocarcinoma*

*CHOL - Cholangiocarcinoma*

*COAD - Colon adenocarcinoma*

*ESCA - Esophageal carcinoma*

*GBM - Glioblastoma multiforme*

*HNSC - Head and Neck squamous cell carcinoma*

*KIRC - Kidney renal clear cell carcinoma*

*KIRP - Kidney renal papillary cell carcinoma*

*LHC - Liver hepatocellular carcinoma*

*LUAD - Lung adenocarcinoma*

*LUSC - Lung squamous cell carcinoma*

*PAAD - Pancreatic adenocarcinoma*

*PCPG - Pheochromocytoma and Paraganglioma*

*PRAD - Prostate adenocarcinoma*

*READ - Rectum adenocarcinoma*



*SARC- Sarcoma*

*SNVs - single nucleotide variants*

*STAD - Stomach adenocarcinoma*

*THCA - Thyroid carcinoma*

*THYM - Thymoma*

*UCEC - Uterine Corpus Endometrial Carcinoma*

# Índice

<b>1) INTRODUÇÃO</b>	<b>21</b>
<b>2) OBJETIVO GERAL</b>	<b>29</b>
2.1 OBJETIVOS ESPECÍFICOS	29
<b>3) MATERIAIS E MÉTODOS</b>	<b>30</b>
3.1 MATERIAL	31
3.1.1 Delineamento de estudo	31
3.1.2 Download dos dados	31
3.2 MÉTODOS	33
3.2.1 Filtragens	33
3.2.2 Filtragens de dados de mutação	34
3.2.3 Conversão de formato de arquivo	36
3.2.4 Seleção somente de single nucleotide variants (SNVs)	36
3.2.5 Proporções de trocas de todo exoma	36
3.2.6 Encontrando a posição de todos os sítios CpG do Genoma	36
3.2.7 Intersecção dos dados de mutação com os sítios CpG.	37
3.2.8 Proporções de trocas em sítios CpG do exoma	37
3.2.9 Separação das amostras de metilação entre normal e tumoral	37
3.2.10 Identificação de trocas C > T em sítios metilados - Estratégia do algoritmo	37
3.2.11 Filtragens a partir do beta value	41
3.2.12 Identificação de Patogenicidade	42
3.2.13 Identificação de Genes	42
3.2.14 Categorias de processos biológicos as quais os genes pertencem	42
3.2.15 Enriquecimento de Vias	43
<b>4) RESULTADOS</b>	<b>44</b>
4.1 Proporções	46
4.2 Identificação de trocas em sítios CpG	47
4.3 Intersecção com dados de metilação	49
4.4 Separação entre metilados e não metilados	50
4.5 Mutações distribuídas pelos genes	55
4.6 Identificação de categorias a que os genes pertencem	57
4.7 Enriquecimiento de vías	60
<b>5) DISCUSSÃO</b>	<b>62</b>
<b>6) CONCLUSÃO</b>	<b>67</b>
<b>7) REFERÊNCIAS</b>	<b>68</b>

<b>8) DADOS SUPLEMENTARES</b>	<b>72</b>
8.1 Tabelas	72
8.2 Programas	90
<b>9. MANUSCRITO</b>	<b>104</b>



# 1) INTRODUÇÃO

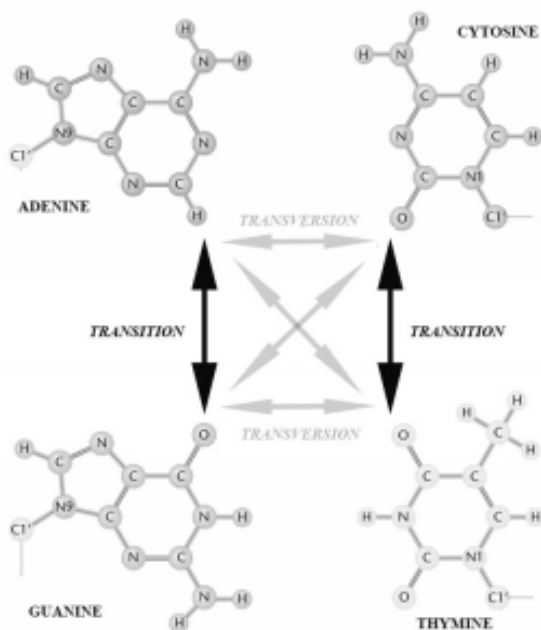
A metilação é um mecanismo epigenético importante de regulação da expressão gênica, que apesar de poder mudar sua expressão, não altera a sequência do DNA. Desta forma, é responsável pelo desencadeamento de diversos tipos de doenças complexas, como o câncer, diabetes, obesidade, dentre outras. Essa metilação pode ocorrer principalmente de duas formas, através da metilação de histonas, onde os grupamentos metil se ligarão as histonas fazendo com que a fita de DNA fique mais compacta, dificultando assim o acesso aos fatores de transcrição e consequente expressão gênica e metilação do DNA (KIM, Somi, Kaang, Bong-Kiun. 2017). No segundo caso ela ocorre em maior abundância sobre as bases de citosina, no quinto carbono da cadeia, gerando a 5-metilcitosina interferindo também na expressão gênica, dificultando a ligação dos fatores de transcrição (Jaenisch R and Bird A., 2003; Felsenfeld G., 2014; Tirado-Magallanes, Roberto et al., 2017).

Em mamíferos estima-se que 80% dos eventos de metilação ocorram em dinucleotídeos CG (sítios CpG), entretanto sua ocorrência no genoma é baixa (em média 1/80 pb). No entanto, a frequência esperada (em média 1/16 pb) para esses dinucleotídeos pode ser encontrada em algumas regiões genômicas denominadas ilhas CpG. Frequentemente as ilhas CpG estão localizadas em regiões gênicas (cerca de 70% dos genes codificadores de proteínas) levando em consideração desde regiões codificadoras até regiões reguladoras (Illingworth, Robert S. and Bird, Adrian P. Long., 2009; Mark D. et al., 2017; Lovkvist, Cecilia et al., 2016). A baixa ocorrência de

dinucleotídeos CpG no genoma é em grande parte atribuída a hipermutabilidade de CpGs metilados para TpGs (ou CpAs na cadeia complementar) que foram se acumulando ao longo da evolução. Logo, regiões de ilha CpG se encontram pouco metiladas devido a pressão seletiva para que esses locais se mantenham conservados (Lander ES, et al., 2001; Zhao Z. and Zhang F., 2006; Laird C.D, et al., 2004).

As estimativas sugerem que 35% das mutações pontuais causadoras de distúrbios genéticos humanos ocorram nos dinucleotídeos CpG, e mais de 90% destes, são transições de C para T ( $C > T$ ) ou transições correspondentes G para A ( $G > A$ ) frente a uma quantidade menor de possibilidades de mutações, de transições T para C ( $T > C$ ) e A para G ( $A > G$ ) e a qualquer tipo de transversão, figura 1 (Rideout III, William M. et al., 1990). A metilação de sítios CpGs em tecidos normais pode aumentar a probabilidade de mutações nesses locais devido à capacidade de 5-metilcitosina sofrer desaminação, resultando em timina (Sassa, Akira et al., 2016).

**Figura 1.** Possibilidades de mutações, transições e transversões. Na figura são apresentadas as possibilidades de mutações (transições e transversões), sendo que as transições C > T e G > A ocorrem em maior frequência.



Fonte: Palero, Ferran; Crandall, Keith A., 2016.

Uma das possibilidades decorrentes da hipermutabilidade C > T é a ocorrência do mecanismo de desaminação hidrolítica. Esse mecanismo é a perda de um grupamento amina (-NH<sub>2</sub>) que certas bases podem sofrer como resultado da reação de hidrólise. Esse tipo de alteração química pode ocorrer nas bases adenina, guanina e citosina (Figura 2).

**Figura 2.** Conversão das bases nitrogenadas ao sofrerem o mecanismo de desaminação. É mostrada a conversão das bases nitrogenadas ao sofrerem o mecanismo de desaminação. A citosina é convertida em uracila, a adenina é convertida em Hipoxantina, e a guanina é convertida em Xantina. A timina não sofre desaminação por já não conter um grupamento amina.

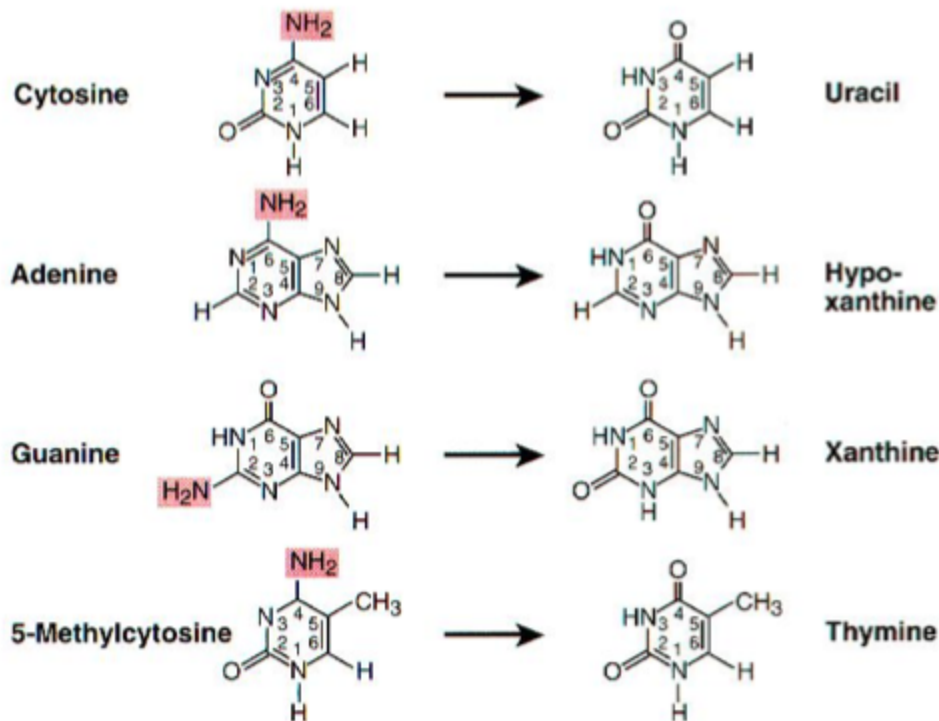


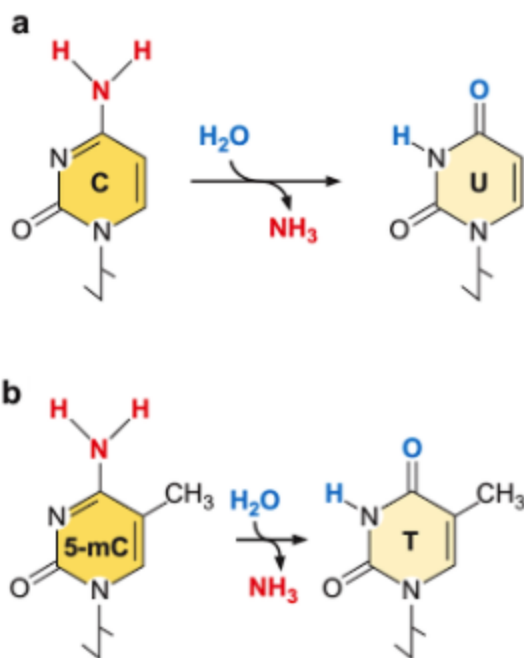
Imagem retirada de: <https://biotechkhan.wordpress.com/tag/deamination/>

A desaminação da citosina é o tipo mais frequente no DNA que ao perder um grupamento amina é convertida em uracila, ocorrendo em uma frequência de 100 transformações por genoma por dia. Entretanto, o sistema de reparo de DNA detecta a uracila como base estranha corrigindo os erros decorrentes da desaminação hidrolítica da citosina, com eliminação de todas as uracilas do DNA, o que evita a elevação da taxa de mutação a níveis insuportáveis, ou seja, sempre que aparecer um par U para G (U > G) o DNA é regenerado. Portanto em condições fisiológicas normais a citosina é



convertida em uracila, base que não é natural ao DNA, sendo assim prontamente reconhecida e corrigida pelos mecanismos de reparo, entretanto quando a desaminação espontânea ocorre sobre uma 5-metilcitosina, ela é convertida em timina, natural ao DNA, podendo assim escapar com maior facilidade da ação dos mecanismos de reparo. Sendo assim, após um ciclo de replicação será fixada no DNA (Watson, James D. et al., 2015) (Figura 3).

**Figura 3.** Mecanismo de desaminação em citosinas metiladas. **a)** É mostrada uma citosina sofrendo desaminação sendo convertida em uma uracila que é prontamente reconhecida pelos mecanismos de repara. **b)** É mostrada uma 5-metilcitosina sofrendo desaminação sendo convertida em timina, não sendo reconhecida pelos mecanismos de reparo, consolidando a mutação no DNA.



Fonte: Watson, James D. et al.2015.

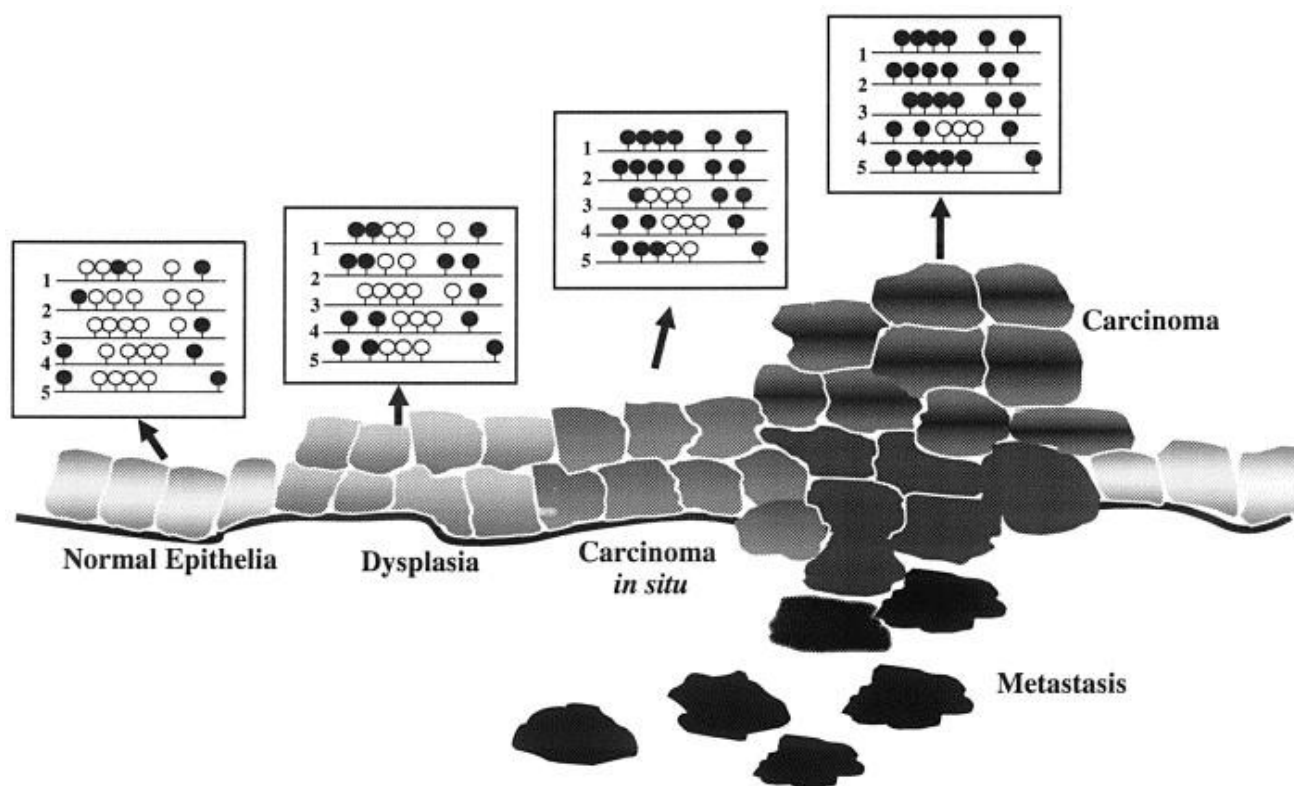
Estes tipos de modificações podem ser pró-mutagênicas podendo contribuir para a formação de hotspots mutacionais nas células (Sassa, Akira et al., 2016). Em

tumores, esse evento de troca de C > T, pode ser melhor observado uma vez que há uma hipometilação relativa geral intercalada com hipermetilação de ilhas CpG. A hipermetilação de sítios CpG com subsequente inativação de genes importantes, como genes supressores de tumor, podem fornecer uma vantagem de crescimento seletivo para células cancerosas (P.A. Jones, S.B. Baylin, 2002). No entanto, a hipermetilação de ilhas CpG pode ocorrer no início do processo neoplásico, mas, também há a possibilidade de que a hipermetilação desses genes em cânceres não familiares ocorram também tardiamente (Wong, David J. et al., 2001). Assim, pode-se destacar que com a progressão tumoral a quantidade de regiões de sítios CpG metilados aumenta. Em geral, a densidade de locais CpG metilados dentro de um locus, bem como o número de locais metilados aumentam em estágios mais avançados de câncer. A figura 4 exemplifica bem esse mecanismo, podemos ver a progressão de um tumor desde o seu início quando vemos um tecido epitelial normal até o estabelecimento de um carcinoma. A figura mostra 5 ilhas CpG onde os pequenos círculos claros representam sítios CpG não metilados e os círculos escuros representam os sítios CpG presentes em cada ilha CpG que estão metilados. Podemos perceber que à medida que o tumor progride, a quantidade de sítios CpG metilados aumenta. Assim a quantidade de sítios CpG metilados no carcinoma é muito maior que a quantidade de sítios CpG metilados nas etapas anteriores da progressão tumoral. Essas regiões metiladas podem contribuir para a progressão tumoral, uma vez que genes importantes como supressores tumorais possam ser silenciados, não desempenhando assim seu papel. Entretanto, nossa hipótese é que por haver uma grande quantidade de sítios CpG metilados, eles podem estar sofrendo também a desaminação hidrolítica, lavando

a mutações que podem também alterar a expressão de genes importantes para a instabilidade tumoral.

(Nephew, Kenneth P. et al., 2003). Isto pode se tornar especialmente relevante quando os genes envolvidos são supressores tumorais (Abdelfatah, Eihab et al., 2016).

**Figura 4.** Demonstração do aumento de metilação em sítios CpG de acordo com a progressão tumoral.



Fonte: Nephew, Kenneth P. et al.2003.

Apesar de alguns autores já terem trabalhado com essa hipótese de mutações associadas a metilação, como no trabalho de Weisenberger, Daniel J. e colaboradores (2006), onde no câncer colorretal, foi demonstrada uma forte associação entre um fenótipo de hipermetilação e mutação do oncogene BRAF que codifica uma proteína

quinase envolvida na via EGFR a jusante de KRAS (Weisenberger, Daniel J. et al., 2006), ainda não existem trabalhos envolvendo uma grande quantidade de tumores e amostras descrevendo qual o verdadeiro impacto do mecanismo de desaminação na instabilidade gênica em tumores.

Desta forma, o entendimento dos mecanismos envolvidos na progressão do câncer são de grande interesse por parte da comunidade científica. Uma vez que, de certa forma a progressão da tumorigênese possa ser diminuída, é fundamentada por diferentes abordagens terapêuticas no processo de cura, ou, inicialmente por estudos utilizando ferramentas em bioinformática de predição de mutações. Sendo assim, apoiados na hipótese de que o mecanismo de desaminação desempenha um papel relevante na promoção da instabilidade gênica em tumores, o presente trabalho foi delineado para investigar o impacto da metilação do DNA na taxa de substituições C > T em tecidos tumorais através do desenvolvimento de uma abordagem computacional. Para este projeto foram inicialmente escolhidos os dados de mutação somática em exons (gerados por next generation sequence) e metilação, pois o principal objetivo foi encontrar trocas C > T em sítios CpG metilados (tanto para amostras normais quanto para tumorais), com essa informação foi possível inferir que essa troca ocorreu devido ao mecanismo de desaminação.

## 2) OBJETIVO GERAL

Investigar o impacto do mecanismo de metilação na instabilidade cromossômica em diferentes tumores, com ênfase nas mutações pontuais intragênicas com auxílio de uma abordagem *in silico*.

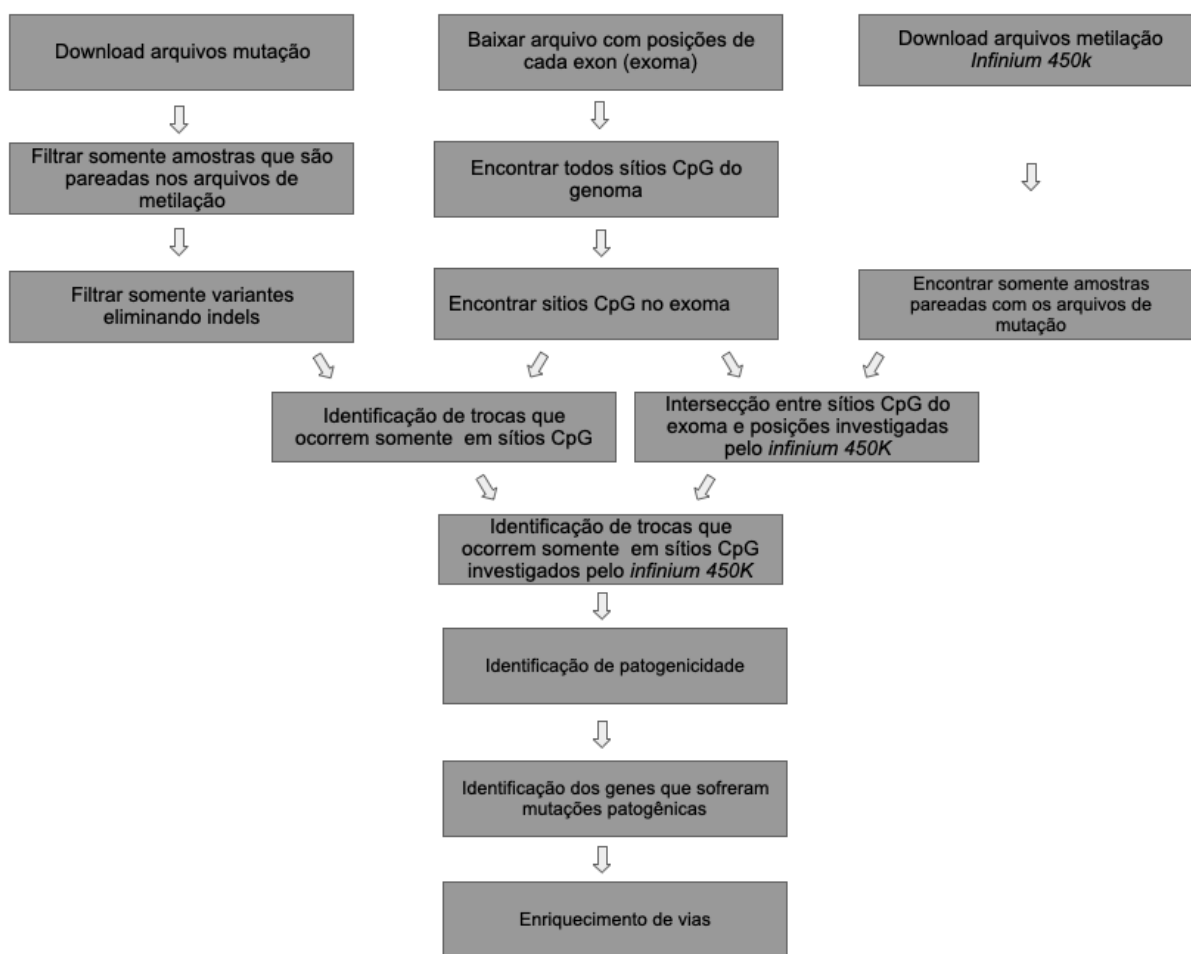
### 2.1 OBJETIVOS ESPECÍFICOS

- Identificar a taxa de substituições C > T causada pelo mecanismo de desaminação em sítios CpG metilados em regiões codificadoras em tecidos normais e tumorais.
- Analisar a taxa de mutação patogênica em regiões codificadoras.
- Analisar as vias biológicas enriquecidas com genes mutados sob ação da metilação do DNA.

### 3) MATERIAIS E MÉTODOS

Primeiramente, foi delineado um fluxograma de abordagem computacional, conforme figura 5, que foi empregado no desenvolvimento deste projeto.

**Figura 5.** Fluxograma detalhando cada passo do projeto.



Fonte: próprio autor.

## 3.1 MATERIAL

### 3.1.1 Delineamento de estudo

Para a realização do estudo foram escolhidos os dados gerados pelo projeto *The Cancer Genome Atlas (TCGA)*. Nesse projeto foram reunidos milhares de pacientes com tumores primários que ocorrem em diferentes locais do corpo, inicialmente cobrindo 12 tipos diferentes de tumores. Foram realizadas seis tipos de caracterizações diferentes incluindo mutações, copy number, expressão gênica, metilação, expressão de microRNAs e dados clínicos, assim para cada paciente podemos ter mais de um experimento ou todos em alguns casos, contando também com amostras pareadas entre tecidos normais e tumorais (Weinstein, John N. et al., 2013).

### 3.1.2 Download dos dados

Foram realizados os downloads dos dados de mutações somáticas de exoma em formato .maf e de metilação em em formato .txt a partir do repositório público GDC (<https://gdc.cancer.gov>). Entretanto, como a quantidade de dados que foram baixadas era muito grande (da ordem de gigabytes), este processo foi realizado com o auxílio de um pacote R chamado *TCGABiolinks* (COLAPRICO, Antonio et al., 2015), este pacote está presente no repositório bioconductor e permite consulta, download e realização de análises integrativas de dados do TCGA de forma rápida e prática. O download dos dados foi realizado em duas etapas. Os primeiros dados baixados foram os de metilação, primeiro foram investigados quais eram todos os tipos de dados de

metilação disponíveis no repositório GDC, são eles *methylation27*, *methylation450* e *methylationHiseq*, os escolhidos para o download foram os dados gerados com o array *methylation450*, por investigarem uma maior quantidade de sítios CpG aproximadamente 99% dos genes do genoma humano e conter uma maior quantidade de dados. Foi investigado também quais eram todos os tipos tumorais contendo informações do array *methylation450* que estavam disponíveis para download no repositório GDC e quantas amostras cada um deles continha (cada amostra corresponde a um arquivo), posteriormente foi investigado quais desses tipos tumorais continham amostras pareadas, ou seja, para um mesmo paciente foi investigada a presença de amostras normais e tumorais, posteriormente somente os dados dos tipos tumorais contendo amostras pareadas com normais foram baixados (programa `download_metylation_files.R`).

A segunda etapa realizada foi o download dos dados de mutação, onde somente foram baixados os dados dos tipos tumorais selecionados na etapa anterior (os tipos tumorais que continham tanto informações de metilação quanto de mutação), ou seja, os tipos tumorais que continham informações de amostras pareadas para o array *methylation450* (programa `download_mafs.R`). O download dos dados de metilação foi realizado no dia 23/09/2016 enquanto que o download dos dados de mutação foi realizado no dia 25/09/2016.



## 3.2 MÉTODOS

### 3.2.1 Filtragens

Os tipos tumorais que continham amostras pareadas (normal/tumoral) para os dados do array *methylation 450* estão descritos na **tabela 1**.

**Tabela 1.** Tabela contendo os tipos tumorais que contém amostras tanto para os dados de mutação quanto para os dados do array *methylation 450*.

Identificador	Tipo Tumoral
<b>TCGA-BLCA</b>	Bladder Urothelial Carcinoma
<b>TCGA-BRCA</b>	Breast invasive carcinoma
<b>TCGA-CESC</b>	Cervical squamous cell carcinoma and endocervical adenocarcinoma
<b>TCGA-CHOL</b>	Cholangiocarcinoma
<b>TCGA-COAD</b>	Colon adenocarcinoma
<b>TCGA-ESCA</b>	Esophageal carcinoma
<b>TCGA-GBM</b>	Glioblastoma multiforme
<b>TCGA-HNSC</b>	Head and Neck squamous cell carcinoma
<b>TCGA-KIRC</b>	Kidney renal clear cell carcinoma
<b>TCGA-KIRP</b>	Kidney renal papillary cell carcinoma
<b>TCGA-LIHC</b>	Liver hepatocellular carcinoma
<b>TCGA-LUAD</b>	Lung adenocarcinoma

<b>TCGA-LUSC</b>	Lung squamous cell carcinoma
<b>TCGA-PAAD</b>	Pancreatic adenocarcinoma
<b>TCGA-PCPG</b>	Pheochromocytoma and Paraganglioma
<b>TCGA-PRAD</b>	Prostate adenocarcinoma
<b>TCGA-READ</b>	Rectum adenocarcinoma
<b>TCGA-SARC</b>	Sarcoma
<b>TCGA-STAD</b>	Stomach adenocarcinoma
<b>TCGA-THCA</b>	Thyroid carcinoma
<b>TCGA-THYM</b>	Thymoma
<b>TCGA-UCEC</b>	Uterine Corpus Endometrial Carcinoma



Fonte: próprio autor.

### 3.2.2 Filtragens de dados de mutação

Apesar dos experimentos terem sido feitos com os mesmos tipos tumorais, nem sempre eles foram realizados com as mesmas amostras (cada amostra representa um paciente). Nesta etapa foram selecionadas somente as amostras presentes em arquivos de metilação e nos arquivos de mutação (nomeadas de amostras pareadas). As amostras do TCGA são identificadas com um Id composto chamado de barcode (Figura 6) onde o terceiro campo indica qual é o Id do paciente, os outros campos são referentes a outras informações referentes a amostra, como projeto a que pertence, tipo tumoral, centro de onde foi realizado o experimento, centro onde foram realizadas as análises, entre outros. Dessa forma independentemente de qual foi o

experimento realizado o terceiro campo do barcode sempre será o identificador do paciente doador da amostra. Nessa etapa o objetivo foi encontrar as amostras oriundas do mesmo paciente submetidas a dois diferentes experimentos (sequenciamento de DNA e identificação de regiões metiladas). Foi gerada uma lista para cada tipo tumoral contendo o identificador de todos pacientes doadores de amostras para os dados de metilação, e a partir dessa lista foram procurados os mesmos ids nos dados de mutação selecionando assim somente as amostras contendo dados de mutação pareadas com os dados de metilação. Essa busca foi realizada com auxílio da ferramenta grep, nativa do bash script.

**Figura 6.** Barcodes do TCGA. Figura mostrando os barcodes do TCGA, o primeiro barcode representa uma amostra de mutação, o segundo barcode representa uma amostra de metilação de tecido normal adjacente e o terceiro barcode representa uma amostra de metilação de tecido tumoral. Em vermelho está o identificador do paciente doador, mostrando que todas as amostras pertencem ao mesmo paciente. Em azul temos o identificador do tipo de amostra sendo 01 referenciando tecidos tumorais e 11 referenciando tecidos normais adjacentes.

TCGA-E2-**A1L7**-01A-11D-A142-09 mutação

TCGA-E2-**A1L7**-11A-33D-A145-05 metilação normal

TCGA-E2-**A1L7**-01A-11D-A145-05 metilação tumor

Fonte: Adaptado de Weinstein, John N. et al., 2013.

### 3.2.3 Conversão de formato de arquivo

Foi também realizado uma conversão no formato dos arquivos, de .maf para .bed e .txt para .bed, para com um formato de dados padrão, amplamente utilizado em análises genômicas seria mais simples de utilizar as ferramentas do bedtools (Quinlan, A. R.; Hall, I. M., 2010) por exemplo para facilitar as análises. Para a conversão de .maf para .bed foram retirados alguns campos do arquivo .maf que não seriam utilizados nas análise e para conversão de .txt para bed, fazendo uso da linguagem nativa do linux awk que facilita o trabalho com tabelas.

### 3.2.4 Seleção somente de single nucleotide variants (SNVs)

Os dados de mutações contém trocas de bases inserções e deleções, no entanto foi confeccionado um programa para somente selecionar as trocas entre as bases, eliminando assim as inserções e deleções, pois não são o objetivo de estudo do projeto. O programa identifica qual o campo da tabela que diz qual o tipo de mutação e somente selecionava as SNVs.

### 3.2.5 Proporções de trocas de todo exoma

Foi desenvolvido um programa para fazer uma análise prévia das proporções de todas as trocas contidas em cada um dos arquivos de cada tipo tumoral (count\_mutations\_v2.perl). No arquivo de mutações existe um campo que diz para cada troca qual é a base identificada no tumor, outro campo que diz qual a base identificada no tecido normal e um campo que diz qual a base identificada no genoma

referencia. Este programa quais as trocas levando em consideração qual a base identificada no tecido normal e qual a base identificada no tecido tumoral.

### 3.2.6 Encontrando a posição de todos os sítios CpG do Genoma

Com a utilização do programa `get_pattern.pl` baixado do link `<script src="https://gist.github.com/stratust/d6f611f050de6f7153d7ac00541ee679.js"></script>` foi encontrada a posição de todos os sítios CpG do genoma estudado e posteriormente do exoma. Este programa identifica padrões, o padrão estabelecido foi `cg`, ao vasculhar o genoma, toda vez que o programa encontrava o padrão `cg`, recuperava a posição e incrementava o contador. Assim, foi possível identificar todas as coordenadas de todos os sítios CpG do genoma Este programa foi desenvolvido por Thiago Yukio Kikuchi Oliveira.

### 3.2.7 Intersecção dos dados de mutação com os sítios CpG.

Foi utilizado então o software `bedtools intersect` (Quinlan, A. R.; Hall, I. M., 2010) para realizar a intersecção entre os dados de mutação e os sítios CpG do exoma. Este software trabalha com coordenadas genômicas, assim, ele pode comparar dois arquivos contendo coordenadas genômicas e gerar um terceiro arquivo com a intersecção dessas coordenadas, ou seja, as regiões genômicas que estão em ambos os arquivos. Assim foi possível identificar quais as mutações que estavam em regiões de sítios CpG.

### 3.2.8 Proporções de trocas em sítios CpG do exoma

O programa `count_mutations_v2.perl` foi novamente utilizado para identificação das proporções de trocas em sítios CpG.

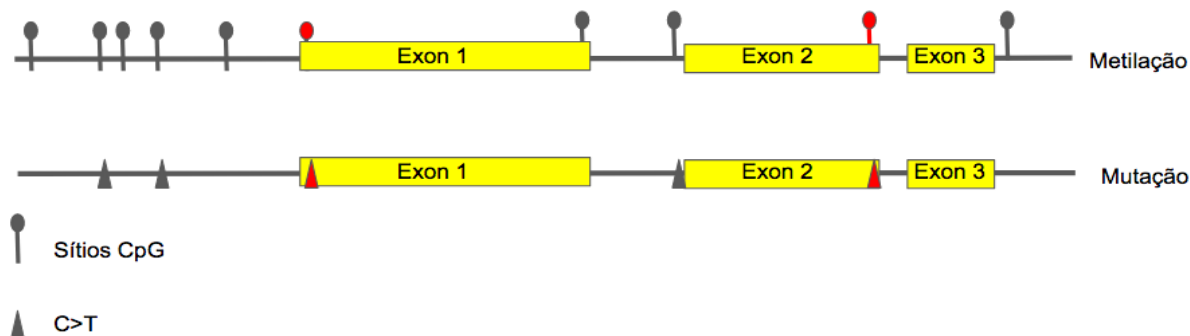
### 3.2.9 Separação das amostras de metilação entre normal e tumoral

Como as amostras de mutação foram baixadas todas de uma vez elas vieram misturadas entre normais e tumorais, então foi confeccionado um programa para separar essas amostras em diferentes diretórios: normais e tumorais tendo como base para essa separação o quarto campo do *barcode*. Foi utilizado o quarto campo do *barcode* para essa separação onde 01 representam amostras tumorais e 11 representa amostras normais. O programa identificava esse campo direcionava cada amostra para um diretório correspondente.

### 3.2.10 Identificação de trocas C > T em sítios metilados - Estratégia do algoritmo

Foi construído um programa que tinha como objetivo encontrar mutações C > T em sítios CpG metilados, para tal era necessário associar informações referentes às coordenadas em que cada mutação C > T ocorria (informação armazenada nos arquivos de mutação) com as coordenadas de cada sítio CpG metilado (informação armazenada nos arquivos de metilação) (Figura 7).

**Figura 7.** Associação das regiões de mutação e metilação. Figura indicando como foi feita a associação das informações de mutação e metilação.

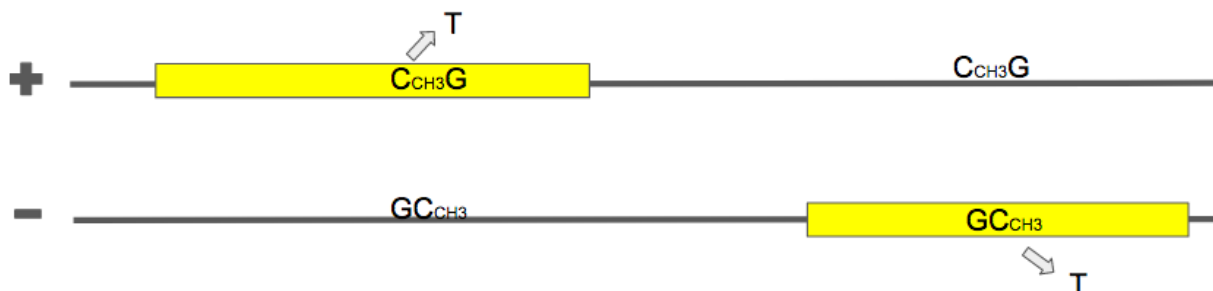


Fonte: Próprio autor.

Na primeira parte da figura 7 indicada por metilação, temos a representação de uma região codificadora no genoma, mostrando na região à esquerda do exon 1 o que seria a região promotora contendo vários sítios CpG metilados e também alguns sítios CpG metilados que invadem os éxons. Na segunda parte da figura temos uma representação da mesma região genômica só que agora indicando a posição de mutações C > T naquela região, pode-se ver que existem algumas mutações na região promotora e também na região codificadora. Pode-se ver também que existem tanto mutações, quanto sítios CpG indicados em vermelho, esses representam mutações e sítios CpG metilados na mesma coordenada genômica, são justamente esses casos que o programa busca.

Assim seria necessário que o programa identificasse 4 diferentes situações, as duas primeiras são as mais evidentes, quando ocorre uma troca C > T em um sítio CpG metilado em uma região codificadora (figura 8).

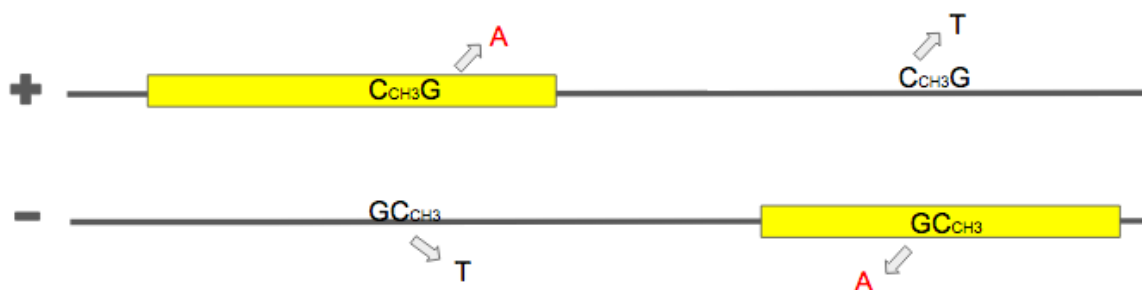
**Figura 8.** Mutações C > T ocorrendo em sítios CpG metilados em regiões codificadoras tanto na fita + quanto na fita -. Figura mostrando mutações C > T ocorrendo em sítios CpG metilados em regiões codificadoras tanto na fita + quanto na fita -.



Fonte: Próprio autor.

As outras duas são mutações menos evidentes pois a mutação em uma região codificadora é consequência de uma troca C > T na fita complementar à fita em que o gene está presente (Figura 9).

**Figura 9.** Mutações em sítio CpG metilado. Figura mostrando quando a mutação C > T em sítio CpG metilado ocorre na fita reversa com relação ao gene e só depois de um ciclo de replicação ela se fixa também onde se encontra o gene, mas como uma troca de G > A.



Fonte: Próprio autor.



Entretanto a anotação tanto dos dados de mutação quanto dos dados de metilação só é feita na fita +, então mesmo que a mutação ocorra na fita - ela será registrada na fita +. Da mesma forma os dados de metilação, pois só são investigadas as metilações em sítios CpG da fita +, dessa forma não é possível saber nenhuma informação sobre sítios metilados na fita *reverse* figura 10.

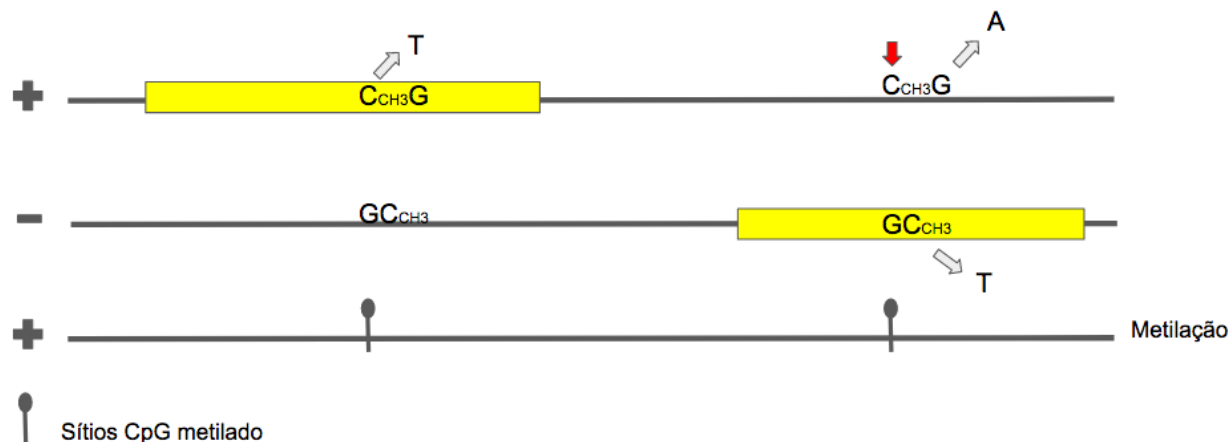
**Figura 10.** Informações referentes a fita +. Figura mostrando que só se tem informações referentes a fita + tendo trocas de C > T e G > A mas sem saber se as trocas de G > A tiveram origem em uma troca C > T.



Fonte: Próprio autor.

Assim o programa foi construído da seguinte maneira, quando uma troca era de C > T, era investigada a posição exata da troca nos arquivos de metilação para tentar encontrar um sítio CpG metilado naquela exata região, mas quando uma troca G > A também era investigada a posição imediatamente anterior, caso fosse um C, significaria que ali anteriormente havia um sítio CpG, então a posição do C anterior era procurada no arquivo de metilação, pois entende-se que se um sítio CpG estiver metilado o seu complementar também o estará (TIRADO-MAGALLANES, Roberto et al., 2017; LAIRD C.D, et al., 2004), e poderia se supor que a troca G > A foi causada por uma troca C > T na fita complementar (Figura 11) (programa **Compare\_mutation\_metilation2.perl**)

**Figura 11.** Mecanismo de busca do algoritmo. Figura mostrando que quando se tem uma troca C > T sua posição exata é procurada no arquivo de metilação, mas quando se tem uma troca G > A, investiga-se a posição anterior a troca, caso seja um C essa posição será procurada no arquivo de metilação, pois supõe-se que o C do sítio CpG na fita complementar também esteja metilado.



Fonte: Próprio autor.

### 3.2.11 Filtragens a partir do beta value

Primeiramente o **count\_mutations\_v2.perl** foi executado para todas trocas C > T, C > G, C > A e G > A, G > C e G > T, para sítios CpG metilados ou não, para ter um panorama da densidade de trocas C > T em sítios CpG cobertos pelo array *methylation 450*. Dessa forma foram necessárias outras filtragens após essa etapa, como seleção somente de trocas C > T e G > A, separação entre sítios CpG metilados (beta value >= 0.3) e sítios não metilados (beta value < 0.3) (SHIAH, Yu-Jia et al., 2017; WARDEN, Charles D. et al., 2013). Identificação de quais sítios contendo mutação estão mudados

tanto em amostras normais quanto em amostras tumorais. E identificação dos tipos mutações divididos em missense, nonsense, silenciosa e em RNA.

### 3.2.12 Identificação de Patogenicidade

Para as variantes identificadas como causadoras de mutações missense e nonsense foi executado o programa *UMD predictor* (Salgado, David et al., 2016), para identificação da patogenicidade das trocas. O *UMD predictor* combina propriedades bioquímicas, impacto em sinais de splicing, localização em domínios de proteínas, frequência de variação na população global e conservação através da matriz de substituição global BLOSUM62 e uma conservação específica de proteína entre 100 espécies para definir a patogenicidade de uma mutação. Apresenta resultados classificando as mutações em polimórficas, provavelmente polimórficas, patogênicas e provavelmente patogênicas rever esta frase.

### 3.2.13 Identificação de Genes

Para a identificação de genes foram geradas lista de genes que sofreram troca e a partir da informação de patogenicidade foram geradas listas de genes que sofreram trocas patogênicas segundo a predição do *UMD predictor*. Essa lista foram geradas utilizando uma combinação dos comandos em bash grep e awk.

### 3.2.15 Enriquecimento de Vias

Novamente foi utilizado o pacote *clusterprofile* presente no repositório Bioconductor (que é um pacote que implementa métodos para analisar e visualizar

perfis funcionais de genes e clusters de genes) (Yu, Guangchuang et al., 2012) para realização do enriquecimento de vias dos genes que sofreram mutações em sítios CpG causadas por metilação.

## 4) RESULTADOS

A quantidade de amostras e variantes após a realização dos primeiros filtros (pareamento entre as mostras de mutação e metilação e seleção apenas de SNVs eliminando os indels) está sintetizada na tabela 2.

**Tabela 2.** Quantidade de amostras e variantes para tumores estudados considerando o pareamento com os dados de metilação.

	Amostras (mutação)		Variantes (mutação)		
	total	pareados (normal/tumoral)	total	pareados com metilação	Somente SNV
<b>TCGA-BLCA</b>	130	15	39442	3451	3300
<b>TCGA-BRCA</b>	90	89	83363	11456	10008
<b>TCGA-CESC</b>	194	3	46741	100	91
<b>TCGA-CHOL</b>	35	9	6790	1608	1511
<b>TCGA-COAD</b>	154	0	62684	0	0
<b>TCGA-ESCA</b>	185	16	58787	9342	8015
<b>TCGA-GBM</b>	290	1	22363	61	56
<b>TCGA-HNSC</b>	279	50	52078	8697	8212
<b>TCGA-KIRC</b>	417	143	26786	10134	8807
<b>TCGA-KIRP</b>	161	43	15746	3720	3250

<b>TCGA-LIHC</b>	198	47	28090	6709	6275
<b>TCGA-LUAD</b>	230	20	72771	6344	5732
<b>TCGA-LUSC</b>	115	21	41636	8758	8611
<b>TCGA-PAAD</b>	150	4	30507	268	257
<b>TCGA-PCPG</b>	184	3	4846	107	82
<b>TCGA-PRAD</b>	332	43	12680	2753	2601
<b>TCGA-READ</b>	69	5	22144	1713	1577
<b>TCGA-SARC</b>	247	4	20623	267	239
<b>TCGA-STAD</b>	289	2	148809	9982	9324
<b>TCGA-THCA</b>	405	47	7863	1344	1241
<b>TCGA-THYM</b>	123	2	3187	15	11
<b>TCGA-UCEC</b>	248	5	185109	9138	9100

Fonte: Próprio autor.

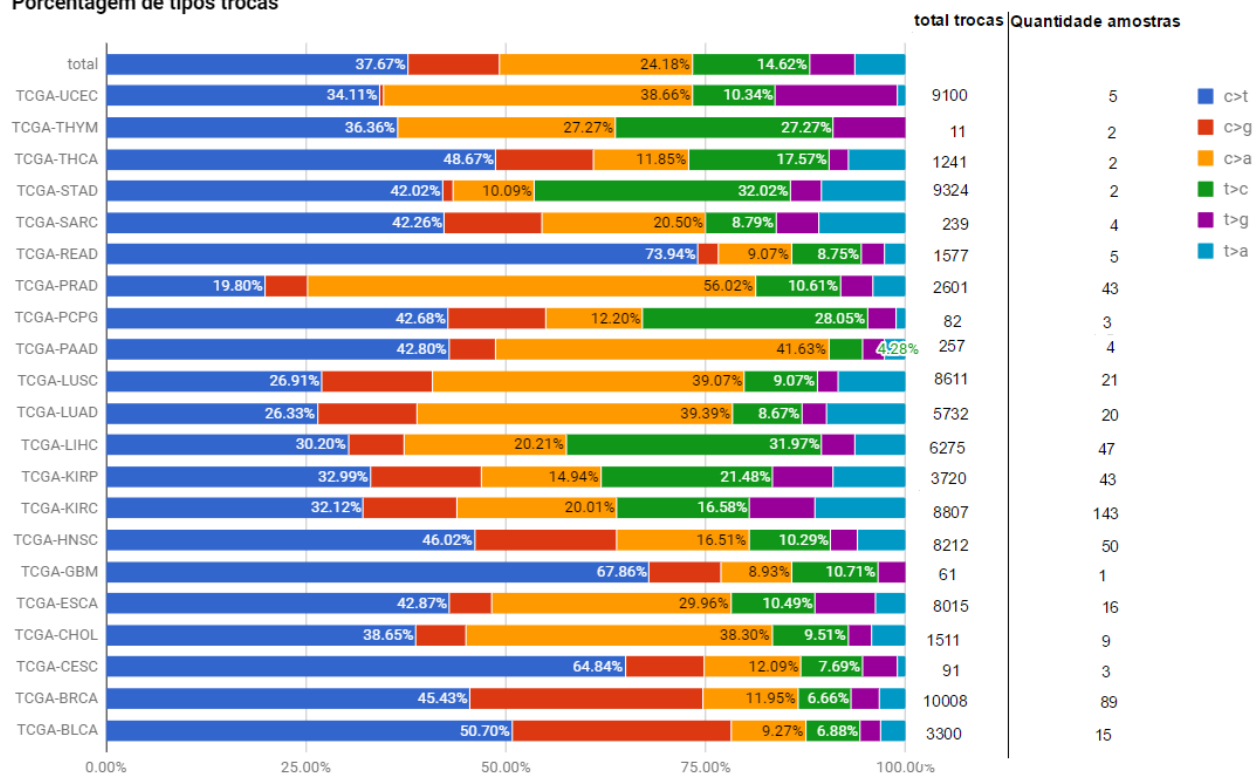
## 4.1 Proporções

Como resultado da execução do programa **count\_mutations\_v2.perl** as proporções de cada tipo de troca foi encontrado e está sumarizado na Figura 12.

**Figura 12.** Proporções de todas as trocas possíveis para cada tumor considerando todo o exoma. Gráfico representando as proporções de todas as trocas possíveis para cada tumor. As trocas complementares tais como: C > T e G > A, C > G e G > C, C > A

e  $G > T$ ,  $T > C$  e  $A > T$ ,  $T > G$  e  $A > C$ ,  $T > A$  e  $A > T$  foram somados pois após um ciclo de replicação representam as mesmas trocas.

Porcentagem de tipos trocas



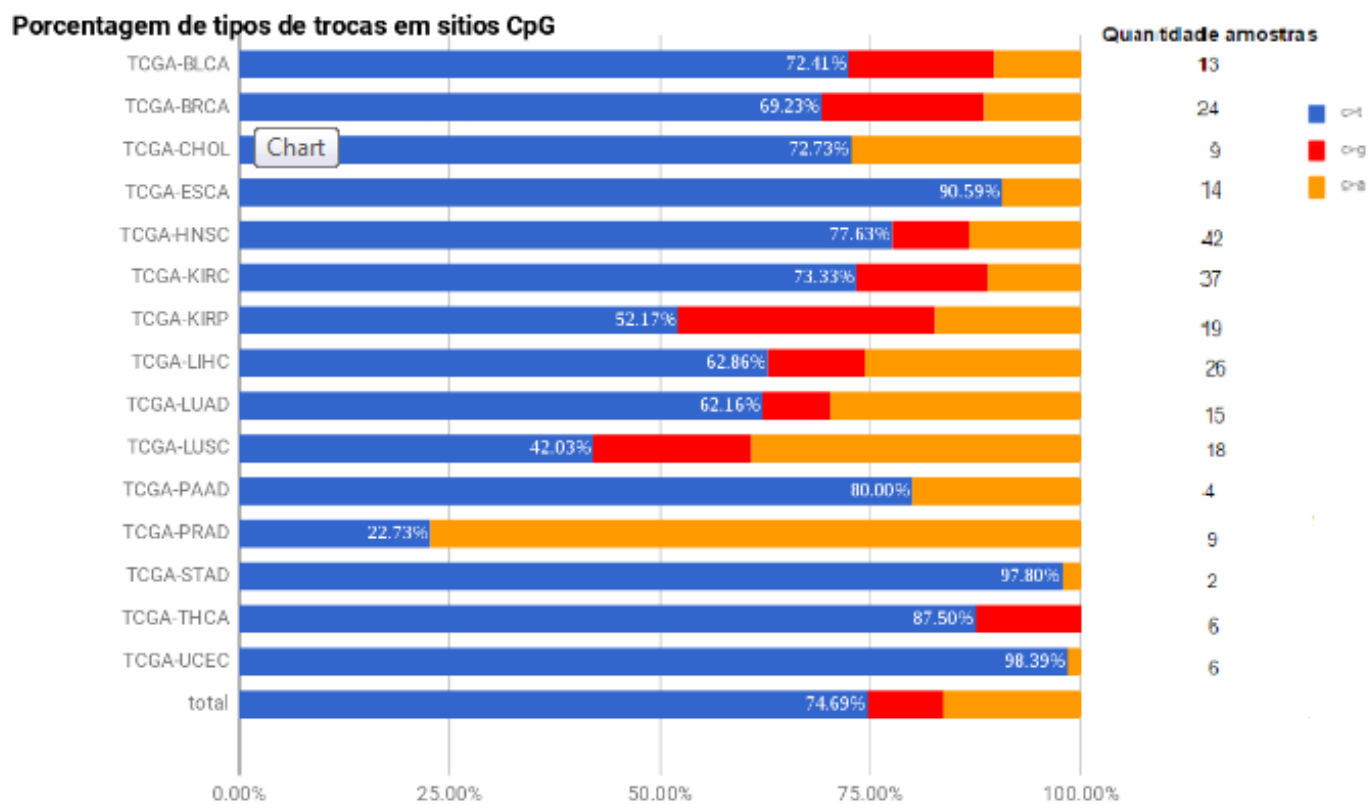
Fonte: Próprio autor.

Nesse gráfico pode ser visto que na maioria dos tumores existe uma maior porcentagem de trocas  $C > T$ , corroborando nossa hipótese. A exceção fica por conta dos tumores: TCGA-LIHC, TCGA-LUAD, TCGA-LUSC, TCGA-PRAD e TCGA-UCEC já que a porcentagem de  $C > T$  foi menor do que a porcentagem de outras trocas tais como  $T > C$  e  $C > A$ . Esse resultado pode estar sendo observado pelo fato de não estarmos olhando somente para sítios CpG, ou seja, as mutações podem estar submetidas a diversos outros mecanismos que não o mecanismo de desaminação.

## 4.2 Identificação de trocas em sítios CpG

O programa `Compare_mutation_metilation2.perl` foi executado em um primeiro momento com o objetivo de encontrar todas as trocas possíveis em sítios CpG metilados ou não, encontrando as seguintes proporções (Figura 13).

**Figura 13.** Proporções de todas as trocas possíveis em sítios CpG para cada tumor. Gráfico demonstrando as proporções de todas as trocas possíveis em sítios CpG para cada tumor. Como o array *metylation450* investiga sempre as mesmas posições, os resultados para amostras normais e tumorais é o mesmo.



Fonte: Próprio autor.



Após a execução do programa **Compare\_mutation\_metilation2.perl** não foi encontrada nenhuma troca em sítios CpG para os tipos tumorais TCGA-COAD, TCGA-PCPG, TCGA-READ, TCGA-SARC, TCGA-THYM, sendo assim esses tipos tumorais foram excluídos das análises posteriores. Os tipos tumorais TCGA-CESC e TCGA-GBM também foram excluídos do prosseguimento das análises pois só tinham uma troca em regiões de sítios CpG e essa troca não se tratava de um C > T.

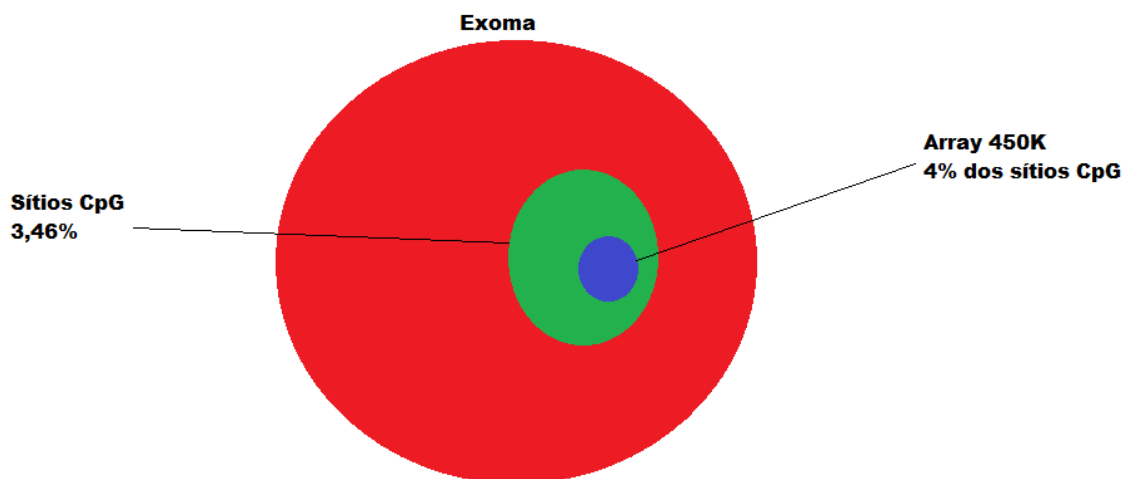
Esses resultados mostram uma predominância das trocas C > T em sítios CpG, entretanto ainda temos uma exceção, nas proporções das trocas de adenocarcinoma de próstata, onde pode ser observado um valor consideravelmente maior de trocas C > A em relação às trocas C > T.

### 4.3 Intersecção com dados de metilação

Com a execução do programa **Compare\_mutation\_metilation2.perl** foi possível fazer a separação das variantes metiladas e não metiladas. As variantes consideradas metiladas foram as que estavam metiladas em ambas amostras (normal e tumoral) ou pelo ao menos em uma delas.

Também foi possível identificar o escopo que estávamos trabalhando, como só estávamos trabalhando com trocas em sítios CpG, identificamos que os sítios CpG correspondem a somente a 3,46% de todo o genoma e que os sítios CpG investigados pelo array 450k correspondiam somente 4% dos sítios CpG. Ou seja, todos os nossos resultados correspondem a uma amostra equivalente a 4% dos sítios CpG de todo o exoma.

**Figura 14.** Sítios investigados pelo array 450k. A figura mostra em azul qual é a quantidade de sítios investigados pelo array 450k, conseqüentemente esse é o campo e busca utilizado no projeto.

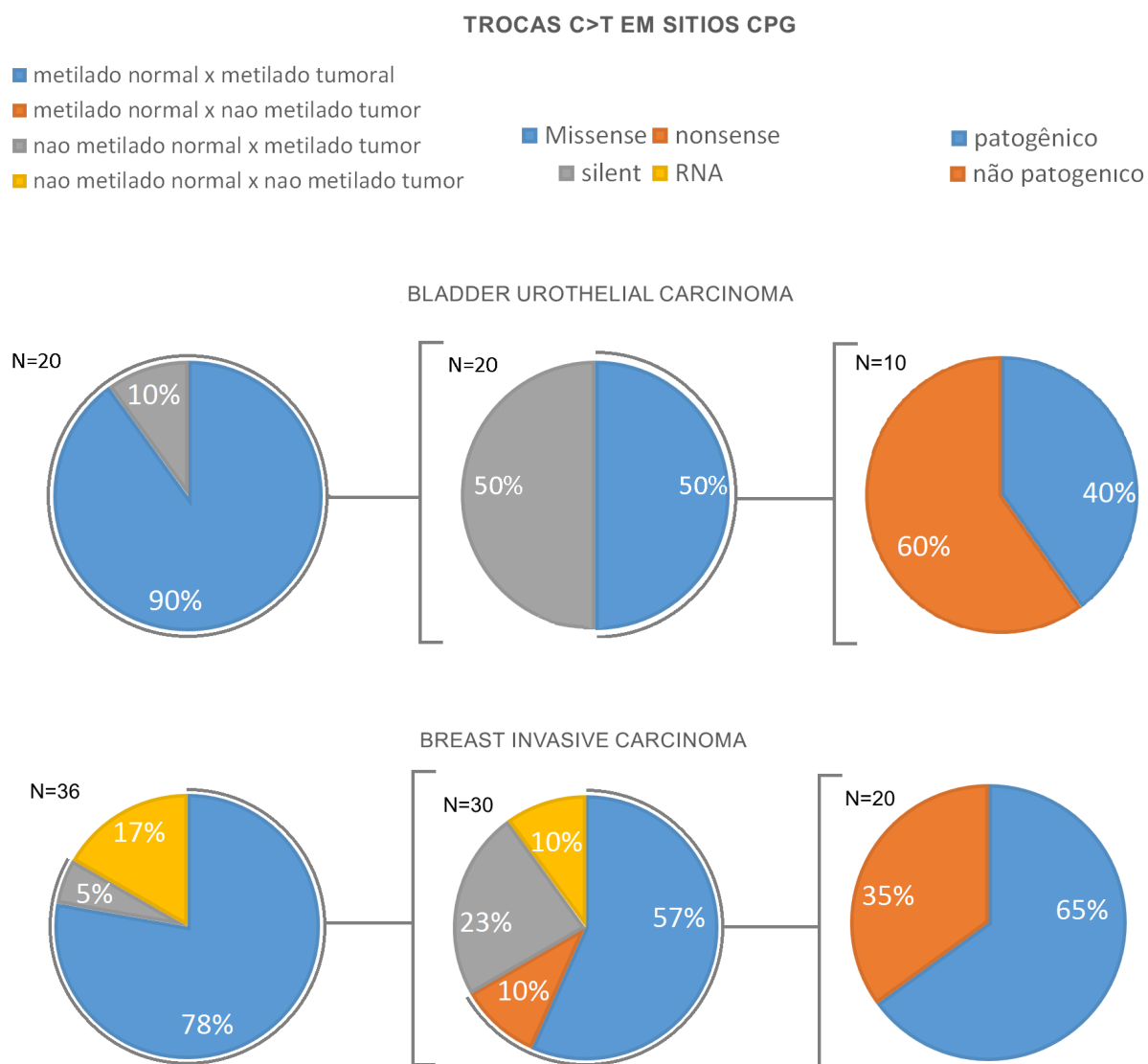


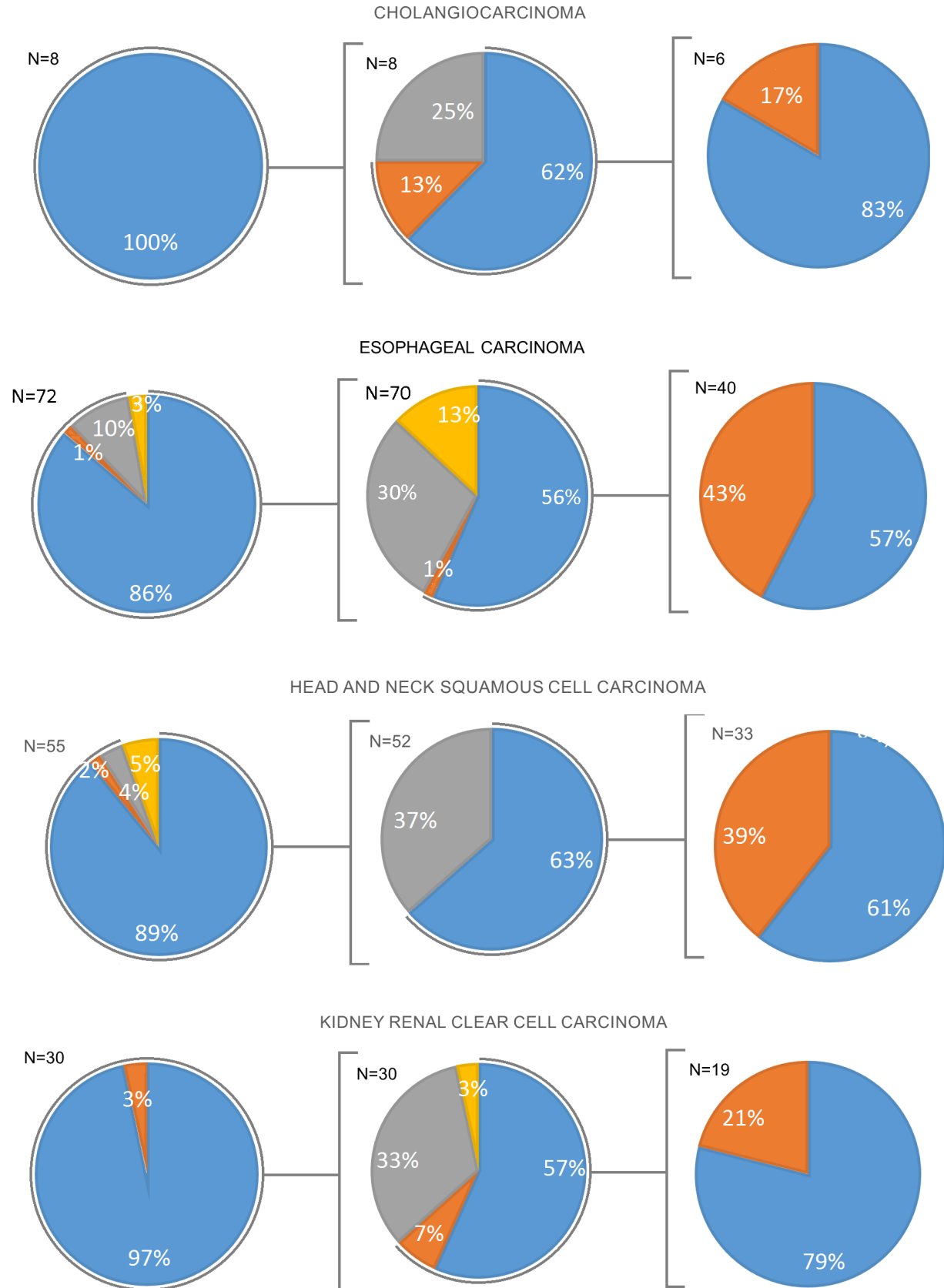
Fonte: Próprio autor.

#### 4.4 Separação entre metilados e não metilados

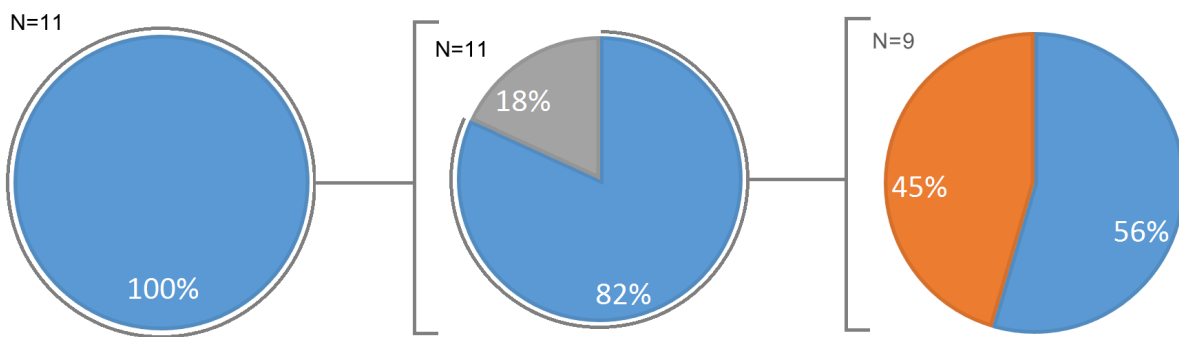
Após termos certeza de qual era o nosso campo de busca, foram selecionadas somente as trocas C > T para as próximas etapas. As variantes foram então separadas entre metiladas e não metiladas, com base nisso foi possível identificar amostras que estavam metiladas em ambas amostras (normal e tumoral) ou pelo ao menos em uma delas. A partir dessa seleção foram identificadas quais tipos de mutação essas trocas provocavam, se eram missense, nonsense, silenciosas e RNA. Desses tipos de mutações foram somente selecionadas as missenses e nonsenses para identificação de patogenicidade. A figura 15 sumariza todos os resultados obtidos nesta etapa.

**Figura 15.** Metilação entre as amostras normais e tumorais versus ao tipo de mutação correspondente a troca. Na primeira coluna estão as informações sobre metilação entre as amostras normais e tumorais. Na segunda coluna é identificado o tipo de mutação correspondente a troca, somente são levadas em consideração mutações em sítios CpG metilados. Na terceira coluna são identificadas mutações patogênicas ou não. Só são consideradas trocas que levam a mutações missense e nonsense.

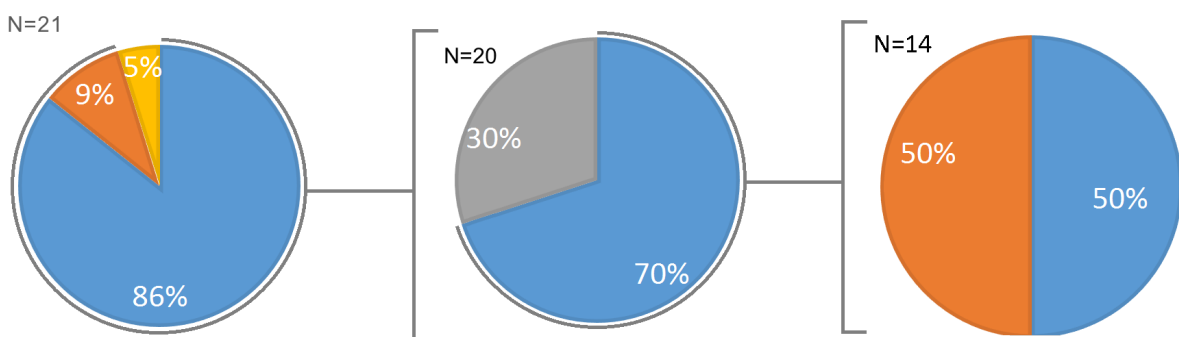




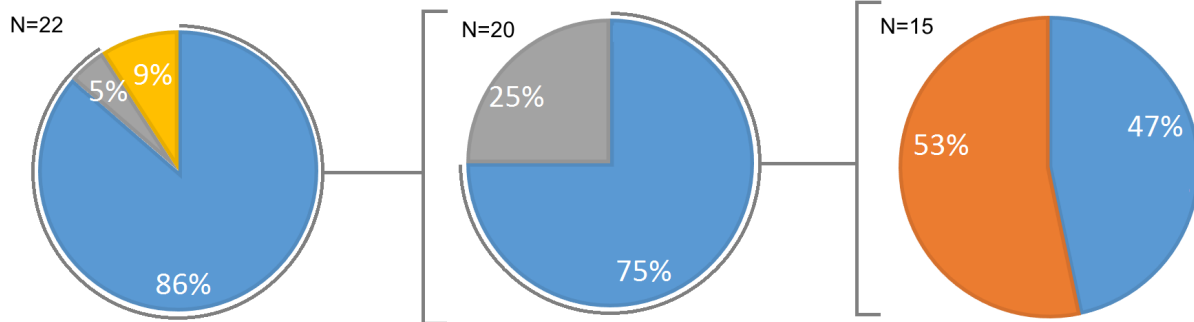
KIDNEY RENAL PAPILLARY CELL CARCINOMA



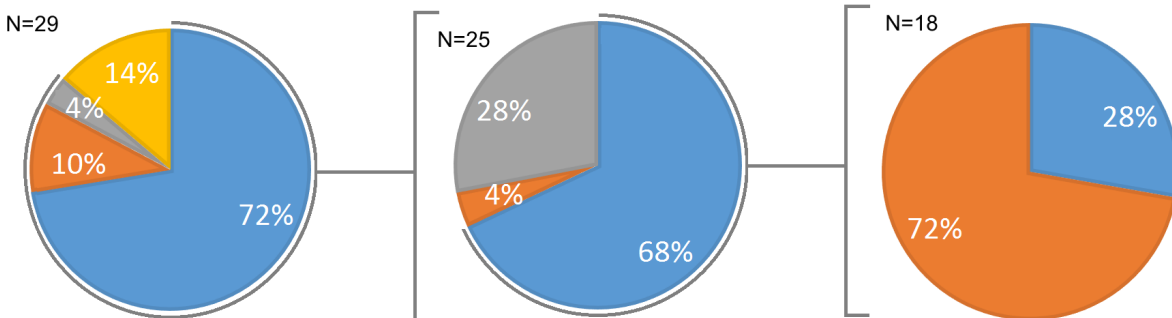
LIVER HEPATOCELLULAR CARCINOMA



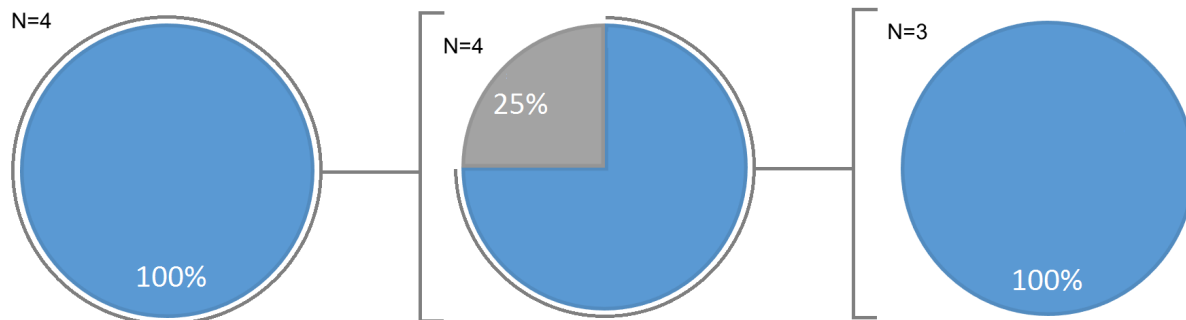
LUNG ADENOCARCINOMA



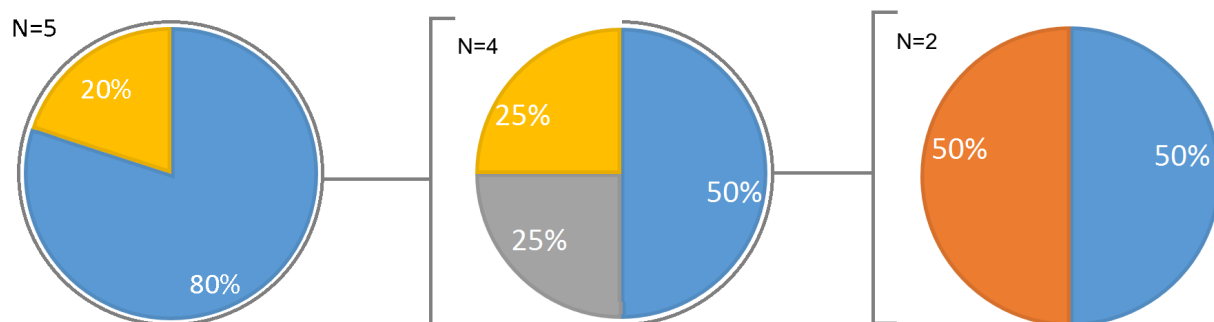
LUNG SQUAMOUS CELL CARCINOMA



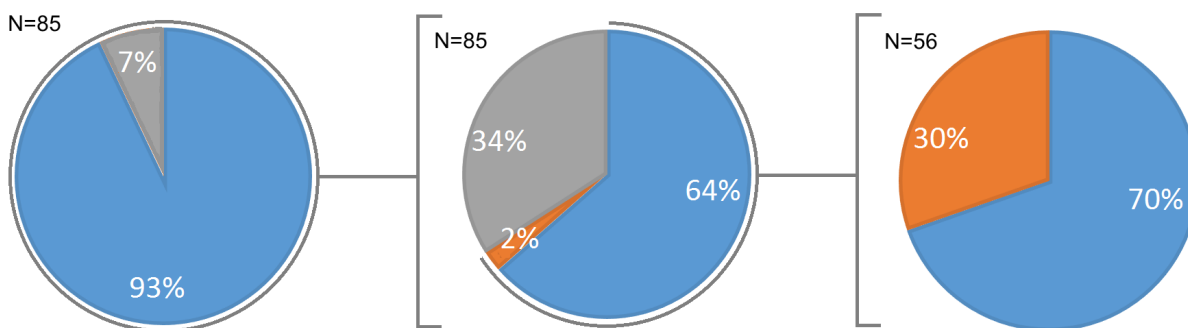
## PANCREATIC ADENOCARCINOMA



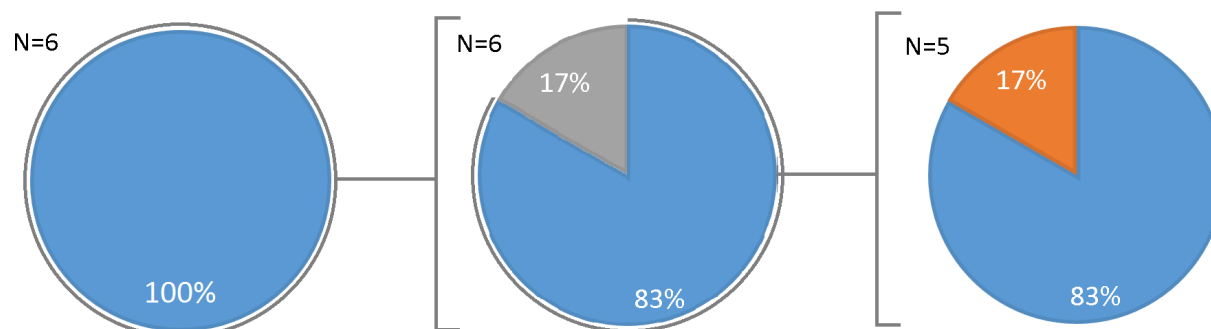
## PROSTATE ADENOCARCINOMA

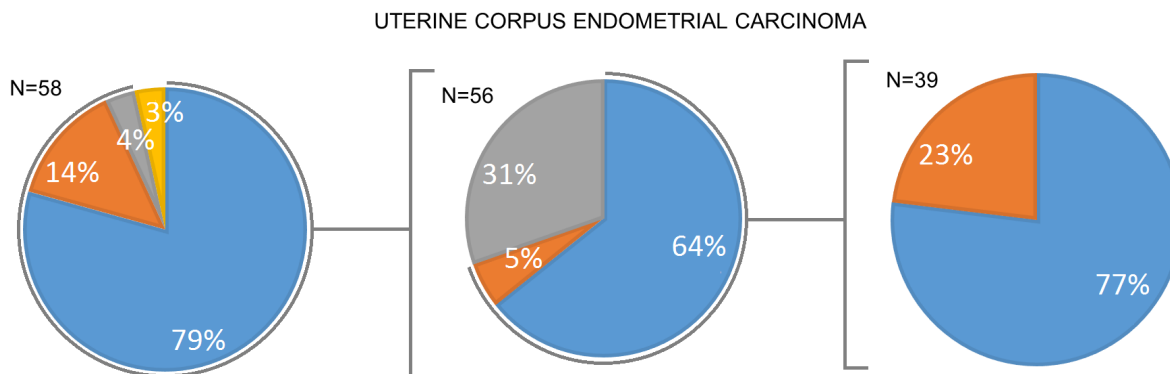


## STOMACH ADENOCARCINOMA



## THYROID CARCINOMA





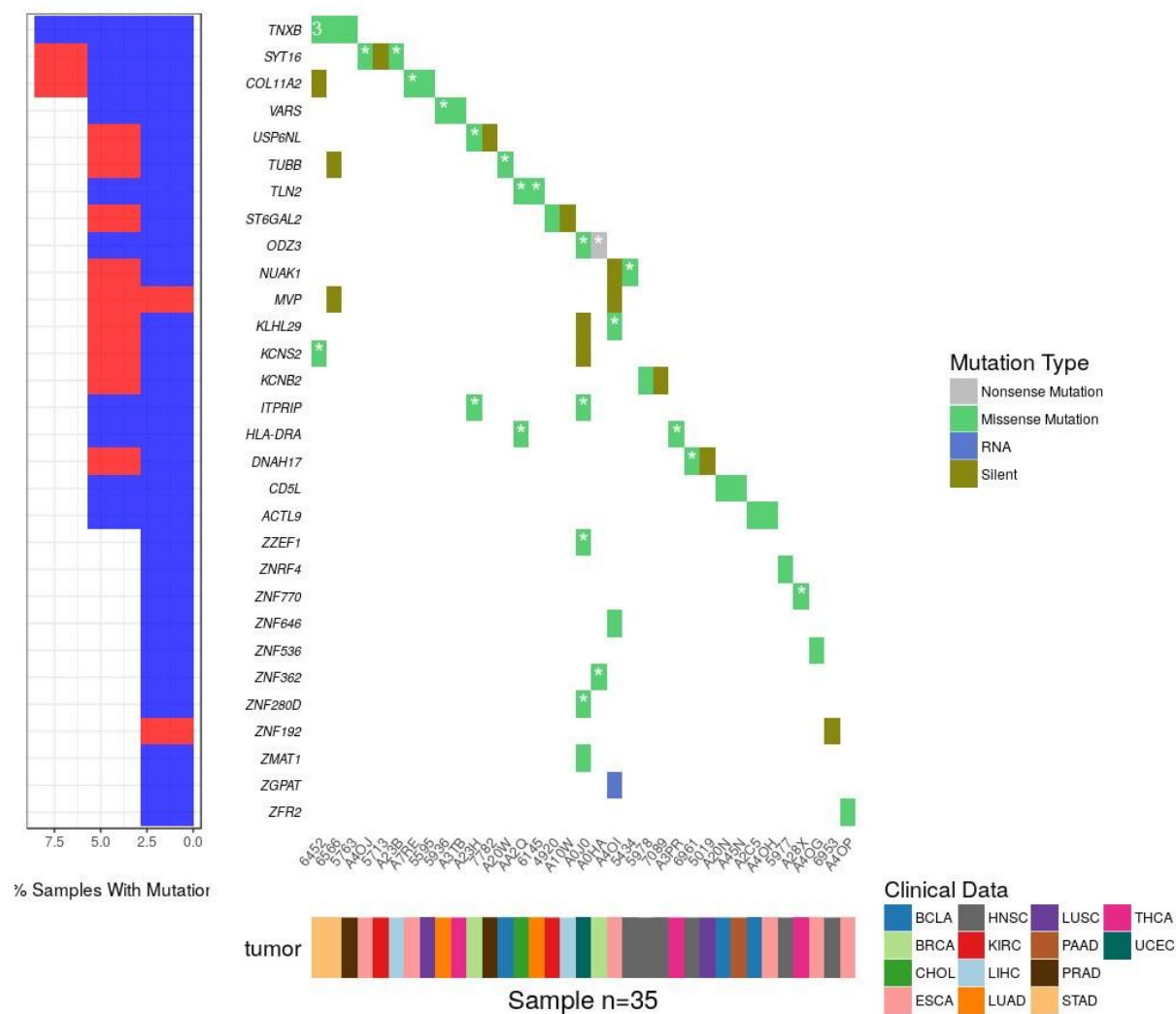
Fonte: Próprio autor.

Entretanto nessa etapa foi possível notar que em alguns sítios CpG onde ocorreram trocas C > T não continham informação sobre a metilação (beta value), assim essas trocas foram excluídas para o prosseguimento das análises (a totalidade das trocas que foram excluídas nessa etapa podem ser vistas na tabela 15 do material suplementar).

## 4.5 Mutações distribuídas pelos genes

A figura 16 sumariza a quantidade de mutações para os 30 genes com mais mutações levando em consideração mutações em diferentes amostras e tipos tumorais, identificando se essas mutações são patogênicas ou não.

**Figura 16.** Oncoprint da quantidade de mutações por gene Este oncoprint mostra a quantidade de mutações por gene, demonstrando também a classificação desta mutações e se são patogênicas (asterisco) segundo o UMD predictor para os 30 genes com mais mutações levando em consideração todos os tipos tumorais.



Fonte: Próprio autor.

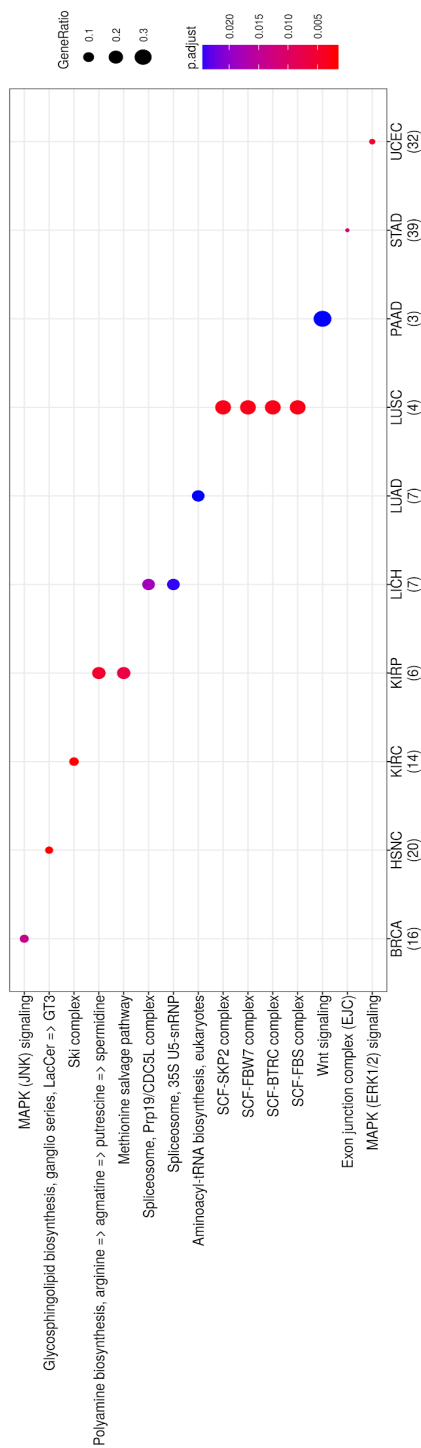
Apenas 20 genes sofreram mais de uma mutação, sendo em sua maioria 2 mutações em amostras diferentes, somente o gene TNXB sofreu 5 mutações sendo que 2 delas foram na mesma amostra e nenhuma delas foi classificada com patogênica segundo a predição do *UMD predictor*.



## 4.7 Enriquecimiento de vías

A figura 17 mostra que somente os tumores TCGA-BRCA, TCGA-HSNC, TCGA-KIRC, TCGA-KIRP, TCGA-LICH, TCGA-LUAD, TCGA-LUSC e TCGA-STAD tiveram vias enriquecidas com valor de  $p$  significativo. É possível notar que diferentes vias foram enriquecidas para cada um dos tumores em TCGA-BRCA a via enriquecida foi MAPK (JNK) signaling .Para TCGA-HSNC a via enriquecida foi a de biossíntese de gangliolipideos, para KIRC foi enriquecido o complexo Ski envolvido na detecção de RNAs, para KIRP foram enriquecidas as vias de biossíntese de poliamina e na via de resgate da metionina , para TCGA-LICH foram enriquecidas as vias spliceossomos Prp19/CDC5L e 35S U5-snRNP, para TCGA-LUAD foi enriquecida a via da aminoacil-tRNA sintetase, para LUSC foram enriquecidas as vias complexo contendo Skp, Cullin, F-box (ou complexo SCF ). Para TCGA-PAAD foi enriquecida a via de sinalização WNT. Para TCGA-STAD foi enriquecida a via de junção de exons e para TCGA-UCEC foi enriquecida via MAPK / ERK.

**Figura 17.** Vias enriquecidas segundo os genes que sofreram mutações em sítios CpG causadas por metilação separadas por tipo tumoral.



Fonte: Próprio autor.

## 5) DISCUSSÃO

Com a realização das análises iniciais em bioinformática foi possível ter um panorama das taxas de tipos de trocas para cada um dos tumores. Já era sabido e pôde ser comparado na literatura, uma maior taxa de trocas  $C > T$  em todo exoma especialmente em regiões de sítios CpG (Zhao, Zhongming. et al., 2006; Abdelfatah, Eihab et al., 2016). Vale destacar que, o único tipo tumoral que não se enquadra a essa regra foi o tumor de próstata, onde foi encontrada uma maior quantidade de trocas  $C > A$  ao invés do esperado  $C > T$ . Entretanto não observamos isso no trabalho de Santosh Yadav onde observaram justamente o contrário, uma maior proporção de trocas  $C > T$  em relação às demais (Yadav, Santosh et al., 2020). Esse resultado pode ter sido encontrado devido a baixa quantidade de trocas encontradas em sítios CpG desse tumor (total de 22 trocas), o que pode ter gerado um viés nessa situação em específico para termos uma proporção maior de  $C > A$ , mas como a quantidade de mutações é baixa talvez ao olharmos para as demais posições do exoma tenhamos um resultado diferente. É interessante notar, também, que a variação de metilação de amostras de tecidos normais para tecidos tumorais em sítios CpG onde houveram trocas  $C > T$  foi pequena, ou seja, na maioria dos casos a metilação se manteve constante em amostras de tecido normal e tumoral. Resultados que contrastam com os encontrados por Zhou, Dan e colaboradores que encontraram 53 mutações somáticas em sítios CpG para tumor de cólon e foi visto que os níveis de metilação eram menores 50 dos 53 em tumores em relação a amostras pareadas normais, sugerindo que essa diminuição foi causada pela perda da citosina (Zhaou, Dan., 2015).

Ao analisarmos os genes que mais sofreram mutações podemos destacar o gene *TNXB* que sofreu 5 mutações (nenhuma delas patogênica), esse gene codifica a proteína Tenascina-X, ela é uma glicoproteína de matriz extracelular associada à deposição de fibrilas de colágeno. Impactando a rigidez ou elasticidade de praticamente todas as células do corpo (Valcourt, Ulrich et al., 2015). Relacionado ao cancer de ovário (Kramer, Marianne et al., 2015). Marcador para mesotelioma (Yuan, Yuan et al., 2009). O segundo gene com maior numero de mutações é o gene *SYT16*, ocorreram 3 mutações nesse gene sendo que 2 delas são patogênicas, ou seja, podem inviabilizar a funcionalidade da proteína. Chen e colaboradores (2020) averiguaram que o aumento da expressão de *SYT16* se correlaciona significativamente com o grau de gliomas (Chen, Jianfeng et al., 2020), este gene codifica uma proteína que pode estar envolvido no tráfego e exocitose de vesículas secretoras em tecidos não neuronais. O gene *COL11A2* teve três mutações sendo que uma delas é patogênica, ele executa a função de fornecer instruções para a produção de um componente do colágeno do tipo XI denominado cadeia pró-alfa2 (XI) (Jakkula, Eveliina et al., 2020). O gene *VARS* sofreu 2 mutações sendo que uma delas é patogênica. Este gene codifica a proteína valil-tRNA sintetase ela pertence ao grupo das aminoacil-tRNA sintetases, elas catalisam a aminoacilação do tRNA por seu aminoácido cognato. Podemos notar também que esta proteína é expressa em câncer colorretal, mamario, de próstata, pulmão e fígado. (Bonfond, Luc et al., 2005).

Quando olhamos para as categorias a que os genes que sofreram mutações supostamente causadas pelo mecanismo de desaminação, podemos mostrar que alguns genes relacionados de alguma forma a categorias de processos biológicos que podem de alguma forma contribuir para a tumorigênese, podemos destacar: processos

metabólicos, crescimento e proliferação celular. Esses resultados sugerem que mutações em citosinas metiladas contribuem para a instabilidade genômica de células tumorais com implicações no funcionamento de genes associados com a tumorigênese.

Ao analisarmos as vias gênicas enriquecidas podemos ver que várias delas podem de alguma forma estar envolvidas com a progressão da tumorigênese, nos tumores de mama (TCGA-BRCA) e de ovário TCGA-UCEC a via enriquecida foi a MAPK (JNK) signaling uma das principais vias de sinalização da proteína-quinase responsável pelo controle de vários processos celulares, incluindo proliferação, desenvolvimento embrionário e apoptose (Keshet Y, Seger R, 2010), dessa forma A sinalização anormal de MAPK pode levar a proliferação celular aumentada ou descontrolada e resistência à apoptose, fatores preponderantes para o desenvolvimento e progressão do câncer, sendo que alterações nessa via já foram relatadas em câncer humano como resultado da ativação anormal de tirosina quinases receptoras ou mutações de ganho de função principalmente nos genes RAS ou RAF (Santarpia L et. al., 2012). Sendo possível também identificar evidências que sugerem que a modulação da atividade do MKP-1 pode ser uma opção viável para tornar a quimioterapia do câncer de mama mais eficaz. (Haagenson, Kelly K. et., al 2010). Liu e colaboradores (Liu, Ai et al., 2019) sugerem que as alterações nessa via possam estar relacionadas aos receptores de estrogênio, o que faz sentido que essa via seja enriquecida em câncer de mama e ovário. Em TCGA-KIRC foi enriquecida a via do complexo Ski que é uma proteína oncogênica que atua como um repressor de TGF- $\beta$  e impede a transcrição de genes relacionados. Esta proteína atuaria como uma oncoproteína no melanoma e no câncer de pancreático. (Heider, T. Ryan et al., 2007)

Evidências também sugerem que alterações na via MAPK (JNK) também estejam relacionadas com o câncer renal, assim como encontrado em nossos resultados. Ski aceleraria a progressão do câncer renal ao atenuar a sinalização do fator de crescimento transformador  $\beta$  (Taguchi, Luna et al., 2019.) Em TCGA-KIRP foram enriquecidas as vias de biossíntese de poliamina essas moléculas estão envolvidas em muitos processos fundamentais de crescimento e sobrevivência celular. No câncer, o metabolismo da poliamina é frequentemente desregulado, indicando de maneira geral que níveis elevados de poliamida são necessários para a transformação e progressão do tumor (Casero, Robert A, et al., 2018; Murray-Stewart, et al., 2016). Outra via enriquecida para o tumor TCGA-KIRP foi a de resgate da metionina , onde a dependência das células cancerosas da metionina exógena. As linhas celulares não tumorigênicas têm a mesma taxa de proliferação em meios contendo metionina ou meios em que a metionina é substituída pelo precursor metabólico imediato homocisteína, ocorrendo em diversos tumores entre eles o rim (Stern et al., 1984)

Para TCGA-LUAD foi enriquecida a via da aminoacil-tRNA sintetase esta via forma um complexo de proteínas macromoleculares com três fatores auxiliares, designados proteína multifuncional associadas ao câncer (Kim, Sunghoon, et al., 2011). Para TCGA-LUSC foram enriquecidas as vias do complexo SCF este complexo tem papéis importantes na ubiquitinação de proteínas envolvidas no ciclo celular. (Ou, Young; Rattner, J. B., 2004). É interessante destacar que os complexos SCF se tornaram um alvo anticâncer atraente por causa de sua regulação positiva em alguns cânceres humanos e seus locais ativos bioquimicamente distintos.(Skaar, et al., 2014). Em TCGA-PAAD a via de sinalização Wnt foi enriquecida (Komiya, Y., Habas, R. 2008), essa via regula diversos fenômenos e eventos durante desenvolvimento embrionário,

com a organogênese, diferenciação, polarização e migração celular - recentemente a via da Wnt foi relacionada a renovação de células-tronco na carcinogênese foi descrito de forma mais proeminente para o câncer colorretal. (Zhan, T et al., 2017).

## 6) CONCLUSÃO

Com esse trabalho foi possível constatar que as taxas de trocas C > T são significativamente maiores do que os demais tipos de trocas como já relatado em literatura. Ao analisarmos as trocas que possivelmente foram causadas pelo mecanismo de desaminação, percebemos o impacto, principalmente funcional, que esse mecanismo pode desencadear, ao notar que grande parte das vias gênicas afetadas por trocas patogênicas estão diretamente relacionadas ao câncer. Sendo assim, existem evidências que esse mecanismo impacta no desenvolvimento da tumorigênese.



## 7) REFERÊNCIAS

1. ABDELFAH, Eihab et al. Epigenetic therapy in gastrointestinal cancer: the right combination. **Therapeutic advances in gastroenterology**, v. 9, n. 4, p. 560-579, 2016.
2. BIRD, Adrian P. CpG-rich islands and the function of DNA methylation. **Nature**, v. 321, n. 6067, p. 209-213, 1986.
3. BONNEFOND, Luc et al. Toward the full set of human mitochondrial aminoacyl-tRNA synthetases: characterization of AspRS and TyrRS. **Biochemistry**, v. 44, n. 12, p. 4805-4816, 2005.
4. CASERO, Robert A.; STEWART, Tracy Murray; PEGG, Anthony E. Polyamine metabolism and cancer: treatments, challenges and opportunities. **Nature Reviews Cancer**, v. 18, n. 11, p. 681-695, 2018.
5. CHEN, Jianfeng et al. SYT16 is a prognostic biomarker and correlated with immune infiltrates in glioma: A study based on TCGA data. **International Immunopharmacology**, v. 84, p. 106490, 2020.
6. CLARK, Susan J. et al. DNA methylation: bisulphite modification and analysis. **Nature protocols**, v. 1, n. 5, p. 2353, 2006.
7. COLAPRICO, Antonio et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. **Nucleic acids research**, v. 44, n. 8, p. e71-e71, 2016.
8. DEATON, Aimée M.; BIRD, Adrian. CpG islands and the regulation of transcription. **Genes & development**, v. 25, n. 10, p. 1010-1022, 2011.
9. FELSENFELD, Gary. A brief history of epigenetics. **Cold Spring Harbor perspectives in biology**, v. 6, n. 1, p. a018200, 2014.
10. GOLL, Mary Grace; BESTOR, Timothy H. Eukaryotic cytosine methyltransferases. **Annu. Rev. Biochem.**, v. 74, p. 481-514, 2005.
11. HAAGENSON, Kelly K.; WU, Gen Sheng. The role of MAP kinases and MAP kinase phosphatase-1 in resistance to breast cancer treatment. **Cancer and Metastasis Reviews**, v. 29, n. 1, p. 143-149, 2010.
12. HEIDER, T. Ryan et al. Ski promotes tumor growth through abrogation of transforming growth factor- $\beta$  signaling in pancreatic cancer. **Annals of surgery**, v. 246, n. 1, p. 61, 2007.

13. INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM et al. Initial sequencing and analysis of the human genome. **Nature**, v. 409, p. 860-921, 2001.
14. ILLINGWORTH, Robert S.; BIRD, Adrian P. CpG islands—'a rough guide'. **FEBS letters**, v. 583, n. 11, p. 1713-1720, 2009.
15. JAENISCH, Rudolf; BIRD, Adrian. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. **Nature genetics**, v. 33, n. 3, p. 245-254, 2003.
16. JAKKULA, Eveliina et al. The role of sequence variations within the genes encoding collagen II, IX and XI in non-syndromic, early-onset osteoarthritis. **Osteoarthritis and cartilage**, v. 13, n. 6, p. 497-507, 2005.
17. JONES, Peter A.; BAYLIN, Stephen B. The fundamental role of epigenetic events in cancer. **Nature reviews genetics**, v. 3, n. 6, p. 415-428, 2002.
18. KESHET, Yonat; SEGER, Rony. The MAP kinase signaling cascades: a system of hundreds of components regulates a diverse array of physiological functions. In: **MAP kinase signaling protocols**. Humana Press, Totowa, NJ, 2010. p. 3-38.
19. KIM, Sunghoon; YOU, Sungyong; HWANG, Daehee. Aminoacyl-tRNA synthetases and tumorigenesis: more than housekeeping. **Nature Reviews Cancer**, v. 11, n. 10, p. 708-718, 2011.
20. KIM, Somi; KAANG, Bong-Kiun. Epigenetic regulation and chromatin remodeling in learning and memory. **Experimental & molecular medicine**, v. 49, n. 1, p. e281-e281, 2017.
21. KOMIYA, Y.; HABAS, R. Wnt signal transduction pathways *Organogenesis* 4: 68-75. 2008.
22. KRAMER, Marianne et al. Secretome identifies tenascin-X as a potent marker of ovarian cancer. **BioMed research international**, v. 2015, 2015.
23. LAIRD, Charles D. et al. Hairpin-bisulfite PCR: assessing epigenetic methylation patterns on complementary strands of individual DNA molecules. **Proceedings of the National Academy of Sciences**, v. 101, n. 1, p. 204-209, 2004.
24. LANDER, Eric S. et al. Initial sequencing and analysis of the human genome. 2001.
25. LIU, Ai et al. Estrogen receptor alpha activates MAPK signaling pathway to promote the development of endometrial cancer. **Journal of cellular biochemistry**, v. 120, n. 10, p. 17593-17601, 2019.
26. LONG, Mark D.; SMIRAGLIA, Dominic J.; CAMPBELL, Moray J. The genomic impact of DNA CpG methylation on gene expression; relationships in prostate cancer. **Biomolecules**, v. 7, n. 1, p. 15, 2017.

27. LÖVKVIST, Cecilia et al. DNA methylation in human epigenomes depends on local topology of CpG sites. **Nucleic acids research**, v. 44, n. 11, p. 5123-5132, 2016.
28. MURRAY-STEWART, Tracy R.; WOSTER, Patrick M.; CASERO JR, Robert A. Targeting polyamine metabolism for cancer therapy and prevention. **Biochemical Journal**, v. 473, n. 19, p. 2937-2953, 2016.
29. NEPHEW, Kenneth P.; HUANG, Tim Hui-Ming. Epigenetic gene silencing in cancer initiation and progression. **Cancer letters**, v. 190, n. 2, p. 125-133, 2003.
30. OU, Young; RATTNER, J. B. The centrosome in higher organisms: structure, composition, and duplication. **International review of cytology**, v. 238, p. 119-182, 2004.
31. PALERO, FERRAN; CRANDALL, KEITH A. Phylogenetic inference using molecular data. In: Decapod Crustacean Phylogenetics. CRC Press, 2016. p. 79-100.
32. RIDEOUT, WM 3rd et al. 5-Methylcytosine as an endogenous mutagen in the human LDL receptor and p53 genes. **Science**, v. 249, n. 4974, p. 1288-1290, 1990.
33. SANTARPIA, Libero; LIPPMAN, Scott M.; EL-NAGGAR, Adel K. Targeting the MAPK–RAS–RAF signaling pathway in cancer therapy. **Expert opinion on therapeutic targets**, v. 16, n. 1, p. 103-119, 2012.
34. SALGADO, David et al. UMD-predictor: a high-throughput sequencing compliant system for pathogenicity prediction of any human cDNA substitution. **Human mutation**, v. 37, n. 5, p. 439-446, 2016.
35. SHIAH, Yu-Jia et al. Comparison of pre-processing methods for Infinium HumanMethylation450 BeadChip array. **Bioinformatics**, v. 33, n. 20, p. 3151-3157, 2017.
36. SKAAR, Jeffrey R.; PAGAN, Julia K.; PAGANO, Michele. SCF ubiquitin ligase-targeted therapies. **Nature reviews Drug discovery**, v. 13, n. 12, p. 889-903, 2014.
37. STERN, Peter H.; WALLACE, C. Douglas; HOFFMAN, Robert M. Altered methionine metabolism occurs in all members of a set of diverse human tumor cell lines. **Journal of cellular physiology**, v. 119, n. 1, p. 29-34, 1984.
38. SUZUKI, Miho M.; BIRD, Adrian. DNA methylation landscapes: provocative insights from epigenomics. **Nature Reviews Genetics**, v. 9, n. 6, p. 465-476, 2008.
39. SASSA, Akira et al. Mutagenic consequences of cytosine alterations site-specifically embedded in the human genome. **Genes and Environment**, v. 38, n. 1, p. 17, 2016.
40. TAGUCHI, Luna et al. c-Ski accelerates renal cancer progression by attenuating transforming growth factor  $\beta$  signaling. **Cancer science**, v. 110, n. 6, p. 2063, 2019.

41. TIRADO-MAGALLANES, Roberto et al. Whole genome DNA methylation: beyond genes silencing. **Oncotarget**, v. 8, n. 3, p. 5629, 2017.
42. VALCOURT, Ulrich et al. Tenascin-X: beyond the architectural function. **Cell adhesion & migration**, v. 9, n. 1-2, p. 154-165, 2015.
43. WANG, Jing et al. WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. **Nucleic acids research**, v. 45, n. W1, p. W130-W137, 2017.
44. WARDEN, Charles D. et al. COHCAP: an integrative genomic pipeline for single-nucleotide resolution DNA methylation analysis. **Nucleic acids research**, v. 41, n. 11, p. e117-e117, 2013.
45. WATSON, James D. et al. **Biologia molecular do gene**. Artmed Editora, 2015.
46. WEINSTEIN, John N. et al. The cancer genome atlas pan-cancer analysis project. **Nature genetics**, v. 45, n. 10, p. 1113, 2013.
47. WEISENBERGER, Daniel J. et al. CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer. **Nature genetics**, v. 38, n. 7, p. 787-793, 2006.
48. WONG, David J. et al. p16INK4a lesions are common, early abnormalities that undergo clonal expansion in Barrett's metaplastic epithelium. **Cancer research**, v. 61, n. 22, p. 8284-8289, 2001.
49. YADAV, Santosh et al. Somatic mutations in the DNA repairome in prostate cancers in African Americans and Caucasians. **Oncogene**, v. 39, n. 21, p. 4299-4311, 2020.
50. YUAN, Yuan et al. Tenascin-X is a novel diagnostic marker of malignant mesothelioma. **The American journal of surgical pathology**, v. 33, n. 11, p. 1673, 2009.
51. YU, Guangchuang et al. clusterProfiler: an R package for comparing biological themes among gene clusters. **Omics: a journal of integrative biology**, v. 16, n. 5, p. 284-287, 2012.
52. ZHAN, T.; RINDTORFF, N.; BOUTROS, Michael. Wnt signaling in cancer. **Oncogene**, v. 36, n. 11, p. 1461-1473, 2017.
53. ZHAO, Zhongming; ZHANG, Fengkai. Sequence context analysis of 8.2 million single nucleotide polymorphisms in the human genome. **Gene**, v. 366, n. 2, p. 316-324, 2006.
54. ZHOU, Dan et al. Polymorphisms involving gain or loss of CpG sites are significantly enriched in trait-associated SNPs. **Oncotarget**, v. 6, n. 37, p. 39995, 2015.

## 8) DADOS SUPLEMENTARES

### 8.1 Tabelas

**Tabela 3.** Tabela contendo os valores absolutos da quantidade de cada tipo de troca para cada tumor nos no arquivo .maf

Quantidade de cada uma das trocas por tumor														
Tumor	c > t	c > g	c > a	t > c	t > g	t > a	a > t	a > g	a > c	g > t	g > a	g > c	total trocas	Quantidade amostras
TCGA-BLCA	850	479	141	121	32	54	51	106	52	165	823	426	3300	15
TCGA-BRCA	2182	1395	572	320	160	153	157	307	179	554	2097	1343	10008	89
TCGA-CESC	33	5	6	6	2	0	1	1	2	5	26	4	91	3
TCGA-CHOL	242	40	113	65	16	25	25	46	19	334	209	33	1511	9
TCGA-ESCA	1741	203	588	424	313	159	155	417	289	181 3	1695	218	8015	16
TCGA-GBMLGG	18	2	2	4	1	0	0	2	1	3	20	3	61	1
TCGA-HNSC	1892	758	667	436	150	246	260	409	118	689	1887	700	8212	50
TCGA-KIPAN	1388	513	886	730	342	483	518	730	382	876	1441	518	8807	143
TCGA-KIPAN	523	227	250	341	122	153	143	355	123	234	546	223	3720	43
TCGA-LIHC	990	235	636	100 5	133	206	200	1001	123	632	905	209	6275	47
TCGA-LUAD	747	341	1107	236	95	315	253	261	85	1151	762	379	5732	20
TCGA-LUSC	1142	600	1703	391	102	370	366	390	122	166 1	1175	589	8611	21

TCGA -PAAD	56	6	65	6	3	5	2	5	4	42	54	9	257	4
TCGA -PCP G	21	2	6	12	2	0	1	11	1	4	14	8	82	3
TCGA -PRA D	267	68	991	155	46	55	53	121	61	466	248	70	2601	43
TCGA -REA D	582	20	73	65	23	25	17	73	24	70	584	21	1577	5
TCGA -SAR C	46	18	21	10	8	15	11	11	5	28	55	11	239	4
TCGA -STAD	1936	82	511	157 3	186	638	378	1511	193	461	2111	50	9324	2
TCGA -THCA	233	66	89	131	11	27	62	87	20	58	371	86	1241	2
TCGA -THY M	2	0	0	1	1	0	0	2	0	3	2	0	11	2
TCGA -UCE C	1525	20	1784	495	714	54	50	446	670	173 4	1579	29	9100	5
total	1641 6	5080	1021 1	652 7	246 2	298 3	2703	6292	2473	109 83	1660 4	4929	88775	527

**Tabela 4.** Tabela contendo a soma das trocas complementares para cada um dos tumores.

<b>Contagem de cada uma das trocas por tumor (soma das trocas complementares)</b>							
	<b>c&gt;t</b>	<b>c&gt;g</b>	<b>c&gt;a</b>	<b>t&gt;c</b>	<b>t&gt;g</b>	<b>t&gt;a</b>	<b>total</b>
<b>TCGA-BLCA</b>	1673	905	306	227	84	105	3300
<b>TCGA-BRCA</b>	4279	2738	1126	627	339	310	9419
<b>TCGA-CESC</b>	59	9	11	7	4	1	91
<b>TCGA-CHOL</b>	451	73	447	111	35	50	1167
<b>TCGA-ESCA</b>	3436	421	2401	841	602	314	8015
<b>TCGA-GBM</b>	38	5	5	6	2	0	56
<b>TCGA-HNSC</b>	3779	1458	1356	845	268	506	8212
<b>TCGA-KIRC</b>	2829	1031	1762	1460	724	1001	8807
<b>TCGA-KIRP</b>	1069	450	484	696	245	296	3240
<b>TCGA-LIHC</b>	1895	444	1268	2006	256	406	6275
<b>TCGA-LUAD</b>	1509	720	2258	497	180	568	5732
<b>TCGA-LUSC</b>	2317	1189	3364	781	224	736	8611
<b>TCGA-PAAD</b>	110	15	107	11	7	7	257
<b>TCGA-PCPG</b>	35	10	10	23	3	1	82
<b>TCGA-PRAD</b>	515	138	1457	276	107	108	2601
<b>TCGA-READ</b>	1166	41	143	138	47	42	1577
<b>TCGA-SARC</b>	101	29	49	21	13	26	239
<b>TCGA-STAD</b>	4047	132	972	3084	379	1016	9630
<b>TCGA-THCA</b>	604	152	147	218	31	89	1241
<b>TCGA-THYM</b>	4	0	3	3	1	0	11
<b>TCGA-UCEC</b>	3104	49	3518	941	1384	104	9100
<b>total</b>	33020	10009	21194	12819	4935	5686	87663

**Tabela 5.** Tabela contendo a contagem total de trocas em sítios CpG para cada tumor.

Contagem total de trocas em sítios CpG por tumor								
Tumor	c > t	c > g	c > a	g > t	g > a	g > c	total trocas	Quantidade amostras
TCGA-BL CA	12	2	1	2	9	3	29	13
TCGA-BR CA	18	5	3	3	18	5	52	24
TCGA-CE SC	0	0	0	0	1	0	1	1
TCGA-CH OL	6	0	0	3	2	0	11	9
TCGA-ES CA	33	0	3	5	44	0	85	14
TCGA-GB M	0	1	0	0	0	0	1	1
TCGA-HN SC	30	4	5	5	29	3	76	42
TCGA-KIR C	16	2	2	3	17	5	45	37
TCGA-KIR P	6	2	2	2	6	5	23	19
TCGA-LIH C	12	3	2	7	10	1	35	26
TCGA-LU AD	12	1	5	6	11	2	37	15
TCGA-LU SC	9	3	14	13	20	10	69	18
TCGA-PA AD	2	0	0	1	2	0	5	4
TCGA-PR AD	4	0	12	5	1	0	22	9
TCGA-ST AD	33	0	1	1	56	0	91	2
TCGA-TH	2	0	0	0	5	1	8	6



CA								
TCGA-UC EC	40	0	0	1	21	0	62	6

**Tabela 6.** Tabela contendo a soma das trocas complementares em sítios CpG para cada um dos tumores.

<b>Contagem de cada uma das trocas em sítios CpG por tumor (soma das trocas complementares)</b>				
	c>t	c>g	c>a	total
TCGA-BLCA	21	5	3	29
TCGA-BRCA	36	10	6	52
TCGA-CHOL	8	0	3	11
TCGA-ESCA	77	0	8	85
TCGA-HNSC	59	7	10	76
TCGA-KIRC	33	7	5	45
TCGA-KIRP	12	7	4	23
TCGA-LIHC	22	4	9	35
TCGA-LUAD	23	3	11	37
TCGA-LUSC	29	13	27	69
TCGA-PAAD	4	0	1	5
TCGA-PRAD	5	0	17	22
TCGA-STAD	89	0	2	91
TCGA-THCA	7	1	0	8
TCGA-UCEC	61	0	1	62
total	487	58	107	652

**Tabela 7.** Tabela contendo as trocas em sítios CpG metilados para cada um dos tecidos em amostras normais.

Contagem total de trocas em sítios CpG metilados (amostras normais)								
Tumor	c > t	c > g	c > a	g > t	g > a	g > c	total trocas	Quantidade amostras
TCGA-BL CA	12	1	1	1	6	2	23	11
TCGA-BR CA	14	2	2	3	14	0	35	20
TCGA-C HOL	6	0	0	3	2	0	11	8
TCGA-ES CA	26	0	2	5	39	0	72	14
TCGA-H NSC	23	4	2	4	27	2	62	36
TCGA-KI RC	15	1	1	2	15	4	38	32
TCGA-KI RP	5	1	2	2	6	4	20	17
TCGA-LI HC	11	3	1	5	9	0	29	21
TCGA-LU AD	10	1	5	6	9	1	32	12
TCGA-LU SC	6	1	11	13	18	6	55	16
TCGA-PA AD	2	0	0	1	2	0	5	4
TCGA-PR AD	3	0	10	5	1	0	19	6
TCGA-ST AD	29	0	1	1	50	0	81	2
TCGA-TH CA	1	0	0	0	5	1	7	5
TCGA-U CEC	34	0	0	1	19	0	54	5

<b>Total</b>	197	14	38	52	222	20	543	
--------------	-----	----	----	----	-----	----	-----	--

**Tabela 8.** Tabela contendo a soma das trocas complementares em sítios CpG metilados para cada um dos tecidos em amostras normais.

<b>Contagem total de trocas em sítios CpG metilados (amostras normais complementares somados)</b>					
<b>Tumor</b>	<b>c&gt;t</b>	<b>c&gt;g</b>	<b>c&gt;a</b>	<b>total</b>	
TCGA-BLCA	18	3	2	23	
TCGA-BRCA	28	2	5	35	
TCGA-CHOL	8	0	3	11	
TCGA-ESCA	63	0	7	70	
TCGA-HNSC	50	6	6	62	
TCGA-KIRC	30	5	3	38	
TCGA-KIRP	11	5	4	20	
TCGA-LIHC	20	3	6	29	
TCGA-LUAD	19	2	11	32	
TCGA-LUSC	24	7	24	55	
TCGA-PAAD	4	0	1	5	
TCGA-PRAD	4	0	15	19	
TCGA-STAD	79	0	2	81	
TCGA-THCA	6	1	0	7	
TCGA-UCEC	53	0	1	54	
<b>total</b>	<b>417</b>	<b>35</b>	<b>90</b>	<b>542</b>	

**Tabela 9.** Tabela contendo trocas em sítios CpG não metilados para cada um dos tecidos em amostras normais.

<b>Contagem total de trocas em sítios CpG não metilados (amostras normais)</b>								
<b>Tumor</b>	<b>c &gt; t</b>	<b>c &gt; g</b>	<b>c &gt; a</b>	<b>g &gt; t</b>	<b>g &gt; a</b>	<b>g &gt; c</b>	<b>total trocas</b>	<b>Quantidade amostras</b>
TCGA-BLCA	0	1	0	1	2	1	5	5
TCGA-BRCA	4	3	1	0	4	3	15	8
TCGA-CHOL	0	0	0	0	0	0	0	0
TCGA-ESCA	8	0	1	0	6	0	15	9
TCGA-HNSC	7	0	3	1	2	1	14	14
TCGA-KIRC	1	1	1	1	2	1	7	6
TCGA-KIRP	1	1	0	0	0	1	3	3
TCGA-LIHC	1	0	1	2	1	1	6	6
TCGA-LUAD	2	0	0	0	2	1	5	5
TCGA-LUSC	3	2	3	0	2	4	14	11
TCGA-PAAD	0	0	0	0	0	0	0	0
TCGA-PRAD	1	0	2	0	0	0	3	3
TCGA-STAD	4	0	0	0	6	0	10	2
TCGA-THCA	0	1	0	0	0	0	1	2
TCGA-UCEC	6	0	0	0	2	0	8	1

total	38	9	12	5	30	13	107	
-------	----	---	----	---	----	----	-----	--

**Tabela 10.** Tabela contendo a soma das trocas complementares em sítios CpG não metilados para cada um dos tecidos em amostras normais.

<b>Quantidade de trocas em sítios CpG não metilados (somando as complementares )</b>				
<b>Tumor</b>	<b>c &gt; t</b>	<b>c &gt; g</b>	<b>c &gt; a</b>	<b>total</b>
TCGA-BLCA	2	2	1	5
TCGA-BRCA	8	6	1	15
TCGA-CHOL	0	0	0	0
TCGA-ESCA	9	0	1	10
TCGA-HNSC	9	1	4	14
TCGA-KIRC	3	2	2	7
TCGA-KIRP	1	2	0	3
TCGA-LIHC	2	1	3	6
TCGA-LUAD	3	1	0	4
TCGA-LUSC	5	6	3	14
TCGA-PAAD	0	0	0	0
TCGA-PREAD	1	0	2	3
TCGA-STAD	10	0	0	10
TCGA-THCA	0	1	0	1
TCGA-UCEC	8	0	0	8
total	68	22	17	107

**Tabela 11.** Tabela contendo as trocas em sítios CpG metilados para cada um dos tecidos em amostras tumorais.

<b>Contagem total de trocas em sítios CpG metilados (amostras tumorais)</b>								
<b>Tumor</b>	<b>c &gt; t</b>	<b>c &gt; g</b>	<b>c &gt; a</b>	<b>g &gt; t</b>	<b>g &gt; a</b>	<b>g &gt; c</b>	<b>total trocas</b>	<b>Quantidade amostras</b>
TCGA-BLCA	12	2	0	1	6	2	21	11
TCGA-BRCA	16	1	2	3	14	0	36	21
TCGA-CHOL	6	0	0	3	2	0	12	8
TCGA-ESCA	30	0	2	5	39	0	76	14
TCGA-HNSC	23	4	4	4	28	2	65	35
TCGA-KIRC	15	2	1	2	14	4	38	32
TCGA-KIRP	5	1	2	2	6	4	20	17
TCGA-LIHC	10	3	1	4	8	0	26	19
TCGA-LUAD	10	1	5	6	10	1	33	12
TCGA-LUSC	6	2	10	13	16	6	53	17
TCGA-PAAD	2	0	0	1	2	0	5	4
TCGA-PRAD	3	0	10	5	1	0	19	6
TCGA-STAD	32	0	1	1	53	0	87	2
TCGA-THCA	6	0	0	0	4	1	6	5
TCGA-UCEC	30	0	0	0	18	0	48	4
<b>Total</b>	<b>206</b>	<b>17</b>	<b>38</b>	<b>50</b>	<b>221</b>	<b>20</b>		

**Tabela 12.** Tabela contendo a soma das trocas complementares em sítios CpG metilados para cada um dos tecidos em amostras tumorais.

<b>Contagem total de trocas em sítios CpG metilados (amostras tumorais)</b>				
<b>Tumor</b>	<b>c &gt; t</b>	<b>c &gt; g</b>	<b>c &gt; a</b>	<b>total</b>
TCGA-BLCA	20	4	1	25
TCGA-BRCA	30	1	5	36
TCGA-CHOL	8	0	3	11
TCGA-ESCA	69	0	7	76
TCGA-HNSC	51	6	8	65
TCGA-KIRC	29	6	3	38
TCGA-KIRP	11	5	4	20
TCGA-LIHC	18	3	5	26
TCGA-LUAD	20	2	11	33
TCGA-LUSC	22	8	23	53
TCGA-PAAD	4	0	1	5
TCGA-PRAD	4	0	15	19
TCGA-STAD	85	0	2	87
TCGA-THCA	6	1	0	7
TCGA-UCEC	48	0	48	48
<b>total</b>	<b>425</b>	<b>37</b>	<b>136</b>	



**Tabela 13.** Tabela contendo as trocas em sítios CpG não metilados para cada um dos tecidos em amostras tumorais.

quantidade de trocas em sítios CpG não metilados (amostras tumorais)								
Tumor	c > t	c > g	c > a	g > t	g > a	g > c	total trocas	Quantidade amostras
TCGA-BLCA	0	0	1	1	0	1	3	3
TCGA-BRCA	2	4	1	0	4	3	14	7
TCGA-CHOL	0	0	0	0	0	0	0	0
TCGA-ESCA	3	0	0	0	5	0	8	7
TCGA-HNSC	7	0	1	1	1	1	11	11
TCGA-KIRC	1	0	1	1	3	1	7	6
TCGA-KIRP	1	1	0	0	0	1	3	3
TCGA-LIHC	2	0	1	3	2	1	9	9
TCGA-LUAD	2	0	0	0	1	1	4	4
TCGA-LUSC	3	1	4	0	4	4	16	10
TCGA-PAAD	0	0	0	0	0	0	0	0
TCGA-PRAD	1	0	2	0	0	0	3	3
TCGA-STAD	1	0	0	0	3	0	4	2
TCGA-THCA	0	1	0	0	0	0	1	2
TCGA-UCEC	10	0	0	1	3	0	14	3
<b>Total</b>	33	7	11	7	27	13	98	

**Tabela 14.** Tabela contendo a soma das trocas complementares em sítios CpG não metilados para cada um dos tecidos em amostras tumorais.

<b>quantidade de trocas em sítios CpG nao metilados (somando as complementares, amostras tumorais)</b>				
<b>Tumor</b>	<b>c &gt; t</b>	<b>c &gt; g</b>	<b>c &gt; a</b>	<b>total</b>
TCGA-BLCA	0	1	2	3
TCGA-BRCA	6	7	1	14
TCGA-CHOL	0	0	0	0
TCGA-ESCA	3	0	0	3
TCGA-HNSC	8	1	2	11
TCGA-KIRC	4	1	2	7
TCGA-KIRP	1	2	0	3
TCGA-LIHC	4	1	4	9
TCGA-LUAD	2	1	0	3
TCGA-LUSC	7	5	4	16
TCGA-PAAD	0	0	0	0
TCGA-PRAID	1	0	2	3
TCGA-STAD	4	0	0	4
TCGA-THCA	0	1	0	1
TCGA-UCEC	13	0	1	14
<b>Total</b>	<b>54</b>	<b>20</b>	<b>18</b>	<b>92</b>

**Tabela 15.** Tabela que relaciona os dados metilação entre amostras normais e tumorais.

<b>Comparação metilação amostras normais x amostras tumorais</b>					
<b>Tumor</b>	<b>metilado normal e metilado tumoral</b>	<b>metilado normal x nao metilado tumor</b>	<b>nao metilado normal x metilado tumor</b>	<b>nao metilado normal x nao metilado tumor</b>	<b>sem informacao de beta_value (excluidas)</b>
<b>BLCA</b>	18	0	2	0	1
<b>BRCA</b>	28	0	2	6	0
<b>CHOL</b>	8	0	0	0	0
<b>ESCA</b>	62	1	7	2	5
<b>HNSC</b>	49	1	2	3	4
<b>KIRC</b>	29	1	0	0	3
<b>KIRP</b>	11	0	0	0	1
<b>LIHC</b>	18	2	0	1	1
<b>LUAD</b>	19	0	1	2	1
<b>LUSC</b>	21	3	1	4	1
<b>PAAD</b>	4	0	0	0	0
<b>PRAD</b>	4	0	0	1	0
<b>STAD</b>	79	0	6	0	4
<b>THCA</b>	6	0	0	0	1
<b>UCEC</b>	46	8	2	2	3

**Tabela 16.** Tabela contendo a contagem dos tipos de mutações.

Tipos de mutações para Trocas em sitios CpG metilados				
Tumor	Missense	nonsense	silent	RNA
TCGA-BLCA	10	0	10	0
TCGA-BRCA	17	3	7	3
TCGA-CHOL	5	1	2	0
TCGA-ESCA	39	1	21	9
TCGA-HNSC	33	0	19	0
TCGA-KIRC	17	2	10	1
TCGA-KIRP	9	0	2	0
TCGA-LIHC	14	0	6	0
TCGA-LUAD	15	0	5	0
TCGA-LUSC	17	1	7	0
TCGA-PAAD	3	0	1	0
TCGA-PRA D	2	0	1	1
TCGA-STAD	54	2	29	0
TCGA-THCA	5	0	1	0
TCGA-UCE C	36	3	17	0
total	276	13	138	14

**Tabela 17.** Tabela contendo contagem de trocas patogênicas e não patogênicas.

<b>Impacto das mutações</b>				
<b>missense</b>	<b>patogenico</b>	<b>não patogenico</b>	<b>nonsense</b>	<b>patogenico</b>
<b>TCGA-BLCA</b>	4	6	<b>TCGA-BLCA</b>	0
<b>TCGA-BRCA</b>	10	7	<b>TCGA-BRCA</b>	3
<b>TCGA-CHOL</b>	5	0	<b>TCGA-CHOL</b>	1
<b>TCGA-ESCA</b>	22	17	<b>TCGA-ESCA</b>	1
<b>TCGA-HNSC</b>	19	14	<b>TCGA-HNSC</b>	0
<b>TCGA-KIRC</b>	12	5	<b>TCGA-KIRC</b>	2
<b>TCGA-KIRP</b>	6	3	<b>TCGA-KIRP</b>	0
<b>TCGA-LIHC</b>	7	7	<b>TCGA-LIHC</b>	0
<b>TCGA-LUAD</b>	7	8	<b>TCGA-LUAD</b>	0
<b>TCGA-LUSC</b>	3	14	<b>TCGA-LUSC</b>	1
<b>TCGA-PAAD</b>	3	0	<b>TCGA-PAAD</b>	0
<b>TCGA-PRAD</b>	1	1	<b>TCGA-PRAD</b>	0
<b>TCGA-STAD</b>	36	18	<b>TCGA-STAD</b>	2
<b>TCGA-THCA</b>	2	3	<b>TCGA-THCA</b>	0
<b>TCGA-UCEC</b>	26	10	<b>TCGA-UCEC</b>	3
<b>total</b>	163	113	<b>total</b>	10

## 8.2 Programas

### download\_metylation\_files.R

#R

```
library(TCGABiolinks) #carrega o pacote TCGABioLinks no R
#estes dois comandos definem em qual diretório serão salvos os arquivos
gerados pelo script
dir <- paste0("~/data/cisb-0037/data/metilation/methylation_450/", cancer)
setwd(dir)
```

#Faz a consulta ao banco de dados do GDC e retorna um dataframe com várias informações a cerca de dados de metilação

#É definido o projeto a ser pesquisado e em seguida são definidos vários filtros para a consulta, tais como: categoria dos dados (metilação de DNA), legacy (que dis respeito a dados gerados tendo como genoma referência o Ghr37, caso fosse usado o Ghr38 seria harmonized)

```
query_metilation <- GDCquery(project= cancer,
                             data.category = "DNA methylation",
                             legacy = TRUE)
```

Essa consulta permite o acesso a todos os dados de metilação de diferentes plataformas, entretanto foi observado que podemos ter o mesmo ID de amostra em diferentes plataformas, dessa forma essa consulta deve ser feita especificando a plataforma de interesse.

É possível a utilização de 3 plataformas:

```
query_metilation450 <- GDCquery(project= cancer,
                                data.category = "DNA methylation",
                                platform = "Illumina Human Methylation 450",
                                legacy = TRUE)
```

```
query_metilation27 <- GDCquery(project= cancer,
                                data.category = "DNA methylation",
                                platform = "Illumina Human Methylation 27",
                                legacy = TRUE)
```

```
query_metilation_vcf <- GDCquery(project= cancer,
                                  data.category = "DNA methylation",
                                  platform = "Illumina HiSeq",
                                  legacy = TRUE)
```

#Algo interessante ocorre com a plataforma Illumina HiSeq, já que são disponibilizados dois tipos de arquivos .bed e .vcf. Esses dois arquivos curiosamente

tem o mesmo ID, sendo assim, quando a consulta é feita as amostras parecem estar duplicadas. Para resolver isso, mais um filtro deve ser adicionado, `file.type = "vcf"` (tipo de arquivo, bed ou vcf):

```
query_metilation_vcf <- GDCquery(project= cancer,
                                data.category = "DNA methylation",
                                platform = "Illumina HiSeq",
                                file.type = "vcf",
                                legacy = TRUE)
```

```
query_metilation_bed <- GDCquery(project= cancer,
                                 data.category = "DNA methylation",
                                 platform = "Illumina HiSeq",
                                 file.type = "bed",
                                 legacy = TRUE)
```

Com a consulta para cada uma das plataformas feitas é necessário mais um filtro, a identificação das amostras pareadas, ou seja, a identificação das amostras normais e tumorais retiradas do mesmo paciente, esse dado é importante para a comparação do que está metilado em tecido normal e não está metilado em tecido tumoral. Para tal foram utilizados dois comandos:

Este comando retira da query campo results primeira coluna campo cases todos os IDs (também chamados de barcodes) de todas as amostras para aquela consulta.

```
bar450 <- unlist(query_metilation450$results[[1]]$cases)
#Escreve um arquivo .txt com todas as amostras para a consulta corrente
write(bar450, paste0(cancer,"_Methylation_450.txt"))
```

Este comando seleciona de todos os barcodes recuperados quais são pareados.

```
SSS450 <- TCGAquery_MatchedCoupledSampleTypes(bar450,c("NT","TP"))
```

```
##Escreve um arquivo .txt com todas as amostras pareadas para a consulta corrente
write(SSS450, paste0(cancer,"_Methylation_450_pareados.txt"))
```

Isto é feito analisando O BARcode das amostras olhando especificamente para o \$° campo

TCGA-02-0001-01C-01D-0182-01 -> 01 é o código para tecidos tumorais

TCGA-02-0001-11C-01D-0182-01 -> 11 é o código para tecidos normais

Após esse passo ter sido realizado é hora de fazer o download das amostras. O download é realizado com outra consulta (query) onde somente as amostras pareadas vão ser consultadas e comporão a query.

```
query_metilation450_2 <- GDCquery(project= cancer,
                                data.category = "DNA methylation",
                                platform = "Illumina Human Methylation 450",
                                legacy = TRUE,
                                barcode = SSS450)
#Faz o download das amostras pareadas
GDCdownload(query_metilation450_2)
#Prepara os dados para serem usados pelo TCGABiolinks
fileName <- paste0("metilation450_", cancer, ".rda")
GDCprepare(query_metilation450_2, save=TRUE, save.filename= fileName)
```

### **download\_mafs.R**

```
# Este programa faz o download dos arquivos de mutação .maf
library(TCGABiolinks)
```

```
setwd("/home/erico/Documents/mutation/biolinks")
```

```
query.maf.hg19 <- GDCquery(project = "TCGA-BRCA",
                           data.category = "Simple nucleotide variation",
                           data.type = "Simple somatic mutation",
                           access = "open",
                           legacy = TRUE)
GDCdownload(query.maf.hg19)
maf <- GDCprepare(query.maf.hg19)
```

### **count\_mutations\_v2.perl**

```
#!/usr/bin/perl
```

```
#Este programa faz a contagem de trocas de todas as bases para todas as bases dos
alelos normais para os alelos tumorais
```

```
my $filename = $ARGV[0];#pega o nome do arquivo que é passado como parametro
open(my $fh, $filename)# Abre o arquivo passado como parametro
or die "Não foi possível abrir o arquivo '$filename' $!";
```



```

print "Contagem trocas arquivo $filename \n";
while (my $row = <$fh>) { #percorre todo arquivo linha por linha

    @temp = split("\t",$row); #splita cada linha usando com separador o \t,
guardando em capos do vetor @temp
    $size_temp = scalar(@temp);#encontr o tamanho do vetor @temp
    #O campo 17 contém informação de base no alelo1 normal
    #O campo 18 contém informação de base no alelo2 normal
    #O campo 11 contém informação de base no alelo1 tumoral
    #O campo 12 contém informação de base no alelo2 normal
    chomp $temp[11];#reira caracter \n da string
    chomp $temp[12];
    chomp $temp[17];
    chomp $temp[18];
    $temp[11] =~ s/^\s+|\s+$//g;#retira espaços do inicio e do fim da string
    $temp[12] =~ s/^\s+|\s+$//g;
    $temp[17] =~ s/^\s+|\s+$//g;
    $temp[18] =~ s/^\s+|\s+$//g;

    if($temp[17] eq "C" ){#toda vez que a base C é encontrada no campo 17 a
variavel $base_C é incrementada em 1, assim até o final do arquivo é possivel saber
quantas vezes a base um apareceu
        $base_C = $base_C + 1;
    }
    if($temp[17] eq "G"){
        $base_G = $base_G + 1;
    }
    if($temp[17] eq "T"){
        $base_T = $base_T + 1;
    }
    if($temp[17] eq "A"){
        $base_A = $base_A + 1;
    }
    ##### Troca de C para demais bases #####
    if($temp[17] eq "C" and $temp[11] eq "T"){ #toda vez que a base C é
encontrada no campo 17 e a base T é encontrada no campo 11 a variavel $c_t_allele1
é incrementada em 1, assim até o final do arquivo é possivel saber quantas vezes hobe
a troca de C>T
        $c_t_allele1 = $c_t_allele1 + 1;
    }
    if($temp[18] eq "C" and $temp[12] eq "T"){

```

```

        $c_t_allele2 = $c_t_allele2 +1;
    }
    #####
    if($temp[17] eq "C" and $temp[11] eq "G"){
        $c_g_allele1 = $c_g_allele1 +1;
    }
    if($temp[18] eq "C" and $temp[12] eq "G"){
        $c_g_allele2 = $c_g_allele2 +1;
    }
    #####
    if($temp[17] eq "C" and $temp[11] eq "A"){
        $c_a_allele1 = $c_a_allele1 +1;
    }
    if($temp[18] eq "C" and $temp[12] eq "A"){
        $c_a_allele2 = $c_a_allele2 +1;
    }

    ##### Troca de T para demais bases #####

    if($temp[17] eq "T" and $temp[11] eq "C"){
        $t_c_allele1 = $t_c_allele1 +1;
    }
    if($temp[18] eq "T" and $temp[12] eq "C"){
        $t_c_allele2 = $t_c_allele2 +1;
    }
    #####
    if($temp[17] eq "T" and $temp[11] eq "G"){
        $t_g_allele1 = $t_g_allele1 +1;
    }
    if($temp[18] eq "T" and $temp[12] eq "G"){
        $t_g_allele2 = $t_g_allele2 +1;
    }
    #####
    if($temp[17] eq "T" and $temp[11] eq "A"){
        $t_a_allele1 = $t_a_allele1 +1;
    }
    if($temp[18] eq "T" and $temp[12] eq "A"){
        $t_a_allele2 = $t_a_allele2 +1;
    }

    ##### Troca de A para demais bases #####

```

```

    if($temp[17] eq "A" and $temp[11] eq "T"){
        $a_t_allele1 = $a_t_allele1 +1;
    }
    if($temp[18] eq "A" and $temp[12] eq "T"){
        $a_t_allele2 = $a_t_allele2 +1;
    }
    #####
    if($temp[17] eq "A" and $temp[11] eq "G"){
        $a_g_allele1 = $a_g_allele1 +1;
    }
    if($temp[18] eq "A" and $temp[12] eq "G"){
        $a_g_allele2 = $a_g_allele2 +1;
    }
    #####
    if($temp[17] eq "A" and $temp[11] eq "C"){
        $a_c_allele1 = $a_c_allele1 +1;
    }
    if($temp[18] eq "A" and $temp[12] eq "C"){
        $a_c_allele2 = $a_c_allele2 +1;
    }

    ##### Troca de G para demais bases #####

    if($temp[17] eq "G" and $temp[11] eq "T"){
        $g_t_allele1 = $g_t_allele1 +1;
    }
    if($temp[18] eq "G" and $temp[12] eq "T"){
        $g_t_allele2 = $g_t_allele2 +1;
    }
    #####
    if($temp[17] eq "G" and $temp[11] eq "A"){
        $g_a_allele1 = $g_a_allele1 +1;
    }
    if($temp[18] eq "G" and $temp[12] eq "A"){
        $g_a_allele2 = $g_a_allele2 +1;
    }
    #####
    if($temp[17] eq "G" and $temp[11] eq "C"){
        $g_c_allele1 = $g_c_allele1 +1;
    }

```

```

        if($temp[18] eq "G" and $temp[12] eq "C"){
            $g_c_allele2 = $g_c_allele2 + 1;
        }
    $i = $i + 1;
}

print "\n Quantidade de C:$base_C \n
Quantidade de G:$base_G \n
Quantidade de T:$base_T \n
Quantidade de A:$base_A \n

Troca C para T alelo 1:$c_t_allele1 e alelo 2: $c_t_allele2 \n
Troca C para G alelo 1:$c_g_allele1 e alelo 2: $c_g_allele2 \n
Troca C para A alelo 1:$c_a_allele1 e alelo 2: $c_a_allele2\n
#####\n
Troca T para C alelo 1:$t_c_allele1 e alelo 2: $t_c_allele2 \n
Troca T para G alelo 1:$t_g_allele1 e alelo 2: $t_g_allele2 \n
Troca T para A alelo 1:$t_a_allele1 e alelo 2: $t_a_allele2\n
#####\n
Troca A para T alelo 1:$a_t_allele1 e alelo 2: $a_t_allele2 \n
Troca A para G alelo 1:$a_g_allele1 e alelo 2: $a_g_allele2 \n
Troca A para A alelo 1:$a_c_allele1 e alelo 2: $a_c_allele2\n
#####\n
Troca G para T alelo 1:$g_t_allele1 e alelo 2: $g_t_allele2 \n
Troca G para A alelo 1:$g_a_allele1 e alelo 2: $g_a_allele2 \n
Troca G para C alelo 1:$g_c_allele1 e alelo 2: $g_c_allele2\n
Quantidade de mutacoes:$i\n";

```

### Compare\_mutation\_metilation2.perl

```

#!/usr/bin/perl
#Esse script tem o objetivo de a partir de dois arquivos .maf identificar o numero de
variantes presentes nos dois arquivos.
#Pega o nome do primeiro arquivo que é passado como parâmetro
my $filename = $ARGV[0];
$file1 = `cat $filename`;#le todo o arquivo 2 e o armazena na variavel $file2
@rows_file1 = split("\n",$file1);#guarda cada linha do arquivo em um campo do vetor
@rows_file1
$size_file1 = scalar(@rows_file1);#guarda na variavel $size_file1 a quantidade de
linhas do arquivo1

```

```

for(my $i = 0; $i < $size_file1;$i++){ #percorre todas as linhas do arquivo1
    $rows_file1[$i] =~ s/^\s+|\s+$//g;#remove espaços em branco
    @fields_file1 = split("\t",$rows_file1[$i]); #divide cada linha em campos tendo como
base o caracter \t
    $size_fields1 = scalar(@fields_file1); #encontra quantos campos tem na linha
    for(my $j = 0;$j < $size_fields1 ; $j++){ #percorre todos os campos da linha
        $matriz_file1[$i][$j] = $fields_file1[$j]; # guarda cada um dos campos em
campos de uma linha de uma matiz
    }
}

```

```

#Pega o nome do segundo arquivo que é passado como parâmetro
my $filename2 = $ARGV[1];
$file2 = `cat $filename2`;#le todo o arquivo 2 e o armazena na variavel $file2
@rows_file2 = split("\n",$file2);#guarda cada linha do arquivo em um campo do vetor
@rows_file2
$size_file2 = scalar(@rows_file2);#guarda na variavel $size_file2 a quantidade de
linhas do arquivo2

```

```

for(my $i = 1; $i < $size_file2;$i++){ #percorre todas as linhas do arquivo2
    $rows_file2[$i] =~ s/^\s+|\s+$//g;#remove espaços em branco
    @fields_file2 = split("\t",$rows_file2[$i]);#divide cada linha em campos tendo como
base o caracter \t
    $size_fields2 = scalar(@fields_file2);#encontra quantos campos tem na linha
    for(my $j = 0;$j < $size_fields2 ; $j++){#percorre todos os campos da linha
        $matriz_file2[$i][$j] = $fields_file2[$j];# guarda cada um dos campos em
campos de uma linha de uma matiz
    }
}

```

```

for(my $i = 0; $i < $size_file1; $i++){#percorre todas as linhas da primeira matriz
    #print "linha $i\n";
    $coord_file1 = $matriz_file1[$i][5] - 1;#guarda o valor da coluna 5 da matriz
subtraído de 1 na variavel coord_file1
    for(my $y = 1; $y < $size_file2; $y++){#percorre todas as linhas da matriz2
        $coord_file2 = $matriz_file2[$y][4] + 0;#converte o valor da coluna 4 em
inteiro e o guarda na variavel $coord_file2
        $beta_value = $matriz_file2[$y][1] + 0;#converte o valor da coluna 4 em
inteiro e o guarda na variavel $beta_value
        #if($beta_value => 0.3){

```

```

        if($matriz_file1[$i][4] eq $matriz_file2[$y][3] and $matriz_file1[$i][5] eq
$matriz_file2[$y][4]){#compara o cromossomo e as posições nos dois arquivos
            print"$matriz_file1[$i][0] \t $matriz_file1[$i][4] \t $matriz_file1[$i][5] \t
$matriz_file1[$i][8] \t $matriz_file1[$i][18] \t $matriz_file1[$i][12] \t $matriz_file1[$i][45] \t
$matriz_file2[$y][1] \t $matriz_file1[$i][15] \t $matriz_file1[$i][16]\n "; #imprime os
campos nome do gene, cromossomo, posição, impacto da mutação, base normal, base
tumor, strand, beta_value, barcode_tumor, barcode normal
        }

        if($matriz_file1[$i][4] eq $matriz_file2[$y][3] and $coord_file1 ==
$coord_file2){#compara o cromossomo nos dois arquivos a posição no arquivo de
mutação com a posição anterior no arquivo de metilação
            print"$matriz_file1[$i][0] \t $matriz_file1[$i][4] \t
$matriz_file1[$i][5] \t $matriz_file1[$i][8] \t $matriz_file1[$i][18] \t $matriz_file1[$i][12] \t
$matriz_file1[$i][45] \t $matriz_file2[$y][1] \t $matriz_file1[$i][15] \t
$matriz_file1[$i][16]\n"; #imprime os campos nome do gene, cromossomo, posição,
impacto da mutação, base normal, base tumor, strand, beta_value, barcode_tumor,
barcode normal
        }
    #}
}
}
}

```

Mas como esse programa foi rodado em cluster para várias amostras ao mesmo tempo foi necessária a confecção do programa `run_intersect.sh`

```
#!/bin/bash
```

```
for i in `cat list_of_donnors_metilation`;do
```

```
#echo $i;
```

```
met=`grep "$i" list_metilation_files_CESC_normal`;
```

```
var=`echo $i | awk -F '-' '{print $3}`;
```

```
#echo $var;
```

```
mut=`grep "$var" ../list_mutation_files_CESC`;
```

```
#echo "aqui $mut";
```

```
#echo $met;
```

```
cat modelo.sh | sed 's/NOME_DO_JOB/run_intersect/g' > programs/$i.sh;
```

```

echo          "/compare_mutation_metilation2.perl          $mut          $met          >
/data1/projects/cisbi-0037/TCGA-CESC/output/mut_x_met/normal/$i.intersect"      >>
programs/$i.sh;

```

done

Para cada paciente foi gerado um programa como esse:

```
#!/bin/bash -v
```

```
#PBS -S /bin/bash
```

```
#PBS -N run_intersect
```

```
#PBS -l nodes=1:ppn=1
```

```
#PBS -o saida_torque
```

```
#PBS -e saida_erro_torque
```

```
#PBS -e out
```

```
#PBS -q emu
```

```
### cd to directory where the job was submitted:
```

```
cd $PBS_O_WORKDIR
```

```
### determine the number of allocated processors:
```

```
NPROCS=`wc -l < $PBS_NODEFILE`
```

```
echo "-----"
```

```
echo "PBS job running on: `hostname`"
```

```
echo "in directory:    `pwd`"
```

```
echo "nodes: $NPROCS"
```

```
echo "nodefile:"
```

```
cat $PBS_NODEFILE
```

```
echo "-----"
```

```
#ulimit -s unlimited
```

```
set -m # Enable Job Control
```

```
##Coloque seus comandos apos esta linha
```

```
#####
```

```
./compare_mutation_metilation2.perl
```

```
/data1/projects/cisbi-0037/data/mutation/tcga2bed_original/<amostra>.maf
```

```
/data1/projects/cisbi-0037/data/metilation/tcga2bed_original/normal/j<amostra>.txt      >
```

```
/data1/projects/cisbi-0037/output/mut_x_met/normal/beta_maior_0.3/<amostra>.intesec
```

```
t
```

**select\_metilation\_cutof.perl**

```
#!/usr/bin/perl
```

#este script seleciona dos arquivos de metilação o cutoff de beta value para se considerar uma base C como metilada.

```
my $filename = $ARGV[0];
```

```
open(my $fh, '<:encoding(UTF-8)', $filename)
  or die "Não foi possível abrir o arquivo '$filename' $!";
```

```
$i = 0;
```

```
while (my $row = <$fh>) {
  chomp $row;
  @temp = split("\t", $row);
  if($i <= 1){
    #print "$row \n";
  }
  $beta_value = $temp[7] + 0;
  if($beta_value <= 0.3){
    print "$row\n";
  }
  $i = $i + 1;
}
```

**select\_cs\_e\_gs.perl**

```
#!/usr/bin/perl
```

#este script seleciona somente as trocas C>T e G>A dos arquivos de intersecção.

```
my $filename = $ARGV[0];
```

```
open(my $fh, '<:encoding(UTF-8)', $filename)
  or die "Não foi possível abrir o arquivo '$filename' $!";
```

```
$i = 0;
```

```
while (my $row = <$fh>) {
  chomp $row;
```



```

@temp = split("\t", $row);
$temp[4] =~ s/^\s+|\s+$//g;
$temp[5] =~ s/^\s+|\s+$//g;
    if($temp[4] eq "C" and $temp[5] eq "T"){
        print "$row\n";

    }
    if($temp[4] eq "G" and $temp[5] eq "A"){
        print "$row\n";

    }
    $i = $i + 1;
}

```

Que tinha como nome o próprio identificador da amostra.

## Comandos

A seguir está demonstrado todos os comandos utilizados até agora, lembrando que alguns comandos são a chamada para os programas listados no índice anterior:

Este comando foi utilizado para listar todos os arquivos que começam com a string jhu-usc e terminam .txt Pois todos os arquivos precisavam ser extraídos de suas pastas

```
ls */jhu-usc.**.txt > list_files
```

Este comando extrai todos os arquivos de seus diretórios e os salva no diretório corrente

```
for i in `cat list_files`; do cp $i .; done
```

Este comando envia os arquivos para o diretório correto

```
mv jhu-usc.edu* ../../../../../../tcga2bed_original_pareados/
```

```
cd ../../../../../../tcga2bed_original_pareados/
```

As amostras precisavam ser divididas em dois grupos, normais e tumorais, para isso dois arquivos foram gerados, um contendo barcodes de amostras normais e outro contendo barcodes de amostras tumorais.

```
ls jhu-usc.edu* | awk -F '.' '{ print $6}' | awk -F '-' '{ print $1 "-" $2 "-" $3 "-" $4}' | grep
'-.11[A-Z]' > normal_list
```

```
ls jhu-usc.edu* | awk -F '.' '{ print $6}' | awk -F '-' '{ print $1 "-" $2 "-" $3 "-" $4}' | grep
'-.01[A-Z]' > tumoral_list
```

A partir dessas listas foram geradas listas contendo o nome de arquivos tumorais e normais

```
for i in `cat normal_list`; do ls jhu-usc.edu* | grep "$i" ;done > files_normal_list
```

```
for i in `cat tumoral_list`; do ls jhu-usc.edu* | grep "$i" ;done > files_tumoral_list
```

A partir dessas lista geradas foi possível copiar amostras normais para o diretório de normais e amostras tumorais para o diretório de tumorais.

```
for i in `cat files_normal_list`; do mv $i normal;done
```

```
for i in `cat files_tumoral_list`; do mv $i tumor;done
```

Este comando foi usado para gerar uma lista com o id de todos os doadores a partir dos dados de mutação comparado com os dados de metilação, garantindo que haveria somente amostras iguais sendo comparadas.

```
ls *.txt | awk -F '-' '{print $3}' > list_of_donors
```

Este comando recupera todo o caminho do arquivos de metilação normais ou tumorais

```
ls
/data1/projects/cisbi-0037/TCGA-BLCA/data/metilation/tcga2bed_original_pareados/files/normal/*.txt >
/data1/projects/cisbi-0037/TCGA-BLCA/list_metilation_files_BLCA_normal
```

Este comando recupera somente as amostras de mutação que também tem amostras de metilação

```
for i in `ls
/data1/projects/cisbi-0037/TCGA-BLCA/data/metilation/tcga2bed_original_pareados/files/tumor/*.txt | awk -F '.' '{print $6}' | awk -F '-' '{print $3}'`;do ls
/data1/projects/cisbi-0037/TCGA-BLCA/data/mutation/*.txt | grep "$i"; done | sort -u>
/data1/projects/cisbi-0037/TCGA-BLCA/list_mutation_files
```

Este comando copia os mafs pareados para outro diretório

```
for i in `cat list_mutation_files_LIHC`; do cp $i
mutation/tcga2bed_original_pareados;done
```

Este comando concatena os arquivos maf pareados

```
for i in `ls *.maf.txt`; do cat $i;done > pareados_CESC_total.maf
```

Estes comandos selecionam somente as trocas de bases eliminando as inserções e deleções

Este comando verifica a quantidade de registros de SNPs, INS e DELs

```
awk '{print $10}' total_mut_pareados.maf | grep "SNP" total_mut_pareados.maf | wc -l
```

Este comando verifica se o filtro esta deixando passar mais informações do que o necessário, o parametro -w seleciona somente letras maiusculas

```
grep -w "SNP" total_mut_pareados.maf | wc -l
```

Este comando faz efetivamente o filtro

```
grep -w "SNP" total_mut_pareados.maf > total_mut_pareados_SNP.maf
```

Este comando gera a contagem de cada uma das bases nos arquivos maf

```
./count_mutations_v2.perl
/data1/projects/cisbi-0037/TCGA-BLCA/data/mutation/tcga2bed_original_pareados/pareados_BCLA_total_SNP.maf >
/data1/projects/cisbi-0037/TCGA-BLCA/output/porcentagens_mutacoes/porcentagens_mutacoes_mut_BLCA
```

Este comando coloca os resultados do intersect para todas as amostras em um arquivo só:

```
for i in `ls *.intersect`;do cat $i;done; > total_<TUMOR>_tumor
```

Foi usado o programa count\_intersects.perl para contar a quantidade de c para qualquer coisa e g para qualquer coisa

```
for i in `ls *`;do ../../count_intersects.perl $i > proporcoes_$i;done
```

Foi usado o programa select\_metilation\_cutof.perl para separar entre metilados e não metilados de acordo com o beta value de cutoff 0,3

```
for i in `ls total*`;do ../select_metilation_cutof.perl $i >
beta_maior_03/metilado_`;done
```

## 9. MANUSCRITO

Manuscrito que será submetido à revista PlosOne.

### **Computational approach to assess the impact of DNA methylation on tumor instability.**

Running title: DNA methylation by Computational approaches

Érico Torrieri<sup>1</sup>, Wilson Araujo Silva Junior<sup>1</sup>.

1- medical school of ribeirão preto - USP - genetics department.

Key-words: Deamination, methylation, mutation, tumor instability, cytosine.

### **Abstract**

DNA methylation is an important epigenetic mechanism for regulating gene expression and can trigger several complex diseases, such as cancer, diabetes, and obesity. DNA methylation occurs in greater abundance on cytosine bases, mostly in CG dinucleotides (CpG sites). Estimates suggest that 35% of point mutations causing human genetic disorders have occurred at CpG dinucleotides, and over 90% of these were transitions from C > T or corresponding G > A transitions. Methylation of CpGs in normal tissues might increase the probability of mutations at such sites due to the ability of 5-methylcytosine to undergo deamination, resulting in the exchange of cytosine for thymine. In tumors, this event can be better observed since regions of CpG islands are

usually abundantly methylated. This work aims to locate C > T exchanges at methylated CpG sites and to infer whether these exchanges occurred due to the deamination mechanism. For this purpose, The Cancer Genome Atlas (TCGA) project's mutation and methylation data were used. The results indicate that for exome data, a rate of 37.67% of C > T exchanges were found, with the second most probable exchange being C > A transition at 24.18%. The rate of C > T exchanges goes up to 75.23% for CpG sites and up to 76.98% for methylated CpG sites, and 65.98% exchanges lead to pathogenic mutations. These results suggested that mutations in methylated cytosines contribute to the progression of tumorigenesis.

Keywords: Deamination, methylation, mutation, tumorigenesis, cytosine.

Methylation is an important epigenetic mechanism involved in regulation of gene expression and can trigger several types of complex diseases, such as, cancer, diabetes, and obesity. It occurs primarily in two ways: histone methylation and DNA methylation<sup>1</sup>. DNA methylation occurs in greater abundance on cytosines, on the carbon present at the fifth position of the pyrimidine ring, generating 5-methylcytosine. In mammals, it is estimated that 80% of methylation events occur in CpG dinucleotides (CpG sites)<sup>2,3,4</sup>.

The occurrence of CpG sites in the genome is low (on an average 1/80 bp); however, CpG sites occur with high frequency (1/16 bp) in genomic regions called CpG islands, which are often located in the promoter regions of approximately 70% of protein-coding genes. Thus, CpG islands are associated with the coding regions of genes<sup>5,6</sup>. The low occurrence of CG dinucleotides in the genome is attributed to the hypermutability of methylated CpGs to TpGs (or CpAs in the complementary chain) that

have accumulated over the course of evolution. However, CpG islands are low in methylation due to the selective pressure for these sites to remain conserved<sup>7,8,9</sup>.

Estimates suggest that 35% of point mutations causing human genetic disorders have occurred at CpG dinucleotides, and over 90% of these were transitions from C > T or corresponding G > A transitions<sup>10</sup>. Methylation of CpGs in normal tissues might increase the probability of mutations at these sites due to the ability of 5-methylcytosine to undergo deamination, resulting in a thymine<sup>11</sup>.

Under normal physiological conditions, cytosine is converted to uracil, a base that is not natural to DNA, and is recognized and corrected by the repair mechanisms. However, spontaneous deamination of 5-methylcytosine (methylated cytosine) converts the residue to thymine, which is natural for DNA and is recognized by the repair mechanisms; following a replication cycle, such a mutation becomes permanent as the strand carrying T will be complemented by an A<sup>12</sup>. These types of modifications can be pro-mutagenic and can contribute to the formation of mutational hotspots in cells<sup>11</sup>.

In tumors, this C > T exchange event can be better observed since there is a general relative hypomethylation of DNA in cancer cells interspersed with hypermethylation of CpG islands. This event is more predominant for tumor suppressor genes<sup>13</sup>.

Previous studies by Weisenberger and coworkers have proved that there are mutations associated with methylation. They have shown that in colorectal cancer, a strong association was observed between a hypermethylation phenotype and mutation of BRAF oncogene, a protein kinase involved in EGFR pathway downstream of KRAS<sup>14</sup>. However, there are no known studies involving a large number of tumors and samples describing the true impact of deamination mechanism on genomic instability in tumors.

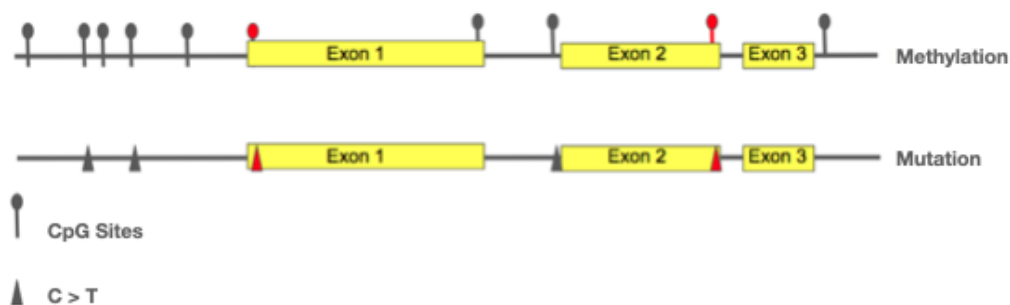
The understanding of the mechanisms involved in the progression of cancer is of great interest to the scientific community. The progression of tumorigenesis can be reduced by different therapeutic approaches during the healing process, and also initially by studies using bioinformatics tools to predict mutations. Thus, supported by the hypothesis that the deamination mechanism plays a relevant role in promoting gene instability in tumors, the present work was designed to investigate the impact of DNA methylation on the rate of C > T substitutions in tumor tissues through the development of a novel computational approach. For this project, the data on somatic mutations present in exons (generated by next generation sequencing) and methylation were chosen, and the main objective was to identify the impact of C > T exchanges in methylated CpG sites (for both normal and tumorigenic samples).

## Methods

To carry out this work, the data generated by The Cancer Genome Atlas (TCGA) project was chosen<sup>15</sup>. Exomatic mutation data in .maf format generated from next generation sequencing and methylation in .txt format generated from experiments using Human Methylation 450 array from the public GDC repository (<https://gdc.cancer.gov>) with the aid of R TCGABiolinks tool<sup>16</sup> were taken for analysis. Only samples that contained data for both mutations and methylation were downloaded. Next, 22 tumor types were selected containing samples from both experiments.

Although the experiments were conducted using the same tumor types, the same set of samples were not used always. Hence, only samples present in both the methylation mutation files were selected for bioinformatics analysis.

A program was then built with the objective of finding C > T mutations in methylated CpG sites, for which it was necessary to associate information regarding the coordinates in which each C > T mutation occurred (information stored in the mutation files) with the coordinates of each methylated CpG site (information stored in the methylation files) (Fig 1).

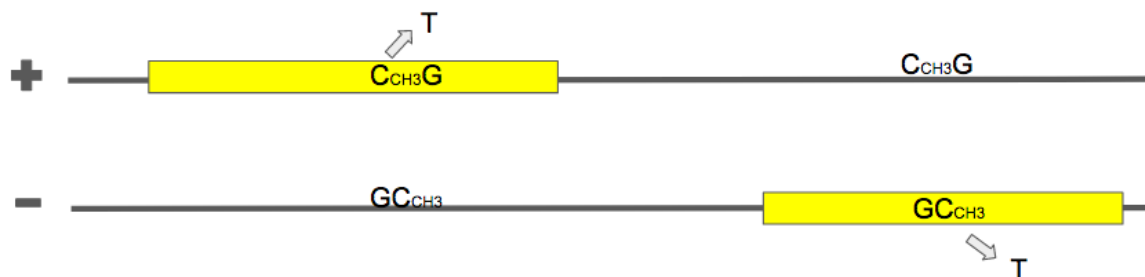


**Figure 1.** Association of mutation and methylation regions. Figure indicating how the mutation and methylation information was associated.

As indicated in Fig 1 representing DNA methylation, the coding region is shown to the left of exon 1, which is the promoter region containing several methylated CpG sites and some methylated CpG sites that invade exons. In the second part of the figure, we have a representation of the same genomic region, only now indicating the position of C > T mutations in that region; some mutations are seen in the promoter region while some also in the coding region. The mutations and CpG sites indicated in red represent mutations and methylated CpG sites in the same genomic coordinate; precisely these cases are the ones that the program seeks.

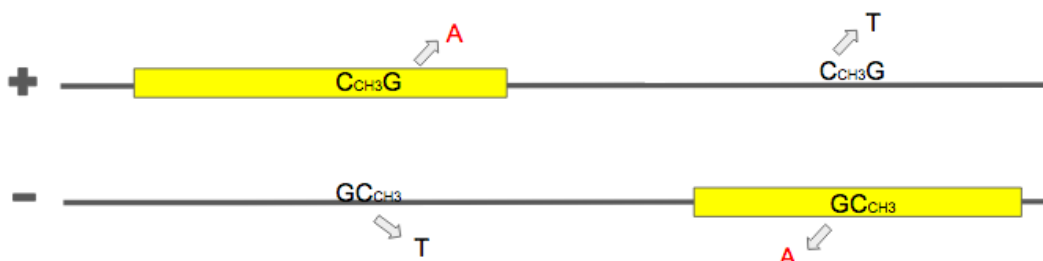
Thus, it would be necessary for the program to identify four different situations, the first two being the most evident when a C > T exchange occurs at a methylated CpG site in a coding region (figure 2).





**Figure 2.** C > T mutations occurring at methylated CpG sites in coding regions on both the + and the - strands. Figure showing C > T mutations occurring at methylated CpG sites in coding regions on both the + and the - strands.

The other two are less evident mutations because the mutation in a coding region is a consequence of a C >T exchange on the complementary strand to the strand in which the gene is present (Figure 3).



**Figure 3.** Mutation at methylated CpG site. Figure showing when the C to T mutation at the methylated CpG site occurs on the reverse ribbon with respect to the gene and only after a cycle of replication does it also fix where the gene is found, but as an exchange from G to A.

However, annotation of both mutation and methylation data is only done on the + tape, so even if the mutation occurs on the tape, it will be recorded on the + tape. In the same way, methylation data, as they only investigated methylation at CpG sites on the + tape, so it is not possible to know any information about methylated sites on the reverse tape (figure 4).



**Figura 4.** Informações referentes a fita +. Figura mostrando que só se tem informações referentes a fita + tendo trocas de C para T e G para A mas sem saber de se as trocas de G para A tiveram origem em uma troca C para T.

Therefore, the program was built as follows: in cases of C > T exchange, the exact position of the exchange in the methylation files was investigated to find a methylated CpG site in that exact region. However, in cases of G > A exchange, the immediately previous position was also investigated. If it was a C, it would mean that there was a CpG site there before, so the position of the previous C was searched in the methylation file, as it is understood that if a CpG site is methylated, so would be its complement also<sup>4,9</sup>. It could be assumed that the G > A exchange was caused by a C > T exchange on the complementary tape (Figure 5) (Compare\_mutation\_metilation2.perl program).



**Figure 5.** Search engine of the algorithm. Figure showing that when you have a C to T exchange your exact position is searched for in the methylation file, but when you have a G to A exchange, the position prior to the exchange is investigated, if it is a C this position will be searched in the file methylation, as it is assumed that the C of the CpG site on the complementary strand is also methylated.

For the variants identified as causing missense and nonsense mutations, the UMD predictor program was run<sup>17</sup>, to identify the pathogenicity of the exchanges. The UMD predictor combines biochemical properties, impact on splicing signals, localization in protein domains, frequency of variation in the global population, and conservation through the BLOSUM62 global substitution matrix, and a specific protein conservation among 100 species to define the pathogenicity of a mutation. It presents results classifying the mutations as polymorphic, probably polymorphic, pathogenic, and probably pathogenic.

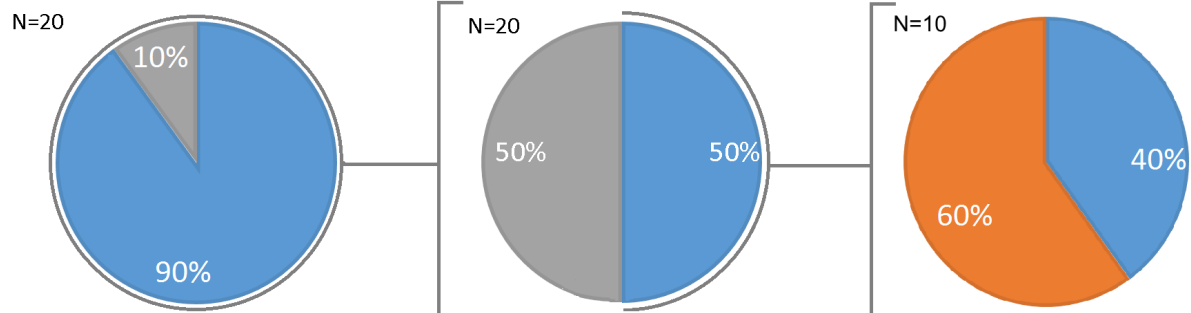
For the enrichment of pathways, the clusterprofile package present in the Bioconductor repository was used to enrich the pathways of genes that have undergone mutations in CpG sites caused by methylation.

## **Results**

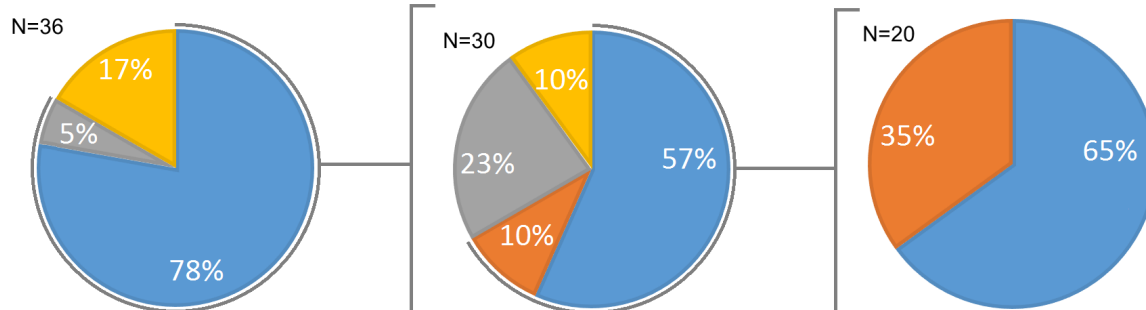
For exome data, a rate of 37.67% of C > T exchanges was found, while the second most likely exchange was C > A with 24.18%. When we look only at CpG sites, this percentage goes up to 75.23% of C > T exchanges, and when we look only at methylated CpG sites, we see that 76.98% are C > T exchanges.

Figure 6 summarizes the results obtained after the execution of the algorithm described in the methods. The results are obtained for tumor and normal samples in relation to the type of mutation as well as the pathogenicity of the mutation.

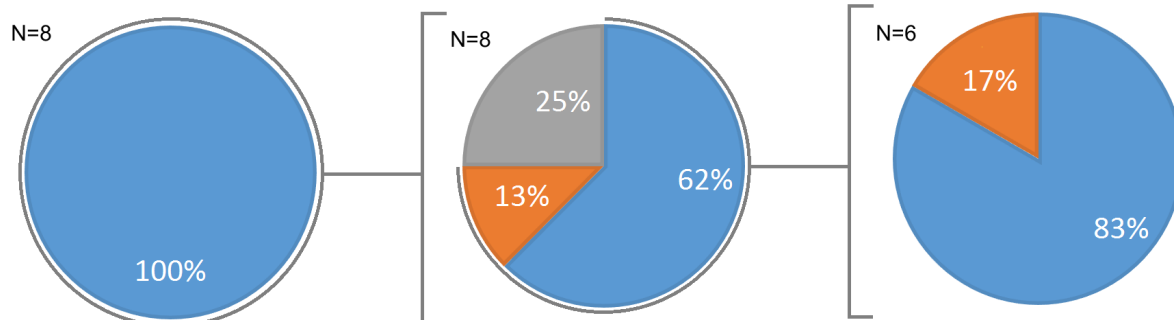
## BLADDER UROTHELIAL CARCINOMA



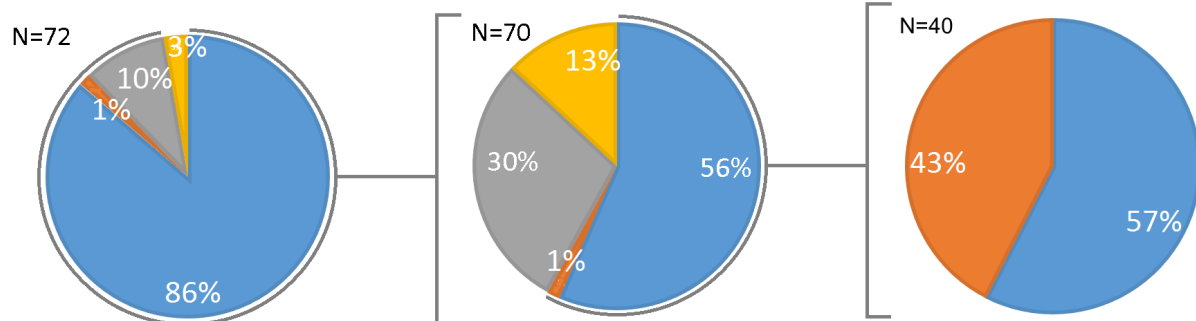
## BREAST INVASIVE CARCINOMA



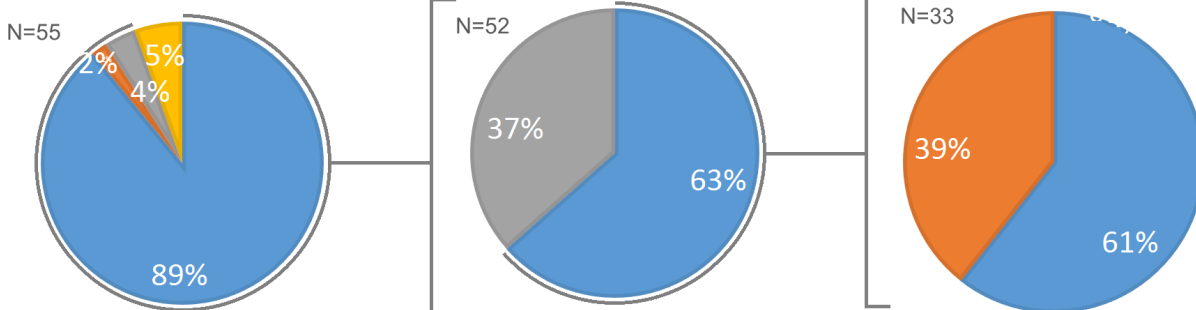
## CHOLANGIOCARCINOMA



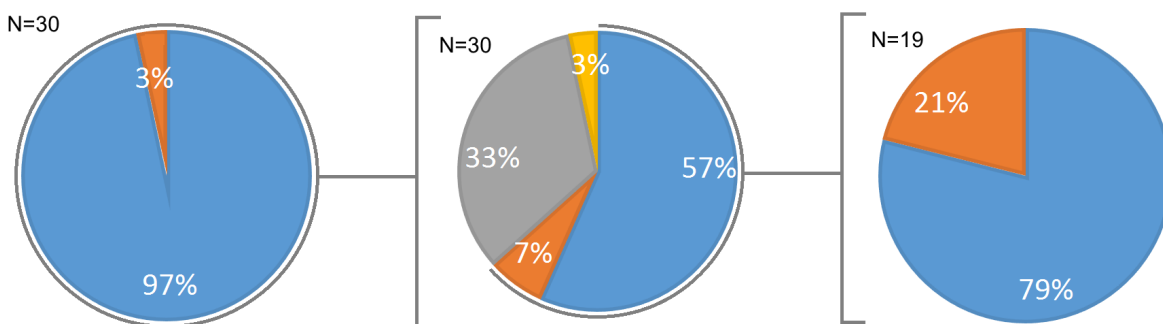
## ESOPHAGEAL CARCINOMA



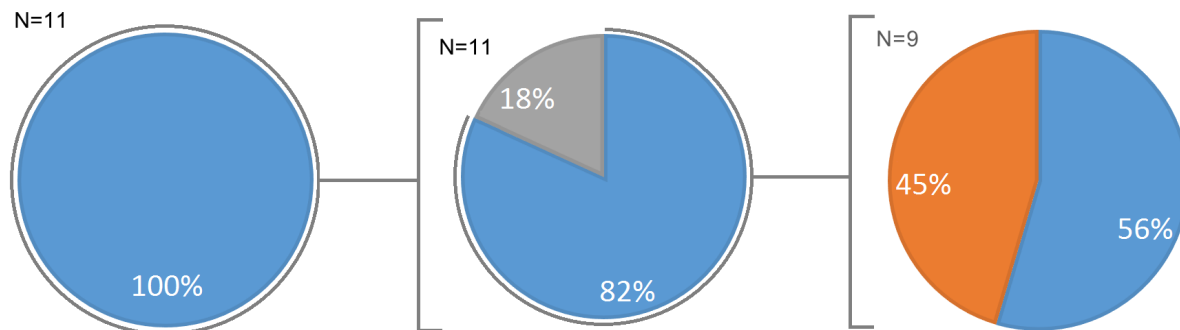
## HEAD AND NECK SQUAMOUS CELL CARCINOMA



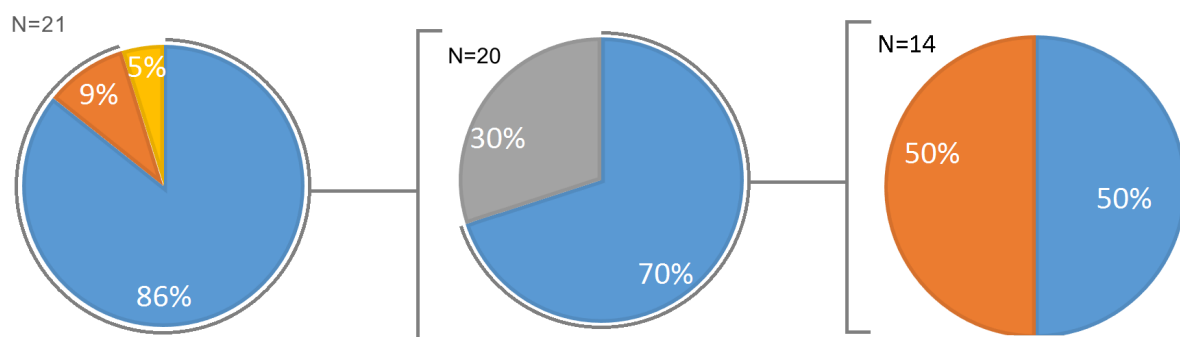
## KIDNEY RENAL CLEAR CELL CARCINOMA



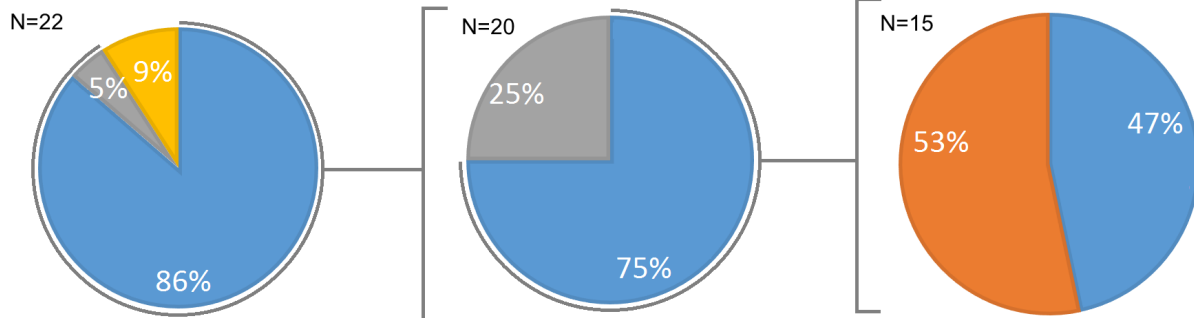
## KIDNEY RENAL PAPILLARY CELL CARCINOMA



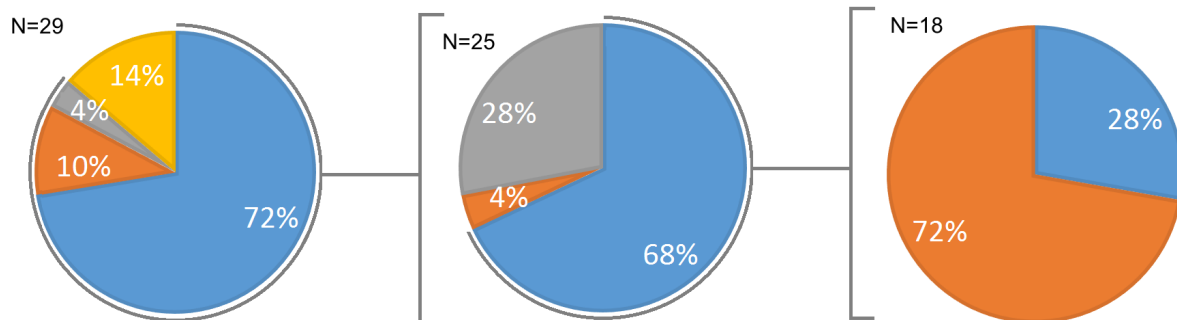
## LIVER HEPATOCELLULAR CARCINOMA



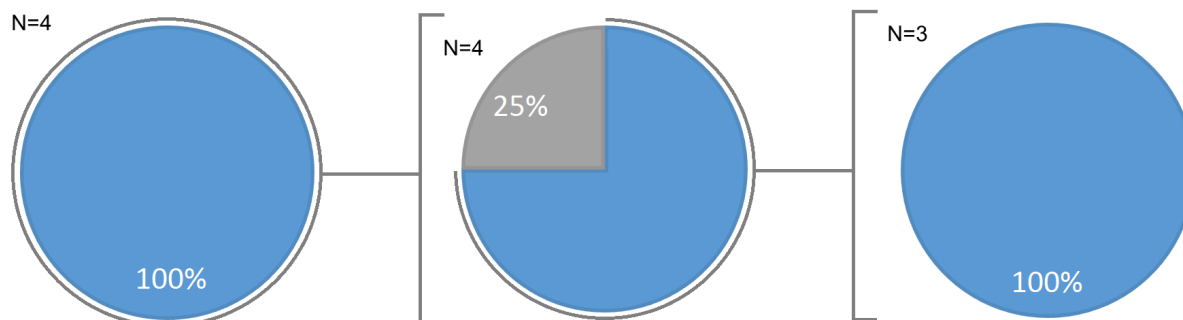
LUNG ADENOCARCINOMA



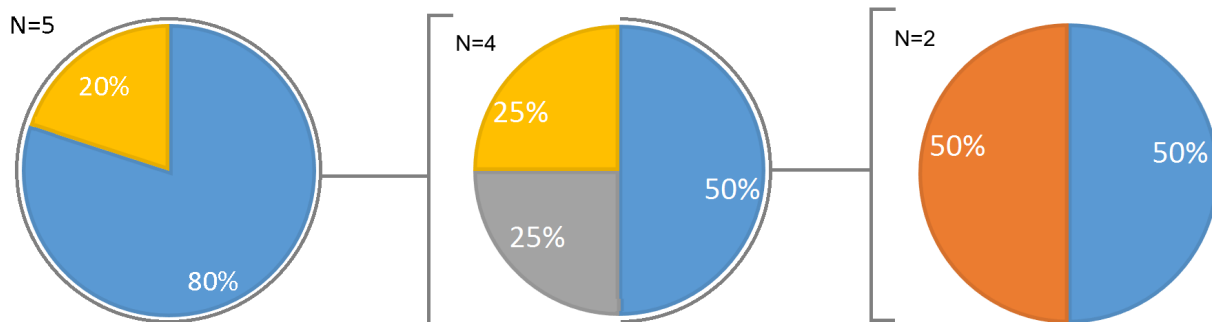
LUNG SQUAMOUS CELL CARCINOMA

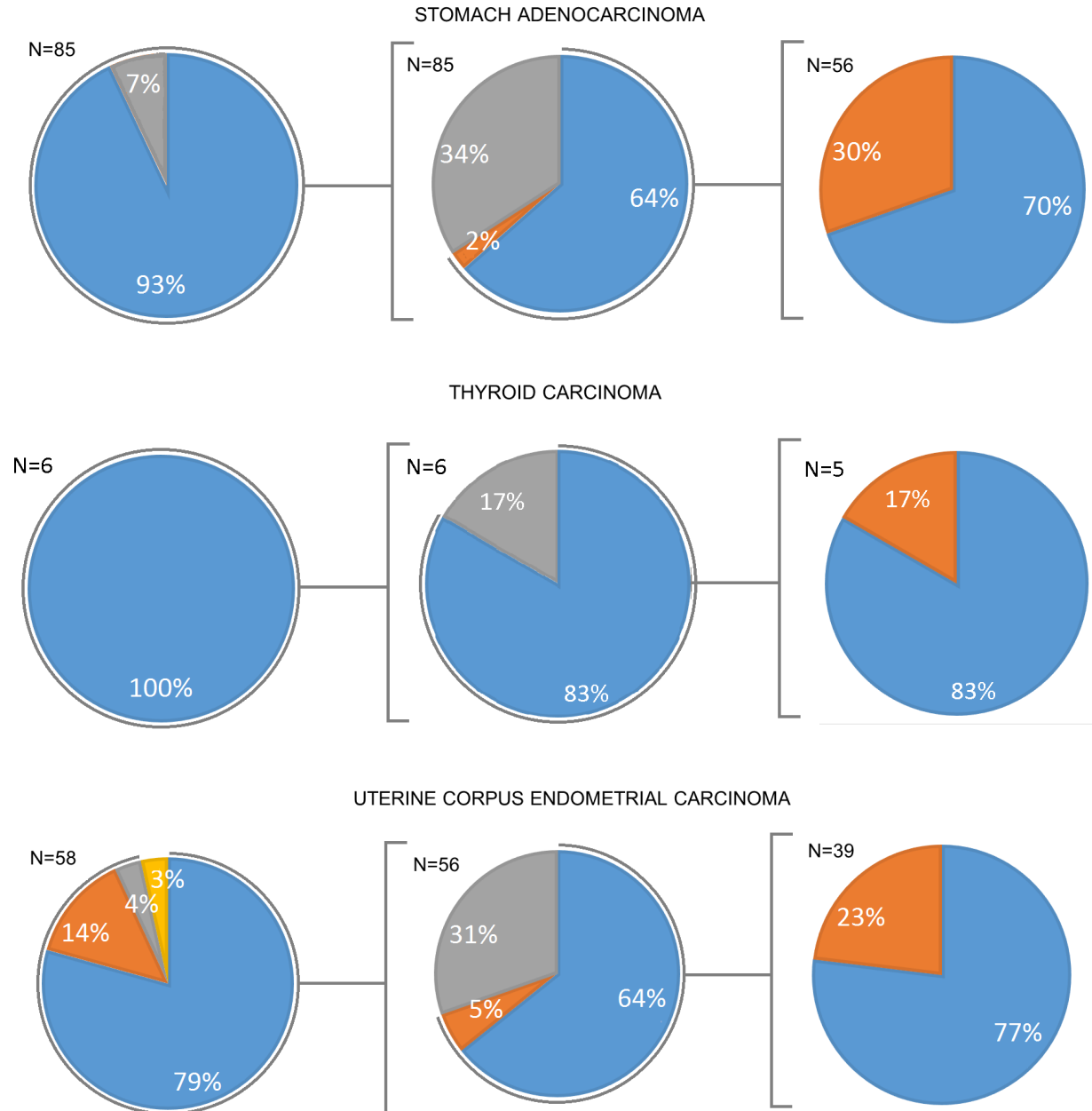


PANCREATIC ADENOCARCINOMA



PROSTATE ADENOCARCINOMA



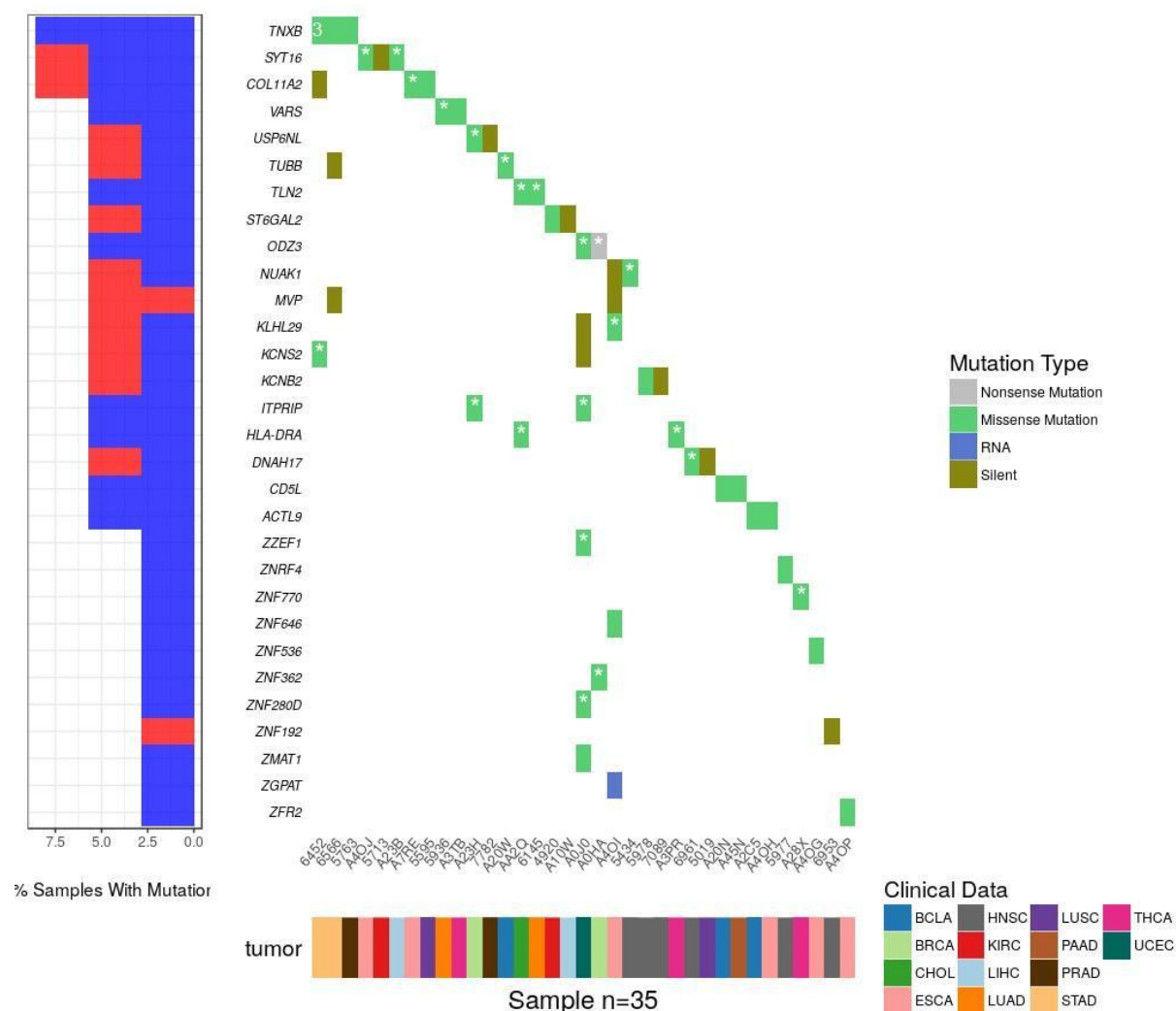


**Figure 6.** Methylation between normal and tumor samples versus the type of mutation corresponding to the exchange. The first column contains information on methylation between normal and tumor samples. In the second column, the type of mutation corresponding to the exchange is identified, only mutations in methylated CpG sites are taken into account. In the third column, pathogenic or not mutations are identified. Only exchanges that lead to missense and nonsense mutations are considered.

Figure 7 summarizes the number of mutations for the 30 genes with the most mutations, taking into account mutations in different samples and tumor types, identifying whether these mutations are pathogenic or not.

Only 20 genes underwent more than one mutation, the majority of which were two mutations in different samples. Only the *TNXB* gene underwent five mutations, two of which were in the same sample and none of them was classified as pathogenic according to the prediction of the UMD predictor.

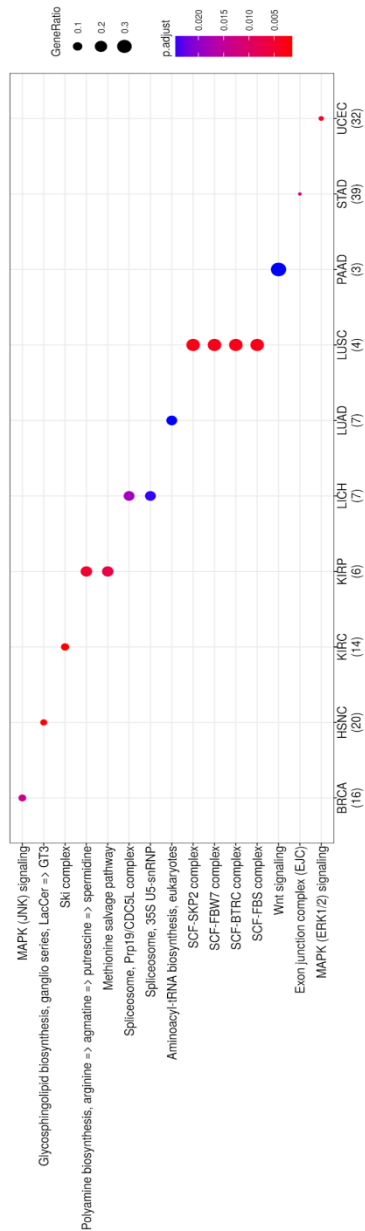
### Mutations distributed by genes





**Figure 7.** Oncoprint of the number of mutations per gene This oncoprint shows the number of mutations per gene, also showing the classification of these mutations and whether they are pathogenic (asterisk) according to the UMD predictor for the 30 genes with the most mutations taking into account all types tumoral.

It is observed that different pathways were enriched for each of the tumors (Figure 8).



**Figure 8.** Pathways enriched according to genes that have mutated at CpG sites caused by methylation separated by tumor type.

## Discussion

With the completion of the initial bioinformatics analysis, it was possible to obtain an overview of the rates of types of exchange for each of the tumors. It was already known and could be compared in the literature, a higher rate of C > T exchanges in all regions, especially in regions of CpG sites<sup>8,13</sup>. It is worth mentioning that the only tumor type that does not fit this rule was the prostate tumor, where a greater amount of C > A exchanges was found instead of the expected C > T. However, we did not observe this in the work of Santosh Yadav, where they observed just the opposite, a greater proportion of C > T exchanges in relation to the others<sup>18</sup>. This result may have been due to the low amount of exchanges found in CpG sites of this tumor (a total of 22 exchanges), which may have generated a bias in this situation in specific terms, having a higher proportion of C > A, but as the number of mutations is low, perhaps when looking at the other positions of the exome, we will have a different result. It is also interesting to note that the variation in methylation of samples from normal tissues to tumor tissues at CpG sites where there were C > T exchanges was small, that is, in most cases, methylation remained constant in samples of normal and tumor tissues. This result can be purchased with the results found by LONG, Mark D. and collaborators who showed that in TCGA prostate cancer data, where positive and negative correlations were found between methylation at CpG sites and gene expression of a set of genes involved in DNA methylation and with the tumor itself, suggesting direct and indirect events, leading to an incomplete understanding of the event<sup>6</sup>. In contrast to the results by ZHOU, Dan, and collaborators who found 53 somatic mutations in CpG sites for colon tumors, it was observed that methylation levels

were lower than 50 of the 53 in tumors in relation to normal paired samples, suggesting that this decrease was caused by the loss of cytosine<sup>19</sup>.

The genes that undergo mutation due to deamination are related to biological processes that can contribute to tumorigenesis, such as metabolic processes, cell growth, and proliferation. These results suggest that mutations in methylated cytosines contribute to the genomic instability of tumor cells, resulting in genes being associated with tumorigenesis.

When analyzing the enriched gene pathways, we can see that several of them may somehow be involved with the progression of tumorigenesis. In breast (TCGA-BRCA) and ovarian TCGA-UCEC tumors, the enriched pathway is MAPK (JNK) signaling, one of the main signaling pathways of the protein kinase responsible for the control of several cellular processes, including proliferation, embryonic development, and apoptosis<sup>20</sup>. Therefore, abnormal MAPK signaling can lead to increased or uncontrolled cell proliferation and resistance to apoptosis, major factors for cancer development and progression, and changes in this pathway have already been reported in human cancer as a result of abnormal activation of receptor tyrosine kinases or gain of function mutations mainly in the RAS or RAF genes<sup>21</sup>. It is also possible to identify evidence that suggests that modulating MKP-1 activity may be a viable option to make breast cancer chemotherapy more effective<sup>22</sup>. Liu et al. (2019) suggested that changes in this pathway may be related to estrogen receptors, which makes sense that this pathway is enriched in breast and ovarian cancer<sup>23</sup>. In TCGA-KIRC, the Ski complex pathway, which is an oncogenic protein that acts as a TGF- $\beta$  repressor, prevents the transcription of related genes. This protein acts as an oncoprotein in melanoma and esophageal cancer<sup>24</sup>.

Evidence also suggests that changes in the MAPK (JNK) pathway are also related to kidney cancer, as found in our results. Ski would accelerate the progression of kidney cancer by attenuating the signaling of transforming growth factor  $\beta$ <sup>25</sup>. In TCGA-KIRP, the polyamine biosynthesis pathways were enriched. These molecules are involved in many fundamental processes of cell growth and survival. In cancer, polyamine metabolism is often deregulated, generally indicating that high levels of polyamide are necessary for tumor transformation and progression<sup>26,27</sup>. Another enriched pathway for the TCGA-KIRP tumor was the rescue of methionine, with the dependence of cancer cells on exogenous methionine. Non-tumorigenic cell lines have the same proliferation rate in media containing methionine or media in which methionine is replaced by the immediate metabolic precursor homocysteine, occurring in several tumors, including that of the kidney<sup>28</sup>. For TCGA-LUAD, the aminoacyl-tRNA synthetase pathway was enriched. This pathway forms a complex of macromolecular proteins with three auxiliary factors, called cancer-associated multifunctional proteins<sup>29</sup>. For TCGA-LUSC, the SCF complex pathways were enriched. This complex has important roles in the ubiquitination of proteins involved in the cell cycle<sup>30</sup>. It is interesting to note that SCF complexes have become an attractive anti-cancer target because of their positive regulation in some human cancers and their biochemically distinct active sites<sup>31</sup>. In TCGA-PAAD, the Wnt signaling pathway has been enriched<sup>32</sup>. This pathway regulates several phenomena and events during embryonic development, with organogenesis, differentiation, polarization, and cell migration. Recently, the Wnt pathway has been linked to stem cell renewal in carcinogenesis and has been described most prominently for colorectal cancer<sup>33</sup>.

## Conclusions

Based on the results shown in this work, we verified the reports present in the existing literature that the exchange rates of C > T were significantly higher. When analyzing the exchanges that were probably caused by the deamination mechanism, we realized that the major functional impact that this mechanism can trigger is that a significant part of the genetic pathways affected by pathogenic exchanges are directly related to cancer. Therefore, it is clear that this mechanism affects the development of tumorigenesis. More studies need to be carried out to better understand and deepen the model established here.

## References

1. Kim, S., & Kaang, B. K. (2017). Epigenetic regulation and chromatin remodeling in learning and memory. *Experimental & molecular medicine*, 49(1), e281-e281.
2. Jaenisch, R., & Bird, A. (2003). Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature genetics*, 33(3), 245-254.
3. FELSENFELD, Gary. A brief history of epigenetics. **Cold Spring Harbor perspectives in biology**, v. 6, n. 1, p. a018200, 2014.
4. Tirado-Magallanes, R., Rebbani, K., Lim, R., Pradhan, S., & Benoukraf, T. (2017). Whole genome DNA methylation: beyond genes silencing. *Oncotarget*, 8(3), 5629.
5. Illingworth, R. S., & Bird, A. P. (2009). CpG islands—'a rough guide'. *FEBS letters*, 583(11), 1713-1720.
6. Long, M. D., Smiraglia, D. J., & Campbell, M. J. (2017). The genomic impact of DNA CpG methylation on gene expression; relationships in prostate cancer. *Biomolecules*, 7(1), 15.
7. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., ... & Funke, R. (2001). Initial sequencing and analysis of the human genome.
8. Zhao, Z., & Zhang, F. (2006). Sequence context analysis of 8.2 million single nucleotide polymorphisms in the human genome. *Gene*, 366(2), 316-324.
9. Laird, C. D., Pleasant, N. D., Clark, A. D., Sneed, J. L., Hassan, K. A., Manley, N. C., ... & Stöger, R. (2004). Hairpin-bisulfite PCR: assessing epigenetic methylation patterns

- on complementary strands of individual DNA molecules. *Proceedings of the National Academy of Sciences*, 101(1), 204-209.
10. Rideout, W. 3., Coetzee, G. A., Olumi, A. F., & Jones, P. A. (1990). 5-Methylcytosine as an endogenous mutagen in the human LDL receptor and p53 genes. *Science*, 249(4974), 1288-1290.
  11. Sassa, A., Kanemaru, Y., Kamoshita, N., Honma, M., & Yasui, M. (2016). Mutagenic consequences of cytosine alterations site-specifically embedded in the human genome. *Genes and Environment*, 38(1), 17.
  12. Watson, J. D., Baker, T. A., Bell, S. P., Gann, A., Levine, M., & Losicke, R. (2015). *Biologia molecular do gene*. Artmed Editora.
  13. Abdelfatah, E., Kerner, Z., Nanda, N., & Ahuja, N. (2016). Epigenetic therapy in gastrointestinal cancer: the right combination. *Therapeutic advances in gastroenterology*, 9(4), 560-579.
  14. Weisenberger, D. J., Siegmund, K. D., Campan, M., Young, J., Long, T. I., Faasse, M. A., ... & Koh, H. (2006). CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer. *Nature genetics*, 38(7), 787-793.
  15. Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., ... & Cancer Genome Atlas Research Network. (2013). The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10), 1113.
  16. Colaprico, A., Silva, T. C., Olsen, C., Garofano, L., Cava, C., Garolini, D., ... & Ceccarelli, M. (2016). TCGAAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic acids research*, 44(8), e71-e71.
  17. Salgado, D., Desvignes, J. P., Rai, G., Blanchard, A., Miltgen, M., Pinard, A., ... & Bérout, C. (2016). UMD-predictor: a high-throughput sequencing compliant system for pathogenicity prediction of any human cDNA substitution. *Human mutation*, 37(5), 439-446.
  18. Yadav, S., Anbalagan, M., Baddoo, M., Chellamuthu, V. K., Mukhopadhyay, S., Woods, C., ... & Makridakis, N. (2020). Somatic mutations in the DNA repairome in prostate cancers in African Americans and Caucasians. *Oncogene*, 39(21), 4299-4311.
  19. Zhou, D., Li, Z., Yu, D., Wan, L., Zhu, Y., Lai, M., & Zhang, D. (2015). Polymorphisms involving gain or loss of CpG sites are significantly enriched in trait-associated SNPs. *Oncotarget*, 6(37), 39995.
  20. Keshet, Y., & Seger, R. (2010). The MAP kinase signaling cascades: a system of hundreds of components regulates a diverse array of physiological functions. In *MAP kinase signaling protocols* (pp. 3-38). Humana Press, Totowa, NJ.
  21. Santarpia, L., Lippman, S. M., & El-Naggar, A. K. (2012). Targeting the MAPK–RAS–RAF signaling pathway in cancer therapy. *Expert opinion on therapeutic targets*, 16(1), 103-119.
  22. Haagenson, K. K., & Wu, G. S. (2010). The role of MAP kinases and MAP kinase phosphatase-1 in resistance to breast cancer treatment. *Cancer and Metastasis Reviews*, 29(1), 143-149.
  23. Liu, A., Zhang, D., Yang, X., & Song, Y. (2019). Estrogen receptor alpha activates MAPK signaling pathway to promote the development of endometrial cancer. *Journal of cellular biochemistry*, 120(10), 17593-17601.

24. Heider, T. R., Lyman, S., Schoonhoven, R., & Behrns, K. E. (2007). Ski promotes tumor growth through abrogation of transforming growth factor- $\beta$  signaling in pancreatic cancer. *Annals of surgery*, 246(1), 61.
25. Taguchi, L., Miyakuni, K., Morishita, Y., Morikawa, T., Fukayama, M., Miyazono, K., & Ehata, S. (2019). c-Ski accelerates renal cancer progression by attenuating transforming growth factor  $\beta$  signaling. *Cancer science*, 110(6), 2063.
26. Casero, R. A., Stewart, T. M., & Pegg, A. E. (2018). Polyamine metabolism and cancer: treatments, challenges and opportunities. *Nature Reviews Cancer*, 18(11), 681-695.
27. Murray-Stewart, T. R., Woster, P. M., & Casero Jr, R. A. (2016). Targeting polyamine metabolism for cancer therapy and prevention. *Biochemical Journal*, 473(19), 2937-2953.
28. Stern, P. H., Wallace, C. D., & Hoffman, R. M. (1984). Altered methionine metabolism occurs in all members of a set of diverse human tumor cell lines. *Journal of cellular physiology*, 119(1), 29-34.
29. Kim, S., You, S., & Hwang, D. (2011). Aminoacyl-tRNA synthetases and tumorigenesis: more than housekeeping. *Nature Reviews Cancer*, 11(10), 708-718.
30. Ou, Y., & Rattner, J. B. (2004). The centrosome in higher organisms: structure, composition, and duplication. *International review of cytology*, 238, 119-182.
31. Skaar, J. R., Pagan, J. K., & Pagano, M. (2014). SCF ubiquitin ligase-targeted therapies. *Nature reviews Drug discovery*, 13(12), 889-903.
32. Komiya, Y., & Habas, R. (2008). Wnt signal transduction pathways *Organogenesis* 4: 68-75.
33. Zhan, T., Rindtorff, N., & Boutros, M. (2017). Wnt signaling in cancer. *Oncogene*, 36(11), 1461-1473.
34. Bird, A. P. (1986). CpG-rich islands and the function of DNA methylation. *Nature*, 321(6067), 209-213.
35. Bonnefond, L., Fender, A., Rudinger-Thirion, J., Giegé, R., Florentz, C., & Sissler, M. (2005). Toward the full set of human mitochondrial aminoacyl-tRNA synthetases: characterization of AspRS and TyrRS. *Biochemistry*, 44(12), 4805-4816.
36. Chen, J., Wang, Z., Wang, W., Ren, S., Xue, J., Zhong, L., ... & Zhang, C. (2020). SYT16 is a prognostic biomarker and correlated with immune infiltrates in glioma: A study based on TCGA data. *International Immunopharmacology*, 84, 106490.
37. Clark, S. J., Statham, A., Stirzaker, C., Molloy, P. L., & Frommer, M. (2006). DNA methylation: bisulphite modification and analysis. *Nature protocols*, 1(5), 2353.
38. Deaton, A. M., & Bird, A. (2011). CpG islands and the regulation of transcription. *Genes & development*, 25(10), 1010-1022.
39. Goll, M. G., & Bestor, T. H. (2005). Eukaryotic cytosine methyltransferases. *Annu. Rev. Biochem.*, 74, 481-514
40. International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409, 860-921.
41. Jakkula, E., Melkonniemi, M., Kiviranta, I., Lohiniva, J., Räinen, S. S., Perälä, M., ... & Ala-Kokko, L. (2005). The role of sequence variations within the genes encoding collagen II, IX and XI in non-syndromic, early-onset osteoarthritis. *Osteoarthritis and cartilage*, 13(6), 497-507.

42. Jones, P. A., & Baylin, S. B. (2002). The fundamental role of epigenetic events in cancer. *Nature reviews genetics*, 3(6), 415-428.
43. Kramer, M., Pierredon, S., Ribaux, P., Tille, J. C., Petignat, P., & Cohen, M. (2015). Secretome identifies tenascin-X as a potent marker of ovarian cancer. *BioMed research international*, 2015.
44. Lövkvist, C., Dodd, I. B., Sneppen, K., & Haerter, J. O. (2016). DNA methylation in human epigenomes depends on local topology of CpG sites. *Nucleic acids research*, 44(11), 5123-5132.
45. Nephew, K. P., & Huang, T. H. M. (2003). Epigenetic gene silencing in cancer initiation and progression. *Cancer letters*, 190(2), 125-133.
46. Shiah, Y. J., Fraser, M., Bristow, R. G., & Boutros, P. C. (2017). Comparison of pre-processing methods for Infinium HumanMethylation450 BeadChip array. *Bioinformatics*, 33(20), 3151-3157.
47. Suzuki, M. M., & Bird, A. (2008). DNA methylation landscapes: provocative insights from epigenomics. *Nature Reviews Genetics*, 9(6), 465-476.
48. Valcourt, U., Alcaraz, L. B., Exposito, J. Y., Lethias, C., & Bartholin, L. (2015). Tenascin-X: beyond the architectural function. *Cell adhesion & migration*, 9(1-2), 154-165.
49. Wang, J., Vasaikar, S., Shi, Z., Greer, M., & Zhang, B. (2017). WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. *Nucleic acids research*, 45(W1), W130-W137.
50. Warden, C. D., Lee, H., Tompkins, J. D., Li, X., Wang, C., Riggs, A. D., ... & Yuan, Y. C. (2013). COHCAP: an integrative genomic pipeline for single-nucleotide resolution DNA methylation analysis. *Nucleic acids research*, 41(11), e117-e117.
51. Wong, D. J., Paulson, T. G., Prevo, L. J., Galipeau, P. C., Longton, G., Blount, P. L., & Reid, B. J. (2001). p16INK4a lesions are common, early abnormalities that undergo clonal expansion in Barrett's metaplastic epithelium. *Cancer research*, 61(22), 8284-8289.
52. Yuan, Y., Nymoén, D. A., Stavnes, H. T., Rossnes, A. K., Bjørang, O., Wu, C., ... & Davidson, B. (2009). Tenascin-X is a novel diagnostic marker of malignant mesothelioma. *The American journal of surgical pathology*, 33(11), 1673.
53. Yu, G., Wang, L. G., Han, Y., & He, Q. Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics: a journal of integrative biology*, 16(5), 284-287.