UNIVERSIDADE DE SÃO PAULO
FACULDADE DE MEDICINA DE RIBEIRÃO PRETO

RODRIGO GUARISCHI MATTOS AMARAL DE SOUSA

ABORDAGEM COMPUTACIONAL PARA IDENTIFICAR NOVOS
SNVs EM BASE DE DADOS DE ESTs

RIBEIRÃO PRETO - SP

2012

RODRIGO GUARISCHI MATTOS AMARAL DE SOUSA

Abordagem computacional para identificar novos SNVs em bases de dados de ESTs

Dissertação apresentada ao Programa de Pós-Graduação
em Genética do Departamento de Genética da Faculdade
de Medicina de Ribeirão Preto da Universidade de São
Paulo para a obtenção do título de Mestre em Ciências.

Área de Concentração: Bioinformática
Orientador: Prof. Dr. Wilson Araújo da Silva Jr

Ribeirão Preto – SP

2012

FICHA CATALOGRÁFICA

DE SOUSA, Rodrigo Guarischi Mattos Amaral
Abordagem computacional para identificar novos SNVs em base de dados de ESTs.
Ribeirão Preto, São Paulo, 2012

79 p.: il ; 30cm

Dissertação de Mestrado, apresentada à Faculdade de Medicina de Ribeirão Preto / USP - Área de concentração: Genética.

Orientação: DA SILVA JR, Wilson Araújo.

1. Mutação de ponto. 2. Polimorfismo. 3. Abordagem computacional em EST (ORESTES).

Dedicatória

Dedico esta dissertação,
A meus pais Eduardo Amaral Carvalho de Sousa e Kátia Guarischi Mattos Silva de Sousa
que em todos os momentos, de felicidade e também nas dificuldades, sempre me
ajudaram com muita paciência e sobriedade e que sempre me incentivam em minhas decisões.
A minha irmã e amiga Renata Guarischi Mattos Amaral de Sousa
que tanta influência tem sobre a minha pessoa.
Aos meus queridos avós maternos
Levy Mattos Silva e Lia Guarischi Mattos Silva
que tanto fizeram e fazem, e sem dúvida são fundamentais para mim.
A memória de meus avós paternos,
Olavo Silva Sousa e Marina do Amaral Carvalho de Sousa
A todos os meus tios e tias, em especial às minhas tias
Mônica Guarischi Mattos Silva de Sousa e
Marina Amaral Carvalho de Sousa
Ao pessoal da Rep. Sparta Ribeirão por todos os momentos loucos
e alegres que passei na república mais top da cidade.

Agradecimentos

Ao Prof. Dr. Wilson Araújo da Silva Junior pela orientação e incentivo.

Ao pessoal do Laboratório de Bioinformática (BiT) por toda a paciência e confiança depositada em mim ao longo desses cinco anos em que lá estive e incalculável conhecimento que me foi passado. Em especial ao Dr Israel Tojal da Silva que me coorientou nas etapas desse projeto realizadas durante a graduação, ao Dr Jorge Estefano S. de Souza que contribuiu intensamente nas etapas finais deste trabalho, ao Dr Daniel G. Pinheiro, ao doutorando Thiago Yukio e ao administrador de sistemas Marcelo Gomes, sem os quais não seria possível a realização do trabalho.

A todos da equipe do Laboratório de Genética Molecular por todos os momentos agradáveis que me proporcionaram, em especial à doutoranda Rafaela de Barros e Lima Bueno que contribuiu diretamente na realização desse projeto.

À Meire Tarla pela ajuda e dedicação nas tarefas burocráticas e administrativas.

A todos os professores das Faculdades de Filosofia e de Medicina que tive durante estes indescritíveis seis anos, quatro anos de graduação e dois de mestrado, que, juntamente com meus colegas do Laboratório são responsáveis por todo o meu conhecimento profissional.

A todos que contribuíram direta ou indiretamente com este projeto, mas não citei, têm meu profundo agradecimento.

Sumário

Lista de Figuras

Lista de Tabelas

Resumo

Abstract

Lista de abreviaturas e siglas

1	Introdução	p. 15
1.1	Introdução a genética molecular	p. 15
1.1.1	Fundamentos e conceitos básicos	p. 15
1.1.2	Fluxo da informação gênica	p. 16
1.1.3	Enovelamento de proteínas	p. 19
1.2	Ensaio baseado em RNA	p. 20
1.2.1	Bibliotecas de cDNA	p. 20
1.2.2	Análise do Transcriptoma	p. 21
1.3	Mutações	p. 22
1.3.1	Definição e tipos	p. 22
1.3.2	Mutação e Polimorfismo	p. 23
1.3.3	Patologias de base genética	p. 23
1.3.4	Identificação de variações genéticas <i>in silico</i>	p. 26

1.3.5	Banco de dados de polimorfismos	p. 27
1.4	Motivação	p. 28
1.5	Organização do trabalho	p. 28
2	Objetivos	p. 29
3	Material e Métodos	p. 30
3.1	Material	p. 31
3.1.1	Base de dados de ORESTES	p. 31
3.1.2	Base de dados de CCDS (<i>Consensus Coding DNA Sequences</i>)	p. 31
3.1.3	Base de dados de polimorfismos	p. 32
3.2	Metodologia	p. 32
3.2.1	Modelagem do Banco de Dados	p. 32
3.2.2	Identificações dos SNVs	p. 33
3.2.3	Validação experimental	p. 35
3.2.4	Padronização da PCR	p. 36
3.2.5	<i>High Resolution Melt</i> - HRM	p. 37
3.2.6	Sequenciamento de DNA	p. 38
3.2.7	Análise do Sequenciamento	p. 38
3.2.8	Análise de interação Proteína-Proteína	p. 39
4	Resultados e Discussão	p. 41
4.1	Banco de dados dbSNP-HCGP	p. 41
4.2	Identificação dos SNVs	p. 41
4.2.1	Validação experimental	p. 44
4.2.2	Análise de interações Proteína-Proteína	p. 46
5	Conclusão	p. 50

Referências Bibliográficas	p. 51
Apêndice A – Triagem por HRM e Sequenciamento	p. 55
Apêndice B – Análise de PPI	p. 67

Lista de Figuras

1	A estrutura do DNA	p. 16
2	Dogma Central da Biologia Molecular	p. 17
3	<i>Splicing</i> alternativo	p. 18
4	Representação esquemática do mRNA	p. 18
5	Esquema da polimerização peptídica em um ribossomo	p. 19
6	Enovelamento proteico	p. 19
7	Etapas de construção de bibliotecas de cDNA	p. 20
8	Distribuição posicional das ORESTES no gene	p. 21
9	Imagens de microscopia dos glóbulos vermelhos	p. 24
10	Estrutura da hemoglobina S	p. 24
11	Cromossomo Filadélfia	p. 25
12	Tratamento farmacológico da Leucemia Mieloide Crônica	p. 26
13	Resistência ao tratamento farmacológico da Leucemia Mieloide Crônica	p. 26
14	Fluxograma do projeto	p. 30
15	<i>Pipeline</i> da identificação de SNVs	p. 33
16	Filtro de regiões parálogas	p. 34
17	Caracterização dos SNVs identificados	p. 35
18	Exemplo de curva de melt	p. 38
19	Diagrama Entidade-Relacional do banco dbSNP-HCGP	p. 42
20	Validação <i>in-silico</i> de acordo número de bibliotecas distintas	p. 43
21	Teste da padronização da PCR	p. 45
22	PPI de mama	p. 47

23	PPI de Sistema Nervoso Central	p. 48
24	Resultado da investigação do SNV no gene ANP32C	p. 57
25	Resultado da investigação do SNV no gene HSPA8	p. 58
26	Resultado da investigação do SNV no gene HTATIP2	p. 59
27	Resultado da investigação do SNV no gene KRT7	p. 60
28	Resultado da investigação do SNV no gene MYL1	p. 61
29	Resultado da investigação do SNV no gene ALDOA	p. 62
30	Resultado da investigação do SNV no gene CTSB	p. 63
31	Resultado da investigação do SNV no gene EEF1A1	p. 64
32	Resultado da investigação do SNV no gene EIF4G1	p. 65
33	Resultado da investigação do SNV no gene EPCAM	p. 66
34	Interações dos genes com SNVs no tecido de mama normal (BN)	p. 68
35	Interações dos genes com SNVs no tecido de mama tumor (BT)	p. 68
36	Interações dos genes com SNVs no tecido de pulmão normal (EN)	p. 69
37	Interações dos genes com SNVs no tecido de pulmão tumor (ET)	p. 69
38	Interações dos genes com SNVs no tecido de próstata normal (FN)	p. 70
39	Interações dos genes com SNVs no tecido de próstata tumor (FT)	p. 70
40	Interações dos genes com SNVs no tecido de sistema nervoso normal (NN)	p. 71
41	Interações dos genes com SNVs no tecido de sistema nervoso tumor (NT)	p. 71
42	Interações dos genes com SNVs em regiões que dão origem aos domínios proteicos no tecido de mama normal (BN)	p. 72
43	Interações dos genes com SNVs em regiões que dão origem aos domínios proteicos no tecido de mama tumor (BT)	p. 72
44	Interações dos genes com SNVs em regiões que dão origem aos domínios proteicos no tecido de pulmão normal (EN)	p. 73
45	Interações dos genes com SNVs em regiões que dão origem aos domínios proteicos no tecido de pulmão tumor (ET)	p. 73

46	Interações dos genes com SNVs em regiões que dão origem aos domínios proteicos no tecido de próstata normal (FN)	p. 74
47	Interações dos genes com SNVs em regiões que dão origem aos domínios proteicos no tecido de próstata tumor (FT)	p. 74
48	Interações dos genes com SNVs em regiões que dão origem aos domínios proteicos no tecido de sistema nervoso normal (NN)	p. 75
49	Interações dos genes com SNVs em regiões que dão origem aos domínios proteicos no tecido de sistema nervoso tumor (NT)	p. 75
50	Interações dos genes com SNVs danosos em regiões que dão origem aos domínios proteicos no tecido de mama normal (BN)	p. 76
51	Interações dos genes com SNVs danosos em regiões que dão origem aos domínios proteicos no tecido de mama tumor (BT)	p. 76
52	Interações dos genes com SNVs danosos em regiões que dão origem aos domínios proteicos no tecido de pulmão normal (EN)	p. 77
53	Interações dos genes com SNVs danosos em regiões que dão origem aos domínios proteicos no tecido de pulmão tumor (ET)	p. 77
54	Interações dos genes com SNVs danosos em regiões que dão origem aos domínios proteicos no tecido de próstata normal (FN)	p. 78
55	Interações dos genes com SNVs danosos em regiões que dão origem aos domínios proteicos no tecido de próstata tumor (FT)	p. 78
56	Interações dos genes com SNVs danosos em regiões que dão origem aos domínios proteicos no tecido de sistema nervoso normal (NN)	p. 79
57	Interações dos genes com SNVs danosos em regiões que dão origem aos domínios proteicos no tecido de sistema nervoso tumor (NT)	p. 79

Lista de Tabelas

1	ESTs geradas no HCGP	p. 31
2	Relação das amostras utilizadas na validação experimental	p. 36
3	Protocolo utilizado na amplificação das regiões de interesse	p. 37
4	Programa de termociclagem utilizado na reação de PCR.	p. 37
5	SNVs selecionados para validação	p. 44
6	Lista de <i>primers</i> desenhados.	p. 44
7	Tabela resumo dos resultados da validação experimental dos SNVs	p. 45
8	Genótipos observados nas amostras de genes normais e tumorais	p. 55
9	Genótipos observados nas amostras de genes tumorais	p. 56

*”Com grande poder,
deve também vir grande responsabilidade”.*

Stanley Martin Lieber

Resumo

Abordagem computacional para identificar novos SNVs em bases de dados de ESTs

Indivíduos não relacionados apresentam apenas 1% de diferenças entre seus genomas. Essas variações ocorrem na forma de substituições, inserções, deleções, rearranjos complexos ou até estruturais. Dentre essas variações, aquelas que apresentam uma frequência populacional acima de 1% são denominadas de polimorfismos. Tais variações são responsáveis por diferenças que vão desde a resposta imunológica até o tratamento com drogas, incluindo sensibilidade das células tumorais, níveis de plasma, efeitos colaterais e toxicidade. A forma mais comum de polimorfismo genético entre humanos são os polimorfismo de base única ou Single Nucleotide Polymorphisms (SNPs), sendo mais de 47 milhões descritos no dbSNP, um banco de dados de pequenos polimorfismos do NCBI. No presente estudo, foi estabelecida uma abordagem computacional, com etapas de exclusão de regiões parálogas ou de baixa qualidade, com o objetivo de identificar variantes genéticas em sequências expressas gerados pelo método de Open Reading Frame ESTs (ORESTES) durante o Projeto Genoma Humano do Câncer. Diferentemente de outros *softwares* de detecção de polimorfismos, a abordagem computacional descrita neste estudo leva em consideração a informação *a priori* do número de bibliotecas distintas que reportaram a mesma variação. Foram identificadas 1900 mutações (853 sinônimas e 1047 não-sinônimas) presentes em duas ou mais bibliotecas distintas, que foram validados *in-silico* contra o dbSNP v130. O resultado da análise identificou 901 mutações já descritas no dbSNP (47,42%). Para confirmação da análise, foram selecionadas 10 mutações (6 novas e 4 já presentes no dbSNP) para validação pelo método de High Resolution Melt (HRM), seguido da caracterização por sequenciamento de DNA. Nesse caso, o resultado foi a validação de 50% das mutações selecionadas. A análise de interação proteica, *Protein-Protein Interaction* (PPI), realizada com as mutações não-sinônimas localizadas em domínios funcionais, revelou redes gênicas mais complexas em tecidos tumorais do que nos tecidos normais. Esta observação ratificou a literatura a respeito da transformação tumorigênica ser desencadeada pela combinação de mutações que ativam uma série de processos biológicos, para isso, afetando genes, vias gênicas e *networks* de vias gênicas relacionados. Em resumo, o presente estudo descreve uma abordagem computacional eficiente para identificação de mutações em dados de sequências expressas, além de avaliar o papel das mutações na tumorigênese.

Abstract

Computational approach to identify new SNVs in ESTs data set

Unrelated humans have only 1% of non-similarity in their genome. These variations occur as substitutions, insertions, deletions, or even complex structural rearrangements. Among these variations, those which show a population frequency above 1% are called polymorphisms. Such variations are responsible for differences ranging from the immune response to treatment with drugs, including sensitivity of tumor cells, plasma levels, toxicity and side effects. The most common form of genetic polymorphism among human are Single Nucleotide Polymorphisms (SNPs), with more than 47 million reported in dbSNP, a database of small polymorphisms from NCBI. In this study, we established a computational approach, with steps to exclude low quality and paralogous regions, aiming to identify genetic variants in expressed sequences generated by the method of Open Reading Frame ESTs (ORESTES) for the Human Cancer Genome Project. Unlike other polymorphisms detection softwares, the computational approach described in this study takes into account the *a priori* information about the number of different libraries that reported the same variation. We identified 1900 mutations (853 synonymous and 1047 non-synonymous) present in two or more different libraries, these mutations were in-silico validated against the dbSNP V130. The analysis result showed 901 mutations already described in dbSNP (47.42%). To confirm the analysis, we selected 10 mutations (six new and four already present in dbSNP) for validation by the method of High Resolution Melt (HRM), followed by characterization by DNA sequencing. In this case, the result was the validation of 50 % of the selected mutations. The Protein-Protein Interaction analysis (PPI), performed with non-synonymous mutations located in functional domains, showed more complex gene networks in tumor tissues than in normal tissues. This observation confirmed the literature regarding the tumorigenic transformation is triggered by the combination of mutations that activate a number of biological processes, thereby, affecting genes, gene pathways and networks of related gene pathways. In summary, this study describes an efficient computational approach to identify mutations in expressed sequence data, besides to evaluate the role of mutations in tumorigenesis.

Keywords: Point mutations, Polymorphism, Computational approach in ESTs

Lista de abreviaturas e siglas

DNA	Ácido desoxirribonucleico
exons	Sequências do RNA primário que compõem o RNA maduro
íntrons	regiões não codificadoras situadas entre dois exons
splicing	retirada de íntrons do RNA primário
mRNA	RNA mensageiro
tRNA	RNA transportador
ESTs	Etiqueta de Sequência Expressa
ORESTES	<i>Open Reading Frame ESTs</i>
SNV	Variação de Base Única
SNP	Polimorfismo de Base Única
LMC	Leucemia mielóide crônica
TKI	Inibidores de tirosino-quinase
gDNA	DNA genômico
NCBI	<i>National Center for Biotechnology Information</i>
CCDS	sequências consenso de DNA codificante
Ensembl	<i>European Bioinformatics Institute & Wellcome Trust Sanger Institute</i>
UCSC	<i>University of California, Santa Cruz</i>
SGBD	Sistema Gerenciador de Banco de Dados
HRM	High Resolution Melt
PCR	Reação em Cadeia da Polimerase

1 Introdução

O estudo apresenta uma abordagem computacional para identificação de variações genéticas de base única em um banco de dados de sequências expressas geradas no Projeto Genoma Humano do Câncer. A Seção 1.1 apresenta os conceitos fundamentais de genética molecular. A seção 1.2 descreve brevemente a natureza e fonte dos dados utilizados. A Seção 1.3 fornece uma visão geral sobre o estudo das variações genéticas e sua importância. A Seção 1.4 descreve a motivação para o desenvolvimento deste projeto. A Seção 1.5 descreve os capítulos dessa dissertação.

1.1 Introdução a genética molecular

1.1.1 Fundamentos e conceitos básicos

Todo organismo vivo possui um material genético que é passado à sua prole, na maioria dos casos esse material se apresenta na forma de ácido desoxirribonucleico (DNA). A compreensão dessa intrigante molécula deve-se ao estudo de inúmeros pesquisadores, os quais merecem destaque: Friedrich Miescher (1869) pioneiro no isolamento dessa molécula que inicialmente designou como "nucleína", Griffith (1928) pela formulação da teorias do princípio transformante, Avery, Macleod e McCarty (1944) e Hershey e Chase (1952) que comprovaram que o DNA era o condutor da informação genética, bem como Chargaff (1950), que observou a proporcionalidade entre purinas e pirimidinas, e Watson e Crick (1953) que definiram a estrutura da molécula de DNA, formada por duas fitas que se ligam em orientações opostas. Cada uma das fitas é composta por sub-unidades simples chamadas **nucleotídeos**, que são compostos por uma molécula de açúcar, uma molécula de fosfatos e uma das quatro possíveis **bases nitrogenadas**: Adenina (A), Timina (T), Citosina (C) ou Guanina (G).

A ligação entre os monômeros adjacentes na fita é feita por ligações fosfodiéster entre os açúcares e fosfatos que se dispõem de maneira intercalada, criando uma espécie de estrutura de sustentação. A ligação entre as fitas dá-se através de pontes de hidrogênio feitas pelos monômeros complementares em cada uma das fitas (Figura 1-A). O DNA normalmente se en-

contra em forma de um espiral dextrogiro onde tipicamente a cada dez bases ocorre um giro completo formando dois sulcos, um sulco maior com 22Å e outro menor com 12Å (Figura 1-B).

Figura 1: A estrutura do DNA. (A) Disposição anti-paralela da fita devido a direção de síntese oposta 5' → 3' e 3' → 5'. (B) Estrutura convencional do DNA, formado por uma espiral dextrógiro com dois sulcos: o maior com 22Å e outro menor com 12Å. Figura adaptada de Strachan e Read (2001).

O material genético de um indivíduo está organizado em longos polímeros de DNA, os **cromossomos**, e o conjunto deles é denominado **genoma**. Dentro de um genoma apenas uma parte é responsável por codificar moléculas de RNA, tais trechos são chamados **genes** e são as unidades fundamentais da hereditariedade.

1.1.2 Fluxo da informação gênica

Com a confirmação de que o DNA era responsável pela transmissão da informação hereditária pelos experimentos de Hershey e Chase em 1952, restava agora determinar qual a sequência com que as moléculas interagem até a produção da proteína, dentro de um circuito gênico. A compreensão deste modelo que caracterizava o fluxo das informações contidas em uma região do DNA coube a Crick (1970), que definiu tal fluxo como **“Dogma Central da Biologia Molecular”**.

Em sua concepção, o Dogma Central considerava que o fluxo da informação genética se dava no sentido onde um trecho do DNA dá origem a uma molécula de RNA e esta por sua vez, origina uma molécula de proteína. No conceito do Dogma Central também está a informação de que o DNA se replica dando origem a duas moléculas idênticas.

Entretanto, este dogma não se manteve inalterado e foi sistematicamente ampliado com os conhecimentos adquiridos graças a avanços de métodos de biologia molecular. Esses avanços

permitiram uma análise mais refinada dos processos biológicos ativos em uma célula e trouxe como consequência a incorporação de modificações tais como: RNA $\rightarrow$ DNA através da enzima transcriptase reversa (BALTIMORE, 1970; TEMIN; MIZUTANI, 1970); RNA $\leftrightarrow$ RNA em alguns vírus em plantas (AHLQUIST, 2002); DNA $\rightarrow$ Proteína (MCCARTHY; HOLLAND, 1965) (Figura 2).

Figura 2: Dogma Central da Biologia Molecular, em preto o dogma como em sua concepção por Crick (1970), em laranja as ampliações que ele sofreu.

Apesar de importantes, tais modificações, fogem ao escopo deste trabalho que se aprofundará no dogma em sua visão clássica.

Transcrição Gênica

A molécula de RNA é produzida no núcleo das células de eucariotos pelo mecanismo de **transcrição gênica**. Durante esse processo, um trecho do DNA complementar a região gênica é lida pela enzima **RNA polimerase** o que leva a produção do **RNA primário**.

Em alguns organismos, *e.g.* eucariotos, o RNA primário precisa ser processado para retirada dos **íntrons** (regiões não codificadoras situadas entre dois **exons**) pelo processo de *splicing*, e este, ocorre simultaneamente a outras duas modificações: i) Adição da 7-metilguanosina na extremidade 5' da fita de RNA, chamado de **capping**, e ii) Adição de uma cauda Poli-Adenina na extremidade 3'. Tais modificações dão origem a molécula de **RNA maduro** pronta para ser exportada para o citosol através dos poros nucleares (KNOERS; MONNENS, 2006).

Alguns dos RNAs primários podem ser processados em mais de um tipo de RNA maduro pelo mecanismo de **splicing alternativo**. Este processo é controlado por complexos de proteínas reguladoras (*splicing regulatory* - SR) que reconhecem junções intron/exon e exon/intron e determinam a retirada de introns específicos pelo **spliceossomo** (Figura 3) (BLACK, 2003).

Tradução gênica

Enquanto para alguns RNAs (*non-coding RNA*) a forma do RNA maduro é o produto final do gene (AMARAL *et al.*, 2008), para outros, é apenas um veículo que carrega as instruções

Figura 3: Representação esquemática do mecanismo de *splicing* alternativo. O uso alternativo de exons do RNA primário é o principal mecanismo de produção de isoformas gênicas, e consequentemente, fonte de variabilidade do transcriptoma.

necessárias na síntese proteica, em outras palavras um **RNA mensageiro (mRNA)**. Este RNA mensageiro pode conter a informação para síntese de uma única proteína (**monocistrônico**, comum em eucariotos) ou várias (**policistrônico**, comum em procariotos). Todo mRNA consiste de três partes: *5' untranslated region* (5' UTR), região codificante de proteína e *3' untranslated region* (3' UTR) (Figura 4).

Figura 4: Representação esquemática do mRNA e suas regiões: *capping*, UTR, codificantes e cauda Poli-A.

A região codificante contém a informação necessária para a síntese de proteínas codificada em trincas de nucleotídeos. Cada trinca é chamada **códon** e corresponde a um sítio de ligação complementar ao anticódon de um RNA transportador (**tRNA**). Cada tRNA leva no anticódon sempre o mesmo tipo de aminoácido. Estes aminoácidos são polimerizados no ribossomo de acordo com a ordem das trincas da região codificante. O ribossomo catalisa a ligação do tRNA no mRNA, fazendo a retirada do aminoácido do tRNA e concatenando-o à sequência peptídica nascente (Figura 5) (HANSEN *et al.*, 2003; BERK; CATE, 2007). Cada molécula de mRNA é traduzida em várias moléculas de proteína, sendo que em mamíferos a média é de 900 moléculas (SCHWANHAUSSER *et al.*, 2011).

Em eucariotos, a tradução pode ocorrer em vários lugares dependendo do local onde a proteína exercerá sua função. Em sua maioria ocorre no citoplasma (proteínas solúveis) ou na

Figura 5: Esquema de um ribossomo catalisando a polimerização dos aminoácidos em uma cadeia peptídica.

membrana do retículo endoplasmático (proteínas de membrana ou secretadas).

1.1.3 Enovelamento de proteínas

Quando um mRNA é traduzido no ribossomo, a cadeia polipeptídica assume uma forma não enovelada ou *random coil* (Figura 6-A). Simultaneamente a transcrição, a proteína nascente sofre um processo físico-químico que, dependendo da sequência de aminoácidos, se enovela até assumir uma estrutura tridimensional estável (Figura 6-B). Esta estrutura, por sua vez, está intimamente relacionada com sua função (ANFINSSEN, 1972).

Figura 6: Enovelamento proteico. **A)** Sequência proteica não enovelada ou *random coil*, e **B)** Sequência proteica enovelada com estruturas tridimensionais bem definidas.

A mutação da sequência do mRNA pode levar a substituição do aminoácido a ser traduzido, e esta mudança por sua vez, pode causar problemas no enovelamento e produzir proteínas não funcionais sendo algumas dessas mutações causadoras de doenças conhecidas, tais como, a Anemia Falciforme e a resistência ao tratamento na Leucemia Mieloide Crônica, que serão discutidas posteriormente.

1.2 Ensaios baseados em RNA

Há uma enorme variedade de métodos de biologia molecular que são utilizados de acordo com a pergunta a ser respondida. Para a análise da região transcrita do genoma, o **transcriptoma**, as abordagens metodológicas mais comuns se baseiam na hibridação da molécula de RNA (*e.g.*: *microarrays*) ou no **sequenciamento de cDNA**.

O sequenciamento de cDNA permite a obtenção de informações como: quantificação da expressão diferencial de genes, detecção de mutações pós-transcricionais e identificação de fusões gênicas. Esta metodologia só é possível graças ao uso do **cDNA**.

1.2.1 Bibliotecas de cDNA

Para que segmentos gênicos possam ser sequenciados são necessárias múltiplas cópias da mesma região, tal procedimento é denominado **amplificação**. No transcriptoma esse procedimento é possível graças a conversão da molécula de RNA em **DNA complementar (cDNA)**.

A construção de uma biblioteca de cDNA, requer os seguintes passos: inicialmente o RNA é extraído e purificado (Figura 7-1) e através de uma reação catalisada pelas enzimas transcriptase reversa e DNA polimerase, as moléculas de cDNA são sintetizadas (Figura 7-2). O material é então convertido em DNA dupla fita (Figura 7-3) e inserido em plasmídeos (Figura 7-4). As células competentes são transformadas (Figura 7-5) para que o material seja amplificado durante a duplicação das bactérias (Figura 7-6). Posteriormente os clones isolados, purificados (Figura 7-7) e sequenciados (BELYAVSKY; VINOGRADOVA; RAJEWSKY, 1989).

Figura 7: Etapas de construção de bibliotecas de cDNA. Figura adaptada de: www.molecularstation.com.

1.2.2 Análise do Transcriptoma

A confecção de bibliotecas de cDNA permitiu a análise parcial do transcriptoma de um tecido. Esses transcritos que eram obtidos pelo método de cDNA foram denominados de **Eti-quetas de Sequências Expressas (ESTs)**. Esta abordagem foi descrita no início da década de 90, quase que simultaneamente por Wilcox *et al.* (1991) e Okubo *et al.* (1992) e possibilitou a identificação dos genes expressos nos tecidos estudados. Contudo, as ESTs eram obtidas somente das extremidades 5' ou 3', faltando informações da região central dos mRNA.

Na década seguinte, uma nova abordagem de sequenciamento de ESTs, descrita por Dias-Neto *et al.* (2000), foi desenvolvida para privilegiar a região central dos mRNAs. Essa metodologia foi denominada **ORESTES** (*Open Reading Frame ESTs*) e usava a amplificação randômica sem ancoragem das extremidades do mRNA.

O método de ORESTES baseia-se na amplificação do mRNA com *primers* não específicos que, por chance, tem a capacidade de aumentar a representatividade dos transcritos raros e da região central dos genes (Figura 8-A).

Figura 8: Distribuição posicional das ORESTES no gene. **A)** Posição relativa das ORESTES dentro do gene, **B)** Comparação da posição relativa, no gene, das ORESTES e outras técnicas de ESTs. Figura obtida de Dias-Neto *et al.* (2000).

A figura 8-B ilustra a comparação da distribuição da posição relativa das ESTs com os métodos de ORESTES e as técnicas de obtenção de ESTs das extremidades 5' e 3', confirmando a preferência das ORESTES pela posição central do gene.

1.3 Mutações

1.3.1 Definição e tipos

O DNA é uma molécula estável. De maneira geral, sua replicação se dá de maneira correta, porém, eventualmente podem ocorrer erros durante esse processo. Alterações hereditárias do material genético de um organismo, decorrentes de erros de replicação durante a divisão celular e não causadas por recombinação ou segregação, são denominadas **mutações** (BERTRAM, 2000).

Dependendo da célula que ocorrem, as mutações são divididas em duas classes:

- **Somáticas:** Podem acarretar prejuízo ao indivíduo portador. Em adultos, podem causar tumores e outras lesões degenerativas. Se atingem a célula-ovo, embrião ou feto, podem causar mosaicismo, sendo que o grau deste último dependerá do período do desenvolvimento em que as mutações ocorreram. Se atingirem as células germinativas primordiais, que originarão os gametas do indivíduo, poderão diminuir sua fertilidade e ainda transmitir o dano a futuras gerações.
- **Gaméticas:** Não causam prejuízo ao portador, no máximo, podem acarretar diminuição de sua fertilidade. Porém são transmitidas às gerações futuras.

A extensão das alterações genômicas pode variar de uma única ou poucas bases (pequena escala) até centenas de milhares (grande escala). Dentro dessas escalas, as variações se diferem quanto ao tipo da seguinte forma:

- **Pequena escala:** Afetam poucos nucleotídeos
 - **Mutações de ponto:** há a troca de um único nucleotídeo por outro, geralmente são causada por substâncias mutagênicas ou erros na replicação do DNA
 - **Inserções:** adição de um ou mais nucleotídeos na sequência de DNA
 - **Deleções:** remoção de um ou mais nucleotídeos, e juntamente com as inserções podem promover uma alteração na janela de leitura do mRNA
- **Grande escala:** Alteram uma grande sequência de nucleotídeos
 - **Amplificações:** criação de várias cópias de uma região gênica
 - **Deleções:** levam a perda dos genes contidos na região

- **Translocações:** troca de segmentos de DNA entre cromossomos não homólogos
- **Inversão:** troca da orientação da sequência gênica
- **Perda de Heterozigose:** Perda de um alelo por deleção ou recombinação

Existem diversas classificações que definem uma mesma mutação, dependendo da característica que se pretende observar, tais como: **efeito funcional**, **efeito adaptativo**, ou ainda quanto ao **impacto na sequência proteica**. Entretanto uma revisão pormenorizada destas classificações foge ao escopo deste trabalho que se concentrará na identificação de mutações de ponto.

1.3.2 Mutação e Polimorfismo

Uma variação pode ser **frequente** em uma dada população, **rara** ou ainda **exclusiva** de um único indivíduo. Para se qualificar como **variação polimórfica** (do grego, *poli* = muitas, *morphos* = formas) tem de haver algum tipo de equilíbrio entre as formas (**alelos**), e este deve ser sustentado por herança. O critério usado para conceituar o termo polimorfismo, se baseia no fato de que a frequência do alelo seja elevado o suficiente para não ser eliminado nas gerações seguintes. Por definição, um alelo é polimórfico quando sua frequência na população é igual ou superior a 1% (SHEPPARD, 1975).

A mutação de ponto, também chamada de **Variação de Base Única** (SNV), constitui a forma mais comum de alteração genética entre humanos, sendo as variantes polimórficas, neste caso chamado de **Polimorfismo de Base Única** (SNP), estimados a perfazerem um total entre 10 e 12 milhões na população humana (BUNZ, 2008). Entretanto, o banco de dados de polimorfismos de base única do NCBI, o dbSNP, em sua versão 135 já contava com mais de 47 milhões mutações de base única em humanos e, portanto, o número total permanece incerto ¹.

Quando comparamos o genoma de dois indivíduos não relacionados, estes apresentam cerca de 99% de similaridade e é o 1% de não similaridade, grande parte composta por SNPs, que nos faz únicos e tão fenotipicamente diferentes. No entanto, alguns desses polimorfismos genéticos são responsáveis por uma infinidade de patologias.

1.3.3 Patologias de base genética

Existe uma lista extensa de doenças causadas por desordens genéticas, algumas são monogênicas (causadas por um só gene) e outras, mais complexas, poligênicas (causadas por vários

¹www.ncbi.nlm.nih.gov/projects/SNP/snp_summary.cgi

genes). Foge ao escopo deste trabalho fazer uma revisão minuciosa destes tipos de patologias, no entanto, serão exemplificadas duas delas como ilustração.

Anemia falciforme

Anemia falciforme é uma doença hereditária comum na África, Europa mediterrânea, Oriente Médio e regiões da Índia, causada por uma mutação no gene beta-globina que leva a deformação da estrutura das hemácias, fazendo com que estas assumam forma semelhante à foice (Figura 9) (PLATT *et al.*, 1994).

Figura 9: Imagens de microscopia eletrônica de varredura dos glóbulos vermelhos. **A)** Eritrócito normal, figura adaptada de www.nisenet.org, **B)** Eritrócito falcizado, figura adaptada de www.sciencephoto.com.

Os pacientes acometidos pela anemia falciforme possuem uma substituição em homozigose de um único nucleotídeo da cadeia β da hemoglobina levando a troca de um Ácido Glutâmico pela Valina (GAG→GTG: Glu6Val), produzindo a hemoglobina S. É uma doença com herança recessiva que só se manifesta em homozigose. Nos heterozigotos, que são assintomáticos, coabitam dois tipos de hemoglobinas, a Hb A e Hb S (Figura 10-A).

Figura 10: Estrutura da hemoglobina S. **A)** Estrutura cristalizada de alta definição da desoxi-Hb S (Azul = cadeias α , laranja = cadeias β , vermelho = grupamentos heme e ciano = mutação Glu6Val nas cadeias β ; figura adaptada da estrutura PDB ID: 2HBS). **B)** Microscopia eletrônica da célula falciforme com grande número de fibras se dispondo no sentido longitudinal da célula, figura obtida em www.ciencianews.com.br.

Em certas condições de baixa tensão de oxigênio ou alta concentração de Hb-S. Estas hemoglobinas se polimerizam em forma de bastões (Figura 10-B), o que torna a hemácia menos maleáveis e a fazem ficar presas dentro de vasos sanguíneos pequenos, obstruindo-os e causando dor e eventuais danos aos órgãos. Nesse estado, as hemácias são destruídas no baço, gerando um quadro anêmico grave no portador (NCBI, 1998).

Leucemia mieloide crônica

Leucemia mielóide crônica (LMC) é uma desordem de células-tronco da medula óssea causada pela expansão clonal de granulócitos maduros e seus precursores.

A patologia é resultado de uma translocação recíproca entre os cromossomos 9 e 22 **t(9;22)**, dando origem a um cromossomo pequeno chamado de cromossomo Filadélfia ou cromossomo Ph1. Esta alteração produz uma proteína híbrida fruto da fusão dos genes **ABL** e **BCR** (Figura 11) (FADERL *et al.*, 1999).

Figura 11: Cromossomo Filadélfia. **A)** Esquema representando a translocação t(9;22), figura adaptada de www.cancer.gov. **B)** Células em metáfase positivas para rearranjo BCR/ABL usando coloração de FISH, figura adaptada de en.wikipedia.org.

A terapia indicada para pacientes portadores de LMC incluem: o transplante de medula óssea, que apesar do potencial curativo ainda apresenta uma taxa de mortalidade alta, e o tratamento farmacológico com drogas inibidoras de tirosino-quinase (TKI - *Tyrosine Kinase Inhibitors*).

A primeira droga TKI foi o **Glivec** (imatinib) desenvolvido no final dos anos 90, que apesar de não erradicar as células LMC, limita muito a expansão clonal do tumor e reduz o risco de evolução para a etapa de “Crise Blástica”. O domínio quinase da proteína BCR/ABL tem um sítio de ligação para o ATP, que é bloqueado pelo imatinib, inibindo a atividade enzimática da proteína por competição (Figura 12).

Figura 12: Tratamento farmacológico da Leucemia Mieloide Crônica. **A)** Esquema representando o mecanismo de ação do imatinib, figura adaptada de en.wikipedia.org. **B)** Estrutura cristalizada de alta definição da proteína fusionada BCR/ABL (azul) em complexo com o imatinib (vermelho), figura adaptada da estrutura PDB ID: 2HYY.

Apesar do tratamento com imatinib ser eficiente, alguns pacientes apresentam resistência ao tratamento com este fármaco, que em muitos casos é explicada por mutações de ponto na região do domínio quinase, com mais de 50 mutações, causadoras de resistência, descritas até o presente momento (GORRE *et al.*, 2001; HOCHHAUS *et al.*, 2002; SHAH *et al.*, 2002; BRANFORD *et al.*, 2002, 2003). Isso levou a indústria farmacêutica a desenvolver novas drogas que são eficazes nos casos de resistência ao imatinib, como por exemplo o dasatinibe e o nilotinibe (Figura 13).

Figura 13: Resistência ao tratamento farmacológico da Leucemia Mieloide Crônica. **A)** Estrutura cristalizada de alta definição da proteína fusionada BCR/ABL (azul) em complexo com o TKI de segunda geração nilotinibe (vermelho), figura adaptada da estrutura PDB ID: 3CS9. **B)** Tabela de resistência aos fármacos de segunda geração de acordo com os genótipos mais comuns de resistência ao imatinib, adaptado de Branford, Melo e Hughes (2009).

1.3.4 Identificação de variações genéticas *in silico*

Descrever e compreender as diferenças genéticas entre os genomas é uma tarefa fundamental para a evolução do conhecimento médico e pavimentar o caminho para estudos que ten-

tam explicar como, pequenas e grandes alterações no genoma são responsáveis por diferentes fenótipos. A busca de mutações nos dados de sequências geradas em larga-escala é um trabalho árduo e oneroso, que é facilitada graças ao uso da computação.

Essencialmente, o processo de identificação de variantes envolve duas etapas: *i*) O alinhamento de sequências, geralmente intercalados por etapas de filtragem, e *ii*) Identificação de variações.

Alinhamento

O **alinhamento**, em geral, envolve pareamento de sequências de cDNA ou DNA genômico (gDNA) com um conjunto de sequências referência: CCDS, Refseq, genomas ou outras sequências de qualidade. Para isso há um amplo conjunto de *softwares* desenvolvidos para tal finalidade, cada um com paradigmas e propósitos diferentes.

Filtros pré e pós alinhamento são muitas vezes necessários a fim de eliminar resultados espúrios, pois tem o objetivo de excluir erros de incorporação ou artefatos gerados pela técnica.

Identificação de variações

A **identificação de variações** compreende uma etapa de mineração das mutações observadas entre as sequências alinhadas. Nesta etapa deve-se combinar todas as informações disponíveis, *e.g.* qualidade da leitura da base, para reduzir ao máximo o número de falsos positivos.

Alguns *softwares* foram desenvolvidos exclusivamente para a detecção de mutações de ponto como por exemplo o **PolyBayes** (MARTH *et al.*, 1999) que é um pacote de programas que permite realizar: *i*) alinhamento múltiplo ancorado, *ii*) filtro de regiões parálogas e, *iii*) identificação de SNPs. O algoritmo de identificação de mutações desse pacote usa inferência Bayesiana para calcular a probabilidade de uma variação ser um polimorfismo verdadeiro em vez de um erro de sequenciamento. Outro programa, o **PolyPhred** (NICKERSON; TOBE; TAYLOR, 1997), identifica as mutações de acordo com a qualidade da leitura da base mutada e das bases flanqueadoras.

1.3.5 Banco de dados de polimorfismos

Com o objetivo de armazenar informações relacionadas as variações já descritas, foram criados alguns bancos de dados de polimorfismos como por exemplo: o **SNPedia** (CARIASO; LENNON, 2012) que é um banco de dados estilo *wiki* que inclui anotações de genomas pes-

soais, interpretação e análise; o **OMIM** (HAMOSH *et al.*, 2002) que descreve as associações entre polimorfismos e doenças, o **HGMD** (COOPER; BALL; KRAWCZAK, 1998) que descreve mutações em genes que causam ou estão associados à doenças humanas hereditárias e SNPs funcionais, ou ainda o **GWAS** (MANOLIO, 2010), onde através do *GWAS Central* é possível consultar resumos de estudos genéticos de associação.

Por último e o mais conhecido, encontra-se o **dbSNP**, um repositório mantido pelo *National Center for Biotechnology Information* (NCBI) que armazena informações de polimorfismos de base única e pequenas deleções e inserções de maneira integrada com outras bases de dados do NCBI. O dbSNP define o termo polimorfismo para qualquer variação, não requerendo assim, a frequência mínima do alelo numa dada população (SHERRY *et al.*, 2001).

1.4 Motivação

Um esforço público e privado substancial está sendo realizado para se caracterizar os SNVs responsáveis pela diversidade genética. A meta atual é identificar, mapear e determinar a frequência alélica de um conjunto de mutações distribuídas ao longo do genoma. Estes SNVs, são usados para identificar genes ligados a desordens genéticas, ao metabolismo de drogas, e no estudo da diversidade das populações (SCHIFFREEN; STORTS; BULLER, 2002).

Seguindo esta linha de pensamento, este trabalho vem propor uma abordagem computacional alternativa àquela utilizada pelos softwares de detecção de mutações de ponto, objetivando a caracterização de novas mutações em sequências de ESTs oriundas do Projeto Genoma Humano do Câncer.

1.5 Organização do trabalho

Visando o melhor entendimento deste trabalho, esta dissertação foi organizada da seguinte maneira: O capítulo 2 enumera os objetivos do projeto. No capítulo 3 será descrito o material utilizado e métodos seguidos. No capítulo 4 serão apresentados resultados práticos da abordagem proposta. Por fim, no capítulo 5, serão apresentadas as conclusões.

2 *Objetivos*

O estudo tem como objetivo central desenvolver uma metodologia alternativa àquelas utilizadas pelos demais *softwares* de detecção de polimorfismos, podendo ser aplicada a qualquer conjunto de dados de ESTs. Para tanto, as seguintes atividades serão desenvolvidas:

- Estabelecer uma abordagem computacional para identificação de SNVs nas ESTs geradas pelo Projeto Genoma Humano do Câncer.
- Modelar uma base de dados relacional para indexar informações funcionais de todos os SNVs.
- Validar os SNVs *in silico* contra a base de dados *dbSNP* e contra os dados gerados por Sjoblom *et al.* (2006) e Wood *et al.* (2007).
- Validar experimentalmente 10 mutações de ponto selecionadas.
- Avaliar a relação dos SNVs identificados através de redes *Protein-Protein Interaction* (PPI).

3 *Material e Métodos*

A figura 14 descreve passo a passo as abordagens usadas no desenvolvimento do projeto, que incluem a identificação dos SNVs, validação contra o dbSNP, validação laboratorial de um *set* de SNVs e finalmente a análise de PPI.

Figura 14: Fluxograma mostrando as etapas realizadas na execução deste projeto.

3.1 Material

3.1.1 Base de dados de ORESTES

A base de dados de ESTs usada para a identificação dos SNVs consiste do conjunto de 1.543.301 sequências provenientes de 29 tecidos normais e tumorais gerados no Projeto Genoma Humano do Câncer (Tabela 1) (CAMARGO *et al.*, 2001). Para a análise de SNVs foram usadas ESTs de todos os tecidos. No caso da análise de PPI, foram usados somente tecidos que tinha seu equivalente normal e que apresentavam quantidades próximas de ESTs.

Tabela 1: ESTs gerados no Projeto Genoma Humano do Câncer separadas por tecido.

Símbolo	Tecido	Total ESTs
BN	Mama Normal	97836
BT	Mama Tumor	117661
CN	Cólon Normal	29632
CT	Cólon Tumor	130503
EN	Pulmão Normal	30803
ET	Pulmão Tumor	52889
FN	Próstata Normal	34035
FT	Próstata Tumor	30471
HN	Cabeça e Pescoço Normal	27236
HT	Cabeça e Pescoço Tumor	215897
NN	Sistema Nervoso Normal	117566
NT	Sistema Nervoso Tumor	72484
SN	Estômago Normal	21319
ST	Estômago Tumor	57087
TN	Testículo Normal	43891
TT	Testículo Tumor	5840
AN	Células Amnióticas Normal	25380
DT	Síndrome Denys-Drash Tumor	18313
GN	Placenta Normal	109736
IT	Epidídimo Tumor	10026
LT	Leiomiossarcoma Tumor	17046
MT	Medula Óssea Tumor	74949
OT	Ovário Tumor	29022
PT	PNET Tumor	9837
RT	Rim Tumor	15496
UM	Útero Mesclado	30781
UT	Útero Tumor	48241
CI	SI*	52853
CS	SI*	16471
Total		1543301

SI* = sem informação do tecido.

3.1.2 Base de dados de CCDS (*Consensus Coding DNA Sequences*)

O CCDS foi utilizado como referência para o alinhamento das ESTs. As informações providas pelo CCDS são provenientes de um esforço colaborativo para formar uma base de dados curada das sequências codificadoras de humanos e camundongos, incluindo suas isoformas, es-

tavelmente representadas no NCBI, Ensembl e UCSC *Genome Browser*. A iniciativa tem por objetivo convergir progressivamente a um conjunto completo de *protein-coding* nestes genomas, mantendo um padrão de confiança e precisão biológica (PRUITT *et al.*, 2009).

No presente estudo, um conjunto de 19.820 CCDS (Outubro 2009) foram obtidos a partir do *UCSC Genome Bioinformatics Site* (HINRICHS *et al.*, 2006) utilizando a ferramenta de recuperação de dados *Table Browser* (KAROLCHIK *et al.*, 2004).

3.1.3 Base de dados de polimorfismos

Para determinar os SNVs já descritos na literatura foram utilizados três bases de dados: i) **dbSNP**, já previamente descrito no item 1.3.5, que possui em sua versão 130, 17.804.034 mutações depositadas, e também dados de ii) **Sjoblom *et al.* (2006)** e iii) **Wood *et al.* (2007)**. Estes dois últimos são formados por um conjunto de mutações 2.496 mutações em genes relacionados com câncer de cólon de útero e mama.

3.2 Metodologia

Para a modelagem do banco de dados utilizou-se a ferramenta de modelagem DBDesigner¹. O sistema de gerenciamento de banco de dados (SGBD) escolhido foi o PostgreSQL².

Foram desenvolvidos alguns scripts na linguagem Perl³ para auxiliar na execução das tarefas deste trabalho. A importância desses scripts residuiu na automação e na rapidez com as quais os resultados puderam ser obtidos.

3.2.1 Modelagem do Banco de Dados

A fim de se armazenar todas as informações geradas no projeto, uma base de dados relacional foi modelada utilizando a ferramenta DBDesigner e o SGBD escolhido foi o PostgreSQL, esta base de dados será doravante denominada **dbSNP-HCGP**.

¹www.fabforce.net/dbdesigner4

²www.postgresql.org

³www.perl.org

3.2.2 Identificações dos SNVs

A figura 15 apresenta um esquema simplificado das quatro etapas do *pipeline* usadas para a identificação dos SNVs. Todo o processo inclui o pré-processamento das ORESTES, o alinhamento pareado, a exclusão de parálogos e finalmente a identificação dos SNVs.

Figura 15: Representação esquemática do *pipeline* executado para a identificação dos SNVs. **A)** Seleção dos melhores alinhamentos, **B)** Processo de identificação e caracterização dos SNVs

Pré-Processamento

Na etapa de pré-processamento (Figura 15-A, etapa 1) foram consideradas apenas regiões com qualidade phred (EWING; GREEN, 1998; EWING *et al.*, 1998) iguais ou maiores a 15.

Esse critério foi adotado para excluir sequências de baixa qualidade que podem gerar falso positivos. Outro parâmetro adotado foi a exclusão de sequências inferiores a 100 bases, visando aumentar a especificidade do alinhamento contra bases de dados de nucleotídeos e proteínas.

Alinhamento Pareado

Seguindo o esquema da figura 15-A, as ORESTES pré-processadas foram alinhadas contra 19.820 sequências de CCDS. Este processo foi executado no programa de alinhamento local BLAT, com o parâmetro *fastMap*, que não permite a inserção de introns.

Exclusão de Parálogos

Três filtros foram aplicados para a exclusão de transcritos parálogos, mapeados no mesmo CCDS (Figura 15-A, etapas 2 e 3):

1. Exclusão de ORESTES mapeadas com menos de 100 *matches* de nucleotídeos.
2. Em seguida, dos alinhamentos aceitos, foram selecionados somente aqueles com identidade igual ou superior a 95%.
3. Casos onde o alinhamento de uma EST resultou em mais de dois alinhamentos ótimos, foi considerado aquele com maior porcentagem de identidade. Em caso de empate é checado se os dois CCDS são isoformas do mesmo gene, caso afirmativo o alinhamento é considerado válido (Figura 16-A), caso negativo a ORESTES é excluída (Figura 16-B).

Figura 16: Ilustração do alinhamento de ESTs em mais de um CCDS. **A)** Figura ilustra um alinhamento aceitável, onde os 2 CCDS pertencem à mesma região genômica. **B)** Figura ilustra uma ORESTES que deve ser excluída, pois gera dois alinhamentos a partir de CCDS de regiões distintas do genoma

Todos os alinhamentos considerados válidos foram inseridos no dbSNP-HCGP desenvolvido no presente estudo.

Identificação de SNVs e validação *in silico*

Na etapa de identificação dos SNVs (Figura 15-B), as informações de alinhamentos inseridas no dbSNP-HCGP foram usadas para identificar pontos de variações (*mismatches*), que indicasse a presença de uma mutação verdadeira. A figura 17 descreve em forma de esquema os tipos de SNVs computados para as análises seguintes.

Figura 17: Representação esquemática do alinhamento de seis ESTs em um único CCDs que resultou na identificação de 4 SNVs (631C>A, 743G>A, 1053T>A, 1053T>C).

Previamente a inserção do SNV no dbSNP-HCGP, foi verificado se o alelo já estava descrito nas bases de dados dbSNP, Sjoblom *et al.* (2006) e Wood *et al.* (2007). Caso afirmativo, o SNV foi considerado válido; caso negativo, uma etapa seguinte de checagem da qualidade da leitura foi realizada, onde foram considerados como novos SNVs aqueles com pelo menos uma leitura maior ou igual a Q20, o que representa uma acurácia durante o *base calling* mínima de 99% (Figura 15-B).

3.2.3 Validação experimental

A fim de avaliar a eficiência da abordagem empregada no presente estudo, foram usadas 32 amostras (Tabela 2), obtidas no Serviço de Cirurgia de Cabeça e Pescoço do Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo, com coordenação e supervisão do Prof. Dr. Rui Celso Martins Mamede e do Dr. David L. Alves Figueiredo, obtidas após a explicação e assinatura do Termo de Consentimento Livre e Esclarecido. Todas estão armazenadas no Banco de Amostras, processo HCRP nº 9371/2003. Os métodos utilizados na validação experimental incluem o *High Resolution Melt* (HRM) para triagem das amostras, seguido do sequenciamento de DNA. Ambos serão discutidos em detalhes

a seguir.

Tabela 2: Relação das amostras utilizadas na validação experimental.

IdPaciente	Sexo	Raça	Idade ao diagnóstico (Ano)	Localização
CP5/0185	Masculino	branco	46 (2005)	laringe (supraglote)
CP5/0186	Masculino	branco	75 (2005)	laringe
CP5/0201	Masculino	branco	58 (2006)	laringe (supraglote)
CP5/0205	Masculino	mulato	51 (2006)	laringe (supraglote)
CP5/0209	Masculino	branco	51 (2006)	laringe (glote e supraglote)
CP5/0213	Masculino	branco	64 (2004)	laringe (glote)
CP5/0224	Masculino	branco	57 (2006)	laringe (glote)
CP5/0226	Masculino	branco	67 (2006)	laringe (glote/subglótica)
CP5/0243	Masculino	branco	78 (2006)	laringe (glote)
CP5/0247	Feminino	branco	57 (2006)	orofaringe
CP5/0252	Masculino	branco	73 (2006)	laringe (supraglote)
CP5/0287	Masculino	mulato	83 (2007)	laringe (glote)
CP5/0297	Masculino	branco	59 (2007)	lábio, cavidade oral e faringe
CP5/0327	Masculino	mulato	62 (2007)	laringe (glote)
CP5/0331	Masculino	branco	68 (2007)	laringe (glote)
CP5/0332	Masculino	branco	58 (2007)	laringe (glote)
CP5/0333	Masculino	branco	71 (2007)	laringe
CP5/0335	Masculino	branco	49 (2008)	laringe
CP5/0336	Masculino	branco	54 (2008)	laringe (glote)
CP5/0344	Masculino	branco	76 (2008)	laringe (supraglote)
CP5/0358	Masculino	branco	50 (2008)	laringe
CP5/0361	Masculino	branco	52 (2008)	laringe
CP5/0365	Masculino	branco	62 (2008)	laringe (glote)
CP5/0368	Masculino	branco	40 (2008)	laringe (glote)
CP5/0371	Masculino	branco	53 (2008)	laringe
CP5/0380	Masculino	branco	74 (2008)	laringe
CP5/0383	Masculino	branco	49 (2008)	laringe (glote)
CP5/0388	Masculino	branco	57 (2007)	laringe
CP5/0393	Masculino	branco	75 (2009)	laringe
CP5/0394	Masculino	branco	53 (2009)	laringe (supraglote)
CP5/0400	Masculino	negro	70 (2009)	orofaringe (epiglote)
CP5/0407	Masculino	branco	64 (2009)	laringe (supraglote)

Foram selecionados 10 SNVs (5 SNVs exclusivos de tecidos tumorais e 5 presentes em tecidos normais e tumorais) para validação experimental, esta seleção ficou restrita a mutações identificadas em mais de 2 bibliotecas distintas, sendo pelo menos uma biblioteca de Cabeça e Pescoço.

3.2.4 Padronização da PCR

As regiões de interesse foram amplificadas pela reação em cadeia da polimerase (PCR). Os primers foram definidos usando o software Primer3⁴, com parâmetros padrões e tamanho médio do amplicom de 250 bp, sendo em média 125bp a jusante e 125bp a montante da posição do SNV.

As condições iniciais de reação e ciclagem da PCR estão descritas nas tabelas 3 e 4. Para cada uma das regiões amplificadas, foi usada a temperatura de anelamento de acordo com a

⁴frodo.wi.mit.edu

sequência dos pares de *primers*.

Tabela 3: Protocolo utilizado na amplificação das regiões de interesse

Reagentes	Volume (μL)
Tampão	2,5
dNTP	2,0
primer Forward	1,5
primer Reverse	1,5
Taq DNA polimerase	0,2
DNA	1,0
H ₂ O MilliQ	16,3
Volume final	25

Tabela 4: Programa de termociclagem utilizado na reação de PCR.

Temperatura ($^{\circ}\text{C}$)	Tempo (s)	Repetições
94	300	1
94	35	39
TA*	40	
72	40	
72	300	1
16	∞	

* TA é a temperatura de anelamento de cada par de *primers*

3.2.5 High Resolution Melt - HRM

O método de HRM é comumente usado para triagem de mutações pontuais ou pequenas deleções. É um método eficiente sendo informativo para mais de 95% das mutações investigadas em um conjunto de pacientes, reduzindo assim, o número de amostras a serem sequenciadas. O método tem como base detectar variações da curva de melting de uma região mutada em relação ao controle normal. Através do emprego desta técnica é possível definir padrões de curva de melting e assim agrupar as amostras em diferentes alelos (Figura 18).

Figura 18: Exemplo de Curva de Melt Normalizada (azul=selvagem, verde=mutante, vermelho=heterozigoto). Figura obtida em www.rocke-applied-science.com

Como ilustrado por Reed, Kent e Wittwer (2007), esta técnica apresenta vantagens sobre as demais técnicas de genotipagem, tais como sequenciamento e *TaqMan® SNP Genotyping Assays*, pois apresenta:

- Menor custo comparativo relativamente as demais técnicas
- Simples e com alta acurácia, capaz de genotipar muitas amostras rapidamente

3.2.6 Sequenciamento de DNA

O sequenciamento foi realizado no sequenciador automático 3500xL Genetic Analyzer da marca life technologies™. Apenas uma amostra de cada padrão de curva de *melting* foi sequenciada para confirmar a mutação a ser validada.

3.2.7 Análise do Sequenciamento

A análise das sequências das amostras selecionadas após a análise de HRM foi realizada usando os *softwares* Phred (EWING; GREEN, 1998; EWING *et al.*, 1998), Phrap (GORDON; DESMARAI; GREEN, 2001) e Consed (GORDON; ABAJIAN; GREEN, 1998). O programa Phred é responsável por analisar os cromatogramas e atribuir um valor de qualidade a cada nucleotídeo da sequência de DNA (“*base-calling*”). O programa Phrap é responsável pelo agrupamento das sequências com base na identidade entre elas, permitindo assim a montagem de uma região consenso. O programa Consed é utilizado para visualizar a saída gerada pelo Phrap.

3.2.8 Análise de interação Proteína-Proteína

A fim de se avaliar a interação entre as proteínas mutadas identificadas no presente estudo foi usado o software Genelist⁵. Com ele é possível verificar diferentes relações entre uma lista de genes tais como: Vias metabólicas do KEGG (KANEHISA *et al.*, 2004), Ontologia Gênia do GO (HARRIS *et al.*, 2008), Gráficos de interações Proteína-Proteína, entre outros.

O objetivo dessa abordagem é de verificar se as interações entre os genes mutados formam um *network* de processos biológicos específica dos tecidos tumorais. Para esta análise foram usados apenas os tecidos de Mama, Pulmão, Próstata e Sistemas Nervoso Central, por possuírem o par normal e tumor e o número de ESTs equivalentes. Esse critério foi adotado para evitar o viés quanto ao número de genes e mutações estar relacionada a quantidade de ESTs.

Análise de PPI de genes mutados

A análise de PPI foi aplicada para identificar interações diretas e indiretas (realizadas a partir de um terceiro gene) dos genes identificados com mutações não-sinônimas presentes em duas ou mais bibliotecas distintas e exclusivas do tecido.

Devido ao fato de alguns tecidos, especialmente os germinativos, apresentarem um grande número de genes expressos, devido a sua natureza, nosso critério de exclusividade do tecido não desconsidera o SNV se este estiver presente no tecido em questão e adicionalmente nos tecidos de Cabeça e Pescoço, Testículo, Medula Óssea, Ovário ou Útero. Caso o SNV estivesse presente além do tecido em questão, em qualquer outro tecido que não estes, o SNV foi considerado não exclusivo e o gene não foi incluído na análise de PPI.

Dois filtros foram aplicados com o intuito de se obter somente as interações informativas. Ou seja, interações entre genes com mutações não-sinônimas danosas (“patogênicas”).

Análise PPI de genes com mutações em domínios

Mutações não-sinônimas localizadas dentro das regiões codificadoras que dão origem aos domínios funcionais devem desempenhar um papel especialmente danoso na atividade da proteína, pois alterações nessas regiões podem afetar a função desempenhada pela proteína.

Com o objetivo de se determinar as regiões do CCDS que dão origem aos domínios, foi utilizada a versão *stand-alone* do software HMMER 3.0⁶, e a base de dados Pfam-A (versão 26.0)

⁵www.bioinformatics-brazil.org/genelist

⁶www.hmmerr.org

(PUNTA *et al.*, 2012). Nesse processo, usou-se o módulo **hmmsearch** do HMMER 3.0 que procura por um ou mais perfis de domínios nas sequências de aminoácidos, no caso, o CCDS traduzido. Os seguintes parâmetros foram usados na análise: *threshold* de sequência (0.01), *threshold* de *hit* (0.03), *bit-score* de sequência (25) e *bit-score* de *hit* (10). Esses parâmetros são os mesmos da versão online do software⁷. Todas as informações de domínios obtidas nessa análise foram armazenadas no dbSNP-HCGP.

Análise de PPI de genes com mutações “patogênicas” em domínios

O programa SIFT (NG; HENIKOFF, 2001) foi usado para restringir ainda mais o conjunto de mutações “patogênicas”. O SIFT baseia-se na premissa de que durante a evolução das proteínas, regiões funcionais são conservadas. Ou seja, regiões importantes para a função de uma proteína devem ser conservadas dentro de uma família de proteínas, enquanto que as posições com menor importância podem variar.

Nesta análise foi usado o módulo de predição de mutações não-sinônimas da versão *online* do *software* SIFT⁸. Igualmente as análises anteriores, o resultado desta predição foi armazenado no dbSNP-HCGP.

⁷hmmer.janelia.org/search/phmmer

⁸siftdna.org

4 *Resultados e Discussão*

O objetivo principal do estudo é propor uma metodologia alternativa àquela utilizada pelos demais softwares de detecção de polimorfismos. Esta metodologia inclui etapas de pré e pós processamento e a identificação de variações propriamente dita (Seção 3.2) e pode ser aplicada a qualquer conjunto de dados de ESTs. No que segue, serão apresentados e discutidos os resultados obtidos em cada uma das etapas.

4.1 Banco de dados dbSNP-HCGP

A figura 19 representa o diagrama Entidade-Relacional da base de dados **dbSNP-HCGP**, que armazena todas as informações geradas pelo projeto.

4.2 Identificação dos SNVs

Pré-Processamento

Durante a etapa de pré-processamento, que considera o tamanho e a qualidade de cada EST (Seção 3.2.2), das 1.543.301 ESTs geradas pelo método ORESTES (Tabela 1), foram excluídas 609.057 ESTs (39,46%), resultando num conjunto de 934.244 ESTs selecionadas para a análise de SNVs.

Alinhamento Pareado

A etapa de alinhamento pareado (Seção 3.2.2) resultou em 329.092 alinhamentos e, destes, 184.330 foram considerados e inseridos no dbSNP-HCGP.

Identificação dos SNVs e validação *in-silico*

Na etapa de identificação de variações (Seção 3.2.2), um total de 69.112 SNVs foram identificados. Esse total foi dividido em dois grupos, aqueles anotados no dbSNP ou nos estudos

Figura 19: Diagrama Entidade-Relacional do banco dbSNP-HCGP

de Sjoblom *et al.* (2006) e Wood *et al.* (2007), e aqueles ainda não anotados em nenhuma destas fontes. Foram depositados no banco os 3.271 (4,73%) já descritos, bem como os 41.427 (59,94%) SNVs que passaram no filtro de qualidade das bases, sendo este o conjunto considerado para as demais análises.

A fim de se delimitar um conjunto de dados de alta confiabilidade, algumas particularidades a respeito dos dados utilizados foram consideradas, como o PCR realizado antes do sequenciamento. Esta etapa tem como objetivo amplificar uma ou poucas cópias de um pedaço de DNA em centenas de milhares, para isso utilizando a enzima conhecida como *Taq DNA polimerase*, capaz de construir uma cadeia de DNA com mil bases em cerca de 30 segundos a 72 °C, porém com uma fidelidade relativamente baixa, ou seja, ao fabricar trechos de DNA durante o processo de PCR pode incorporar nucleotídeos errados, com uma taxa de erro estimada em 1 a cada 9000 nucleotídeos polimerizados (TINDALL; KUNKEL, 1988).

Para que esses artefatos fossem minimizados, adotou-se o critério de restringência onde foi considerada o número de bibliotecas distintas que reportaram o mesmo SNV, parâmetro escolhido pelo fato de bibliotecas distintas serem obtidas a partir de preparações distintas, onde dificilmente o mesmo erro de amplificação acontecerá duas ou mais vezes. Essa informação *a priori* reduz significativamente a taxa de falso-positivos e delimita um conjunto de SNVs altamente confiáveis, entretanto não é considerada pelos demais programas de identificação de SNVs.

Analisando a porcentagem de SNVs validados *in-silico* pelo número de bibliotecas distintas, observamos que o critério se mostra realmente válido uma vez que a taxa de validação aumenta assintoticamente com o aumento de número de bibliotecas que reportaram tal SNV (Figura 20).

Figura 20: Validação *in-silico* de acordo número de bibliotecas distintas. Representação da porcentagem de SNVs validados *in-silico* de acordo com número de bibliotecas distintas em que o mesmo SNV foi encontrado

4.2.1 Validação experimental

Foram selecionados 10 SNVs (cada um em genes distintos) para validação experimental. A seleção dos genes seguiu os critérios discutidos no item 3.2.3. Nos SNVs tumor específicos seguimos o critério de mutações de pontos que encontravam-se inseridos em genes já relacionados ao processo tumorigênico (Tabela 5).

Tabela 5: Relação dos genes selecionados para validação.

Tipo de tecido	Símbolo do Gene	Alelos	Fita	Posição genômica (hg18)	Mutação	dbSNP
Normal/Tumor	ANP32C	C>A	-	chr4:165337995	Q107K	-
	HSPA8	A>G	-	chr11:122435087	T405A	-
	HTATIP2	T>G	+	chr11:20361189	S197R	rs3824886
	KRT7	G>C	+	chr12:50925569	G364A	rs2608009
	MYL1	C>T	-	chr2:210867219	T158I	-
Tumor	ALDOA	C>T	+	chr16:29988735	P295S	-
	CTSB	A>T	-	chr8:11740681	M274L	-
	EEF1A1	A>C	-	chr6:74285151	K395Q	-
	EIF4G1	A>G	+	chr3:185522360	M432V	rs2178403
	EPCAM	T>C	+	chr2:47454610	M115T	rs1126497

Amplificação por PCR

Como descrito no item 3.2.4, a partir dos 10 SNVs selecionados (Tabela 5) foram desenhados *primers* visando um amplicon médio de 250 bp. A tabela 6 descreve as relação de *primers* desenhados e as respectivas condições de reação da PCR.

Tabela 6: Lista de *primers* desenhados.

Gene	Primer Forward	Primer Reverse	Tamanho do amplicon (bp)	Temperatura de anelamento (°C)
ANP32C	CACCTCAATCTCAGACTTAC	CTGTCGAGATATGTGAGTTG	267	58
HSPA8	GATGTCTGAGCGAGTGATAG	CAGGCTGGTTGTCAGAATAG	275	60
HTATIP2	AGTTCTGTTATGTGATAGGC	TGGTGATGGGTGTTAAGG	274	59
KRT7	AAGGCAGACTGGTGAGC	TTTCCAGGTTCCCTTGTC	290	60
MYL1	AGCCATTTCCAACAACAAGG	CCCTTTGTAGTCTTTCCTCC	229	62
ALDOA	GATCACCTTCCTGTCTGGAG	CACCTCCTGCCTATCCTTAC	219	61
CTSB	GAAGCTGTTTCTCCTTCCCG	AGCCACTCACCATTGTCACC	244	64
EEF1A1	GTGATTATCCTGAACCATCC	GTAGTCATCCTTACCCAAAG	249	58
EIF4G1	CCATCTGAAGATCTGGAACC	CTTCTTCTTCCATTTCTCTCC	286	60
EPCAM	CAGTGGCTGCCAAATGTTTG	CACCTACTAGGTTCTCACTC	251	62

Os *primers* foram diluídos a concentração de 2.5 M, em seguida, testada a amplificação dos segmentos alvo. O resultado da PCR de cada gene selecionado para validação está exposto na figura 21.

Figura 21: Teste da padronização da PCR. Gel de agarose mostrando a correta amplificação dos 10 segmentos alvos utilizados para validar os SNVs

Triagem por HRM e Sequenciamento

Amostras de DNA de sangue periférico das 32 amostras de câncer de cabeça e pescoço (Tabela 2) foram analisadas para cada um dos 10 SNVs (Tabela 5) utilizando o método de HRM para triagem das amostras a partir dos diferentes perfis de curvas de melting.

Além da triagem feita a partir de amostras de sangue periférico, foram utilizadas amostras de DNA do tumor. O resultado revelou que 100% dos SNVs validados são do tipo germinativo, pois estão presentes no tumor e no sangue periférico (Tabela 7).

Tabela 7: Tabela resumo dos resultados da validação experimental dos SNVs

Tipo de tecido	Símbolo do Gene	Alelos	Mutação	dbSNP	Validação p/ HRM (Sangue periférico)*	Validação p/ HRM (DNA de tumor)*	ESP
Normal/Tumor	ANP32C	C>A	Q107K	-	-	-	-
	HSPA8	A>G	T405A	-	-	-	-
	HTATIP2	T>G	S197R	rs3824886	32/32	32/32	sim
	KRT7	G>C	G364A	rs2608009	27/32	27/32	sim
	MYL1	C>T	T158I	-	-	-	-
Tumor	ALDOA	C>T	P295S	-	-	-	-
	CTSB	A>T	M274L	-	-	-	sim
	EEF1A1	A>C	K395Q	-	1/32	1/32	-
	EIF4G1	A>G	M432V	rs2178403	32/32	32/32	sim
	EPCAM	T>C	M115T	rs1126497	22/32	22/32	sim

*número de amostras positivas para cada mutação.

Nas tabelas 8 e 9 presentes no Apêndice-A deste trabalho é possível observar o genótipo observado para cada uma das 32 amostras analisadas no projeto.

Adicionalmente, os 10 SNVs foram analisados contra o banco do projeto *NHLBI GO Exome Sequencing Project* (ESP), este projeto tem o objetivo de descobrir novos genes associados com

doenças pulmonares, cardíacas e sanguíneas. Para isso utiliza o sequenciamento do exoma de diferentes populações (Euro e Afro Americanos)¹.

O resultado da análise validou a mutação M274L, que não estava presente no dbSNP e que não tinha sido validada experimentalmente (Tabela 7). Por outro lado, a mutação K395Q, que foi validada experimentalmente, não foi detectada no ESP.

Podemos dizer que a abordagem proposta no presente estudo foi eficiente por ter validado 60% dos SNVs selecionadas (6/10). Considerando apenas os SNVs específicos do tecido tumoral, a percentagem de validação subiu para 80% (4/5). Portanto a abordagem provou ser útil para identificar mutações em dados do Projeto Genoma do Humano do Câncer.

As ilustrações 25 a 33 presentes do apêndice A deste trabalho mostram de forma mais detalhada os resultados de HRM e sequenciamento obtidos para cada um dos 10 SNVs.

4.2.2 Análise de interações Proteína-Proteína

Seguindo a metodologia descrita no item 3.2.8, foi avaliada a interação do conjunto de genes com mutações identificadas nas ESTs nos tecidos de mama, pulmão, próstata e sistemas nervoso central.

A análise de PPI foi realizada inicialmente com o conjunto de genes mutados selecionados com base no critérios de inclusão e filtros a fim de observar os padrões de interações entre o tecido normal e tumoral. Em seguida, foi aplicado, no mesmo conjunto de genes, um filtro para mutações localizadas em regiões de domínios proteicos, e finalmente, um outro filtro para mutações preditas como patogênicas segundo o *software* SIFT.

Com a aplicação desses filtros é possível observar uma redução progressiva no número de interações e consequentemente dos chamados *hub-genes* (genes centrais de interação). Interessantemente, mesmo sem a redução discrepante no número de genes mutados, observou-se redes tumorais mais conectadas em tumores do que em seus respectivos tecidos normais. As figuras 22 e 23 ilustram, como exemplificação, respectivamente os PPIs obtidos nos tecidos de mama e sistema nervoso central com a sucessiva aplicação dos filtros.

Nosso resultado reforça a ideia de que os processos biológicos que atuam durante a transformação tumorigênica podem ser conduzidos não pela frequência das mutações mas pela combinação de mutações envolvendo proteínas que interagem entre si. Este resultado é similar ao de outro estudo realizado pelo nosso grupo onde foi feito a comparação entre uma linhagem tumoral de

¹ evs.gs.washington.edu/EVS

Figura 22: Análise PPI do tecido de Mama, seguido da aplicação dos filtros de SNVs em domínios e SNVs danosos em domínios

Figura 23: Análise PPI de Sistema Nervoso Central, seguido da aplicação dos filtros de SNVs em domínios e SNVs danosos em domínios

mama (HCC1954) e outra normal de tecido linfoblastóide (HCC1954BL), derivadas do mesmo paciente (GALANTE *et al.*, 2011).

O apêndice B deste trabalho mostra de forma detalhada a lista de genes mutados e os respectivos PPIs de cada um dos tecidos analisados (Figura 34 a 57).

5 *Conclusão*

No presente estudo foi proposta uma abordagem computacional alternativa àquelas utilizadas nos demais *softwares* de detecção de polimorfismos disponíveis como: Polyphred, PolyScan, etc. A abordagem foi empregada no conjunto de dados de ESTs geradas pelo Projeto Genoma Humano do Câncer e o resultado do estudo permite-nos concluir que:

- O emprego de informações *a priori*, tal como o número de bibliotecas distintas, utilizado neste estudo, mostrou-se importante, pois possibilitou a discriminação entre observações reais e possíveis resultados espúrios. Juntamente com os demais parâmetros utilizados (tamanho mínimo da sequência de qualidade, exclusão de parálogos, qualidade mínima da leitura) mostraram-se bastante informativos para este tipo de metodologia.
- A abordagem apresentou uma alta taxa de validação, confirmando seis dos dez SNVs selecionados.
- A aplicação do filtro de mutações patogênicas localizados em regiões que dão origem aos domínios proteicos gerou redes de interações gênicas mais conectadas em tecidos tumorais do que nos respectivos normais, provando ser um critério valioso e indicando que os processos biológicos que atuam durante a transformação tumorigênica de células podem ser conduzidos não pela frequência das mutações mas pela combinação de mutações envolvendo proteínas próximas, que interagem entre si.

Referências Bibliográficas

- AHLQUIST, P. RNA-dependent RNA polymerases, viruses, and RNA silencing. *Science*, v. 296, p. 1270–1273, May 2002.
- AMARAL, P. P. *et al.* The eukaryotic genome as an RNA machine. *Science*, v. 319, p. 1787–1789, Mar 2008.
- ANFINSEN, C. B. The formation and stabilization of protein structure. *Biochem. J.*, v. 128, p. 737–749, Jul 1972.
- AVERY, O. T.; MACLEOD, C. M.; MCCARTY, M. Studies on the chemical nature of the substance inducing transformation of pneumococcal types: Induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. *J. Exp. Med.*, v. 79, p. 137–158, Feb 1944.
- BALTIMORE, D. RNA-dependent DNA polymerase in virions of RNA tumour viruses. *Nature*, v. 226, p. 1209–1211, Jun 1970.
- BELYAVSKY, A.; VINOGRADOVA, T.; RAJEWSKY, K. PCR-based cDNA library construction: general cDNA libraries at the level of a few cells. *Nucleic Acids Res.*, v. 17, p. 2919–2932, Apr 1989.
- BERK, V.; CATE, J. H. Insights into protein biosynthesis from structures of bacterial ribosomes. *Curr. Opin. Struct. Biol.*, v. 17, p. 302–309, Jun 2007.
- BERTRAM, J. S. The molecular biology of cancer. *Mol. Aspects Med.*, v. 21, p. 167–223, Dec 2000.
- BLACK, D. L. Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.*, v. 72, p. 291–336, 2003.
- BRANFORD, S.; MELO, J. V.; HUGHES, T. P. Selecting optimal second-line tyrosine kinase inhibitor therapy for chronic myeloid leukemia patients after imatinib failure: does the BCR-ABL mutation status really matter? *Blood*, v. 114, n. 27, p. 5426–5435, Dec 2009.
- BRANFORD, S. *et al.* High frequency of point mutations clustered within the adenosine triphosphate-binding region of BCR/ABL in patients with chronic myeloid leukemia or Ph-positive acute lymphoblastic leukemia who develop imatinib (STI571) resistance. *Blood*, v. 99, p. 3472–3475, May 2002.
- BRANFORD, S. *et al.* Detection of BCR-ABL mutations in patients with CML treated with imatinib is virtually always accompanied by clinical resistance, and mutations in the ATP phosphate-binding loop (P-loop) are associated with a poor prognosis. *Blood*, v. 102, p. 276–283, Jul 2003.

- BUNZ, F. *Principles of Cancer Genetics*. 1. ed. [S.l.]: Springer Science, 2008. ISBN 978-1-4020-6784-6.
- CAMARGO, A. A. *et al.* The contribution of 700,000 ORF sequence tags to the definition of the human transcriptome. *Proc. Natl. Acad. Sci. U.S.A.*, v. 98, p. 12103–12108, Oct 2001.
- CARIASO, M.; LENNON, G. SNPedia: a wiki supporting personal genome annotation, interpretation and analysis. *Nucleic Acids Res.*, v. 40, p. D1308–1312, Jan 2012.
- CHARGAFF, E. Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. *Experientia*, v. 6, p. 201–209, Jun 1950.
- COOPER, D. N.; BALL, E. V.; KRAWCZAK, M. The human gene mutation database. *Nucleic Acids Res.*, v. 26, p. 285–287, Jan 1998.
- CRICK, F. Central dogma of molecular biology. *Nature*, v. 227, p. 561–563, Aug 1970.
- DIAS-NETO, E. *et al.* Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. *Proc. Natl. Acad. Sci. U.S.A.*, v. 97, p. 3491–3496, Mar 2000.
- EWING, B.; GREEN, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.*, v. 8, p. 186–194, Mar 1998.
- EWING, B. *et al.* Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.*, v. 8, p. 175–185, Mar 1998.
- FADERL, S. *et al.* Chronic myelogenous leukemia: biology and therapy. *Ann. Intern. Med.*, v. 131, p. 207–219, Aug 1999.
- GALANTE, P. A. *et al.* Distinct patterns of somatic alterations in a lymphoblastoid and a tumor genome derived from the same individual. *Nucleic Acids Res.*, v. 39, n. 14, p. 6056–6068, Aug 2011.
- GORDON, D.; ABAJIAN, C.; GREEN, P. Consed: a graphical tool for sequence finishing. *Genome Res.*, v. 8, n. 3, p. 195–202, mar. 1998. ISSN 1088-9051.
- GORDON, D.; DESMARAIS, C.; GREEN, P. Automated finishing with autofinish. *Genome Res.*, v. 11, n. 4, p. 614–25, abr. 2001. ISSN 1088-9051.
- GORRE, M. E. *et al.* Clinical resistance to STI-571 cancer therapy caused by BCR-ABL gene mutation or amplification. *Science*, v. 293, p. 876–880, Aug 2001.
- GRIFFITH, F. The Significance of Pneumococcal Types. *J Hyg (Lond)*, v. 27, p. 113–159, Jan 1928.
- HAMOSH, A. *et al.* Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, v. 30, p. 52–55, Jan 2002.
- HANSEN, T. M. *et al.* Maintenance of the correct open reading frame by the ribosome. *EMBO Rep.*, v. 4, p. 499–504, May 2003.
- HARRIS, M. *et al.* The Gene Ontology project in 2008. *Nucleic Acids Res.*, v. 36, n. Database issue, p. D440–444, Jan 2008.

- HERSHEY, A. D.; CHASE, M. Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J. Gen. Physiol.*, v. 36, p. 39–56, May 1952.
- HINRICHS, A. S. *et al.* The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.*, v. 34, p. D590–598, Jan 2006.
- HOCHHAUS, A. *et al.* Molecular and chromosomal mechanisms of resistance to imatinib (STI571) therapy. *Leukemia*, v. 16, p. 2190–2196, Nov 2002.
- KANEHISA, M. *et al.* The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, v. 32, n. Database issue, p. D277–280, Jan 2004.
- KAROLCHIK, D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.*, v. 32, p. D493–496, Jan 2004.
- KNOERS, N. V.; MONNENS, L. A. Teaching molecular genetics: Chapter 1–Background principles and methods of molecular biology. *Pediatr. Nephrol.*, v. 21, p. 169–176, Feb 2006.
- MANOLIO, T. A. Genomewide association studies and assessment of the risk of disease. *N. Engl. J. Med.*, v. 363, p. 166–176, Jul 2010.
- MARTH, G. T. *et al.* A general approach to single-nucleotide polymorphism discovery. *Nat. Genet.*, v. 23, p. 452–456, Dec 1999.
- MCCARTHY, B. J.; HOLLAND, J. J. Denatured DNA as a direct template for in vitro protein synthesis. *Proc. Natl. Acad. Sci. U.S.A.*, v. 54, p. 880–886, Sep 1965.
- NCBI. *Genes and Disease [Internet]*. [S.l.]: Bethesda (MD): National Center for Biotechnology Information (US), 1998. Anemia, sickle cell.
- NG, P. C.; HENIKOFF, S. Predicting deleterious amino acid substitutions. *Genome Res.*, v. 11, n. 5, p. 863–874, May 2001.
- NICKERSON, D. A.; TOBE, V. O.; TAYLOR, S. L. PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Res.*, v. 25, p. 2745–2751, Jul 1997.
- OKUBO, K. *et al.* Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. *Nat. Genet.*, v. 2, p. 173–179, Nov 1992.
- PLATT, O. S. *et al.* Mortality in sickle cell disease. Life expectancy and risk factors for early death. *N. Engl. J. Med.*, v. 330, p. 1639–1644, Jun 1994.
- PRUITT, K. D. *et al.* The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res.*, v. 19, p. 1316–1323, Jul 2009.
- PUNTA, M. *et al.* The Pfam protein families database. *Nucleic Acids Res.*, v. 40, n. Database issue, p. 290–301, Jan 2012.
- REED, G. H.; KENT, J. O.; WITTWER, C. T. High-resolution DNA melting analysis for simple and efficient molecular diagnostics. *Pharmacogenomics*, v. 8, n. 6, p. 597–608, Jun 2007.

- SCHIFREEN, R. S.; STORTS, D. R.; BULLER, A. M. The challenge of using SNPs in the understanding and treatment of disease. *BioTechniques*, Suppl, p. 14–16, Jun 2002.
- SCHWANHAUSSER, B. *et al.* Global quantification of mammalian gene expression control. *Nature*, v. 473, p. 337–342, May 2011.
- SHAH, N. P. *et al.* Multiple BCR-ABL kinase domain mutations confer polyclonal resistance to the tyrosine kinase inhibitor imatinib (STI571) in chronic phase and blast crisis chronic myeloid leukemia. *Cancer Cell*, v. 2, p. 117–125, Aug 2002.
- SHEPPARD, P. M. *Natural Selection and Heredity*. 4. ed. [S.l.]: London:Hutchinson, 1975. ISBN 9780091226114.
- SHERRY, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, v. 29, p. 308–311, Jan 2001.
- SJOBLOM, T. *et al.* The consensus coding sequences of human breast and colorectal cancers. *Science*, v. 314, p. 268–274, Oct 2006.
- STRACHAN, T.; READ, A. P. *Genética Molecular Humana*. 2. ed. [S.l.]: Artmed, 2001. ISBN 1-85996-202-5.
- TEMIN, H. M.; MIZUTANI, S. RNA-dependent DNA polymerase in virions of Rous sarcoma virus. *Nature*, v. 226, p. 1211–1213, Jun 1970.
- TINDALL, K. R.; KUNKEL, T. A. Fidelity of DNA synthesis by the *Thermus aquaticus* DNA polymerase. *Biochemistry*, v. 27, n. 16, p. 6008–6013, Aug 1988.
- WATSON, J. D.; CRICK, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, v. 171, p. 737–738, Apr 1953.
- WILCOX, A. S. *et al.* Use of 3' untranslated sequences of human cDNAs for rapid chromosome assignment and conversion to STSs: implications for an expression map of the genome. *Nucleic Acids Res.*, v. 19, p. 1837–1843, Apr 1991.
- WOOD, L. D. *et al.* The genomic landscapes of human breast and colorectal cancers. *Science*, v. 318, p. 1108–1113, Nov 2007.

APÊNDICE A – Triagem por HRM e Sequenciamento

Tabela 8: Relação dos genótipos observados em cada uma das amostras analisadas na validação experimental dos genes identificados como normais/tumorais.

IdPaciente	ANP32C (C>A)	HSPA8 (A>G)	HTATIP2 (T>G)	KRT7 (G>C)	MYL1 (C>T)
CP5/0185	CC	AA	GG	GC	CC
CP5/0186	CC	AA	TG	CC	CC
CP5/0201	CC	AA	GG	GG	CC
CP5/0205	CC	AA	GG	GC	CC
CP5/0209	CC	AA	GG	GG	CC
CP5/0213	CC	AA	GG	GC	CC
CP5/0224	CC	AA	GG	CC	CC
CP5/0226	CC	AA	GG	CC	CC
CP5/0243	CC	AA	GG	CC	CC
CP5/0247	CC	AA	GG	CC	CC
CP5/0252	CC	AA	GG	CC	CC
CP5/0287	CC	AA	TG	CC	CC
CP5/0297	CC	AA	GG	CC	CC
CP5/0327	CC	AA	TG	GC	CC
CP5/0331	CC	AA	TG	GC	CC
CP5/0332	CC	AA	GG	CC	CC
CP5/0333	CC	AA	GG	CC	CC
CP5/0335	CC	AA	GG	CC	CC
CP5/0336	CC	AA	GG	CC	CC
CP5/0344	CC	AA	GG	CC	CC
CP5/0358	CC	AA	GG	GG	CC
CP5/0361	CC	AA	GG	GG	CC
CP5/0365	CC	AA	GG	GG	CC
CP5/0368	CC	AA	GG	CC	CC
CP5/0371	CC	AA	GG	CC	CC
CP5/0380	CC	AA	TG	CC	CC
CP5/0383	CC	AA	TG	CC	CC
CP5/0388	CC	AA	GG	CC	CC
CP5/0393	CC	AA	GG	CC	CC
CP5/0394	CC	AA	TG	CC	CC
CP5/0400	CC	AA	TG	CC	CC
CP5/0407	CC	AA	TG	CC	CC

Tabela 9: Relação dos genótipos observados em cada uma das amostras analisadas na validação experimental nos genes identificados como tumorais.

IdPaciente	ALDOA (C>T)	CTSB (A>T)	EEF1A1 (A>C)	EIF4G1 (A>G)	EPCAM (T>C)
CP5/0185	CC	AA	AA	GG	TC
CP5/0186	CC	AA	AA	GG	CC
CP5/0201	CC	AA	AA	GG	TT
CP5/0205	CC	AA	AA	AG	CC
CP5/0209	CC	AA	AA	AG	TT
CP5/0213	CC	AA	AA	AG	CC
CP5/0224	CC	AA	AA	GG	TT
CP5/0226	CC	AA	AA	GG	TT
CP5/0243	CC	AA	AA	AG	TC
CP5/0247	CC	AA	AA	GG	TC
CP5/0252	CC	AA	AA	GG	TC
CP5/0287	CC	AA	AA	GG	CC
CP5/0297	CC	AA	AA	AG	TT
CP5/0327	CC	AA	AC	AG	CC
CP5/0331	CC	AA	AA	GG	TC
CP5/0332	CC	AA	AA	GG	TT
CP5/0333	CC	AA	AA	AG	TC
CP5/0335	CC	AA	AA	AG	TT
CP5/0336	CC	AA	AA	GG	TC
CP5/0344	CC	AA	AA	GG	TT
CP5/0358	CC	AA	AA	GG	TC
CP5/0361	CC	AA	AA	AG	TC
CP5/0365	CC	AA	AA	GG	TT
CP5/0368	CC	AA	AA	GG	TC
CP5/0371	CC	AA	AA	GG	TC
CP5/0380	CC	AA	AA	GG	TC
CP5/0383	CC	AA	AA	GG	TC
CP5/0388	CC	AA	AA	AG	TT
CP5/0393	CC	AA	AA	GG	CC
CP5/0394	CC	AA	AA	AG	TC
CP5/0400	CC	AA	AA	GG	CC
CP5/0407	CC	AA	AA	AG	TC

ANP32C

Figura 24: Resultado da abordagem utilizada para investigação do SNV no gene ANP32C, foram identificados somente amostras homozigóticas referência (G), portanto este SNV não foi confirmado nas amostras analisadas. Entretanto foi observado um sítio 108bp a montante do ponto, onde foram encontrados alelos em homozigose referência (C), homozigose mutante (T) e heterozigose, confirmando os padrões diferentes de curvas representados no HRM

HSPA8

Figura 25: Resultado da abordagem utilizada para investigação do SNV no gene HSPA8, foram identificados somente amostras homozigóticas referência (A), portanto este SNV não foi confirmado nas amostras analisadas.

HTATIP2

Figura 26: Resultado da abordagem utilizada para investigação do SNV no gene HTATIP2, foram identificados amostras homozigóticas mutadas (G) e heterozigóticas confirmando o SNV nas amostras.

KRT7

Figura 27: Resultado da abordagem utilizada para investigação do SNV no gene KRT7, foram identificados amostras homozigóticas referência (G) e mutadas (C) e heterozigóticas confirmando o SNV nas amostras.

Figura 28: Resultado da abordagem utilizada para investigação do SNV no gene MYL1, foram identificados somente amostras homozigóticas referência (C), portanto este SNV não foi confirmado nas amostras analisadas.

ALDOA

Figura 29: Resultado da abordagem utilizada para investigação do SNV no gene ALDOA, foram identificados somente amostras homozigóticas referência (C), portanto este SNV não foi confirmado nas amostras analisadas.

CTSB

Figura 30: Resultado da abordagem utilizada para investigação do SNV no gene CTSB, foram identificados somente amostras homozigóticas referência (A), portanto este SNV não foi confirmado nas amostras analisadas.

Figura 31: Resultado da abordagem utilizada para investigação do SNV no gene EEF1A1, foram identificados amostras homozigóticas referência (T) e heterozigóticas confirmando o SNV nas amostras.

EIF4G1

Figura 32: Resultado da abordagem utilizada para investigação do SNV no gene EIF4G1, foram identificados amostras homozigóticas mutadas (G) e heterozigóticas confirmando o SNV nas amostras.

EPCAM

Figura 33: Resultado da abordagem utilizada para investigação do SNV no gene EPCAM, foram identificados amostras homozigóticas referência (T) e mutadas (C) e heterozigóticas confirmando o SNV nas amostras.

APÊNDICE B – Análise de PPI

Figura 36: Interações dos genes com SNVs no tecido de pulmão normal (EN)

Figura 37: Interações dos genes com SNVs no tecido de pulmão tumor (ET)

Figura 38: Interações dos genes com SNVs no tecido de próstata normal (FN)

Figura 39: Interações dos genes com SNVs no tecido de próstata tumor (FT)

Figura 40: Interações dos genes com SNVs no tecido de sistema nervoso normal (NN)

Figura 41: Interações dos genes com SNVs no tecido de sistema nervoso tumor (NT)

Figura 42: Interações dos genes com SNVs em regiões que dão origem aos domínios proteicos no tecido de mama normal (BN)

Figura 43: Interações dos genes com SNVs em regiões que dão origem aos domínios proteicos no tecido de mama tumor (BT)

Figura 44: Interações dos genes com SNVs em regiões que dão origem aos domínios proteicos no tecido de pulmão normal (EN)

Figura 45: Interações dos genes com SNVs em regiões que dão origem aos domínios proteicos no tecido de pulmão tumor (ET)

Figura 46: Interações dos genes com SNVs em regiões que dão origem aos domínios proteicos no tecido de próstata normal (FN)

Figura 47: Interações dos genes com SNVs em regiões que dão origem aos domínios proteicos no tecido de próstata tumor (FT)

Figura 48: Interações dos genes com SNVs em regiões que dão origem aos domínios proteicos no tecido de sistema nervoso normal (NN)

Figura 49: Interações dos genes com SNVs em regiões que dão origem aos domínios proteicos no tecido de sistema nervoso tumor (NT)

Figura 50: Interações dos genes com SNVs danosos em regiões que dão origem aos domínios proteicos no tecido de mama normal (BN)

Figura 51: Interações dos genes com SNVs danosos em regiões que dão origem aos domínios proteicos no tecido de mama tumor (BT)

Figura 52: Interações dos genes com SNVs danosos em regiões que dão origem aos domínios proteicos no tecido de pulmão normal (EN)

Figura 53: Interações dos genes com SNVs danosos em regiões que dão origem aos domínios proteicos no tecido de pulmão tumor (ET)

Figura 54: Interações dos genes com SNVs danosos em regiões que dão origem aos domínios proteicos no tecido de próstata normal (FN)

Figura 55: Interações dos genes com SNVs danosos em regiões que dão origem aos domínios proteicos no tecido de próstata tumor (FT)