SIMTar: Uma ferramenta para predição de SNPs interferindo em sítios alvos de microRNAs

Amanda Rusiska Piovezani

Dissertação Apresentada
ao
Programa Interunidades
de
Pós-Graduação
em
Bioinformática
da
Universidade de São Paulo
para
Obtenção do Título
de
Mestre em Ciências

Área de Concentração: **Bioinformática**

Orientadora: **Profª. Drª. Ariane Machado Lima**

Durante a elaboração deste trabalho, a autora recebeu auxílio financeiro do CNPq

– São Paulo, SP – Novembro de 2013 –

SIMTar: Uma ferramenta para predição de SNPs interferindo em sítios alvos de microRNAs

Este exemplar corresponde à redação final
da dissertação devidamente corrigida e
defendida por Amanda Rusiska Piovezani e
aprovada pela comissão julgadora.

São Paulo, Novembro de 2013

Comissão Julgadora:

- Profª. Drª. Ariane Machado Lima - EACH-USP (orientadora)
- Prof. Dr. Eduardo Moraes Rego Reis - IQ-USP
- Prof. Dr. Luciano Antonio Digiampietri - EACH-USP

Dedico esta dissertação a você mãe,
como agradecimento por sua dedicação
e amor incondicional.

Se você é...

Se você é um Vencedor,
terá alguns falsos amigos
e alguns amigos verdadeiros.
Vença assim mesmo.

Se você é Honesto e Franco,
as pessoas podem enganá-lo.
Seja honesto e franco assim mesmo.

O que você levou anos para Construir
Alguém pode destruir de uma hora para outra.
Construa assim mesmo.

Se você tem Paz e é Feliz,
As pessoas podem sentir inveja.
Seja feliz assim mesmo.

Dê ao mundo o Melhor de Você,
mas isso pode nunca ser o bastante.
Dê o melhor de você assim mesmo.

Veja você que, no final de tudo
Será Você ... e Deus.
E não você ... e as pessoas!

Madre Tereza de Calcutá

Agradecimentos

GOSTARIA de agradecer primeiramente a Deus e Nossa Senhora, pela oportunidade de eu estar aqui, parte de uma família tão maravilhosa e por se fazerem presentes em todos os instantes da minha vida.

Mãe, obrigada por estar sempre ao meu lado me apoiando em todas as decisões. Em você me inspiro para buscar sempre mais e ser sempre melhor a cada dia. Aqui está mais um 'café pequeno', mas não se esqueça, eu ainda preciso dos 'suquinhos da vitória';-) Te Amo Mãe!

Beto, irmão de ouro e muito fofinho sempre. Obrigada pela força e paciência. Você mais que ninguém presenciou a maior parte do meu mestrado, minha ansiedade e nervoso. Desculpe qualquer coisa, eu Te Amo!

Gê, você é parte importante da minha história. Tomamos várias decisões juntas e sou a pessoa mais feliz do mundo. Obrigada pela paciência quando eu não podia fazer qualquer coisa por conta do mestrado. 'Essa'fase acabou, rsrs. Obrigada por ser sempre muito companheira em Tudo. Amo Você!

Dra. Ariane Machado Lima, a minha orientadora. Obrigada pela amizade e pela excelente experiência durantes esses anos e por conduzir os meus aprendizados. Você é uma pessoa fantástica. Foi um imenso prazer trabalhar com você, Orientadora!

Dra. Helena Brentani, extra-oficialmente co-orientadora deste trabalho, desde a sua concepção até o presente momento. Você é peça chave da minha formação, uma pessoa incrível que tive o prazer de conhecer e trabalhar junto. Saiba que tenho muita admiração por você, por sua pessoa e seu profissi-

onalismo. Obrigada por todos os minutos que estivemos juntas, pelos ensinamentos de todos os níveis.

Embora muitas pessoas maravilhosas tenham passado em minha vida, me ensinando muito e de várias formas, algumas sempre estiveram em todos os momentos, a minha família. Neste ponto é muito difícil escolher por onde começar os agradecimentos. Em especial as tias e tios Rusiskas foram positivamente responsáveis por muitas coisas chaves da minha história. À minha família Piovezani obrigada pela confiança.

Obrigada Tia Helena e Tio Marcos por Sempre estarem por perto, desde o tempo da natação, sempre me apoiando em tudo, literalmente em Tudo. Sou eternamente grata.

Tia Salete, tio Val, Michelle e Marcelo, obrigada por sempre serem tão presentes em tudo, por abrirem as portas de sua casa quando eu precisei. Pelos cafezinhos da tia Salete em dia de prova. Nossa, obrigada por tudo!

Tia Marta e tio Jair, tia Andréa e tio Ademir obrigada pelo incentivo que também sempre deram desde o início.

Agradeço a Dna Elinha Barros pelo carinho e pelas orações, que sempre chegaram fiéis até mim!!!

Agradeço a todos os meus amigos do laboratório de visão computacional e bioinformática da USP pela amizade durante esses anos. Obrigada especialmente Dr. David Pires, Dra. Talita Persiano, Doutorandos Sérgio Nery, Leandro Lima e Jihan e a mestranda Suzzana Rezende, adoro vocês. Obrigada primeiramente pela amizade e depois pela ajuda com tudo, desde o início até agora na reta final, vocês foram sempre muito companheiros, eu diria pessoas ímpares.

Ao Dr. Roberto Hirata, uma pessoa incrível que eu admiro muito, como pessoa e também como profissional. Por vezes disse palavras tão especiais de incentivo. Obrigada por cruzar o meu caminho.

Agradeço o carinho da Dna. Jovita, Lucia e Sandra, as mocinhas que cuidadosamente acordaram meus dias e minhas tardes com um cafezinho.

Agradeço à CPG da bioinformática, em especial à Patrícia pela gentileza, disponibilidade e amizade sempre.

Ao CNPq pelo auxílio financeiro na realização deste trabalho e FAPESP # 2011/50761-2, CNPq, CAPES, NAP eScience - PRP - USP pela estrutura eScience.

RESUMO

PIOVEZANI, AMANDA RUSISKA. **SIMTar: Uma ferramenta para predição de SNPs interferindo em sítios alvos de microRNAs**. 71 p. Dissertação (Mestrado em Bioinformática) - Interunidades em Bioinformática, Universidade de São Paulo, São Paulo, 2013.

Polimorfismos de um nucleotídeo (SNPs) podem alterar, além de códons de leitura da tradução de genes, posições genômicas importantes do processo de regulação gênica, como sítios de iniciação de transcrição, de *splicing* e, mais especificamente, sítios alvos de microRNAs (miRNAs). Em particular, a identificação de SNPs alterando sítios alvos de miRNAs é um problema em aberto, embora venha ganhando um importante destaque nos últimos anos decorrente dos avanços descobertos sobre a capacidade dos miRNAs como elementos reguladores do genoma, associados inclusive a muitas doenças como o câncer e vários transtornos psiquiátricos. Os recursos computacionais atualmente disponíveis para esta finalidade (alguns bancos de dados e uma ferramenta) estão restritos à análise de SNPs na região 3'UTR (*UnTranslated Regions*) de RNAs mensageiros, onde miRNAs geralmente se ligam para reprimir sua tradução. No entanto, essa é uma simplificação do problema, dado que já se conhece a regulação por miRNAs ativando ou reprimindo a transcrição gênica quando ligados à sua região promotora, aumentando a efetividade da regulação negativa da tradução quando ligados à região codificante do gene ou ainda miRNAs se ligando a RNAs não codificantes. Esses recursos se limitam também à identificação de SNPs na região *seed* dos sítios de miRNAs, e portanto só identificam criação ou ruptura de sítios. Porém, SNPs localizados fora dessa região não só podem colaborar na criação e ruptura de sítios como também interferir na estabilidade de ligação de miRNAs e, portanto, na efetividade da regulação. Além disso, considerando toda a extensão do sítio, não somente a *seed*, é possível ocorrer mais de um SNP e, sendo assim, a combinação desses SNPs pode ter uma influência ainda maior na ligação com o miRNA. Os recursos atuais também não informam quais alelos dos SNPs, muito menos quais combinações deles, estão causando qual efeito. Por fim, tais recursos estão restritos a *Homo sapiens* e *Mus musculus*. Assim, este trabalho apresenta a ferramenta computacional SIMTar (*SNPs Interfering in MicroRNA Targets*), desenvolvida para identificar SNPs que alteram sítios alvos de miRNAs e que preenche as lacunas mencionadas. Além disso, é descrita uma aplicação de SIMTar na análise de 114 SNPs associados à esquizofrenia, na qual todos foram preditos interferindo em sítios alvos de miRNAs.

Palavras-chave: microRNAs, predição de alvos de microRNAs, SNPs, SNPs em sítios alvos de microRNAs, ligação de microRNAs em DNA e RNAs não codificantes.

ABSTRACT

PIOVEZANI, AMANDA RUSISKA. **SIMTar: A tool for prediction of SNPs Interfering on MicroRNA Target sites**. 71 p. Master Thesis (Master of Science on Bioinformatics) - Inter-institutional Grad Program on Bioinformatics, University of São Paulo, São Paulo, 2013.

Single nucleotide polymorphisms (SNPs) can be involved in alteration of not only open reading frame but also important genomic positions of gene regulation process such as transcription initiation sites, splicing sites and microRNA target sites. In particular, the identification of SNPs interfering on microRNA target sites is still an open problem, despite its increasing prominence in recent years due to the discoveries about the microRNA abilities as regulatory elements in the genome and association with severals diseases such as cancer and psychiatric disorders. The computational resources currently available for this purpose (four databases and one tool) are restricted to the analysis of SNPs in the 3'UTR (UnTranslated Regions) of mRNAs, where the microRNAs typically bind in order to repress their translation. However, this is a simplification of the problem, since it is already known the gene transcription activation by microRNAs bound to its promoter region, increasing of the effectiveness of negative regulation of translation when microRNAs are bound to the coding region of the gene or binding of microRNA into non-coding RNAs. These resources are also limited to the identification of SNPs in the seed region of miRNAs, and therefore they can only identify sites creation or disruption. However, SNPs located outside this region can not only create and disrupt target sites but also interfere on the stability of miRNAs binding and therefore on the regulation effectiveness. Moreover, considering the target site length, more than one SNP can occur inside of a site and thus, the combination of these SNPs can have an even greater influence on the microRNA binding. Also, current resources do not display which alleles of SNPs or what combinations of them are causing which effect. Finally, these features are restricted to the *Homo sapiens* and *Mus musculus* species. This work presents the computational tool SIMTar (SNPs Interfering on MicroRNA Targets), developed to identify SNPs that alter miRNA target sites and fills the mentioned gaps. Finally, it is described an application of SIMTar on the analysis of 114 SNPs associated with schizophrenia, all of them being predicted interfering with miRNA target sites.

Keywords: microRNAs, microRNAs target prediction, SNPs, SNPs in microRNA target sites, microRNAs binding on DNA and Non-coding RNAs.

Sumário

Resumo **7**

Abstract **8**

Lista de Figuras **11**

Lista de Tabelas **12**

1 Introdução **1**

2 Fundamentos **4**
2.1 Fluxo da informação genética 4
2.2 Polimorfismos de um nucleotídeo (SNPs) 6
2.3 MicroRNAs 8
2.3.1 Biogênese dos microRNAs 8
2.3.2 Nomenclatura dos microRNAs 9
2.3.3 Pareamento do microRNA com o sítio alvo 10
2.3.4 Formas de regulação 14
2.4 Predição computacional de alvos de microRNAs 16
2.4.1 Características do pareamento miRNAs-alvo 16
2.4.2 Ferramentas de predição de alvos de microRNAs 18
2.5 Estudos de SNPs em sítios alvos de microRNAs 18
2.6 Doenças complexas e esquizofrenia 21

3 Objetivos **23**

4 Metodologia **24**
4.1 Características da ferramenta SIMTar 24
4.2 Estrutura do SIMTar 26
4.3 Implementação 32
4.4 Aplicação em esquizofrenia 32
4.4.1 Seleção de SNPs associados a esquizofrenia 32
4.4.2 Execução do SIMTar 33
4.4.3 Seleção de miRNAs associados à esquizofrenia para análise dos resultados do estudo de caso 33

5 Resultados e Discussão **34**
5.1 SIMTar 34
5.1.1 Arquivos de saída da ferramenta SIMTar 36
5.2 Aplicação do SIMTar em esquizofrenia 42
5.2.1 SNPs associados a esquizofrenia interferindo em sítios de miRNAs 42
5.2.2 MicroRNAs associados à esquizofrenia com sítios interferidos por SNPs . . 44

6 Conclusões e trabalhos futuros **56**
6.1 Conclusões 56
6.2 Trabalhos futuros 58

A SNPs utilizados no estudo de caso de SIMTar **59**

Referências Bibliográficas **60**

Lista de Figuras

2.1 Esquema geral da disposição estrutural de um gene. 5
2.2 Dogma central da biologia molecular. 6
2.3 Biogênese de miRNAs. 9
2.4 Nomenclatura de miRNAs. 10
2.5 Nomenclatura de sequências maduras e de famílias de miRNAs. 11
2.6 Pareamento estritamente perfeito da seed com o alvo e da região 3' do miRNA compensando o pareamento da primeira. 12
2.7 Tipos de pareamentos seed observados em animais. 13
2.8 Trecho do alinhamento múltiplo da 3'UTR do gene ANKK1 humano, contendo um sítio de miRNA. 17

4.1 Estrutura da ferramenta SIMTar. 27

5.1 Tabela gerada pela ferramenta SIMTar. 39
5.2 Contexto genômico dos SNPs utilizados na aplicação de SIMTar. 43
5.3 Relação de sítios alvos por SNP associado à esquizofrenia. 44
5.4 Rede de interação entre miRNAs e os seus respectivos alvos identificados a partir da predição genômica de SIMTar utilizando dados de esquizofrenia. 53
5.5 Destaque da interação observada a partir dos resultados de SIMTar para o miR-4761-5p e gene *DTNBP1* 55

A.1 Relação dos 23 genes e 114 SNPs utilizados no estudo de caso. 59

Lista de Tabelas

2.1 Lista de algumas ferramentas de predição de alvos de miRNAs. 19

5.1 Relação de recursos computacionais existentes para análise de SNPs em sítios alvos de miRNAs. 37
5.2 MiRNAs associados à esquizofrenia identificados por SIMTar. 45
5.3 Relação de SNPs e miRNAs associados à esquizofrenia preditos por SIMTar e suas respectivas regiões genômicas. 46
5.4 Genes com sítios de miRNAs interferidos por SNPs, para miRNAs e SNPs associados a esquizofrenia. 52

Capítulo 1

Introdução

MICRORNAS (miRNAs) são pequenas moléculas de RNAs não codificadores, com tamanho aproximado de 22 nucleotídeos (nts), que têm revelado uma importante participação no processo de regulação de vários organismos, atuando em nível pré e pós-transcricional (Büssing *et al.* [2008]; Chen *et al.* [2004]; Frankel *et al.* [2008]; Si *et al.* [2006]). Em animais este processo de regulação é realizado através de um pareamento imperfeito com o sítio alvo. Os miRNAs são mais conhecidos pelo papel que desempenham na regulação negativa da tradução, efetuada quando estão pareados à porção 3'UTR de sequências de mRNAs. No entanto, o estudo de miRNAs mostrou recentemente várias novas maneiras de regulação pelas quais os miRNAs podem regular, além da tradução, também a transcrição dos genes. Estes novos estudos identificaram que a regulação pode ser positiva ou negativa, não só quando estão ligados à região 3'UTR mas também quando se ligam inclusive no DNA e sequências de outros RNAs não codificadores (Huang *et al.* [2012]; Jackson *et al.* [2010]; Lee *et al.* [2005]; Lin *et al.* [2011]; Shen *et al.* [2002]; Toscano-Garibay & Aquino-Jarquin [2012]; Vasudevan *et al.* [2007]; Zisoulis *et al.* [2012]). Muitos deles estando associados a doenças complexas como o câncer, diabetes e transtornos psiquiátricos.

Dado que o pareamento em animais ocorre geralmente de forma imperfeita e a região do pareamento abrange por volta de 22 nts, pequenas alterações como é o caso de polimorfismos de um nucleotídeo (*Single Nucleotide Polymorphisms* - SNPs), podem criar um novo sítio, romper um sítio

pré-existente ou mesmo interferir na estabilidade do pareamento dada a sua variação alélica. Isso ocorre porque a variação alélica do SNP pode implicar no aumento ou redução da complementaridade da sequência do miRNA com a sequência alvo e consequentemente na efetividade da regulação pelo miRNA.

SNPs com menor frequência alélica (*Minor Allele Frequence* - MAF) $\geq 1\%$ correspondem a 98% da variabilidade humana [McVean *et al.*, 2012]. Embora estudos de associação (*Genome-Wide Association Study* – GWAS) [Hindorff, 2009] tenham revelado centenas destes SNPs associados a muitas doenças (Jia *et al.* [2010]; Thomas *et al.* [2009]; Zhang *et al.* [2007]), muitos ainda não possuem nenhuma associação. Aproximadamente 90% destes SNPs estão localizados em regiões não codificantes do genoma [Hindorff, 2009], tal como a região promotora dos genes [Savinkova *et al.*, 2013], onde podem alterar sítios de elementos reguladores dessa região e consequentemente alterar muitas vias metabólicas relacionadas à transcrição destes genes. Podem também ocorrer em sítios de excisão de íntrons (*splicing*) e alterar o uso de éxons, implicando na síntese de diferentes isoformas de um mesmo gene (transcritos). Outra possibilidade é que tais SNPs podem criar, romper ou interferir na estabilidade de sítios de ligação de miRNAs [Lu & Clark, 2012].

O estudo computacional de identificação de SNPs que estejam interferindo em sítios alvos de miRNAs é uma área recente de pesquisa e constitui um problema em aberto. Entre as lacunas existentes nos estudos realizados até o momento está o fato desses estudos não considerarem que miRNAs podem se ligar a regiões distintas das 3'UTRs, em RNAs não codificantes ou até mesmo em DNA. Além disso, não relatam o efeito de cada variação alélica, nem tampouco analisam a variação de combinação alélica de mais de um SNP localizado no mesmo sítio alvo, informações importantes, uma vez que a frequência alélica de um SNP pode estar associada a uma doença. Outra limitação é que tais estudos analisam SNPs localizados somente em uma pequena região dos possíveis sítios alvos, conhecida como *seed*. Finalmente, os recursos computacionais utilizados nesses estudos não estão disponíveis como ferramentas de uso local e portanto não podem, por exemplo, serem aplicados a espécies distintas das analisadas no estudo.

Diante disso, este trabalho objetivou o desenvolvimento de uma ferramenta computacional *open-source* capaz de superar as lacunas existentes aqui mencionadas. A ferramenta desenvolvida chama-se

SIMTar (*SNPs Interfering in MicroRNAs Targets*). Para exemplificar o uso da ferramenta, SIMTar foi aplicado a SNPs associados a esquizofrenia.

O restante desta dissertação está organizado da seguinte forma. O capítulo 2 revisa os conceitos fundamentais importantes para o entendimento deste trabalho. O capítulo 3 detalha os objetivos gerais e específicos do projeto. O capítulo 4 descreve a metodologia empregada tanto para o desenvolvimento do SIMTar como para a sua aplicação em dados de esquizofrenia. O capítulo 5 apresenta e discute os resultados obtidos, onde o próprio SIMTar é considerado como resultado do trabalho, além dos resultados de sua aplicação em esquizofrenia. Finalmente o capítulo 6 traz as conclusões desta dissertação e trabalhos futuros. Está também incluso neste documento, em forma de apêndice, a lista de SNPs utilizados na etapa de aplicação.

Capítulo 2

Fundamentos

ESTE capítulo descreve os fundamentos teóricos dos principais tópicos utilizados neste trabalho. Entre eles estão o fluxo da informação genética, SNPs, miRNAs, doenças complexas e ferramentas computacionais para o estudo de SNPs em sítios alvos de miRNAs.

2.1 Fluxo da informação genética

Ácidos nucleicos são biomoléculas que contêm a informação genética dos organismos. Existem dois tipos de ácidos nucleicos nas células: o DNA e o RNA.

O DNA (*DeoxyriboNucleic Acid*) é uma molécula biológica replicativa presente em quase todos os organismos vivos como principal constituinte dos cromossomos. A sua estrutura é formada por duas cadeias helicoidais [Watson & Crick, 1953], cuja unidade estrutural básica é o nucleotídeo. Cada nucleotídeo é composto por: uma base nitrogenada, que pode ser uma adenina (A), citosina (C), guanina (G) ou timina (T), mais um grupo fosfato e uma molécula de açúcar (*deoxyribose*) [Saenger, 1984].

O RNA (*RiboNucleic Acid*) difere do DNA pela alteração da base nitrogenada Uracila (U) ao invés de Timina, e pela alteração da molécula de açúcar, que no RNA é a *ribose*. A síntese de RNA ocorre a partir de um molde de DNA através de um processo chamado transcrição.

Os cromossomos possuem sequências específicas de DNA que formam os genes. Em termos moleculares, um gene é comumente definido como uma sequência de DNA necessária para a síntese de RNA. Existem diferentes tipos de RNAs, aqueles que poderão atuar funcionalmente dentro da célula logo após serem transcritos (RNAs não codificadores - ncRNAs) e aqueles que passarão pelo processo de tradução para a síntese proteica (RNAs mensageiros - mRNAs) [Lodish *et al.*, 2000].

Internamente, um gene pode ser constituído por um éxon ou por um conjunto de éxons e íntrons (Figura 2.1). No caso de genes compostos de éxons e íntrons, os íntrons são removidos através de um processo denominado *splicing*. Neste processo podem ocorrer diferentes combinações de éxons e íntrons, dando origem a diferentes isoformas de um mesmo gene.

A sequência processada de um mRNA representa o conteúdo para síntese proteica – o quadro aberto de leitura – e nas extremidades duas sequências que não são traduzidas, a 5' e 3'UTRs (*Un-Translated Regions*). Dentro do quadro aberto de leitura, cada três nucleotídeos consecutivos – códon – é traduzido em um aminoácido. As extremidades 5' e 3'UTRs podem apresentar sequências com funções reguladoras e são responsáveis pela estabilidade da sequência do mRNA.

Inicialmente, o dogma central da biologia molecular defendia que o produto funcional das células eram somente as proteínas e o transcritoma eucariótico era composto por mRNAs, como elementos intermediários das proteínas, e por alguns ncRNAs: os RNAs ribossomais, transportadores e pequenos RNAs nucleolares, com funções estruturais e catalíticas envolvidos na síntese proteica [Crick, 1970] (Figura 2.2). Contudo, anos mais tarde, descobertas sucessivas revelaram uma grande variedade de ncRNAs (Lander *et al.* [2001]; Venter *et al.* [2001]). Hoje já são conhecidas mais de 2.200 famílias

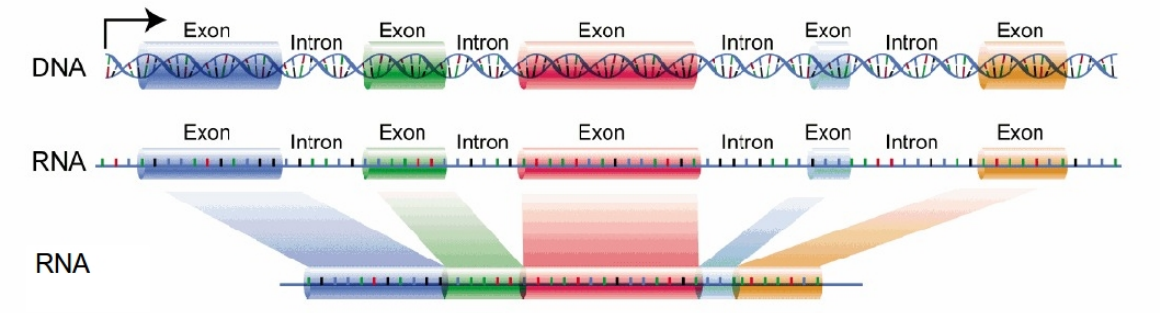

Figura 2.1: *Esquema geral da disposição estrutural de um gene.*
Adaptado de: `http://stochasticscientist.blogspot.com.br/`.

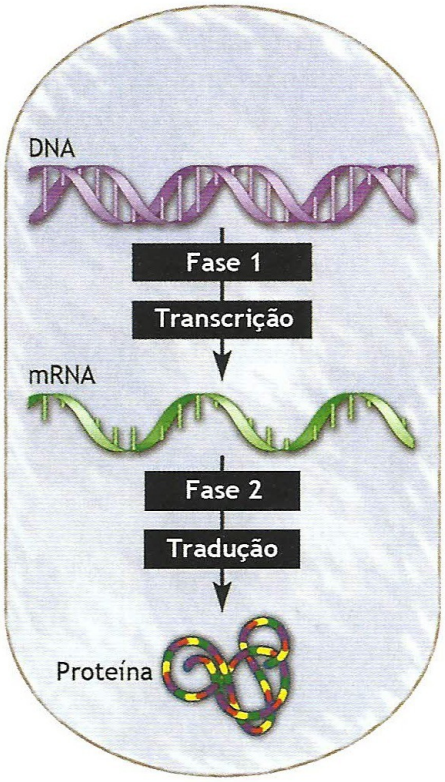

Figura 2.2: *Dogma central da biologia molecular.*
Fonte: `http://www.academiamalhacao.com.br/nikolascte/?paged=28`.

distintas de ncRNAs (Rfam versão 11.0) [Burge *et al.*, 2012]. Em especial, os miRNAs constituem uma parcela importante dessas famílias, com destacado papel em processos de regulação. Existem atualmente 2.578 sequências conhecidas de miRNAs em humanos (miRBase versão 20) [Griffiths-Jones, 2006].

2.2 Polimorfismos de um nucleotídeo (SNPs)

Por definição, polimorfismo é uma variante genética que ocorre em pelo menos 1% da população. SNPs são polimorfismos que ocorrem na sequência de DNA através da alteração de apenas uma posição nucleotídica (A, T, C ou G).

A genética clássica define o termo alelo como sendo a variação de um gene. Essa definição foi formulada em um período em que o gene era a unidade relacionada ao fenótipo. Entretanto, a genética molecular moderna observou que mesmo as variações que não são intragênicas podem interferir na expressão do fenótipo e convenciona-se chamar de alelo qualquer variação do genoma. Neste trabalho

o termo alelo é utilizado baseando-se nesta última definição. Assim, cada variação de um SNP é chamada de alelo e por definição pode ser bi-, tri- ou tetra-alélico, que corresponde ao número de possibilidades de bases nucleotídicas a serem trocadas, sendo duas, três ou quatro respectivamente para cada posição do SNP. A forma bi-alélica é a mais comum no genoma humano [Brookes, 1999].

A frequência de um SNP é dada em termos da sua menor frequência alélica (*Minor Allele Frequency* - MAF). Por exemplo, se a frequência do alelo menos frequente (G) é 0,40, dizemos que a frequência desse SNP é 40%. Desta forma, SNPs com frequência superior a 5% são chamados de comuns, e SNPs com frequência entre 0,5% e 5% são chamados SNPs raros [Bush & Moore, 2012].

Apenas 10% dos SNPs comuns associados às doenças estão localizados em regiões codificadoras de proteínas. Estes SNPs podem alterar o códon de leitura de tradução dos genes, o que irá depender de quais bases estão envolvidas na troca alélica e, neste caso, a troca pode ser sinônima ou não sinônima. Sinônima é quando a variação não implica na mudança de aminoácido para o qual o códon é traduzido, o que ocorre por conta da propriedade degenerada do código genético. Não sinônima é quando a variação altera o aminoácido codificado pelo códon. Estas alterações podem implicar na modificação da estrutura e estabilidade dessas regiões.

Para os demais 90% dos SNPs comuns associados às doenças, localizados em regiões não codificadoras de proteínas, muitos podem ser também os impactos. Tais SNPs, quando localizados na região promotora de um gene, podem alterar o sítio de ligação de fatores de transcrição ou de RNAs polimerases e, consequentemente, alterar vias metabólicas nas quais este gene está envolvido [Savinkova *et al.*, 2013]. Quando localizados em sítios de *splicing*, podem alterar o uso de éxons e consequentemente produzir diferentes isoformas de um gene (Faustino & Cooper [2013]; Kwan *et al.* [2008]; Nissim-Rafinia & Kerem [2005]). Quando localizados nas regiões 3' UTR de mRNAs, os SNPs podem modificar ou romper sítios de ligação de miRNAs [Lu & Clark, 2012]. Adicionalmente, uma vez que miRNAs podem se ligar em outras regiões distintas da 3'UTR, codificadoras ou não, transcritas ou não (revisado na seção 2.3), SNPs nestas regiões podem também interferir em seus sítios de ligação.

2.3 MicroRNAs

Os microRNAs (miRNAs) possuem um importante papel de regulação em muitos organismos eucariotos. Foram inicialmente reportados em *Caenorhabditis elegans* durante estudos com estágios do desenvolvimento larval [Lee *et al.*, 1993]. Entretanto, já estão descritos em muitos outros organismos (Brennecke *et al.* [2003]; Lagos-Quintana *et al.* [2003]; Lai *et al.* [2003]; Lim *et al.* [2003]; Reinhart *et al.* [2002]; Sempere *et al.* [2003]). Muitos destes miRNAs são conservados filogeneticamente, enquanto outros são espécie-específicos (Lagos-Quintana *et al.* [2002]; Pasquinelli *et al.* [2000]).

2.3.1 Biogênese dos microRNAs

Em animais os genes de miRNAs podem tanto ser expressos individualmente como também em pequenos grupos de 2 a 7 genes (Figura 2.3) (Bartel [2009]; Baskerville & Bartel [2005]; Lau *et al.* [2001]; Lee *et al.* [2002]). Estes genes são sintetizados pelas enzimas RNA polimerase II e III (Borchert *et al.* [2006]; Lee *et al.* [2004]). Podem ainda ter origem a partir de íntrons de genes codificadores de proteínas, íntrons e éxons de genes não codificadores [Rodriguez *et al.*, 2004], ou mesmo a partir da região 3' UTR de genes codificadores de proteínas [Cai *et al.*, 2004].

A síntese de miRNAs ocorre de maneira diferente em plantas e animais. Em plantas o processo ocorre inteiramente no interior do núcleo, desde o processamento do transcrito primário até a formação da sua sequência madura, que é a forma ativa. Já em animais a síntese ocorre no núcleo a partir de um longo transcrito primário, representado pelo Pri-miRNA na Figura 2.3 e que apresenta uma estrutura de um ou mais grampos. Este transcrito é então pré-processado pela enzima Drosha gerando um ou mais Pré-miRNA com tamanho aproximado de 70 nts, em estrutura de um grampo. O Pré-miRNA é então transportado para o citoplasma através de partículas ribonucleoproteicas (Exportina 5), onde é processado pelo complexo enzimático Dicer, que elimina a estrutura de grampo, abre a fita dupla de RNA e gera duas sequências simples de miRNAs. Cada sequência gerada tem em média 22 nts de comprimento e geralmente apenas uma é ativa e a outra é degradada, mas em alguns casos, ambas são funcionais (Bohnsack *et al.* [2004]; Lee *et al.* [2002]; Lund *et al.* [2004]; Yi *et al.* [2003]; Zeng &

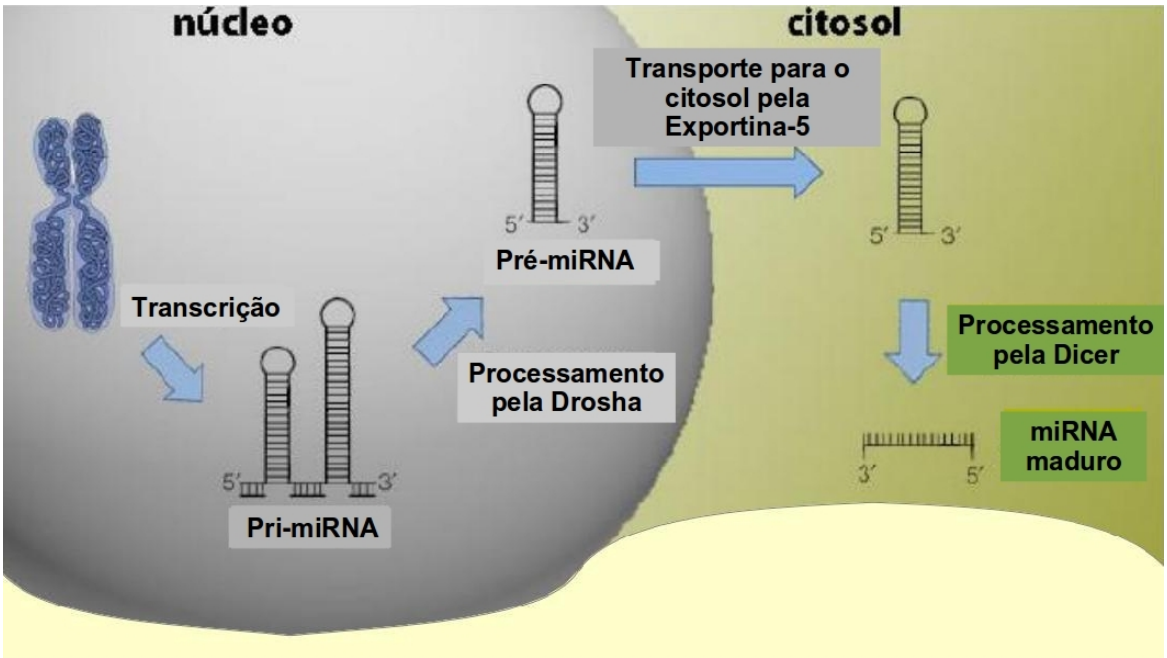

Figura 2.3: *Biogênese de miRNAs em animais.* *Adaptado de[Oliveira-Carvalho* ET AL.*, 2012]*

Cullen [2003]). A sequência madura é a forma ativa do miRNA, aquela que desempenha a regulação sobre os genes alvos.

2.3.2 Nomenclatura dos microRNAs

A atribuição do nome para um miRNA ocorre depois que este miRNA já foi validado experimentalmente e possui uma publicação relacionada. Basicamente a nomenclatura utilizada para nomear os miRNAs contém: um código que indica a espécie de origem, a informação sobre o tipo de sequência (informando se é a sequência precursora ou madura) e a origem da sequência no processo de biogênese.

Utilizando como exemplo um miRNA identificado em *Homo sapiens* (Figura 2.4), o *hsa* indica a espécie (*Homo sapiens*). O prefixo mir com a letra 'r' minúscula indica que a sequência é um Pré-miRNA e portanto não é a forma madura, já o prefixo miR com a letra 'R' maiúscula indica que é uma sequência madura. Embora a sequência madura seja a forma ativa do miRNA, também são nomeadas e estão disponíveis no banco de dados de miRNAs as suas formas precursoras (miRBase). O último campo contém a origem da sequência. Esta origem é atribuída em função da sequência precursora do miRNA, por ser uma dupla fita, pode dar origem a uma ou duas sequências funcionais. Neste caso,

cada sequência funcional vai ter origem de uma extremidade da fita, podendo ser da porção 3' (3p) ou 5' (5p) (Figura 2.5a). Esta é a nomeclatura atualmente utilizada para a origem da sequência, no entanto ela é sucessora do uso do asterisco (*), que era utilizada para indicar a forma menos frequente quando as duas sequências do miRNA são funcionais (Figura 2.5a).

MiRNAs altamente relacionados (com sequências altamente simulares) são agrupados em famílias de acordo com a sua sequência precursora. No exemplo mencionado da Figura 2.5, o código da família é mostrado pelo miR-24, sendo as formas maduras miR-24-3p, miR-24-1-5p e miR-24-2-5p todos membros da mesma família e diferentes em apenas alguns nucleotídeos [Griffiths-Jones, 2004]. Letras também podem ser utilizadas, como por exemplo a família hsa-miR-121, na qual os miRNAs maduros hsa-miR-121a e hsa-miR-121b são expressos dos precursores hsa-mir-121a e hsa-mir-121b, respectivamente.

2.3.3 Pareamento do microRNA com o sítio alvo

Para que a sequência madura do miRNA possa se ligar à sequência alvo e efetuar a regulação, primeiro ela deve ser reconhecida por um complexo proteico de ação reguladora chamado RISC (*RNA-Induced Silencing Complex*) [Lee *et al.*, 2003]. O complexo RISC possui como principal elemento catalítico as proteínas da família Argonauta (Ago), já observado mediando a regulação de vários sítios alvos. Em especial as proteínas argonautas têm permitido a identificação de novas vias de regulação por miRNAs (descritas na subseção 2.3.4).

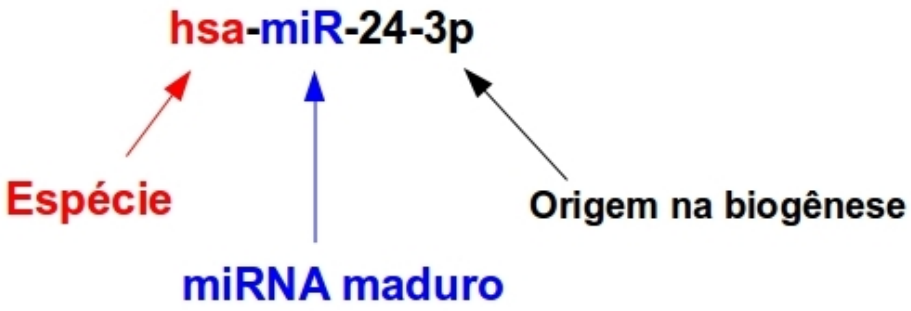

Figura 2.4: *Nomenclatura de miRNAs.*

Em plantas, o pareamento do miRNA com o sítio alvo ocorre geralmente através de uma perfeita complementaridade facilitando estudos computacionais de predição de genes alvos [Dai & Zhao, 2011]. Já em animais a complementaridade perfeita é incomum enquanto o pareamento imperfeito ocorre frequentemente, permitindo que um miRNA possa se ligar a diferentes sequências. Assim, um mesmo miRNA pode se ligar em diferentes regiões de um mesmo alvo, bem como, um gene ou RNA pode ser alvo de vários miRNAs (Ambros [2003]; Baskerville & Bartel [2005]; Reinhart *et al.* [2002]; Vella *et al.* [2004]). Tal fato requer sofisticadas ferramentas para predição de alvos de miRNAs em animais.

Inicialmente em estudos experimentais realizados para analisar a relação do pareamento e a efetividade da regulação, observou-se que existe uma correlação entre a efetividade e a energia livre da ligação do alvo com uma pequena região do pareamento chamada de semente (*seed*) [Doench & Sharp, 2004]. Trata-se de uma região com pareamento Watson-Crick (A-U, C-G) entre as duas moléculas, localizada a partir da extremidade 5' do miRNA, entre as posições 1–8 do mesmo. Esta região é extremamente conservada filogeneticamente entre miRNAs de uma mesma família. Em outros estudos mostrou-se que é através da região *seed* que o complexo RISC faz o reconhecimento do

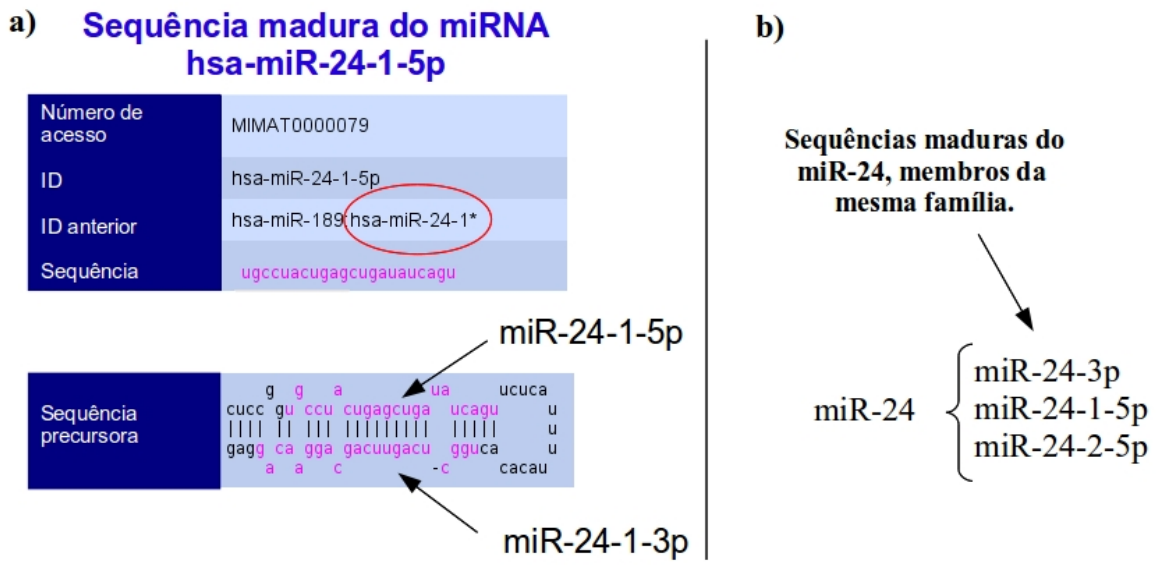

Figura 2.5: *Nomenclatura de sequências maduras e de famílias de miRNAs.*
a) sequência madura do miRNA miR-24-1-5p. A sequência precursora mostra a origem das duas sequências maduras desse miRNA. A extensão 5p usada para a sequência gerada a partir da extremidade 5' e 3p para sequência gerada a partir da extremidade 3'. O círculo vermelho destaca a nomenclatura em asterisco '' utilizada anteriormente.*
b) miRNAs membros de uma mesma família. *Fonte: miRBase [Griffiths-Jones, 2006].*

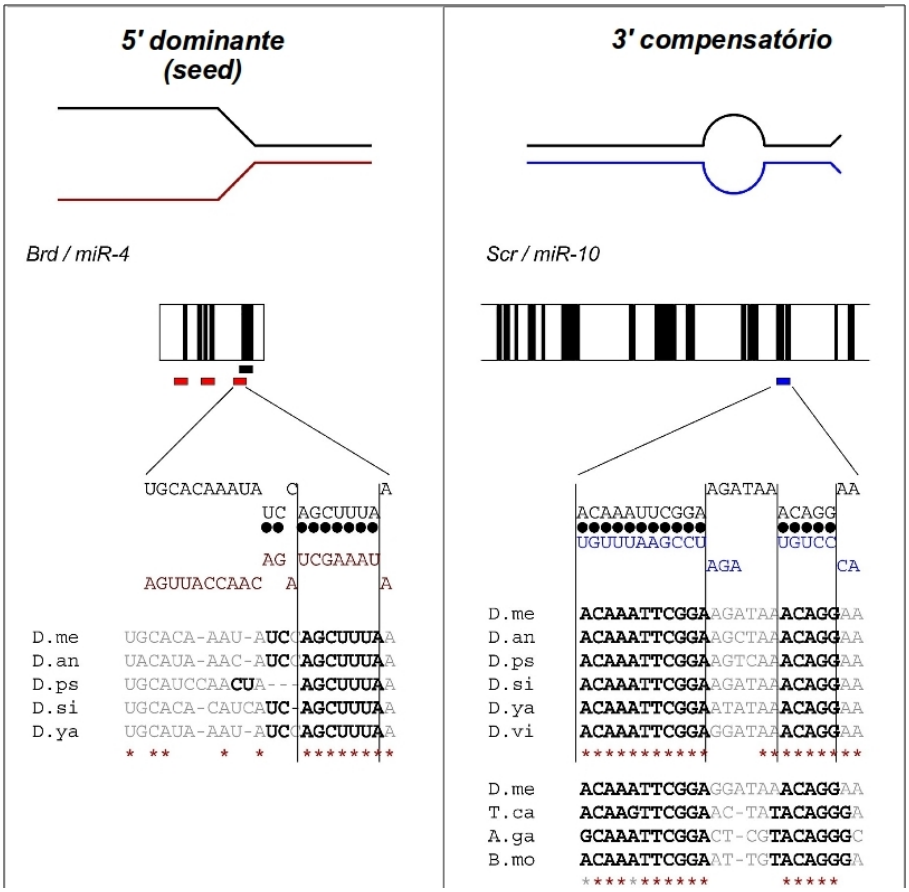

Figura 2.6: *Pareamento estritamente perfeito da região seed com o sítio alvo no primeiro quadro e da região 3' do miRNA compensando o pareamento imperfeito com a seed.* *Adaptado de:[Brennecke* ET AL.*, 2005]*

sítio alvo (Lewis *et al.* [2003]; Rajewsky & Socci [2004]; Stark *et al.* [2003]). Hoje já se sabe que não somente esta região é fundamental para o pareamento, mas ao invés disso, pareamentos ocorridos na extremidade 3' do miRNA podem compensar pareamentos imperfeitos com a *seed*. A Figura 2.6 mostra pareamentos nestas duas regiões.

O primeiro tipo de pareamento observado entre a *seed* e o sítio alvo ocorre através de uma complementaridade perfeita entre todas as bases. Entretanto, pareamentos imperfeitos nessa região, incluindo aqueles entre as bases G:U ([Doench & Sharp, 2004]), também são identificados pelo complexo RISC como funcionais. Este último é conhecido como não estritamente perfeito (Johnston & Hobert [2003]; Moss *et al.* [1997]).

Considerando o pareamento perfeito, dependendo da configuração do pareamento entre as posições de 1 a 8 do miRNA com o sítio alvo (Figura 2.7), a *seed* pode ocorrer de 4 diferentes maneiras:

`6mer`, `7mer-A1`, `7mer-m8` e `8mer` [Lewis *et al.*, 2005]. Uma característica comum dessas 4 maneiras é que todas conservam o pareamento de pelo menos 6 pares de bases (pb) entre as posições 2–7 do miRNA. A região `7mer-A1` conserva o pareamento de 6 pb e ainda possui uma adenina (A) na posição 1 do miRNA. A *seed* `7mer-m8` possui, além dos 6 pb, um pareamento com o alvo na posição 8. A região `8mer` apresenta o pareamento `6mer` mais uma adenina na posição 1 e um pareamento na posição 8 do alvo. E por fim, a região `6mer` possui somente o pareamento de 6 pb [Grimson *et al.*, 2007].

Medir o nível de eficiência de cada um destes tipos de *seed* não é uma tarefa fácil. Entretanto alguns estudos experimentais conseguiram identificar uma hierarquia de eficiência relativa entre elas: `8mer` > `7mer-m8` > `7mer-A1` > `6mer`, onde '>' indica a relação de 'apresenta eficiência maior que'(Bartel [2009]; Hammell *et al.* [2008]).

Ainda que a presença de uma possível *seed* seja em si uma forte característica utilizada na predição de sítios alvos de um dado miRNA, ela não garante alvos funcionais. Além disso, existem outras características que contribuem na ligação como o pareamento da extremidade 3' do miRNA com o alvo, que pode compensar pareamentos imperfeitos com a *seed* (Brennecke *et al.* [2005]; Doench & Sharp [2004]) e a acessibilidade do sítio alvo, que inclui entre outras: a energia livre gerada no pareamento do miRNA com o sítio alvo, uma característica que permite medir a espontaneidade da ligação e a presença de uma região nucleotídica rica em AU que reside preferencialmente ao redor de sítios efetivos [Grimson *et al.*, 2007].

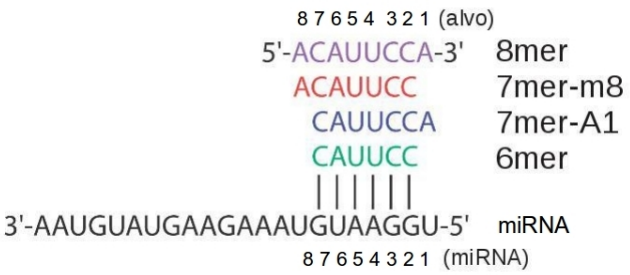

Figura 2.7: *Descrição dos diferentes tipos de seed observados no pareamento de um miRNAs com o sítio alvo em animais.* Adaptado de: http://hollywood.mit.edu/

2.3.4 Formas de regulação

A complementaridade perfeita do miRNA com o sítio alvo implica geralmente na clivagem e o enfraquecimento da sequência alvo [Valencia-Sanchez *et al.*, 2006]. Já o pareamento imperfeito, pode levar à regulação positiva ou negativa através de vias pré ou pós-transcricionais como descritas abaixo.

Em eucariotos a mais conhecida via de regulação por miRNAs ocorre pós-transcricionalmente com regulação negativa de genes codificadores de proteínas. Neste processo, mesmo após ter sido transcrito, o gene tem sua tradução reprimida pela ação do miRNA que se liga na extremidade 3'UTR da sua sequência de mRNA. Este processo, além de reprimir a tradução também leva à desestabilização do gene alvo [Jackson *et al.*, 2010].

A ativação da tradução foi observada tanto por miRNAs se ligando à 3'UTR como à 5'UTR. Como exemplo o miR-369-3p e miR-206, que se ligam à 3'UTR e ativam a tradução das sequências de mRNA dos genes TNFα e KLF4 respectivamente, ambos genes envolvidos na quiescência celular em mamíferos (Lin *et al.* [2011]; Vasudevan *et al.* [2007]). Um outro exemplo de ativação da tradução ocorre em tipos celulares específicos, como observado pelo miR-122 que é um miRNA específico do fígado e estimula a tradução da sequência de RNA do vírus da Hepatite C através da ligação direta em dois sítios alvos na região 5'UTR [Henke *et al.*, 2008].

Uma outra via de regulação por miRNAs ocorre através da sua ligação a regiões de sequências codificadoras de proteínas (*Coding Sequence* - CDS). A ligação a CDS é mais comum em plantas, no entanto foi observada em humanos pela família do miR-148 regulando a tradução do gene *Dnmt3b* (*DNA methyltransferase 3b*). Um dos sítios alvos do miR-148 é conservado evolutivamente em três de quatro variantes de *splicing* deste gene (*Dnmt3b1*, *Dnmt3b2* e *Dnmt3b4*) [Duursma *et al.*, 2008]. Assim, o variante de *Dnmt3b3* tem sua tradução aumentada em relação às outras variantes. O gene *Dnmt3b* está envolvido na metilação do DNA, que é uma modificação epigenética que pode envolver o silenciamento gênico, remodelagem da cromatina e estabilidade genômica [Jones & Baylin, 2007]. O miRNA agindo via CDS contribui para a regulação específica de variantes de *splicing*.

MiRNAs também regulam sequências de DNA pela sua ligação à região promotora do gene. Essa ligação pode induzir ao silenciamento gênico via pré-transcricional como exemplificado pelo miR-

138 e miR-27a envolvidos na regulação do gene MDR1. Neste exemplo, o miRNA pode agir direta ou indiretamente sobre o sítio do promotor. O mecanismo de regulação do DNA por miRNAs ocorre através da importação da maquinaria de regulação do miRNA presente no citoplasma para dentro do núcleo da célula. Este processo foi observado em células de adenocarcinoma de cérvix (HeLa), uma linhagem celular bastante utilizada na comunidade científica, onde foi observado a translocação da Ago2 para dentro do núcleo pela proteína de transmembrana *Importin* 8. Com isso, mostrou-se que a *Importin* 8 é essencial no silenciamento gênico como cofator da proteína Ago2 que faz a mediação da regulação [Weinmann *et al.*, 2009]. O mecanismo de regulação direta envolve a proteína Ago2 translocada do citoplasma, que conduz o miRNA a se ligar a uma fita simples do DNA obtida durante a bolha de transcrição, uma etapa em que o promotor está ativo. Já a regulação indireta está relacionada à estabilização de um triplex de sequências. Neste último caso, a Ago2 leva o miRNA a inativar a sequência do promotor através da sua ligação à dupla fita de DNA. Em ambos os casos a RNA polimerase II é impedida de efetuar a transcrição. A proteína codificada pelo gene exemplificado MDR1 está envolvida na resistência a multidrogas, frequentemente observada mediando o desenvolvimento de resistência a drogas anticâncer [Toscano-Garibay & Aquino-Jarquin, 2012].

Em um outro exemplo de regulação por miRNAs ligados à região promotora, a expressão gênica é ativada em um fenômeno chamado de ativação de RNA (regulação positiva por miRNAs). Tal fato foi observado pelos miRNAs miR-744 e miR-1186 ligados ao promotor do gene Ccnb1 (*Cyclin B1*) [Huang *et al.*, 2012]. O produto deste gene é uma proteína reguladora envolvida no processo de mitose.

Diferente da regulação de DNA e mRNAs, os miRNAs também podem se ligar em sequências de outros RNA não codificadores, inclusive que servem de substrato para biogênese de miRNAs, como observado em *C. elegans* pelo miRNA let-7. Neste caso, a proteína argonauta foi observada se ligando em um sítio específico da extremidade 3' do let-7 (a extremidade oposta da região *seed*), promovendo eventos de processamento à jusante. Este fato amplia ainda mais as formas de regulação por miRNAs, bem como o papel da argonauta nessa autoregulação de miRNAs [Zisoulis *et al.*, 2012].

Desta forma, os miRNAs podem atuar ativando ou reprimindo tanto a transcrição como a tradução, podendo se ligar em regiões codificantes ou não, nos RNAs ou no DNA. Além disso, estão envolvidos

em uma ampla gama de processos biológicos como a diferenciação [Chen *et al.*, 2004], proliferação e morte celular (Frankel *et al.* [2008]; Si *et al.* [2006]). Consequentemente, um mau funcionamento na regulação por miRNAs pode estar associado com doenças [Büssing *et al.*, 2008].

2.4 Predição computacional de alvos de microRNAs

Existem várias ferramentas disponíveis para predição de alvos de miRNAs. Tais ferramentas podem ser distinguidas através de algumas características utilizadas na predição de alvos.

2.4.1 Características do pareamento miRNAs-alvo

Conservação evolutiva

Na região *seed* estão pareados apenas alguns poucos nucleotídeos e como consequência deste fato são preditos muitos alvos falsos-positivos. Uma forma de melhorar a especificidade dos resultados de predição é selecionar apenas alvos evolutivamente conservados, como mostrado na Figura 2.8.

Pelo fato de muitos miRNAs e seus alvos serem conservados evolutivamente, algumas estratégias de predição de sítios alvos exigem a presença dessa característica (Friedman *et al.* [2009]; Lall *et al.* [2006]; Lewis *et al.* [2005]). A vantagem de utilizar essa restrição é reduzir o número de resultados falsos-positivos. No entanto, esta estratégia deixa de recuperar por exemplo os resultados espécie-específicos (Carè *et al.* [2007]; Lu *et al.* [2008]; Yang *et al.* [2007]).

Pareamento com a *seed*

Como descrito na subseção 2.3.3, a *seed* é uma região importante para o reconhecimento do miRNA pelo complexo RISC no processo de regulação. Esta região pode ocorrer perfeitamente ou imperfeitamente pareada com o sítio alvo e ainda assim ser efetiva no processo de regulação. A vantagem de se usar diferentes tipos de região *seed* é possuir um maior controle de especificidade da predição. A *seed* `8mer` é mais rara de se encontrar ao acaso do que a `6mer` e, portanto, procurar apenas por *seed* `8mer` contribui para uma maior especificidade. No entanto, a predição terá uma sensibilidade

prejudicada, uma vez que alvos funcionais com os outros tipos de *seed* não serão encontrados. Por outro lado, aceitar outros tipos de *seed* aumenta a sensibilidade mas diminui a especificidade.

Acessibilidade e contexto do sítio alvo

Para que o sítio alvo possa ser reconhecido pelo complexo RISC e a ligação seja feita, a estrutura secundária da molécula alvo deve ser tal que permita que a região do sítio esteja acessível. Também foi identificado uma composição nucleotídica rica em AU localizada em aproximadamente 30 nts antes e depois do sítio alvo e o posicionamento do sítio a uma distância de pelo menos 15 nts após o códon de parada (para alvos na 3'UTR de mRNAs) [Grimson *et al.*, 2007].

Número e proximidade de sítios alvos

Uma mesma molécula de RNA ou um trecho de DNA pode possuir vários sítios de miRNAs [John *et al.*, 2004]. Foi observado que a presença desses múltiplos sítios tende a apresentar maior eficiência na regulação comparado a sítios únicos. No entanto, para serem efetivos estes múltiplos sítios devem estar dentro de uma distância mínima. De acordo com Sætrom *et al.* (2007), uma distância ótima está preferencialmente entre 17 e 35 nts, mas a distância de 14 e 46 nts também mostrou um aumento

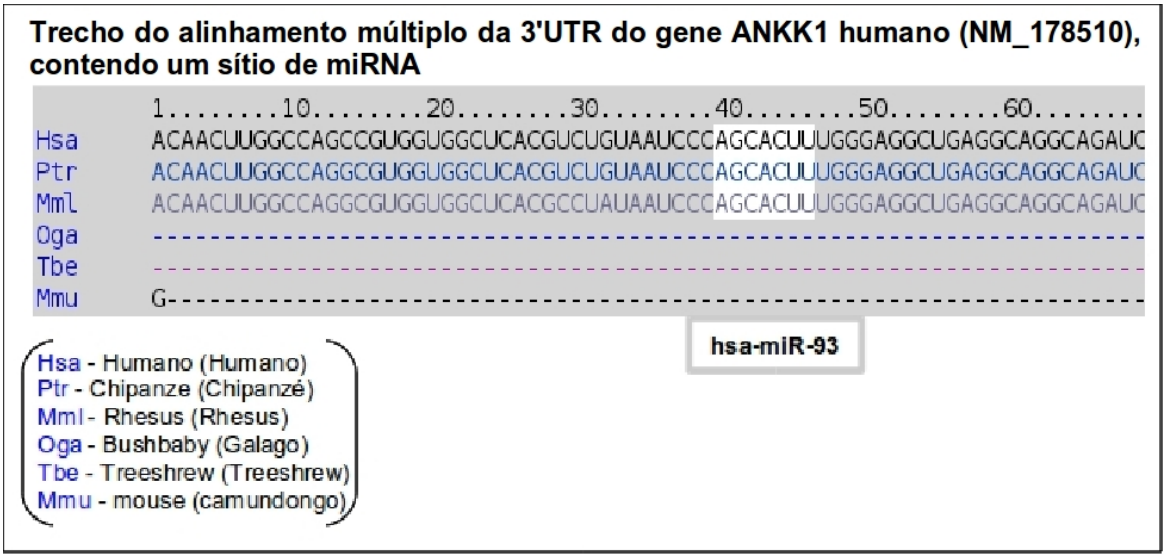

Figura 2.8: *Trecho do alinhamento múltiplo da 3'UTR do gene ANKK1 humano, com a região seed do miRNA miR-93 se ligando próximo a posição 40 da 3'UTR (destacado em branco). Essa região está conservada em Humanos, Chipanzé e Rhesus.*

dessa eficiência. A cooperação entre mais de um sítio é efetiva para um mesmo miRNA e também para miRNAs diferentes.

2.4.2 Ferramentas de predição de alvos de microRNAs

Existem várias ferramentas de predição de alvos de miRNAs, contudo uma descrição de cada uma seria muito extensa e por isso escolheu-se listar algumas delas apontando o uso ou não do critério de conservação da região alvo e a disponibilidade para execução local (Tabela 2.1).

As ferramentas disponíveis não aplicam todas as características de predição de alvos mencionadas na seção 2.4.1, mas apenas uma ou algumas delas. Em um estudo de revisão de literatura, Saito & Sætrom (2010) mostram que a comparação entre as ferramentas de predição é uma difícil tarefa, pois os genes alvos preditos por elas não se sobrepõem muito bem. Além disso, um outro limitante é a disponibilidade para download das ferramentas, um requesito importante para pesquisadores que desejam utilizar seus próprios dados de miRNAs e sequências alvos, bem como aplicar as predições sobre uma espécie que não é padrão das ferramentas. Neste caso, apenas as ferramentas disponíveis podem ser utilizadas.

Dessa maneira, a escolha de uma ferramenta é feita em função das características que melhor se enquadram no estudo em questão, além de se considerar a disponibilidade para execução local. Neste estudo, os autores apontaram as três melhores ferramentas quanto ao desempenho de predições já analisadas, são elas: TargetScan, PicTar e miRanda.

2.5 Estudos de SNPs em sítios alvos de microRNAs

O miRNA se liga em um pequeno sítio da sequência alvo e qualquer alteração nucleotídica que ocorra neste sítio alvo pode interferir diretamente no pareamento, seja contribuindo ou desfavorecendo a sua afinidade, ambos diretamente relacionados à efetividade da regulação. Neste sentido, têm-se investigado as implicações que polimorfismos de um nucleotídeo (SNPs) podem exercer sobre sítios alvos de miRNAs, uma vez que a sua variação alélica pode criar ou romper um sítio alvo ou ainda interferir na estabilidade da ligação. Uma vez que muitos SNPs e miRNAs isoladamente já

Tabela 2.1: *Lista de ferramentas de predição de alvos de miRNAs e algumas características.*
[a]conservação do sítio alvo. •: sim, '-': não.
[b]disponibilidade de execução local. •:sim, '-':não

Ferramenta	Conserv[a]	Disp[b]	Referência
TargetScan	•	•	Friedman *et al.* [2009]; Grimson *et al.* [2007]; Lewis *et al.* [2005]
PicTar	•	-	Chen & Rajewsky [2006]; Lall *et al.* [2006]; Grün *et al.* [2005]; Krek *et al.* [2005]
miRanda	-	•	Betel *et al.* [2008]; John *et al.* [2004]
MicroCosm Targets	-	-	Griffiths-Jones *et al.* [2008]; Griffiths-Jones [2006]; Enright *et al.* [2003]
RNAhybrid	-	•	Krüger & Rehmsmeier [2006]; Rehmsmeier *et al.* [2004]
PITA	•	•	Kertesz *et al.* [2007]
STarMir	-	-	Long *et al.* [2007]
Rajewsky & Socci	-	-	Rajewsky & Socci [2004]
Robins	-	-	Robins *et al.* [2005]
mirWIP	-	•	Hammell *et al.* [2008]
MicroInspector	-	-	Rusinov *et al.* [2005]
MicroTar	-	•	Thadani & Tammi [2006]
MirTarget2	•	-	Wang & El Naqa [2008]
miTarget	-	-	Kim *et al.* [2006]
TargetMiner	-	-	Bandyopadhyay & Mitra [2009]
EIMMo	-	-	Gaidatzis *et al.* [2007]
NbmiRTar	-	-	Yousef *et al.* [2007]
TargetBoost	-	-	Sætrom *et al.* [2005]
RNA22	-	-	Miranda *et al.* [2006]
TargetRank	•	-	Nielsen *et al.* [2007]
MovingTarget	-	-	Burgler & Macdonald [2005]
DIANA-microT	-	-	Kiriakidou *et al.* [2004]

foram associados a várias doenças, encontrar uma relação entre eles pode ajudar na compreensão das bases moleculares dessas doenças e auxiliar no desenvolvimento de novas abordagens terapêuticas e métodos diagnósticos.

Inicialmente, os recursos computacionais para estudos de SNPs e miRNAs focaram principal-

mente em SNPs associados a doenças e localizados em sítios alvos de miRNAs preditos na 3'UTR de mRNAs.

Neste sentido, Ziebarth *et al.* (2012) criaram o banco de dados PolymiRTS para disponibilizar um repertótio de SNPs localizados em sítios alvos de miRNAs. Este banco disponibiliza dados de humano e camundongo, bem como integra dados de expressão de locos de características quantitativas (eQTLs) junto com resultados de GWAS. Entre as principais características do PolymiRTS estão: 1) análise da região 3'UTR de mRNAs, 2) uso de genes codificadores associados a doenças humanas, 3) polimorfismos em sítios alvos de miRNAs validados experimentalmente e 4) SNPs em região *seed*.

Em um outro trabalho, Bruno *et al.* (2012) desenvolveram o banco de dados miRdSNP, o qual também disponibiliza dados de SNPs associados às doenças. miRdSNP fornece uma fonte de dados para exploração da distância que estes SNPs estão de sítios alvos de miRNAs dentro de regiões de 3'UTR. Além disso, tais dados estão disponíveis apenas para a espécie humana.

Liu *et al.* (2012) desenvolveram o banco de dados MirSNP. Este banco contém informações semelhantes comparado aos anteriores, mas neste caso os dados de eQTL e GWAS podem ser consultados separadamente. MirSNP difere dos anteriores por disponibilizar dados de SNPs localizados também em genes de miRNAs.

Também na forma de um banco de dados, Gong *et al.* (2012) desenvolveram o miRNASNP, o primeiro recurso computacional a disponibilizar dados de SNPs criando sítios alvos de miRNAs. No entanto, o resultado não detalha as condições obtidas de todos os alelos. Ademais, os dados disponibilizados são restritos à 3'UTR de mRNAs e os SNPs localizados no interior da região *seed*.

Diferente dos trabalhos anteriores, Barenboim *et al.* (2010) desenvolveram uma ferramenta web chamada MicroSNiPer. Esta ferramenta disponibiliza a busca por SNPs que estão localizados na região *seed* dos miRNAs baseando-se nos critérios de alvos em 3'UTR de mRNAs. Embora permita a execução de novas buscas de pares SNP-miRNAs, MicroSNiPer não está disponível para execuções locais e é restrito para *Homo sapiens* e *Mus musculus*.

Em resumo, os estudos realizados até o momento no sentido de identificar SNPs interferindo em alvos de miRNAs assumem apenas o mecanismo de regulação negativa de tradução por miRNAs, assumindo que os sítios dos miRNAs estejam nas regiões 3' UTR de mRNAs. Nenhum dos recursos

está disponível para execução local de novas predições de sítios alvos de miRNAs.

O fato é que foram descobertas novas vias de regulação por miRNAs e formas alternativas do pareamentos entre o miRNA e o sítio alvo e, portanto, um sítio alvo de miRNA pode ser alterado tanto pela variação alélica de um SNP, quanto pela combinação de mais de um alelo, como pode ocorrer quando SNPs vizinhos estão localizados em um mesmo sítio alvo. No entanto, os atuais recursos computacionas não mostram possíveis combinações alélicas de vários SNPs localizados dentro de um mesmo sítio. Além disso, os resultados obtidos a partir dessas fontes atualmente disponíveis não incluem dados das novas formas de regulação e pareamentos de miRNAs, como também não mostram a implicação da variação alélica do SNP sobre cada sítio alvo, dado que um alelo pode criar um novo sítio enquanto o outro pode romper um sítio previamente existente ou ao invés disso, a variação alélica do SNP pode interferir na estabilidade do pareamento levando a maior ou menor efetividade de regulação. Tais análises somente são possíveis quando se têm os resultados de predição para cada alelo do SNP. Por fim, os recursos disponíveis atualmente são restritos a humano e camundongo, embora os miRNAs já tenham sido observados em muitos outros eucariotos.

2.6 Doenças complexas e esquizofrenia

Doenças humanas sofrem interferência de fatores genéticos. As doenças de forma geral são categorizadas como simples ou complexas. Enquanto nas doenças simples o fenótipo está relacionado à ação de apenas um gene, nas complexas o fenótipo é o resultado da interação de vários genes, daí sua classificação como poligênica, além da interferência de fatores não genéticos relacionados ao meio ambiente.

Enquanto nas doenças simples o padrão de herança genética é facilmente definido como mendeliano, ou seja, apenas um gene é responsável pela transmissão hereditária do conteúdo genético responsável pela doença, em doenças complexas é mais difícil estabelecer este padrão de herança, pois não somente estão envolvidos vários genes como também estão envolvidos alguns fatores não genéticos. Isso significa que embora o indivíduo contenha predisposição genética a uma doença, o produto dos genes (expressão gênica) relacionados estarão ainda sob influência de fatores não genéti-

cos, que então combinados, exibirão o fenótipo suscetível. Neste grupo de doenças complexas estão inclusas: câncer, asma, diabetes e a maioria das doenças psiquiátricas como Parkinson, Alzheimer e esquizofrenia [Hunter, 2005].

Esquizofrenia é um transtorno psiquiátrico de grande impacto social e que afeta aproximadamente uma a cada cem pessoas no mundo [McGrath *et al.*, 2008]. Estudos realizados com gêmeos monozigóticos representam a principal evidência da influência não genética para a doença. Considerando dois irmãos gêmeos monozigóticos, seria esperado que ambos apresentassem igualmente a expressão do fenótipo, quando um deles é afetado. O fato é que ocorrem casos em que somente um deles é afetado, mostrando que uma parte da influência é decorrente de fatores não genéticos. Em esquizofrenia, os padrões de herança genética são variáveis e não consistentes com os traços de doenças mendelianas [Harrison & Weinberger, 2005].

A herdabilidade de esquizofrenia é de aproximadamente 80%, valor que indica que a sua causa é pelo menos parcialmente decorrente de fatores genéticos [Sullivan, 2005]. Mesmo essa porcentagem sendo alta, muitos dos seus componentes genéticos ainda continuam sendo identificados. Dados gerados a partir de estudos de associação sobre todo o genoma (*Genome Wide Association Study* - GWAS) em esquizofrenia mostram que os SNPs comuns correspondem a 23% da variação de suscetibilidade em esquizofrenia [Lee *et al.*, 2012].

Uma alternativa para recuperar os resultados estatisticamente mais relacionados em cada caso em particular são os estudos de meta-análise. O termo meta-análise foi inicialmente definido por Glass (1976), como uma análise de integração (estatística) de uma grande coleção de resultados de análises individuais. Dessa maneira é possível obter os resultados de meta-análise de GWAS.

Capítulo 3

Objetivos

O objetivo geral deste trabalho foi implementar uma ferramenta computacional para identificar polimorfismos de um nucleotídeo (SNPs) que, através de sua variação alélica, interferem em sítios alvos de miRNAs, preencheendo as lacunas dos recursos computacionais existentes no estudo de SNPs em sítios alvos de miRNAs.

Como objetivos específicos tem-se:

a- Desenvolver uma ferramenta computacional capaz de:

i) analisar SNPs localizados em RNAs ou em qualquer região genômica;

ii) analisar os efeitos de todas as combinações alélicas de SNPs próximos localizados em um mesmo sítio alvo;

iii) identificar o efeito do SNP mesmo quando fora da região *seed*;

iv) gerar os resultados para cada alelo dos SNPs identificados no item a-i;

v) ser aplicada a qualquer espécie desde que o usuário forneça certos arquivos específicos da espécie de interesse;

b- Disponibilizar a ferramenta para execuções locais como um programa *open-source*;

c- Exemplificar a utilização da ferramenta através de sua aplicação em dados de esquizofrenia.

Capítulo 4

Metodologia

ESTE capítulo descreve as características da ferramenta SIMTar desenvolvida neste trabalho, bem como a sua estrutura e implementação. Ademais também descreve a obtenção dos dados de SNPs associados à esquizofrenia e sua utilização como estudo de caso da ferramenta.

4.1 Características da ferramenta SIMTar

SIMTar (*SNPs Interfering in MicroRNA Targets*) é uma ferramenta para a predição de SNPs interferindo em sítios alvos de miRNAs. Assim, a fim de superar as limitações dos atuais recursos computacionais nesta área, SIMTar foi concebido de forma a apresentar as seguintes características:

- **Diferentes abordagens na predição de sítios alvos de miRNAs:**

  Na etapa de predição de sítios alvos de miRNAs, SIMTar utiliza duas ferramentas já existentes. Na escolha das ferramentas objetivou-se disponibilizar ao usuário mais de uma abordagem de predição. As três ferramentas identificadas com melhor desempenho no trabalho de Saito & Sætrom (2010) são: TargetScan, PicTar e miRanda. TargetScan e PicTar utilizam a característica de conservação evolutiva do sítio alvo como critério de execução, diferente da ferramenta miRanda que é capaz de analisar regiões não necessariamente conservadas permitindo assim a identificação de alvos espécie-específico. Quanto à disponibilidade, apenas TargetScan e miRanda possuem código

aberto. Assim, TargetScan e miRanda foram selecionadas para serem utilizadas por SIMTar.

A ferramenta MiRanda é composta de um programa executável e TargetScan é composto de três programas executáveis. Cada ferramenta faz uso de alguns critérios para predição de sítios alvos.

MiRanda considerada os valores de escore e energia-livre. O escore é o valor dado em função da complementaridade de bases do miRNA com a sequência alvo e a energia-livre é calculada a partir desse pareamento.

Para o TargetScan são considerados: a) a probabilidade de conservação evolutiva do alvo (*probability of conserved targeting* – pct), b) valor de contexto do sítio alvo (*context score*) e c) o tipo da região *seed*. Valores de pct próximos de 1 indicam alvos mais conservados (para mais detalhes ver Friedman *et al.* (2009)). O valor de contexto do sítio alvo (*context score*) considera o tipo do sítio (relacionado ao tipo da *seed*), conteúdo de bases AU ao redor do sítio predito e contribuição de posição em relação a outros sítios preditos.

Por fim, para o tipo da região *seed* TargetScan utiliza '`1a`' para `7mer-1a`, '`m8`' para `7mer-m8` e `8mer`.

- **Predição de sítios alvos de miRNAs em regiões de SNPs ao invés de simples detecção de SNPs em potenciais regiões seed:**

  Para cada SNP, uma janela centrada na sua posição, por padrão de 25 nt para ambos os lados, define a sequência que irá ser analisada pelas ferramentas de predição de alvos de miRNAs. Atualmente, o usuário pode escolher por uma ou as duas ferramentas disponíveis: miRanda e TargetScan.

- **Análise de todos os alelos, incluindo a combinação deles quando mais de um SNP está próximo um do outro:**

  Para cada SNP, uma sequência para cada alelo é analisada pelas ferramentas de predição de alvos de miRNAs e SIMTar reporta o efeito desse alelo sobre o sítio alvo predito. Quando mais de um SNP está presente na janela, sequências contendo todas as combinações alélicas de todos os SNPs são geradas e analisadas.

- **SNPs localizados em RNA ou DNA:**

  O usuário pode restringir a análise de SNPs localizados apenas em sequências de RNA, incluindo mRNA e RNAs não codificadores. Neste caso, SIMTar constrói a janela ao redor de cada SNP considerando somente regiões de éxons, permitindo a análise sobre todas as possíveis variantes de *splicing* dessa região. Se por outro lado o usuário tiver interesse em investigar miRNAs se ligando em DNA, a janela de predição circundante do SNP irá ser construída sobre a sequência genômica e não somente em regiões de éxons.

- **Independência da espécie:**

  Embora os parâmetros padrões sejam para analisar SNPs humanos, SIMTar pode ser aplicado para qualquer espécie, desde que o usuário configure os locais dos arquivos específicos na entrada da ferramenta. Um programa foi também desenvolvido para auxiliar o usuário a obter um desses arquivos, o de mapeamento genômico de RNAs. A lista desses arquivos assim como o uso de programas auxiliares estão descritos no tutorial do SIMTar.

4.2 Estrutura do SIMTar

SIMTar está estruturado como um *workflow* composto por módulos, dos quais alguns são específicos para as ferramentas de predição de alvos de miRNAs. A Figura 4.1 mostra como estes módulos estão estruturados e abaixo são descritas as atividades executadas por cada um. A única entrada obrigatória é um arquivo contendo a lista dos SNPs a serem analisados.

a) **Obtenção das coordenadas genômicas dos SNPs:**

A primeira etapa é a obtenção das coordenadas genômicas de todos os SNPs dados pelo usuário (Figura 4.1a). Cada SNP da lista do usuário deve estar representado pelo seu identificador da sequência de referência (*reference sequence* – rs). Assim, para cada SNP o módulo `generateSNPGFF` recupera a partir do dbSNP: o nome do cromossomo no qual está localizado, a sua posição no cromossomo e variação alélica.

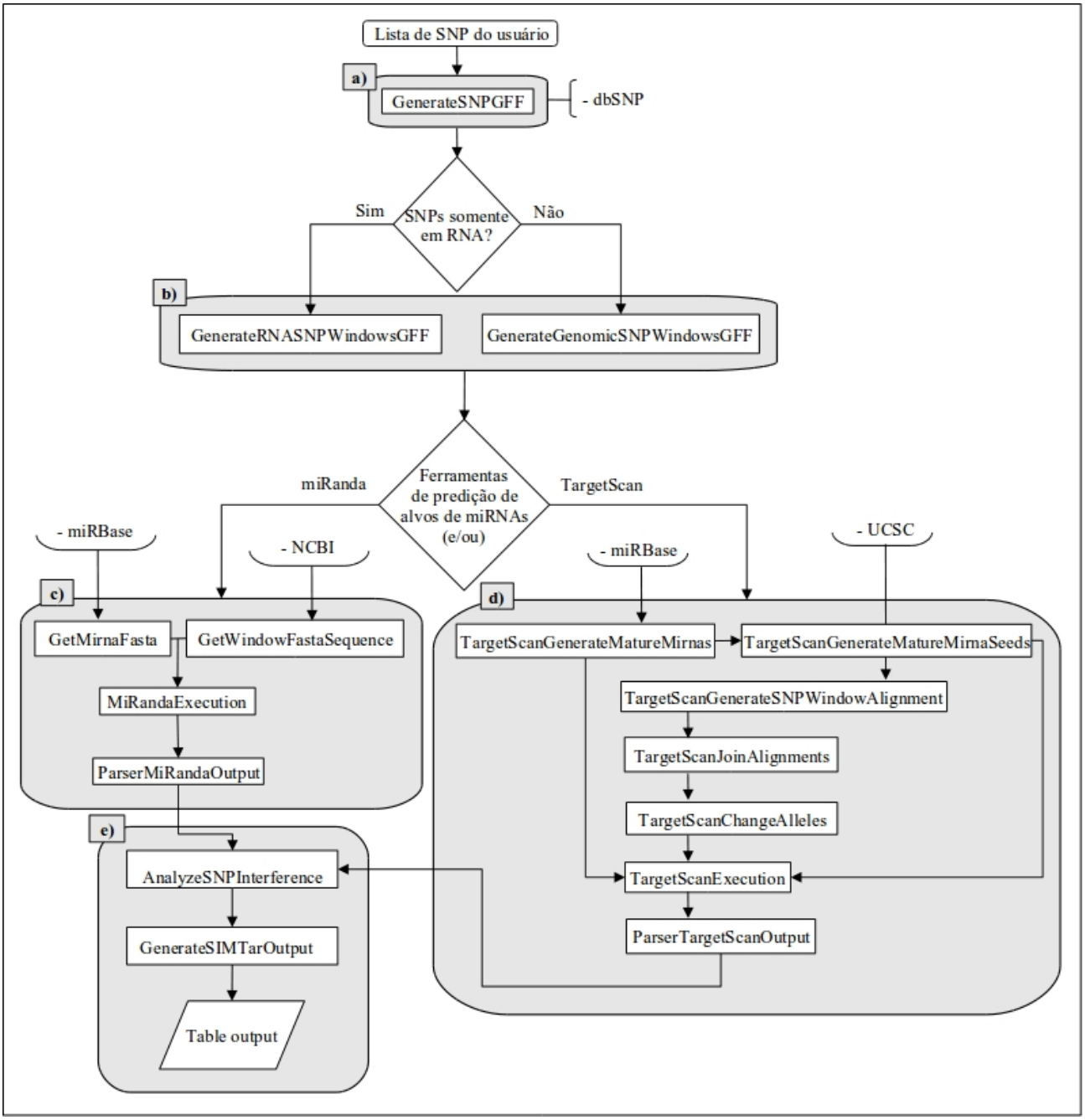

Figura 4.1: *Estrutura da ferramenta SIMTar.*

Essas informações são armazenadas em arquivos no formato GFF (*Generic Feature Format*) (`http://www.sanger.ac.uk/resources/software/gff/spec.html`).

b) **Construção das coordenadas das janelas a serem analisadas:**

A próxima etapa é a construção das coordenadas das janelas sobre as sequências alvos a serem analisadas pelos preditores de sítios alvos de miRNAs (Figura 4.1b). Essa janela é definida por um número de bases à esquerda e à direita do SNP (parâmetro *length* com valor padrão igual a 25 para ambos os lados). Se o usuário escolher analisar SNPs em potenciais sítios no DNA, o módulo `GenerateGenomicSNPWindowsGFF` define as janelas que circundam a posição do SNP nas sequências genômicas das duas fitas, independente da presença ou não de éxons na região.

Se o usuário escolher analisar SNPs localizados somente em RNAs, o módulo `GenerateRNA-SNPWindowsGFF` define as coordenadas das janelas considerando somente regiões de éxon. Isso significa que haverá uma janela de análise para cada isoforma de RNA contendo aquele determinado SNP[1]. Isso é feito baseando-se em um arquivo GFF contendo as coordenadas genômicas de todos os genes da espécie que está sendo analisada. Como padrão é utilizado um arquivo contendo as coordenadas dos genes humanos baseados na montagem hg19 GRCh37 do genoma, disponibilizado com o programa SIMTar. No entanto, caso o usuário queira utilizar o SIMTar para análise de SNPs em outras espécies, SIMTar também disponibiliza um programa auxiliar para gerar as coordenadas genômicas de cada gene da espécie de referência. Este programa, desenvolvido neste projeto, utiliza a ferramenta BLAT [Kent, 2002] para realizar o mapeamento das sequências dos RNAs a partir das sequências FASTA dos RNAs e das sequências genômicas da espécie de referência.

Para ambos os casos, análise de alvos em DNA ou RNA, os programas utilizados na construção das coordenadas das janelas de predição também verificam a existência de mais de um SNP localizado dentro de uma determinada janela, para posterior combinação de alelos. Todas essas informações são armazenadas em um novo arquivo em formato GFF.

c) **MiRanda - preparação dos dados e execução:**

[1]Diferentemente da análise de DNA, a análise de RNA só analisa a fita que contém um transcrito.

Na etapa representada pela Figura 4.1c ocorre o processamento dos dados de entrada para a ferramenta miRanda bem como a sua execução. MiRanda utiliza o formato FASTA como entrada para ambas as sequências, do alvo e do miRNA. O módulo `GetWindowFastaSequence` utiliza o arquivo GFF com as coordenadas das janelas vindo do módulo anterior e um arquivo FASTA para obtenção de cada sequência alvo. No caso de análise de SNPs genômicos, o arquivo FASTA deve conter as sequências genômicas da espécie sendo analisada[1]. No caso de análise de SNPs localizados em RNAs, recomenda-se que o arquivo FASTA seja composto por sequências RefSeq de RNAs[2] [Pruitt *et al.*, 2009]. O módulo `GetMirnaFasta` recupera todas as sequências FASTA de miRNAs maduros da espécie sendo analisada e que estão disponíveis no banco de dados miRBase ou somente um subgrupo delas, caso o usuário defina que deseja analisar apenas uma lista de miRNAs (lista passada por parâmetro). Assim, as sequências alvos geradas e as sequências dos miRNAs são utilizadas na chamada do programa miRanda (`MiRandaExecution`). O resultado do miRanda é processado pelo módulo `ParserMiRandaOutput`, com o objetivo de selecionar apenas aqueles resultados com valores de escore e energia-livre acima de um certo limiar, definidos pelo usuário. Como padrão são utilizados os valores '90' para escore e '-17' para energia-livre, valores estes definidos pelos autores do miRanda para predição de sítios de microRNAs em sequências humanas [John *et al.*, 2004].

d) **TargetScan – preparação dos dados e execução:**

Esta etapa está representada na Figura 4.1d. Os módulos `TargetScanGenerateMatureMirnas` e `TargetScanGenerateMatureMirnasSeeds` geram os arquivos de miRNAs nos formatos exigidos pela ferramenta TargetScan, os quais incluem informações de *seed* e de taxonomia da espécie do miRNA. TargetScan recebe como entrada da sequência alvo um alinhamento múltiplo de sequências de várias espécies, que então permite analisar a conservação evolutiva do sítio alvo. Dessa maneira, SIMTar deve gerar um alinhamento múltiplo para cada sequência alvo a ser analisada. Primeiro SIMTar gera uma sequência de alinhamento múltiplo para cada janela de um SNP e não ainda para cada alelo. Para isso, o módulo `TargetScanGenerateSNPWindowAlignment`

[1]Como padrão, SIMTar disponibiliza e utiliza a versão hg19 GRCh37 do genoma humano.
[2]Como padrão, SIMTar disponibiliza e utiliza sequências de RNAs humanos obtidos do RefSeq versão 46.

lê um arquivo de alinhamento múltiplo da espécie sendo analisada e então recupera blocos de alinhamentos múltiplos para cada janela de um SNP. O arquivo de alinhamento múltiplo utilizado deve estar em formato MAF (*Multiple Alignment Format*). Em seguida o módulo TargetScanJoinAlignment tenta fazer a concatenação dos blocos de alinhamentos recuperados no alinhamento múltiplo mantendo um número mínimo de espécies que devem ser mantidas no alinhamento final[1], e pelo número máximo de lacunas (*gaps*) que podem ser inseridas entre cada um desses segmentos[2]. Finalmente o módulo `TargetScanChangeAlleles` gera um novo alinhamento para cada combinação alélica. Todos os programas de execução de TargetScan são executados pelo módulo `TargetScanExecution` e os arquivos de saída da ferramenta são processados pelo módulo `ParserTargetScanOutput`, o qual gera um único arquivo com os resultados da predição por TargetScan relevantes para o SIMTar.

e) **Identificação de interferência do SNP:**

A Figura 4.1e mostra a etapa de análise de interferência de SNPs em sítios alvos de miRNAs e de geração dos resultados. O módulo `AnalyseSNPInterference` lê os resultados gerados da(s) ferramenta(s) de predição de alvos de miRNAs e analisa quais os SNPs que, em decorrência de sua variação alélica, estão interferindo no pareamento dos miRNAs com seus sítios alvos. Para execução de SIMTar no DNA, a interferência é analisada para cada par de SNP-miRNA, enquanto que em execuções sobre regiões de RNA, a interferência é analisada para cada trio SNP-miRNA-RefSeq. O efeito de interferência pode ocorrer de duas formas: 1) *Creation/Disruption*, quando o SNP possui pelo menos dois alelos onde o sítio é predito em um alelo mas não no outro e

2) *Binding affected*, quando todos os alelos do SNP possuem sítios preditos para aquele miRNA mas a variação alélica altera as propriedades pareamento. Neste segundo caso, essa alteração é analisada em termos dos valores de escore e energia-livre do miRanda e pelo tipo da *seed* dado pelo TargetScan. Se nenhum sítio alvo é predito para um determinado SNP, este SNP é omitido do arquivo final.

[1]O número mínimo de espécies que devem estar presentes no alinhamento final é outro parâmetro de SIMTar, com valor padrão igual a 2.

[2]Parâmetro definido pelo usuário com valor padrão igual a 0.

O módulo `GenerateSIMTarOutput` lê este arquivo com os dados de interferência obtidos na etapa anterior, faz a quantificação dos resultados analisados, obtém o contexto genômico dos SNPs e gera a saída da ferramenta SIMTar.

A quantificação dos resultados gera três arquivos, um com a lista dos SNPs do usuário onde não foram identificadas interferências, um segundo arquivo contendo os SNPs ordenados decrescentemente pelo número de sítios interferidos e um terceiro arquivo contendo os SNPs que estão envolvidos na criação/ruptura (*Creation*/*Disruption*) de sítios alvos de miRNAs juntamente com o número de sítios envolvidos. Além destes arquivos adicionais, é feita uma quantificação do número de sítios envolvidos em cada efeito e do número de miRNAs preditos, que são então impressos no arquivo com os dados dos resultados.

A obtenção do contexto genômico de cada SNP é feita utilizando como referência a tabela refGene do *UCSC Genome Bioinformatics Site*, tomando como referência cada transcrito próximo ao SNP. Em análise de SNPs no DNA, este módulo obtém o contexto do SNP para as duas fitas genômicas.

O SNP pode estar localizado no interior de um gene, próximo a um gene ou em uma região intergênica. No caso de estar localizado no interior de um gene, é obtido o nome deste gene e a região em que o SNP se encontra (CDS, íntron, 5'UTR, 3'UTR). Um SNP é considerado próximo a um gene se a distância entre ele e o gene é menor ou igual a 2.000 pares de base (pb) a montante ou 500 pb a jusante de um gene (valores padrões dos parâmetros do SIMTar). Neste caso, o contexto deste SNP é anotado com essa distância seguido da localização, como por exemplo '2000B-upstream'. Se a distância exceder um desses valores o contexto genônico é anotado como intergênico. No caso de análise de SNPs em RNA, mesmo sabendo que ele está localizado em éxon por conta da construção da janela de predição, este módulo obtém o seu contexto genômico, informando se a região é codificadora (CDS), 5' ou 3'UTR.

Por fim, é gerado um arquivo tabular final com os resultados de SIMTar. A figura mostrando a saída da ferramenta bem como a descrição dos campos utilizados estão na seção de resultados (capítulo 5).

4.3 Implementação

SIMTar foi desenvolvido em linguagem Perl como um *workflow* científico, definido para ser executado usando o gerenciador de *workflows* Bpipe [Sadedin *et al.*, 2012]. Bpipe foi escolhido como gerenciador de execução por permitir que tarefas em paralelo pudessem ocorrer ao mesmo tempo, como é o caso da execução das ferramentas de predição de sítios alvos de miRNAs, que independem uma da outra. A parametrização do SIMTar é inteiramente feita em um arquivo de configuração, baseado na ferramenta eGene [Durham *et al.*, 2005], através do qual o usuário pode configurar todos os parâmetros necessários.

A lista de SNPs é o único parâmetro obrigatório, todos os outros parâmetros possuem um valor padrão que, embora pensados para análise de SNPs humanos, podem ser configurados para a análise de SNPs de outras espécies. Uma completa descrição sobre todos os parâmetros, bem como orientações para análise de SNPs de outras espécies que não a humana estão disponíveis no tutorial da ferramenta SIMTar. SIMTar é uma ferramenta de código aberto e pode ser obtida em `http://www.vision.ime.usp.br/~amanda/software/simtar/`.

4.4 Aplicação em esquizofrenia

A fim de exemplificar o uso do programa SIMTar, o mesmo foi aplicado para a identificação de SNPs associados a esquizofrenia que pudessem estar interferindo em alvos de microRNAs.

4.4.1 Seleção de SNPs associados a esquizofrenia

Primeiramente, foi obtida uma lista de 38 genes candidatos do banco de genes de esquizofrenia (*Schizophrenia Gene Resource* - SZGR) [Jia *et al.*, 2010]. Os genes contidos nesta lista são provenientes de dados de literatura e de meta-análise de estudos de associação em esquizofrenia. Posteriormente, foi realizada a busca por SNPs também associados à doença. Para isso, utilizou-se o banco de dados SzGene [Allen *et al.*, 2008] que possui SNPs associados à esquizofrenia e agrupados por genes. Assim, para cada gene da lista inicial foram obtidos todos os SNPs nele localizados. Entretanto,

alguns genes não tinham nenhum SNP associado à esquizofrenia e por isso foram excluídos da lista final. Ao todo foram obtidos 114 SNPs para 23 genes (Apêndice A).

4.4.2 Execução do SIMTar

A lista de 114 SNPs serviu como arquivo de entrada para a ferramenta SIMTar. O parâmetro de tipo de molécula analisada foi selecionado para DNA (análise genômica). Os demais parâmetros do arquivo de configuração do SIMTar foram deixados com os valores padrões. Estes valores incluem: espécie humana como referência, predição de alvos para todos os miRNAs maduros da espécie de referência, tamanho da janela de predição utilizada com 25 nt para ambos os lados do SNP, ferramentas de predição de sítios alvos de miRNAs configurados para miRanda e TargetScan, entre outros.

4.4.3 Seleção de miRNAs associados à esquizofrenia para análise dos resultados do estudo de caso

Com o objetivo de enriquecer a análise dos resultados obtidos pelo SIMTar, foi obtida uma lista de 18 miRNAs também associados à esquizofrenia a partir do mesmo banco de dados onde foi obtida a lista de genes candidatos (SZGR). Esta lista é composta pelos seguintes miRNAs: miR-106b, miR-20b, miR-212, miR-181b, miR-29a, miR-29b, miR-29c, miR-195, miR-92b, miR-9, let-7g, miR-7, miR-206, miR-24, miR-30d, miR-30b, miR-198 e miR-26b.

Capítulo 5

Resultados e Discussão

5.1 SIMTar

O principal resultado deste trabalho é a ferramenta computacional SIMTar. Ela cumpri com os objetivos esperados de superar as limitações dos atuais recursos computacionais da área e permite que SNPs antes não estudados sejam agora analisados, como é o caso de SNPs em regiões promotora, CDS, intrônica e 5'UTR, não se limitando apenas à região 3'UTR.

Identificação do alelo ou combinação alélica que interfere em sítios alvos de miRNAs

A ferramenta SIMTar não somente identifica qual SNP está localizado no sítio alvo do miRNA, como também indica a variação alélica ou combinação alélica de SNPs vizinhos que estão interferindo no sítio alvo. SIMTar mostra ainda o efeito predito para cada alelo, dizendo se o alelo cria um novo sítio, rompe um sítio pré-existente ou afeta a estabilidade do pareamento em função de sua variação.

Identificar o alelo do SNP que está interferindo neste pareamento contribui no entendimento sobre como estes polimorfismos podem estar agindo no processo de regulação do genoma, uma vez que centenas deles estão associados a fenótipos de doença, como a esquizofrenia, e apresentam frequências alélicas diferentes entre os casos saudáveis e doentes.

Através dos miRNAs, estes SNPs podem estar agindo tanto na regulação transcricional dos genes

envolvidos, podendo ativar ou silenciar o processo, como podem também agir ativando ou reprimindo o processo de tradução.

Abordagem evolutiva da predição de sítios alvos

Estudos experimentais e de predição computacional realizados inicialmente com miRNAs revelaram a sua forte participação em eventos de regulação negativa da tradução através da sua ligação em sítios alvos localizados na região 3'UTR, que é uma região altamente conservada entre as espécies. Além disso, centenas de miRNAs também foram identificados conservados evolutivamente.

Portanto, para sequências conhecidamente conservadas, a utilização de uma ferramenta específica contribui para o aumento da especificidade da predição. Dessa forma, SIMTar utiliza a TargetScan para disponibilizar ao usuário esta abordagem de predição de sítios de alvos de miRNAs.

Essa conservação entre os genomas de várias espécies deve estar assegurando que processos biológicos importantes sejam regulados efetivamente. No entanto têm-se descoberto cada vez mais novos miRNAs espécie-específicos, mostrando a continuidade da evolução dos miRNAs e que além dos miRNAs conservados existe um grande número de miRNAs emergentes [Berezikov *et al.*, 2006]. Dessa maneira, SIMTar também disponibiliza a ferramenta miRanda de predição de sítios alvos de miRNAs, que permite analisar as regiões não conservadas.

Espécies de aplicação

A ferramenta foi inicialmente aplicada na espécie humana e por isso utiliza e disponibiliza os arquivos específicos para esta espécie, como os arquivos de alinhamentos múltiplos do genoma, sequências genômicas, coordenadas de SNPs, sequências maduras dos miRNAs e coordenadas dos genes. Os valores padrões de outros parâmetros do arquivo de configuração foram também definidos para com esta espécie como alvo.

Embora o seu desenvolvimento tenha sido realizado utilizando dados da espécie humana, SIMTar pode ser aplicado em SNPs de outras espécies desde que o usuário disponibilize os arquivos necessários. A lista destes arquivos estão descritos no tutorial da ferramenta, bem como quais são específicos das ferramentas de predição. Assim, para aplicação em uma nova espécie, SIMTar adicionalmente

disponibiliza alguns programas e um tutorial de execução para que o usuário possa gerar o arquivo com coordenadas dos genes para qualquer espécie. Este programas adicionais apenas precisam ser utilizados para executar o SIMTar em outra espécie que não a humana e com a opção de analisar SNPs localizados em região de RNAs.

É importante mencionar que somente faz sentido utilizar o SIMTar em espécies que respeitam o padrão de pareamento imperfeito ocorrido em animais, pois os algoritmos implementados nas ferramentas de predição utilizadas são específicos para esse padrão.

Por fim, SIMTar é a primeira ferramenta de código aberto capaz de aplicar tais características na identificação de SNPs interferindo em sítios alvos de miRNAs. A Tabela 5.1 mostra resumidamente as principais diferenças observadas entre SIMTar e os recursos computacionais disponíveis para o estudo de SNPs em sítios alvos de miRNAs.

5.1.1 Arquivos de saída da ferramenta SIMTar

Cada execução de SIMTar pode ser realizada para apenas um SNP ou uma lista contendo vários SNPs, onde ao final de cada execução são gerados quatro arquivos: um contendo os resultados positivos, que correspondem aos dados dos SNPs que estão interferindo em sítios alvos de miRNAs em função de sua variação alélica e três arquivos de quantificação dos resultados preditos (descritos a seguir). Todos estes arquivos são gerados ao final de cada execução da ferramenta, independentemente se o usuário escolheu analisar SNPs localizados em DNA ou RNA.

Tabela 5.1: *Relação de recursos computacionais existentes para análise de SNPs em sítios alvos de miRNAs.*

Nome do recurso	Tipo de tecnologia	Região estudada	Posição do SNP	Estratégia de predição	Referência
SIMTar	programa *standalone*	DNA, RNA	todo o sítio alvo	miRanda TargetScan	Artigo em fase final de preparação
MicroSNiPer	aplicação web	3'UTR	seed	MicroSNiPer	Barenboim *et al.* [2010]
PolymiRTS	banco de dados	3'UTR	seed	TargetScan	Ziebarth *et al.* [2012]
MirSNP	banco de dados	3'UTR	seed	MiRanda	Liu *et al.* [2012]
miRNASNP	banco de dados	3'UTR	seed	MiRanda, TargetScan	Gong *et al.* [2012]
MirR*dSNP*	banco de dados	3'UTR	seed	PicTar, TargetScan	Bruno *et al.* [2012]

Quantificação dos dados analisados

Os três arquivos de quantificação gerados ao final de cada execução correspondem a: um arquivo contendo o número de interferências preditas para cada par SNP-miRNA (em análise de DNA) ou para cada trio SNP-miRNA-RefSeq (em análise de RNA), um segundo arquivo contendo os nomes dos SNPs que foram preditos com pelo menos um sítio de efeito *Creation/Disruption* juntamente com o número correspondente de sítios alvos interferidos e um último arquivo contendo os nomes dos SNPs do usuário que não foram identificados pelo SIMTar.

Os resultados positivos do SIMTar são impressos em um arquivo tabular (Figura 5.1). Também neste arquivo são impressos na porção superior: o número de SNPs passados pelo usuário, o número de SNPs com resultados positivos gerados pela ferramenta, o número de miRNAs com sítios interferidos e o número de sítios interferidos para cada efeito (*Creation/Disruption* ou *Binding affected*).

Forma de ordenação dos resultados

Nas colunas abaixo da primeira linha de informações gerais estão inseridos os campos referentes ao alelo e miRNA predito. Os resultados inseridos em cada campo são ordenados de forma diferente para análise de SNPs localizados no DNA e no RNA. No primeiro caso, a chave de ordenação é o SNP seguido do miRNA, no segundo caso é o SNP seguido do nome do transcrito (RefSeq) e depois o miRNA.

Conteúdo do arquivo de predição gerado

a) Dados dos SNPs:

A primeira coluna é nomeada como 'RefSNP ID' e informa o identificador do SNP dentro do SIMTar.

A segunda coluna é nomeada 'Chr' então preenchida com o nome do cromossomo no qual o SNP está localizado. A terceira coluna é chamada de 'Chr Pos', preenchida com a posição do SNP no cromossomo.

SIMTar output: Total number of unique SNPs given by user = 114; total number of SNPs predicted affecting miRNA target sites = 114; Creation/Disruption effects = 1854; Binding affected = 723; total number of miRNAs predicted = 1503.

RefSNP ID	Chr	Chr Pos	Genomic context	Gene symbol	RefSeq	Strand	Alleles	Interacting miRNA	miRanda-Effect	TargetScan-Effect	Pairing attributes						
rs2391191	chr13	106119446	UTR-5	DAOA	NM_001161814	+	A/G	miR-10b-3p		TargetScan-Binding affected	TargetScan | pct [0.00919]	TargetScan | allele[A] | contextScore [too_close]	TargetScan | allele[A] | siteType[1a]	TargetScan | allele[A] | outside the seed	TargetScan | allele[G] | contextScore [too_close]	TargetScan | allele[G] | siteType [8mer]	TargetScan | allele[G] | outside the seed
rs1000731	chr1	231963491	intron	TSNAX-DISC1	NR_028393	+	A/G	miR-371b-3p	miRanda-Creation/ Disruption	TargetScan-Creation/ Disruption	miRanda | allele[G] | score [154.00]	miRanda | allele[G] | energy [-19.15]	TargetScan | pct[0]	TargetScan | allele[G] | contextScore [m8]	TargetScan | allele[G] | siteType [m8]	TargetScan | allele[G] | inside the seed	
rs1000731	chr1	231963491	intron	DISC1	NM_018662	+	A/G	miR-371b-3p	miRanda-Creation/ Disruption	TargetScan-Creation/ Disruption	miRanda | allele[G] | score [154.00]	miRanda | allele[G] | energy [-19.15]	TargetScan | pct[0]	TargetScan | allele[G] | contextScore [m8]	TargetScan | allele[G] | siteType [m8]	TargetScan | allele[G] | inside the seed	
rs1000731	chr1	231963491	intron	DISC1	NM_001164547	+	A/G	miR-371b-3p	miRanda-Creation/ Disruption	TargetScan-Creation/ Disruption	miRanda | allele[G] | score [154.00]	miRanda | allele[G] | energy [-19.15]	TargetScan | pct[0]	TargetScan | allele[G] | contextScore [m8]	TargetScan | allele[G] | siteType [m8]	TargetScan | allele[G] | inside the seed	
rs2619538	chr6	15665209	1920B-upstream	DTNBP1	.	-	T/A	miR-4761-5p	miRanda-Creation/ Disruption		miRanda | allele[T] | score [144.00]	miRanda | allele[T] | energy [-20.77]					
rs2619538	chr6	15665209	1920B-upstream	DTNBP1	.	-	T/A	let-7g-3p	miRanda-Creation/ Disruption		miRanda | allele[T] | score [140.00]	miRanda | allele[T] | energy [-18.64]					
rs2492367	chr1	231906589	UTR-5	TSNAX-DISC1	NR_028394	+	C/T	miR-9-5p	miRanda-Creation/ Disruption		miRanda | allele[C] | score [154.00]	miRanda | allele[C] | energy [-20.07]					
rs2492367	chr1	231906589	UTR-5	TSNAX-DISC1	NR_028395	+	C/T	miR-9-5p	miRanda-Creation/		miRanda | allele[C] | score	miRanda | allele[C] | energy					

Figura 5.1: *Tabela gerada pela ferramenta SIMTar.*

A quarta (*Genomic context*), quinta (*Gene symbol*) e sexta (*RefSeq*) colunas estão relacionadas às informações do gene e transcrito dos SNPs analisados[1]. *Genomic context* contém o contexto genômico do SNP em questão, que pode ser intergênico, CDS, íntron, 5'UTR, 3'UTR, *upstream* ou *downstream* de um gene, sendo nestes últimos mostrados também a distância que eles estão do gene. *Gene symbol* é preenchida com o nome do gene e *RefSeq* é preenchida com o identificador RefSeq do transcrito, a menos que o SNP seja intergênico.

A sétima coluna é nomeada 'Strand' e mostra a fita genômica analisada.

A oitava coluna é chamada 'Alleles'e mostra a variação alélica do SNP na fita genômica em questão.

b) Dados do pareamento com o(s) miRNA(s):

A partir da nona coluna são impressas as informações estritamente relacionadas ao pareamento do SNP com um miRNA específico, descrito na coluna 'Interacting miRNA'.

A coluna 'miRanda-Effect' irá ser preenchida apenas se o usuário escolheu utilizar esta ferramenta na predição de alvos de miRNAs e se foram preditos resultados positivos para o respectivo miRNA::alelo. Neste caso os campos dessa coluna mostrarão o nome da ferramenta seguido do efeito identificado sobre o sítio alvo, podendo aparecer das seguintes formas: miRanda-*Creation*/-*Disruption* ou miRanda-*Binding affected*. O mesmo ocorre na coluna 'TargetScan-Effect' que apresenta as seguintes opções: TargetScan-*Creation*/*Disruption* e TargetScan-*Binding affected*.

Através dos atributos do pareamento fornecidos por cada ferramenta de predição de sítios é possível identificar o alelo com maior afinidade sobre a alteração da estabilidade do pareamento. Na região da tabela nomeada 'Pairing attributes' estão inseridas as colunas de informações correspondentes aos pareamentos preditos. Assim, para cada ferramenta, são mostrados os seguintes campos:

[1]Essas três colunas estão também presentes na análise de DNA pelo fato do contexto genômico ser dependente do RNA de referência.

Em miRanda:

- miRanda|alelo|valor do escore;
- miRanda|alelo|valor da energia-livre.

Em TargetScan:

- TargetScan|pct;
- TargetScan|alelo|context-score;
- TargetScan|alelo|tipo da *seed*;
- TargetScan|alelo|posição do SNP em relação à *seed*.

Destas colunas, apenas a 'TargetScan|pct' é impressa para cada par miRNA-SNP e não para cada miRNA-alelo, isso porque o valor do pct é atribuído à região alvo conservada e é obtido independe da variação alélica do SNP. Todas as outras colunas mencionadas são impressas para cada par miRNA-alelo, possibilitando a comparação da estabilidade do pareamento observada para cada alelo.

A coluna 'TargetScan|alelo|posição do SNP em relação à *seed*' foi criada para mostrar que também a alteração alélica ocorrida fora da região *seed* pode interferir no tipo de *seed* predita. Este contexto do SNP somado à informação da coluna 'TargetScan|alelo|tipo da *seed*' permite a visualização clara sobre este impacto. Um exemplo disso pode ser visualizado na primeira linha da Figura 5.1, onde o SNP rs2391191 é predito interferindo o sítio alvo do miRNA miR-10b-3p, segundo a execução da ferramenta TargetScan. Neste exemplo, ambos os alelos do SNP (A/G) foram preditos fora da região *seed* do miRNA, no entanto o alelo G foi predito com a *seed* do tipo '`8mer`', portanto com um pareamento altamente complementar e que consequentemente implica em uma alta efetividade da regulação do alvo. Já o alelo A foi predito com uma *seed* do tipo '`1a`', que possui uma menor eficiência (Bartel [2009]; Hammell *et al.* [2008]).

5.2 Aplicação do SIMTar em esquizofrenia

5.2.1 SNPs associados a esquizofrenia interferindo em sítios de miRNAs

SIMTar identificou que todos os 114 SNPs da lista inicial estão interferindo em sítios alvos de miRNAs, os quais inclusive estão envolvidos na criação/ruptura de pelo menos um sítio alvo de miRNA. Como o SIMTar foi executado com o parâmetro de análise de DNA, para cada SNP as duas fitas genômicas foram analisadas. Ao todo, foram preditos 2.553 sítios de miRNAs interferidos por esses SNPs, sendo estes sítios de 1.503 miRNAs distintos (58,3% de todos os miRNAs maduros humanos). Destes 2.553 sítios, 1854 sítios sofrem efeito de criação/ruptura de sítio pela variação alélica, e 723 sítios sofrem o efeito de alteração da força de ligação[1].

Como cada fita pode conter isoformas de *splicing* de um mesmo gene, cada SNP pode possuir um número distinto de contextos genômicos, dependendo dos transcritos que se colocalizam com o SNP. Para analisar a distribuição dos contextos genômicos desses SNPs, foram somados todos os contextos de todos os SNPs. De um total de 672 contextos, apenas 18 (2,7%) estão localizados em região 3'UTR (Figura 5.2), que correspondem a apenas 8 SNPs dos 114 SNPs iniciais (7%). Desta forma, SIMTar identificou uma grande quantidade de interferências de SNPs que não seriam identificados por estratégias tradicionais.

Os SNPs utilizados nesta aplicação são associados à esquizofrenia mas ainda não possuem um claro papel biológico e por isso ainda precisam ser investigados. A maior porcentagem deles está localizada em íntrons (58,4% dos contextos, representados por 65 SNPs).

Aproximadamente 60 a 70% do genoma humano é transcrito em sequências de RNAs (Carninci *et al.* [2005]; Frith *et al.* [2005]). Nesta porção de transcritos estão inclusos os íntrons, que são regiões de importantes funções biológicas já confirmadas para o genoma, diferente do que foi defendido até não muito tempo atrás onde os íntrons eram desconsiderados como parte funcional do genoma [Ohno, 1972]. Muitos íntrons são precursores de sequências de RNAs não codificadores de proteínas [Mattick & Makunin, 2006].

[1] Alguns sítios envolvendo SNPs tri- ou tetra-alélicos apresentam os dois efeitos.

Entre as moléculas de RNAs não codificadores que podem estar inseridas nas regiões de íntron, tem-se por exemplo as longas sequências de RNAs não codificadores, que podem atuar na regulação do genoma através do pareamentos com a fita sense de mRNAs e assim regular genes codificadores de proteínas, como podem também se ligar a sítios de proteínas específicas e consequentemente regular a atividade proteica, ou ainda atuar como sequência precursora de pequenos RNAs não codificadores de proteínas, como os próprios miRNAs (Mattick & Makunin [2006]; Wilusz *et al.* [2009]). Interessantemente, o miRNA let-7 já foi reportado em processo de autoregulação da biogênese através do pareamento em sequências de RNAs não codificadores [Zisoulis *et al.*, 2012].

Por outro lado, pode ser que as posições dos SNPs identificados em íntrons correspondam a sítios de *splicing* e o pareamento com o miRNA esteja afetando este processo e consequentemente interferindo na síntese de isoformas dos genes.

Apesar dos SNPs utilizados como entrada terem sido obtidos a partir de 23 genes envolvidos em esquizofrenia, o SIMTar identificou que esses SNPs estão localizados também em mais 11 RNAs não codificadores, interferindo em sítios de miRNAs nessas sequências. Isso foi possível pelo fato do SIMTar ter sido executado com o parâmetro de análise em DNA, que avalia as duas fitas onde se encontra cada SNP.

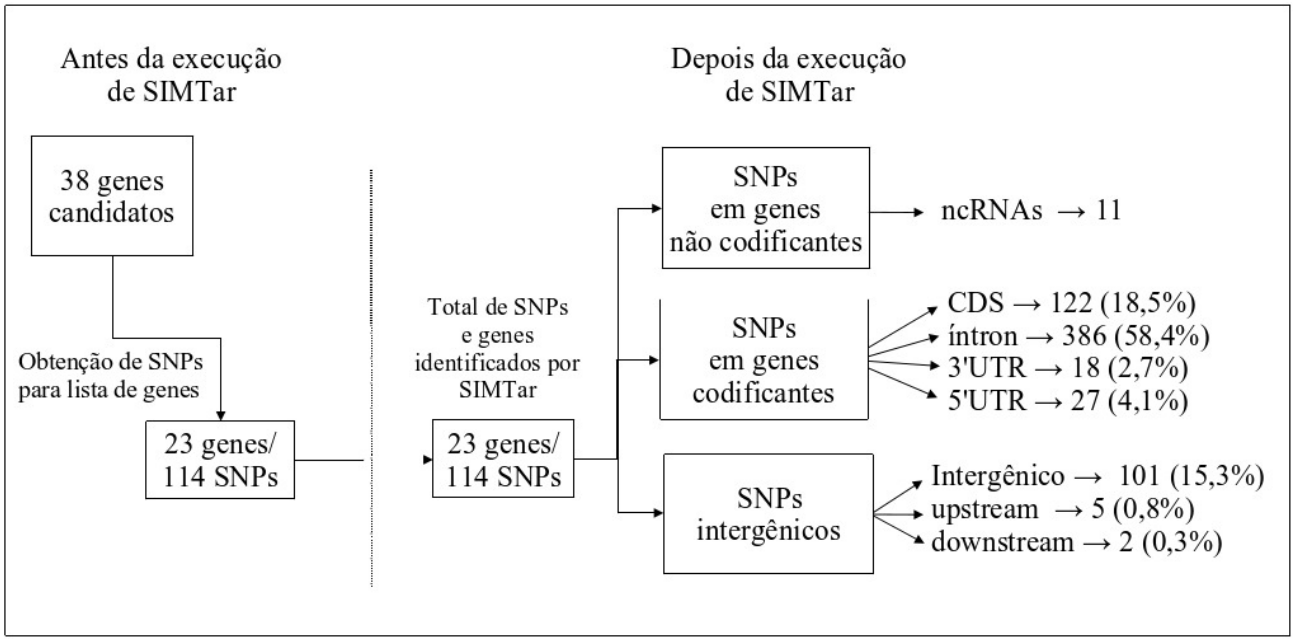

Figura 5.2: *Contexto genômico dos SNPs utilizados na aplicação de SIMTar.*

Dos resultados totais foram preditos, em média, 22 sítios alvos interferidos por SNP, cada SNP interferindo em 1 a 72 sítios (Figura 5.3). O resultado completo dessa aplicação está disponível em `http://www.vision.ime.usp.br/~amanda/software/simtar/applications`.

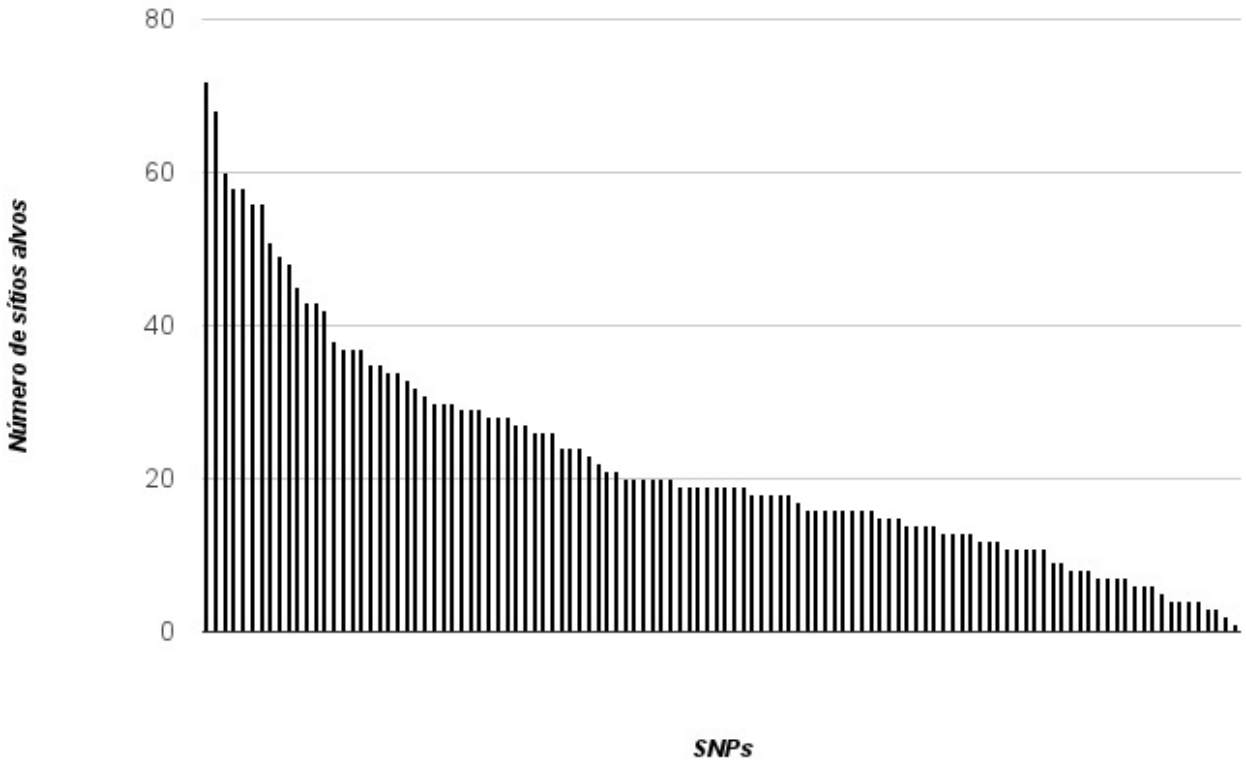

Figura 5.3: *Relação de sítios alvos por SNP associado à esquizofrenia. Os resultados de SIMTar para os dados analisados mostram que um SNP está afetando em média 22 sítios alvos de microRNAs.*

5.2.2 MicroRNAs associados à esquizofrenia com sítios interferidos por SNPs

Como já mencionado, o SIMTar foi executado considerando todos os miRNAs maduros da espécie humana. Os SNPs analisados são provenientes de estudos de associação com esquizofrenia, no entanto, os miRNAs utilizados na predição, embora sejam da espécie de referência, não são oriundos de estudos específicos com este fenótipo. Dessa forma, objetivou-se investigar a existência de conexões dos resultados gerados pelo SIMTar com dados de miRNAs também associados ao fenótipo de referência. Para isso, dos 18 precursores de miRNAs associados a esquizofrenia (obtidos con-

Tabela 5.2: *Relação de miRNAs associados à esquizofrenia identificados por SIMTar. A primeira coluna mostra a lista de miRNAs já associados à esquizofrenia e a segunda coluna mostra aqueles que foram identificados por SIMTar. Os nomes dos miRNAs podem diferir entre as duas colunas devido à precisão da metodologia SIMTar, que analisa todas as sequências maduras de cada miRNA e mostra a sequência exata predita.*

MicroRNAs SZGR	MicroRNAs SIMTar
let-7g	let-7g-3p
miR-106b	miR-106b-5p
miR-181b	-
miR-195	miR-195-5p
miR-198	miR-198
miR-20b	-
miR-206	miR-206
miR-212	miR-212-5p
miR-24	miR-24-3p miR-24-1-5p miR-24-2-5p
miR-26b	miR-26b-3p
miR-29a	-
miR-29b	miR-29b-1-5p
miR-29c	-
miR-30b	miR-30b-3p
miR-30d	miR-30d-3p
miR-7	-
miR-9	miR-9-5p
miR-92b	miR-92b-5p

forme descrito na seção 4.4.3), SIMTar predisse interferência de sítios de 13 deles, representando 15 miRNAs maduros, que estão dentre os 1.503 miRNAs com sítios interferidos identificados inicialmente pelo SIMTar (Tabela 5.2). Estes 15 miRNAs possuem sítios interferidos por 24 dos SNPs de entrada.

Na literatura, geralmente os autores referenciam os miRNAs apenas indicando o nome da sua forma precursora, como é mostrado na primeira coluna da Tabela 5.2 para os miRNAs obtidos do banco de dados SZGR. Já a metodologia utilizada por SIMTar analisa cada sequência madura de um miRNA, tornando possível a identificação da sequência exata que esta sendo afetada pela variação alélica do SNP.

A Tabela 5.3 mostra a relação dos 15 miRNAs associados à esquizofrenia com os respectivos SNPs. A tabela está ordenada pelos identificadores dos SNPs, permitindo assim que sejam visualizados todos os miRNAs ligados a cada SNP. As três últimas colunas dessa tabela são mostradas adicionalmente, para informar se na região predita (posição do SNP) já existem transcritos conhecidos.

Tabela 5.3: *Relação de SNPs e miRNAs associados à esquizofrenia preditos por SIMTar e suas respectivas regiões genômicas.*

SNP	microRNA	Contexto genômico	Transcrito	Gene
rs1011313	miR-24-1-5p	intron	NM_032122	DTNBP1
rs1011313	miR-24-2-5p	intron	NM_183040	DTNBP1
rs1011313	miR-24-2-5p	intron	NR_036448	DTNBP1
rs1011313	miR-24-2-5p	intron	NM_032122	DTNBP1
rs1011313	miR-24-2-5p	intron	NM_183040	DTNBP1
rs1011313	miR-24-2-5p	intron	NR_036448	DTNBP1
rs1018381	miR-106b-3p	intergenic	.	.
rs1322783	miR-206	intron	NM_001012957	DISC1
rs1322783	miR-206	intron	NM_001012959	DISC1
rs1322783	miR-206	intron	NM_001164537	DISC1
rs1322783	miR-206	intron	NM_001164538	DISC1
rs1322783	miR-206	intron	NM_001164539	DISC1
rs1322783	miR-206	intron	NM_001164540	DISC1
rs1322783	miR-206	intron	NM_001164541	DISC1
rs1322783	miR-206	intron	NM_001164542	DISC1
rs1322783	miR-206	intron	NM_001164544	DISC1
rs1322783	miR-206	intron	NM_001164545	DISC1

Tabela 5.3 – Continuação

SNP	microRNA	Contexto genômico	Transcrito	Gene
rs1322783	miR-206	intron	NM_001164546	DISC1
rs1322783	miR-206	intron	NM_001164547	DISC1
rs1322783	miR-206	intron	NM_001164548	DISC1
rs1322783	miR-206	intron	NM_001164556	DISC1
rs1322783	miR-206	intron	NM_018662	DISC1
rs1322783	miR-206	intron	NR_028393	TSNAX-DISC1
rs1322783	miR-206	intron	NR_028394	TSNAX-DISC1
rs1322783	miR-206	intron	NR_028395	TSNAX-DISC1
rs1322783	miR-206	intron	NR_028396	TSNAX-DISC1
rs1322783	miR-206	intron	NR_028397	TSNAX-DISC1
rs1322783	miR-206	intron	NR_028398	TSNAX-DISC1
rs1327175	miR-92b-5p	intron	NM_025179	PLXNA2
rs155333	miR-212-5p	intergenic	.	.
rs165774	miR-195-5p	intergenic	.	.
rs1800497	miR-106b-5p	intergenic	.	.
rs1801028	miR-92b-5p	CDS	NM_000795	DRD2
rs1801028	miR-92b-5p	CDS	NM_016574	DRD2
rs1934909	miR-29b-1-5p	intergenic	.	.
rs1997679	miR-9-5p	intron	NM_032122	DTNBP1
rs1997679	miR-9-5p	intron	NM_183040	DTNBP1
rs1997679	miR-9-5p	intron	NR_036448	DTNBP1
rs2492367	miR-9-5p	CDS	NM_001012957	DISC1

Tabela 5.3 – Continuação

SNP	microRNA	Contexto genômico	Transcrito	Gene
rs2492367	miR-9-5p	CDS	NM_001012959	DISC1
rs2492367	miR-9-5p	CDS	NM_001164537	DISC1
rs2492367	miR-9-5p	CDS	NM_001164538	DISC1
rs2492367	miR-9-5p	CDS	NM_001164539	DISC1
rs2492367	miR-9-5p	intron	NM_001164540	DISC1
rs2492367	miR-9-5p	CDS	NM_001164541	DISC1
rs2492367	miR-9-5p	CDS	NM_001164542	DISC1
rs2492367	miR-9-5p	CDS	NM_001164544	DISC1
rs2492367	miR-9-5p	CDS	NM_001164545	DISC1
rs2492367	miR-9-5p	CDS	NM_001164546	DISC1
rs2492367	miR-9-5p	CDS	NM_001164547	DISC1
rs2492367	miR-9-5p	CDS	NM_001164548	DISC1
rs2492367	miR-9-5p	CDS	NM_001164549	DISC1
rs2492367	miR-9-5p	CDS	NM_001164556	DISC1
rs2492367	miR-9-5p	CDS	NM_018662	DISC1
rs2492367	miR-9-5p	UTR-5	NR_028393	TSNAX-DISC1
rs2492367	miR-9-5p	UTR-5	NR_028394	TSNAX-DISC1
rs2492367	miR-9-5p	UTR-5	NR_028395	TSNAX-DISC1
rs2492367	miR-9-5p	UTR-5	NR_028396	TSNAX-DISC1
rs2492367	miR-9-5p	UTR-5	NR_028397	TSNAX-DISC1
rs2492367	miR-9-5p	UTR-5	NR_028398	TSNAX-DISC1
rs2619538	let-7g-3p	1920B-upstream	.	DTNBP1

Tabela 5.3 – Continuação

SNP	microRNA	Contexto genômico	Transcrito	Gene
rs2738864	miR-26b-3p	intergenic	.	.
rs3738401	miR-198	CDS	NM_001012957	DISC1
rs3738401	miR-198	CDS	NM_001012958	DISC1
rs3738401	miR-198	CDS	NM_001012959	DISC1
rs3738401	miR-198	CDS	NM_001164537	DISC1
rs3738401	miR-198	CDS	NM_001164538	DISC1
rs3738401	miR-198	CDS	NM_001164539	DISC1
rs3738401	miR-198	CDS	NM_001164540	DISC1
rs3738401	miR-198	CDS	NM_001164541	DISC1
rs3738401	miR-198	CDS	NM_001164542	DISC1
rs3738401	miR-198	CDS	NM_001164544	DISC1
rs3738401	miR-198	CDS	NM_001164545	DISC1
rs3738401	miR-198	CDS	NM_001164546	DISC1
rs3738401	miR-198	CDS	NM_001164547	DISC1
rs3738401	miR-198	CDS	NM_001164548	DISC1
rs3738401	miR-198	CDS	NM_001164549	DISC1
rs3738401	miR-198	CDS	NM_001164550	DISC1
rs3738401	miR-198	CDS	NM_001164551	DISC1
rs3738401	miR-198	CDS	NM_001164552	DISC1
rs3738401	miR-198	CDS	NM_001164553	DISC1
rs3738401	miR-198	CDS	NM_001164554	DISC1
rs3738401	miR-198	CDS	NM_001164555	DISC1

Tabela 5.3 – Continuação

SNP	microRNA	Contexto genômico	Transcrito	Gene
rs3738401	miR-198	intron	NM_001164556	DISC1
rs3738401	miR-198	CDS	NM_018662	DISC1
rs3738401	miR-198	UTR-5	NR_028393	TSNAX-DISC1
rs3738401	miR-198	UTR-5	NR_028394	TSNAX-DISC1
rs3738401	miR-198	UTR-5	NR_028395	TSNAX-DISC1
rs3738401	miR-198	UTR-5	NR_028396	TSNAX-DISC1
rs3738401	miR-198	UTR-5	NR_028397	TSNAX-DISC1
rs3738401	miR-198	UTR-5	NR_028398	TSNAX-DISC1
rs3738401	miR-198	UTR-5	NR_028399	TSNAX-DISC1
rs3738401	miR-198	UTR-5	NR_028400	TSNAX-DISC1
rs383964	miR-212-5p	intergenic	.	.
rs4680	miR-24-3p	CDS	NM_000754	COMT
rs4680	miR-24-3p	CDS	NM_001135161	COMT
rs4680	miR-24-3p	CDS	NM_001135162	COMT
rs4680	miR-24-3p	CDS	NM_007310	COMT
rs4680	miR-24-3p	4B-upstream	.	MIR4761
rs6267	miR-206	intergenic	.	.
rs6277	miR-26b-3p	CDS	NM_000795	DRD2
rs6277	miR-26b-3p	CDS	NM_016574	DRD2
rs6277	miR-29b-1-5p	CDS	NM_000795	DRD2
rs6277	miR-29b-1-5p	CDS	NM_016574	DRD2
rs6994992	miR-26b-5p	intergenic	.	.

Tabela 5.3 – Continuação

SNP	microRNA	Contexto genômico	Transcrito	Gene
rs737865	miR-30d-3p	762B-upstream	.	TXNRD2
rs742105	miR-198	intergenic	.	.
rs742105	miR-30b-3p	intergenic	.	.
rs751229	miR-30d-3p	intergenic	.	.
rs9432024	miR-212-5p	intergenic	.	.
rs951436	miR-198	intergenic	.	.

Estes resultados mostram que um mesmo SNP está alterando vários sítios alvos de miRNAs. Quando os miRNAs correspondem a membros de uma mesma família, geralmente a região *seed* é conservada, tornando as sequências semelhantes e consequentemente aumentando a chance de estarem ligados a um mesmo sítio alvo. Como exemplo desta situação tem-se o SNP rs1011313 afetando os sítios dos miRNAs miR-24-1-5p e miR-24-2-5p e também o rs6277 afetando os sítios dos miRNAs miR-26b-3p e miR-26b-1-5p (Tabela 5.3).

Já o fato de existirem diferentes miRNAs podendo se ligar em um mesmo sítio pode significar um modo de segurança (*backup mode*) do organismo, uma vez que se um miRNA falhar, teriam outros que poderiam se ligar e regular o alvo [Balaga *et al.*, 2012]. Nestes casos, esses dois SNPs estão interferindo na ligação dos dois miRNAs.

Fazendo um recorte dos resultados que envolvem apenas estes 15 miRNAs também associados a esquizofrenia, é possível agrupar os genes alvos em comum para esta lista de miRNAs, como mostra a Tabela 5.4. O aumento da quantidade de sítios alvos é diretamente proporcional a efetividade de regulação. Entretanto, os efeitos de regulação por par SNP e miRNA mostrados nesta tabela ainda não tinham sido mostrados por nenhum outro trabalho, como mostram as duas fontes de dados validados consultadas: miRecords [Xiao *et al.*, 2009] e TarBase [Vergoulis *et al.*, 2012] (setembro de 2013).

Tabela 5.4: *Genes com sítios de miRNAs interferidos por SNPs, para miRNAs e SNPs associados a esquizofrenia.*

MicroRNAs	Gene	Descrição dos genes
miR-24-3p	COMT	catechol-O-methyltransferase
miR-206, miR-9-5p,miR-198	DISC1	disrupted in schizophrenia 1
miR-92b-5p, miR-26b-3p,miR-29b-1-5p	DRD2	dopamine receptor D2
miR-24-1-5p,miR-24-2-5p, miR-9-5p,let-7g-3p	DTNBP1	dystrobrevin binding protein 1
miR-24-3p	MIR4761	microRNA 4761
miR-92b-5p	PLXNA2	plexin A2
miR-206, miR-9-5p,miR-198	TSNAX-DISC1	TSNAX-DISC1 readthrough
miR-30d-3p	TXNRD2	thioredoxin reductase 2

A não identificação de tais resultados na base de dados validados deve-se ao fato de que, os atuais recursos computacionais disponíveis para a identificação de SNPs que interferem em sítios alvos de miRNAs apenas analisam SNPs localizados na região *seeds* e com sítios na região 3'UTR de mRNAs, diferentemente do SIMTar.

Uma outra forma de visualizar os resultados preditos por SIMTar para os 15 miRNAs associados a esquizofrenia é através de uma rede de interação (Figura 5.4) que mostra, neste caso, os diferentes miRNAs e os seus respectivos alvos. Para exemplificar como pode ser realizada a interpretação biológica a partir desses dados, escolheu-se os genes *COMT* e *MIR4761*, ambos alvos do miR-24-3p e para o mesmo SNP, o rs4680.

Primeiramente, o rs4680 é um SNP associado a esquizofrenia e que apresenta o alelo G com maior frequência nos pacientes suscetíveis (SzGene [Allen *et al.*, 2008]).

O gene *COMT* codifica a proteína catecol-O-metiltransferase, que por sua vez participa da regulação da via dopaminérgica. A dopamina é um neurotransmissor com função estimulante do sistema nervoso central que está envolvida no controle de movimentos, aprendizado, cognição e memória. Em pacientes com esquizofrenia são observadas concentrações alteradas dessa subtância, o que pode

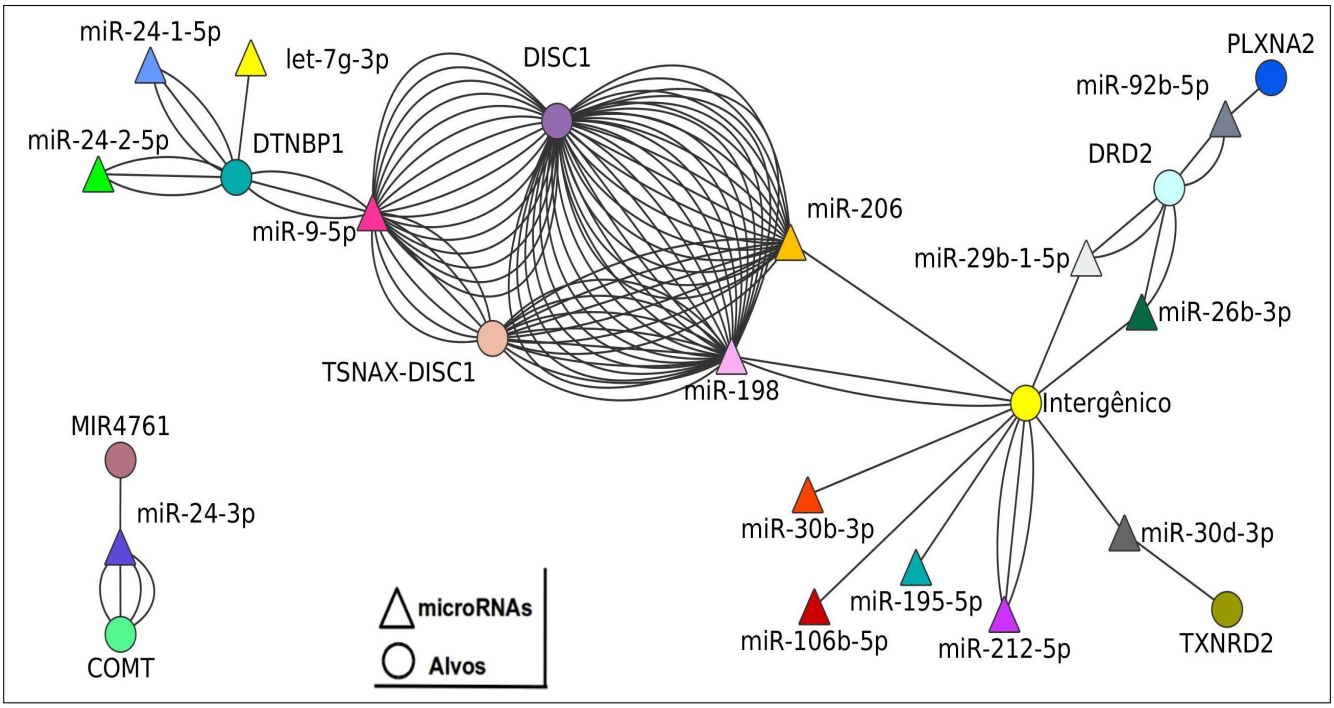

Figura 5.4: *Rede de interação entre miRNAs e os seus respectivos alvos identificados a partir da predição genômica de SIMTar, utilizando dados de esquizofrenia. Triângulos representam os miRNAs, círculos representam os alvos e cada aresta representa um transcrito para sítios alvos em região gênica, já os sítios alvos em região intergênica, cada aresta representa um SNP.* *Software utilizado: cytoscape [Shannon* ET AL.*, 2003].*

estar ocorrendo em função da alteração da expressão do gene *COMT*. De acordo com os resultados totais gerados por SIMTar, o miRNA-24-3p foi predito se ligando em quatro transcritos desse gene, o NM_001135161, NM_001135162, NM_007310 e NM_000754, todos com sítios alvos localizados na região CDS. SIMTar mostrou ainda que a interferência da variação alélica do rs4680 foi observada pelo o alelo A (A/C/G/T) na criação do sítio alvo (dbSNP versão 135 para humano). Isso significa que segundo a predição computacional, na presença do alelo G (o mais frequente em pacientes com esquizofrenia), o sítio alvo do miR-24-3p está sendo rompido.

Como descrito na subsessão 2.3.4, quando ligado à região CDS, um miRNA pode regular negativamente a tradução do gene. Para confirmar essa regulação tais resultados devem ser submetidos a experimentos biológicos, que permitem além de validar a predição computacional, também mostrar se a variação alélica de A para G está influenciando na síntese da proteína catechol-O-metiltransferase e portanto, afetando o metabolismo dopaminérgico. Além disso, para melhor compreensão da regulação por miRNAs deve-se considerar também a existência de outros sítios alvos, seja do mesmo miRNA ou

de miRNAs diferentes, localizados ao longo do mesmo gene e que podem estar agindo sinergicamente na regulação do alvo, como mostrou Gong *et al.* (2013).

O outro alvo do miR-24-3p é o gene *MIR4761*, que é um gene de miRNA. Embora tenha sido predito como alvo de um miRNA associado a esquizofrenia, a(s) sua(s) forma(s) madura(s) não está(ão) associada(s) à doença, segundo o banco de dados *Schizophrenia Gene Resource* - SZGR [Jia *et al.*, 2010]. Aqui, o sítio alvo está localizado 4 bases-upstream do gene, indicando uma possível regulação transcricional do mesmo.

Considerando que a expressão deste gene é de fato influenciada pela ligação com o miR-24-3p, consequentemente as suas formas maduras também seriam. Além disso, também poderiam ser influenciados os alvos que tal(is) miRNA(s) estaria(m) se ligando. Neste caso, verificou-se a existência dessas formas maduras no conjunto de resutados totais, para então ver o possível gene que estaria sendo afetado. Observou-se que o miR-4761-5p está se ligando 1920 bases-upstream do gene *DTNBP1*. Neste pareamento a presença do alelo T (T/A) do SNP rs2619538 está criando o sítio alvo. Esse resultado mostra uma potencial regulação ao nível transcricional deste gene.

O gene *DTNBP1* codifica uma proteína cujo papel está relacionado a biogênese de organelas. Em esquizofrenia ele está associado ao declínio intelectual dos pacientes [Straub *et al.*, 2002].

Com a predição de SIMTar é possível visualizar uma possível ligação entre os genes *MIR4761* e *DTNBP1*, como pode ser visualizado na Figura 5.5, que não seria possível através da utilização dos atuais recursos computacionais da área, dado que a abordagem por eles realizada apenas considera SNP localizados na região 3'UTR de mRNAs.

O SNP no gene *DTNBP1* que influencia o sítio alvo do miR-4761-5p também influencia o sítio do miRNA let-7g-3p.

Diferente dos sítios alvos preditos na região upstream do gene *DTNBP1*, também foram preditos resultados para os transcritos NM_032122, NM_183040 e NR_036448 deste gene como alvos dos miRNAs miR-24-1-5p e miR-24-2-5p, pela interferência do SNP rs1011313, que está localizado no íntron deste gene e como alvos do miRNA miR-9-5p, pela interferência do SNP rs1997679, um SNP também localizado na região intrônica do gene. Quanto a forma de regulação por estes miRNAs ligados em região intrônica ainda não existe nada relatado, embora tenha sido revelado recentemente

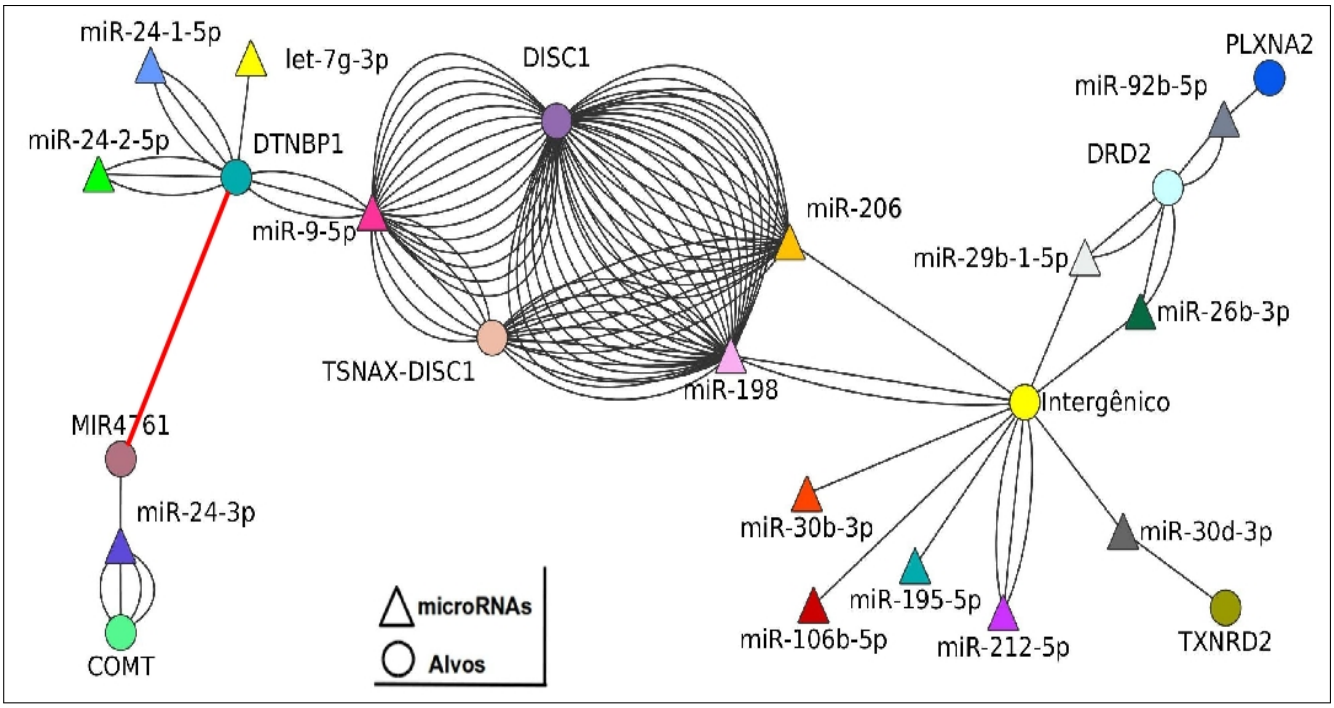

Figura 5.5: *Rede de interação entre miRNAs e os seus respectivos alvos identificados a partir da predição genômica de SIMTar utilizando dados de esquizofrenia. Destaque para a interação criada entre o miR-4761-5p (forma madura do gene MIR4761) e o gene DTNBP1 (aresta em vermelho).* *Software utilizado: cytoscape [Shannon* ET AL.*, 2003].*

que os íntrons compreendem uma importante região de atividade de RNAs [ENCODE Project Consortium *et al.*, 2012]. Cabe dessa forma, explorar futuramente a existência de sequências de RNAs não codificantes nestas regiões preditas como alvos de tais miRNAs.

Devido à sua capacidade de regulação, os miRNAs têm sido foco de cada vez mais estudos com doenças. O seu estudo em esquizofrenia, considerando a interferência de SNPs no DNA, pode contribuir ainda mais para o entendimento dos processos que estão envolvidos na expressão do fenótipo suscetível. Os resultados totais dessa aplicação estão disponíveis em: `http://www.vision.ime.usp.br/~amanda/software/simtar/applications/`.

Capítulo 6

Conclusões e trabalhos futuros

6.1 Conclusões

Em estudos de regulação por miRNAs, têm crescido as investigações quanto à presença de polimorfismos de um nucleotídeo (SNPs) ocorrendo em sítios alvos de miRNAs. Isso porque a sua variação alélica pode interferir no pareamento com o sítio alvo. Além disso, se tanto o SNP quanto o miRNA estiver associado a algum traço fenotípico, a predição de interferência entre eles aumenta ainda mais a conexão de regulação.

O fato é que a investigação envolvendo estes dois elementos deve incluir as várias formas de regulação por miRNAs e, portanto, todas as possíveis regiões em que eles puderem se ligar. No entanto, os recursos computacionais existente para predizer essa relação apenas a fazem sobre a primeira via de regulação conhecida, que é sobre a região 3'UTR de mRNAs. Com isso todas as outras regiões não estão sendo analisadas. Além disso, não está sendo considerada a possibilidade de mais de um SNP estar localizado em um mesmo sítio alvo e também estar interferindo no pareamento.

Dessa forma, a principal questão levantada nesse trabalho foi a metodologia utilizada na identificação de SNPs que podem alterar sítios alvos de miRNAs em função de sua variação alélica. Para isso foram discutidos os fatores que podem estar relacionados a esta ligação. Em linhas gerais foram considerados e discutidos: a escolha da ferramenta de predição de alvos de miRNAs, que pode res-

tringir a análise para alvos conservados evolutivamente ou não, o delineamento para análise de cada alelo do SNP e também de SNPs vizinhos inseridos em um mesmo sítio alvo, o contexto genômico da região alvo, a posição do SNP em relação ao sítio alvo, que considera o SNP dentro ou fora da seed, a análise de sequência de DNA (genômica) ou RNA (apenas éxons) a existência e disponibilidade para uso local de uma ferramenta de predição de tais SNPs e a espécie analisada.

Com o levantamento realizado ficou claro que as tecnologias disponíveis no estudo de SNPs e miRNAs resolvem apenas uma parte do problema. Neste sentido, este trabalho desenvolveu a ferramenta computacional **SIMTar: SNPs Interfering on MicroRNA Target sites**, a fim de superar as principais lacunas existentes.

A ferramenta analisa SNPs localizados tanto em regiões codificadoras como não codificadoras de proteínas. Os resultados são gerados para cada alelo ou combinação de alelos do SNP, caso existam mais um localizados em um mesmo sítio alvo. Para análise de SNPs localizados em DNA, os resultados mostram se naquela região existem sequências transcritas de RNAs codificadores e não codicadores de proteínas. Por fim, SIMTar é uma ferramenta de código aberto e que pode ser aplicada a outras espécies além da humana. No presente trabalho, SIMTar foi aplicado a dados de esquizofrenia como estudo de caso, mas esta pode ser utilizada para qualquer lista de SNPs.

No estudo de caso utilizando SNPs associados à esquizofrenia, 95% dos SNPs foram preditos interferindo em sítios alvos de miRNAs. Além disso, nestes resultados foram identificados 8 de 18 miRNAs também associados à esquizofrenia. Com a análise dos resultados foi possível identificar quais os alelos dos SNPs estão afetando os sítios alvos e os nomes dos miRNAs envolvidos. Quando todos os dados dos SNPs e miRNAs preditos são plotados em uma rede de interação, é possível identificar a conexão entre mais de um gene, quando ambos já foram observados em estudos de associação como é o caso da esquizofrenia, seria interessante investigar se estes miRNAs não estariam agindo sobre os sítios alvos preditos e regulando a região, seja em nível transcricional ou traducional. Esta regulação pode estar envolvendo processos biológicos e vias metabólicas importantes para o indivíduo e que pode estar danificada no estado de doença.

Neste estudo, a lista de SNPs foi obtida a partir de 38 genes candidatos em esquizofrenia, partindo assim de uma lista consideravelmente pequena, dado que os SNPs correspondem à forma mais

frequente de variação do genoma humano. Por outro lado, se for considerado o conjunto de dados gerado por exemplo por estudos de associação, onde centenas de SNPs são obtidos para cada doença e a maioria ainda não possui papel biológico definido, SIMTar pode ser aplicado para indicar um subconjunto mais promissor de SNPs para alguma validação biológica. Por exemplo, é possível selecionar facilmente os SNPs que interferem em um grande número de sítios. Uma das saídas do SIMTar é justamente a lista de SNPs ordenada decrescentemente pelo número de sítios alvos interferidos pelo SNP.

Um artigo descrevendo o SIMTar e sua aplicação em esquizofrenia está em fase final de escrita. Além disso está sendo solicitado o registro do SIMTar como software aberto.

6.2 Trabalhos futuros

SIMTar é executado em linha de comando. No entanto, isso não representa um impedimento para usuários que não possuem familiaridade com programação, pois está implementado para ser executado através de um gerenciador, facilitando o processo de execução. Além disso, o tutorial SIMTar explica passo a passo o que deve ser feito para sua execução e aplicação na espécie padrão (humana) ou em outras espécies. No entanto, o desenvolvimento de uma interface gráfica para configuração dos parâmetros e execução da ferramenta, respeitando critérios de usabilidade, deve facilitar o uso SIMTar para usuários não habituados com interfaces baseadas em texto e linha de comando. Além disso, o desenvolvimento de métodos de vizualização gráfica dos resultados contribuirão para uma melhor interpretação dos mesmos.

Além disso, como já mencionado, um bom experimento para o SIMTar seria executá-lo tendo como entrada um grande número de SNPs vindos de um estudo de associação.

Apêndice A

SNPs utilizados no estudo de caso de SIMTar

Relação de genes e SNPs associados à esquizofrenia utilizados na execução de SIMTar.												
Genes	MTHFR	RGS4	PLXNA2	DISC1	TPH1	DRD4	DRD2	GRIN2B	DAO	HTR2A	DAOA	CHRNA7
ID SNP	rs1801133	rs2661319	rs841865	rs3737597	rs1800532	rs1800955	rs1079597	rs1019385	rs4623951	rs6311	rs778293	rs3826029
	rs1801131	rs10917670	rs752016	rs11122359	rs1799913	rs747302	rs1799732	rs1805502	rs2070586	rs6313	rs2391191	
		rs951436	rs2498028	rs1322783			rs1800497	rs1806201	rs2070587	rs6314	rs3916965	
		rs951439	rs1327175	rs1322784			rs1801028		rs2111902		rs3916966	
				rs1934909			rs6275		rs3741775		rs3916967	
				rs1984895			rs6277		rs3825251		rs3916970	
				rs2038636					rs3918346		rs3916971	
				rs2255340					rs3918347		rs3918342	
				rs2492367					rs4623951		rs778293	
				rs2738864							rs778294	
				rs3738401							rs947267	
				rs6675281								
				rs751229								
				rs821597								
				rs821616								
				rs843979								
				rs9432024								
				rs999710								
				rs3737597								
				rs1000731								

Genes	TP53	IL1B	GAD1	ZNF804A	ERBB4	PRODH	COMT	DRD1	DTNBP1	RELN	NRG1
ID SNP	rs1042522	rs16944	rs3749034	rs1344706	rs839523	rs383964	rs737865	rs4532	rs3213207	rs7341475	rs10503929
		rs1143634				rs372055	rs165599		rs1011313	rs155333	rs2439272
							rs165774		rs1018381	rs262355	rs2466058
							rs4633		rs1047631	rs39339	rs2954041
							rs4646316		rs1474605		rs35753505
							rs4680		rs17470454		rs3924999
							rs4818		rs1997679		rs4733376
							rs6267		rs2005976		rs6988339
							rs737865		rs2619522		rs6994992
							rs740603		rs2619528		rs776401
							rs769224		rs2619538		
									rs2619539		
									rs3213207		
									rs742105		
									rs742106		
									rs760666		
									rs760761		
									rs909706		

Figura A.1: *Relação dos 23 genes e 114 SNPs utilizados no estudo de caso. Em preto estão os SNPs identificados por SIMTar. Em vermelho estão destacados apenas aqueles SNPs que não foram identificados alterando sítios alvos de miRNAs.*

Referências Bibliográficas

Allen, N. C., Bagade, S. & McQueen, M. B. (2008). Systematic Meta-Analyses and Field Synopsis of Genetic Association Studies in Schizophrenia: The SzGene Database. *Nature Genetics* **40**, 827–34.

Ambros, V. (2003). MicroRNA pathways in flies and worms: Growth, death, fat, stress, and timing. *Cell* **113**, 673–676.

Balaga, O., Friedman, Y. & Linial, M. (2012). Toward a combinatorial nature of microRNA regulation in human cells. *Nucleic Acids Research* **40**, 9404–9416.

Bandyopadhyay, S. & Mitra, R. (2009). TargetMiner: microRNA target prediction with systematic identification of tissue-specific negative examples. *Bioinformatics* **25**, 2625–2631.

Barenboim, M., Zoltick, B. J., Guo, Y. & Weinberger, D. R. (2010). MicroSNiPer: a web tool for prediction of snp effects on putative microRNA targets. *Human Mutation* **31**(11), 1223–1232.

Bartel, D. P. (2009). MicroRNAs: target recognition and regulatory functions. *Cell* **136**(2), 215–233.

Baskerville, S. & Bartel, D. P. (2005). Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA* **11**(3), 241–247.

Berezikov, E., Thuemmler, F., Van Laake, L. W. & et al (2006). Diversity of microRNAs in human and chimpanzee brain. *Nature Genetics* **38**, 1375–1377.

Betel, D., Wilson, M., Gabow, A. & et al (2008). The microRNA.org resource: targets and expression. *Nucleic Acids Research* **36**, D149–D153.

Bohnsack, M. T., Czaplinski, K. & Gorlich, D. (2004). Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs. *RNA* **10**(2), 185–191.

Borchert, G. M., Lanier, W. & Davidson, B. L. (2006). RNA polymerase iii transcribes human microRNAs. *Nature Structural & Molecular Biology* **13**(12), 1097–1101.

Brennecke, J., Hipfner, D. R., Stark, A. & et al (2003). bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in drosophila. *Cell* **113**, 25–36.

Brennecke, J., Stark, A., Russell, R. B. & Cohen, S. M. (2005). Principles of MicroRNA–Target Recognition. *PLoS Biology* **3**(3), e85+.

Brookes, A. J. (1999). The essence of snps. *Gene* **234**, 177–186.

Bruno, A. E., Li, L., Kalabus, J. L. & et al. (2012). miRdSNP: a database of disease-associated SNPs and microRNA target sites on 3'UTRs of human genes. *BMC Genomics* **13:14**.

Burge, S., Daub, J., Eberhardt, R. & et al (2012). Rfam 11.0: 10 years of RNA families. *Nucleic Acids Research* .

Burgler, C. & Macdonald, P. M. (2005). Prediction and verification of microRNA targets by MovingTargets, a highly adaptable prediction method. *BMC Genomics* **6**, 88.

Bush, W. S. & Moore, J. H. (2012). Chapter 11: Genome-Wide Association Studies. *PLoS Computational Biology* **8**, e1002822.

Büssing, I., Slack, F. J. & Gross, H. (2008). let-7 microRNAs in development, stem cells and cancer. *Trends in Molecular Medicine* **14**(9), 400–409.

Cai, X., Hagedorn, C. H. & Cullen, B. R. (2004). Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *RNA* **10**(12), 1957–1966.

Carè, A., Catalucci, D., Felicetti, F. & et al (2007). MicroRNA-133 controls cardiac hypertrophy. *Nature Medicine* **13**, 613–618.

Carninci, P., Kasukawa, T., Katayama, S. & et al (2005). The transcriptional landscape of the mammalian genome. *Science* **309**, 1559–1563.

Chen, C.-Z., Li, L., Lodish, H. F. & et al (2004). MicroRNAs Modulate Hematopoietic Lineage Differentiation. *Science* **303**(5654), 83–86.

Chen, K. & Rajewsky, N. (2006). Natural selection on human microRNA binding sites inferred from SNP data. *Nature Genetics* **38**(12), 1452–1456.

Crick, F. (1970). Central dogma of molecular biology. *Nature* **227**, 561–563.

Dai, X. & Zhao, P. X. (2011). psRNATarget: a plant small RNA target analysis server. *Nucleic Acids Research* **39**, W155–W159.

Doench, J. G. & Sharp, P. A. (2004). Specificity of microRNA target selection in translational repression. *Genes & Development* **18**(5), 504–511.

Durham, A. M., Kashiwabara, A. Y., Matsunaga, F. T. G. & et al (2005). Egene: a configurable pipeline generation system for automated sequence analysis. *Bioinformatics* **21**, 2812–2813.

Duursma, A. M., Kedde, M., Schrier, M. & et al (2008). mir-148 targets human DNMT3b protein coding region. *RNA* , 872–877.

ENCODE Project Consortium, Bernstein, B. E., Birney, E., Dunham, I., Green, E. D., Gunter, C. & Snyder, M. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**(7414), 57–74.

Enright, A., John, B., Gaul, U. & et al (2003). MicroRNA targets in drosophila. *Genome Biology* **5**, R1+.

Faustino, N. A. & Cooper, T. A. (2013). Pre-mRNA splicing and human disease. *GENES & DEVELOPMENT* **17**, 419–437.

Frankel, L. B., Christoffersen, N. R., Jacobsen, A. & et al (2008). Programmed Cell Death 4 (PDCD4) Is an Important Functional Target of the Microrna miR-21 in Breast Cancer Cells. *Journal of Biological Chemistry* **283**(2), 1026–1033.

Friedman, R. C., Farh, K. K., Burge, C. B. & et al (2009). Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research* , 92–105.

Frith, M. C., Pheasant, M. & Mattick, J. S. (2005). The amazing complexity of the human transcriptome. *European Journal of Human Genetics* **13**, 894–897.

Gaidatzis, D., van Nimwegen, E., Hausser, J. & Zavolan, M. (2007). Inference of miRNA targets using evolutionary conservation and pathway analysis. *BMC Bioinformatics* **8**, 69+.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher* **5**, 3–8.

Gong, J., Tong, Y., Zhang, H.-M. & et al (2012). miRNASNP: a database of miRNA related SNPs and their effects on miRNA function. *BioMed Central Bioinformatics* **13**, A2.

Gong, Y., Wu, C. N., Xu, J. & et al (2013). Polymorphisms in microRNA target sites influence susceptibility to schizophrenia by altering the binding of miRNAs to their targets. *European Neuropsychopharmacology* **23**(10), 1182–1189.

Griffiths-Jones, S. (2004). The microRNA registry. *Nucleic Acids Research* **32**, D109–11.

Griffiths-Jones, S. (2006). mirbase: MicroRNA sequences, targets and gene nomenclature. *Nucleic Acids Research* **34**, D140–D144.

Griffiths-Jones, S., Saini, H. K., van Dongen, S. & et al (2008). miRBase: tools for microRNA genomics. *Nucleic Acids Research* **36**, D154–D158.

Grimson, A., Farh, K. K., Johnston, W. K. & et al (2007). MicroRNA Targeting Specificity in Mammals: Determinants beyond Seed Pairing. *Molecular Cell* **27**(1), 91–105.

Grün, D., Wang, Y.-L., Langenberger, D. & et al (2005). microRNA Target Predictions across Seven Drosophila Species and Comparison to Mammalian Targets. *PLoS Computational Biology* **1**(1), e13+.

Hammell, M., Long, D., Zhang, L. & et al (2008). mirWIP: microRNA target prediction based on microRNA-containing ribonucleoprotein-enriched transcripts. *Nature methods* **5**(9), 813–819.

Harrison, P. J. & Weinberger, D. R. . (2005). Schizophrenia genes, gene expression, and neuropathology: on the matter of their convergence. *Molecular Psychiatry* **10**, 40–68.

Henke, J. I., Goergen, D., Zheng, J. & et al (2008). microRNA-122 stimulates translation of hepatitis C virus RNA. *The EMBO Journal* **27**, 3300–3310.

Hindorff, L. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. In: *Proceedings of the National Academy of Sciences*, vol. 106.

Huang, V., Place, R. F., Portnoy, V. & et al (2012). Upregulation of cyclin B1 by miRNA and its implications in cancer. *Nucleic Acids Research* **40**, 1695–1707.

Hunter, D. J. (2005). Gene-environment interactions in human diseases. *Nature Reviews Genetics* **6**(4), 287–298.

Jackson, R. J., Hellen, C. U. & Pestova, T. V. (2010). The mechanism of eukaryotic translation initiation and principles of its regulation. *Nature Reviews Molecular Cell Biolology* **11**, 113–127.

Jia, P., Sun, J., Guo, A. Y. & et al (2010). SZGR: a comprehensive schizophrenia gene resource. *Molecular psychiatry* **15**(5), 453–462.

John, B., Enright, A. J., Aravin, A. & et al (2004). Human MicroRNA Targets. *PLoS Biology* **2**(11).

Johnston, R. J. & Hobert, O. (2003). A microRNA controlling left/right neuronal asymmetry in Caenorhabditis elegans. *Nature* **426**, 845–849.

Jones, P. A. & Baylin, S. B. (2007). The epigenomics of cancer. *Cell* **128**, 683–692.

Kent, W. J. (2002). BLAT – the BLAST-like alignment tool. *Genome Research* **12**, 656–664.

Kertesz, M., Iovino, N., Unnerstall, U. & et al (2007). The role of site accessibility in microRNA target recognition. *Nature Genetics* **39**(10), 1278–1284.

Kim, S.-K., Nam, J.-W., Rhee, J.-K. & et al (2006). miTarget: microRNA target gene prediction using a support vector machine. *BMC Bioinformatics* **7**, 411.

Kiriakidou, M., Nelson, P. T., Kouranov, A. & et al (2004). A combined computational-experimental approach predicts human microRNA targets. *Genes Development* **18**, 1165–1178.

Krek, A., Grün, D., Poy, M. N. & et al (2005). Combinatorial microRNA target predictions. *Nature genetics* **37**(5), 495–500.

Krüger, J. & Rehmsmeier, M. (2006). RNAhybrid: microRNA target prediction easy, fast, and flexible. *Nucleic Acids Research* , 451–454.

Kwan, T., Benovoy, D., Dias, C. & et al (2008). Genome-wide analysis of transcript isoform variation in humans. *Nature genetics* **40**(2), 225–231.

Lagos-Quintana, M., Rauhut, R., Meyer, J. & et al (2003). New microRNAs from mouse and human. *RNA* **9**(2), 175–179.

Lagos-Quintana, M., Rauhut, R., Yalcin, A. & et al (2002). Identification of tissue-specific microRNAs from mouse. *Current biology* **12**(9), 735–739.

Lai, E., Tomancak, P., Williams, R. & et al (2003). Computational identification of Drosophila microRNA genes. *Genome Biology* **4**(7), R42+.

Lall, S., Grün, D., Krek, A. & et al (2006). A Genome-Wide Map of Conserved MicroRNA Targets in C. elegans. *Current Biology* **16**(5), 460–471.

Lander, E. S., Linton, L. M., Birren, B. & et al (2001). Initial sequencing and analysis of the human genome. *Nature* **409**(6822), 860–921.

Lau, N. C., Lim, L. P., Weinstein, E. G. & et al (2001). An abundant class of tiny RNAs with probable regulatory roles in caenorhabditis elegans. *Science* **294**(5543), 858–862.

Lee, R. C., Feinbaum, R. L. & Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* **75**(5), 843–854.

Lee, S. H., DeCandia, T. R. & Ripke, S. (2012). Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nature Genetics* **44**, 247–250.

Lee, S. J., Jeon, H. S., Jang, J. S. & et al (2005). DNMT3B polymorphisms and risk of primary lung cancer. *Carcinogenesis* **26**, 403–409.

Lee, Y., Ahn, C., Han, J. & et al (2003). The nuclear RNase iii Drosha initiates microRNA processing. *Nature* **425**(6956), 415–419.

Lee, Y., Jeon, K., Lee, J. T. & et al (2002). MicroRNA maturation: stepwise processing and subcellular localization. *EMBO Journal* **21**(17), 4663–4670.

Lee, Y., Kim, M., Han, J. & et al (2004). MicroRNA genes are transcribed by RNA polymerase ii. *The EMBO journal* **23**(20), 4051–4060.

Lewis, B. P., Burge, C. B. & Bartel, D. P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120**, 15–20.

Lewis, B. P., Shih, I.-h. H., Jones-Rhoades, M. W. & et al (2003). Prediction of mammalian microRNA targets. *Cell* **115**(7), 787–798.

Lim, L. P., Glasner, M. E., Yekta, S. & et al (2003). Vertebrate MicroRNA Genes. *Science* **299**(5612), 1540.

Lin, C.-C., Liu, L.-Z., Addison, J. B. & et al (2011). A KLF4–miRNA-206 Autoregulatory Feedback Loop Can Promote or Inhibit Protein Translation Depending upon Cell Context. *Molecular Cell Biolology* **31**, 2513–2527.

Liu, C., Zhang, F., Li, T. & et al (2012). MirSNP, a database of polymorphisms altering miRNA target sites, identifies miRNA-related SNPs in GWAS SNPs and eQTLs. *BMC Genomics* **13**(1), 661.

Lodish, H., Berk, A., Zipursky, S. L. & et al (2000). *Section 9.1, Molecular Definition of a Gene*. New York: W. H. Freeman, Molecular Cell Biology. 4th edition ed.

Long, D., Lee, R. & Williams, P. (2007). Potent effect of target structure on microRNA function. *Nature Structure Molecular Biology* **14**, 287–94.

Lu, J. & Clark, A. G. (2012). Impact of microRNA regulation on variation in human gene expression. *Genome Research* **22**(7), 1243–1254.

Lu, M., Zhang, Q., Deng, M. & et al (2008). An Analysis of Human MicroRNA and Disease Associations. *PLoS ONE* **3**(10), e3420+.

Lund, E., Güttinger, S., Calado, A. & et al (2004). Nuclear export of microrna precursors. *Science* **303**(5654), 95–98.

Mattick, J. S. & Makunin, I. V. (2006). Non-coding RNA. *Human Molecular Genetics* **15**, R17–R29.

McGrath, J., Saha, S., Chant, D. & et al (2008). Schizophrenia: a concise overview of incidence, prevalence, and mortality. *Epidemiologic reviews* **30**, 67–76.

McVean, G. A., Donnelly, P., Lunter, G. & et al (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65.

Miranda, K. C., Huynh, T., Tay, Y. & et al (2006). A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell* **126**(6), 1203–1217.

Moss, E. G., Lee, R. C. & Ambros, V. (1997). The cold shock domain protein lin-28 controls developmental timing in c. elegans and is regulated by the lin-4 RNA. *Cell* **88**, 637–646.

Nielsen, C. B., Shomron, N., Sandberg, R. & et al (2007). Determinants of targeting by endogenous and exogenous microRNAs and siRNAs. *RNA* **13**(11), 1894–1910.

Nissim-Rafinia, M. & Kerem, B. (2005). The splicing machinery is a genetic modifier of disease severity. *Trends Genet* **21**(9), 480–483.

Ohno, S. (1972). So much "junk"DNA in our genome. *Brookhaven Symp Biol* **23**, 366–370.

Oliveira-Carvalho, V., Carvalho, V. O., Silva, M. M. & et al (2012). MicroRNAs: um novo paradigma no tratamento e diagnóstico da insuficiência cardíaca? *Arquivos Brasileiros de Cardiologia* **98**, 362–370.

Pasquinelli, A. E., Reinhart, B. J., Slack, F. & et al (2000). Conservation of the sequence and temporal expression of let-7 heterochronic regulatory {RNA}. *Nature* **408**, 86–89.

Pruitt, K. D., Tatusova, T., Klimke, W. & et al (2009). NCBI reference sequences: current status, policy and new initiatives. *Nucleic Acids Research* **37**, D32–36.

Rajewsky, N. & Socci, N. D. (2004). Computational identification of microRNA targets. *Developmental biology* **267**(2), 529–535.

Rehmsmeier, M., Steffen, P., Hochsmann, M. & et al (2004). Fast and effective prediction of microRNA/target duplexes. *RNA* **10**(10), 1507–1517.

Reinhart, B. J., Weinstein, E. G., Rhoades, M. W. & et al (2002). MicroRNAs in plants. *Genes Development* **16**, 1616–26.

Robins, H., Li, Y. & Padgett, R. W. (2005). Incorporating structure to predict microRNA targets. *Proceedings of the National Academy of Sciences of the United States of America* **102**(11), 4006–4009.

Rodriguez, A., Griffiths-jones, S. & Ashurst, J. L. (2004). Identification of Mammalian microRNA Host Genes and Transcription Units. *Genome Research* , 1902–1910.

Rusinov, V., Baev, V., Minkov, I. N. & et al (2005). MicroInspector: a web tool for detection of miRNA binding sites in an RNA sequence. *Nucleic Acids Research* **33**(suppl_2), W696–W700.

Sadedin, S. P., Pope, B. & Oshlack, A. (2012). Bpipe: a tool for running and managing bioinformatics pipelines. *Bioinformatics* **28**(11), 1525–1526.

Saenger, W. (1984). *Principles of Nucleic Acid Structure*. New York : Springer-Verlag.

Sætrom, O., Snøve, O. & Sætrom, P. (2005). Weighted sequence motifs as an improved seeding step in microRNA target prediction algorithms. *RNA* **11**, 995–1003.

Sætrom, P., Heale, B. S. E., Snøve, O. & et al (2007). Distance constraints between microRNA target sites dictate efficacy and cooperativity. *Nucleic Acids Research* **35**(7), 2333–2342.

Saito, T. & Sætrom, P. (2010). MicroRNAs–targeting and target prediction. *Nature Biotechnology* **27**(3), 243–249.

Savinkova, L., Drachkova, I., Arshinova, T. & et al (2013). An Experimental Verification of the Predicted Effects of Promoter TATA-Box Polymorphisms Associated with Human Diseases on Interactions between the TATA Boxes and TATA-Binding protein. *PLoS ONE* **8**(2), e54626.

Sempere, L. F., Sokol, N. S., Dubrovsky, E. B. & et al (2003). Temporal regulation of microRNA expression in Drosophila melanogaster mediated by hormonal signals and broad-Complex gene activity. *Developmental biology* **259**, 9–18.

Shannon, P., Markiel, A., Ozier, O. & et al (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research* **13**, 2498–504.

Shen, H., Wang, L., Spitz, M. & et al (2002). A novel polymorphism in human cytosine DNA-methyltransferase-3B promoter is associated with an increased risk of lung cancer. *Cancer Research* **62**, 4992–4995.

Si, M. L., Zhu, S., Wu, H. & et al (2006). miR-21-mediated tumor growth. *Oncogene* **26**(19), 2799–2803.

Stark, A., Brennecke, J., Russell, R. B. & et al (2003). Identification of Drosophila MicroRNA Targets. *PLoS Biology* **1**(3), e60+.

Straub, R. E., Jiang, Y., MacLean, C. J. & et al (2002). Genetic variation in the 6p22.3 gene dtnbp1,

the human ortholog of the mouse dysbindin gene, is associated with schizophrenia. *American Journal Human Genetics* **71**, 337–348.

Sullivan, P. F. (2005). The genetics of schizophrenia. *PLoS Medicine* **2**, e212.

Thadani, R. & Tammi, M. T. (2006). MicroTar: predicting microRNA targets from RNA duplexes. *BMC Bioinformatics* **7**.

Thomas, G., Jacobs, K. B., Kraft, P. & et al (2009). A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11$\dot{2}$ and 14q24$\dot{1}$ *rad*51*l*1. *Nature Genetics* **41**, 579–84.

Toscano-Garibay, J. D. & Aquino-Jarquin, G. (2012). Regulation exerted by miRNAs in the promoter and UTR sequences: MDR1/P-gp expression as a particular case. *DNA Cell Biology* **31**, 1358–64.

Valencia-Sanchez, M. A., Liu, J., Hannon, G. J. & et al (2006). Control of translation and mRNA degradation by miRNAs and siRNAs. *Genes & Development* **20**(5), 515–524.

Vasudevan, S., Tong, Y. & Steitz, J. A. (2007). Switching from repression to activation: microRNAs can up-regulate translation. *Science* **318**(5858), 1931–1934.

Vella, M. C., Choi, E. Y., Lin, S. Y. & et al (2004). The c. elegans microRNA let-7 binds to imperfect let-7 complementary sites from the lin-41 3'utr. *Genes Development* **18**, 132–137.

Venter, J. C., Adams, M. D., Myers, E. W. & et al (2001). The sequence of the human genome. *Science* **291**(5507), 1304–1351.

Vergoulis, T. I., Vlachos, P., Alexiou, G. & et al (2012). Tarbase 6.0: Capturing the exponential growth of mirna targets with experimental suppor. *Nucleic Acids Research* **40**(D1), D222–D229.

Wang, X. & El Naqa, I. M. (2008). Prediction of both conserved and nonconserved microRNA targets in animals. *Bioinformatics* **24**, 325–332.

Watson, J. D. & Crick, F. H. C. (1953). A structure for deoxyribose nucleic acid. *Nature* **171**, 737–738.

Weinmann, L., Hock, J., Ivacevic, T. & et al (2009). Importin 8 is a gene silencing factor that targets argonaute proteins to distinct mRNAs. *Cell* **136**, 496–507.

Wilusz, J. E., Sunwoo, H. & Spector, D. L. (2009). Long noncoding RNAs: functional surprises from the RNA world. *Genes & Development* **23**(13), 1494–1504.

Xiao, F., Zuo, Z., Cai, G. & et al (2009). miRecords: an integrated resource for microRNA-target interactions. *Nucleic acids research* **37**, D105–D110.

Yang, B., Lin, H., Xiao, J. & et al (2007). The muscle-specific microRNA miR-1 regulates cardiac arrhythmogenic potential by targeting GJA1 and KCNJ2. *Nature Medicine* **13**(4), 486–491.

Yi, R., Qin, Y., Macara, I. G. & et al (2003). Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. *Genes & development* **17**(24), 3011–3016.

Yousef, M., Jung†, S. & Kossenkov, A. V. (2007). Naïve bayes for microrna target predictions—machine learning for microrna targets. *Bioinformatics* **23**, 2987–2992.

Zeng, Y. & Cullen, B. R. (2003). Sequence requirements for microRNA processing and function in human cells. *RNA* **9**(1), 112–123.

Zhang, H., Liu, L., Wang, X. & et al (2007). Guideline for data analysis of genomewide association studies. *Cancer genomics & proteomics* **4**(1), 27–34.

Ziebarth, J. D., Bhattacharya, A., Chen, A. & et al (2012). PolymiRTS Database 2.0: linking polymorphisms in microRNA target sites with human diseases and complex traits. *Nucleic Acids Research* **40**(D1), D216–D221.

Zisoulis, D. G., Kai, Z. S. & Chang, R. K. (2012). Autoregulation of microRNA biogenesis by let-7and Argonaute. *Nature* **486**, 541–544.