

Avaliação meta-classificatória de ferramentas de predição de alvos de
microRNAs e análise de enriquecimento funcional de alvos utilizando
Homo sapiens como modelo biológico

Arthur Casulli de Oliveira

Botucatu, SP

2017

UNIVERSIDADE ESTADUAL PAULISTA

“Julio de Mesquita Filho”

INSTITUTO DE BIOCIÊNCIAS DE BOTUCATU

Avaliação meta-classificatória de ferramentas de predição de alvos de
microRNAs e análise de enriquecimento funcional de alvos utilizando
Homo sapiens como modelo biológico

Candidato: Arthur Casulli de Oliveira

Orientador: Danilo Pinhal

Dissertação apresentada ao Instituto de
Biociências, Câmpus de Botucatu,
UNESP, para obtenção do título de
Mestre pelo Programa de Pós-Graduação
em Ciências Biológicas (Genética).

Botucatu, SP

2017

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉC. AQUIS. TRATAMENTO DA INFORM.
DIVISÃO TÉCNICA DE BIBLIOTECA E DOCUMENTAÇÃO - CÂMPUS DE BOTUCATU - UNESP
BIBLIOTECÁRIA RESPONSÁVEL: ROSEMEIRE APARECIDA VICENTE-CRB 8/5651

Oliveira, Arthur Casulli de.

Avaliação meta-classificatória de ferramentas de predição de alvos de microRNAs e análise de enriquecimento funcional de alvos utilizando Homo sapiens como modelo / Arthur Casulli de Oliveira. - Botucatu, 2017

Dissertação (mestrado) - Universidade Estadual Paulista "Júlio de Mesquita Filho", Instituto de Biociências de Botucatu

Orientador: Danillo Pinhal

Capes: 20205007

1. MicroRNAs. 2. Genética humana. 3. Bioinformática.
4. Regulação da expressão gênica.

Palavras-chave: Bioinformática; Genética humana; RNAs não-codificadores; Regulação gênica.

Dedico este trabalho aos meus pais e avós, que me forneceram todo carinho e apoio necessário no decorrer do meu mestrado e à minha namorada, que sempre esteve ao meu lado me apoiando durante todos estes anos.

Agradeceço:

Ao meu orientador, Prof. Dr. Danillo Pinhal, pelo suporte, orientação científica e amizade durante todos estes anos em que fui seu aluno e por sempre me incentivar a encarar novos desafios nesta jornada acadêmica.

À minha mãe, Fernanda Bressanelli Casulli, por todo amor e carinho e por sempre estar ao meu lado me apoiando e educando, nos momentos felizes e tristes da minha vida. Sem ela, não estaria perto de onde estou agora.

Aos meus avós, Ida Maria e Pilade, por todo amor, carinho e suporte que sempre me deram durante todos os anos da minha vida e, principalmente, durante anos de graduação e pós-graduação.

Ao meu irmão, Victor Casulli de Oliveira, por todos estes anos de amizade, brincadeiras e brigas e por sempre estar ao meu lado, nas risadas e nas tristezas.

Ao amor da minha vida, minha namorada Karina Gabriele Alves Dias (Nuros), por compartilhar comigo todos os momentos, bons e ruins, dos últimos quatro anos, pelos momentos divertidos que passamos juntos: jantares, parques, sessões de *Netflix*, cinema, etc e por tudo que o futuro ainda nos reserva. Te amo.

A todos os colegas do Laboratório Genômica e Evolução Molecular pelas conversas e discussões sobre os mais variados temas relacionados a pesquisa e ciência.

Aos amigos de trabalho Pedro (Batata), Marcos, e Luiz (Chokito) por todas as brincadeiras, risadas e colaborações realizadas durante todo o período do meu mestrado e por sempre me auxiliarem durante o desenvolvimento deste projeto.

Ao Prof. Dr. Ney Lemke, por todo suporte fornecido na área da computação e bioinformática, essenciais durante a realização deste projeto.

Ao Dr. Simon Moxon, pelo estágio realizado em Norwich, Inglaterra e por todo suporte e conhecimento transferido durante minha estadia no TGAC.

Aos meus grandes amigos e parceiros de vida, Julinho, Dox e Marsal, por todos os anos de grande amizade, jogatina e conversas e por estarem sempre comigo nos principais momentos da minha vida

Aos funcionários da seção de Pós-graduação por todas as dúvidas esclarecidas.

Ao Laboratório de Genômica e Evolução Animal, ao Departamento de Genética, à Pós-graduação em Genética, ao Instituto de Biociências de Botucatu e à Universidade Estadual Paulista pela estrutura cedida para a realização deste trabalho.

À CAPES pela bolsa de mestrado concedida no período de estudo.

A todas as pessoas que de alguma forma, direta ou indiretamente, auxiliaram na realização e finalização deste trabalho.

*Se, a princípio, a ideia não é absurda,
então não há esperança para ela.*

- Albert Einstein

Resumo

MicroRNAs (miRNAs) são pequenos RNAs não codificadores que regulam uma ampla gama de vias biológicas. Esta regulação ocorre através do pareamento complementar entre o miRNA e seu RNA mensageiro (mRNA) alvo, geralmente na região 3'UTR, inibindo a síntese proteica. Diversos trabalhos têm buscado determinar as funções biológicas desempenhadas pelos miRNAs por meio da identificação de seus alvos e posterior análise de enriquecimento funcional. Entretanto, as ferramentas de predição de alvos *in silico* disponíveis atualmente apresentam resultados pouco robustos e não há um consenso sobre a melhor ferramenta e estratégia para análise dos dados. Adicionalmente, a metodologia de enriquecimento funcional atual não leva em conta diversos fatores fundamentais atuantes na regulação dos alvos dos miRNAs, retornando resultados inconsistentes que culminam em experimentos de validação desnecessários e pouco específicos, com conseqüente desperdício de tempo e recursos. Desta maneira, o presente trabalho tem como objetivos (i) elaborar metodologia de predição de alvos com alta eficiência utilizando as ferramentas de bioinformática disponíveis e (ii) avaliar a regulação dos processos biológicos controlados pelos miRNAs através da análise de enriquecimento funcional, considerando o *fold-change* de seus mRNA alvo. Para tal, comparou-se as performances das três ferramentas de predição de alvos atualmente mais utilizadas (TargetScan, miRanda-mirSVR, e Pita), assim como testou-se todas as possibilidades de combinação dos dados gerados por cada ferramenta (uniões e/ou intersecções). A metodologia de união das ferramentas TargetScan + miRanda-mirSVR resultou na melhor performance, com o melhor balanço entre sensibilidade e especificidade. Posteriormente, dados de expressão de genes alvos obtidos por *microarray* após a superexpressão de onze miRNAs foram utilizados para as análises de enriquecimento funcional. Os alvos dos miRNAs foram agrupados manualmente em cinco *clusters* de acordo com seu *fold-change*. Os *clusters* foram então submetidos à análise de enriquecimento funcional. Os processos biológicos enriquecidos por esta análise foram distintos em cada *cluster*, sugerindo que os miRNAs regulam com intensidade semelhante genes associados a uma mesma função biológica, mas funções biológicas distintas são reguladas com intensidades diferentes. Os resultados obtidos neste projeto aprimoram significativamente a qualidade das análises *in silico* de predição de alvos, o que permitirá aos pesquisadores obterem resultados mais robustos durante a identificação de alvos dos miRNAs. Adicionalmente, a análise de enriquecimento funcional realizada sugere uma nova complexidade dos miRNAs, podendo justificar o fato de um único miRNA ser capaz de regular processos biológicos distintos com a especificidade demandada para cada processo dentro de um contexto celular.

Abstract

MicroRNAs (miRNAs) are short non-coding RNAs that regulates a wide range biological pathways. This regulation occurs by the complementary binding between miRNA and its target Messenger RNA (mRNA), mainly at 3'UTR region, blocking the protein synthesis. Several works tries to identify the biological functions that miRNAs are assign by detecting its mRNA targets and performing functional enrichment analysis using bioinformatic tools. However, *in silico* target prediction tools available nowadays often return little robust results and there is no consensus about a tool that highlights from the others or if combining the results from more than one tool improves the quality of the analysis. Moreover, the functional enrichment methodology used nowadays do not take in account several important aspects of the regulation of the miRNA targets, thus generates inconsistent results. This way, the objectives of this project are (i) to elaborate a target prediction method with high efficiency using the available tools and (ii) to evaluate the regulation of the biological process controlled by the miRNAs by functional analysis considering the fold-change levels of the target mRNAs. To do this, we compared the performances of the three most used target prediction tools (TargetScan, miRanda-mirSVR and Pita), as well as all combinatorial possibilities of these tools (unions and intersections). The union of TargetScan + miRanda-mirSVR returned the greatest performance, with the best balance between sensitivity e specificity. After, microarray data from gene expression after super-expression of eleven miRNAs were used for the functional enrichment analysis. The miRNA targets were grouped in five clusters according to their fold-change levels after the superexpression of the miRNAs. The clusters were individually submitted to functional enrichment analysis. The enriched biological process were distinct in each cluster, suggesting that miRNAs control genes assign with one function with similar intensity, but distinct biological process are controlled with distinct intensities. The results obtained in this project improved the quality of *in silico* target prediction analysis, which can help researchers obtaining results with more quality when performing miRNA target prediction. Moreover, the functional enrichment analysis suggests a new complexity of miRNAs, and could justify the fact of an unique miRNA be capable of control several biological process with the specificity required for each one within the cellular context.

Sumário

1. Introdução geral.....	11
1.1. A predição de alvos e seus principais atributos.....	14
1.1.1. Ferramentas de predição de alvos.....	16
1.2. O enriquecimento funcional dos alvos de miRNAs	18
1.2.1. Os problemas do enriquecimento funcional	19
2. Objetivos.....	21
3. Capítulo I: Análise de Predição de Alvos	22
3.1. Material e métodos	22
3.1.1. Resumo do workflow.....	22
3.1.2. Obtenção dos dados de predição de alvos	22
3.1.3. União e intersecção dos resultados das ferramentas.....	24
3.1.4. Cálculo da sensibilidade, especificidade, precisão e performance	24
3.1.5. Análise estatística	27
3.2. Resultados e discussão – Artigo Científico	28
4. Capítulo II: Análise de enriquecimento funcional.....	49
4.1. Material e Métodos.....	50
4.1.1. Resumo do Workflow.....	50
4.1.2. Obtenção dos dados de microarray.....	50
4.1.3. Agrupamento dos alvos em clusters de mRNA fold-change	50
4.1.4. Análise de enriquecimento funcional	51
4.1.5. Análise de conservação evolutiva.....	51
4.2. Resultados e discussão – Artigo Científico	53
5. Considerações finais	69
6. Referências bibliográficas	71

1. Introdução geral

MicroRNAs (miRNAs) são pequenos RNAs não-codificadores (~22 nucleotídeos) presentes no genoma de animais, plantas e, inclusive, vírus (Lee et al., 1993; Sunkar et al., 2005; Jia et al., 2008). Descobertos pioneiramente há mais de duas décadas em *Caenorhabditis elegans* (Lee et al., 1993), possuem atualmente a reconhecida importância de participar da regulação de uma vasta gama de processos biológicos, tais como diferenciação e proliferação celular, carcinogênese, resposta imune, morte celular, dentre outros (Ambros, 2004; Flynt et al., 2007, 2009; Liu e Olson, 2010; Shkumatava et al., 2009; Christodoulou et al., 2010; Takacs e Giraldez, 2011).

A via canônica da biogênese de um miRNA (Figura 1) tem início com a transcrição de uma longa molécula (~110 nucleotídeos) conhecida como miRNA primário (pri-miRNA) (Borchert et al., 2006). Essa molécula dobra-se em uma estrutura secundária em forma de grampo de cabelo (estrutura em *hairpin*), passando então a ser reconhecida pela enzima *Drosha*. A *Drosha* cliva o pri-miRNA na região caudal do *hairpin*, formando o miRNA precursor (pré-miRNA; ~70 nucleotídeos; Lee et al., 2006). O pré-miRNA é, então, exportado para o citoplasma através da proteína *exportina-5* (Lund et al., 2004) e processado pela enzima *Dicer*. Essa enzima cliva o pré-miRNA na região *loop*, formando uma molécula de RNA fita dupla (dsRNA) de aproximadamente 22 pares de bases. Proteínas da família argonauta associam-se com uma das fitas do dsRNA para formar o complexo de silenciamento induzido por RNA (RISC), dando origem ao miRNA maduro, ou canônico (Krutzfeldt et al., 2006), enquanto que a outra fita (miRNA*) pode ser degradada ou também se associar a um complexo RISC (Rand et al., 2005; Bang et al., 2014).

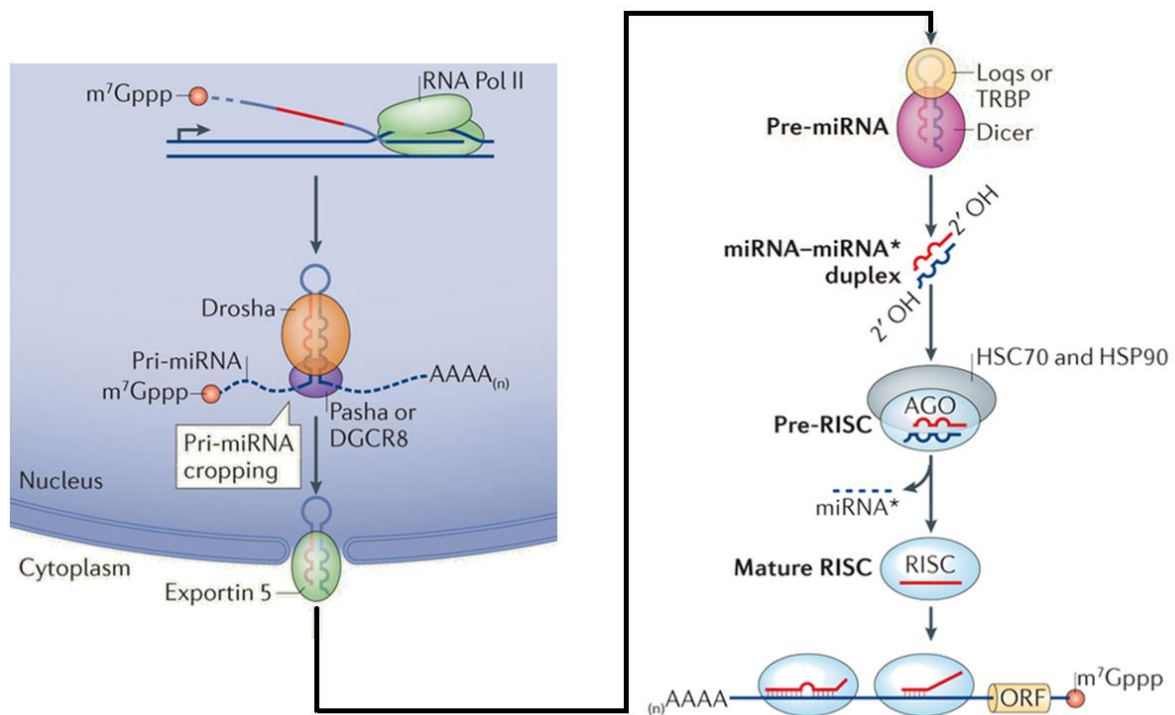


Figura 1: Via canônica da biogênese de miRNAs. Figura editada de Ameres e Zamore (2013).

Funcionalmente, os miRNAs atuam pelo pareamento simples de "Watson e Crick" com a sequência complementar presente na molécula do RNA mensageiro (mRNA) alvo. O miRNA maduro interage preferencialmente com a região 3'UTR do RNA mensageiro (mRNA) alvo por complementaridade total (nas plantas) ou parcial (nos animais) levando à inibição de sua síntese proteica (Lee e Dutta, 2009). Entretanto, adicionalmente à região 3'UTR, diversos trabalhos também detectaram sítios de interação em éxons (Tay et al., 2008; Reckzo et al., 2012; Schnall-Levin et al., 2010; Hausser et al., 2013) e na região 5'UTR (Lytle et al., 2007; Orom et al., 2008; Devlin et al., 2010; Zhou e Rigoutsos, 2014) da molécula de mRNA. Nas plantas, a alta complementaridade dos miRNAs tende a desencadear a clivagem do mRNA pelas proteínas da família argonauta (Tang et al., 2003; Lanet et al., 2009), enquanto que, a baixa complementaridade dos miRNAs com seu mRNA alvo, nos animais, geralmente não permite a clivagem do mRNA, sugerindo que neste grupo, a regulação da expressão gênica seja feita de maneira alternativa à clivagem (Karginov et al., 2010; Shin et al., 2010), embora haja exceções em ambos os grupos. Estudos realizados no *Danio rerio* (zebrafish) e em *Drosophila melanogaster* (drosófila) apontam que nos animais, os miRNAs

tendem a primeiramente inibir a tradução impedindo o acoplamento do ribossomo ao mRNA e posteriormente levar à degradação prematura deste mRNA (Bazzini et al., 2012; Mathonnet et al., 2007; Zdanowicz et al., 2009).

A interação de um miRNA com seu alvo é guiada principalmente por uma sequência de 7 nucleotídeos na região 5' do miRNA (nucleotídeos 2 a 8), chamada de sequência *seed*, embora importância cada vez maior esteja sendo atribuída ao pareamento 3' complementar nesse processo (Broughton et al., 2016). Cinco tipos de pareamentos da sequência *seed* são descritos atualmente: 8mer, 7mer-m8, 7mer-A1, 6mer e *offset*-6mer, apresentados na ordem do mais para o menos efetivo (Agarwal et al., 2015). O pareamento 8mer, se caracteriza pelo pareamento de sete nucleotídeos (2-8) com um "A" na posição "1" do 3'UTR. Estudos mostram que há uma preferência no reconhecimento de MREs que apresentam este nucleotídeo na posição "1" do 3'UTR, devido ao fato de este "A" ser uma região de assentamento das proteínas argonautas (Baek et al., 2008; Schirle et al., 2014). O pareamento 7mer-m8 representa um pareamento de sete nucleotídeos (2-8), porém sem a presença da "A" na posição "1". O pareamento 7mer-A1, é um pareamento de seis nucleotídeos (2-7), que contém um "A" na posição "1" do 3'UTR. O pareamento 6mer representa um pareamento de 6 nucleotídeos (2-7), enquanto que a *seed offset*-6mer caracteriza-se por um pareamento deslocado de 6 nucleotídeos (3-8), ambos sem a presença do "A" na posição "1". Entretanto, estes dois últimos tipos de pareamento apresentam baixa eficiência de regulação e são pouco conservados (Agarwal et al., 2015). Adicionalmente a estas interações baseadas na sequência *seed*, diversas outras interações não baseadas na *seed* foram detectadas (Clark et al., 2012; Clark et al., 2014; Chi et al., 2012), assim como foi demonstrada que a porção 3' dos miRNAs pode ser tão relevante quanto a região *seed* na detecção de seus alvos (Broughton et al., 2016).

Análises computacionais de predição de alvos de miRNAs indicam que um único miRNA pode ligar-se a centenas de mRNAs. Assim, cada miRNA regula uma gama extensa de processos biológicos distintos. Portanto, os miRNAs devem ser capazes de regular de forma específica cada processo biológico.

Desta maneira, a predição de genes alvo e a caracterização dos processos biológicos regulados são etapas fundamentais em diversas pesquisas envolvendo miRNAs e seus papéis biológicos. A predição de alvos de miRNAs atualmente é realizada através de ferramentas computacionais que avaliam diversos parâmetros envolvendo a interação miRNA-alvo. Já a caracterização dos processos biológicos regulados pelos miRNAs geralmente é feita através

de análises de enriquecimento funcional dos alvos preditos computacionalmente ou dos genes cuja expressão foi alterada através de técnicas de manipulação do miRNA estudado.

Entretanto, a predição computacional de alvos ainda está distante do ideal, pois as ferramentas exibem uma alta quantidade de interações falso-positivas ou são falhas em detectar interações genuínas. Adicionalmente, as análises de enriquecimento funcional dos alvos de miRNAs são realizadas sem que sejam considerados diversos fatores relevantes para que haja a interação miRNA-alvo (ex., pareamento da *seed*, acessibilidade do sítio de ligação, tamanho do 3'UTR). Desta maneira, diversas funções biológicas desempenhadas pelos miRNAs podem estar sendo equivocadamente propostas ou descartadas a priori.

1.1. A predição de alvos e seus principais atributos

Nos últimos anos, diversas propriedades importantes para o reconhecimento de um mRNA como alvo de um miRNA foram identificadas em animais, melhorando a compreensão destas interações e, conseqüentemente, aprimorando a predição computacional de alvos. Dentre todos os parâmetros descritos atualmente, alguns recebem maior destaque devido sua grande influência tanto na regulação quanto no reconhecimento de um mRNA como pleno alvo de um miRNA (Figura 2).

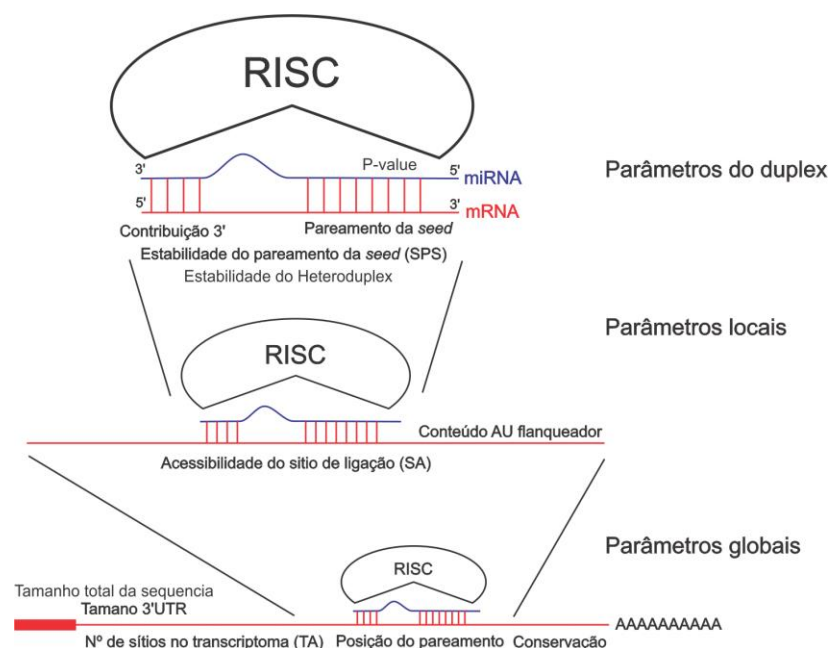


Figura 2: Principais parâmetros utilizados durante a predição de alvos. Figura editada de Betel et al. (2010).

Estes parâmetros podem ser divididas em três grandes grupos: os parâmetros do duplex, os parâmetros locais e os parâmetros globais (Betel et al., 2010). Os parâmetros do duplex contêm os parâmetros pareamento da *seed*, contribuição 3', estabilidade do pareamento da *seed* (*seed pairing stability* – SPS; Betel et al., 2010), energia livre do heteroduplex e P-value (Miranda et al., 2010). Estes parâmetros aferem a interação direta entre o miRNA e seus alvos. O pareamento da *seed* avalia quantos nucleotídeos da região *seed* do miRNA estão pareado com o mRNA alvo. A contribuição 3' avalia se além do pareamento da *seed* ocorre também um pareamento da porção 3' do miRNA e o quão ele auxilia na regulação (Witkos et al., 2011). O SPS avalia quais os nucleotídeos que compõe a sequencia *seed* (Garcia et al., 2008). A energia livre do heteroduplex avalia se a mínima energia livre formada entre o miRNA e o mRNA alvo é suficiente para estabelecer uma hibridização. Finalmente, o P-value avalia a probabilidade da interação miRNA-alvo ter sido predita de maneira aleatória.

Os parâmetros locais aferem propriedades da 3'UTR do mRNA com influência direta no reconhecimento deste como um alvo de um miRNA qualquer. Estes parâmetros perfazem a acessibilidade do sítio de ligação (*site accessibility* – SA) e o conteúdo AU flanqueador. A SA avalia a capacidade do miRNA em desdobrar a potencial estrutura secundária formada na região de interação do miRNA (Kertsz et al., 2007), conhecida como elemento de reconhecimento de miRNAs (*miRNA recognition element* – MRE). O conteúdo AU flanqueador avalia o número de nucleotídeos “A” e “U” que flanqueiam os MREs, uma vez que altas concentrações destes nucleotídeos nestas regiões aumentam a eficiência da regulação (Grimsom et al., 2007).

Os parâmetros globais aferem propriedades da 3'UTR do mRNA alvo com influência indireta no reconhecimento deste como um alvo. Estes parâmetros são o tamanho total da sequencia e do 3'UTR, o número de sítios de ligação no transcriptoma (*transcriptome abundance* – TA), a posição do pareamento e o grau de conservação do 3'UTR. O tamanho do 3'UTR é importante de ser avaliado, uma vez que 3'UTRs mais longos sofrem maior grau de regulação (Sandberg et al., 2008), já o tamanho total da sequencia é relevante pois há maior probabilidade de predições falsas em sequencias maiores (Miranda et al., 2006). O TA avalia a quantidade de sítios de ligação de um mesmo miRNA em todo o transcriptoma, uma vez que quanto mais alvos esse miRNA regular, maior vai ser a diluição de seu efeito. A posição

do pareamento avalia a posição do MRE no 3'UTR, uma vez que MREs localizados nas porções terminais do 3'UTR apresentam um maior potencial regulatório (Grimsom et al., 2007). A conservação avalia a conservação dos MREs entre as espécies, uma vez que miRNAs mais conservados tendem a apresentar maior potencial regulatório (Grimsom et al., 2007).

1.1.1. Ferramentas de predição de alvos

A partir dos avanços da bioinformática nos últimos anos, diversas ferramentas de predição foram elaboradas na tentativa de se otimizar a busca por interações de genes alvos relacionados às vias regulatórias diversas nas quais os miRNAs estejam atuando. Atualmente, dezenas de ferramentas encontram-se disponíveis. Dentre elas, quatro têm sido amplamente utilizadas pela comunidade científica: TargetScan, miRanda-mirSVR, Pita e RNA22. As ferramentas Targetscan, miRanda-mirSVR e Pita consideram predições baseadas na sequência *seed* e nas regiões 3'UTR, enquanto a ferramenta RNA22 também considera interações não baseadas na *seed* e em todo o transcrito.

Apesar de todas estas ferramentas terem como objetivo a identificação de uma ampla gama de interações miRNA-alvo genuínas, elas são constituídas de parâmetros distintos (Tabela 1), proporcionando resultados divergentes entre elas.

Contudo, mesmo essas ferramentas mais avançadas, ainda geram predições de alvo pouco robustas, pois retornam (i) uma alta quantidade de interações falso-positivas ou (ii) são falhas em detectar interações genuínas. Adicionalmente, como demonstrado, essas ferramentas utilizam uma série de parâmetros divergentes, o que produz resultados inconsistentes quando comparadas entre si. Este fato acentua-se principalmente nas predições de alvos de miRNAs de animais, devido à possibilidade de pareamento incompleto entre o miRNA e seu alvo.

Desta maneira, apesar da disponibilidade de uma série de ferramentas de predição de alvos não há um consenso sobre a melhor maneira de utilizá-las. De fato, diversos experimentos de validação revelaram muitos resultados falso positivos e falso negativos, demonstrando que ainda há necessidade de futuras melhoras nas ferramentas. Na tentativa de minimizar estes resultados pesquisadores tem usado diversas estratégias para selecionar os alvos preditos, incluindo a utilização da intersecção ou união dos resultados de diversas

ferramentas. Entretanto, estas metodologias vêm sendo usadas de maneira indiscriminada, sem um critério bem definido e um teste comparativo para determinar a qualidade de tais estratégias. Desta maneira, ainda não se sabe a intersecção ou união dos resultados de mais de uma ferramenta de fato melhora a qualidade das análises de predição de alvo.

Tabela 1: Principais parâmetros utilizados pelas ferramentas TargetScan, miRanda-mirSVR, PiTa e RNA22.

Grupos	Atributos	TargetScan	miRanda-miRSVR	PiTa	RNA22
Parâmetros do duplex	Pareamento <i>seed</i>	X	X	X	X
	Contribuição 3' SPS	X	X	X	X
	Energia livro do heteroduplex				X
	P-value				X
	SA	X	X	X	
Parâmetros Locais	Conteúdo AU flanqueador	X	X	X	
Parâmetros Globais	TA	X			
	Posição do pareamento	X	X		
	Tamanho do 3' UTR	X	X		
	Tamanho total da sequencia				X
	Conservação	X	X	X	
-	Outros	X			X

1.2. O enriquecimento funcional dos alvos de miRNAs

Após a obtenção dos resultados provenientes da predição de alvos dos miRNAs, uma das principais análises realizadas com estes dados visa à busca e identificação dos papéis e

vias biológicas que cada miRNA participa (Bleazard et al., 2015). O método mais utilizado nesta análise é o enriquecimento funcional *in silico* dos alvos regulados pelos miRNAs. Este método consiste em três etapas: (i) identificar os genes regulados pelos miRNAs analisados, (ii) associar estes alvos com suas funções biológicas e (iii) calcular a super-representação estatística dos processos biológicos dos alvos dos miRNAs (Gusev et al., 2007).

A primeira etapa é geralmente realizada com a utilização de ferramentas de predição de alvo ou de *datasets* experimentais que avaliaram interação mRNA-alvo. Uma vez que tais experimentos em larga escala da interação mRNA-alvo, como por exemplo *chips* de *microarray* e *CLIP-seqs*, ainda são escassos a abordagem de predição de alvos permanece como a mais utilizada. Durante esta etapa os pesquisadores podem optar por analisar os alvos provenientes de um único miRNA ou de uma lista de miRNAs que possuem uma característica em comum, como por exemplo serem enriquecidos em determinado tecido ou estarem sub/super-expressos em pacientes com determinada doença.

A segunda etapa é geralmente realizada utilizando-se as anotações do *Gene Ontology* (GO; Ashburner et al., 2000), ou as vias biológicas do *Kyoto Encyclopedia of Genes and Genome* (KEGG; Kanehisa and Goto, 2000). O GO é um consórcio que agrega termos funcionais dos genes de diversas espécies de animais, plantas e microorganismos, dividindo-os em três categorias: processos biológicos, componentes celulares e funções moleculares. Os termos associados aos processos biológicos se referem às vias nas quais o gene contribui, como por exemplo *generation of neurons* (GO:0048699) e *response to stress* (GO:0006950). Os termos associados aos componentes celulares referem-se às partes da célula ou ambiente extracelular em que os genes atuam, como por exemplo *cytoplasmic vesicle* (GO:0031410) e *synapse part* (GO:0044456). Os termos associados às funções moleculares referem-se às atividades bioquímicas do gene, como por exemplo *hydrolase activity* (GO:0016787) e *protein complex binding* (GO:0032403) (Ashburner et al., 2000).

O KEGG, assim como o GO, é um banco de dados que agrega termos funcionais de diversas espécies de animais, plantas e microorganismos. Entretanto, diferentemente do GO, o KEGG fornece um mapa de vias biológicas, agrupando os genes segundo as grandes vias biológicas das quais participam, por exemplo, *fatty acid metabolism* e *lipid metabolism* (Kanehisa and Goto, 2000), ao invés de agrupá-los por processos biológicos relacionados a eventos específicos.

A terceira etapa consiste no teste de distribuição hipergeométrica, ou o teste de Fisher, utilizado para o enriquecimento dos dados. Neste contexto, a distribuição hipergeométrica calcula a probabilidade de um miRNA regular n genes num determinado processo biológico dado um total de x genes presentes na amostra. Por essa estratégia é possível testar se os genes alvos de miRNAs são controlados aleatoriamente ou se estão preferencialmente associados a determinados processos biológicos (Bleazard et al., 2015).

1.2.1. Os problemas do enriquecimento funcional

Apesar de este ser o método mais utilizado para determinação dos processos e vias biológicas controladas pelos miRNAs, ele não leva em consideração diversos fatores relevantes durante a interação miRNA-alvo, como os demonstrados na Tabela 1.

Bleazard et al. (2015) questiona a eficiência do modo pelo qual o enriquecimento funcional dos alvos de miRNAs é atualmente realizado, alertando que esta abordagem possui diversos problemas e enviesamento metodológico. Estes autores discutem que esta abordagem gera diversos resultados inespecíficos, como por exemplo retorna processos biológicos enriquecidos semelhantes independentemente da lista de alvos utilizada. Isso ocorre, pois, uma vez que os genes podem estar associados a mais de uma função biológica, isto pode tendenciar o aparecimento de determinados processos biológicos gerais, como por exemplo *regulation of biological process* (GO:0050789), *single-organism process* (GO:0044699), *multicellular organismal process* (GO:0032501), dentre outros, em diversas listas independentes.

Na tentativa de atenuar estes problemas, estes e outros pesquisadores têm buscado identificar abordagens alternativas à metodologia padrão de enriquecimento funcional. Bleazard et al. (2015), por exemplo, elaborou um cálculo empírico que também leva em consideração o número de MREs presentes em cada 3'UTR. Entretanto, além desta, diversas outras características influenciam a regulação final dos alvos dos miRNAs. Ignorar tais características ou o potencial regulatório de um miRNA sobre seus alvos como um todo durante a etapa de enriquecimento pode ocultar diversos padrões regulatórios e propriedades biológicas recorrentes dos miRNAs. Em outras palavras, uma vez que um único miRNA é capaz de regular diversos processos biológicos dentro de um mesmo contexto celular, é provável que cada processo seja regulado de modo diferencial e particular e que os genes alvo

atuantes num mesmo processo biológico sejam regulados com intensidades semelhantes, diferentemente de genes atuantes em outros processos biológicos cuja intensidade de regulação não estaria correlacionada (aqui, nos definimos “intensidade de regulação” como o grau de fold-change apresentado pelos mRNA após a super-expressão dos miRNAs através de mimetizadores). Tal categorização regulatória hipotética não pode ser avaliada com o uso das metodologias atuais de enriquecimento funcional, porém caso verdadeira, ajudaria a justificar a capacidade dos miRNAs em regular vias biológicas distintas com alta especificidade, atendendo às demandas próprias de cada via, e assim promovendo a homeostasia celular.

2. Objetivos

2.1. Objetivo geral

Tendo em vista as problemáticas referentes (I) à falta de consenso sobre a melhor estratégia de predição de alvos e (II) ao viés metodológico das análises de enriquecimento funcional, os objetivos deste trabalho são:

- I) Elaborar metodologia de predição de alvos com alta eficiência utilizando as ferramentas disponíveis;
- II) Avaliar a regulação dos processos biológicos controlados pelos miRNAs de acordo com a intensidade da interação miRNA-alvo.

2.2. Objetivos específicos

- I-1) Avaliar a especificidade, sensibilidade, precisão e performance das ferramentas de predição de alvos TargetScan, miRanda-mirSVR, Pita e RNA22;
- I-2) Comparar a performance da utilização da intersecção e união dos resultados destas ferramentas;
- I-3) Identificar a abordagem com a melhor performance, produzindo uma metodologia de melhor qualidade;
- II-1) Avaliar a ocorrência do possível padrão regulatório determinado pela intensidade da regulação dos mRNA alvos;
- II-2) Analisar os processos biológicos regulados por três dos miRNAs identificados através da técnica de enriquecimento funcional na qual levou-se em consideração o padrão regulatório identificado.

----- **3. Capítulo I:**
Análise de Predição de Alvos

3.1. Material e métodos

3.1.1. Resumo do workflow

Para as análises de predição de alvos dos miRNAs, foram obtidos os dados de predição de alvos de quatro ferramentas computacionais. Posteriormente, foi realizada a união e intersecção dos resultados destas ferramentas para comparação de diversos parâmetros de qualidade (veja figura 3 para um resumo completo das atividades realizadas).

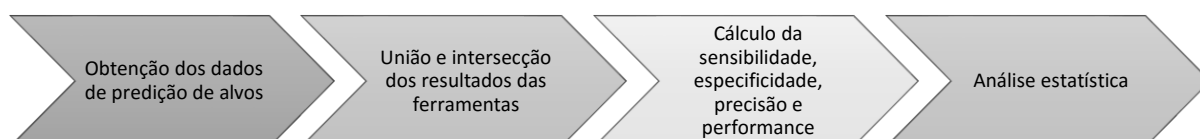


Figura 3: Fluxograma apresentando resumo das atividades realizadas referentes as análises de predição de alvos.

3.1.2. Obtenção dos dados de predição de alvos

Para esta análise, forem utilizadas as predições disponíveis para download pelas ferramentas TargetScan (TS), miRanda-mirSVR (MR), Pita (PT) e RNA22 (R22). A seleção destas ferramentas foi realizada considerando a reconhecida qualidade de seus dados e sua política de atualizações. Apenas as melhores predições consideradas por cada ferramenta foram utilizadas, visando comparar os melhores dados de cada uma (Tabela 2).

Table 2: Resumo das ferramentas de predições de alvos analisadas

	TargetScan	miRanda-mirSVR	Pita	RNA22
Website	targetscan.org	microrna.org	genie.weizmann.ac.il	https://cm.jefferson.edu/rna22/
Version	v7.1 (06/2016)	V3.3a (08/2010)	V6 (08/2008)	V2 (04/2015)
Predictions downloaded	Conserved sites	Good mirSVR score, Conserved	Seed 7- or 8-mer and conservation score 0.9 or higher	Base pair: > 12 Folding energy: <= -12kcal/mol

	miRNA			p-value: ≤ -0.1 miRbase21/Ensembl78
Reference	Lewis et al., 2005	Enright et al., 2003	Kertesz et al., 2007	Miranda et al., 2006

Para o TS as melhores predições são as classificadas como alvos conservados de miRNAs conservados. O TS utiliza diferentes cortes para considerar um sítio como conservado, de acordo com o tipo de *seed* que estes sítios possuem: 8mer ≥ 0.8 , 7mer-m8 ≥ 1.3 , 7mer-A1 ≥ 1.6 e sítios 6mer e *offset* 6-mer são sempre considerados não-conservados (http://www.targetscan.org/faqs.Release_7.html). Para o miRanda-mirSVR, as melhores predições são as classificadas como conservadas com mirSVR *score* ≤ -0.1 . O miRanda-mirSVR considera como conservada as interações com *PhastCOons score* > 0.57 (Betel et al., 2008). Pita, as melhores predições são aquelas com *seed match* 7- ou 8mer e *score* de conservação 0.9 ou maior (https://genie.weizmann.ac.il/pubs/mir07/mir07_data.html). Para o RNA22, as melhores predições são aquelas com pelo menos 12 nucleotídeos pareados, mínima energia livre do heteroduplex de -12 kcal/mol e máximo P-value 0.1.

3.1.3. União e intersecção dos resultados das ferramentas

Os nomes dos genes preditos foram convertidos para os termos do Ensembl Gene ID, para padronizar as anotações de todas as ferramentas. Finalmente, os resultados das ferramentas foram combinados utilizando as técnicas de união e intersecção. As uniões utilizadas foram TS + MR + PT + R22, TS + MR + PT, TS + MR + R22, TS + PT + R22, MR + PT + R22, TS + MR, TS + PT, TS + R22, MR + PT, MR + R22 e PT + R22. As intersecções utilizadas foram TS + MR + PT + R22, TS + MR + PT, TS + MR + R22, TS + PT + R22, MR + PT + R22, TS + MR, TS + PT, TS + R22, MR + PT, MR + R22, PT + R22 e majority vote. Esta última consiste na seleção de qualquer alvo predito por pelo menos duas das quatro ferramentas.

3.1.4 Cálculo da sensibilidade, especificidade, precisão e performance

Primeiramente, foi realizado o download dos dados de interações miRNA-alvo validadas para o genoma humano do banco de dados miRTarBase

(<http://mirtarbase.mbc.nctu.edu.tw>; v6 – 09/2015; Chou et al., 2016). Etão, foram selecionados os dez miRNAs com o maior número de alvos validados. Para tal, foram consideradas apenas as validações caracterizadas como *Strong* pelo miRTarBase (*Reporter assays, Western blot e/ou qPCR*). Desta maneira, os miRNAs utilizados nesta análise foram: miR-155-5p (224 alvos validados), miR-145-5p (129 alvos validados), miR-21-5p (115 alvos validados), miR-34a-5p (101 alvos validados), miR-29a-3p (96 alvos validados), miR-125b-5p (83 alvos validados), miR-124-3p (83 alvos validados), miR-24-3p (83 alvos validados), miR-17-5p (74 alvos validados), e miR-1-3p (73 alvos validados).

Posteriormente, foi calculada a sensibilidade, especificidade, precisão e performance (através do cálculo do Coeficiente de Correlação de Matthews (Matthews Correlation Coefficient – MCC)) de cada ferramenta e método de combinação de resultados (Figura 4).

A)

		Condição Preditada		
		Condição Positiva Preditada	Condição Negativa Preditada	
Condição Real	Condição Positiva	Verdadeiro Positivo (VP)	Falso Negativo (FN)	Sensibilidade
	Condição Negativa	Falso Positivo (FP)	Verdadeiro Negativo (VN)	Especificidade
		Precisão		Performance (MCC)

B)

$$(1) \textit{Sensitivity} = \frac{VP}{VP+FN}; (2) \textit{Specificity} = \frac{VN}{VN+FP}$$

$$(3) \textit{Precision} = \frac{VP}{VP+FP}; (4) \textit{MCC} = \frac{VP*VN-FP*FN}{\sqrt{(VP+FP)(VP+FN)(VN+FP)(VN+FN)}}$$

Figura 4: Cálculo da sensibilidade, especificidade, precisão e performance das estratégias de predição. A) Matriz de confusão caracterizando as condições dos alvos preditos. B) Fórmulas utilizadas para os cálculos.

Neste cálculo verdadeiro positivo (VP) é o número de alvos validados que foram preditos, falso negativo (FN) é o número de alvos validados não preditos, falso positivo (FP) é o número de alvos preditos que não foram validados e verdadeiro negativo (VN) é o número de genes que não foram preditos nem validados.

A sensibilidade e a especificidade são funções matemáticas que medem a qualidade de classificações binárias. Uma vez que predições *in silico* de alvo de miRNAs são classificadores binários (alvo/não alvo), estas funções podem ser utilizadas para medir a qualidade das predições. No caso do trabalho apresentado, a sensibilidade mede a capacidade da ferramenta em corretamente identificar alvos verdadeiros, enquanto a especificidade mede a capacidade da ferramenta em corretamente excluir genes que não são regulados pelos miRNAs (Parikh et al., 2008). Adicionalmente, a precisão calcula quantas predições erradas estão presentes dentro do total de genes preditos como alvo (Powers et al., 2007). Finalmente, o MCC combina todas estas medidas para fornecer um único número que consegue ser comparado entre as ferramentas e metodologias. O MCC é uma medida reconhecida que calcula a qualidade de classificadores binários e usualmente é utilizado para classificar ferramentas de predição de alvos (Bandyopadhyay and Mitra, 2009; Fan and Kurgan et al., 2015).

Os valores de sensibilidade e especificidade variam de 0 a 1, nos quais valores próximos a zero indicam baixa qualidade, enquanto valores próximos a um indicam alta qualidade. Já o MCC apresenta valores de -1 a 1, nos quais representam baixa e alta eficiência, respectivamente. Além destes, valores próximos a zero indicam predições de qualidade idêntica às predições aleatórias. Para calcular estes valores, foram selecionados aleatoriamente 70 alvos validados e 70 alvos não validados para cada miRNA, para os conjuntos de dados verdadeiros e falsos, respectivamente, com cinco replicatas para cada. Desta maneira, foram avaliados um total de 1400 genes (700 verdadeiros e 700 falsos) para cada replicata. Finalmente, a média de cada replicata foi calculada e submetida à análise estatística.

Para confirmar se as especificidades das ferramentas não estão enviesadas pela falta de dados na literatura de predições de alvos confirmadas como errôneas, foi realizado um teste controle, no qual realizou-se a predição de alvos em sequências aleatórias. Desma maneira, foram quatro grupos com 1.000 sequencias aleatórias cada, totalizando 4.000 sequencias. Os grupos variaram em tamanho (500, 1.000, 2500 e 5.000 nucleotídeos) uma vez que o tamanho da sequencia influencia nas chances de ocorrer predições falso-positivas (Miranda et al., 2006). Desta maneira, foi feito o download do código fonte de cada ferramenta e realizou-se a predição de alvos dos dez miRNAs utilizando-se os mesmos parâmetros descritos para os dados pre-computados (Tabela 2), com excessão para os valores de conservação para o

TargetScan, miRanda e Pita e o *score* mirRVS para o miRanda (uma vez que esse score não está presente em seu código fonte).

3.1.5. Análise estatística

Para comparar as performances de cada ferramenta e método de combinação foi utilizado o teste one-way ANOVA juntamente com o teste de Tuckey para comparações múltiplas (P-value < 0.05), uma vez que os dados apresentaram distribuição normal.

3.2. Resultados e Discussão

Os resultados e a discussão estão apresentados a seguir no formato de artigo científico, em revisão na revista *Frontiers in Genetics*:

Combining results from distinct microRNA target prediction tools enhances the performance of analyses

Arthur C. Oliveira¹, Luiz A. Bovolenta², Pedro G. Nachtigall¹, Marcos E. Herkenhoff¹, Ney Lemke² and Danilo Pinhal¹

¹Laboratory of Genomics and Molecular Evolution, Institute of Biosciences of Botucatu, Department of Genetics, Sao Paulo State University - UNESP, Botucatu, Brazil

²Laboratory of Bioinformatics and Computational Biophysics, Institute of Biosciences of Botucatu, Department of Physics and Biophysics, Sao Paulo State University - UNESP, Botucatu, Brazil

Keywords: in silico prediction, TargetScan, miRanda-mirSVR, Pita, RNA22, non-coding RNA, bioinformatics

Abstract

Target prediction is generally the first step towards recognition of bona fide microRNA-target interactions in living cells. Several target prediction tools are now available, which use distinct criteria and stringency to provide the best set of candidate targets for a single microRNA or a subset of microRNAs. However, there are many false-negative predictions, and consensus about the optimum strategy to select and use the output information provided by the target prediction tools is lacking. We compared the performance of four tools cited in literature—TargetScan, miRanda-mirSVR, Pita, and RNA22, and we determined the most effective approach for analyzing target prediction data (individual, union or intersection). For this purpose, we calculated the sensitivity, specificity, precision, and correlation of these approaches using 10 microRNAs (miR-1-3p, miR-17-5p, miR-21-5p,

miR-24-3p, miR-29a-3p, miR-34a-5p, miR-124-3p, miR-125b-5p, miR-145-5p, and miR-155-5p) and 1,400 genes (700 validated and 700 non-validated) as targets of these microRNAs. The four tools provided a subset of high-quality predictions and returned few false-positive predictions; however, they could not identify several known true targets. We demonstrate that union of TargetScan/miRanda-mirSVR and TargetScan/miRanda-mirSVR/RNA22 enhanced the quality of *in silico* prediction analysis of microRNA targets. We conclude that the union rather than the intersection of the aforementioned tools is the best strategy for maximizing performance while minimizing the loss of time and resources in subsequent *in vivo* and *in vitro* experiments for functional validation of microRNA-target interactions.

1. Introduction

MicroRNAs (miRNAs) are a large class of small non-coding RNAs (~22 nucleotides) that post-transcriptionally regulate gene expression. They were first identified in the context of *Caenorhabditis elegans* development (Lee et al., 1993), and they are now known to regulate most biological process in animals, plants, and even certain viruses (Lee et al., 1993; Sunkar et al., 2005; Jia et al., 2008). Their function ranges from cellular proliferation and differentiation to response to environmental stimuli and diseases such as cancer (Shenoy and Billeloch, 2014; Qiu et al., 2012; Reddy, 2015). Therefore, identification of their target genes is important for understanding their role in the complex biological regulatory pathways regulated by miRNA-target interactions.

In animals, a sequence of approximately seven nucleotides (nts) in the 5' region of the miRNA (ranging from nts 2-8), known as the seed region, guides the miRNA to its target mRNA. Five types of perfect Watson-Crick pairing of seed matches have been described so far, namely, 8-mer, 7-mer-m8, 7-mer-A1, 6-mer, and offset-6-mer in the descending order of the strength of their matches (Agarwal et al., 2015). The 8-mer site is a perfect match for nts 2-8, with an adenine at relative nt 1 in the mRNA. The 7-mer-m8 is a perfect match for nts 2-8, whereas the 7-mer-A1 is a perfect match for nts 2-7, with an adenine at relative nt 1 in the mRNA. The weaker 6-mer and offset-6-mer are perfect matches for nts 2-7 and 3-8, respectively. The adenosine at relative nt position 1 of the mRNA supports the miRNA-mediated regulation, even if the opposing nt does not form a Watson-Crick pairing (Baek et al., 2008). In addition to the seed-based interactions, recent studies also reported miRNA

regulation through non-seed interactions, demonstrating that the 3' region of the miRNA transcript might be equally important as the seed sequence for securing target recognition (Tay et al., 2008; Nelson et al., 2011; Clarke et al., 2012; Chi et al., 2012; Broughton et al., 2016).

Irrespective of seed or non-seed match, miRNA pairing is largely prevalent with elements at the 3' untranslated region (UTR) of target genes. However, studies have identified miRNA pairing to sites outside the 3'UTR, both in the coding region (Tay et al., 2008; Schnall-Levin et al., 2010; Gartner et al., 2013; Hausser et al., 2013) and in the 5'UTR (Lytle et al., 2007; Orom et al., 2008; Devlin et al., 2010; Zhou et al., 2014) of the mRNA. Such findings showed that although the 3'UTR is the main site of miRNA pairing, the whole mRNA transcript should be inspected when predicting miRNA-target interactions.

Currently, several *in silico* tools are available for identifying putative miRNA targets. The main parameters used by these tools can be gathered and divided into three groups: duplex features, local context features, and global context features (Betel et al., 2010). Duplex features encompass seed match, 3' contribution, seed pairing stability (SPS; Betel et al., 2010), heteroduplex free energy, and p-value (Miranda et al., 2006). These parameters evaluate the hybridization of the miRNA to its target gene. Seed match evaluates the number of nts that can bind to the mRNA target in the seed region. The 3' contribution evaluates the possibility of binding at the 3' position of the miRNA (Witkos et al., 2011). The SPS evaluates the types of nts compose the seed region (Garcia et al., 2011). The heteroduplex free energy evaluates whether the minimum free energy between the miRNA and its target is sufficient to establish hybridization, and the p-value evaluates whether the probability of a selected interaction has been predicted by chance.

Local context features include mRNA sequence properties that directly influence target recognition, such as site accessibility (SA) and presence of flanking AU. SA evaluates the capacity of the mRNA to unfold into a potential secondary structure in the region containing the miRNA cognate sequence, which is known as the miRNA recognition element (MRE) (Kertesz et al., 2007). The flanking AU corresponds to the number of A and U nts flanking the MRE region. High concentrations of flanking A and U nts enhance miRNA regulation (Grimsom et al., 2007).

Global context features aggregate mRNA sequence properties with indirect influence on target recognition, such as whole transcript length, 3'UTR length, transcriptome abundance, pairing position at the 3'UTR, and sequence conservation. Sequence length evaluates the total length of the string analyzed, since the chances of false prediction increases with target length (Miranda et al., 2006). The 3'UTR length, as the name suggests, evaluates the length of the 3'UTR of the potential miRNA targets, since larger 3'UTRs are regulated more stringently than shorter ones (Sandberg et al., 2008). Transcriptome abundance evaluates the number of MREs of a miRNA within the transcriptome. Pairing position evaluates the position of the MRE within the 3'UTR, because MREs near the ends of the 3'UTR have stronger regulatory potential (Grimsom et al., 2007). Finally, sequence conservation evaluates the extent of conservation of the MREs among species. Together, all these binding metrics decisively regulate the determination of potential miRNA-target pairs.

Despite the availability of several target prediction tools that use distinct parameters and strategies to search for putative targets, consensus about the best tool is lacking. In fact, experimental validation (the usual step after target prediction) has revealed many false-negative predictions, implying that further improvement of prediction tools is required. To circumvent this caveat, researchers use diverse strategies for determining putative miRNA targets, including intersection and union of predictions. However, this approach is being used indiscriminately, without well-defined criteria and rigorous comparative tests to assess the performance of the prediction strategies. Thus, whether union or intersection of results obtained from multiple tools improves the overall quality of target prediction is yet unknown.

Here, we compared the performance of four widely used target prediction tools to identify the strategy that best predicts miRNA targets. Our results would assist researchers in selecting the correct candidates for subsequent experimental validation of miRNA-target interactions.

2. Material and Methods

2.1. Target prediction tools data

We used TargetScan (TS), miRanda-mirSVR (MR), Pita (PT), and RNA22 (R22) pre-computed predictions, which are freely available online. TS, MR, and PT consider seed-based

interactions in the 3'UTR, whereas R22 also considers non-seed based interactions (full-length matches) in the whole transcript.

These tools were selected based on their recognized popularity among researchers and the presence of an update policy (i.e., data is updated when new miRNAs and/or parameters are reported). We exclusively used the best predictions from each database to maximize the quality of predictions (summarized in Table 1).

Table 1: Summary of the target prediction tools analyzed.

	TargetScan	miRanda-mirSVR	Pita	RNA22
Website	targetscan.org	microrna.org	genie.weizmann.ac.il	https://cm.jefferson.edu/rna22/
Version	v7.1 (06/2016)	V3.3a (08/2010)	V6 (08/2008)	V2 (04/2015)
Predictions downloaded	Conserved sites	Good mirSVR score, Conserved miRNA	Seed 7- or 8-mer and conservation score 0.9 or higher	Base pair: > 12 Folding energy: <= -12kcal/mol p-value: <= -0.1 miRbase21/Ensembl78
Reference	Lewis et al., 2005	Enright et al., 2003	Kertesz et al., 2007	Miranda et al., 2006

In detail, the best predictions were those with conserved sites for TS. TS considers different cutoffs for conservation, according to seed match; for example, it is ≥ 0.8 for site 8-mer, ≥ 1.3 for site 7-mer-m8, and ≥ 1.6 for site 7-mer-A1, whereas sites 6-mer and offset 6-mer are always classified as non-conserved (http://www.targetscan.org/faqs.Release_7.html). Best predictions of MR present good mirSVR score (≤ -0.1) and conserved sites (PhastCons score > 0.57 ; Betel et al., 2008). PT ranks those with seed match to 7- or 8-mer and conservation score ≥ 0.9 (https://genie.weizmann.ac.il/pubs/mir07/mir07_data.html) as best predictions, whereas R22 best predictions comprise those with base pair minimum value of 12, folding energy max value of -12 kcal/mol, max p-value of 0.1, and miRbase 21/Ensembl 78 databases.

Gene names predicted were converted to the Ensembl gene ID to standardize the annotations from all tools. We also combined the outputs of the tools to evaluate union and intersection approaches. The unions tested were TS + MR + PT + R22, TS + MR + PT, TS +

MR + R22, TS + PT + R22, MR + PT + R22, TS + MR, TS + PT, TS + R22, MR + PT, MR + R22, and PT + R22. The intersections tested were TS + MR + PT + R22, TS + MR + PT, TS + MR + R22, TS + PT + R22, MR + PT + R22, TS + MR, TS + PT, TS + R22, MR + PT, MR + R22, PT + R22, and majority vote. The majority vote consists of counting any target that was predicted by at least two of the four tools.

2.2. Performance evaluation

In order to evaluate the performance of each tool and the combinatorial method, we downloaded the validated miRNA target dataset for the human genome from miRTarBase (<http://mirtarbase.mbc.nctu.edu.tw>; v6 – 09/2015; Chou et al., 2016). Then, we selected 10 miRNAs with the highest number of validated targets, including miR-155-5p (224 validated targets), miR-145-5p (129 validated targets), miR-21-5p (115 validated targets), miR-34a-5p (101 validated targets), miR-29a-3p (96 validated targets), miR-125b-5p (83 validated targets), miR-124-3p (83 validated targets), miR-24-3p (83 validated targets), miR-17-5p (74 validated targets), and miR-1-3p (73 validated targets). This analysis was limited to these 10 miRNAs due to the few number of validated targets available to the other miRNAs, which inclusion would prejudice the power of the statistical analysis. We analyzed only "strong validations" assigned by miRTarBase, which refer to miRNA-target interactions validated using reporter assays, western blot and/or quantitative polymerase chain reaction (qPCR). We did not include "less strong validations", such as those reported using microarray, pSILAC, and next generation sequencing (NGS)-based experiments (e.g., Ago HITS-CLIP, degradome-seq, CLASH, PAR-CLIP and iPAR-CLIP) to enforce maximum stringency.

We calculated the sensitivity, specificity, precision, and performance of each target prediction tool and their combinations. The performance was calculated using Matthews correlation coefficient (MCC):

$$(1) \textit{Sensitivity} = \frac{TP}{TP+FN}$$

$$(2) \textit{Specificity} = \frac{TN}{TN+FP}$$

$$(3) \textit{Precision} = \frac{TP}{TP+FP}$$

$$(4) \textit{MCC} = \frac{TP*TN-FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

where TP (true positive) is the number of validated targets predicted, FN (false negative) is the number of validated targets not predicted, FP (false positive) is the number of predicted targets that were not validated, and TN (true negative) is the number of genes that were neither predicted nor validated. Sensitivity and specificity are mathematical functions that measure the quality of binary classifications. Since *in silico* target prediction tools are binary classifiers, these two functions can be used to evaluate the quality of each tool. Sensitivity measures a tool's ability to identify bona fide miRNA targets, while specificity measures the capacity of the tool to correctly exclude a gene target that is not regulated by the miRNA (Parikh et al., 2008).

Knowledge about the proportion of true predictions within the total number of miRNA targets predicted is also important. Therefore, precision is calculated to evaluate the number of true targets among all predicted targets (Powers et al., 2007). Finally, MCC can combine all these values to generate a unique comparable number. MCC is a recognized measure that is used to evaluate the quality of binary classifiers (*i.e.*, true targets/false targets), and it is often used to classify miRNA target prediction tools (Bandyopadhyay and Mitra, 2009; Fan and Kurgan et al., 2015).

The sensitivity, specificity, and precision values range from 0 to 1, with near zero values indicating low quality results and values near one representing high quality results. MCC ranges from -1 to 1, which represent low quality and high quality predictions, respectively. Values near zero indicate predictions that are similar to random predictions. To calculate these values, we randomly selected 70 validated targets as the true set and 70 non-validated genes as the negative set for each miRNA, with reposition. This generated 1,400 genes (700 true and 700 false) for each replicate (N = 5). Finally, the average of the five replicates was calculated and subjected to statistical analysis (see supplementary Table 1 for individual values from each replicate and each miRNA).

To confirm whether the specificity values of the tools were biased due to the lack of false predictions in literature, we performed a control test by predicting putative targets in a random strings analysis. Towards this objective, we generated four different groups with 1,000 random strings each, totaling to 4,000 random strings. Groups of variable length (500, 1,000, 2,500, and 5,000 nts each) were tested because length highly influences the chances of

false prediction (Miranda et al., 2006). Then, we downloaded the source code of each tool and locally ran the predictions of the ten miRNAs with the same parameters used for the best predictions of the pre-computed data, with the exception of the "conservation score" for TS, MR, and PT, and the "mirSVR score" for MR, which was not available on the miRanda source code (Table 1).

2.3. Statistical analysis

To compare the performance of each tool and the combinatorial method, we used one-way analysis of variance (ANOVA) and the Tukey test for multiple comparisons (p -value < 0.05) since the data presented a Gaussian distribution.

3. Results

3.1. Target prediction outputs

Each target prediction tool noticeably generated different results. TargetScan and miRanda-mirSVR had the highest number of mutual targets (310 predicted and 303 validated). The number of targets predicted by TS, MR, and PT (99 validated and 2 non-validated) was equivalent to the number predicted by all tools together (97 validated and 2 non-validated) (Figure 1a). MiRanda-mirSVR itself predicted the highest number of targets among all the tools, of which 433 were validated and 33 were non-validated. Pita predicted the lowest number (234 validated and 6 non-validated), Targetscan predicted 366 validated and 11 non-validated targets, and RNA22 predicted 325 validated and 60 non-validated. RNA22 predicted the highest number of validated targets not identified by any other tool (81; with 60 non-validated), followed by miRanda-mirSVR (61 validated and 15 non-validated), TargetScan (28 validated and 1 non-validated), and Pita (2 validated and 0 non-validated). Interestingly, of the 81 validated targets predicted exclusively by RNA22, 56 possessed non-canonical sites, 34 of which had sites only outside the 3'UTR (either seed-based or full-length), and R22 targets with sites inside the 3'UTR but with a mismatch in the seed region.

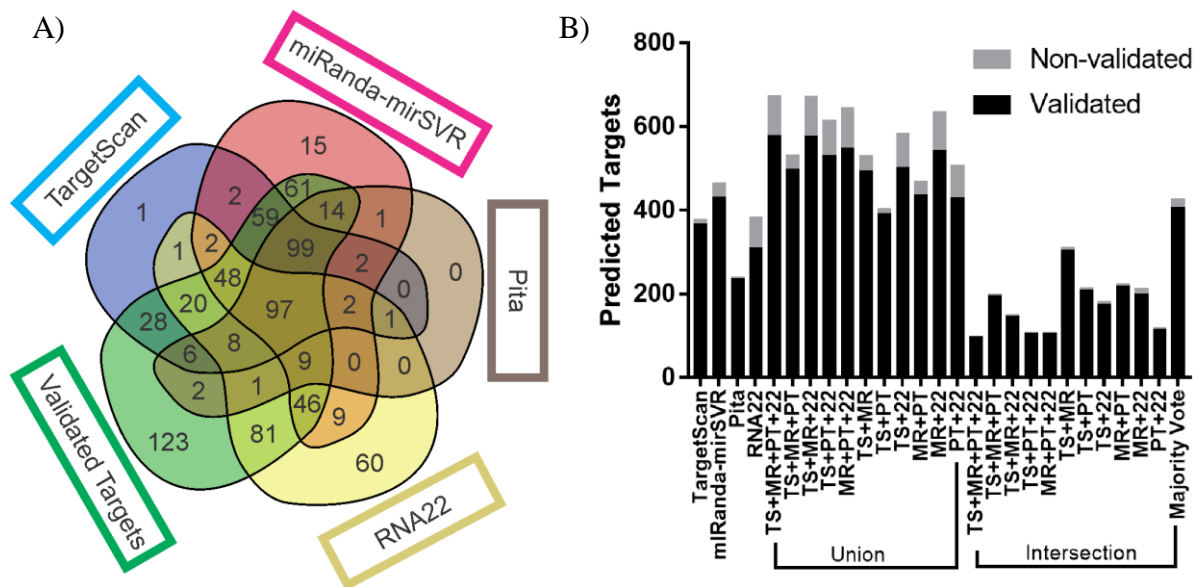


Figure 1: Target prediction output comparison. A) Venn diagram of the number of validated and non-validated targets predicted by each tool, as well as the number targets that were not predicted. B) Total number of validated and non-validated targets predicted by each tool and combinatorial approach. Venn diagrams from all replicates are available at Supplementary Data Sheet 1.

The majority of intersections consistently returned a lower number of predicted targets than any other approach, with the exception of the intersection of TS + MR (304 validated and 7 non-validated) and majority vote (406 validated and 19 non-validated), which predicted more targets than Pita. All the unions predicted more targets than any other approach (Figure 1b), except for TS + PT (390 validated and 12 non-validated), which predicted less targets than miRanda-mirSVR and majority vote.

The four tools were able to recover much more true predictions than false predictions (366 validated and 11 non-validated for TargetScan; 433 validated and 33 non-validated for miRanda-mirSVR; 234 validated and 6 non-validated for Pita; 325 validated and 60 non-validated for RNA22). However, approximately 18% of the validated targets (123) were not predicted by any tool. Supplementary Table 2 shows the predicted targets of the 10 miRNAs by each tool.

3.2. Sensitivity, specificity, and precision of the methods

Table 2 summarizes the sensitivity, specificity, and precision of all tools and methods. All methods showed striking specificity (> 0.85) and precision (> 0.80), but variable sensitivity. Considering the four tools individually, miRanda-mirSVR showed the highest

sensitivity (0.62) and RNA22 showed the lowest specificity (0.89) and precision (0.81). TargetScan and Pita showed similar values of specificity and precision, but Pita showed a significantly lower sensitivity.

Table 2: Sensitivity, specificity and precision of the target prediction methods.

Method	Tool	Sensitivity	Specificity	Precision
Individual tool	TargetScan	0.524±0.004	0.984±0.005	0.971±0.004
	miRanda-mirSVR	0.617±0.012	0.954±0.006	0.930±0.010
	Pita	0.336±0.009	0.992±0.004	0.977±0.011
	RNA22	0.336±0.009	0.893±0.019	0.805±0.027
Union	TS+MR+PT+22	0.825±0.008	0.862±0.020	0.857±0.017
	TS+MR+PT	0.710±0.006	0.949±0.007	0.933±0.009
	TS+MR+22	0.822±0.007	0.862±0.020	0.857±0.017
	TS+PT+22	0.757±0.038	0.879±0.028	0.863±0.023
	MR+PT+22	0.784±0.006	0.865±0.019	0.853±0.018
	TS+MR	0.706±0.006	0.949±0.008	0.932±0.009
	TS+PT	0.558±0.007	0.983±0.005	0.970±0.009
	TS+22	0.716±0.004	0.885±0.21	0.862±0.21
	MR+PT	0.624±0.016	0.954±0.006	0.931±0.010
	MR+22	0.773±0.008	0.865±0.019	0.851±0.019
Intersection	PT+22	0.613±0.005	0.889±0.020	0.847±0.023
	TS+MR+PT+22	0.139±0.006	0.998±0.003	0.984±0.018
	TS+MR+PT	0.279±0.014	0.994±0.003	0.980±0.010
	TS+MR+22	0.201±0.007	0.995±0.005	0.976±0.022
	TS+PT+22	0.150±0.006	0.997±0.002	0.976±0.016
	MR+PT+22	0.151±0.005	0.997±0.003	0.978±0.018
	TS+MR	0.435±0.013	0.989±0.004	0.976±0.009
	TS+PT	0.298±0.012	0.993±0.003	0.978±0.009
	TS+22	0.249±0.006	0.992±0.004	0.969±0.014
	MR+PT	0.312±0.013	0.993±0.003	0.977±0.010

MR+22	0.285±0.008	0.981±0.001	0.940±0.005
PT+22	0.164±0.004	0.996±0.003	0.974±0.018
Majority Vote	0.581±0.11	0.973±0.005	0.955±0.008

The union of the four tools undoubtedly returned the best sensitivity, with TS + MR + PT + R22 and TS + MR + R22 returning values above 0.80. Interestingly, the increase in sensitivity had no negative impact in the specificity and precision indexes. By contrast, the intersection of tools resulted in low levels of sensitivity, except for the intersection of TS+MR and majority vote that showed higher sensitivity than Pita alone. Overall, there was no improvement in specificity and precision upon using the intersection approach, with values closely resembling to those obtained by Pita or TargetScan alone.

For the random strings control analysis, the specificity of all tools decreased with increase in string length (Table 3), which corroborates the data from Miranda et al. (2006).

Table 3: Specificity values of random string predictions. Each length contain 1000 strings with random sequence of ATCG nucleotides.

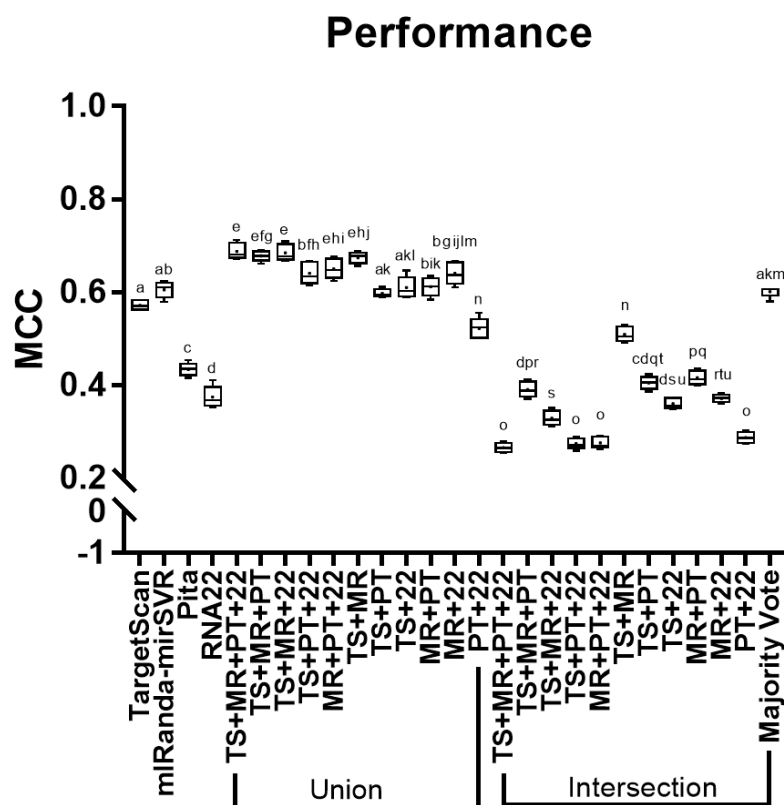
Strings length	TargetScan	miRanda-mirSVR	Pita	RNA22
500 nt	0.9489	0.9446	0.8874	0.9338
1000 nt	0.8966	0.8903	0.7917	0.8766
2500 nt	0.7668	0.7550	0.5562	0.7466
5000 nt	0.5846	0.5596	0.3081	0.6008
500 + 1000 + 2500 + 5000	0.7992	0.7874	0.6359	0.7895
500 + 100 + 2500	0.8708	0.8633	0.7451	0.8523

When the results from all string lengths were summed, the tools had worse specificity values than their pre-computed data. However, the human 3'UTR ranges from 200 to 2,500 nucleotides in length (average = 1,040 nts; Kotogama et al., 2015). Thus, when we summed the results only from 500, 1,000, and 2,500 nts, the specificity of TS, MR, and PT approached, while R22 equaled, to those observed in their pre-computed data. Therefore, these results suggest that the differences in specificities observed for TS, MR, and PT are

more likely to be related to the lack of the conservation parameter (mirSVR score parameter for MR), which were not evaluated by RNA22, than to a bias in validation experiments.

3.3. Evaluating the performance of the methods

All methods showed performance score values higher than zero, although intersections of TS + MR + PT + R22, TS + PT + R22, MR + PT + R22, and PT + R22 showed the lowest performance scores among all methods (i.e., below 0.3; Figure 2). Majority vote showed the best performance among all intersections, which were similar to TargetScan and miRanda-mirSVR. MiRanda-mirSVR showed slightly higher performance than TargetScan, whereas RNA22 showed the lowest performance among the individual tools, followed by Pita. The unions TS + MR + PT + R22, TS + MR + PT, TS + MR + R22, MR + PT + R22, and TS + MR achieved the highest performances with no statistical differences between them (see Supplementary Table 1 for a detailed data of MCC score for each individual miRNA and the combinatorial approach).



4. Discussion

4.1. Similarities and singularities of each tool

The four tools predicted a considerable number of similar targets, with only a few exclusive targets. As expected, the number of correct predictions was more than those of false predictions, which indicated the elevated accuracy of the tools. However, a considerable subset of validated targets (around 18%) was not recovered in the outputs of any tool, indicating the existence of biological and/or methodological aspects that have not yet been addressed by prediction strategies and algorithms. These results also demonstrated that prediction tools have a tendency of identifying certain interactions between miRNA and target genes, but lack the ability to predict other putative interactions.

Comparison of the strategies showed that intersection of TargetScan, miRanda-mirSVR, and Pita predicted twice the number of targets than the intersection of the four tools, whereas RNA22 predicted the highest number of exclusive interactions. These findings are related to the similarity in features of the three seed-based and 3'UTR-specific algorithms for target prediction, especially between TargetScan and miRanda-mirSVR than with RNA22. Moreover, most of the targets predicted exclusively by RNA22 were predicted either outside the 3'UTR or they referred to non-seed based interactions, showing that the use of approaches with distinct search strategies may provide valuable information about miRNA-target interactions.

Similarly, the low number of targets predicted by Pita may be due to the low number of features used by its algorithm. Also, TargetScan, miRanda-mirSVR, and RNA22 use input from the human genome version hg19 (released in 02/2009), whereas Pita uses an older version (hg18, released in 03/2006). This may also explain the lower sensitivity of Pita and demonstrates the importance of regularly updating the database.

Each tool has a unique set of learning attributes (see Table 4 for more details); however, we noticed that for TS, MR, and PT (that focus on the 3'UTR) the missing features in one tool appeared after an update. For instance, the most recent update of TargetScan (TargetScan 7.1, 2016) uses 16 features that are considered important for miRNA target recognition, which generates a score called the “Weighted Context++Score” (WCS). The mirSVR score (version 3.3a, 2010) is a new ranking system that scores targets predicted by miRanda using seven features to improve the miRanda-mirSVR approach. Pita was last

updated in 2008 (version 6) and considers only five parameters to perform the target prediction. RNA22 (version 2, 2015), which focuses on the entire transcript and full-length matches, uses a completely different subset of features, which may explain the differences in the targets identified by RNA22 and the other tools.

Table 4: Summary of the learning attributes of each tool.

Groups	Attributes	TargetScan	miRanda-miRSVR	PiTa	RNA22
Duplex Features	Seed match	X	X	X	X
	3' contribution	X	X	X	X
	SPS	X			
	Heteroduplex free energy p-value				X
Local Context Features	SA	X	X	X	
	Flanking AU	X	X	X	
Global Context Features	TA	X			
	Paring position	X	X		
	3' UTR length	X	X		
	Sequence length				X
	Conservation	X	X	X	
-	Others	X			X

Individually, miRanda-mirSVR showed the best performance with the best balance between sensitivity, specificity, and precision, thus making it the optimal individual choice in most cases. However, TargetScan and Pita showed better precision with the lowest number of false positives. Thus, they could also be used if the objective is to select only few target genes for validation. In this case, TargetScan provides a larger amount of predicted genes that can be selected for further analysis than Pita. RNA22 showed inferior performance compared to those of other tools owing to its slightly lower specificity and precision. However, RNA22 is a unique tool that takes into account non-seed based matches and sites outside the 3'UTR, making it a valuable choice for searching putative non-canonical interactions. It is noteworthy that this analysis is somewhat limited by the number of miRNAs investigated and increasing the number of miRNAs might give a more comprehensive picture of miRNA-target predictions.

Throughout the analysis, all tools demonstrated both positive and negative aspects (summarized in Table 5).

Table 5: Positive and negative aspects of the target prediction tools analyzed

Tool	Positive aspects	Negative aspects
TargetScan	<ul style="list-style-type: none"> - Friendly-user database - Highest number of organism available 	<ul style="list-style-type: none"> - Predictions are the similar for all members of a miRNA family - Not possible to change parameter cutoffs
miRanda-mirSVR	<ul style="list-style-type: none"> - Possible to change parameter cutoffs in source code only 	<ul style="list-style-type: none"> - Database not so friendly - mirSVR score not available in source code
Pita	<ul style="list-style-type: none"> - Possible to change parameter cutoffs - Enable online predictions of users miRNA and 3'UTR 	<ul style="list-style-type: none"> - Not shows interactive view of miRNA-target pairing
RNA22	<ul style="list-style-type: none"> - Friendly-user database - Allows predictions in multiple sources - Possible to change parameter cutoffs 	<ul style="list-style-type: none"> - Source code takes too long to run

For instance, TargetScan has a practical and user-friendly online database, containing the highest number of species that can be analyzed among all tools. However, TargetScan assigns the same targets for miRNAs with similar seed (miRNAs of the same family), which is a drawback considering that the 3' region of the miRNA has an important impact on target recognition (Broughton et al., 2016). Additionally, TargetScan does not allow users to change the parameter cutoffs neither in the online data nor in the source code. Miranda-mirSVR offers the possibility of changing input parameter cutoffs in the source code, although it is not possible to do so in the pre-computed data. However, the miRanda database is less user-

friendly than TargetScan, which causes difficulty in simultaneous visualization of several targets. Moreover, the mirSVR scores are not available in the source code. Pita allows the user to manipulate input parameter cutoffs in both online data and source code. This tool also enables online predictions of user 3'UTR and miRNA queries that were not pre-computed. However, Pita's online applications do not possess any interactive view of the miRNA-target pairing, relying only on the statistical numbers of the predictions. Finally, RNA22 has a user-friendly database that allows predictions of distinct RNA classes and database versions. Additionally, users are allowed to manipulate input parameter cutoffs in both online data and source code, although it is possible to only filter a miRNA sequence but not an mRNA target online. The disadvantage of RNA22 is that its source code has an increased the run time compared to those of other tools (data not shown).

4.2. Intersection versus union

There is no consensus regarding the gold standard for miRNA target prediction. The main questions are whether a tool that is superior to the existing tools exists and whether the intersection or union of two or more tools should be used to acquire more reliable results. According to Witkos et al. (2011), mixing the results from distinct tools decreases the performance of the prediction. They also indicate that the intersection of the results from two or more tools improves specificity at the cost of decreasing sensitivity, whereas the union of two or more tools increases the number of true targets as well the number of false targets detected, which decreases the specificity. Therefore, they suggest using a single target prediction tool. However, several researchers use the intersection approach (D'aurizio et al., 2016; Wang et al., 2016) to avoid false-positive prediction regardless of the loss in sensitivity. Therefore, the use of single tool and an intersection of distinct tools are currently the most common methods of target prediction.

Our analysis revealed that the intersection strategy showed the lowest performance. All intersections showed results that were inferior to the predictions of the individual tools (Figure 2). The lowest performance was obtained by intersections of Pita and RNA22. This may be due to the low sensitivity level of Pita (Figure 1a; Table 2) in addition to the differences in the true targets identified by RNA22 and the other tools. Thus, intersections involving these tools exclusively identify few overlapping targets. Conversely, the intersections of TS+MR and majority vote, which do not depend on Pita and RNA22, showed

better performance, although they were inferior to those of TargetScan and miRanda-miSVR alone.

The methods with the best performance were the unions of TS + MR + PT + R22, TS + MR + PT, TS + MR + R22, MR + PT + R22, and TS + MR, with no significant difference between them (Figure 2). Interestingly, all validated targets predicted by Pita (with the exception of two targets) were also predicted by one of the other tools (Figure 1a). Thus, inclusion of Pita is not required for the union approach. The main difference between the unions of TS+MR+R22 and TS+MR was in the balance of sensitivity and specificity/precision. The union of TS+MR+R22 has high sensitivity (0.82) but lower specificity and precision (0.86 for both), whereas the union of TS+MR has lower sensitivity (0.71) but high specificity and precision (0.95 and 0.93, respectively). Therefore, the choice of the best approach depends on the intended use of the target prediction output.

Researchers perform target prediction analysis for two main reasons. First, to support the subsequent experimental validation of the miRNA-mRNA interaction predicted *in silico*. Second, to select the best candidates for gene ontology enrichment analysis and to identify biological processes that require the activity of these miRNAs. Both objectives demand caution during target prediction analysis. Experimental validation of miRNAs is time-consuming and costly, and therefore, selection of correct positively predicted targets is fundamental for this functional analysis. On the other hand, the quality of gene ontology enrichment analysis strongly depends on the number of inputs. The use of low number of genes as input often does not return results since the data is too scanty to obtain statistically significant values. The TS + MR union provides greater specificity and precision levels, and is recommended for the majority of analyses related to experimental validation of target sites. The TS + MR + R22 union has greater sensitivity, and is appropriate for performing subsequent functional enrichment analysis. Additionally, the TS + MR + R22 union can detect non-canonical interactions (outside 3'UTR and/or full-length match) and is also recommended for exploratory analysis or when most of the targets of the studied miRNA have been validated (although the last option has not yet been fully accomplished). The only disadvantage of using the union of two or more tools is that the scores of these tools (WCS from TS, mirSVR from MR, and minimum free energy and p-value for R22) are composed of different parameters and do not correlate with each other. Therefore, this approach cannot be

used if the final predictions require ranking. In such cases, a single tool should be selected according to the experimental design. For most cases, miRanda-mirSVR offers the best performance.

The poor performance of the intersection approach demonstrates the importance of sensitivity in miRNA target prediction. Until recently, target prediction tools provided outputs with hundreds of false-positive targets per miRNA, which fuelled efforts for enhancing the overall quality of predictions. However, our data shows that the last available updates of the tools have high specificity and precision levels, independent of the method used to combine the data. Thus, the new challenge is to improve the sensitivity of the analysis without decreasing specificity and precision. Since all the parameters governing miRNA-target interaction are not known, the tools use severe cutoffs in the existing parameters (e.g. no mismatch in the seed region) to eliminate false positive predictions, which results in the exclusion of several correct targets. Identification of new features involving miRNA-target recognition may allow these tools to attenuate these cutoffs and increase the range of putative true targets. For example, it is well known that the 3'UTR undergoes alternative polyadenylation (aPa), resulting in transcripts with distinct 3'UTR length in different tissues (Giammartino et al., 2011; Yeh and Yong et al., 2016), which may affect miRNA recognition and regulation. Recent studies showed that conserved miRNA sites are preferentially enriched immediately after aPa sites, and thus, 3'UTR shortening is a potential escape mechanism from miRNA-mediated regulation (Hoffman et al., 2016). TargetScan has already considered aPa sites in its last update; however, since data for the majority of species is still scarce, researchers consider the data for only few cell types and extrapolate those results to a whole organism (Agarwal et al., 2015). In addition, certain interactions are influenced by chromosomal architecture. Considering that chromosomes reside in specific locations inside the nucleus (called chromosome territories; Cremer and Cremer, 2001, 2010) that vary among cell types (Marella et al., 2009), the miRNA-mediated regulation of a gene can fluctuate depending on the proximity of these two mature molecules in the cytoplasm. Study of these and other unknown properties of cellular and genomic parameters can improve the sensitivity of target prediction tools.

5. Conclusions

Current versions of the miRNA target prediction tools evaluated in this study possess high specificity and precision, generating results with negligible false positive rate. This shows that further use of the intersection strategy to obtain high quality predictions is not required. We also found that several true targets were not identified by these tools, necessitating the union of several tools for improving sensitivity. Thus, improvement of sensitivity should be the objective of the next updates.

Overall, the unions of TargetScan + miRanda-mirSVR, as well as that of TargetScan + miRanda-mirSVR + RNA22 provided better results in miRNA target prediction in terms of higher specificity and precision, whereas the latter offers remarkable sensitivity. Therefore, we recommend using these approaches prior to designing target validation experiments. However, the union approach should be avoided when ranking of the output is required. In this scenario, miRanda-mirSVR provided the best performance.

6. References

- Agarwal, V., Bell, G.W., Nam, J., Bartel, D.P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. *eLife*. **4**:e05005. doi: 10.7554/eLife.05005
- Bandyopadhyay, S., Mitra, R. (2009). TargetMiner: microRNA target prediction with systematic identification of tissue-specific negative examples. *Bioinformatics*. **25(20)**:2625-31. doi: 10.1093/bioinformatics/btp503
- Baek, D., Villen, J., Shin, C., Camargo, F.D., Gygi, S.P., Bartel, D.P. (2008). The impact of microRNAs on protein output. *Nature*. **455(7209)**:64–71. doi: 10.1038/nature07242
- Betel, D., Koppal, A., Agius, P., Sander, C., Leslie, C. (2010). Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biol*. **11(8)**:R90. doi: 10.1186/gb-2010-11-8-r90.
- Betel, D., Wilson, M., Gabow, A., Marks, D.S., Sander, C. (2008). The microRNA.org resource: targets and expression. *Nucleic Acids Res*. **36**: D149-53. doi: 10.1093/nar/gkm995.
- Broughton, J.P., Lovci, M.T., Huang, J.L., Yeo, G.W., Pasquinelli, A.E. (2016). Pairing beyond the Seed Supports MicroRNA Targeting Specificity. *Mol Cell*. **64(2)**:320-333. doi: 10.1016/j.molcel.2016.09.004
- Chi, S.W., Hannon, G.J., Darnell, R.B. (2012). An alternative mode of microRNA target recognition. *Nat Struct Mol Biol*. **19(3)**:321-7. doi: 10.1038/nsmb.2230.
- Chou, C.H., Chang, N.W., Shrestha, S., Hsu, S.D., Lin, Y.L., Lee, W.H. et al. (2016) miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. *Nucleic Acids Res*. **44(D1)**:D239-47. doi: 10.1093/nar/gkv1258.
- Clarke, C., Henry, M., Doolan, P., Kelly, S., Aherne, S., Sanchez, N., Kelly, P., Kinsella, P., Breen, L., Madden, S.F., Zhang, L., Leonard, M., Clynes, M., Meleady, P., Barron, N. (2012). Integrated miRNA, mRNA and protein expression analysis reveals the role of post-transcriptional regulation in controlling CHO cell growth rate. *BMC Genomics*. **13**:656. doi: 10.1186/1471-2164-13-656.

- Cremer, T., Cremer, C. (2001). Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat Rev Genet.* **2(4)**:292-301.
- Cremer, T., Cremer, M. (2010). Chromosome Territories. *Cold Spring Harb Perspect Biol.* **2(3)**:a003889. doi:10.1101/cshperspect.a003889
- D'Aurizio, R., Russo, F., Chiavacci, E., Baumgart, M., Groth, M., D'Onofrio, M. et al. (2016). Discovering miRNA Regulatory Networks in Holt–Oram Syndrome Using a Zebrafish Model. *Front Bioeng Biotechnol.* **4**: 60. doi: 10.3389/fbioe.2016.00060.
- Devlin, A.H., Thompson, P., Robson, T., McKeown, SR. (2010). Cytochrome P450 1B1 mRNA untranslated regions interact to inhibit protein translation. *Mol Carcinog.* **49(2)**:190-9. doi: 10.1002/mc.20589.
- Di Giammartino, D.C., Nishida, K., Manley, J.L. (2011). Mechanisms and consequences of alternative polyadenylation. *Mol Cell.* **43(6)**:853-66. doi: 10.1016/j.molcel.2011.08.017.
- Fan, X., Kurgan, L. (2015) Comprehensive overview and assessment of computational prediction of microRNA targets in animals. *Brief Bioinform.* **16(5)**:780-94. doi: 10.1093/bib/bbu044.
- Garcia, D.M., Baek, D., Shin, C., Bell, G.W., Grimson, A., Bartel, D.P. (2011). Weak seed-pairing stability and high target-site abundance decrease the proficiency of *lscy-6* and other microRNAs. *Nat Struct Mol Biol.* **18(10)**:1139-1146. doi: 10.1038/nsmb.2115.
- Gartner, J.J., Parker, S.C., Prickett, T.D., Dutton-Regester, K., Stitzel, M.L., Lin, J.C., et al. (2013). Whole-genome sequencing identifies a recurrent functional synonymous mutation in melanoma. *Proc Natl Acad Sci U S A.* **110(33)**:13481-6. doi: 10.1073/pnas.1304227110
- Grimson, A., Farh, K.K., Johnston, W.K., Garrett-Engele, P., Lim, L.P., Bartel, D.P. (2007). MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell.* **27(1)**:91–105. doi: 10.1016/j.molcel.2007.06.017.
- Hausser, J., Syed, A.P., Bilén, B., Zavolan, M. (2013). Analysis of CDS-located miRNA target sites suggests that they can effectively inhibit translation. *Genome Res.* **23(4)**:604-15. doi: 10.1101/gr.139758.
- Hoffman, Y., Bublik, D.R., Ugalde, A.P., Elkon, R., Biniashvili, T., Agami, R., Oren, M., Pilpel, Y. (2016) 3'UTR Shortening Potentiates MicroRNA-Based Repression of Pro-differentiation Genes in Proliferating Human Cells. *PLoS Genet.* **12(2)**:e1005879. doi: 10.1371/journal.pgen.1005879.
- Jia, W., Li, Z., Lun, Z. (2008). Discoveries and functions of virus-encoded MicroRNAs. *Chinese Science Bulletin* **53**:169–177. doi: 10.1007/s11434-008-0106-y
- Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U., Segal, E. (2007). The role of site accessibility in microRNA target recognition. *Nat Genet.* **39(10)**:1278-84. doi:10.1038/ng2135
- 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**:843–854. doi:10.1016/0092-8674(93)90529-Y.
- Lytle, J.R., Yario, T.A., Steitz, J.A. (2007). Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5' UTR as in the 3' UTR. *Proc Natl Acad Sci U S A.* **104(23)**:9667-72.
- Marella, N.V., Bhattacharya, S., Mukherjee, L., Xu, J., Berezney, R. (2009). Cell type specific chromosome territory organization in the interphase nucleus of normal and cancer cells. *J Cell Physiol.* **221(1)**:130-8. doi: 10.1002/jcp.21836.
- Miranda, K.C., Huynh, T., Tay, Y., Ang, Y.S., Tam, W.L., Thomson, A.M., Lim, B., Rigoutsos I. 2006. A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell.* **126(6)**:1203-17. doi: 10.1016/j.cell.2006.07.031.

- Nelson, P.T., Wang, W.X., Mao, G., Wilfred, B.R., Xie, K., Jennings, M.H., Gao, Z., Wang, X. (2011) Specific sequence determinants of miR-15/107 microRNA gene group targets. *Nucleic Acids Res.* **39(18)**:8163–8172. doi: 10.1093/nar/gkr532.
- Orom, U.A., Nielsen, F.C., Lund, A.H. (2008). MicroRNA-10a binds the 5'UTR of ribosomal protein mRNAs and enhances their translation. *Mol Cell.* **30(4)**:460-71. doi: 10.1016/j.molcel.2008.05.001.
- Parikh, R., Mathai, A., Parikh, S., Chandra Sekhar, G., & Thomas, R. (2008). Understanding and using sensitivity, specificity and predictive values. *Indian J Ophthalmol.* **56(1)**:45–50. doi: 10.4103/0301-4738.37595
- Powers, D.M.W. (2007). Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies.* **2(1)**: 37-63
- Qiu, C., Chen, G., Cui, Q. (2012). Towards the understanding of microRNA and environmental factor interactions and their relationships to human diseases. *Sci Rep.* **2**:318. doi: 10.1038/srep00318
- Reddy, K. B. (2015). MicroRNA (miRNA) in cancer. *Cancer. Cell. Int.* **15**:38. doi: 10.1186/s12935-015-0185-1
- Sandberg, R., Neilson, J. R., Sarma, A., Sharp, P. A., & Burge, C. B. (2008). Proliferating cells express mRNAs with shortened 3' UTRs and fewer microRNA target sites. *Science.* **320(5883)**:1643–1647. doi: 10.1126/science.1155390
- Shenoy, A., Blelloch, R.H. (2014). Regulation of microRNA function in somatic stem cell proliferation and differentiation. *Nat. Rev. Mol. Cell. Biol.* **15(9)**:565-576. doi: 10.1038/nrm3854.
- Schnall-Levin, M., Zhao, Y., Perrimon, N., Berger, B. (2010). Conserved microRNA targeting in *Drosophila* is as widespread in coding regions as in 3'UTRs. *Proc Natl Acad Sci U S A.* **107(36)**:15751-6. doi: 10.1073/pnas.1006172107
- Sunkar, R., Girke, T., Jain, P.K., Zhu, J.K. (2005). Cloning and Characterization of MicroRNAs from Rice. *Plant Cell.* **17**:1397–1411. doi: 10.1105/tpc.105.031682.
- Tay, Y., Zhang, J., Thomson, A.M., Lim, B., Rigoutsos, I. (2008) MicroRNAs to Nanog, Oct4 and Sox2 coding regions modulate embryonic stem cell differentiation. *Nature.* **455(7216)**:1124-8. doi: 10.1038/nature07299
- Wang, J., Liu, H., Tian, L., Wang, F., Han, L., Zhang, W. et al. (2016). miR-15b Inhibits the Progression of Glioblastoma Cells Through Targeting Insulin-like Growth Factor Receptor 1. *Horm Cancer.* doi: 10.1007/s12672-016-0276-z
- Witkos, T.M., Koscianska, E., Krzyzosiak, W.J. (2011). Practical Aspects of microRNA Target Prediction. *Curr Mol Med.* **11(2)**: 93–109. doi: 10.2174/156652411794859250.
- Yeh, H.S., Yong, J. (2016). Alternative Polyadenylation of mRNAs: 3'-Untranslated Region Matters in Gene Expression. *Mol Cells.* **39(4)**:281-5. doi: 10.14348/molcells.2016.0035.
- Zhou, H., Rigoutsos, I. (2014). MiR-103a-3p targets the 5' UTR of GPRC5A in pancreatic cells. *RNA.* **20(9)**:1431-9. doi: 10.1261/rna.045757.114.

----- **4. Capítulo II:**
Análise de enriquecimento funcional

4.1. Material e métodos

4.1.1. Resumo do workflow

Para as análises de enriquecimento funcional dos alvos dos miRNAs, foram utilizados dados previamente publicados de *microarray* pós-superexpressão de 11 miRNAs. Adicionalmente, para a realização da análise os alvos de cada miRNA foram agrupados em *clusters* de acordo com o grau de *fold-change* apresentado após a superexpressão do miRNA (veja figura 5 para um resumo completo das atividades realizadas).



Figura 5: Fluxograma apresentando resumo das atividades realizadas referentes as análises funcionais.

4.1.2. Obtenção dos dados de *microarray*

Foram selecionados dados previamente publicados de *microarray* que avaliaram a diferença de expressão de mRNAs pós-perturbação de onze miRNAs em *HeLa cells* (GSM210897 – miR-7, GSM210898 – miR-9, GSM210901 – miR-122a, GSM210903 – miR-128a, GSM210904 – miR-132, GSM210907 – miR-133a, GSM210909 – miR-142, GSM210911 – miR-148b, GSM210913 – miR-181a, GSM37599 – miR-1, GSM37601 – miR-124). Estes dados foram selecionados por terem sido analisados por uma mesma plataforma Agilent, apresentarem um claro sinal de repressão por miRNAs e estarem pré-processados e normalizados (Garcia et al., 2011), minimizando assim, efeitos não-específicos e erros metodológicos.

4.1.3. Agrupamento dos alvos em clusters de mRNA *fold-change*

Para cada miRNA analisado os alvos com variações de expressão semelhantes (*fold-change* \log_2) foram manualmente agrupados em cinco *clusters* de acordo com a intensidade da variação. *Cluster* -0.1 contendo variações de mRNA *fold-change* [-0.1 à -0.2], *cluster* -0.2

contendo variações de $[-0.2 \text{ à } -0.3[$, *clusters* -0.3 contendo variações de $[-0.3 \text{ à } -0.4[$, *clusters* -0.4 contendo variações de $[-0.4 \text{ à } -0.5[$ e *clusters* -0.5 contendo variações de $[-0.5 \text{ à } -0.6[$.

4.1.4. Análise de enriquecimento funcional

A análise de enriquecimento funcional foi realizada utilizando a ferramenta ToppCluster (Kaimal et al., 2010). Esta ferramenta é capaz de realizar análises de enriquecimento multi-*cluster* integrando e comparando os dados apresentados por cada *cluster* individualmente. Para tal, as opções *default* foram utilizadas e foram selecionadas as anotações de Processos Biológicos (Biological Process – BP), Contexto Celular (Cellular Context – CC) e Função Molecular (Molecular Function – MF).

Para o grupo controle, em cada miRNA todos genes com variações na expressão (mRNA fold-change -0.1) foram embaralhados e gerou-se cinco *clusters* aleatórios (-0.1r, -0.2r, -0.3r, -0.4r e -0.5r) com dez replicatas, contendo o mesmo número de genes que seu *cluster* correspondente (ex: miR-1: *cluster* -0.1 e *cluster* -0.1r possuem 1488 genes, enquanto *cluster* -0.2 e -0.2r possuem 725 genes). O enriquecimento funcional do grupo controle foi realizado da mesma forma descrita anteriormente.

Para a análise dos resultados do enriquecimento funcional apenas os termos BP foram considerados, evitando-se assim o viés de se comparar termos distintos. Os termos CC e MF foram utilizados para a análise mais aprofundada das funções biológicas dos miRNAs.

4.1.5 Análise de conservação evolutiva

Para avaliar se existe uma possível conservação do agrupamento dos alvos dos miRNAs nos distintos *clusters* entre as espécies de vertebrados, comparou-se o Context++Score (CS) da interação miRNA-alvo entre dez espécies de vertebrados (*Homo sapiens*, *Pan troglodytes*, *Macaca mulata*, *Mus musculus*, *Ratus norvegicus*, *Bos taurus*, *Canis familiaris*, *Monodelphis domestica*, *Gallus gallus* e *Xenopus tropicalis*). Para isso, primeiramente foi verificado se há uma correlação entre o mRNA *fold-change* apresentado pelos dados de microarray e o CS apresentado pelo TargetScan. Nesta análise apenas foram considerados os alvos presentes em todas dez as espécies, com exceção dos miRNAs miR-132 e miR-142 que não possuíam dados para as espécies *P. troglodytes* e *Gallus gallus* (miR-132) e *Canis familiaris* (miR-142).

4.2. Resultados e discussão

Os dados da análise de enriquecimento funcional são apresentados a seguir no formato de artigo científico:

GO enrichment analysis suggests that vertebrate microRNAs distinctly regulate their targets according to their biological function

Arthur C. Oliveira¹, Luiz A. Bovolenta², Pedro G. Nachtigall¹, Marcos E. Herkenhoff¹, Ney Lemke² and Danilo Pinhal^{1,*}

¹ Department of Genetics, Institute of Bioscience of Botucatu, Botucatu, Sao Paulo, 18618-689, Brazil

² Department of Physics and Biophysics, Institute of Bioscience of Botucatu, Botucatu, Sao Paulo, 18618-689, Brazil

ABSTRACT

MicroRNAs (miRNAs) are non-coding RNAs that regulate a wide range of biological pathways by post-transcriptionally controlling gene expression. Since they regulate several biological functions within one cell, they should be able to distinctly regulate several groups of genes assign with distinct biological process in order to correctly control cell homeostasis. We performed GO enrichment analysis of the genes with altered expression caused by the injection of eleven miRNAs after clustering the genes by the intensity of the expression variation. By this, we show that miRNAs differentially regulates gene expression according to the biological process they are assign. Moreover, we show that this is a conserved phenomenon among vertebrates. Finally, we in-depth analysed the biological pathways regulated by three miRNAs in order to better understand the relevance of this occurrence at cellular context.

INTRODUCTION

MicroRNAs (miRNAs) are large class of short non-coding RNAs that regulates virtually every biological process described, from cell differentiation and proliferation (Shenoy and Blelloch, 2014) to diseases such as cancer (Reddy, 2015). They act post-transcriptionally, modulating gene expression by binding with the complementary region of its mRNAs targets, mainly at 3'UTR region (Lai, 2002). In animals, miRNAs regulate its targets by firstly blocking its translation and then promoting the premature degradation of the mRNA (Ameres and Zamore).

Once miRNAs regulates several distinct biological process within the same cell, they should be able to differentially regulate each one in order to provide the specific modulation demanded for each individually case. Several studies described features involved in the miRNA-target recognition, such as seed match (Lai, 2002) and site accessibility (Kertesz et al., 2007), and how its affect the intensity of the regulation delivered by the miRNA. However, these studies only demonstrate variations between each miRNA-target interactions individually, apart from the biological context they are assigned.

Here we revealed that miRNAs regulate most of their targets under related biological functions with similar intensity and that this intensity varies between distinct biological functions, making miRNAs able to precisely control several biological process acting in the same cellular context. We also show that this categorized regulation is conserved at least throughout vertebrates, suggesting that such ranked grouping may have been positively selected by natural selection. Finally, we in-depth analysed the biological pathways regulated by three known miRNAs taking into account these new regulatory mechanism in order to understand its biological impact.

MATERIAL AND METHODS

Microarray data

We selected for data mining a previously published microarray dataset reporting mRNA level changes after the overexpression of 11 miRNAs induced by miRNA mimics injection into HeLa cells (GSM210897 – miR-7, GSM210898 – miR-9, GSM210901 – miR-122a, GSM210903 – miR-128a, GSM210904 – miR-132, GSM210907 – miR-133a, GSM210909 – miR-142, GSM210911 – miR-148b, GSM210913 – miR-181a, GSM37599 –

miR-1, GSM37601 – miR-124; Garcia et al., 2011). This dataset was selected due to its clear signal for miRNA-based repression, acquired using the same Agilent array platform, whose data was pre-processed and normalized, minimizing computational bias.

GO Enrichment Analysis

For each miRNA we manually grouped their respective target genes with changes in expression (fold-changes \log_2) in five clusters according to the intensity of regulation induced by treatment with miRNA mimics. Herewith, we defined “intensity of regulation” by the fold-change of the mRNAs after the overexpression of the miRNA. Therefore, the clusters was grouped as follows: cluster “-0.1” for variations of mRNA fold-change $[-0.1$ to $-0.2[$, cluster “-0.2” for variations of $[-0.2$ to $-0.3[$, cluster “-0.3” for variations of $[-0.3$ to $-0.4[$, cluster “-0.4” for variations of $[-0.4$ to $-0.5[$, and cluster “-0.5” for variations of $[-0.5$ to $-0.6[$.

Later, we performed GO enrichment analysis using the ToppCluster tool (Kaimal et al., 2010). ToppCluster is a tool capable of performing multi-cluster enrichment analysis on large-scale data such as microarray datasets and is able to correlate and compare the output. We used the default option selecting Biological Process (BP), Cellular Component (CC) and Molecular Function (MF) annotations.

In the control groups, for each miRNA we randomized the genes with changes in mRNA expression (fold-change ≤ -0.1) and generated five random clusters (-0.1r, -0.2r, -0.3r, -0.4r and -0.5r) containing the same number of genes of each correspondent cluster. To ensure confidence in results this analysis was performed with ten replicates. The enrichment analyses of the control groups were performed using the same parameters as above.

To compare the results of the GO enrichment, we only consider the BP terms, in order to reduce the bias of working with distinct annotations. The CC and MF were used to an in-depth analysis of the miRNA biological functions.

RESULTS

GO enrichment analysis of miRNA targets suggests a categorized regulation of biological process according to its intensity.

For all miRNAs analysed, the number of perturbed genes decreased as the intensity of the regulation increased. Therefore, cluster “-0.1” was always the cluster with the highest

number of genes while the cluster “-0.5” was the cluster with the lowest number. In addition, we observed a high amount of enriched BP terms with little-to-no overlap between clusters for almost all miRNAs analysed (Figure 1). An exception to that was miR-181a, which showed no enriched term in any cluster and thus was excluded from the subsequent analysis.

Moreover, half of the miRNAs showed little correlation between the number of enriched BP terms (or its absence) and the number of perturbed genes used during the GO enrichment analysis. (Figure 2; $R^2 < 0.7$).

For instance, miR-7 had four times more enriched BP terms in cluster -0.2 than cluster -0.1 (40 and 12 terms respectively), with no overlap between them, despite of having three times less genes than cluster -0.1 (462 and 1468 genes respectively). Another example, miR-9 showed the higher number of enriched BP terms at cluster -0.3 (17 terms and 397 genes) despite having two times less perturbed genes than cluster -0.2 (12 terms and 777 genes) and four times less genes than cluster -0.1 (0 terms 1575 genes), with only one term overlapping between clusters -0.2 and -0.3. Interestingly, cluster -0.1 from miR-9 did not showed any BP enriched term at all (Table 1). These data suggest that miRNAs modulate the intensity of the regulation of their target genes according to which biological process they are assign, with different but defined levels of intensity for each group of genes, rather than distinctly regulating several genes within the same biological process.

Table 1: Number of perturbed genes and enriched terms in each cluster. For each miRNA, “perturbed genes” represents the number of genes contained in each cluster, while “enriched terms” represents the number of of enriched BP terms in the same cluster.

Cluster	Perturbed genes	Enriched terms	Perturbed genes	Enriched terms	Perturbed genes	Enriched terms	Perturbed genes	Enriched terms	Perturbed genes	Enriched terms
--	miR-1		miR-7		miR-9		miR-122a		miR-124	
-0.1	1488	17	1438	12	1575	0	1533	1	1326	8
-0.2	725	13	462	40	777	12	722	0	714	0
-0.3	337	0	151	2	397	17	328	1	418	0
-0.4	201	0	89	0	229	11	191	2	231	1
-0.5	104	0	58	1	121	0	121	0	148	0
--	miR-128		miR-132		miR-133a		miR-142		miR-148b	
-0.1	1491	37	1377	33	1623	12	1511	4	1631	5
-0.2	670	3	559	5	758	0	485	0	856	3
-0.3	302	0	291	2	408	8	174	0	417	0
-0.4	205	3	136	1	204	3	92	2	239	0
-0.5	95	0	102	0	117	3	58	0	164	1

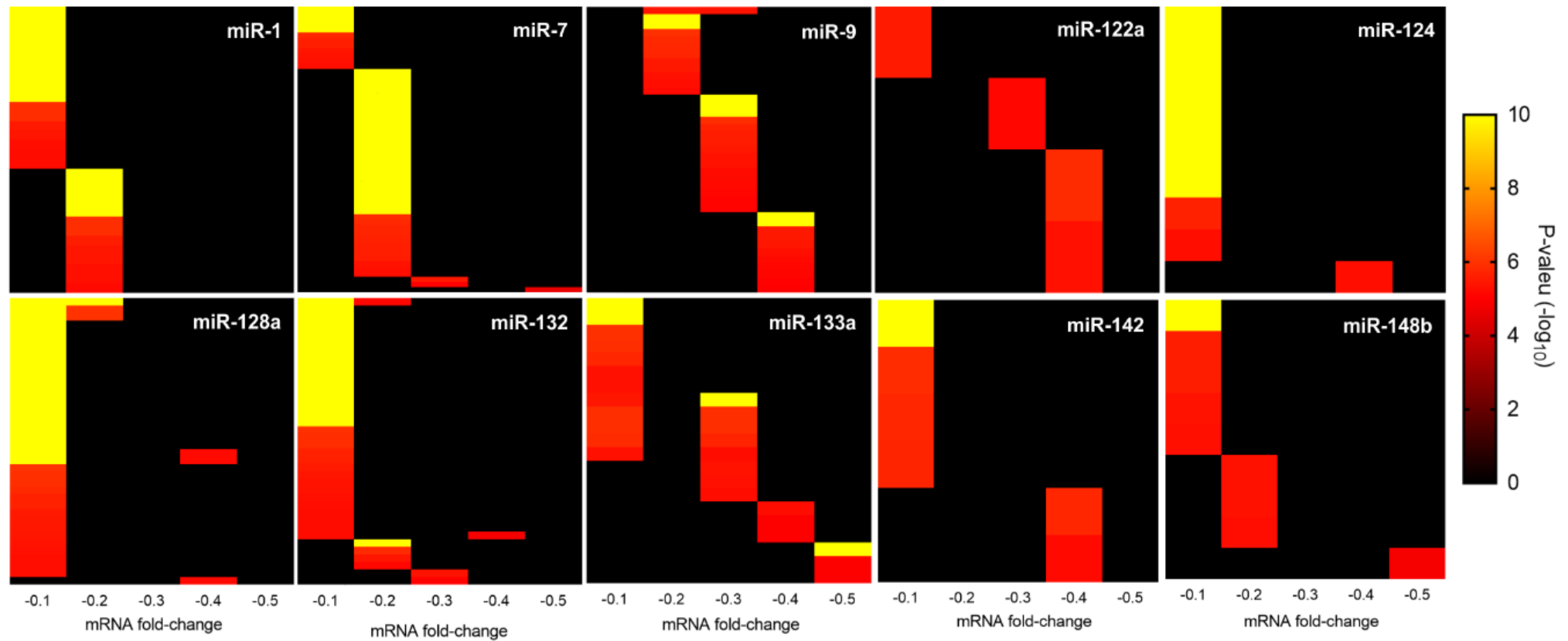


Figure 1: Heatmap of the P-values ($-\log_{10}$) of the enriched terms. The coluns contains the enriched BP terms of each cluster. The values range from 0 to 10 where 0 indicates no term enriched

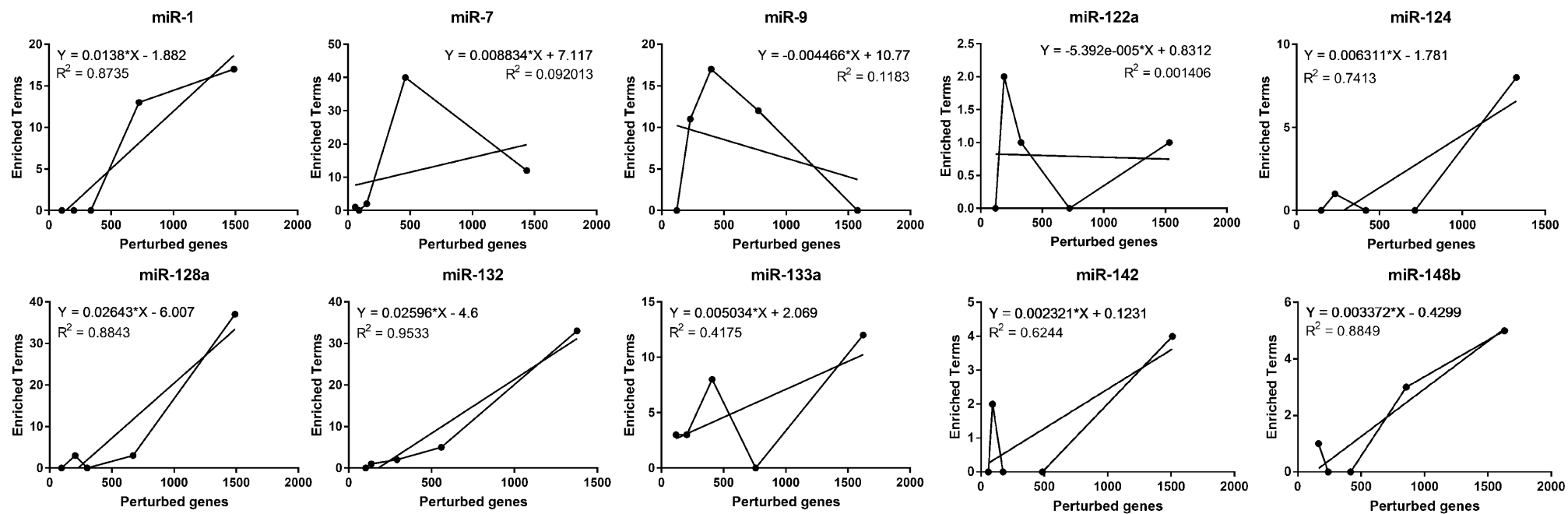


Figure 2: Correlation between the number of enriched terms and perturbed targets. Each dot represents one cluster (-0.1 to -0.5 from right to left). The x values represents the number of genes, while y values represents the number of enriched terms contained in each cluster.

To confirm these results, we performed the same enrichment analysis for the random sets (Table 2). Unlike the studied data, the enriched GO terms of the random sets decreased with the increase of the cluster intensity, reinforcing the idea of the non-random segregation of the BP terms of the studied groups. Moreover, most clusters “-0.1” showed higher number of enriched GO terms in random sets than in studied groups while the other clusters showed a lower number of GO enriched terms in random sets.

Table 2: Number of enriched BP terms of studied and random groups. For each miRNA, “studied group” represents the number of enriched BP terms detected, while “random group” represents the average number of enriched BP terms obtained from 10 random sets.

Cluster	Studied group	Random group	Studied group	Random group	Studied group	Random group	Studied group	Random group	Studied group	Random group
--	miR-1		miR-7		miR-9		miR-122a		miR-124	
-0.1	17	30.6±13.45	12	22.6±9.89	0	49.9±21.55	1	11.4±9.3	8	12.8±12.54
-0.2	13	3.7±4.67	40	2.1±2.96	12	10.4±8.95	0	1.1±1.60	0	6.1±10.28
-0.3	0	2.9±4.57	2	0±0.00	17	2.5±3.60	1	0.3±0.68	0	1.4±2.12
-0.4	0	0.5±0.85	0	0.1±0.32	11	0.9±1.10	2	0.1±0.32	1	0.3±0.67
-0.5	0	0.2±0.63	1	0±0.00	0	0±0.00	0	0.3±0.94	0	0.4±0.97
--	miR-128		miR-132		miR-133a		miR-142		miR-148b	
-0.1	37	38.3±16.41	33	37.6±21.91	12	7.7±5.62	4	21.3±8.87	5	5.2±3.43
-0.2	3	8.5±7.17	5	2.3±2.64	0	2.6±3.53	0	2.4±3.37	3	1.3±1.95
-0.3	0	0.5±2.27	2	1.4±1.76	8	0.6±1.27	0	0.1±0.32	0	0.7±1.90
-0.4	3	0.4±0.53	1	0.1±0.32	3	0.4±0.70	2	0.2±0.63	0	0.3±0.67
-0.5	0	0±0.97	0	0.5±1.08	3	0±0.00	0	0.4±1.27	1	0.1±0.32

To verify whether such regulatory categorization could be evolutionary conserved, we compared the prediction scores from the targets of the miRNAs analysed between ten vertebrate species (*Homo sapiens*, *Pan troglodytes*, *Macaca mulata*, *Mus musculus*, *Ratus norvegicus*, *Bos taurus*, *Canis familiaris*, *Monodelphis domestica*, *Gallus gallus* e *Xenopus tropicalis*). We only consider those target genes predicted into all ten species, with exception of miR-132 and miR-142 that did not had any prior predicted target for the species *P. troglodytes* and *Gallus gallus* (miR-132) and *Cannis familiaris* (miR-142). The Context++Score (CS) of the interaction of the miRNA and its predicted targets provided from TargetScan prediction tool is directly correlated with the mRNA fold-change of the target genes after the overexpression of the miRNA ($R^2 = 0.9737$; Figure 3), being trustworthy to be used for this correlation.

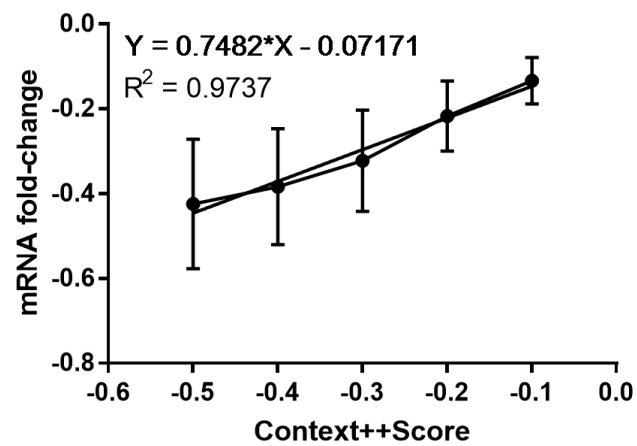


Figure 3: Correlation between the Context++Score and the mRNA fold-change of the perturbed genes.

Our results showed that most miRNA targets are regulated with similar intensity between vertebrates (Figure 4). Therefore, suggesting that the segregation of the biological function regulation promoted by the miRNAs is conserved at least through vertebrates.

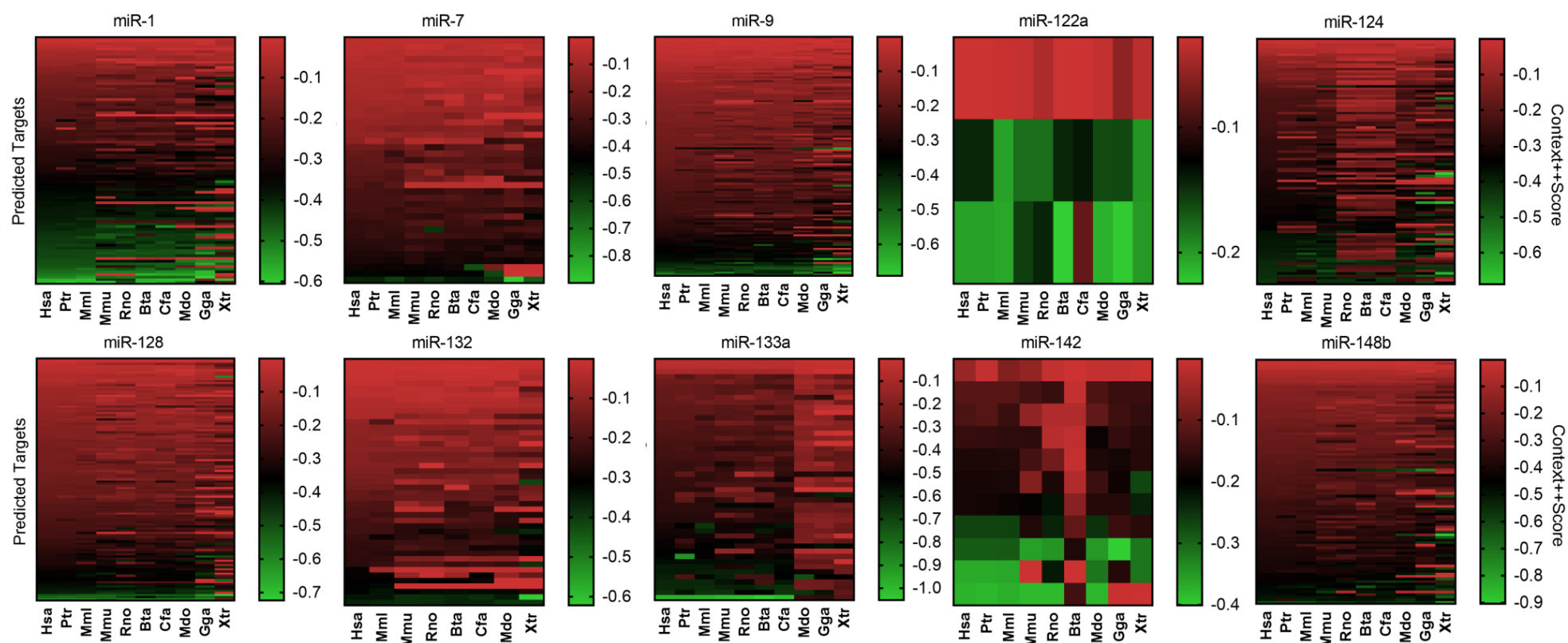


Figure 4: Heatmap of the Context++Score of the target genes from ten vertebrate species. Each column contain Context++Score values from the predicted target of one of the ten species analyzed. Red color represents lower quality Context++Score values while green color represents greater quality Context++Score values.

In-depth analysis of miRNA target GO enrichment

Based on the results obtained, we selected three miRNAs to further scrutinize biological functions using the cluster-based GO enrichment analysis in order to better understand the influence of miRNA-target variable regulatory intensity for the cellular context.

MiR-1. MiR-1 has exhibited an enrichment for BP only in clusters “-0.1” and “-0.2”, whereas MF was enriched in clusters “-0.1”, “-0.2” and “-0.3”, and CC was enriched in all five clusters (Figure 5). Cluster “-0.1” showed 17 enriched BP terms that could be grouped into six biological functions: receptor protein signalling pathway and signalling transduction (3 BP), regulation of multicellular organismal process (3 BP), regulation of transport and secretion (4 BP), regulation of cell motility (4 BP), regulation of cell adhesion (2 BP) and urogenital system development (1 BP). The CC terms enriched in cluster “-0.1” were all associated with membrane region, more specifically regarding the synaptic region. This comes in agreement with the biological functions ascribed, that were mainly related to cell signalling, adhesion and transport/secretion, processes highly associated with synaptic functions, indicating a refinement at this regulation intensity for genes acting in this cellular region.

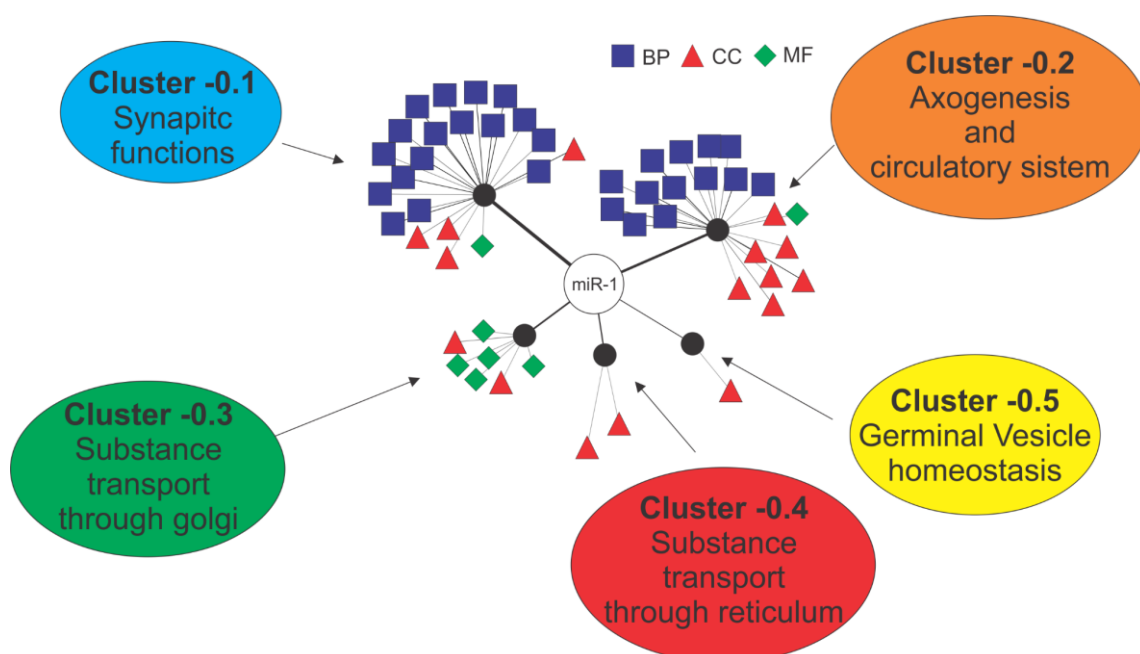


Figure 5: Functional Enrichment map of miR-1. BP = Biological Process, CC = Cellular Context and MF = Molecular Function.

Cluster “-0.2”, showed 13 biological process terms that could be grouped into four biological functions: axonogenesis and neuron projection (4 terms), phosphorylation (4 terms), cell growth (3 terms) and circulatory system development (2 terms). According to that, the enriched CC terms were mainly associated with axon formation and elongation. Interestingly, although both clusters “-0.1” and “-0.2” showed enriched terms associated with neuronal activity, they had distinct biological functions. While cluster “-0.1” was enriched in the synaptic membrane, cluster “-0.2” was associated with the axon formation and phosphorylation processes, known to play important roles in somatodendritic region of the neuron (Walaas and Greengard, 1991). This shows that miR-1 is capable of regulating distinct portions of the neuron body with distinct intensities, promoting a heterogeneous gene expression and thus refining such complex structure. Also, miR-1 is recognized for its influence on cardiac tissue (Deng et al., 2014). Our results showed that most target genes of miR-1 biologically assigned with cardiac functions were regulated with similar intensities of mRNA fold-change of “-0.2”, thus promoting a specific regulatory intensity for these function.

Cluster “-0.3”, despite not having any BP term enriched, has target genes mainly acting in Golgi complex (GC), regulating the movement of substances through membrane. In the same context, cluster “-0.4” contain target genes expressing on endoplasmic reticulum (ER). These two clusters, although do not showing direct biological terms assigned with neuron functions, embrace genes acting in indispensable organelles for this system, since the GC and ER are critical during the production of the neurotransmitters (Purves et al., 2001).

Cluster “-0.5” has target genes enriched in the germinal vesicle. Germinal vesicle is the name given to the nucleus at oocyte stage. In mammals, at this stage, the cell remains for a long period at prophase I of meiosis and re-enter the meiotic maturation after a surge of luteinizing hormone (Brunet and Maro, 2007). Thus, miR-1 regulates genes associated with the maintenance of the oocyte at prophase I stage with a high intensity (“-0.5”) of mRNA fold-change.

The aforementioned analysis showed a complex neuronal heterogeneous regulation provided by miR-1 in nervous system. Moreover, revealed that most cardiac associated target genes were regulated under a specific intensity, suggesting they promote a stable control of this organ. Finally, miR-1 showed to be potentially indispensable during embryonic

development regulating the stability of the oocyte at prophase I of meiosis with the high intensity of mRNA fold-change “-0.5”.

MiR-7. MiR-7 had enriched BP in clusters “-0.1”, “-0.2”, “-0.3 and -0.5”, enriched MF in clusters “-0.1”, “-0.2” and “-0.3”, with one term shared by clusters “-0.1” and “-0.2” (poly(A) RNA binding), and enriched CC in clusters “-0.1”, “-0.2” and “-0.5” (Figure 6).

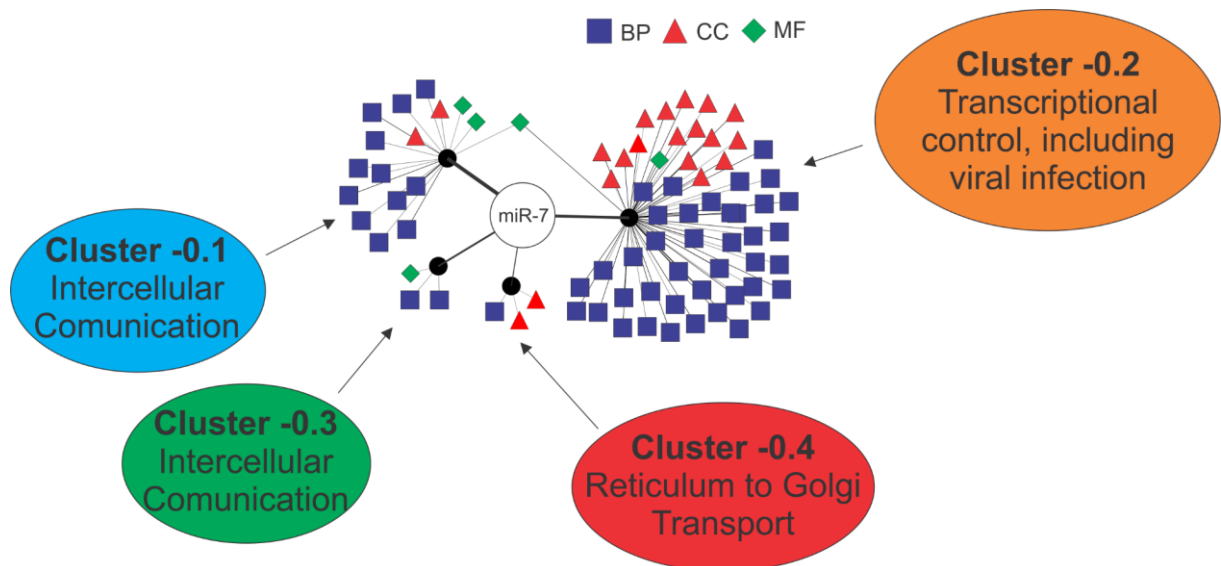


Figure 6: Functional Enrichment map of miR-7. BP = Biological Process, CC = Cellular Context and MF = Molecular Function.

Cluster “-0.1” showed 12 enriched BP terms that could be grouped into five biological functions: transport and secretion (6 terms), transmembrane receptor and signalling (3 terms), ornanonitrogen compound biosynthesis (1 term), response to endogenous stimulus (1 term), and neuron differentiation (1 term). The CC terms enriched in this cluster were cell-cell junction and extracellular space, whereas the MF terms were molecular binding and structure. This suggests that miR-7 regulation target genes important for cellular intercommunication with the intensity of mRNA fold-change -0.1.

Cluster “-0.2” showed 40 enriched BP terms that could be grouped into six biological functions: translation processes (12 terms), protein localization/transport and targeting (12 terms), viral process (8 terms), amide and peptide biosynthesis and metabolic process (4 terms), heterocycle metabolic process (3 terms), and locomotory behaviour (1 term). The CC terms enriched in this cluster were mainly assigned to ribosome compounds, while MF terms were poly(A) RNA binding and structural constituent of ribosome. This indicates a

convergence of targets involved in translation pathways at this regulatory intensity. Interestingly, several target genes in this cluster are important during viral infection. This fact comes in accordance with the main core of the process regulated at this intensity of mRNA fold-change -0.2, since the main influence of the virus inside the cell is to manipulate the transcriptional-translational pathway in order to produce its own proteins and DNA/RNA.

Cluster “-0.3” showed two BP terms associated to regulation of neuron death. Moreover, its MF “lactate transmembrane transporter activity” is important on apoptosis controlling (Romero-Garcia et al., 2016), which suggest that miR-7 regulatory functions are related to genes involved on the control cell death process, given the current intensity in mRNA fold change.

Cluster “-0.5” showed one BP assign with “Golgi endosome transport”, with two CC assign with the pathway nucleus-endoplasmic reticulum-Golgi complex, suggesting a selective regulation of these pathways with mRNA-fold change of -0.5. Additionally, this cluster might work in tandem with cluster -0.1, both containing genes associated to the control of cellular intercommunication.

MiR-9. MiR-9 had enriched BP in clusters “-0.2”, “-0.3” and “-0.4”, with one term shared by clusters “-0.2” and “-0.3” (cellular amide metabolic process). Enriched CC in clusters were found at “-0.2”, -0.3” “-0.4” and “-0.5”, with four overlapping terms between clusters “-0.2” and “-0.3” and one term shared by clusters “-0.2”, “-0.3” and “-0.5”. The MF terms are present in clusters “-0.2”, “-0.3”, “-0.4” and “-0.5”, with one term shared by all these clusters (poly(A) RNA binding) (Figure 7).

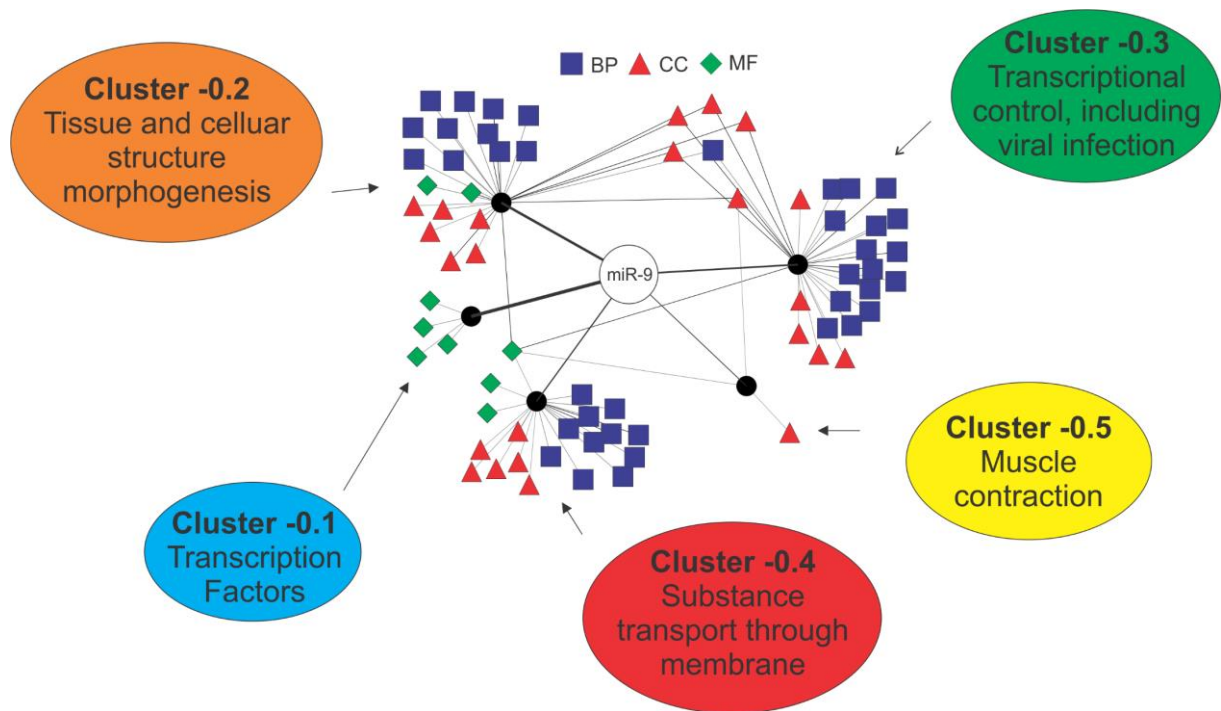


Figure 7: Functional Enrichment map of miR-9. BP = Biological Process, CC = Cellular Context and MF = Molecular Function.

Cluster “-0.1” despite not having any BP term has four MF terms assigned to transcription factors activity. Since transcription factors are essential for gene transcription, this suggests that this cluster might control global gene expression.

Cluster “-0.2” showed 12 BP that could be grouped into five biological functions: morphogenesis (5 terms that can be subdivided in four groups: anatomical morphogenesis, brain development, cardiac system development and epithelium development), organelle organization (3 terms), neuron death (2 terms), organonitrogen compound biosynthesis (1 term) and cellular amide metabolic process (1 term). The CC terms enriched in this cluster are mainly assign with cell junction and structure. The MF terms are assign with cell, RNA and ATP binding functions. This suggests that this cluster main regulation is at the development of some tissues and the establishment of its cells components and structure.

Cluster “-0.3” showed 17 BP that could be grouped into six biological functions: translation process (6 terms), viral process (4 terms), negative regulation of signalling (3 process), amide metabolic process (3 terms) and maintenance of location (1 term). The CC enriched terms are mainly assign with cell adhesion. This suggests that this cluster focus on regulating targets assign with translational pathways. Similar to cluster “-0.2” from miR-7

there are several target genes acting during viral infection, showing the correlation of the regulation between cellular transcription and viral process.

Cluster “-0.4” showed 10 BP that could be grouped into six biological functions: membrane organization (4 terms), protein transport (2 terms), viral life cycle (1 term), cell localization (1 term), androgen receptor signalling (1 term), and response to organic cyclic compound (1 term). The CC terms are main assign with the membrane region and vesicle, and the MF terms are related with protein and RNA binding and transmembrane transporter. This suggests that this cluster main control processes related with transport of cellular compounds through membrane.

Cluster “-0.5” do not have any enriched BP term, but the enriched CC terms are sarcoplasmic reticulum membrane and adherens junctions, whereas the single MF term was poly(A) RNA binding. The CC terms suggests that this cluster control genes in regions responsible for promoting muscular contraction, since sarcoplasmic reticulum balances the calcium storage and release during contraction (Rossi and Dirksen, 2006) and adherens junctions keep the skeletal muscle cells attached (Hartsock and Nelson, 2007).

DISCUSSION

Each miRNA potentially regulates hundreds of target genes with distinct biological functions. Moreover, each of these targets are predicted to be regulated by more than one miRNA. Thus, miRNAs activity can impact dozens of biological pathways that may greatly differ from each other. For a miRNA to be capable of properly and simultaneously regulate distinct pathways, some sort of intrinsic mechanism to ensure specificity is required. In this sense, several properties of the miRNA-target interaction have been described, like seed match (Lai, 2002), site accessibility (Kertesz et al., 2007), 3' contribution (Broughton et al., 2008), and others. Those properties influence in the intensity of the regulation promoted by the miRNA and are distinct from interaction to interaction. However, besides the distinct miRNA-target regulation, the regulation of the targets belonging to the same biological pathways should be orchestrated, otherwise it would disrupt the homeostasis of the tissue or cell.

Our findings suggests that this organization is accomplished by regulating several targets with the same biological functions with a very similar intensity, being the cell able to regulate distinct biological pathways with high degree of specificity for each one. Moreover,

almost no enriched BP term is shared between two or more clusters implying that a well-defined segregation of the target genes exists.

The conservation of the regulatory intensity promoted by miRNAs among vertebrates suggests that this mechanism might have been selected during evolution as it brings advantage for a proper regulation of distinct biological process within the same cell. This fact also suggests that other mechanisms promotes the specificity between species rather than do changes on the regulatory intensity of miRNAs over target genes products. Such mechanism may involve the regulation of distinct targets and the presence of species-specific miRNAs (Mor and Shomron, 2013).

The in-depth analysis performed on miR-1, miR-7 and miR-9 indicates that the miRNA regulatory clusters although have several enriched BP terms, mainly focus on the regulation of one biological function. For example, miR-1 cluster “-0.1” genes are related to the regulation of processes of the synaptic region; miR-7 cluster “-0.2” genes are involved in processes of the transcription-translation pathways, including virus-encoded protein; and miR-9 cluster “-0.2” genes control development and establishment of tissues and cell types. Moreover, miR-7 cluster “-0.2” and miR-9 cluster “-0.3” share several enriched BP terms suggesting distinct miRNAs can combine their regulatory intensity in order to provide the necessary regulation demanded for control each biological function.

The comparison between the studied clusters and the random sets comes in agreement with this idea. Most miRNAs analyzed, with exception of miR-128, miR-133a and miR-148b, had fewer enriched BP terms in cluster “-0.1” that would be expected, according to the number of target genes, while for the other clusters the number of enriched BP terms (exception for when there is no term enriched) were higher than the obtained in the random sets. The differences observed in clusters “-0.1” supports the ideia that target genes with BP related to each other converged into similar regulatory intensity, rather than several genes with non-related BP being under this control intensity. The same idea can be applied to the other clusters. While random sets returned low number of enriched BP terms, probably due to the number of genes used far below than cluster “-0.1”, the studied group showed a higher number of enriched BP terms in this clusters, suggesting that the genes that make up these clusters also corresponds to BP related to each other.

It is important to remark though, that this not mean that all target genes with a biological process are regulated under the same intensity. The enrichment analysis shows BP

terms that are overrepresented in the sample, meaning that there are more targets with these functions than would be in a random selection. Given that, there also should be target genes inside a BP that is regulated with distinct levels, but most of them are controlled by the same intensity.

Our results suggest a potential novel route to explain the specificity and robustness of gene regulation by miRNAs. The segregation of regulatory intensities according to the BP of their targets allows miRNAs to efficiently regulate several biological pathways inside one cellular context with the required degree demanded for each one. In this sense, understanding how miRNAs differentiate their targets according to its BP could allow researches manipulate a biological pathway by perturbing a miRNA without interfering in the others, minimizing the off-target effects generated by this approach nowadays.

REFERENCES

- Ameres and Zamore (2013) Diversifying microRNA sequence and function. *Nature Reviews Molecular Cell Biology*. 2013;14:475–488.
- Brunet S and Maro B (2007) Germinal vesicle position and meiotic maturation in mouse oocyte. *Reproduction*. 133(6):1069-72.
- Deng F, Xu X and Chen Y (2014) The Role of miR-1 in the Heart: From Cardiac Morphogenesis to Physiological Function. *Human Genet Embryol* 4:119.
- Hartsock A and Nelson WJ (2007) Adherens and tight junctions: structure, function and connections to the actin cytoskeleton. *Biochim Biophys Acta*. 1778(3):660-9.
- Lai EC (2002) MicroRNAs are complementary to 3' UTR sequence motifs that mediate negative post-transcriptional regulation. *Nat Genet*. 30:363–4.
- Kaimal V, Bardes EE, Tabar SC, Jegga AG, Aronow BJ (2010) ToppCluster: a multiple gene list feature analyzer for comparative enrichment clustering and network-based dissection of biological systems. *Nucleic Acids Res*. 38:W96-102.
- Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E (2007) The role of site accessibility in microRNA target recognition. *Nat Genet*. 39:1278–84.
- Mor E, Shomron N (2013) Species-specific microRNA regulation influences phenotypic variability: perspectives on species-specific microRNA regulation. *Bioessays*. 35(10):881-8.
- Purves D, Augustine GJ, Fitzpatrick D, et al., editors. *Neuroscience*. 2nd edition. Sunderland (MA): Sinauer Associates; 2001. Neurotransmitter Synthesis. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK11110/>
- Reddy, K. B. (2015). MicroRNA (miRNA) in cancer. *Cancer. Cell. Int*. 15:38.
- Rossi AE, Dirksen RT (2006) Sarcoplasmic reticulum: the dynamic calcium governor of muscle. *Muscle Nerve*. 33(6):715-31.
- Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB (2008) Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. *Science*. 320(5883):1643-7.
- Shenoy, A., Belloch, R.H. (2014). Regulation of microRNA function in somatic stem cell proliferation and differentiation. *Nat. Rev. Mol. Cell. Biol*. 15(9):565-576.
- Walaas SI, Greengard P (1991) Protein phosphorylation and neuronal function. *Pharmacol Rev*. 43(3):299-349.

5. Considerações finais

A quantidade de trabalhos envolvendo o estudo do papel biológico dos miRNAs vem crescendo nos últimos anos. Nestes trabalhos, a predição *in silico* de alvos e o respectivo enriquecimento funcional dos mesmos estão quase sempre presentes. Entretanto, muito ainda se discute sobre a real eficiência de tais métodos, uma vez que em diversos casos trazem resultados pouco eluzivos ou de baixa qualidade. Neste contexto, os resultados apresentados neste trabalho poderão auxiliar os pesquisadores na realização destas duas etapas fundamentais durante a caracterização dos papéis biológicos desempenhados pelos miRNAs.

No capítulo I demonstramos que as últimas atualizações das ferramentas de predição de alvos forneceram valores de alta qualidade de especificidade e precisão, gerando predições quase totalmente livres de resultados falso-positivos. Estes dados, mostram que a técnica de intersecção, apesar de ainda ser muito adotada, atualmente não se faz necessária, trazendo prejuízos devido à perda da sensibilidade. Entretanto, os *cutoffs* utilizados pelas ferramentas excluem diversos alvos verdadeiros. Desta maneira, os desenvolvedores das ferramentas de predição de alvos devem agora investir na melhora da sensibilidade.

A união dos resultados das ferramentas TargetScan e miRanda-mirSVR obteve a elevada performance de 0.7 (em uma escala de 1 a -1), por agregar uma alta sensibilidade sem prejuízo de especificidade e precisão. Desta maneira, os resultados provenientes desta análise foram capazes de aprimorar a qualidade das análises *in silico* de predição de alvos, o que permitirá aos pesquisadores obterem resultados mais robustos durante a identificação de alvos dos miRNAs, economizando tempo e dinheiro em experimentos funcionais subsequentes.

No capítulo II, as análises de enriquecimento funcional, levando em consideração a intensidade da regulação fornecida pelos miRNAs, sugerem um novo nível de complexidade da atuação destas moléculas. Regular diversos genes associados a um mesmo processo biológico com intensidades semelhantes, enquanto processos biológicos distintos são regulados com intensidades diferentes é uma habilidade dos miRNAs que os torna capaz de regular de maneira precisa uma grande gama de funções biológicas distintas dentro de um mesmo contexto celular. Adicionalmente, nossas análises apontam que este é um fenômeno conservado entre as espécies de vertebrados analisadas, sugerindo que tal segregação regulatória vem sendo positivamente selecionada durante a evolução dos vertebrados.

O estudo aprofundado das funções biológicas reguladas pelos miRNAs miR-1, miR-7 e miR-9 trouxe também contribuições para a compreensão do comportamento dos miRNAs dentro da célula ou mesmo do organismo. Assim, a partir dos direcionamentos trazidos, novos estudos podem tentar identificar os padrões que causam a segregação da intensidade da regulação, o que possibilitaria a manipulação de uma via de ação dos miRNAs sem afetar as demais, minimizando os efeitos *off-target* comuns nos experimentos atuais.

6. Referências bibliográficas

- Agarwal V, Bell GW, Nam J, Bartel DP (2015) Predicting effective microRNA target sites in mammalian mRNAs. *eLife*. 4: e05005.
- Ambros, V (2004) The functions of animal microRNAs. *Nature*. 431: 350-5.
- Ameres & Zamore (2013) Ameres SL, Zamore PD. Diversifying microRNA sequence and function. *Nature Reviews Molecular Cell Biology*. 2013;14:475–488.
- Ashburner M., et al.. (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, 25, 25–29.
- Baek D, Villen J, Shin C, Camargo FD, Gygi SP, Bartel DP 2008 The impact of microRNAs on protein output. *Nature*.;455(7209):64–71.
- Bandyopadhyay, S., Mitra, R. (2009). TargetMiner: microRNA target prediction with systematic identification of tissue-specific negative examples. *Bioinformatics*. **25(20)**:2625-31.
- Bang C, Batkai S, Dangwal S, Gupta SK, Foinquinos A, Holzmann A, Just A, Remke J, Zimmer K, Zeug A, Ponimaskin E, Schmiedl A, Yin X, Mayr M, Halder R, Fischer F, Engelhardt S, Wei Y, Schober A, Fiedler J, Thum T (2014) Cardiac fibroblast–derived microRNA passenger strand-enriched exosomes mediate cardiomyocyte hypertrophy. *J Clin Invest*;124(5):2136–46.
- Bazzini AA, Lee MT, Giraldez AJ (2012) Ribosome profiling shows that miR-430 reduces translation before causing mRNA decay in zebrafish. *Science* 336: 233–237.
- Betel D, Wilson M, Gabow A, Marks DS, Sander C (2008) The microRNA.org resource: targets and expression. *Nucleic Acids Res*. 36: D149-53.
- Betel D, Koppal A, Agius P, Sander C, Leslie C (2010) Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Gen Biol*. 11:R90.
- Bleazard T, Lamb JA, Griffiths-Jones S (2015) Bias in microRNA functional enrichment analysis. *Bioinformatics*, 31, 1592–1598.
- Borchert GM, Lanier W, Davidson BL (2006) RNA polymerase III transcribes human microRNAs. *Nat Struct Mol Biol*. 13(12): 1097-1101.
- Broughton JP, Lovci MT, Huang JL, Yeo GW, Pasquinelli AE (2016) Pairing beyond the Seed Supports MicroRNA Targeting Specificity. *64(2)*:320-333.
- Clarke C, Henry M, Doolan P, Kelly S, Aherne S, Sanchez N, Kelly P, Kinsella P, Breen L, Madden SF, Zhang L, Leonard M, Clynes M, Meleady P, Barron, N (2012) Integrated

- miRNA, mRNA and protein expression analysis reveals the role of post-transcriptional regulation in controlling CHO cell growth rate. *BMC Genomics*. 13:656.
- Chi SW, Hannon GJ, Darnell RB (2012) An alternative mode of microRNA target recognition. *Nat Struct Mol Biol*. 19(3):321-7.
- Chou CH, Chang NW, Shrestha S, Hsu SD, Lin YL, Lee WH, Yang CD, Hong HC, Wei TY, Tu SJ, Tsai TR, Ho SY, Jian TY, Wu HY, Chen PR, Lin NC, Huang HT, Yang TL, Pai CY, Tai CS, Chen WL, Huang CY, Liu CC, Weng SL, Liao KW, Hsu WL, Huang HD (2016) miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. *Nucleic Acids Res*. 44(D1):D239-47
- Christodoulou F, Raible F, Tomer R, Simakov O, Trachana K, Klaus S, Snyman H, Hannon GJ, Bork P, Arendt D 2010. Ancient animal microRNAs and the evolution of tissue identity. *Nature*. 463: 1084–88.
- Devlin AH, Thompson P, Robson T, McKeown SR (2010) Cytochrome P450 1B1 mRNA untranslated regions interact to inhibit protein translation. *Mol Carcinog*. 49(2):190-9.
- Enright AJ, John B, Gaul U, Tuschl T, Sander C and Marks DS (2003) MicroRNA targets in *Drosophila*. *Genome Biology*. 5;R1.
- Fan, X., Kurgan, L. (2015) Comprehensive overview and assessment of computational prediction of microRNA targets in animals. *Brief Bioinform*. 16(5):780-94.
- Flynt, AS N; Li, EJ; Thatcher, L; Solnica-Krezel, JG; Patton, JG (2007) Zebrafish miR-214 modulates Hedgehog signaling to specify muscle cell fate. *Nat Gen*. 39: 259-63.
- Flynt, AS; Thatcher, EJ; Burkewitz, K; Li, N; Liu, Y; Patton, JG (2009) *miR-8* microRNAs regulate the response to osmotic stress in zebrafish embryos *J Cell Biol*. 185(1): 115-27.
- Gaidatzis D, Nimwegen E, Hausser J, Zavolan M (2007) Inference of miRNA targets using evolutionary conservation and pathway analysis. *BMC Bioinformatics*. 8:69.
- Garcia DM, Baek D, Shin C, Bell GW, Grimson A, Bartel DP (2011) Weak seed-pairing stability and high target-site abundance decrease the proficiency of *lscy-6* and other microRNAs. *Nature Structural & Molecular Biology*. 18:1139–1146.
- Grimson A, Farh KK, Johnston WK, Garrett-Engle P, Lim LP, Bartel DP (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell*. 27(1):91–105.
- Grimson A, Srivastava M, Fahey B, Woodcroft BJ, Chiang HR, King N, Degan BM, Rokhsar DS, Bartel DP (2008) Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. *Nature*. 455:1193-1197.

- Gusev Y, Schmittgen TD, Lerner M, Postier R, Brackett D (2007) Computational analysis of biological functions and pathways collectively targeted by co-expressed microRNAs in cancer. *BMC Bioinformatics*. 8:S16.
- Hausser J, Syed AP, Bilén B, Zavolan M. (2013) Analysis of CDS-located miRNA target sites suggests that they can effectively inhibit translation. *Genome Res*. 23(4):604-15.
- Jia W, Li Z, Lun Z (2008) Discoveries and functions of virus-encoded MicroRNAs. *Chinese Science Bulletin*. 53:169–177.
- John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS (2004). Human MicroRNA targets. *PLoS Biol*. 2:e363.
- Jones-Rhoades MW, Bartel DP, Bartel B (2006) MicroRNAs and their regulatory roles in plants. *Annu Rev Plant Biol*. 57:19-53.
- Kanehisa M and Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. 28: 27–30.
- Karginov FV, Cheloufi S, Chong MM, Stark A, Smith AD, Hannon GJ (2010) Diverse endonucleolytic cleavage sites in the mammalian transcriptome depend upon microRNAs, Drosha, and additional nucleases. *Mol. Cell* 38, 781–788.
- Kertész M, Iovino N, Unnerstall U, Gaul U, Segal E (2007) The role of site accessibility in microRNA target recognition. *Nature Genetics*;39:1278–1284.
- Krek A, Grun D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, MacMenamin P, da Piedade I, Gunsalus KC, Stoffel M, Rajewsky N. (2005) Combinatorial microRNA target predictions. *Nat Genet*.;37(5):495–500.
- Krutzfeldt J, Poy MN, Stoffel M (2006) Strategies to determine the biological function of microRNAs. *Nat Genet, Suppl*:S14-9.
- Lanet E, Delannoy E, Sormani R, Floris M, Brodersen P, Crete P, Voinnet O, Robaglia C (2009) Biochemical evidence for translational repression by Arabidopsis microRNAs. *Plant Cell*: 21, 1762–8.
- Lee RC, Feinbaum RL, Ambros V (1993) The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell* 75:843–54.
- Lee Y, Han J, Yeom KH, Jin H, Kim VN (2006) Drosha in primary microRNA processing. *CSH Symp Quant Biol*; 71:51-7.
- Lee YS, Dutta A (2009) MicroRNAs in cancer. *Annu Rev Pathol*, 4:199-227.
- Lewis BP, Burge CB, Bartel DP (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*;120:15–20.

- Liu N, Olson EN (2010) MicroRNA Regulatory Networks in Cardiovascular Development. *Dev Cell.* 18(4):510-25.
- Loher P and Rigoutsos I (2012) Interactive exploration of RNA22 microRNA target predictions. *Bioinformatics* 28, 3322–23.
- Lund E, Güttinger S, Calado A, Dahlberg JE, Kutay U (2004) Nuclear export of microRNA precursors. *Science*, 303(5654):95-98.
- Lytle JR, Yario TA, Steitz JA (2007) Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5' UTR as in the 3' UTR. *Proc Natl Acad Sci U S A.* 104(23):9667-72.
- Mathonnet G, Fabian MR, Svitkin YV, Parsyan A, Huck L, Murata T, Biffo S, Merrick WC, Darzynkiewicz E, Pillai RS, Filipowicz W, Duchaine TF, Sonenberg N. (2007) MicroRNA inhibition of translation initiation in vitro by targeting the cap-binding complex eIF4F. *Science* 317, 1764–1767.
- Molnár A, Schwach F, Studholme DJ, Thuenemann EC, Baulcombe DC (2007) miRNAs control gene expression in the single-cell alga *Chlamydomonas reinhardtii*. *Nature.* 447(7148):1126-9.
- Nielsen CB, Shomron N, Sandberg R, Hornstein E, Kitzman J, Burge CB (2007) Determinants of targeting by endogenous and exogenous microRNAs and siRNAs. *RNA.* 13:1894–1910.
- Orom UA, Nielsen FC, Lund AH (2008) MicroRNA-10a binds the 5'UTR of ribosomal protein mRNAs and enhances their translation. *Mol Cell.* 30(4):460-71.
- Parikh, R., Mathai, A., Parikh, S., Chandra Sekhar, G., & Thomas, R. (2008). Understanding and using sensitivity, specificity and predictive values. *Indian J Ophthalmol.* **56(1)**:45–50.
- Powers, D.M.W. (2007). Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies.* 2(1): 37-63.
- Rand TA, Petersen S, Du F, Wang X (2005) Argonaute2 cleaves the anti-guide strand of siRNA during RISC activation. *Cell*, 123(4):621-9.
- Rehmsmeier M, Steffen P, Hochsmann M, Giegerich R (2004). Fast and effective prediction of microRNA/target duplexes. *RNA* 10, 1507–1517.
- Reczko M, Maragkakis M, Alexiou P, Grosse I, Hatzigeorgiou AG (2012) Functional microRNA targets in protein coding sequences. *Bioinformatics.* 28:771–776.
- Reddy, K. B. (2015). MicroRNA (miRNA) in cancer. *Cancer. Cell. Int.* **15**:38.

- Schirle NT, Sheu-Gruttadauria J, MacRae IJ (2014) Structural basis for microRNA targeting. *Science*. 346:608–613.
- Shenoy, A., Belloch, R.H. (2014). Regulation of microRNA function in somatic stem cell proliferation and differentiation. *Nat. Rev. Mol. Cell. Biol.* **15**(9):565-576.
- Shin C, Nam J, Farh KK, Chiang HR, Shkumatava A, Bartel DP (2010) Expanding the microRNA targeting code: functional sites with centered pairing. *Mol Cell*; 38, 789–802.
- Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB (2008) Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. *Science*. 320(5883):1643-7.
- Schnall-Levin M, Zhao Y, Perrimon N, Berger B (2010) Conserved microRNA targeting in *Drosophila* is as widespread in coding regions as in 3'UTRs. *Proc Natl Acad Sci U S A*. 107(36):15751-6.
- Shkumatava A, Stark A, Sive H, Bartel DP (2009) Coherent but overlapping expression of microRNAs and their targets during vertebrate development. *Genes Dev* 23: 466–481.
- Sturm M, Hackenberg M, Langenberger D, Frishman D (2010) TargetSpy: a supervised machine learning approach for microRNA target prediction. *BMC Bioinformatics*. 11.
- Sunkar R, Girke T, Jain PK, Zhu JK (2005) Cloning and Characterization of MicroRNAs from Rice. *Plant Cell*. 17:1397-1411.
- Takacs CM and Giraldez AJ (2011) miR-430 regulates oriented cell division during neural tube development in zebrafish. *Dev Biol*. 409(2):442-450.
- Tang G, Reinhart BJ, Bartel DP, Zamore PD (2003) A biochemical framework for RNA silencing in plants. *Genes Dev*; 17, 49–63.
- Tay Y, Zhang J, Thomson AM, Lim B, Rigoutsos, I (2008) MicroRNAs to Nanog, Oct4 and Sox2 coding regions modulate embryonic stem cell differentiation. *Nature*. 455(7216):1124-8.
- Witkos TM, Koscianska E, Krzyzosiak WJ (2011) Practical Aspects of microRNA Target Prediction. *Curr Mol Med*. 11(2): 93–109.
- Zdanowicz A, Thermann R, Kowalska J, Jemielity J, Duncan K, Preiss T, Darzynkiewicz E, Hentze MW (2009) *Drosophila* miR2 primarily targets the m⁷GpppN cap structure for translational repression. *Mol. Cell* 35, 881–888.
- Zhou H, Rigoutsos I (2014) MiR-103a-3p targets the 5' UTR of GPRC5A in pancreatic cells. *RNA*. 20(9):1431-9. doi: 10.1261/rna.045757.114.