Computational approaches for the study of micro-RNA annotation, function, and evolution

MICKAEL LECLERCQ

Doctor of Philosophy

School of computer Science

McGill University
Montréal, Quebec, Canada
March 2016

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

DEDICATION

This thesis is dedicated to my grandfather, Jean-Marc Leclercq, alias Papinou, died in 2007. Without him, I'm not sure that I would be in Quebec today.

AKNOWLEDGMENTS

I would like to thank everyone whom I had the chance to share new ideas during this PhD. I'm deeply grateful to my supervisor Mathieu Blanchette, an amazing scientist and a great professor, for his precious insights and encouragement. He always sees things in a positive way, which helped me a lot to go forward. I will never be thankful enough to him, considering the huge amount time he spent to drive me through this PhD, where he taught me so much. For the record, I'll always remember our meetings at the coffee Expression on Mont-Royal Avenue to prepare the comprehensive exam. Finally, he was also a great support in my personal life. I had two child during this PhD and Mathieu always tried to arrange our schedule by taking this situation into account, especially around the birth periods!

I also want to thank Abdoulaye Banire Diallo for his support and encouragement. It is because he believed in me that he introduced me to Mathieu, his previous mentor when he was a student at McGill. He taught me a lot during my master degree, and helped me to obtain the PhD scholarship. He continued to follow me as co-director of my doctorate, providing me amazing advises for improving my work.

For my funding, I thank the "fond de recherche nature et technologies" and Mathieu's funding resources for their support.

Finally, I want to thank my family, especially my wife, my parents, my grandparents and my godfather. Without their support, this work wouldn't exist. And for my children, Juliette and Ludovik, I hope one day this work will inspire them to go as far as they can in life!

TABLE OF CONTENTS

LIST OF TABLES

Table **Page**

Supplementary Data table

LIST OF FIGURES

Figure **Page**

Supplementary Data figures

LIST OF ABBREVIATIONS

DNA	Deoxyribonucleic acid
LCA	Last common ancestor
MFE	Minimum Folding Energy
miRNA	MicroRNA
MTG	MicroRNA target gene
ncRNA	Non coding RNA
Pre-miRNA	MicroRNA precursor
RNA	Ribonucleic acid
SSTSP	Single-sequence target site predictors
SVM	Support vector machine
tRNA	Transfert RNA

ABSTRACT

MicroRNAs (miRNAs) are short RNA species derived from hairpin-forming miRNA precursors (pre-miRNA) and acting as key post-transcriptional regulators by silencing specific messenger RNAs (mRNAs). They are involved in virtually every biological process of multicellular eukaryotes and are highly conserved throughout evolution. In this thesis, I first present miRdup, a computational approach that accurately predicts the position of the mature miRNA sequence on its precursor in a species-dependent manner. MiRdup is not only more accurate than the few tools that existed at the time of its publication, but it is also significantly more flexible and applicable to a wide range of species. Then, I study when and how human miRNAs originated among primate ancestors. Existing studies based on last common ancestor (LCA) analysis show miRNA accrual in metazoan genomes through time and a rarity of loss but cannot provide the ancestral genomes required to identify evolutionary pathways leading to their creation. To address this, I inferred the ancestral genomes of the mammal's ancestors and predicted ancient pre-miRNAs and mature miRNAs to classify them in various types of mechanisms of origination. Remarkably, I found that a large fraction of primate-specific miRNAs is due to the accumulation of substitutions and small insertions. Finally, in continuity to the spirit of improving existing research in the field of the miRNAs, I present a new method that increases the miRNA target gene prediction accuracy of existing tools in human with the help of ancestral reconstruction. Many approaches already exist for this purpose, but none has yet exploited the potential of ancestral genomes. My results exceed the recall of the best existing tools. This whole work brings new ideas that could be applied in the prediction of other DNA or RNA functional elements, and improve the understanding of miRNAs' evolution.

Keywords: miRNAs prediction, target gene prediction, machine learning, ancestral reconstruction

ABRÉGÉ

Les microARNs sont de courts ARNs dérivés de précurseurs en forme d’épingle à cheveux qui agissent comme régulateurs post-transcriptionnels par inactivation des ARNs messagers. Ils sont impliqués dans quasiment tous les processus biologiques des eucaryotes multicellulaires et sont très conservés tout au long de l’évolution. Dans cette thèse, je présente miRdup, une approche algorithmique qui prédit avec précision la position de la séquence de microARNs matures sur leur précurseur en prenant compte les caractéristiques spécifiques de chaque espèce. MiRdup est non seulement plus précis que les outils qui existaient au moment de sa publication, mais il est aussi beaucoup plus polyvalent et applicable à un large éventail d'espèces. Puis, j'étudie quand et comment microARNs humains sont apparus chez les ancêtres des primates. Les études existantes basées sur l’analyse du plus petit ancêtre commun montrent que les microARNs se sont accumulés dans les génomes de métazoaires à travers le temps, en étant rarement éliminés, mais cette technique ne peut produire les génomes ancestraux nécessaires pour identifier les mécanismes d'évolution menant à leur création. Pour y remédier, j’ai reconstruit les génomes ancestraux des ancêtres de mammifères et prédit les anciens pré-microARNs et microARNs matures pour les classer dans différents types de mécanismes d'origine. Remarquablement, j’ai constaté qu'une grande partie des microARNs spécifiques aux primates est due à l'accumulation de substitutions et de petites insertions. Enfin, dans le but de continuer à améliorer la recherche dans le domaine des microARNs, je présente une nouvelle méthode qui augmente la précision des outils de prédiction des gènes cibles des microARNs chez l'homme, avec l'aide de la reconstruction ancestrale. De nombreuses approches existent déjà à cet effet, mais aucune n'avait encore exploité le potentiel des génomes ancestraux. Nos résultats dépassent le taux de rappel des meilleurs outils existants. L'ensemble de ce travail apporte de nouvelles idées qui pourraient être appliquées pour la prédiction d'autres éléments fonctionnels de l’ADN ou de l’ARN, et améliore la compréhension de l'évolution de microARNs.

Mots clés: Prédiction de microARNs, prédiction de gènes cibles, apprentissage machine, reconstruction ancestrale

CHAPTER I : INTRODUCTION

Until the beginning of the 1990s, molecular biologists had an understanding of cells where proteins were the key active players. In 1989 Ambros's team discovered a new type of small RNA molecule, known since 2001 as microRNAs, which appeared to affect the expression of other genes. Today, more than 2500 of such molecules are known in human, and they play key roles in most of physiological processes. However, many aspects of the annotation of miRNAs in genomes, the identification of the genes they regulate, and the way they evolved remains poorly understood. In this thesis, I propose novel computational approaches to improve the prediction of mature miRNAs and their target genes, and bring a better understanding of the human microRNA evolution through the study of ancestral mammal genomes.

In this first chapter, we present the central dogma of biology required to understand what microRNAs are and what their role in living organisms and evolution is. It also describes what is known about miRNA function and their role in human health and disease. Since this thesis focuses on the computational prediction of microRNAs and their target genes, three sections of this chapter are devoted to existing bioinformatics approaches to calculate RNA secondary structures and predict microRNAs and their target genes[1].

[1] It is important to note that biology described is this chapter is based on 2016 knowledge. Thus, some parts may change in the future. This is the meaning of science, it can evolve. In biology, everything is possible and rare cases almost always exist, especially in genetic. Consequently, every statement must be interpreted as "generally".

1.1 Genomic principles and microRNAs

To understand the topics covered in this thesis, this section briefly introduces the basic notions related to DNA, RNA, non-coding RNAs and proteins, and describes the current knowledge on microRNAs.

1.1.1 DNA and genomes

The essence of life is encoded in a molecule, present in every living organism's cells, called deoxyribonucleic acid (DNA). This molecule is organized into long antiparallel double-stranded chains of nucleotides, also called nitrogenous bases (or nucleotides), which compose the alphabet of life, A: adenine, G: guanine, C: cytosine and T: Thymine. Nucleotides are organized such that a specific genome exists for each living organism. This code includes regions of variable lengths, called genes, which contains the instructions required for guiding the production of proteins. "A gene is a union of genomic sequences encoding a coherent set of potentially overlapping functional products" (Gerstein et al. 2007), implying that the gene encodes not only the amino acid sequence (i.e. the protein), located in exons, but also contains introns and regulatory regions, including enhancers or silencer sites, promoter, 5' and 3' untranslated regions (non-coding sequence), and poly-A tail (Figure I-1).

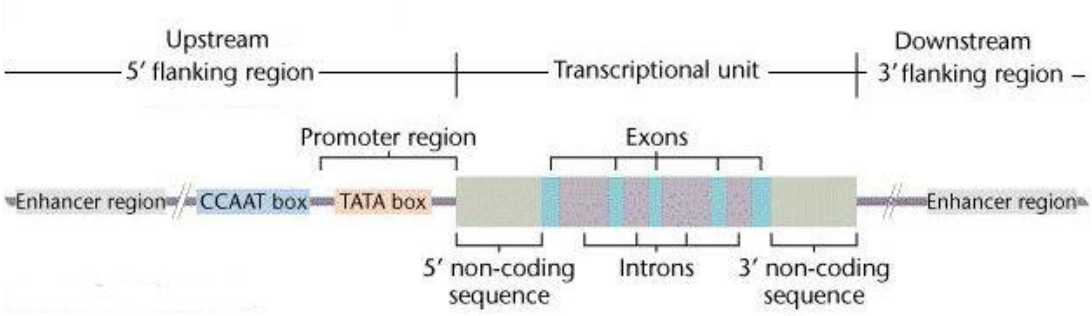


Figure I-1: Organization of a eukaryotic protein-coding gene region (Klug and Cummings 1997)

An organism is then morphologically and physiologically defined by its genome, which is the genetic material of each cell. The size of a genome may vary from few thousand

nucleotides for unicellular organisms or viruses, to billions for others. The size of the human genome is approximately 3.0×10^9 base pairs (bp). Note that genome size is not directly linked to organism complexity, as some unicellular species, like *Amoeba dubia*, or some plant species, such as *Paris japonica*, have genomes 50 to 200 times larger than that of humans (Pellicer et al. 2010). Nucleotides assembled one after the other form long DNA sequences called chromosomes, whose number varies from species to species (e.g. human has 23 pairs of chromosomes). Chromosome structure is presented in Figure I-2.

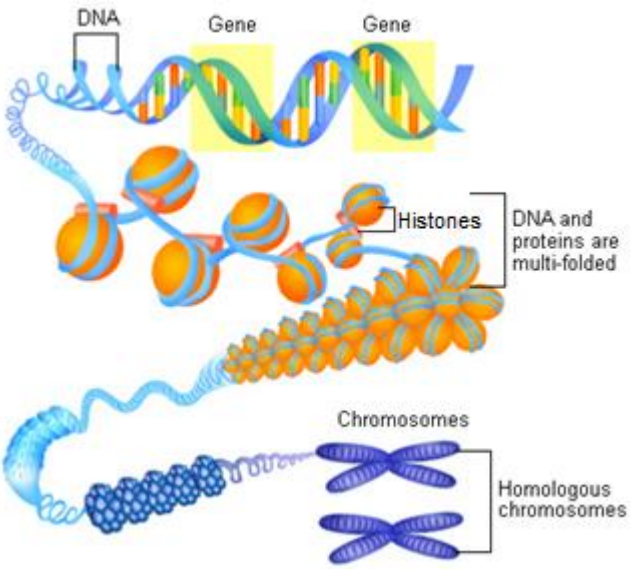


Figure I-2: Chromosome structure (Alberts 2002)

1.1.2 From DNA to proteins

Proteins are large and complex biological molecules consisting of an assembly of one of more chains of amino acids, which are organic compounds composed of carbon, hydrogen, oxygen, and nitrogen. Proteins are involved in virtually all cells functions in living organisms, including DNA replication, structural support, stimulus response, stress response, molecule transport, storage, etc. Specialized proteins include antibodies, enzymes or hormones (Alberts 2002).

Proteins are synthesized in cells through a process called gene expression. The main steps leading to the production of a protein from DNA in a eukaryotic cell are summarized in Figure I-3. In the nucleus of a cell, certain parts of chromosomes called promoters, usually located close to genes, allow the attachment of RNA polymerases. These polymerases transcribe[2] the DNA in RNA molecules, called pre-messenger RNAs (pre-mRNA). After splicing[3] (RNA processing), pre-mRNA becomes a mature messenger RNA (mRNA), which serve as a template readable by ribosomes. Once exported from the nucleus, i.e. in the cytoplasm of the cell, nucleotides are read by ribosomes three by three (codon) and translated into amino acids chains (polypeptides). This is the translation step. Finally, by a complex series of folding and other chemical modifications, called protein maturation, amino acids chains become functional proteins.

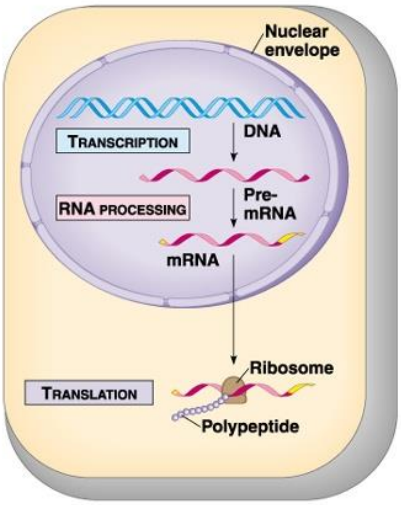


Figure I-3: From DNA to protein (Alberts 2002)

[2] Pre-messenger RNA (pre-mRNA) transcript modification in which introns are removed and exons are joined.
[3] Transcription is the process in which a particular DNA segment is copied into a messenger RNA by the enzyme RNA polymerase.

1.1.3 Ribonucleic Acids (RNAs) biology and structure

RNA molecules are implicated in various biological roles, including protein coding (mRNA), chemical modifications guiding of other RNAs (small nucleolar RNA), translation (transfer RNAs) and gene expression regulation (microRNAs). In eukaryotes, a very small part of the genome codes for proteins. In mammals, exons account for less than 2% of the genome. Portions of the remaining DNA codes for short and long non-coding RNAs (Taft et al. 2010). After years of research accomplished successfully to annotate the protein-coding regions of the genomes (i.e. the genes), research turned to the annotation of the non-coding regions (Alexander et al. 2010).

In eukaryotes, RNA is produced by the transcription of the DNA in the cell nucleus. As DNA, RNA molecules are long nucleotide chains. The chemical structure of both molecules is somewhat similar but differs on three points: (i) RNA molecules are single-stranded, (ii) Thymine is replaced by Uracile (U), although it has the same affinity properties for adenine as thymine and (iii) while DNA nucleotides contains deoxyribose, RNA nucleotides contains ribose, which is less stable and prone to hydrolysis.

RNA has the capacity to fold onto itself due to chemical linkages naturally occurring between nucleotides, leading to complex secondary and tertiary structures. Secondary structure of a nucleic acid molecule is the representation of the set of base pairing interactions within single or interacting molecules. Tertiary structure of a nucleic acid is a three-dimensional structure representation defined by the atomic coordinates of all bases. Folded molecules are near-stable structures maintained by hydrogen bonds between pairs of nucleotides. A molecule can have a dynamic landscape of various structures with probabilities that are dictated by their free energies, plus interaction with other molecules. Usually one or a few structures are more abundant due to favorable free energies. Canonical base pairs are: A-U, G-C, and G-U. Pairs A-U and G-C are called Watson-Crick (Watson and Crick 1953), and G-U is called Wobble (Crick 1966). From the most to the least robust linkage stability, there is the G-C pair,

followed by A-U and G-U pairs. Some exceptions exist, called non-canonical base pairs, in RNA structures. Many structural elements can coexist in a secondary structure, as shown in Figure I-4. These include helices (or stacks), hairpins (or terminal) loops, internal loops, exterior loops, bulge loops, multi-branch loops (junction), and pseudo-knots (Dardel and Kapas 2002). For example, a well-known family of non-coding RNA, the transfer RNA, has a clover shape similar to Figure I-4, capable of attaching a codon from an anti-codon loop (located at the top of the hairpin loop in the Figure I-4) and delivering an amino acid (located at the exterior loop in the Figure I-4).

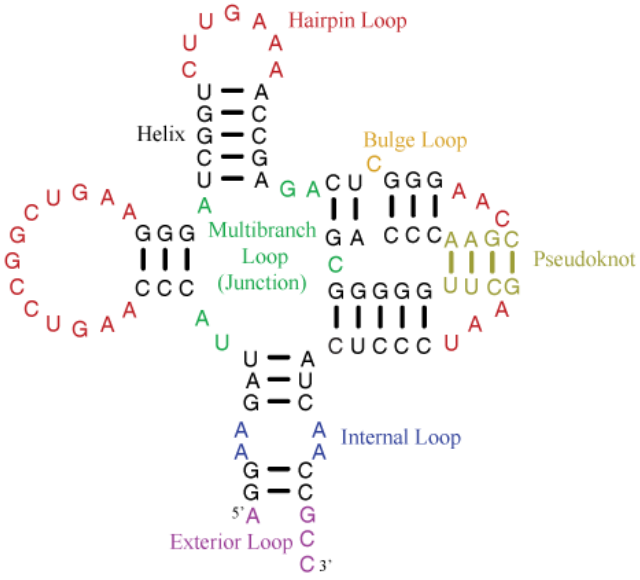


Figure I-4: Structural elements composing a RNA secondary structure. Source: Nearest Neighbor Database Help (Turner and Mathews 2009)

1.1.4 Non-coding RiboNucleic Acids

Since almost 50 years, numerous regulatory RNAs of all shapes and sizes have been discovered (Figure I-5). In eukaryotes, non-coding RiboNucleic Acids (ncRNAs) are involved in various epigenetic processes (Kaikkonen et al. 2011), including transcriptional and post transcriptional silencing (Malecová and Morris 2010), germ cell reprogramming (Guan et al. 2013), germinal maintenance (Saxe and Lin 2011),

development and differentiation (Fatica and Bozzoni 2014), antiviral defence (Ouellet and Provost 2010), transposon silencing (Costa 2008), chromatin remodeling (Saxena and Carninci 2011), X chromosome inactivation (Gontan et al. 2011), etc. They typically are tissue and developmental stages specific (Koerner et al. 2009; Szymanski et al. 2005).

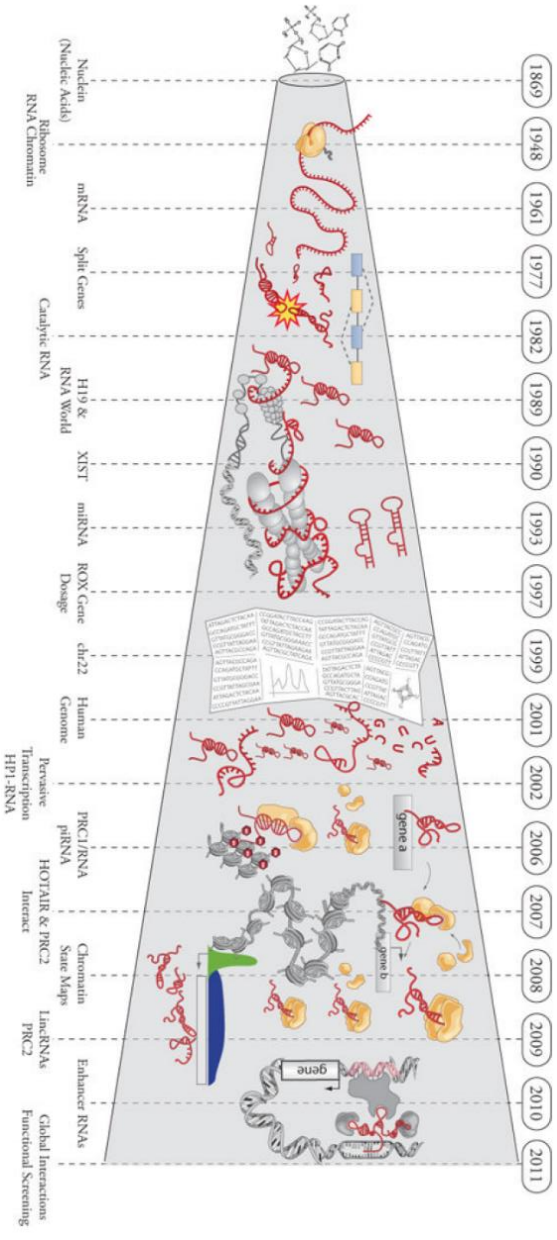


Figure I-5: Non-coding RNAs timeline discoveries (Rinn and Chang 2012)

There exist about twenty types of ncRNAs (Taft et al. 2010; Esteller 2011) (examples in Table I-1), which differ in their sequence length, structure and function. One of the key roles of some of the ncRNAs types is gene silencing, a process called RNA interference (RNAi). NcRNAs classes known to be involved in RNAi are small interfering RNAs (siRNAs), microRNAs (miRNAs) and piwi-interacting RNAs (piRNAs), which are all present in both plants and animals. New types of interfering RNAs have been recently discovered, including promoter-associated small RNAs (PASRs) (Kapranov et al. 2007) and transcription initiation RNAs (tiRNAs) (Taft et al. 2009). To give an idea of the scale, scientists have currently identified more than 2500 miRNAs, hundreds of siRNAs, and millions of piRNAs sequences in the human genome alone. All those sequences are unique, often tissue and developmental stage specific, and testify that those RNAs have a wide range of regulatory functions facilitated by sequence-specific interactions. Due to the fact that some ncRNA families, such as miRNAs, have members that are expressed in very particular conditions, it is assumed that new ones are still to be discovered.

Table I-1: Examples of RNA types

Type	Full name	Role	Definition	Distribution	Reference of discovery
lncRNA	Long non-coding RNA	Gene transcription, epigenetic and post-transcriptional regulation. X-chromosome inactivation in mammals miRNAs sponges	ncRNA that mirror protein-coding genes by sharing common characteristics such as the length (2 to 100 kb) and the presence of polyadenylation signals. Several thousand lncRNAs are currently identified in mammals, where they are abundantly transcribed (Taft et al. 2010). LncRNAs perform epigenetic modifications by recruiting chromatin remodeling complexes to specific loci, and regulate chromatin accessibility by recruiting histone modification enzymes and RNA polymerases (Esteller 2011). Finally, lncRNAs can function as endogenous miRNA sponges, to compete with inhibition of mRNAs (Xia et al. 2014).	Eukaryotes	(Ponting et al. 2009)
miRNA	microRNA	Post-transcriptional regulation	ncRNA of about 22 nt encoded in almost all genomic regions, estimated to regulate the translation of two-thirds of protein-coding genes, making them an important regulator of many physiological processes (Esteller 2011). See section 1.1.5 for more details.	Eukaryotes	(Lee et al. 1993)
mRNA	Messenger RNA	Protein coding	Large family of RNA molecules of various lengths (dozen to thousands nt) created after DNA transcription that carry the genetic information that specifies the amino acid sequence of proteins during translation steps (Voet and Voet 2010)	All organisms	(Brenner et al. 1961; Jacob and Monod 1961; Gros et al. 1961)
PASRs	Promoter-associated small RNAs	Regulation of gene transcription	ncRNA of 22-200 nt located in 5′ regions of protein-coding genes and associated with the transcriptional start sites of genes (Esteller 2011).	Eukaryotes	(Kapranov et al. 2007)
piRNA	Piwi-interacting RNA	Epigenetics by blocking retro-transposons, DNA methylation	ncRNA of 24-31 nt located in intragenic regions, dicer-independent and bind the PIWI subfamily of Argonaute family proteins, which maintain the genome stability in germline cells (Esteller 2011).	Animals	(Seto et al. 2007)
rRNA	Ribosomal RNA	Translation	RNA component of the ribosome that contains two major rRNAs, a large and a small subunit of few hundred to thousands bases, with more than 50 proteins. It is an essential element of protein synthesis in living organisms (Voet and Voet 2010)	All organisms	(Nissen et al. 2000)

siRNA	Silencing RNA	Post-transcriptional regulation	ncRNA of about 20-25 nt, double-stranded, act as guide RNA to induce specific RNA degradation by breaking down the target after its transcription (Agrawal et al. 2003).	Eukaryotes	(Hamilton and Baulcombe 1999)
snoRNA	Small nucleolar RNAs	rRNA modifications (i.e. methylation) Gene regulation	ncRNA of 60-300 bp located in intronic regions that comprise two families, the C/D and H/ACA RNAs (Matera et al. 2007). They are components of small nucleolar ribonucleoproteins responsible for methylation and pseudouridylation of rRNAs in the nucleus to facilitate its folding and stability (Esteller 2011).	Eukaryotes	(Kiss 2001)
snRNA	Small nuclear RNA	Splicing, maintaining telomeres, transcription factor regulation	ncRNA of about 150 nt highly abundant, that function in the nucleoplasm divided in two classes which differ in their LSm (antigen) protein binding sites: Sm-class and Lsm-class RNAs (Matera et al. 2007).	Eukaryotes	(Birnstiel and Schaufele 1988)
tRNA	Transfer RNA	Translation	ncRNA of 75-90 nt, which serves as an adaptator molecule between RNA and protein amino acids during the translation process by ribosomes (Voet and Voet 2010).	All organisms	(Plescia et al. 1965)
tiRNAs	Transcription initiation RNAs	Regulation of gene transcription	Short ncRNA of 17 to 18 nt located downstream of transcriptional start sites and associated with highly expressed transcripts and sites of RNA polymerase II binding (Taft et al. 2009; Esteller 2011).	Eukaryotes	(Taft et al. 2009)
tmRNA	Transfer-messenger RNA	Recycling of stalled ribosomes	ncRNA of about 300 bp found in all eubacteria, chloroplasts and mitochondria. It is recruited by arrested ribosomes in the middle of protein biosynthesis (e.g. end of mRNA with lost stop codon), leading to degradation of aberrant mRNAs (Keiler and Ramadoss 2011; Keiler 2008).	Prokaryotes	(Wower and Zwieb 2000)

1.1.5 Micro Ribonucleic Acids (microRNAs)

MicroRNA were discovered in 1993, when Dr Ambros' team discovered single-stranded non-protein-coding regulatory RNA molecules in the nematode *Caenorhabditis elegans* (Lee et al. 1993). The term "microRNA" was set in 2001 (Ambros 2001) and since then, research on this topic increases every year, with more than 37,000 published papers containing the "microRNA" keyword (Figure I-6).

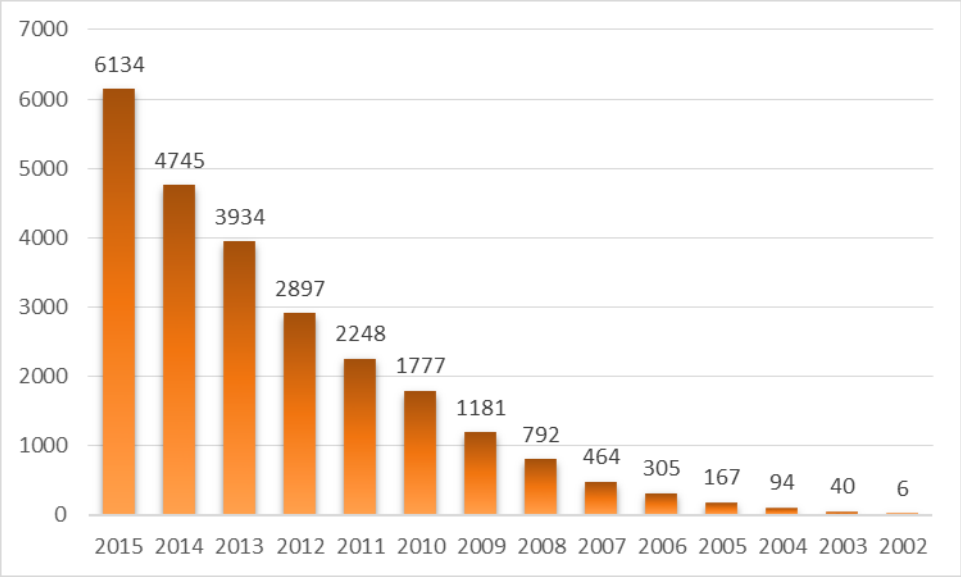


Figure I-6: Number of published papers per year referenced in PubMed containing the keyword "microRNA" in the abstract (Data obtained in January 2016).

These molecules are characterized by a short sequence, generally 19 to 24 nucleotides, involved in post-transcriptional regulation by targeting messenger RNAs in eukaryotes (Ruvkun 2001; Swami 2010). A microRNA gene is first transcribed to a primary miRNA, cleaved by an enzyme called Drosha which results in a sequence of about 100 nt, the precursor (pre-miRNA), which contains the mature miRNA. The precise mechanisms of biogenesis of miRNAs are discussed in the next subsection. MiRNA genes can be located in the exons or intron of protein-coding genes, or in intergenic regions, in which case they need their own promoter (Lu et al. 2005). Once separated

from its precursor, a miRNA has the capacity to hybridize to an mRNA, blocking the production of the corresponding protein by translational repression or target degradation (more details in subsection 1.1.5.3). In humans, miRNAs control the activity of at least ~50% of all protein-coding genes and a given miRNA can target several hundred mRNAs (Krol et al. 2010). Finally, a large portion of the miRNA families are shared between species, reflecting one of most important characteristic of the miRNAs: their conservation across evolution and co-evolution with their target genes due to negative selective pressure (Berezikov 2011).

1.1.5.1 Structure of microRNAs

MicroRNAs are distinguished by their particular structure. MiRNA precursors have the characteristic to fold into a hairpin shape (Example of has-miR-1 in Figure I-7), i.e. two arms containing a miRNA duplex, an extension and a pri-extension, one head loop (rarely two) and multiple bulges/internal loops due to mismatches (Figure I-8). This shape is explained by the presence of an antiparallel motif, i.e. inverted repeated sequence. Experimental observation of such structures are detected by X-ray crystallography and NMR spectroscopy (Brünger et al. 1998).

Figure I-7: Example of mature human microRNA miR-1 in its precursor

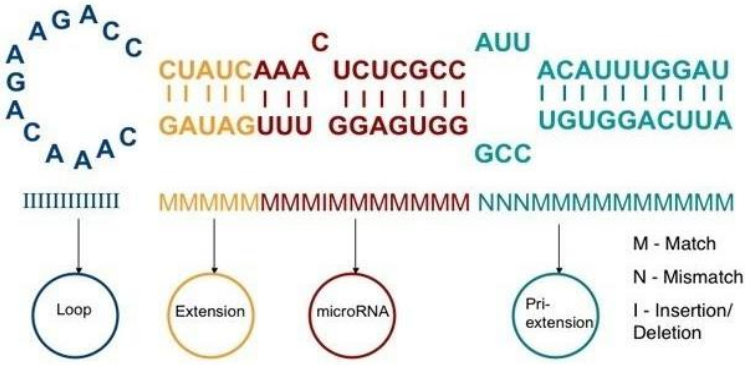


Figure I-8: General representation microRNA precursor's regions (Kadri et al. 2009)

1.1.5.2 Biogenesis of microRNAs

MiRNAs biogenesis is a multi-step process (Figure I-9). In the nucleus, a gene coding for a miRNA is transcribed by RNA polymerase II or III[4] to become a primary microRNA (pri-miRNA). The pri-miRNA fold into hairpin to act as a substrate to a protein complex, which includes Drosha III (ribonuclease) and DGCR8 (binding protein also named Pasha), that cleave it to become a miRNA precursor (pre-miRNA). Once exported in the cytoplasm by the protein complex composed of Exportin-5, Ran and GTP[5] (Ha and Kim 2014), the pre-miRNA is cleaved by Dicer (endonuclease with N-terminal[6] helicase domain interacting with the pre-miRNA terminal loop) and the cofactor TRBP to generate a duplex miRNA 5':miRNA 3' (also called miRNA:miRNA* in the official nomenclature before 2012) (Krol et al. 2010). In plants, Drosha and Dicer are replaced by Dicer Like 1 (DCL1) (Cuperus et al. 2011). Each strand of the duplex is about 21 nt. The miRNA in 3' of the duplex is approximately the complementarity sequence of the miRNA in 5', with a shift of

[4] Polymerase III specifically synthesizes small non-coding RNAs that are linked to regulating cell cycle (Dumay-Odelot et al. 2010) and growth (Goodfellow and White 2007)

[5] GTP (Guanosine triphosphate) is hydrolyzed during the passage through the nuclear membrane to become a GDP. Once in the cytoplasm, Exportin-5 and Ran-GDP are separated from the pri-miRNA (Ha and Kim 2014)

[6] N-terminal is the start of a protein or polypeptide terminated by an amino acid with a free amine group ($-NH_2$)

generally two nucleotides in 3’. This complementarity is not always perfect, since the duplex often includes bulges. The functional miRNA is generally the one in the 5’ portion of the duplex, although there are cases where it is the 3’ portion and even cases where both the 5’ and 3’ portions are functional. The two strands of the duplex are separated by the miRNA induced silencing complex (RISC), and the non-functional strand is degraded (Hackenberg et al. 2009). Rules for strand selection (i.e. 3’or 5’ duplex strand) are determined by AGO2, part of RISC complex, during its loading step, including unstable terminus at the 5′ side and a presence of uracil at position 1 of the strand (Ha and Kim 2014). Finally, RISC guides the miRNA to a targeted messenger RNA (mRNA) to repress its expression by one of two ways, which are discussed in 1.1.5.3 : translation inhibition and mRNA degradation, which includes mRNA target cleavage, and mRNA deadenylation (Krol et al. 2010).

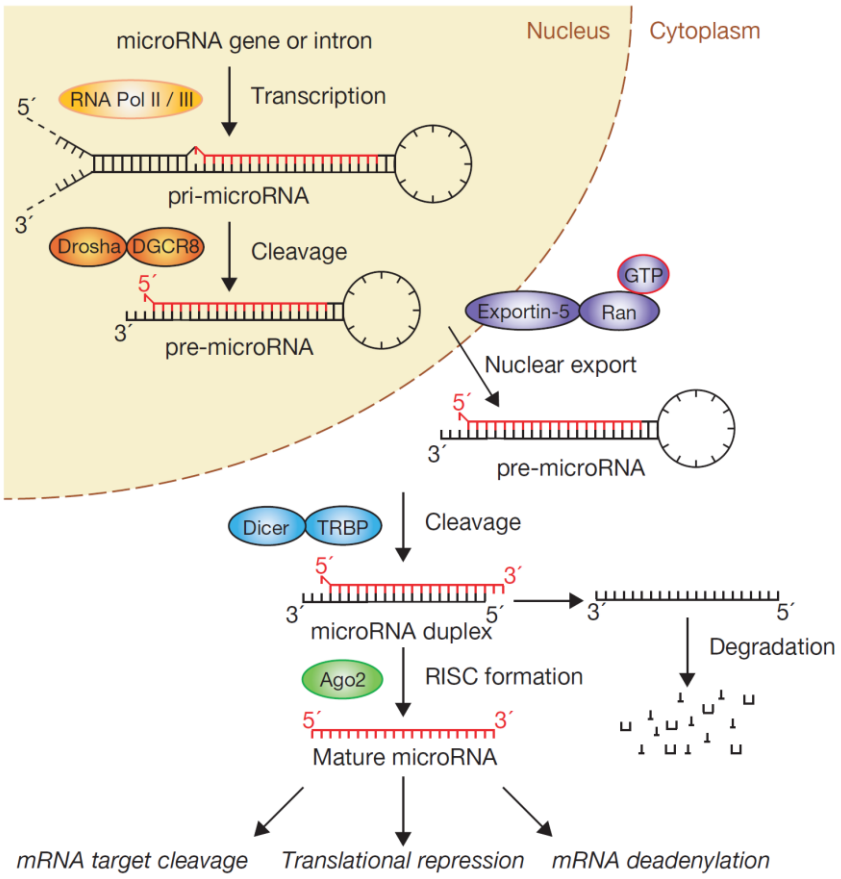


Figure I-9: microRNA pathway processing in animals (Winter et al. 2009)

1.1.5.3 MicroRNAs target genes

The role of miRNAs is to silence specific genes called target genes, i.e. to reduce the amount of protein copies that will be produced from them. One single miRNA has the potential to silence many genes, and silenced genes can be targeted by one or more miRNAs (Gennarino et al. 2012). Unlike in plants, where the silencing requires a near-perfect complementarity between the miRNA and its targeted mRNA, the repression of mRNA expression in animals is determined by complementarity of a very short region of the miRNA, called the seed (Figure I-10). The seed sequence in animals is generally defined as the nucleotides from positions 2 to 8 of the miRNA, from 5' to 3', called the 7-mer seed, but exceptions exist such as 6-mer and 8-mer, presence of mismatches, mer flanked by specific nucleotides (Bartel 2009; Wang 2014; Menor et

al. 2014; Peterson et al. 2014). MiRNA target binding sites (MTBS) are generally located in the 3'UTR (un-translated region) of genes, but also, in a lower proportion, in their 5'UTR and open reading frame (ORF) (Lytle et al. 2007).

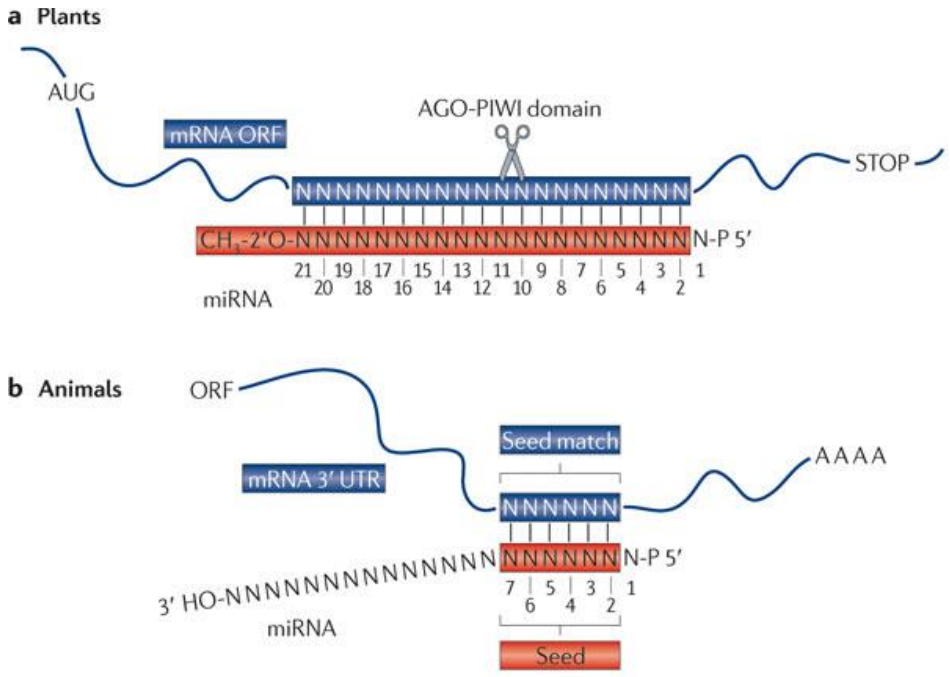


Figure I-10: Typical messenger RNA target recognition by miRNAs in plants and animals (Huntzinger and Izaurralde 2011)

The miRNA:mRNA attachment inhibits the protein synthesis by either repressing translation, or promoting mRNA deadenylation and decay (Krol et al. 2010). The translational repression of an mRNA can occur at all stages of the translation, including inhibition during initiation or elongation, co-translational protein degradation, and premature termination of translation (Wu and Belasco 2008). It has been shown that miRNA-induced changes at the translational level are by far the most common process (more than 84%) leading to the reduction of protein products (Guo et al. 2010).

Messenger RNA degradation involving a shortening of the mRNA poly(A) tail eventually leads to a deadenylation[7] of the molecule, followed by decapping[8] and exonucleolytic digestion[9] (Filipowicz et al. 2008). This mechanism generally occurs in processing bodies (P-bodies), where mRNAs that are targeted for deadenylation and degradation by the decapping pathway (e.g. microRNA induced mRNA silencing) are recruited (Kulkarni et al. 2010).

Finally, the miRNA-induced mRNA cleavage, generally occurring in plants but recently discovered to operate also in animals (Shin et al. 2010; Karginov et al. 2010), has a repressive effect on target mRNA expression, especially during rapid developmental transitions. This mechanism is executed through Argonaute (AGO) endonuclease activity, which acts as a slicer when specific domains in the mRNA are present and form a tertiary structure, allowing unwinding and facilitates the formation of a catalytically competent Argonaute (Park and Shin 2014).

1.1.5.4 Classification and annotations of microRNAs

Due to their growing importance, databases of miRNAs were created to offer an easy access to researchers. The main one is miRBase (Griffiths-Jones 2004), whose latest version, number 21 (since June 2014), contains more than 35,800 miRNAs distributed among 223 organisms. MiRBase is a repository of all published miRNAs/pre-miRNAs sequences and annotation. Each entry in the database is a mature miRNA transcript, grouped by species, from which various elements are provided, such as its pre-miRNA sequence, genomic location, some type of experimental validation, or target genes (Griffiths-Jones 2004). Other databases exist, such as miRMaid (Jacobsen et al. 2010),

[7] Removal of the poly(A) tail from the mRNA 3′ end
[8] Removal of the m7G (7-methyl-guanosine-triphosphate) cap, which is a structure at the 5’ end of mRNAs that promotes the translation and protects from degradation (Krol et al. 2010)
[9] Removal of single nucleotides from the end of a nucleic acid chain

a user-friendly clone of miRbase, or PMRD (Zhang et al. 2010) which contains only plant miRNAs, most being predicted by bioinformatics tools.

Each species contains a variable number of annotated miRNAs. Many mature miRNAs have a homologous sequence in several species. To keep track of this conservation, newly discovered miRNAs are indexed based on a specific nomenclature (Griffiths-Jones et al. 2006). The objective of miRNA classification essentially lies in grouping into families similar sequences conserved in various species and having potentially many of the same (orthologous) target genes. This facilitates attribution of functions associated to a miRNA family. The choice of a group number is performed by miRbase: from a sequence alignment on Rfam database (Griffiths-Jones et al. 2005), the consortium attributes a family to new miRNAs. Some tools are dedicated to this task, such as miRClassify (Zou et al. 2014) which identifies with a machine learning approach miRNAs from their mature sequence and classifies them to an existing family. The nomenclature is set as follows: the name always starts with a concise name of a species (ex: *hsa* for *Homo sapiens*, *tae* for *Triticum aestivum*), followed by "miR" if we are talking about the mature miRNA, or "mir" if it is the precursor. Then a number is added (ex: hsa-mir-31), which is the family number incremented by order of discovery. A letter {a-z} is added at the end of the name if we have many members of one family in a species (ex: hsa-mir-33a, hsa-mir-33b). For mature miRNAs, the arm where it belongs in the precursor hairpin is specified: 5' or 3' (ex: hsa-mir-33a-5p and hsa-mir-33a-3p). If the arm is not specified, it is 5' by default.

1.1.5.5 **Formation and evolution of** miRNA genes

Mechanisms leading to the formation of new miRNA genes are multiple (Berezikov 2011; Fahlgren et al. 2010). One of them is a duplication, which copies an existing gene by homologous or non-homologous recombination, transposable elements, gene duplication, or large-scale segmental duplication events. Copies eventually take other functions after mutation events, avoiding sur-expression of the duplicated gene (Clancy and Shaw 2008). New miRNA-like hairpins can also be formed after mutations within

a transcription unit, i.e. which contains a promoter (e.g. introns, pseudogenes). The chapter 4 investigate in detail the mechanisms of miRNAs' origins.

MiRNAs were formed gradually along evolution of species. Current mature miRNAs exist in closely related species, but rarely between distant species or kingdoms (e.g. plant vs animals). In plants, unlike in animals, the proportion of species-specific miRNAs is high, meaning that most of known miRNA genes arose relatively recently in evolutionary time. These young miRNAs are usually weakly expressed, processed imprecisely and a majority lack functional targets, suggesting they are under neutral evolution. It also happens that arising of new miRNAs by mutations can have deleterious effects and impacts fitness by perturbing existing regulatory networks (Cuperus et al. 2011). Such deleterious miRNAs are usually deleted by natural selection. Conversely, new miRNA genes that improve fitness are maintained in the genome. It has been shown that, by comparing precursors of conserved miRNAs among a set a species, the nucleotide divergence occurs essentially outside the miRNA-miRNA* region, i.e. within the loop and loop distal stem (Cuperus et al. 2011). Targeting ability of the miRNA is then preserved by selective pressure. All these characteristics give the opportunity to perform evolutionary studies using miRNAs.

1.1.5.6 MicroRNAs in physiology and pathology

1.1.5.6.1 Roles in human physiology

MiRNAs are involved in virtually all physiological processes in animals (Osman 2012; Lawrie 2013; Teruel-Montoya et al. 2014). In the last few years, many studies analyzed the roles of miRNAs in human, either by knocking out miRNA genes or by experimentally screening target genes. For example, it has been discovered that miRNA activity influences many biological functions in animals, such as pluripotency maintenance (Suh et al. 2004), germinal maturation (Murchison et al. 2007), cellular differentiation in many tissues (Chen et al. 2006), immunity (Xiao and Rajewsky 2009), cholesterol homeostasis (Zampetaki and Mayr 2012), cardiogenesis, cardiac

conduction or cell cycle (Zhao et al. 2007). More recently, miRNAs were found to regulate foetal gene expression (Li et al. 2015b), and reproduction and sex determination (Li et al. 2015c).

1.1.5.6.2 Roles in Human diseases

Due to their important participation in the development and the physiology of the host, dysfunctional ncRNA often lead to diseases. MiRNAs play an important role in pathogenesis (Cooper et al. 2009) (Figure I-11), as the number of studies discovering their implication in cancers and many other diseases grows every year. A pathology involving miRNAs can be the result of a loss of function or the deregulation of their expression (Clop et al. 2006), caused by mutations in a target gene or in the miRNA itself. Furthermore, a loss of function in one of the proteins involved in the machinery of miRNA processing, for example Dicer, leads to disastrous effects, such as developmental defects (Bernstein et al. 2003; Wienholds et al. 2003; Taft et al. 2010). In humans, studies found aberrantly expressed miRNAs, globally down regulated, in a long list of cancers (more than 170 cancers referenced in miRCancer database in September 2015 (Xie et al. 2013)), but also involved in central nervous troubles (e.g. schizophrenia, Alzheimer, Parkinson, Huntington) (Meza-Sosa et al. 2012), cardiovascular diseases (Dangwal et al. 2012) and various other syndromes (Chang and Mendell 2007; Goodall et al. 2013). Considering the growing list of diseases associated with ncRNAs, it is not impossible that almost every common disease could present a direct or indirect link with particular ncRNAs dysfunction or deregulation.

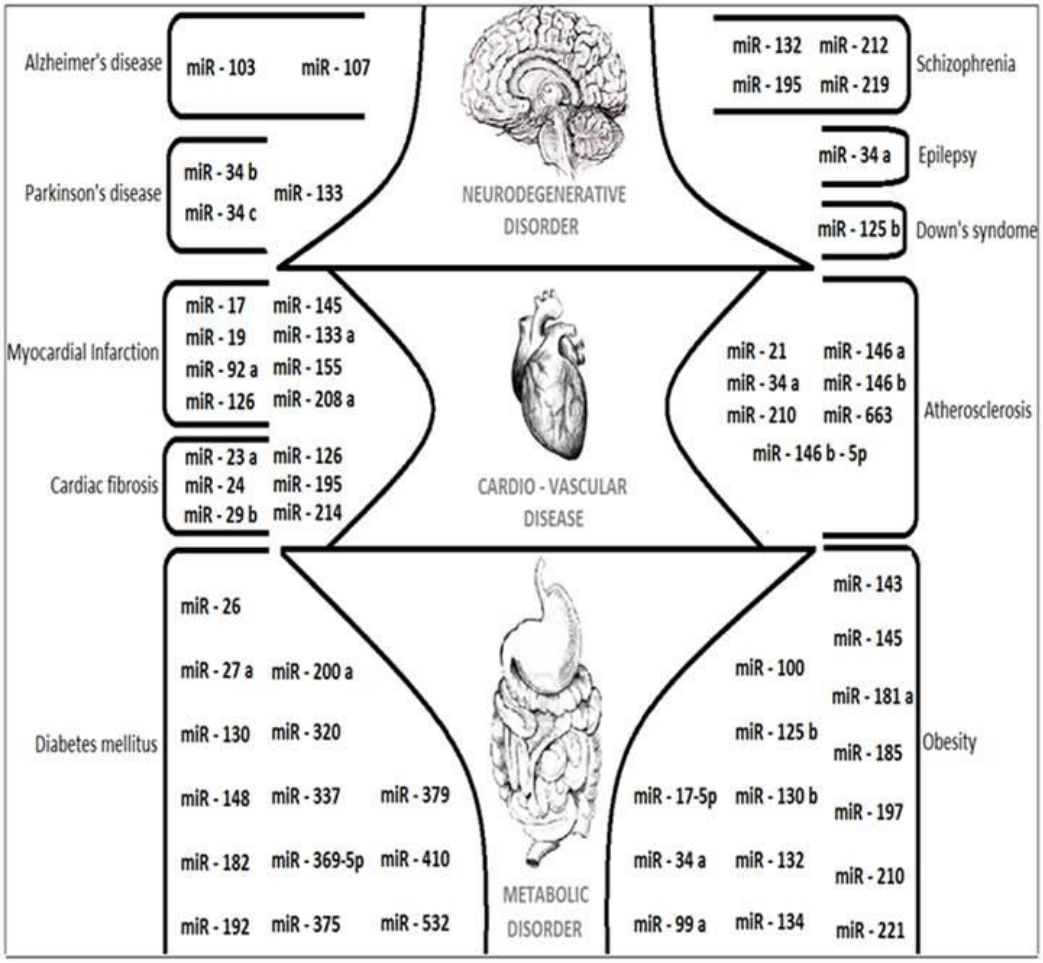


Figure I-11: Example of miRNAs implication in human diseases (Srinivasan et al. 2013)

1.1.5.6.3 Roles in Plants

As in animals, thousands of miRNAs have already been identified in plants, and virtually all physiological processes involve miRNAs. Many studies show that miRNAs have an impact in response to stress and development, where various species can tolerate or resist abiotic stresses by regulating specific genes targeted by miRNAs (Sunkar et al. 2006, 2012; Khraiwesh et al. 2012), such as extreme temperatures or presence of pollutants (Agharbaoui et al. 2015; Lv et al. 2010; Chen et al. 2012). MiRNAs are also required in the development of roots (Boualem et al. 2008), vessels (Kim et al. 2005), flowers (Chen 2004) and leaves (Palatnik et al. 2003). Researches

also demonstrated how production of miRNAs by cells is suppressed by bacteria (Navarro et al. 2008) and how, on the opposite, miRNAs contribute to anti-bacterial resistance (Navarro et al. 2006).

1.1.5.7 MicroRNAs in medicine and biotechnology

Research on miRNA helps improve medicine advances, and therapeutic applications are in development (Lawrie 2013; Hammond 2015). Many patents exist for treatments involving miRNAs (Figure I-12). Currently, miRNAs are used as diagnostic biomarkers or therapeutic targets. In cancer research, expression profiling of a particular circulating miRNA can accurately identify the localization of a tumor, based on miRNA deregulation studies reported for specific cancers. In recent years, miRNAs in blood serum are becoming a novel class of biomarkers for diagnosis of cancer (Chen et al. 2008) and cardiovascular diseases (Creemers et al. 2012). The profiling can also be done directly from saliva and tissues samples. Clinical trials are in progress for specific miRNAs biomarkers involved in lung, breast, colorectal, prostate and other cancers (Nana-Sinkam and Croce 2013; Andorfer et al. 2011; Lianidou et al. 2015; Redova et al. 2013; Barh et al. 2014), but also in atherosclerosis (link with miR-33) or fibrosis (link with miR-21) (Van Rooij et al. 2012). For therapeutic applications, it is possible to restore artificially the expression of a deregulated miRNA (Taft et al. 2010), and recent studies also revealed the potential of miRNA-based drugs for the treatment of cardiovascular diseases (cardiac regeneration (Wu et al. 2013a; Porrello 2013), cardiac calcium signaling restoration (Harada et al. 2014) and cardiac repair after myocardial infarction (Sahoo and Losordo 2014)).

In plants, due to the requirement of full complementarity (~21 nt) between the miRNA and its targets, experiments aiming to silence highly specific genes by artificial miRNAs (amiRNAs) have been proposed in rice (Warthmann et al. 2008). This opens a new field in genetically modified plants, in order to improve agronomic performance and nutritional value, and even virus resistance (Qu et al. 2007; Bahadur et al. 2015).

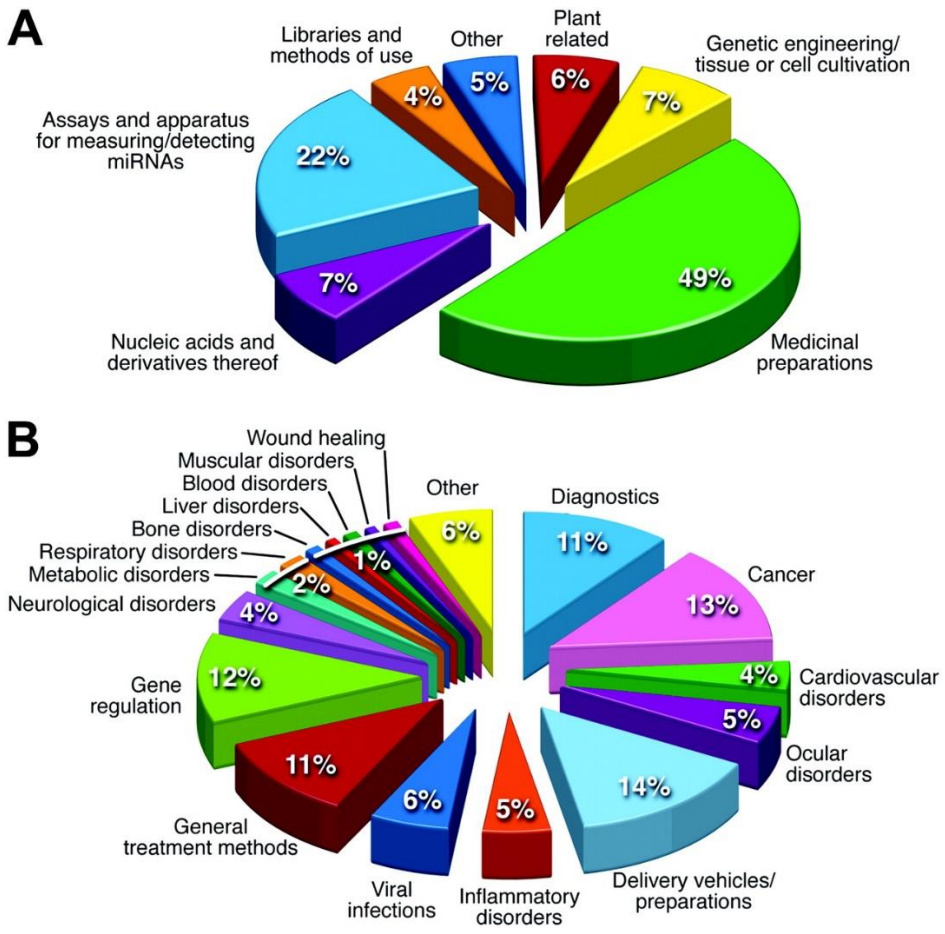


Figure I-12: microRNAs in biotechnology. A, Distribution of technological fields for US miRNA-related patents (determined by International Patent Classification codes). B, Distribution of medicine fields targeted by medicinal preparations in A. (Van Rooij et al. 2012)

1.2 RNA secondary structure prediction

Unlike coding regions of genomes, non-coding RNA lacks particular signatures in their nucleotidic sequence that would help their identification. However, in living organisms, RNAs fold into three-dimensional structures, defining their functionality. Thus, structural approaches have been developed to reveal functional RNAs (Washietl et al. 2005). In this section, I define what the secondary structure of RNA is, and how it can be predicted based on energy minimization principles or the analysis of evolutionary data.

1.2.1 Definition and representation of RNA secondary structure

Each type of non-coding RNA has a particular secondary (Figure I-13) and tertiary structure. Although tertiary structure prediction algorithms exist (Laing and Schlick 2010), they nearly all rely on calculating first the secondary structure. By definition, a secondary structure of a nucleic acid sequence is the set of Watson–Crick (A:U, C:G) or wobble base pairs (G:U) present in the structure. The secondary structure describes the set of base pairing interactions within a molecule, and can be represented as a graph containing all connections between paired bases on a polymer backbone (Figure I-14). Secondary structures can be represented in a parenthesis format, where a pair of parentheses corresponds to a pairing between two nucleotides, and a dot is an unmatched nucleotides.

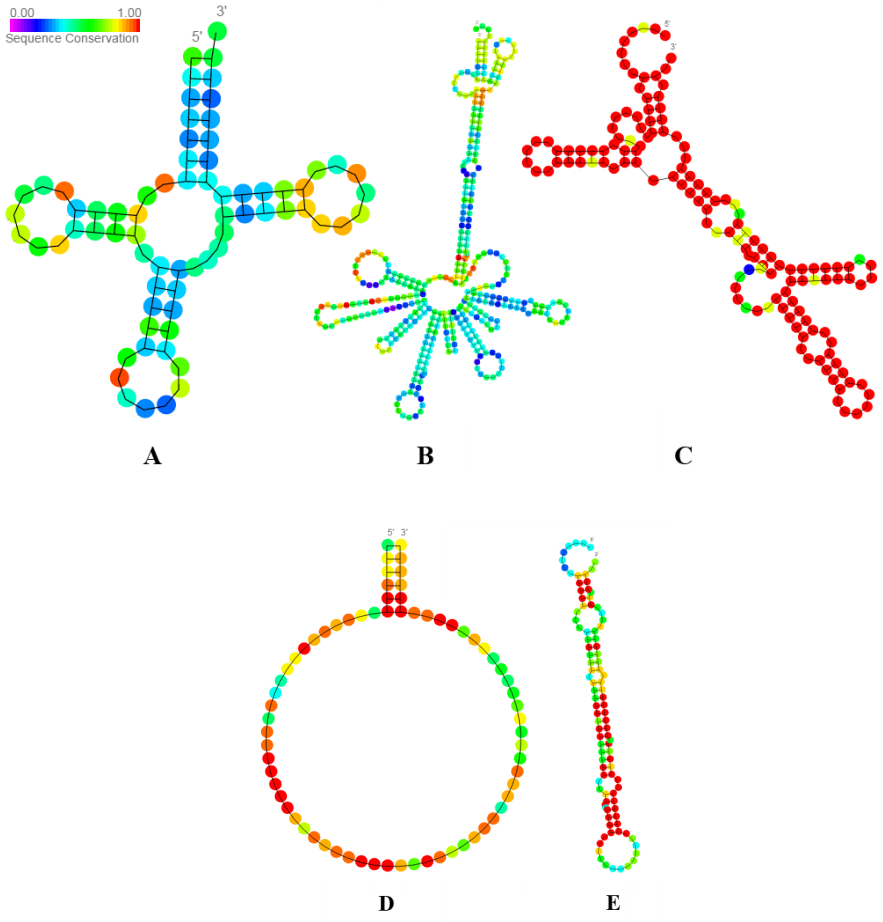


Figure I-13: Examples of RNA secondary structures (Source: Rfam). A: tRNA, B: tm-RNA, C: snRNA, D: snoRNA, E: miRNA. Color corresponds to the structure conservation.

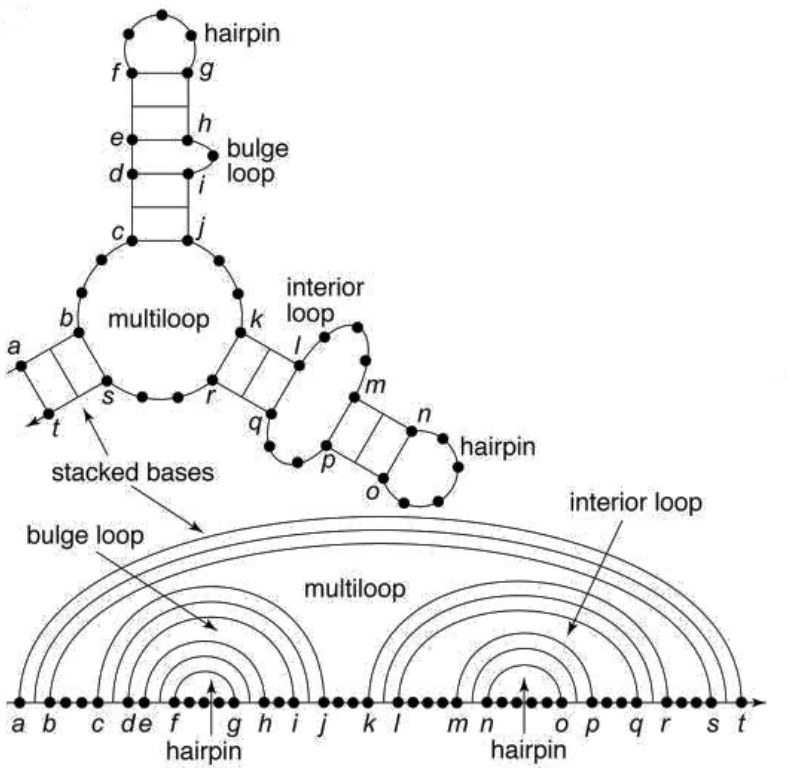


Figure I-14: Secondary structure of an RNA molecule represented as canonical loops (hairpin loops, stacked base pairs, bulge loop, interior loop, a multiloop) and on a backbone (Dirks et al. 2004).

1.2.2 Estimating the free energy of an RNA secondary structure

The principle of minimum energy is based on the second law of thermodynamics, where "the internal energy will decrease and approach a minimum value at equilibrium for a closed system with constant external parameters and entropy[10]" (Calvin 2013). "Free energy" refers to Gibbs energy, which describes the amount of energy obtainable by a thermodynamic system at a constant temperature and pressure at its initial state. The free energy is then minimized when the system, here the RNA, reaches its

[10] For a closed thermodynamic system, entropy is a quantitative measure of the amount of thermal energy not available to do work (Hellweg 2012).

equilibrium in the cell. In other words, the absolute value of MFE can be seen as a measure of energy required to unfold a folded RNA molecule at its equilibrium.

The free energy of a given secondary structure can by approximated by summing over each base pair's and loop's energy contributions, at fixed temperature and ionic concentration, where stacked pairs contribute to decreasing the free energy (Sankoff 1985). These energies are pre-computed in tables used to calculate the energy of a complete structure. The MFE is a negative value, or equals zero when an RNA structure contains no pair. Among a series of RNA sequences of same size, the one having the most negative value of free energy is considered as the most stable. Nevertheless, finding the optimal structure of a RNA molecule is difficult, and such a structure is not necessarily unique. Many optimal structures can be produced, their number increasing exponentially with sequence length (Durbin et al. 1998), and it is challenging to select the native one (Parisien and Major 2008). Furthermore, MFE structures sometimes do not match experimental data, because of a certain lack of biological realism in the energy calculation model (Zuker and Stiegler 1981). Finally, it is important to note that most algorithms based on thermodynamic rules do not take in account the presence of pseudo-knots[11], due the complexity to predict such structures: it is an NP-complete problem[12] (Lyngso 2004). Thus, most secondary structure prediction algorithms consider that RNA structures are pseudo-knots-free.

[11] RNA structure that is minimally composed of two helical segments connected by single-stranded regions or loops (Staple and Butcher 2005).
[12] Non-deterministic polynomial time problem (i.e. cannot be solved in a polynomial time)

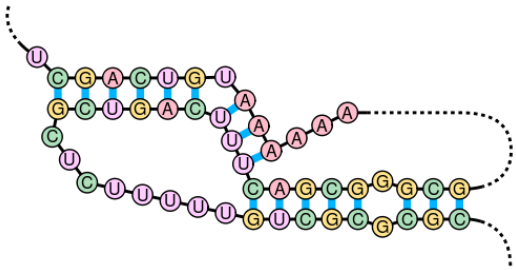


Figure I-15: Example of pseudo-knot structure (Chen et al. 2005)

1.2.3 Algorithms for pairing maximization and free energy minimization

Researchers use various approaches to calculate secondary structures from primary sequences. The Energy Minimization Secondary Structure Prediction problem (Zuker and Stiegler 1981) is often defined as follows: Given a RNA sequence, find the secondary structure that minimizes the free energy.

There exist many methods in the literature to predict secondary structures of RNAs. Among the best known, we briefly present in this section the Nussinov approach (Nussinov et al. 1978), the oldest approach that solved the base pair maximization problem, and MFold, created by Zuker and Stiegler (Zuker and Stiegler 1981) which improved the later by minimizing a more refined approximation of the free energy of the calculated structure. We also present a technique to compute base pairs binding probabilities, proposed by McCaskill (McCaskill 1990). These three methods, with various other optimizations, are used in RNAfold, included in Vienna package (Hofacker et al. 1994), the most widely used RNA folding tool in the literature.

1.2.3.1 Nussinov algorithm

The Nussinov approach (Nussinov et al. 1978) maximizes the number of base pairing of a given RNA sequence S of length n with a dynamic programming algorithm that runs in time $O(n^3)$. Dynamic programing solves complex problems by breaking them into simpler subproblems; here it finds optimal structures for substrings of the input sequence. Only canonical or wobble base pairs are considered. The algorithm proceeds

as follows (Durbin et al. 1998): Given S, a matrix W is calculated, where $W(i,j)$ is the maximal number of base pairs among all possible folds of $S[i \ldots j]$ and $W(1,n)$ the number of base pairs in the maximally base-paired structure. Let $\delta(i,j) = 1$ if i,j is a complementary or wobble base pair, 0 if it is not. The algorithm works in two steps: the matrix fill stage and backtracking stage. In the first stage, the matrix is initialized with zeros on the main diagonal and bottom one as follows:

$$W(i,i) = 0 \text{ for } i = 1 \text{ to } n$$
$$W(i,i-1) = 0 \text{ for } i = 2 \text{ to } n$$

Then, the triangle upper part of the matrix is filled diagonal by diagonal from 2 to n:

$$W(i,j) = \max_{2 \le i,j \le n} \begin{cases} W(i+1,j) \\ W(i,j-1) \\ W(i+1,j-1) + \delta(i,j) \\ \max_{i<k<j} [W(i,k) + W(k+1,j)] \end{cases}$$

In the second stage, to get the optimal structure, a backtracking is done through W, beginning from $W_{i,n}$.

This technique has many limitations, such as the lack of consideration of base pairs stacking, and the absence of special scoring of multiloops, and in inability to produce suboptimal structures. The algorithm delivers only one structure, based on base pair maximization, which is not always biologically relevant. There are, however, ways to overcome the problem of extracting sub-optimal solutions from the calculated matrices (Wuchty et al. 1999).

1.2.3.2 Zuker algorithm

The Zuker algorithm (Zuker and Stiegler 1981) is a method for folding a given RNA sequence. It uses previous work of Salser (Salser 1978) and Nussinov et al. (Nussinov et al. 1978), and includes new features compared to previous folding algorithms in the literature. It is based on dynamic programming and computes the optimal minimum free energy secondary structure of a sequence *S* of length *n* in time O(n^3). The algorithm

uses a defined group of specific substructures (loop, bugles, stacks, etc.) assigned with free energy values depending on nucleotides composition, linked to the reactivity of nucleotides to chemical modification or enzymatic influence on the RNA. The total energy of *S* is the sum of the energy of its substructures (example in Figure I-16).

The recursive algorithm runs as follows: the nucleotides of the RNA molecule are numbered from 5' to 3', denoting by S_i the i^{th}nucleotide for $1 \leq i \leq n$. The technique is to compute two possible energies for each subsequence S_{ij}. For all pairs (i, j) in $1 \leq i < j \leq n$, let matrices $W(i,j)$ and $V(i,j)$ be the MFE of all possible structures formed from S_{ij}, except $V(i,j)$ is set only in case of a base pairing of i and j. When (S_i, S_j) cannot form base pair, $V(i,j) = \infty$. If distance d between i and j is lower or equal to 4, $W(i,j) = 0$, otherwise $V(i,j)$ and $W(i,j)$ are computed in terms of $V(i',j')$ and $W(i',j')$, with $i < i' < j' < j$. $V(i,j)$ is calculated as follows:

$$V(i,j) = \min \begin{cases} E(FH(i,j)) \\ \min\limits_{i<i'<j'<j} \{E\big(FL(i,j,i',j')\big) + V(i',j')\} \\ \min\limits_{i+1<i'<j-2} \{W(i+1,i') + W(i'+1,j-1)\} \end{cases}$$

with $E(FH(i,j))$ the energy of the hairpin loop of subsequence S_{ij} and $E(FL(i,j,i',j'))$ the energy of either stacking region, bugle or internal loop. The entries of the W table are calculated as follows:

$$W(i,j) = \min \begin{cases} W(i+1,j) \\ W(i,j-1) \\ V(i,j) \\ \min\limits_{i<i'<j-1} \{W(i,i') + W(i'+1,j)\} \end{cases}$$

Compared to Nussinov algorithm, Zuker is more accurate, essentially because it integrates the concept of minimum free energy, which involves far more biological knowledge than Nussinov.

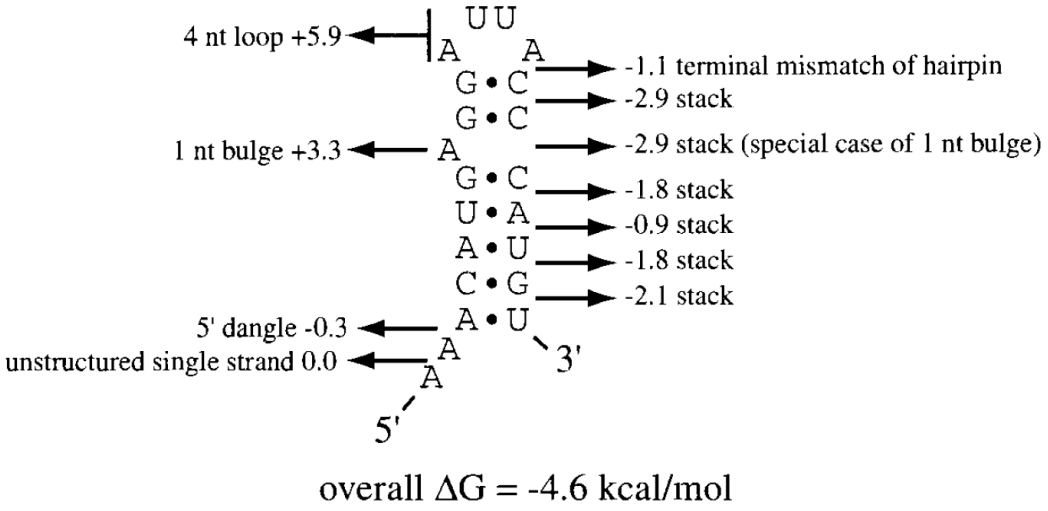


Figure I-16: Example of calculation of free energy (ΔG) for an RNA stem loop (Durbin et al. 1998)

1.2.3.3 McCaskill

The McCaskill dynamic programming algorithm (McCaskill 1990) is based on the calculation of the partition function Z over the canonical ensemble of all possible secondary structures of a given sequence. The canonical ensemble refers to the probability distribution P_i of secondary structure states of an RNA molecule, characterized by the probability of finding the molecule in a particular structure state i with a particular free energy E_i at a temperature T, given by the Boltzmann distribution: $P_i = \frac{1}{Z}\, e^{-E_i/kT} = e^{-\left(E_i - A/kT\right)}$ where $Z = e^{-A/kT}$ is a normalizing constant, A is the Helmholtz free energy function and k is the Boltzmann constant. The resulting partition function is calculated by $Z = \sum_{S \in Q} e^{-E(S)/RT}$, where Q is the set of all possible structures, S is a particular structure, $E(S)$ the energy of the structure, R is the gaz constant in joules/Kelvin and T the temperature in Kelvin. The McCaskill algorithm works by calculating base pair probabilities of a given RNA sequence in the thermodynamic ensemble, using four kinds of partition functions (PF) that describe various possible substructures.

1.2.3.4 RNAfold

RNAfold (Denman 1993), from Vienna package (Lorenz et al. 2011), is a fast and easy to use program developed to calculate minimum free energy (MFE) secondary structures and partition function of RNAs by using various state-of-art algorithms together. The MFE is computed using the Zuker and Stiegler algorithm (Zuker and Stiegler 1981) and the partition function algorithm based on McCaskill algorithm (McCaskill 1990). Energy parameters come from the work of Mathews and Turner (Mathews et al. 2004; Turner and Mathews 2009).

1.2.4 Nucleotide cyclic motifs, covariation analysis and stochastic context-free grammars

Other ideas have been proposed to infer RNA secondary structures. We briefly present here MC-FOLD (Parisien and Major 2008), which uses nucleotide cyclic motifs associated to probabilistic models, and covariation analysis combined with stochastic context-free grammars (SCFGs).

1.2.4.1 MC-fold algorithm

Parisien and Major have created MC-Fold to infer RNA secondary and 3D structures from nucleotide sequences (Parisien and Major 2008). In previous algorithms, models use canonical base pairs: Watson Crick (A-U, G-C) and Wobble (G-U). In MC-Fold, all non-canonical base pairs (e.g. A-A) are accepted, since their contribution to the energy of the structure is actually non-negligible. To find the optimal structure, they use nucleotide cyclic motifs (NCM) and a scoring function. These motifs, stored in a database, include all nucleotides interactions that form loops, bulges and base pairs of fixed lengths. The algorithm computes secondary structures by trying all possible NCM constructions. From a given sequence, all possible hairpins are generated by determining a list of initiation sites assigned to lone-pair NCMs. Then the rest of the sequence is matched recursively to double stranded NCMs. The algorithm also

calculates multibranch structures. All sub-optimal solutions are ranked by a score corresponding to their probability of occurrence.

1.2.4.2 Covariation analysis

Covariation analysis method has been implemented in COVE (Eddy and Durbin 1994). It is based on the principle that base pairing interactions are often better conserved than the genetic sequence itself, due to the presence of correlated compensatory mutations (Durbin et al. 1998). Comparative analysis using conservation to infer a structure is an efficient technique, but difficult to set up. It requires knowing a structurally correct multiple alignment, and this alignment requires knowing the correct structure. The structure is obtained by an iterative refinement process of guessing the structure based on the best multiple alignment, then realigning based on the obtain structure. Also, compared sequences must be relatively similar to obtain a good alignment, but sufficiently dissimilar to detect covarying substitutions. Pairwise sequence covariation can be measured as follows:

$$M_{ij} = \sum_{x_i, x_j} f_{xixj} \log_2 \frac{f_{xixj}}{f_{xi} f_{xj}}$$

where M_{ij} is the mutual information between columns i and j of an alignment. Frequency of bases (A, C, G, U) is represented by f_{xi} and f_{xj} in columns i and j respectively, and f_{xixj} is the frequency of one of the 16 possible base pairs. This measure reveals dependent columns which vary together, i.e. the covariation to maintain base pairing complementarity.

1.2.4.3 Stochastic context-free grammars

Stochastic context-free grammars (SCFGs) consist of a number of symbols and productions rules with probabilities (Durbin et al. 1998). They can be used to predict RNA secondary structures without pseudo-knots. In SCFGs, the symbols are either terminal or non-terminal, where non-terminal symbols are transformed into terminal or

non-terminal ones by production rules. The finality is a string of terminal symbols attached to a probability obtained by the product of the probabilities rules used in the derivation (parsing). To know if a string belongs to particular grammar (set of symbols and production rules), algorithms such as CYK, which employs bottom-up parsing and dynamic programming, are used.

1.2.5 Limitations of secondary structure prediction

Ideally, it would be better to directly predict the tertiary structure, but they are very complex and existing algorithm take a long time to compute them. Thus, computational methods to predict RNA secondary structures have been developed, whose most efficient are based on dynamic programming algorithms for MFE folding, which guaranty to find the energetically best structure. Moreover, McCaskill, who shows how efficiently calculate the partition function over all secondary structures of an RNA molecule, made it possible to get structure and base pair probabilities. These algorithms avoid costly experimental structure determination but have limits: they cannot handle pseudo-knots and interaction with other molecules. Moreover, they do not take in account kinetic properties of RNA molecules, i.e. how easily is a state accessible from other states (Zhao et al. 2010b). Also, according to a study from Doshi et al. (Doshi et al. 2004), the prediction accuracy of secondary structure tested on many rRNAs varies from 20 to 70%. In 2013, Hajdin et al. (Hajdin et al. 2013) reported that prediction of secondary structure of various RNAs whose sequence was experimentally validated. Conventional algorithms based on sequence alone reached a sensitivity (in terms of fraction of base pairs in the accepted structure predicted correctly) of 72%. It is low, but fortunately partially correct predictions still allowed conducting numerous studies implying RNA structures. Research continues to improve the field of structure prediction, new approaches being published on a regular basis.

1.3 MicroRNAs identification and prediction

MicroRNAs form one of the most investigated classes of ncRNAs. Thus, research has been done on the identification of miRNAs in genomes in order to understand their gene regulation networks. But, because detecting miRNAs by experimental techniques is expensive, computational methods have been developed. In this section, we will present briefly some experimental techniques and bioinformatics methods existing to discover new miRNAs.

1.3.1 Experimental identification of miRNAs

In the early 2000s, the first experimentally identified miRNAs were discovered by sequencing small RNAs from mammals, flies, and worms (Lagos-Quintana et al. 2001; Lau et al. 2001; Lee and Ambros 2001) by Sanger sequencing of cloned small RNA-derived cDNAs (Lagos-Quintana et al. 2002, 2003; Berezikov et al. 2006b). Today, experimental identification or validation of previously annotated miRNAs is generally performed from high throughput sequencing of the miRNAome, i.e. all miRNAs coded by the genome, by massively parallel signature sequencing (MPSS (Brenner et al. 2000)) and miRNA serial analysis of gene expression (miRAGE, derived from SAGE protocol (Velculescu et al. 1995)). MPSS and SAGE techniques measure the expression level of mRNAs through the generation of a 17-20 bp and 9-10 bp tags respectively. While MPPS use ligation-based approach by attaching sequences to millions of microbeads, SAGE uses short tags of cDNA, made from all mRNAs in a cell, linked so they can be cloned in large groups. Once genes are identified by the sequencing step, they are counted to determine their relative expression. The massive amounts of data generated by such methods are then processed in bioinformatics pipelines for identification of miRNAs (Vigneault et al. 2012; Wang et al. 2009; Agharbaoui et al. 2015). Nevertheless, one of the limitations of these experimental approaches is the unequal expression of miRNAs depending cell type, tissue and experimental conditions. Unknown lowly expressed miRNAs have a high probability of not being detected by sequencing methods. Also, it is easier to validate specific

miRNA computational predictions (Chiang et al. 2010). Experimental validation of specific miRNAs can be performed by rapid amplification of cDNA ends (RACE) (Xie et al. 2005), hybridization to RNA blots (Berezikov et al. 2005), microarrays (Bentwich et al. 2005) and RNA-primed array-based Klenow extension (RAKE) (Berezikov et al. 2006b) (Figure I-17).

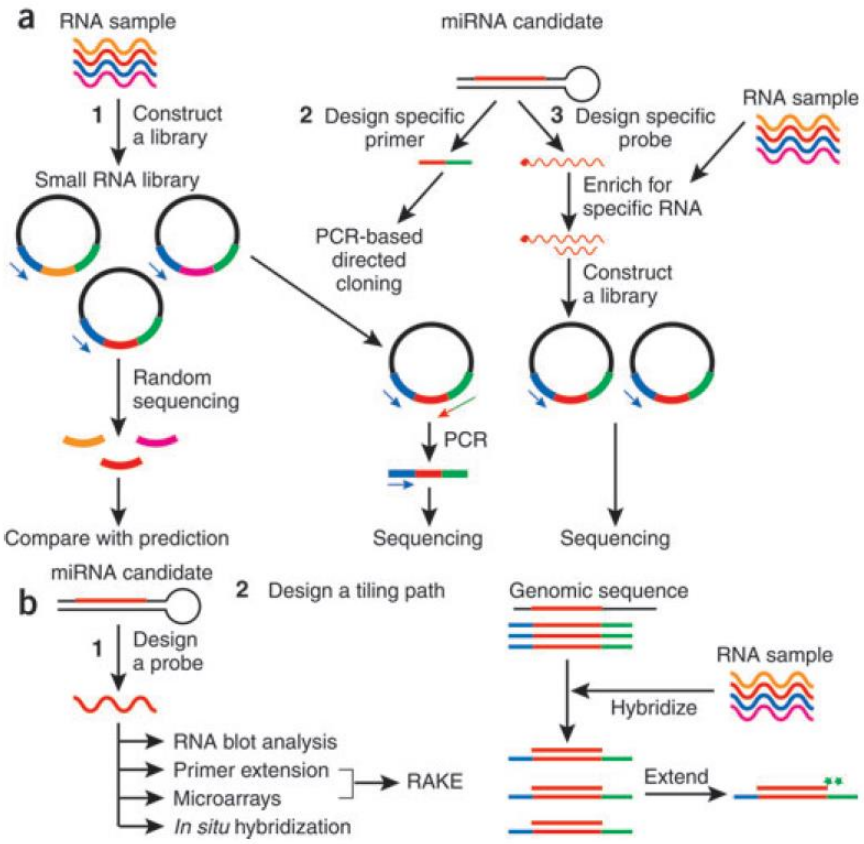


Figure I-17: Approaches to experimental validation of miRNA candidates (Berezikov et al. 2006a). (a) Cloning-based methods, where miRNAs are validated (1) by random sequencing of a collection of small RNA samples (library), or (2) by designing specific primers and adapters designed to amplify a predicted miRNA, or (3) by enriching RNA samples with biotinylated probes. (b) Hybridization-based methods, where validation is done (1) by using specific probes before RNA blot or microarray analysis, primer extension, or in situ hybridization. (2) By designing a tiling array (overlapping predicted 3' ends of mature miRNAs) used in an RNA-primed, array-based Klenow enzyme (RAKE) assay.

1.3.2 Prediction of miRNAs by bioinformatics methods

Computational identification of novel microRNA (miRNA) genes is a challenging task. Today, huge amounts of non-coding RNA transcripts have to be analyzed through prediction pipelines. The secondary structure of RNAs is the main type of evidence used by miRNA prediction algorithms. Various characteristics of the secondary structure, calculated by one of existing technique described in section 1.2 are used to identify miRNAs sequences. The hairpin shape, the number, length and position of bulges, the number of matches and mismatches between hairpin arms are examples of structural properties used by predictor programs. These parameters are set based on experimentally validated miRNAs stored in databases such as miRbase. Nevertheless, one of the limitations of relying only on structural properties is the resulting high rate of false positive (Leclercq et al. 2013; Friedländer et al. 2008). Thus, to improve the accuracy of miRNA prediction, researchers started in recent years combining structural features with sequencing information and conservation properties among species. Indeed, sequencing helps to detect functional miRNAs by revealing those presenting high levels of expression or differential expression between different conditions. Also, functionality can be identified by measuring conservation between species, revealing miRNAs that are maintained by selective pressure in evolution. Finally, another limitation of predictor tools is that some of them are species or clade specific, and all lack of update capabilities: they are trained only once at time of publication (Leclercq et al. 2013).

We present in the next section the various techniques to predict miRNAs, including machine learning and analysis of deep sequencing data. Some are specialized in the prediction of pre-miRNAs and others in the localization of the mature miRNAs within their precursor. We present both these categories in 1.3.2.1 and 1.3.2.2 These prediction techniques are generally using prior knowledge of experimentally validated miRNAs from miRbase (Griffiths-Jones et al. 2006).

1.3.2.1 Prediction of microRNA precursors

Many programs have been developed since the discovery of miRNAs. Among the techniques to predict miRNA precursors (pre-miRNAs), most use machine learning, combined or not with cross-species comparison and RNA sequencing data analysis.

Regardless of their approach, most of the prediction tools have focused on predicting human pre-miRNAs and/or mature miRNAs, although some other have been tailored to specific clades, such as plants (Jones-Rhoades and Bartel 2004), insects (Lai et al. 2003) or viruses (Li et al. 2008). While methods using filters can be used to predict pre-miRNAs precursors [Vmir (Grundhoff et al. 2006), DIANA-mirExTra (Alexiou et al. 2010), miRcheck (Jones-Rhoades 2010)], most of the tools use machine learning classifiers to complete this task. The following subsections present the programs developed to predict pre-miRNA by approaches involving machine learning, sequence conservation and deep sequencing data analysis.

1.3.2.1.1 Machine learning approach

In genome, many regions correspond to sequences that, if transcribed to RNA, could fold to form pre-miRNA-like hairpin structures. Another challenge is to detect special cases, such as multiloop hairpins or pre-miRNAs with large bulges.

For every program dedicated to the prediction of pre-miRNAs, pre-computed secondary structures of RNA molecules are submitted to an algorithm that extracts various features and tries to determine if they contain pre-miRNAs hairpins. Machine learning approaches to predict the miRNA precursors include support vector machine, random forest, hidden Markov models, covariance/SCFG model and Naive Bayes. A non-exhaustive list of existing tools utilizing these types of classifiers is presented in Table I-2. They usually use miRbase as positive training dataset, which contains experimentally validated miRNAs and pre-miRNAs. Negative datasets may be generated from random sequences that fold into hairpin, or are hairpins encoded in exons.

Program using machine learning extracts various features before using classifiers. Among the most used features, there are the minimum free energy, number of triplets (defined later in this section) normalized by hairpin length, sequence and structure characteristics (nucleotides location, length, number of bulges, loop size etc.) and conservation among species.

Table I-2: Non-exhaustive list of pre-miRNAs prediction programs using machine learning classifiers

Type of classifier	Programs
Support vector machine (SVM)	iMiRNA-PseDPC (Liu et al. 2015b) MicroPred (Batuwita and Palade 2009) miRanalyzer (Hackenberg et al. 2009, 2011) miRBoost (Tran et al. 2015) MiRenSVM (Ding et al. 2010) miRfinder (Huang et al. 2007) miRNA-dis (Liu et al. 2015a) miRPara (Wu et al. 2011) PMirP (Zhao et al. 2010a) RNAmicro (Hertel and Stadler 2006) RNAz (Washietl et al. 2005) Triplet-SVM (Xue et al. 2005) YamiPred (Kleftogiannis et al. 2015)
Random Forest	MiPred (Jiang et al. 2007)
Hidden Markov Model	HMMMiR (Kadri et al. 2009) Novomir (Teune and Steger 2010) ProMiR (Nam et al. 2005) SSCprofiler (Oulas et al. 2009)
covariance/SCFG model	Infernal (Nawrocki et al. 2009)
Naïve Bayes classifier	BayesMiRNAfind (Yousef et al. 2006) MatureBayes (Gkirtzou et al. 2010)

Once the program's classifier is trained, its efficiency is measured by four values: Sensitivity (Sn) and specificity (Sp), which measure respectively the proportion of positives and negative instances respectively that are correctly classified, total prediction accuracy (Acc) and the Matthew's correlation coefficient (MCC), calculated with the formulas:

$$\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \qquad \mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN+FP}} \qquad \mathrm{Acc} = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}}$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP}) \times (\mathrm{TN} + \mathrm{FN}) \times (\mathrm{TP} + \mathrm{FN}) \times (\mathrm{TN} + \mathrm{FP})}}$$

where TP=true positives, FP=false positives, TN=true negatives and FN=false negatives. The correlation coefficient MCC returns a value between -1 and +1. A coefficient of -1 indicates a total disagreement between predictions and observations, 0 means random predictions, and +1 a perfect prediction.

MiPred is an example of a random forest based pre-miRNA prediction tool (Jiang et al. 2007). It takes as input an RNA sequence and uses a combination of local contiguous structure-sequence composition called triplets, secondary structure MFE and P-value of the randomization test in order to distinguish pseudo and real pre-miRNAs. All these features are used as attributes in a random forest machine learning algorithm.

A triplet informs on the status of a nucleotide in the structure: paired, represented by a parenthesis in the secondary structure, or unpaired, represented by a dot. Adjacent nucleotides structures are included, giving a total of 8 possible compositions: '(((', '((.', '(..', '…', '.((', '..(', '.(.' and '(.('. Then, the middle nucleotide is added, giving a total of 32 combinations (e.g. 'A(((', 'C.(.'). Finally, the local contiguous triplet structure-sequence composition is calculated by the percentage of appearance.

The P-value of the randomization test determines if the MFE of a given sequence is significantly different from a randomly generated RNA sequence. A Monte Carlo randomization test is set to obtain this P-value:

1) Calculate the MFE *M* of the original given sequence
2) Shuffle the sequence while keeping dinucleotide distribution constant and re-compute the MFE. Repeat this step *N* times (*N* should be equal to 1000)
3) P-value $= \frac{R}{N+1}$ where R is the number of randomized sequences having a MFE $\leq M$

HHMMiR is a good example of how Markov models are used in the miRNA prediction. The Hierarchical Hidden Markov Model is composed of six states and three levels, described in Figure I-18. States starts with *hairpin* nodes, followed by the regions composing the hairpin loop: *Loop*, the hairpin head, *Extension*, the duplex strand between the loop and the mature sequence, *miRNA*, the mature sequence, *pri-Ext*, the rest of the sequence. A last state, *End*, is added as a transition to come back to *hairpin* state. These states follow each other in a one-way directed graph. Every state, except hairpin and end, has a lower level probability model. A *Loop* can only have indel lower-states (I) and *Extension*, *miRNA* and *pri-Ext* can have a match (M), a mismatch (N) or indel at any place, except *Extension* that starts with a match. All lower-states ends with an *End* state.

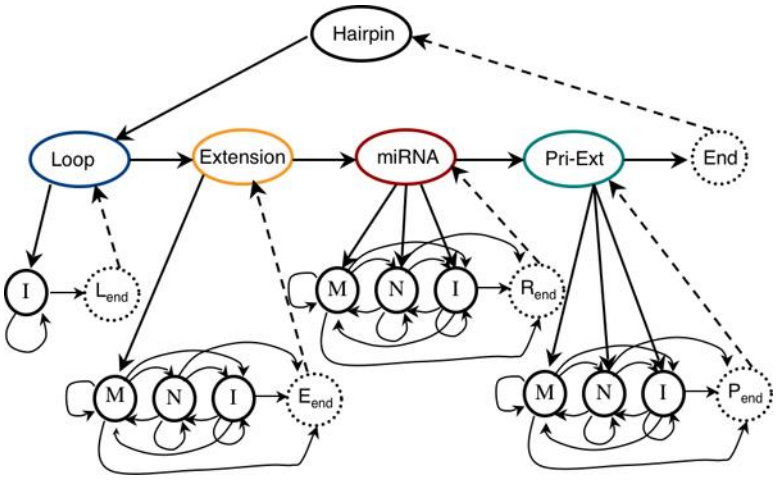


Figure I-18: The HHMM state model of HHMMiR (Kadri et al. 2009), based on the microRNA hairpin template of Figure I-8. M: match, N: mismatch, I: indel. Dashed circles are ending states, and plain circles are production states.

1.3.2.1.1 Analysis of deep sequencing data

Non-coding RNAs, and especially miRNAs, tend to have a highly variable expression due to various factors such as cell type origin, developmental phase and environmental influences. The expression level of a miRNA varies from a few to tens of thousands copies per cell (Hackenberg et al. 2011), and Dicer products amounts, such as miRNA-5p and miRNA-3p sequences, are tissue or developmental stage specific (Krol et al.

2010). To identify known or novel miRNAs and measure their expression, deep sequencing data analysis experiments are performed. Such experiment that focus on the sequencing of miRNAs is called a miRNA-seq, a type of RNA sequencing (RNA-seq), where only RNA within miRNAs size range, i.e. 16 to 30 nt (Leclercq et al. 2013), is isolated. About 10 million reads (expressed sequences) are usually produced for each experiment, although read numbers greater than 5 million contribute very little to the detection of new microRNAs (Metpally et al. 2013). The deep sequencing experiment is mainly defined by the depth of sequencing, which refers to the number of times a nucleotide can be read during the sequencing process. The resulting coverage is many times larger than the original length of the sequence of reference. Then, one of the challenges here is to analyze gigabytes of data generated by each sequencing experiment, another is to discriminate miRNAs from other non-coding RNAs or degradation products. Both miRdeep (Friedlander et al. 2008, 2012) and miRanalyzer (Hackenberg et al. 2011, 2009), the most common used tools, and more recent programs, such as miRTRAP (Hendrix et al. 2010), mirTools (Zhu et al. 2010; Wu et al. 2013b), mirExplorer (Guan et al. 2011), miRdentify (Hansen et al. 2014) and Mirinho (Higashi et al. 2015), take as input deep sequencing data and output predictions of pre-miRNAs and mature miRNAs.

After sequencing, low quality reads are discarded, and mapped (i.e. aligned) to a genome of reference. The mapping procedure of high throughput sequencing data reveal stacks of reads that cover the sequences of reference. Expression profiles and reads stacks allow the identification of potential miRNAs and other Dicer products. MiRdeep (Friedlander et al. 2008, 2012) has been developed to process this information. It uses a probabilistic model of RNA biogenesis to score the position and the frequency of a sequenced small RNAs compared to the secondary structure of the precursor. Among the degradation products produced by the experiment, we have the Dicer residues, such as hairpin loops and precursor extensions around the duplex miRNA and miRNA*. Except for the mature miRNA itself, everything is partially degraded. The position and frequencies of these elements are identified by miRdeep as

a signature to discover miRNAs. The pipeline works as follows: 1) Mapping reads against the reference genome. 2) Extracting and folding precursor around mapped reads. Non-hairpin sequences are discarded. 3) Precursors are then submitted to the miRdeep model based on the positions and frequencies of Dicer residues mapped on the precursor. 4) Identification of conserved miRNAs in miRbase by Blast (Altschul et al. 1990) and false positive rate estimation by permuting structure and signature pairing.

Another approach uses Dicer products to predict miRNAs: miRanalyzer (Hackenberg et al. 2011, 2009). The difference with miRdeep is that it performs mapping against various known RNA stored in databases such as Rfam, a database of RNA families, and Repbase, a database of repetitive DNA elements. This step removes known ncRNA that are not miRNAs, hence reducing the number of sequenced reads and lowering the false positives miRNAs. Besides, instead of a probabilistic model, miRanalyzer implements a machine learning method based on random forest generated from a broad variety of features associated with nucleotide sequence, structure and energy.

1.3.2.1.2 Approach based on structure conservation

An efficient method to detect functional RNAs based on structure conservation is implemented in RNAz (Washietl et al. 2005; Lu et al. 2011). This program predicts structurally conserved and thermodynamically stable RNA secondary structures in multiple sequence alignments. It takes as input an alignment of sequences, usually orthologous genome fragments from several related species, and consists in measuring the RNA secondary structure conservation based on computed structure consensus and their thermodynamic stability by RNAALIFOLD from Vienna package (Hofacker 2004). Regions with high structure conservation will be annotated as potentially functional. RNAz then determines whether the secondary structure consensus corresponds to a known ncRNA family. This approach classifies a candidate genomic region as functional RNAs by calculating two values: (i) the structure conservation index (SCI) and (ii) the normalized z-score. The SCI is obtained by comparing the

consensus structure MFE E_{cons} with the average MFEs $\bar{E}$ of every independently calculated structure (by RNAfold) of the alignment, where SCI $= \frac{E_{cons}}{\bar{E}}$. E_{cons} is calculated by RNAALIFOLD, which uses a combination of phylogenetic information (based on sequence covariation) and thermodynamic methods to predict RNA secondary structure. If the sequences that compose the alignment fold into a conserved common structure, then $\bar{E} \approx E_{cons}$, or SCI$\approx$1, which indicates a perfectly conserved secondary structure. At the opposite, when RNAALIFOLD can't find a consensus structure the SCI is close to zero. Normalized z-score *z* is calculated to measure the significance of MFE predicted value *m*, assessed by comparison with a large set of randomly generated sequences of same length and single or dinucleotide composition. This method relies on the fact that structure MFE alone is not sufficient to detect functional RNAs. Moreover, studies show that functional RNAs are more stable than random sequences, hence the comparison against random sequences. Mean μ and standard deviation σ of those random sequences are calculated to get *z*, where $z = \frac{(m-\mu)}{\sigma}$. Finally, a support vector machine (SVM) classifier, trained on all classes of ncRNAs, is used to classify the aligned sequences in the SCI/z-score plane. An advantage to this approach is the fact that the SVM is not trained on structure and composition characteristics, so it does not contain specific information about particular ncRNAs. Machine learning is used here as a help to interpret SCI and z-score. The time complexity is O($N \times n^3$), where N is the amount of input sequences and n the alignment length. Finally, the accuracy of this approach depends greatly on the type of ncRNA, where low accuracy is obtained for poorly conserved ncRNA family's members, such as tmRNAs, and high accuracy for other classes such as Hammerhead III ribozyme and tRNAs classes.

1.3.2.2 Prediction of microRNAs mature sequence

Search of mature miRNA is the identification of the portion of the pre-miRNA that is processed by Drosha and eventually target genes. Although there are many so-called "miRNA predictors", most are actually pre-miRNA predictors and do not specifically

identify the position of the miRNA within the pre-miRNA. However, a few tools have been developed specifically for this problem, and some of them are even capable of predicting both pre-miRNAs and mature miRNAs (Table I-3). It is important to note that deep sequence analysis pipelines dedicated to the identification of miRNAs predict both pre-miRNAs and mature miRNAs, revealed by read stacks analysis (see 411.3.2.1.1 and 2.3).

Table I-3: List of mature microRNA sequence predictors

Mature miRNA program	Also predict pre-miRNA?	Method	Constraints on mature miRNA length	Lineage
HHMMiR (Kadri et al. 2009)	Yes	Hierarchical hidden Markov models (Figure I-18)	No	Any (user training possible)
MatureBayes (Gkirtzou et al. 2010)	No	Sequence and secondary structure features classified with a Naive Bayes classifier	Yes, 22 nt	Human and mouse
MaturePred (Xuan et al. 2011)	No	miRNA:miRNA* duplex features classified with a SVM	Yes, user defined	Any
MiRalign (Wang et al. 2005)	Yes	Alignment with miRbase	Yes, 22 nt	Any
MIRcheck (Jones-Rhoades and Bartel 2004)	Yes	Set of rules and constraints	Yes, 20 nt	Plants
MiRmat (He et al. 2012)	No	Identification of Drosha and Dicer processing sites using random forest	No	Vertebrates
novoMIR (Teune and Steger 2010)	Yes	Set of filtering steps and statistical model	No	Plants
ProMir (Nam et al. 2005)	Yes	Combination of sequence and structural features in a paired hidden Markov model	No	Human

1.4 Target genes identification and prediction

One of the most challenging tasks in the study of miRNAs is the identification of the set of genes targeted by each miRNA. The target gene repertoire associated to a miRNA determines its function. The search for target genes has a dual purpose: validate a predicted miRNA and discover new potential targets.

To be functional, a miRNA has to target one or more messenger RNAs to stop their translation in protein, hence the importance of validating the presence of target genes after the identification of a putative miRNA. Identifying miRNA target genes is a challenging task because:

- Understanding of the biological processes associated to the binding of a miRNA to a target mRNA is limited (Wang 2014),
- The mRNA structure and the presence of RNA-binding proteins affect target site accessibility (Ameres et al. 2007; Kedde and Agami 2008),
- Target sites inhibiting translation can exist outside UTRs (Hausser et al. 2013a), which considerably increase the regions to analyze,
- Some miRNAs targets lack a complete 6-mer match to the seed portion of the miRNA (Shin et al. 2010; Didiano and Hobert 2006; Wang 2014; Lal et al. 2009; Fasanaro et al. 2012).

As a consequence, the false positive and false negative rate of miRNA target genes prediction programs remains high (Hamzeiy et al. 2014; Zheng et al. 2013). Whether target genes are identified experimentally or by bioinformatics, the two types of approaches are dependent on each other. Bioinformatics approaches rely on knowledge obtained from experimental studies, but are also essential to make sense of the data generated. In the rest of this section, I describe in more details the experimental and computational approaches for miRNA target gene prediction.

1.4.1 Experimental identification of miRNA target genes

The experimental identification of target genes is subdivided into three types of problems: (1) validation of specific miRNA targets, (2) identification of the targets of a given miRNA on a genome-wide scale, and (3) identification of the miRNAs that target a given mRNA.

(1) The objective here is to validate the hypothesis that a particular mRNA is targeted by a specific miRNA under specific experimental conditions. Several approaches are possible. First, candidate miRNA:mRNA interactions can be validated by luciferase reporter assay. This technique uses a vector that is composed of a luciferase gene, a target gene to test, and a poly-A tail. The vector is transfected in a model cell line and the luciferase expression is monitored by measuring its luminescence (Nicolas 2011). Second, taking in account that a miRNA and its targeted mRNA must be co-expressed in order that the repression of expression exists, then the co-expression is measured. This can be performed by Northern blot analysis, quantitative real-time PCR (qRT-PCR[13]) using total RNA isolated from a specific cell type, and probes or primers specific for a given miRNA and mRNA target. In situ hybridization can also demonstrate co-expression. Third, if a mRNA is a real target of a miRNA, then decreasing (by using antisense oligo-ribonucleotides) that miRNA's expression should change the amount of protein produced from the targeted mRNA, which can be measured by Western Blot analysis using specific antibodies against the protein (Kuhn et al. 2008).

(2) For genome-wide identification of the targets of a given miRNA, experimental identification of miRNA:mRNA interactions involves various target screening techniques (Thomson et al. 2011) associating large scale analyses and laboratory methods. It is possible to categorize these techniques in several categories:

[13] Real time polymerase chain reaction. Laboratory technique used to amplify and simultaneously detect levels of a DNA molecule.

Immunoprecipitation-based methods, labelling or tagging based methods, degradome analysis, and DNA synthesis by miRNAs.

1. **Immunoprecipitation** is a method that uses specific antibodies to target a molecule of interest. In this category several techniques to identify miRNA target genes exist:
   - Co-immunoprecipitation[14] of RISC components linked to miRNA:mRNA complex identified by microarrays or RNA-seq
   - High-throughput sequencing of RNA isolated by HITS-CLIP[15]
   - PAR-CLIP[16], which identifies the binding sites of cellular RNA-binding proteins and microRNA-containing ribonucleoprotein complexes (Hafner et al. 2010)
   - AGO-CLIP[17], where samples contain miRNA:target chimeras generated by an endogenous ligase (Grosswendt et al. 2014)
   - AGO2-PAR-CLIP[18], a similar technique than PAR-CLIP that uses a photoactivatable molecule to detect crosslinking interactions (Farazi et al. 2014)
2. **Isotope labelling and tagged sequences**, which consist in adding a marker on molecules of interest. With this approach, it is possible to identify target genes by:
   - Transfecting cells by biotinylated miRNA duplexes followed by microarrays analysis,
   - Parallel analysis of RNA ends (PARE), i.e. identification of mRNA cleavage products on a global scale by high-throughput sequencing of products from a modified 5' RLM-RACE[19],

---

[14] Technique using specific antibodies to isolate protein complexes.
[15] Crosslinking immunoprecipitation. This technique gives the opportunity to locate the targeted site precisely on the mRNA by the miRNA.
[16] Photoactivatable-ribonucleoside-enhanced crosslinking and immunoprecipitation.
[17] Argonaute protein crosslinking and immunoprecipitation.
[18] Argonaute-2 photoactivatable-ribonucleoside-enhanced crosslinking and immunoprecipitation0
[19] RNA ligase mediated-rapid amplification of cDNA ends.

- Measuring protein abundance by mass spectrometry of samples labelled by specific isotopes, i.e. SILAC[20] method,
- Separate miRNA-regulated proteins by electrophoresis on gel labelled by fluorescent colored substances, followed by SDS–PAGE[21] and mass spectrometry, e.g.2D-DIGE[22] method.

3. **Degradome analysis,** where it is possible to measure degraded mRNAs following ectopic[23] miRNA expression. Their expression is then analyzed on genome-wide scale by microarrays[24] or RNA-seq[25].
4. **DNA synthesis by miRNAs**, which performs a reverse transcription of target genes, where endogenous miRNAs serve as primers for cDNA[26] synthesis of targeted mRNAs.

All these methods identify novel targets with a relatively high accuracy, especially the most recent ones. But despite these performances and the decreasing cost of these experimental techniques, they are limited by the miRNAs expression levels, which are tissue and conditions dependent. Therefore, computational methods to predict miRNA target genes have been developed since several years.

1.4.2 Prediction of miRNAs target genes by bioinformatics methods

After the discovery of the effects of miRNAs on gene regulation, many research teams have focused on predicting miRNAs target genes. Many methods have been developed so far (Oulas et al. 2015; Ekimler and Sahin 2014; Hamzeiy et al. 2014) (non-exhaustive list in Table I-4), relying on various features describing miRNA:mRNA

---

[20] Stable isotope labelling with amino acids in cell culture.
[21] Polyacrylamide gel electrophoresis.
[22] Two-dimensional differentiation in-gel electrophoresis.
[23] Abnormal gene expression occurring because of a disease or artificial production. Also a technique to determine the function of a gene.
[24] Collection of microscopic DNA or RNA molecules attached to a solid surface.
[25] RNA sequencing, also called whole transcriptome shotgun sequencing. Sequencing technique that quantifies and sequence RNA from a genome at a given moment.
[26] Complementarity DNA. It is a DNA copy of a synthesized mRNA molecule.

interactions. Among the most common features used by target genes predictors, the seed match, part of miRNA:mRNA sequence complementarity, is one of the most used. In animals, a seed match contains Watson-Crick (A-U and G-C) or Wobble (G-U) pairs; one mismatch is possible. Another feature is the conservation, since the miRNA seed is in general more conserved that the non-seed region, as mRNA targeted sites (Lewis et al. 2003; Friedman et al. 2009) and miRNAs promoters (Fujiwara and Yada 2013). Free energy of the miRNA:mRNA duplex is also a widely used feature (Zheng et al. 2013), and finally target site accessibility (Long et al. 2007), i.e. nucleotides unpaired after folding, which requires to calculate the secondary structure of the putative targeted UTRs.

All programs dedicated to target genes prediction take as input miRNA sequences and references sequences, usually 3' and 5' UTRs (Witkos et al. 2011). Tools that rely on conservation, such as TargetScan, require alignments of UTRs (Friedman et al. 2009). All tools have their specific prediction technique to find target genes and do not offer same accessibility: online prediction, stand-alone software or precomputed predictions download (Peterson et al. 2014). Most of them use prior knowledge from databases of experimentally and/or validated targets (Oulas et al. 2015), including, for the best known, miRbase (Griffiths-Jones et al. 2008), TarBase (Sethupathy et al. 2006; Vlachos et al. 2015), miRDB (Wang 2008), microRNA.org (Betel et al. 2008), miRecords (Xiao et al. 2009), Mirwalk (Dweep et al. 2011) and Mirtarbase (Hsu et al. 2011).

Finally, recent comparison studies show recall rates between 6 to 20% depending on predictor programs (Witkos et al. 2011). In plants, where the target sites are longer than animals, i.e. full miRNA length compared to 6-8 nt resp. (see section 1.1.5.3), recall rates vary between 46 and 97% depending on species and programs (Srivastava et al. 2014).

Table I-4: Non-exhaustive list of microRNA target genes predictors

Target gene prediction program	Features used to predict target genes	Adaptable parameters	Organisms
DIANA-microT (Maragkakis et al. 2009; Reczko et al. 2012) Last update: 2012	Seed match, conservation, free energy, site accessibility, target-site abundance	None	Humans, mice, flies, and worms
miRanda (Enright et al. 2003) Last update: 2010	Seed match, conservation, and free energy	Free energy threshold, alignment threshold, weight of seed region, gap penalty	Animals
mirMark (Menor et al. 2014) Last update: 2014	700 features relating to site and seed match, free energy, conservation, target site accessibility and others	Miranda score	Humans
MirTarget2 (miRDB) (Wang 2008; Wang and El Naqa 2008) Last update: 2012	131 features, including seed match, conservation, free energy, site accessibility and others	Adjustable and default screening options are available for the target mining option	Humans, mice, rats, dogs, and chickens
PicTar (Krek et al. 2005) Last update: 2007	seed match, pairing stability	None	Mammals, fly, worm
PITA (Kertesz et al. 2007) Last update: 2008	Seed match, conservation, free energy, site accessibility and target-site abundance	Seed size, wobble or mismatch, conservation, and inclusion of a flank region	Humans, mice, flies, and worms
psRNATarget (Dai and Zhao 2011; Zhang 2005) Last update: 2011	Seed match, target site accessibility	Target accessibility, central mismatch range, false positive prediction rate	Plants
RNAhybrid (Krüger and Rehmsmeier 2006) Last update: 2006	Seed match, free energy, and target-site abundance	Advanced parameters relative to the miRNA:mRNA hybridization (i.e. energy	Animals

		threshold, internal loop length etc.), hits per targets, max p-value	
SVMicrO (Liu et al. 2010) Last update: 2010	Seed match, conservation, free energy, site accessibility and target-site abundance	None	Animals
TAPIR (Bonnet et al. 2010) Last update: 2010	Free energy, seed match	Score threshold, free energy ratio	Plants
TargetMiner (Bandyopadhyay and Mitra 2009) Last update: 2009	Seed match, conservation, free energy, site accessibility, target-site abundance and others	None	Animals
TargetScan (Lewis et al. 2005; Friedman et al. 2009; Grimson et al. 2007; Garcia et al. 2011) Last update: 2012	seed match, conservation	None	Mammals, flies, and worms

1.5 Thesis Outline, publications and contributions

The thesis is composed of five chapters. This first chapter was a review of the background knowledge related to our research. The next three chapters comprise the full text and figures of published, submitted, or ready to be submitted papers.

- **Chapter II**

  Leclercq M, Diallo AB, Blanchette M (2013) Computational prediction of the localization of microRNAs within their pre-miRNA. Nucleic Acids Research, 41:7200–11.

  The design and implementation of the computational tool in this publication was performed by me under Prof. Mathieu Blanchette's and Prof. Abdoulaye Baniré Diallo's supervision. The manuscript was written by me with input from my supervisors.

- **Chapter III**

  Leclercq M, Diallo AB, Blanchette M (2016) Evolutionary mechanisms leading to the creation of new miRNAs in primates revealed by the analysis of inferred ancestral sequences. To be submitted to Genome Biology and Evolution.

  The design and implementation of the computational tool in this publication was performed by me under Prof. Mathieu Blanchette's supervision. Prof. Abdoulaye Baniré Diallo gave us feedback and reviewed our final work. . The manuscript was written by me with input from my supervisors.

- **Chapter IV**

  Leclercq M, Diallo AB, Blanchette M (2016) Prediction of Human miRNA Target Genes using Computationally Reconstructed Ancestral Mammalian Sequence. Paper submitted in February 2016 at Nucleic Acids Research.

- *The design and implementation of the computational tool in this publication was performed by me under Prof. Mathieu Blanchette's supervision. Prof. Abdoulaye Baniré Diallo gave us feedback and reviewed our final work. The manuscript was written by me with input from my supervisors.*

- **Chapter V**

  This chapter summarize the main conclusions and highlights the contributions made by the work presented in this thesis, and opens future perspectives.

Finally, here is a list of papers I co-authored during my PhD but that aren't included in my thesis:

- Agharbaoui Z., **Leclercq M.**, Remita M.A.., Badawi M., Lord E., Houde M., Danyluk J., Diallo A.B. and Sarhan F. (2015) An integrative approach to identify hexaploid wheat miRNAome associated with development and tolerance to abiotic stress. BMC Genomics 16(1), 339
- Remita MA, Lord E, Agharbaoui Z, Leclercq M, Badawi M, Makarenkov V, Sarhan F, Diallo AB. 2015. WMP: A novel comprehensive wheat miRNA database, including related bioinformatics software. BioRxiv, 024893.
- Haudry A., Platts A.E., Vello E., Hoen D.R., **Leclercq M.**, Williamson R.J. et al. (2013) An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. Nature Genetics 45: 891–8.

CHAPTER II : COMPUTATIONAL PREDICTION OF THE LOCALIZATION OF MICRORNAS WITHIN THEIR PRECURSOR

2.1 Preface

This chapter present miRdup, a tool that has been created to fulfill three objectives: (1) improve the detection of real miRNAs from RNA sequencing by identifying likely false positive candidates, (2) predict the position of the mature miRNA within a precursor, and (3) create an auto-adaptive model based on miRNA's species-specific characteristics.

The first objective was defined at the end of my master thesis. My goal was to identify the miRNAome in the wheat cereal from deep sequencing data (the results were published in (Agharbaoui et al. 2015)). One of our main complications was the very high number of predicted miRNAs and lack of computational methods to validate a combination of a mature and a precursor sequence. That is the problem addressed by miRdup. Applying miRdup to candidate pre-miRNA/miRNA pairs obtained from sequencing experiments reduce the false positive rate of predicted miRNAs. Moreover, the miRNAs that have been experimentally validated in this study were flagged as true positives by miRdup. Later in my thesis, miRdup played a key role in allowing us to predict whether ancestral sequences were likely to be functional miRNA (see CHAPTER III).

The second objective (predicting the position of the mature miRNA within a precursor) will be useful to biologists to create new miRNAs for molecular engineering purposes as well as for fundamental research. Finally, the third objective (create an auto-adaptive model based on miRNA's species-specific characteristics) was a very new approach in the field of miRNA prediction. We have created a program that can automatically remain up-to-date with respect to newly published data, and it is species-customizable. When a user runs miRdup, it retrieves the latest version of miRbase to train on it, discarding non-experimentally validated miRNAs. The user can specify the clade on which the model should be trained, thus increasing its accuracy depending on the species he/she is working on. To support the importance of training species-specific models, we reported the main differences between five chosen clades.

MiRdup model is trained with a random forest classifier, which was first developed by Leo Breiman (Breiman 2001). It is an ensemble of decision tree predictors trained on a random subset of features sampled independently. This classifier has the ability to learn only important features and ignore irrelevant ones, reducing the need for feature selection. This is achieved by randomly selecting subsets of features during the tree construction.

The rest of this chapter is reprinted from:

> *Leclercq M, Diallo AB, Blanchette M (2013) Computational prediction of the localization of microRNAs within their pre-miRNA. Nucleic Acids Res, 41:7200–11.*
> *Copyright (2013) Oxford university press*

2.2 Abstract

MicroRNAs (miRNAs) are short RNA species derived from hairpin-forming miRNA precursors (pre-miRNA) and acting as key post-transcriptional regulators. Most computational tools labelled as miRNA predictors are in fact pre-miRNA predictors and provide no information about the putative miRNA location within the pre-miRNA. Sequence and structural features that determine the location of the miRNA, and the extent to which these properties vary from species to species, are poorly understood. We have developed miRdup, a computational predictor for the identification of the most likely miRNA location within a given pre-miRNA or the validation of a candidate miRNA. MiRdup is based on a random forest classifier trained with experimentally validated miRNAs from miRbase, with features that characterize the miRNA-miRNA* duplex. Since we observed that miRNAs have sequence and structural properties that differ between species, mostly in terms of duplex stability, we trained various clade-specific miRdup models and obtained increased accuracy. MiRdup self-trains on the most recent version of miRbase and is easy to use. Combined with existing pre-miRNA predictors, it will be valuable for both de novo mapping of miRNAs and filtering of large sets of candidate miRNAs obtained from transcriptome sequencing projects. MiRdup is open source under the GPL and available at http://www.cs.mcgill.ca/~blanchem/mirdup/.

2.3 Introduction

MicroRNAs (miRNAs) are short (generally 19 to 24 nucleotides) non-coding single-stranded RNA molecules that are involved in post-transcriptional regulation by targeting messenger RNAs (Ambros 1989; Ruvkun 2001; Swami 2010). In animals, miRNAs expression is a multi-step process (Lee et al. 2004): (i) transcription of the primary miRNAs (pri-miRNAs) by RNA polymerase II, (ii) cleavage of the pri-miRNA by Drosha and the RNAse III enzyme to isolate long hairpins called miRNA

precursors (pre-miRNAs), and (iii) extraction by Dicer of the miRNA-miRNA* duplex from the pre-miRNA. In plants, Drosha and Dicer are replaced by Dicer Like 1 (DCL1) (Cuperus et al. 2011). The miRNA* is the complementary region of the miRNA on the other arm of the hairpin with a shift of 2 nucleotides in the 5' direction (Friedländer et al. 2008)). After separation of the two strands of the duplex, the miRNA is mature and ready to be attached to the RISC complement. It then targets mRNAs by perfect or imperfect complementarity (Schwarz et al. 2003). In some cases, both miRNA and miRNA* are functional (Lagos-Quintana et al. 2002).

Over the past years, a number of studies have shown the involvement of miRNAs in most biological process (Lim et al. 2005). They are involved in developmental and physiological roles in animals and plants (Carrington and Ambros 2003; Cuellar and McManus 2005), such as differentiation of embryonic (Suh et al. 2004), muscle (Ritchie et al. 2009), skeletal (Chen et al. 2006), hematopoietic (Shivdasani 2006) and many other types of cells. They are also known to control cell death (Ambros 2004) and proliferation (Brennecke et al. 2003), insulin secretion (Poy et al. 2004) or lipid metabolism (Wilfred et al. 2007). Loss (Miska et al. 2007) and misregulation (Clop et al. 2006) of microRNAs also play an important role in several diseases (Castanotto and Rossi 2009; Cooper et al. 2009), such as cancers (Murchison et al. 2007; Osada and Takahashi 2007). Finally, several studies revealed that organisms under various stress have a responsive miRNAs signature pattern, allowing resistance and adaptation(Guy 1990; Jones-Rhoades and Bartel 2004; Fujii et al. 2005; Sunkar et al. 2006; Saqib et al. 2008). MiRNAs are even used by viruses to infect hosts (Pfeffer et al. 2005; Sarnow et al. 2006; Nelson 2007).

Although experimental techniques for unambiguous identification of miRNAs exist (Berezikov et al. 2006a), they remain slow and expensive. Sequencing of short RNAs followed by mapping to a reference genome has become an approach of choice (Sunkar et al. 2008; Zhang et al. 2010; Schulte et al. 2010), but many small RNA molecules are unlikely to be miRNAs, while many true miRNAs are likely to be expressed only

under rare circumstances not easily covered experimentally. For those reasons, computational prediction of miRNAs continues to play a very important role in genomics.

Most miRNA prediction approaches rely, at least in part, on the specific hairpin shape of the secondary structure of the pre-miRNA (Grey et al. 2005). These include ProMir (Nam et al. 2005, 2006), TripletSVM (Xue et al. 2005), miRabela (Sewer et al. 2005), miPred (Jiang et al. 2007), SSCprofiler (Oulas et al. 2009), microPred (Batuwita and Palade 2009), HHMMiR (Kadri et al. 2009), SplamiR (Thieme et al. 2011), miRFinder (Bonnet et al. 2010), MiRenSVM, the only tools that handle multiloop hairpins (Ding et al. 2010), and many others. All these tools are trained on known miRNAs stored in MiRbase (Griffiths-Jones et al. 2006), a repository of miRNAs (mostly) experimentally validated. The prediction of the hairpin can be combined with comparative genomics approaches that posit that, in addition to their typical secondary structure, pre-miRNAs exhibit high sequence and structure conservation across species (Lindow et al. 2007; Griffiths-Jones et al. 2008). However, most computational approaches labelled as miRNA predictors are actually pre-miRNA predictors, in the sense that they identify candidate genomic regions that may form pre-miRNAs but rarely attempt to determine the position of the miRNA itself within them.

Computationally predicted pre-miRNAs are often combined with high-throughput short-RNA sequencing data, in an attempt to determine which of the large number of expressed small RNAs may indeed be microRNAs. This kind of approach is challenging, though, as short reads may be incorrectly mapped, or may come from degradation products from the pre-miRNA, especially from the miRNA*, or from other types of RNA molecules. Predictions from deep sequencing can be obtained by considering the abundance and distribution of reads mapped to a candidate pre-miRNA, where read stacks and Dicer products mapped on a reference inform about the location of the miRNA. This strategy is used by miRdeep (Friedländer et al. 2008; Friedlander et al. 2012), miRdeep* (An et al. 2013), MIReNA (Mathelier and Carbone

2010) and miRanalyzer (Hackenberg et al. 2009, 2011). However, lowly expressed miRNA, often lineage-specific (Fahlgren et al. 2010) or condition-specific (Breakfield et al. 2012) ones, will be difficult to detect because Dicer products and the miRNA* are completely degraded.

To the best of our knowledge, only six mature miRNA predictors have been proposed to date. MIRcheck (Jones-Rhoades and Bartel 2004) identifies 20-nt regions of a given plant pre-miRNA using a predetermined set of rules and constraints. MiRalign (Wang et al. 2005) finds miRNAs positions by aligning pre-miRNAs with miRbase, thereby preventing from finding new miRNAs. ProMir (Nam et al. 2005) identifies human pre-miRNAs and their mature miRNAs by combining sequence and structural features in a paired hidden Markov model. MatureBayes (Gkirtzou et al. 2010) identifies 22-nt regions that are likely mature miRNA candidates based on sequence and secondary structure information using a Naive Bayes classifier. MaturePred (Xuan et al. 2011) locates fixed length miRNAs in plants based on miRNA-miRNAs* features and a support vector machine predictor. Finally, MiRmat (He et al. 2012) seeks Drosha and Dicer processing sites in vertebrates using a random forest predictor.

Although the recent research activity related to miRNA prediction shows the importance of the problem, existing tools have severe limitations. First, most tools are trained specifically on data from certain phyla (e.g. plants (Jones-Rhoades and Bartel 2004), humans (Xue et al. 2005; Jiang et al. 2007), or viruses (Pfeffer et al. 2005)), which limits their applicability. Second, most mature miRNA prediction tools seek mature miRNA of a fixed length, although in most species miRNAs lengths vary from 19 to 24 nt. Third, tools are typically trained once, at the time of publication, based on the training data available at that time. This means that they do not benefit from the rapid increase in the quality and quantity of experimentally verified miRNAs available. Finally, accessibility remains an issue, with ProMir 2 being unavailable and MaturePred, MiRalign and MiRmat being only available as web servers, which limits that usability for large-scale analyses.

In this paper, we introduce miRdup (miRNA duplex), a tool for the validation of a candidate mature miRNA or the prediction of the precise position and length of the mature miRNA within a candidate pre-miRNA, based on a combination of sequence and structural features. We trained models separately on data from 5 lineages (mammals, fishes, arthropods, nematodes and plants), which increases species-specificity and allows the discovery of features that distinguishes miRNAs from different species. The algorithm works on both single-hairpin and multiloop pre-miRNAs. Finally, miRdup automatically downloads and trains on the latest miRbase release, to ensure it benefits from the most up-to-date data.

2.4 Material and methods

2.4.1 Datasets

MiRNAs and pre-miRNAs sequences were downloaded from miRbase (http://www.mirbase.org/) (Griffiths-Jones et al. 2008) (Griffiths-Jones et al. 2008) release 19, which contains 19,823 unique mature miRNAs/pre-miRNAs pairs. We note that until recently, miRNAs and miRNA* used to be annotated separately in mirBase and were thought to be functionally distinct, with the former playing a functional role and the latter being a non-functional by-product. This view has changed now due to reports of functional activity of miRNAs* (Yang et al. 2011), and miRbase has stopped distinguishing between the miRNAs and miRNAs* (miRbase blog 27 April 2011). We chose to follow this direction by considering all miRNAs and miRNAs* as functional, labelling them as either 3 prime or 5 prime depending on their location on the pre-miRNA hairpin. We note, however, that for more than 78% of cases, only one miRNA is annotated in a given pre-miRNA, with the complementary region not being annotated as functional.

For the purpose of training classifiers, negative sets of non-miRNAs were generated as follows. For each positive example (pair of miRNA and pre-miRNA), a negative example was generated by randomly relocating the miRNA along the same pre-

miRNA sequence, preserving the miRNA length, but excluding the exact position of the true miRNA or of any other known miRNAs. Note that because of the non-deterministic selection of the negative examples, training results vary very slightly from run to run. The complete training dataset consisted of 19,823 positive examples and an equal number of negative examples.

2.4.2 Feature vectors and training

Each training example was represented as a set of 100 features listed in Supplementary Tables (section 2.7). The minimum free energy (MFE) and the secondary structure of the pre-miRNAs and the miRNA-miRNA* candidate duplexes were obtained with RNAfold and RNAduplex, from Vienna package (Hofacker et al. 1994), using default parameters. To perform the ranking of attributes and classifier training and evaluation, we used Weka and its libraries (Hall et al. 2009). All classifiers were trained using 10-fold cross-validation. Attributes ranking was performed using information gain evaluator (*InfoGain evaluator*) (Dash and Liu 1997) with the *Ranker* search method (Hong 1997) in Weka with default parameters and 10-fold cross validation. Ranker ranks attributes by their individual evaluations in conjunction with other attribute evaluators such as ReliefF (Robnik-Sikonja and Kononenko 2003), GainRatio (Quinlan 1986) and Entropy (Shannon et al. 1949).

MiRdup uses a random forest classifier (a combination of decision tree predictors trained on a random subset of features sampled independently (Breiman 2001)), combined with the Adaboost M1 method (Freund and Schapire 1995). Adaboost is a machine learning meta-algorithm that is used in combination of many other machine learning algorithms in order to improve their performance (Osman and Kelly 1996). The random forest was trained with an unlimited maximum depth of the trees and 50 generated trees (Weka options: -I 50 -K 0 -S 1). Adaboost used 10 iterations, reweighting, and a weight threshold of 100 (Weka options: -P 100 -S 1 -I 10). The other classifiers considered were (i) a support-vector machine (SVM) classifier (Cortes and Vapnik 1995), working with libSVM library (Chang and Lin 2001), using with

radial kernel (Weka options: -S 1 -K 2 -D 3 -G 0.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.0010 -P 0.1) and (ii) the C4.5 decision tree classifier (J48) (Quinlan 1993), trained with Adaboost (Weka options: AdaBoostM1 -P 100 -S 1 -I 10 -W trees.J48 -- -C 0.25 -M 2).

The efficiency of a given classifier was measured as a function of its number of true positive (TP), false positive (FN), true negative (TN), and false negative (FN) predictions. A classifier performance is typically measured by its sensitivity $Sn = TP/(TP + FN)$ and specificity $Sp = TN/(TN + FP)$, as well as by its total prediction accuracy $ACC = (TP + TN) / (TP + TN + FP + FN)$ (Jiang et al. 2007) and its Matthew's correlation coefficient (Matthews 1975) $\mathrm{MCC} = \frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})\times(\mathrm{TN}+\mathrm{FN})\times(\mathrm{TP}+\mathrm{FN})\times(\mathrm{TN}+\mathrm{FP})}}$.

2.4.3 MiRNA prediction

MiRdup can be used in two modes. In the validation mode, miRdup takes as input a pre-miRNA sequence and the position of a candidate miRNA, and returns a score that reflects the likelihood that the candidate is a true miRNA. In the prediction mode, the only input to miRdup is a pre-miRNA sequence, and it evaluates all possible miRNAs and reports the most likely miRNA-containing duplex. For each candidate starting position p and length $16 \leq l \leq 30$ on a pre-miRNA of length n, miRdup calculates score(p, l) using the random forest classifier, as described above. Although candidate miRNAs could simply be ranked based on these scores, we found that the following post-processing approach produced more accurate predictions. We first calculate, for each starting position p, the consensus scores for starting position score $S(p)$ and ending position $E(p)$:

$$S(p) = \sum_{\substack{16\leq l\leq 30 \text{ s.t.} \\ score(p,l)>0.99}} score(p, l)$$

$$E(p) = \sum_{\substack{16 \le l \le 30 \ s.t. \\ score(p-l+1,l) > 0.99}} score(p - l + 1, l)$$

We then identify the position p and length l that results in the largest combined start and end position scores:

$$Predicted\ miRNA = \underset{\substack{16 \le l \le 30 \\ 1 \le p \le n-l+1}}{\text{argmax}} \{S(p) + E(p + l - 1)\}$$

2.5 Results and discussion

We developed miRdup, a classifier for the mature miRNA validation and identification in a given pre-miRNA sequence (Methods). In the former case, miRdup assigns a score to a given candidate mature miRNAs within its pre-miRNA sequence. In the latter, it determines the most likely position of a mature miRNA within a given pre-miRNA sequence. MiRdup is based on a random forest binary classifier using a set of sequence and structural features of the candidate miRNA-miRNA* duplex. By training mirDup on lineage-specific subsets of miRbase, one obtains classifiers that can take advantage of miRNA features that are specific to that clade, which helps improve the accuracy of predictions. Here, we report on the accuracy of miRdup predictions in various settings, and contrast sequence and structure features that are informative for five selected clades: Mammals (mostly primates, rodents, and carnivores), plants (mostly crucifers, maize, and rice), fish (mostly zebrafish, fugu, etc.), arthropods (insects, crustaceans, etc.) and nematodes (Caenorhabditis, *P. pacificus*, etc.).

2.5.1 Evaluation of individual predictive features

We evaluated a set of sequence and structural features (Supplementary Tables (section 2.7), summarized in Table II-1 that may potentially help characterize the position of the miRNA on the pre-miRNA hairpin. They were chosen based on previous studies focusing on miRNA prediction (Xue et al. 2005; Jiang et al. 2007) and on the many properties that could characterize the duplex. These include numerical features

describing the position and length of particular structural elements in the putative miRNA, such as bulges and bases pairs, or distance of the miRNA from the start/end of the hairpin (Figure II-1). We also included summary statistics on the primary miRNA sequence (e.g. mononucleotide and dinucleotide frequencies) and the predicted secondary structure of the miRNA/miRNA* duplex (frequency of base pairs types (G-U, C-G or A-U), frequency of local sequence/structure triplets[27] (Xue et al. 2005), and minimum free energy of the duplex). We note that we also considered adding structural features based on ensembles of structures rather than minimum-free energy structures. However, these features did not prove more informative than their MFE-based counterparts and were not retained.

Features vary in their power to distinguish positive from negative examples. Identifying and removing uninformative features is often important to avoid overfitting and improve computational time (Zhou et al. 2006), although this problem is less of an issue for algorithms based on decision trees and forests of random decision trees (Robnik-Sikonja 2004) than for SVMs (Guyon et al. 2002). Features were ranked based on the information gain they provide (Table II-2). We observe that the most influential features are those related to structural aspects of the miRNA-miRNA* duplex (number of base pairs, MFE, number/size of bulges, position of miRNA in the pre-miRNA hairpin loop). On the opposite, primary sequence features and triplet frequencies showed little discriminative power. We note that because our positive and negative examples were size-matched, miRNA length was not considered informative.

---

[27] A sequence/structure triplet corresponds to a nucleotide coupled by the sequence of presence/absence of base-pairing at that position and the two flanking positions. For example, “A(.(“ represents a case where a nucleotide A is in a bulge surrounded by two base pairs, and “U.(.” means that a U is paired but its two neighbours are not.

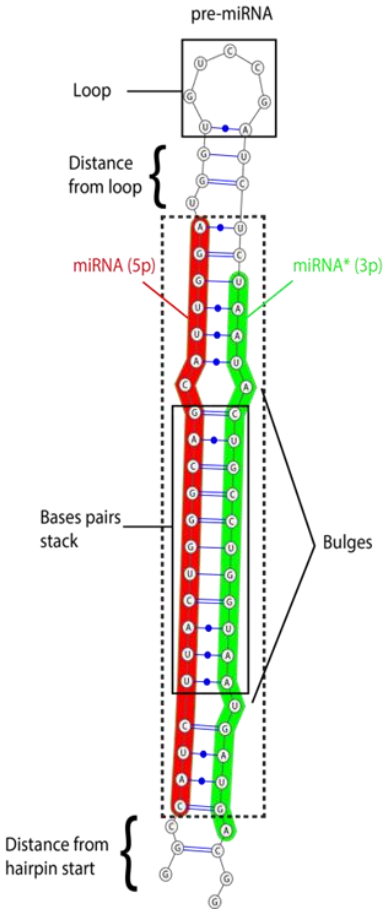


Figure II-1: Pre-miRNA hairpin

Table II-1 : Features used in miRdup

Features	Number	Description
miRNA primary sequence		
Single nucleotide frequency	4	Frequency of each nucleotide
Dinucleotide frequency	16	Frequency of each dinucleotide
GC content	1	Frequency of C or G
First/last nucleotide	8	Nucleotide type at the miRNA start and end
Length	1	miRNA length
miRNA-miRNA duplex		
Triplets	32	Frequency of each sequence/structure triplet (Xue et al. 2005)
Bulges	22	Bulge(s) at positions -4 to +4 nt around start and end of the miRNA. Bulges lengths and number of bulges in the miRNA.
Base pairing	10	Average number of base pairs in duplex and in a sliding window of length 3, 5 and 7 nt. Presence and start position of a perfect 5, 10 and 20 nt base pairs.
Pairs type	3	Percentage of bases forming each type of canonical/wobble base pairs (C-G, A-U, G-U) in the duplex
Loop	2	Percentage of the miRNA overlapping the hairpin loop
Minimum free energy	1	Minimum free energy of the duplex

Table II-2 : Attribute ranking scores evaluated on all miRbase, mammals, and plants datasets with Information Gain ranker. Scores are based on the information gain between the attribute and the class (Hall et al. 2009). Best score is bold. Features with substantially different scores (>0.05) in mammals vs plants are underlined. Full ranking values are in Supplementary tables (section 2.7 for miRbase, mammals and plants.

Features (22)	↓ miRbase Rank score		Mammals Rank score		Plants Rank score		Arthropods Rank score		Nematodes Rank score		Fishes Rank score	
Average number of paired bases in 3 bp sliding widow	**0.186**	**(1)**	0.181	(2)	**0.218**	**(1)**	0.165	(2)	0.190	(5)	0.220	(5)
Length of the longest bulges (% of miRNA length)	0.185	(2)	0.176	(3)	0.203	(5)	0.153	(4)	0.190	(3)	0.193	(3)
Length of the longest bulges (nt)	0.183	(3)	0.175	(4)	0.197	(7)	0.147	(6)	0.196	(2)	0.189	(2)
Average number of paired bases in 5 bp sliding widow	0.174	(5)	0.171	(5)	0.21	(4)	0.163	(3)	0.168	(6)	0.201	(6)
Distance to the terminal loop	0.174	(4)	**0.248**	**(1)**	0.151	(9)	**0.190**	**(1)**	**0.306**	**(1)**	**0.274**	**(1)**
Number of paired bases in the miRNA-miRNA* duplex	0.165	(6)	0.151	(8)	0.213	(3)	0.137	(7)	0.182	(4)	0.188	(4)
Average number of paired bases in 7 bp sliding widow	0.159	(7)	0.156	(7)	0.2	(6)	0.136	(8)	0.146	(7)	0.181	(7)
Length of miRNA overlap within the hairpin loop	0.147	(8)	0.167	(6)	0.107	(14)	0.115	(9)	0.145	(8)	0.150	(8)
Minimum free energy of the duplex	0.137	(9)	0.112	(10)	0.214	(2)	0.162	(5)	0.102	(12)	0.196	(12)
Percentage of GC base pairs in the duplex	0.122	(10)	0.09	(14)	0.102	(15)	0.060	(16)	0.059	(17)	0.068	(17)
Percentage of AU base pairs in the duplex	0.118	(11)	0.068	(18)	0.083	(19)	0.027	(22)	0.058	(18)	0.046	(18)
Triplet U...	0.117	(12)	0.114	(9)	0.124	(10)	0.106	(10)	0.107	(11)	0.128	(11)
Distance to the start of the hairpin	0.112	(13)	0.094	(13)	0.155	(8)	0.077	(14)	0.144	(9)	0.107	(9)
Triplet A...	0.111	(14)	0.099	(12)	0.113	(11)	0.085	(12)	0.126	(10)	0.114	(10)
miRNA included in loop (yes/no)	0.107	(15)	0.105	(11)	0.076	(20)	0.058	(17)	0.088	(14)	0.067	(14)
Triplet C...	0.082	(16)	0.074	(17)	0.09	(17)	0.063	(15)	0.091	(13)	0.101	(13)
Percentage of GU base pairs in the duplex	0.074	(17)	0.076	(16)	0.084	(18)	0.034	(20)	0.045	(19)	0.055	(19)
Triplet G...	0.068	(18)	0.08	(15)	0.069	(21)	0.082	(13)	0.082	(15)	0.110	(15)
Position of the first 5 nt bulge-free region	0.066	(19)	0.059	(19)	0.098	(16)	0.103	(11)	0.076	(16)	0.124	(16)
Triplet G(((	0.059	(20)	0.029	(22)	0.058	(22)	0.027	(21)	0.022	(21)	0.040	(21)
Maximum length without bulges (nt)	0.058	(22)	0.05	(21)	0.112	(12)	0.038	(19)	0.037	(20)	0.056	(20)
Maximum length without bulges (% of the miRNA length)	0.058	(21)	0.051	(20)	0.11	(13)	0.049	(18)	0.033	(22)	0.063	(22)

2.5.2 Mature miRNAs exhibit species-specific properties

We then assessed the power of each feature at distinguishing true miRNA from negative examples in specific lineages. Figure II-2 A-H shows distribution of feature values for some of those that vary significantly between lineages based on

Kolmogorov–Smirnov test (p-value <0.05 for at least one of the comparisons between lineage-specific distribution and the distribution obtained from all MirBase). The length of miRNAs varies significantly between species, where plant miRNA are generally 21 nt long and almost never 23nt, while animal miRNAs have a broader, more regular miRNA length distribution with a mode at 22 nt (Figure II-2 A). Plant miRNAs also stand out with duplexes that are on average more stable (lower free energy) than animals (Figure II-2 B), while arthropods and, to a lesser extent, nematodes, are often less stable. This is also reflected in various structural properties such as the presence of fewer and shorter bulges (Figure II-2 C, D). In fact, more than 13% of plant miRNAs have no bulge at all (100% base-paired positions, Figure II-2 E) and more than 33% have at least 10 consecutive base pairs starting at positions 0 (start) or 1 (Figure II-2 F), two properties that are much more rare in animals. 60% to 90% of animal miRNAs are located within 10 bp of the terminal loop of the pre-miRNA (Figure II-2 G), whereas plant miRNAs are often found much further, in agreement with the fact that plants usually have longer precursors (Zhang et al. 2006). The GC content of miRNAs exhibits significant variations between species (Figure II-2 H), with fish miRNAs being notably less GC-rich than those of other species. Finally, we noted a remarkable nucleotide composition bias at the first position of the miRNA with 40% (in mammals) to 60% (in fish) of miRNAs starting with a U nucleotide (Figure II-2 I).

Feature ranking was then repeated on each set of species separately. While certain structural features such as those relating to the number of base pairings ranked consistently high for all lineages, others, in agreement with the results presented in Figure II-2, are ranked quite differently for different species (Table II-2 and Supplementary tables (section 2.7). In particular, the distance to and overlap with the terminal loop showed decreased informativity in plants, while the minimum free energy and the number of base pairs in the duplex were more informative in plants than animals.

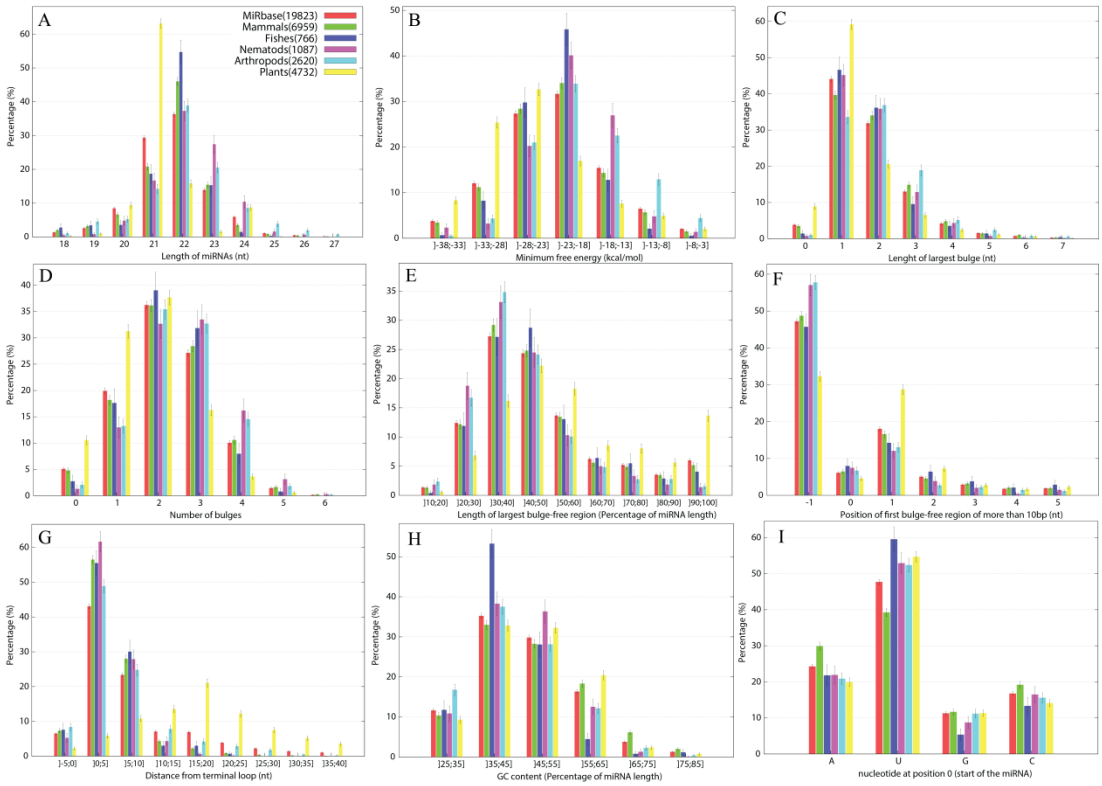


Figure II-2: Properties of microRNAs from six different lineages : all eukaryotes (19,823 miRNAs), mammals (6,959), fish (766), nematodes (1,087), arthropods (2,620) and plants (4,732). Each panel shows the distribution of a selected feature. (A) MiRNA length (nt). (B) Minimum free energy of the miRNA-miRNA* duplex (kcal/mol). (C) Length of the largest bulge in the miRNA (nt). (D) Number of bulges in the miRNA-miRNA* duplex. (E) Length of longest bulge-free stem in the miRNA-miRNA* duplex. (F) Start position of the first 10 nt bulge-free stem in the miRNA-miRNA* duplex; -1 means no such region is present. (G) Distance to the terminal loop of the hairpin (nt). (H) miRNA GC-content. (I) Nucleotide type (A, U, G or C) at the first position of the miRNA.

2.5.3 Training and evaluation of miRNAs classifiers

We first evaluated the classification accuracy of various binary classifiers that, when presented with a candidate miRNA and its pre-miRNA, determine whether the candidate is a positive or negative example. Classifiers were first trained on a balanced data set consisting of 19,823 miRNAs from MirBase (irrespective of species) and the same number of negative instances (randomly selected regions of actual pre-miRNAs, with lengths matched with positive examples; see Methods) and were evaluated using 10-fold cross-validation (Table II-3). Classifiers included a support vector machine

(SVM, using a radial basis kernel), a decision tree classifier (C4.5 with Adaboost), and a random forest classifier (with Adaboost). Other learning algorithms were also considered but were found to be less accurate (data not shown); these included RIPPER (Cohen 1995), a feed-forward artificial neural network (Collobert and Bengio 2004), and a logistic regression classifier (Cessie et al. 1992). Each classifier was trained on either the full set of 100 features or on the subset of 22 best features of the Table II-2. The best overall prediction accuracies were obtained by the random forest classifier using all features (Figure II-3, Table II-3), with an accuracy of 80.6%, an area under the ROC curve of 89.2%, and a Matthews correlation coefficient (MCC) of 61.4%. Boosting on C4.5 tree produced similar but slightly inferior results. The SVM classifier trained using all features performed poorly, with an AUC at 75.8%. The SVM's accuracy improved slightly when restricted to only the 22 most informative features but it remained inferior to that of the random forest classifier.

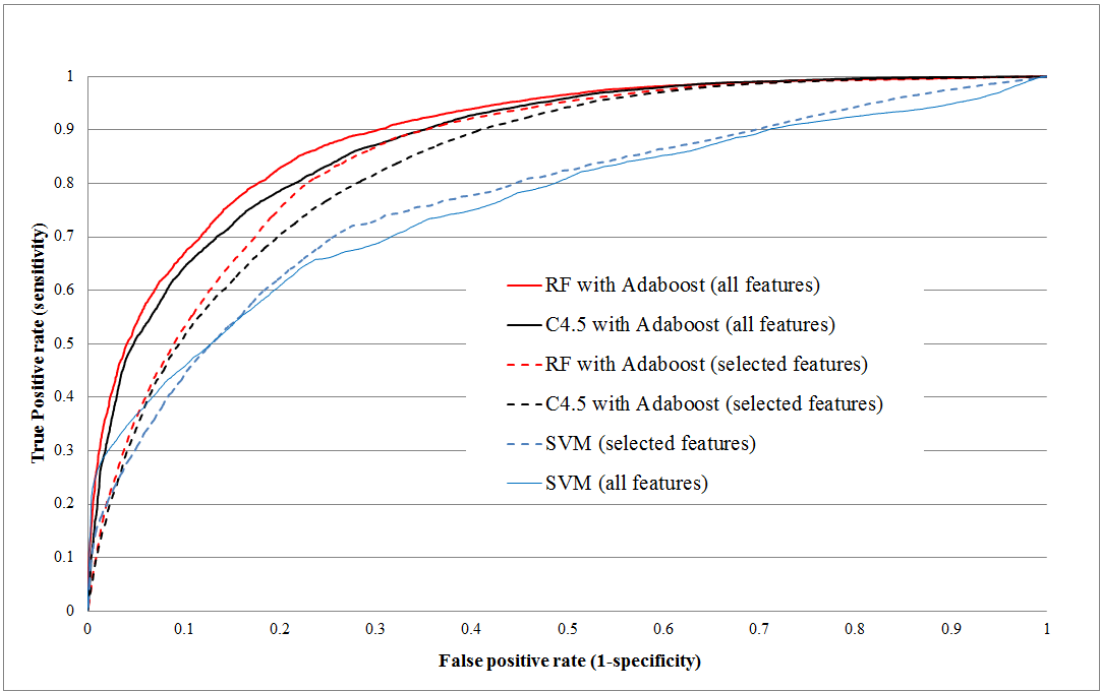


Figure II-3: Receiver-operating characteristic (ROC) curves of classifiers trained on the complete mirBase dataset. See selected features in Table 2.

Table II-3: Results of various classifiers trained on all features of miRbase (all lineages) evaluated using 10-fold cross-validation.

Classifier	Correctly Classified Instances (out of 39,646)	Sensitivity	Specificity	Accuracy	MCC	AUC
Random Forest with AdaBoost	31,940	0.863	0.748	0.806	0.614	0.892
C4.5 decision tree with AdaBoost	31,317	0.809	0.771	0.79	0.58	0.875
SVM with radial basis kernel	25,878	0.344	0.962	0.653	0.385	0.653

Knowing that miRNAs properties are different between species, the training and evaluation steps were repeated separately on each of the five clades. We chose to train lineage-specific classifiers using the random forest classifier with no feature selection, as this is the approach that worked best on the full data set. Results are presented in Table II-4. Accuracy levels were generally improved as compared to the multi-lineage classifier, ranging from 81.7% to 86.4%, but with the exception of arthropods, for which the predictions were only 77.7% accurate. For most of lineages, the accuracy of the lineage-specific predictor (measured using 10-fold cross-validation) is also higher than that of predictors trained on another lineage (Table II-5). The inferior performance of the arthropod-specific predictor is likely due to a combination of the small size of the dataset, the variability of features within the dataset, and large diversity of species within the dataset.

Table II-4: Prediction accuracy of lineage-specific miRdup predictors (Random forest with Adaboost, evaluated using 10-fold cross-validation).

Classifier	Number of instances	Correctly Classified Instances	Sensitivity	Specificity	ACC	MCC	AUC
Mammals	13,918	11,415	0.868	0.772	0.82	0.642	0.897
Plants	9,464	7,734	0.866	0.768	0.817	0.636	0.904
Nematods	2,174	1,789	0.882	0.764	0.823	0.649	0.898
Arthropods	5,240	4,071	0.833	0.721	0.777	0.557	0.857
Fish	1,530	1,323	0.905	0.824	0.864	0.731	0.918

Table II-5 : Accuracy of lineage-specific and non-lineage-specific miRdup predictors (rows) for the prediction of miRNAs from each lineage (columns). The highest accuracy for each column is in bold. For cases where a predictor is applied to data from the lineage it is trained on, the numbers reported are obtained by 10-fold cross-validation.

Test set Training set	miRbase	Nematods	Arthropods	Fish	Mammals	Plants
miRbase	0.806	0818	0.807	0.852	0.808	0.790
Nematods	0.74	0.823	0.755	0.82	0.768	0.618
Arthropods	0.768	0.812	0.777	0.806	0.765	0.712
Fish	0.716	0.808	0.72	0.864	0.741	0.606
Mammals	0.793	0.834	0.766	0.846	0.820	0.655
Plants	0.700	0.662	0.644	0.681	0.645	0.817

To illustrate an important use of miRdup, we used it to reanalyze a set of 1670 miRNA predicted by MiRdeep2 (Friedlander et al. 2012) from short-RNA sequencing data in human cancer lines (SRA SRR029124). MiRdup-mammals validated only 755 (45%) of these candidate miRNAs. There are multiple lines of evidence that suggest that the candidates that were rejected by miRdup were indeed MiRdeep2 false-positives. First, only 3% of the candidate miRNAs that were rejected by miRdup overlapped annotated miRNAs from MiRbase, whereas this fraction was of 47% among candidate miRNAs that were validated by miRdup. Second, we observe that among the miRNAs predicted by MiRdeep2 and validated by miRdup, a large proportion (46.5%) overlap highly conserved sequences among mammals (based on PhastCons highly conserved elements (Siepel et al. 2005)), whereas this proportion drops to only 19.2% among MiRdeep2 miRNA predictions that were rejected by miRdup. These results suggest that the candidate miRNAs rejected by miRdup are either non-functional, or are atypical, unannotated, and poorly conserved miRNAs. Finally, we also reanalyzed the pool of pre-miRNAs and their mature miRNAs predicted and published by the authors of miRdeep (Friedländer et al. 2008), and miRdeep2 (Friedlander et al. 2012). MiRdup confirmed 89% (201 on 226) and 84% (98 on 117) of the identified miRNAs respectively.

2.5.4 Prediction of a miRNA position within a pre-miRNA

MiRdup can be used to predict the most likely miRNA duplex location, i.e. the most likely miRNA in 5 prime (5p) and 3 prime (3p), within a given pre-miRNA. Given a pre-miRNA sequence and a trained classifier for the binary decision problem, miRdup computes prediction scores for every possible combination of miRNA length (16-30 nt) and starting position and then identifies the pair of starting and ending positions, located within 16 to 30 nt of each other, for which the total evidence is highest (see Methods). We finally return the predicted miRNA and its miRNA*. Figure II-4 shows an example of the prediction made for a typical pre-miRNA, drosophila melanogaster's dme-mir-10.

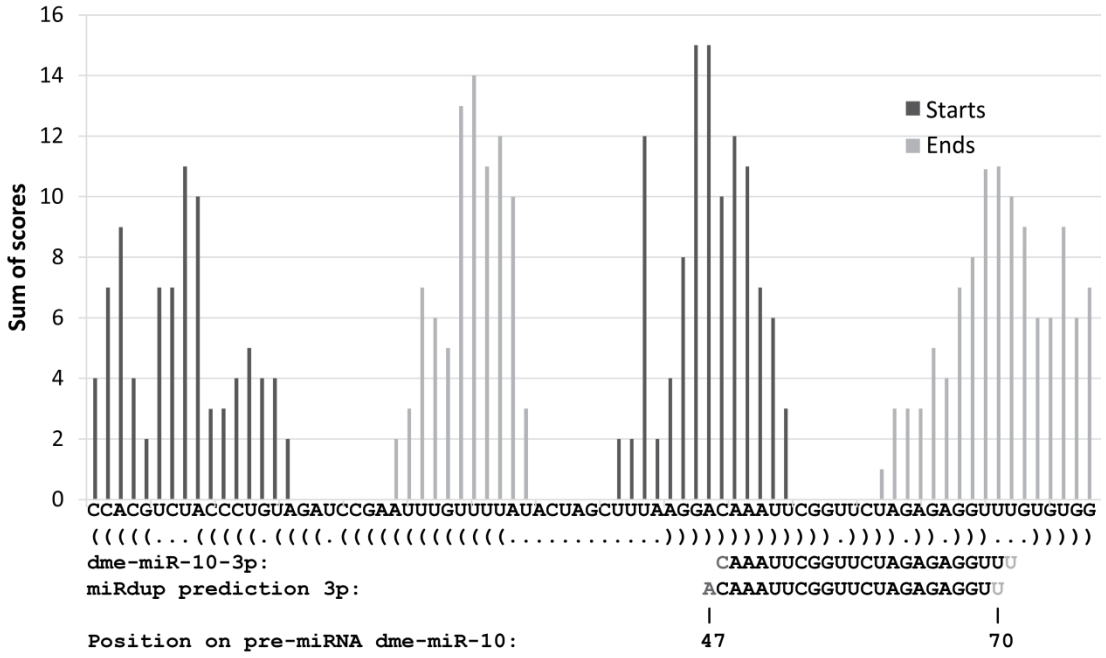


Figure II-4: Example of miRdup prediction on the Drosophila melanogaster dme-mir-10 pre-miRNA. The actual miRNA predicted in 3p extends from positions 48 to 71 of the pre-miRNA, and the predicted miRNA in 5p, the miRNA*, is CUACCCUGUAGAUCCGAAUUUGUU. Dark columns show the score assigned by miRdup-arthropods to each possible starting position (summed over all possible lengths), and in grey columns are scores of each possible ending position. MiRdup predicts that the miRNA extends from positions 47 to 70, off by one position compared to the true miRNA.

To estimate the accuracy of miRdup at locating miRNAs within pre-miRNAs, we calculated the minimum distance between the true and predicted miRNA/miRNA*, for both the start and end positions (Figure II-5 and Supplementary Figure (section 2.8). When trained and evaluated on data from all five lineages combined, miRdup made perfect predictions of start and end positions in 28.7% and 20.18% of the cases respectively, and was within 3 nt in 68.9% and 68.3% of cases respectively. This is significantly better than MatureBayes, miRalign and ProMir 1, the only miRNA predictors we were able to compare to. When evaluated on the same data set, MatureBayes yields only 18.8% and 13.3% exact miRNA duplex start and end position predictions, while MiRalign yields 18.8% and 7.9%, MaturePred 10.34% and 9.14%, and ProMir1 6.13% and 8.01%. The results also indicate that about 10% of the predictions are off by more than 10 nucleotides with miRdup, versus about 20% for the best of the competitors (Figure II-5).

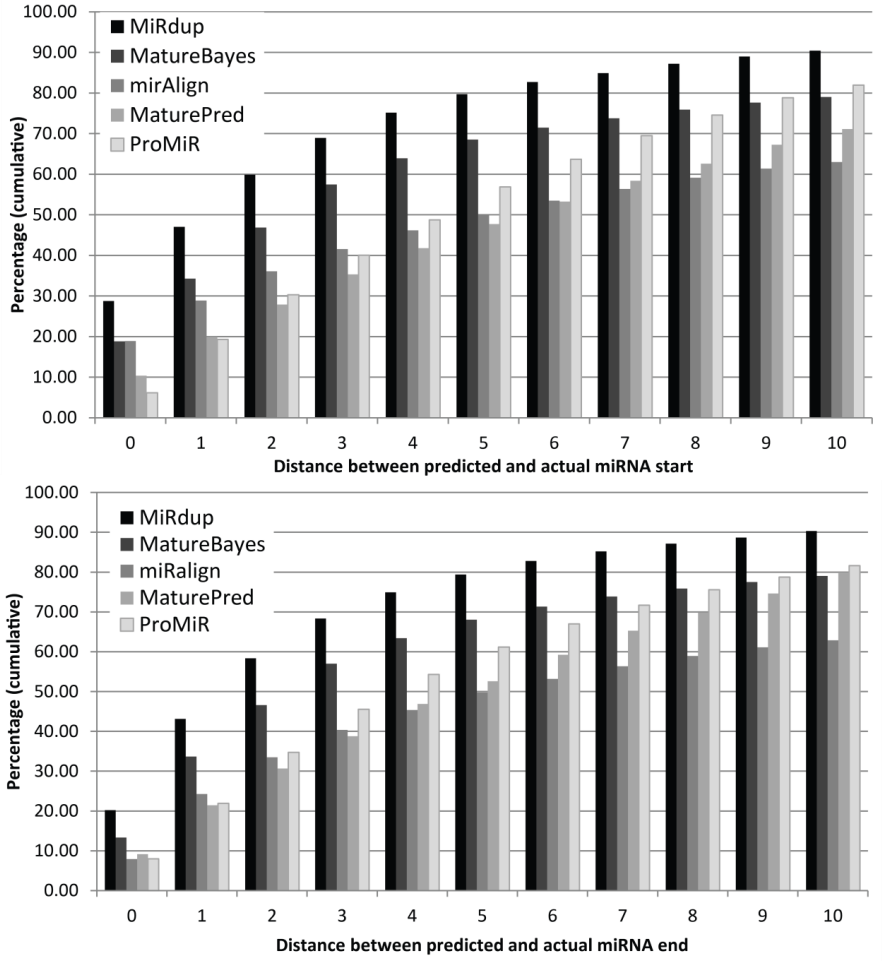


Figure II-5: Cumulative distribution of the minimum distance between the true and predicted miRNAs or miRNAs* starts (up) and ends (down), i.e. the proportion of cases where the prediction is within x bases of the true start/end positions. Multi-lineage miRdup predictions are compared to MatureBayes (Gkirtzou et al. 2010), MiRalign (Wang et al. 2005), MaturePred (Xuan et al. 2011) and PromiR1 (Nam et al. 2005) for all experimentally validated pre-miRNAs from miRbase, except for MaturePred, where our analysis was limited to only 2400 miRNAs submitted due to web server constraints. For MatureBayes and Promir, a small number of queries were rejected by the web server and were thus excluded from the results. We only show distances of up to 10 nt, but in some rare cases errors are substantially larger (up to 250 nt). Results for lineage-specific miRdup compared to MatureBayes for mammals, arthropods, nematods, fish and plants are shown in Supplementary figure (section 2.8).

Results obtained using the appropriate lineage-specific version of miRdup generally improve on the multi-lineage predictor, with 25.9-34.0% (resp. 20.7-24.6%) of start (resp. end) positions predicted exactly correctly (Supplementary Figure (section 2.8)).

Again, fish miRNAs stand out as being the easiest to predict, with 51.5% (resp. 29.6%) of start (resp. end) positions correctly predicted.

2.5.5 The miRdup program

MiRdup is distributed as a java program making use of libraries from the Weka (Hall et al. 2009) and ViennaRNA (Hofacker et al. 1994) packages. The workflow is schematized in Figure II-6. MiRdup can either be trained on a user-provided dataset of known miRNAs and pre-miRNAs, or can automatically download the latest version of mirBase and be trained on all of it or on a lineage-specific subset. For example, if "ruminantia" is specified as clade of interest, the predictor will be trained only on *Bos Taurus* and *Ovis aries*, which are (currently) the only two species present in miRbase in this clade. The set of negative examples is constructed on the fly by randomizing the position of miRNAs on the pre-miRNA. A minimum free energy secondary structure is obtained for each pre-miRNAs and features are calculated. Finally, the random forest predictor is trained. MiRdup can be run in two modes. In the first, miRdup takes as input a pre-miRNA sequence (with or without predicted secondary structure) and a candidate miRNA position, and assigns a score reflecting the likelihood that the candidate is a real miRNA. In the second case, miRdup evaluates every possible combination of miRNA position and length, and reports the most likely pair.

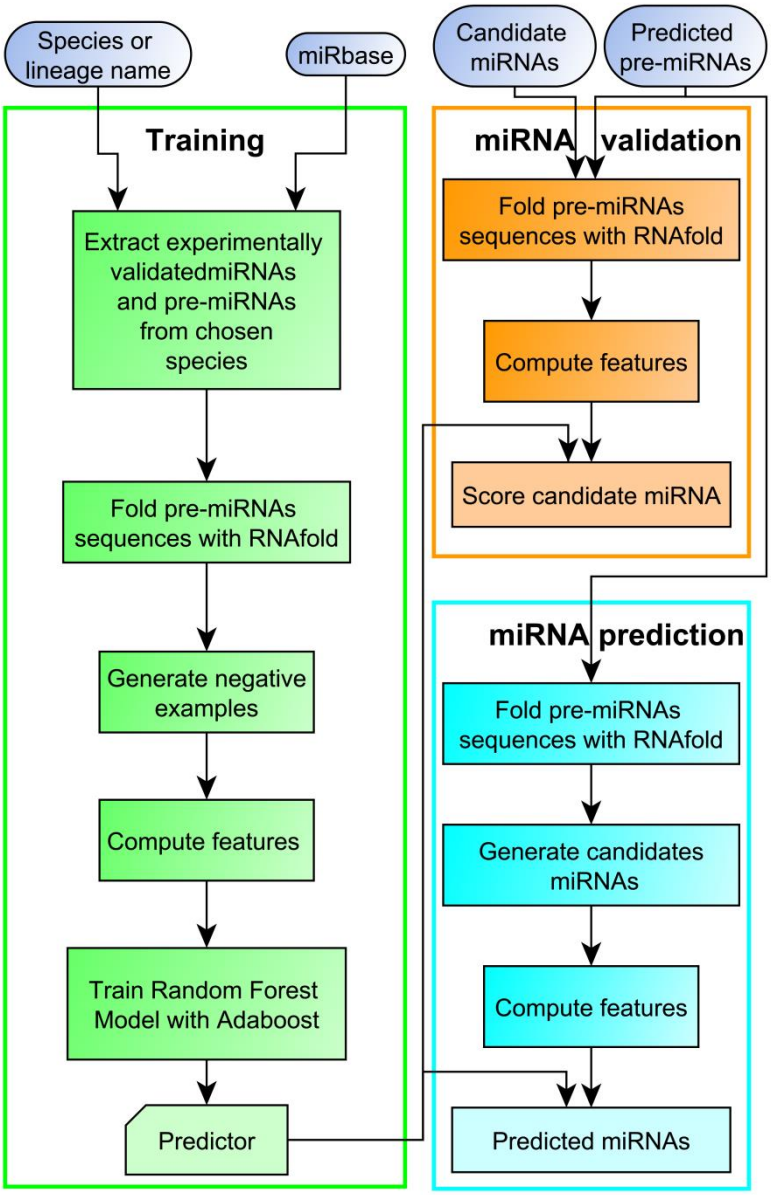


Figure II-6: Workflow of the miRdup algorithm.

Thanks to its relative simplicity, miRdup is fast. On a computer with a single 2.93GHz CPU, the training phase on the complete mirBase database v19 requires less than 80 minutes, and the miRNA prediction phase takes around 10 seconds for a given pre-miRNA of 100 nt.

2.6 Conclusions

Although the structural properties of pre-miRNAs are well characterized (Krol et al. 2004) and have largely been exploited for their predictions (Mendes et al. 2009), the sequence and structure properties that allow Dicer to recognize the exact position of the mature miRNA remains poorly understood (Park et al. 2011). For this reason, computational approaches for the identification of miRNAs within pre-miRNA are rare and relatively inaccurate. Such predictors are, however, of great importance. First, working hand in hand with pre-miRNA predictors, they are essential for the de novo computational miRNA annotation of new genomes. Second, they play an important role even for miRNA annotation projects that have the benefit of short-RNA sequencing data. Indeed, from our experience, the classical approach of identifying likely miRNAs by retaining only reads that map to a genomic regions with strong pre-miRNA potential (as predicted by miPred (Jiang et al. 2007) or HHMMiR (Kadri et al. 2009), for example) still yields tens of thousands predictions. Considering only candidates overlapping pre-miRNAs predicted by more than one tool can reduce this number, but the consequences on sensitivity and specificity are hard to quantify. A more reasonable number of predictions can be obtained by more recent tools such as miRDeep (Friedländer et al. 2008; Friedlander et al. 2012), although even it often produced unlikely miRNA predictions. MiRdup then offers the opportunity to discard these likely false-positives while retaining a high sensitivity.

MiRdup is a flexible, accurate, fast, and user-friendly tool for the localization of mature miRNAs in pre-miRNA. It complements a wide array of computational tools that aim to identify pre-miRNAs and should be used as a post-treatment of predicted hairpins or to validate the miRNA function of short RNA reads mapped to a reference genome. MiRdup's speed and flexibility let it to be trained on data from specific lineages, which allows it to take advantage of species-specific miRNA properties. Because it is automatically trained on the latest version of mirBase, it remains up-to-date and can take advantage of increasingly large and accurate sets of miRNA

annotations. The multi-lineage version of MiRdup outperforms the only other miRNA predictor available for download, matureBayes (Gkirtzou et al. 2010). The lineage-specific version is even more accurate, as it is able to take advantage of features such as the presence of an Uracyl at the first position of the vast majority of fish miRNAs, or the increased stability of the miRNA-miRNA* duplex in plants.

2.7 Supplementary tables: Attribute rankings

2.7.1 Attribute ranking output for miRbase

Search Method: Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 101 Class): Information Gain Ranking Filter

Rank scores	ID	Features
0.183809	40	Average number of paired bases in a sliding widow of 3 nt along the miRNA
0.178933	59	Length of the biggest bulges in percentage of the miRNA length
0.177118	58	Length of the biggest bulges in nucleotides
0.174022	39	Average number of paired bases in a sliding widow of 5 nt along the miRNA
0.173801	34	Distance of the miRNA from the terminal loop of the hairpin
0.163464	7	Bases pairs in the duplex of the miRNA and its complementarity region
0.159278	38	Average number of paired bases in a sliding widow of 7 nt along the miRNA
0.143582	37	Length of the miRNA which overlap in the hairpin loop
0.127442	2	Minimum free energy of the duplex
0.114706	63	Triplet U...
0.108587	60	Triplet A...
0.105129	35	Distance of the start of miRNA from the start of the hairpin
0.100427	93	Percentage of GC base pairs in the duplex
0.094015	92	Percentage of AU base pairs in the duplex
0.092867	36	miRNA included in loop
0.084723	94	Percentage of GU base pairs in the duplex
0.079484	61	Triplet C...
0.072825	62	Triplet G...
0.070982	13	Start of perfect 5 nt base pair in the miRNA
0.059999	4	Maximum length without bulges
0.059992	5	Maximum length without bulges in percentage of the miRNA length
0.042843	45	Bulge at start position 2
0.038907	95	nucleotide at start position 0
0.037558	47	Bulge at start position 3
0.0335	78	Triplet G(((
0.031529	11	Start of perfect 10 nt base pair in the miRNA
0.027518	6	Length without bulges from miRNA start
0.027212	55	Bulge at end position -3
0.024292	10	PresenceOfPerfect10MerBasePair
0.023319	12	PresenceOfPerfect5MerBasePair
0.021186	67	Triplet U(..
0.018154	76	Triplet A(((
0.017903	43	Bulge at position 1
0.017436	3	Percentage of GC content
0.015656	79	Triplet U(((
0.015532	66	Triplet G(..
0.013398	64	Triplet A(..
0.013038	53	Bulge at end position -2
0.011222	44	Bulge at start position -2
0.011043	77	Triplet C(((

0.00878	42	Bulge at start position -1
0.008604	97	nucleotide at start position +1
0.008093	16	Percentage of G
0.00776	9	Start of perfect 20 nt base pair in the miRNA
0.007667	65	Triplet C(..
0.007279	8	PresenceOfPerfect20MerBasePair
0.007188	56	Bulge at end position +4
0.007093	49	Bulge at end position 0
0.006354	51	Bulge at end position -1
0.006253	26	Percentage of dinucleotides AG
0.00574	22	Percentage of dinucleotides AU
0.005649	96	nucleotide at start position -1
0.005625	71	U((.
0.004176	23	Percentage of dinucleotides UU
0.004003	28	Percentage of dinucleotides GG
0.003838	57	Number of bulges
0.003728	27	Percentage of dinucleotides UG
0.003705	18	Percentage of dinucleotides AA
0.003683	99	nucleotide at end position -1
0.003494	14	Percentage of dinucleotides A
0.003112	20	Percentage of dinucleotides GA
0.003001	15	Percentage of U
0.002952	69	Triplet C((.
0.002788	32	Percentage of dinucleotides GC
0.002381	87	Triplet U..(
0.002362	48	Bulge at start position -4
0.002161	19	Percentage of dinucleotides UA
0.00194	46	Bulge at start position -3
0.001851	70	Triplet G((.
0.001651	68	Triplet A((.
0.001602	85	Triplet C..(
0.001488	86	Triplet G..(
0.001233	41	Bulge at start position 0
0.0012	31	Percentage of dinucleotides UC
0.001184	84	Triplet A..(
0.00115	74	Triplet G.((
0.001109	72	Triplet A.((
0.001057	82	Triplet G.(.
0.001009	81	Triplet C.(.
0.000973	100	nucleotide at end position +1
0.000869	17	Percentage of C
0.000819	54	Bulge at end position +3
0.000743	83	Triplet U.(.
0.000637	88	Triplet A(.(
0.000523	25	Percentage of dinucleotides CU
0.000421	24	Percentage of dinucleotides GU
0.000412	50	Bulge at end position +1
0.000376	75	Triplet U.((
0.000369	52	Bulge at end position +2
0.000306	80	Triplet A.(.
0.000271	89	Triplet C(.(
0.000208	98	nucleotide at end position 0
0	1	Length
0	21	Percentage of dinucleotides CA
0	29	Percentage of dinucleotides CG
0	30	Percentage of dinucleotides AC
0	33	Percentage of dinucleotides CC
0	73	Triplet C.((
0	90	Triplet G(.(
0	91	Triplet U(.(

2.7.2 Attribute ranking output for Mammals

Rank scores	ID	Features
0.247652	34	Distance of the miRNA from the terminal loop of the hairpin
0.1811	40	Average number of paired bases in a sliding widow of 3 nt along the miRNA
0.176017	59	Length of the biggest bulges in percentage of the miRNA length
0.175173	58	Length of the biggest bulges in nucleotides
0.171369	39	Average number of paired bases in a sliding widow of 5 nt along the miRNA
0.166916	37	Length of the miRNA which overlap in the hairpin loop
0.156208	38	Average number of paired bases in a sliding widow of 7 nt along the miRNA
0.151227	7	Bases pairs in the duplex of the miRNA and its complementarity region
0.114016	63	Triplet U...
0.112425	2	Minimum free energy of the duplex
0.10498	36	miRNA included in loop
0.099441	60	Triplet A...
0.094377	35	Distance of the start of miRNA from the start of the hairpin
0.090307	93	Percentage of GC base pairs in the duplex
0.079723	62	Triplet G...
0.076208	94	Percentage of GU base pairs in the duplex
0.073829	61	Triplet C...
0.068416	92	Percentage of AU base pairs in the duplex
0.058998	13	Start of perfect 5 nt base pair in the miRNA
0.050537	5	Maximum length without bulges in percentage of the miRNA length
0.04952	4	Maximum length without bulges
0.037305	47	Bulge at start position 3
0.030906	95	nucleotide at start position 0
0.029182	45	Bulge at start position 2
0.028684	78	Triplet G(((
0.024654	11	Start of perfect 10 nt base pair in the miRNA
0.022153	55	Bulge at end position -3
0.020967	76	Triplet A(((
0.019976	66	Triplet G(..
0.019604	10	PresenceOfPerfect10MerBasePair
0.019576	12	PresenceOfPerfect5MerBasePair
0.019012	67	Triplet U(..
0.018733	6	Length without bulges from miRNA start
0.016738	79	Triplet U(((
0.01174	56	Bulge at end position +4
0.011673	43	Bulge at position 1
0.011558	53	Bulge at end position -2
0.011164	64	Triplet A(..
0.009842	96	nucleotide at start position -1
0.009452	77	Triplet C(((
0.009243	42	Bulge at start position -1
0.009063	97	nucleotide at start position +1
0.008585	65	Triplet C(..
0.008039	51	Bulge at end position -1
0.007622	44	Bulge at start position -2
0.007532	99	nucleotide at end position -1
0.006332	3	Percentage of GC content
0.006258	49	Bulge at end position 0
0.004997	98	nucleotide at end position 0
0.004951	8	PresenceOfPerfect20MerBasePair
0.004951	9	Start of perfect 20 nt base pair in the miRNA
0.004827	71	U((.
0.003586	16	Percentage of G
0.003445	69	Triplet C((.
0.003346	22	Percentage of dinucleotides AU
0.002345	86	Triplet G..(
0.002264	82	Triplet G.(.
0.002256	54	Bulge at end position +3
0.002032	74	Triplet G.((
0.001912	26	Percentage of dinucleotides AG

0.001907	17	Percentage of C
0.001902	72	Triplet A.((
0.001868	57	Number of bulges
0.001827	70	Triplet G((.
0.001761	83	Triplet U.(.
0.001581	100	nucleotide at end position +1
0.001552	14	Percentage of dinucleotides A
0.00143	31	Percentage of dinucleotides UC
0.001339	28	Percentage of dinucleotides GG
0.001323	68	Triplet A((.
0.001287	87	Triplet U..(
0.001166	23	Percentage of dinucleotides UU
0.001134	48	Bulge at start position -4
0.001092	27	Percentage of dinucleotides UG
0.000989	88	Triplet A(.(
0.000912	15	Percentage of U
0.000905	25	Percentage of dinucleotides CU
0.00088	18	Percentage of dinucleotides AA
0.000861	46	Bulge at start position -3
0.000767	85	Triplet C..(
0.000765	81	Triplet C.(.
0.000754	33	Percentage of dinucleotides CC
0.000705	80	Triplet A.(.
0.000491	41	Bulge at start position 0
0.000415	50	Bulge at end position +1
0.000373	52	Bulge at end position +2
0	1	length
0	19	Percentage of dinucleotides UA
0	20	Percentage of dinucleotides GA
0	21	Percentage of dinucleotides CA
0	24	Percentage of dinucleotides GU
0	29	Percentage of dinucleotides CG
0	30	Percentage of dinucleotides AC
0	32	Percentage of dinucleotides GC
0	73	Triplet C.((
0	75	Triplet U.((
0	84	Triplet A..(
0	89	Triplet C(.(
0	90	Triplet G(.(
0	91	Triplet U(.(

2.7.3 Attribute ranking output for Plants

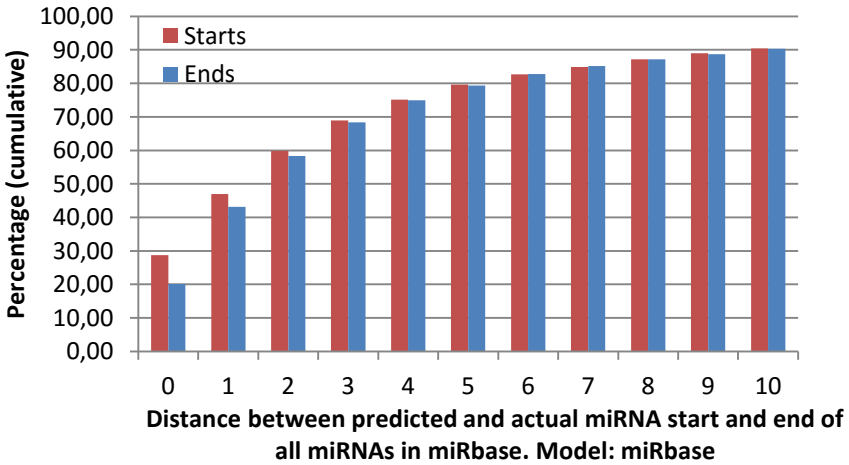
Rank scores	ID	Features
0.2183596	40	Average number of paired bases in a sliding widow of 3 nt along the miRNA
0.2136904	2	Minimum free energy of the duplex
0.2134001	7	Bases pairs in the duplex of the miRNA and its complementarity region
0.2098282	39	Average number of paired bases in a sliding widow of 5 nt along the miRNA
0.2029981	59	Length of the biggest bulges in percentage of the miRNA length
0.2001381	38	Average number of paired bases in a sliding widow of 7 nt along the miRNA
0.1973837	58	Length of the biggest bulges in nucleotides
0.1552158	35	Distance of the start of miRNA from the start of the hairpin
0.1508973	34	Distance of the miRNA from the terminal loop of the hairpin
0.1241754	63	Triplet U...
0.1127512	60	Triplet A...
0.1121255	4	Maximum length without bulges
0.1101996	5	Maximum length without bulges in percentage of the miRNA length
0.1072775	37	Length of the miRNA which overlap in the hairpin loop
0.1022462	93	Percentage of GC base pairs in the duplex
0.0981264	13	Start of perfect 5 nt base pair in the miRNA

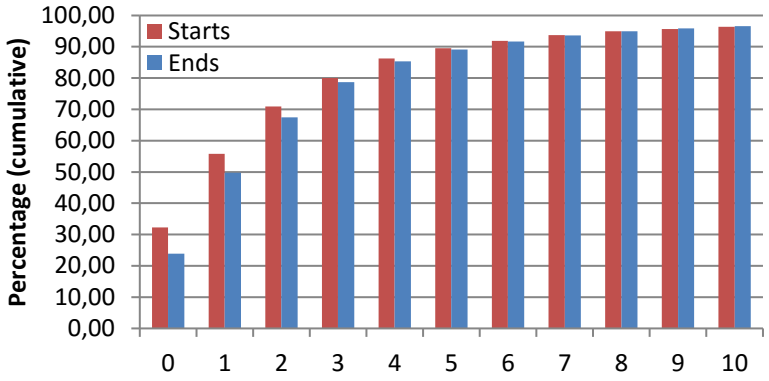
0.0897533	61	Triplet C...
0.0835688	94	Percentage of GU base pairs in the duplex
0.0826534	92	Percentage of AU base pairs in the duplex
0.0763566	36	miRNA included in loop
0.0709132	11	Start of perfect 10 nt base pair in the miRNA
0.0689285	62	Triplet G...
0.0583529	10	PresenceOfPerfect10MerBasePair
0.0581855	45	Bulge at start position 2
0.0581738	78	Triplet G(((
0.0527604	3	Percentage of GC content
0.0487617	6	Length without bulges from miRNA start
0.0485925	55	Bulge at end position -3
0.044935	95	nucleotide at start position 0
0.0396789	12	PresenceOfPerfect5MerBasePair
0.0373814	43	Bulge at position 1
0.0372496	57	Number of bulges
0.0366634	47	Bulge at start position 3
0.0305678	16	Percentage of G
0.0295102	53	Bulge at end position -2
0.0272025	77	Triplet C(((
0.0267036	64	Triplet A(..
0.0263878	67	Triplet U(..
0.0257003	9	Start of perfect 20 nt base pair in the miRNA
0.0246115	8	PresenceOfPerfect20MerBasePair
0.0224664	19	Percentage of dinucleotides UA
0.0220973	15	Percentage of U
0.0210338	18	Percentage of dinucleotides AA
0.020926	23	Percentage of dinucleotides UU
0.0199528	76	Triplet A(((
0.0195947	20	Percentage of dinucleotides GA
0.0181526	26	Percentage of dinucleotides AG
0.0175972	22	Percentage of dinucleotides AU
0.0161085	97	nucleotide at start position +1
0.015936	32	Percentage of dinucleotides GC
0.0155202	49	Bulge at end position 0
0.0148667	28	Percentage of dinucleotides GG
0.0146388	98	nucleotide at end position 0
0.0142824	44	Bulge at start position -2
0.0142462	87	Triplet U..(
0.0136328	51	Bulge at end position -1
0.0133243	65	Triplet C(..
0.0130869	100	nucleotide at end position +1
0.0120069	14	Percentage of dinucleotides A
0.011292	79	Triplet U(((
0.0101996	41	Bulge at start position 0
0.0092681	27	Percentage of dinucleotides UG
0.0082792	71	U((.
0.0080999	84	Triplet A..(
0.0071784	66	Triplet G(..
0.0064138	42	Bulge at start position -1
0.0063921	17	Percentage of C
0.0062835	33	Percentage of dinucleotides CC
0.0052242	21	Percentage of dinucleotides CA
0.0049397	24	Percentage of dinucleotides GU
0.0048082	85	Triplet C..(
0.0046388	68	Triplet A((.
0.0040023	29	Percentage of dinucleotides CG
0.0035412	86	Triplet G..(
0.0033861	69	Triplet C((.
0.0020496	99	nucleotide at end position -1
0.0020327	81	Triplet C.(.
0.001623	82	Triplet G.(.
0.0015835	73	Triplet C.((
0.0015628	75	Triplet U.((
0.001333	83	Triplet U.(.

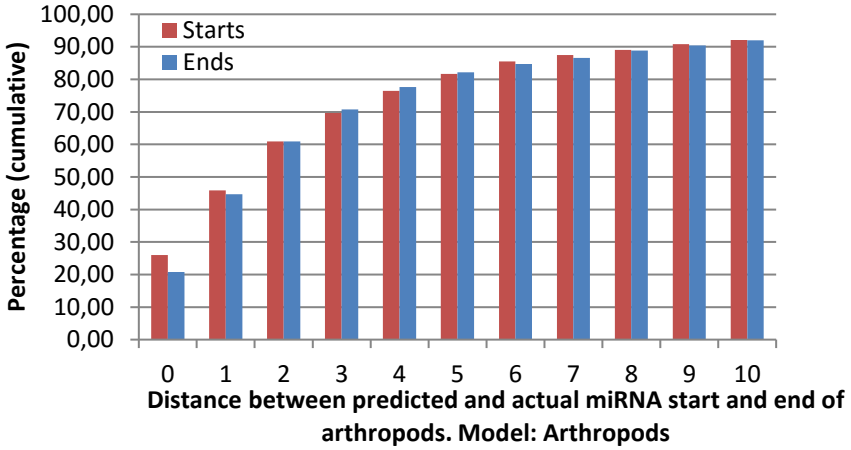
0.0012765	52	Bulge at end position +2
0.0012623	89	Triplet C(.(
0.0011736	54	Bulge at end position +3
0.0010792	72	Triplet A.((
0.0008841	80	Triplet A.(.
0.000736	56	Bulge at end position +4
0.0006831	96	nucleotide at start position -1
0.0005589	46	Bulge at start position -3
0.0002641	50	Bulge at end position +1
0.0000113	48	Bulge at start position -4
0	1	length
0	25	Percentage of dinucleotides CU
0	30	Percentage of dinucleotides AC
0	31	Percentage of dinucleotides UC
0	70	Triplet G((.
0	74	Triplet G.((
0	88	Triplet A(.(
0	90	Triplet G(.(
0	91	Triplet U(.(

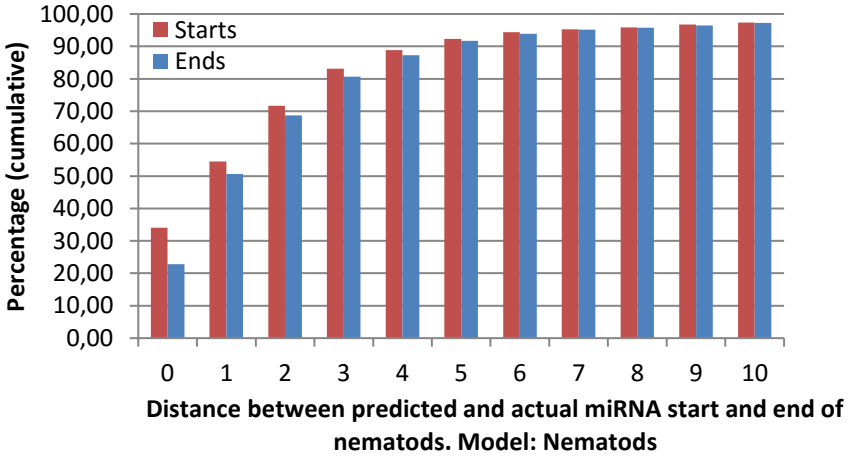
2.8 Supplementary figures

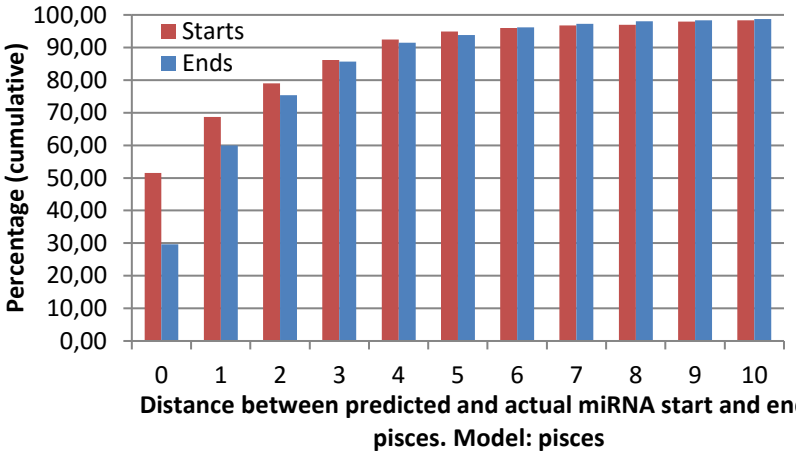
Cumulative distribution of the distance between the true and predicted miRNAs starts and ends, i.e. the proportion of cases where the prediction is within *x* bases of the true start/end positions. We only show distances of up to 10 nt, but in some rare cases errors are substantially larger (up to 250 nt). Results are presented for lineage-specific miRdup for mammals, arthropods, nematods, fish and plants.

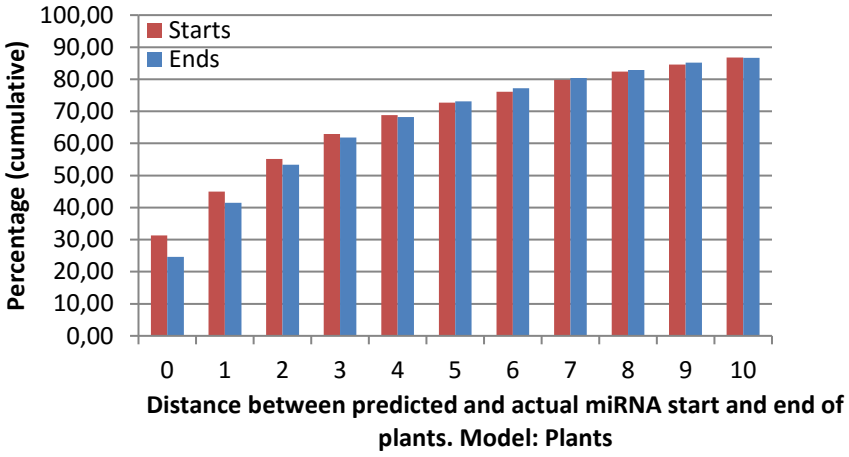



100,00
90,00
80,00
70,00
60,00
50,00
40,00
30,00
20,00
10,00
0,00
Percentage (cumulative)
Starts
Ends
0 1 2 3 4 5 6 7 8 9 10
Distance between predicted and actual miRNA start and end of mammals. Model: Mammals


100,00
90,00
80,00
70,00
60,00
50,00
40,00
30,00
20,00
10,00
0,00
Percentage (cumulative)
Starts
Ends
0 1 2 3 4 5 6 7 8 9 10
Distance between predicted and actual miRNA start and end of arthropods. Model: Arthropods


100,00
90,00
80,00
70,00
60,00
50,00
40,00
30,00
20,00
10,00
0,00
Percentage (cumulative)
Starts
Ends
0 1 2 3 4 5 6 7 8 9 10
Distance between predicted and actual miRNA start and end of nematods. Model: Nematods


100,00
90,00
80,00
70,00
60,00
50,00
40,00
30,00
20,00
10,00
0,00
Starts
Ends
Percentage (cumulative)
0 1 2 3 4 5 6 7 8 9 10
Distance between predicted and actual miRNA start and end of pisces. Model: pisces


100,00
90,00
80,00
70,00
60,00
50,00
40,00
30,00
20,00
10,00
0,00
Starts
Ends
Percentage (cumulative)
0 1 2 3 4 5 6 7 8 9 10
Distance between predicted and actual miRNA start and end of plants. Model: Plants

CHAPTER III : EVOLUTIONARY MECHANISMS LEADING TO THE CREATION OF NEW MIRNAS IN PRIMATES REVEALED BY THE ANALYSIS OF INFERRED ANCESTRAL SEQUENCES

3.1 Preface

Since Darwin's work on theory of evolution (Darwin 1859), an effort to classify the species has been performed by numerous scientists, in order to design the genealogic tree of life. Deep sequencing and alignment algorithms allow constructing trees based on genetic information, which is highly reliable compared to evolutionary trees were elaborated based on morphology traits. We are now able to reconstruct, to a certain extent, lost species genomes. The results presented in this chapter use this concept and bring new insights in the comprehension of the human genome evolution history.

The study presented in this chapter aims to estimate the period of origin of human miRNAs and determine the mechanisms that lead to their creation in the genome. MiRNAs are the result of million years of evolution, undergoing various mechanisms of new miRNA genes origination and selective pressure. A miRNA gene exists because several conditions were met in the past: (1) the original sequence became a Drosha/Dicer compatible sequence, (2) the mature miRNA(s) had an accessible target gene repertoire, and (3) the miRNA gene was preserved in the genome by natural selection. The first statement implies that a sequence, after many genetic modifications, once transcribed, is able to fold into a hairpin shape recognizable and processed by Drosha and Dicer enzymes. Then, the functionality of the resulting miRNAs is determined by its capacity to target messenger RNAs. Finally, the

conservation of the new miRNA gene in the genome will depend on natural selection and fitness. If the miRNA has unwanted targets (i.e. detrimental) or its silencing control has no impact on the organism (i.e. neutral), then selective pressure is likely to eliminate it from the population. Otherwise, it is more likely to be driven to fixation.

To identify the mechanisms of origination, the genome sequences of the mammalian ancestors were reconstructed computationally, for a total of 12 ancestral genomes. Extracting from this data the ancestral sequences of human pre-miRNAs allows us to identify the period of origin of the miRNA, i.e. the first time it became functional. Then, by comparing the sequence at and before the period of origin, we identify the type of genetic modifications the ancestral sequence underwent to become a functional miRNA. This project makes directly use of miRdup, presented in the CHAPTER II , because we need to validate if a mature miRNA is compatible with a pre-miRNA at a given ancestral state. Before miRdup, there was no tool to perform this task, which now opens new horizons for miRNA research.

The rest of this chapter is reproduced from:

> *Leclercq M, Diallo AB, Blanchette M (2016) Evolutionary mechanisms leading to the creation of new miRNAs in primates revealed by the analysis of inferred ancestral sequences.* Paper in finalization for submission.

3.2 Abstract

MicroRNAs (miRNA) are short single stranded RNA molecules derived from hairpin-forming precursors that play a crucial role as key post-transcriptional regulators in eukaryotes and viruses. In recent years, period of origin and birth/death rates of many miRNAs have been estimated in various plant and animal clades, but our understanding of the evolutionary mechanisms leading to the creation of new miRNAs remains poor. Here, we propose an approach that uses ancestral genome reconstruction to determine the period of origin and mechanism of creation of a set of 488 primate-specific human miRNAs. This is achieved by first computationally reconstructing

ancestral mammalian genome sequences, based on an alignment of 42 sequenced mammalian genomes. Extant and ancestral sequences orthologous to human miRNA precursors are then analyzed using pre-miRNA and mature mRNA prediction tools to determine the period of origin of these miRNAs. Finally, we have created a classification pipeline to analyze the evolutionary history of each miRNA in order to identify their mechanisms of origination. A total of nine mechanisms have been observed. The results show that half of human miRNA genes are primate-specific, and a large proportion of them were derived from transposable elements or were created *De novo* by random mutations.

3.3 Introduction

MicroRNAs (miRNAs) are short non-coding single-stranded RNA molecules that are involved in post-transcriptional regulation (Ambros 1989; Ruvkun 2001; Swami 2010). In animals, primary miRNAs transcripts are produced by RNA polymerase II and then cleaved by Drosha and the RNAse III enzyme to isolate long hairpins called miRNA precursors (pre-miRNAs). Subsequently, the miRNA-5p/3p duplex is separated from the hairpin by Dicer, and one of the two strands (or sometimes both (Lagos-Quintana et al. 2002)), called the mature miRNA, is attached to a RISC (RNA-induced silencing complex) complement to target messenger RNAs (Lee et al. 2004).

The human genome encodes approximately 2500 miRNAs, divided in 1500 families, according to the most recent version of miRNA repository, miRbase v21 (Griffiths-Jones et al. 2006). Some of these families are very ancient and are shared among most vertebrates. For example, miR-100 is believed to the oldest known animal miRNA, shared by cnidarians and bilaterians (Christodoulou et al. 2010), arisen about 550 million years ago (Martin et al. 2000). However, not all of the miRNAs of the human genome are as old, and nearly half of human miRNA genes have previously been reported being primate specific. The rate of creation of new miRNAs is thus quite high compared to that of protein-coding genes. Estimation of miRNAs birth and death rates have been obtained in recent studies (Iwama et al. 2013; Soumillon et al. 2013;

Heimberg et al. 2010; Lyu et al. 2014; Taylor et al. 2014). The most common approach to identify the period of origin of a miRNA is based on the lowest common ancestor (LCA) algorithm. This method gave insights into the period of origination of human miRNAs, and their gains and losses during mammalian evolution. However, this type of approach is sensitive to incomplete genomes or annotations.

Several mechanisms leading to the creation of new miRNAs have been proposed (Berezikov 2011). These include the complete duplication of an existing miRNA (e.g. through a tandem or segmental duplication). Just like for protein-coding genes, newly formed copies can then evolve to acquire new functions through changes in their miRNA sequence or their regulation (neofunctionalization) (Clancy and Shaw 2008; Hertel et al. 2006). Typically, miRNAs with a high sequence similarity and identical seed regions are clustered into families (Ambros 2003), but the miRNA genes are not always conserved in primary sequence or secondary structure (Zou et al. 2014), making family organization an unreliable mean to classify miRNAs originated as duplication events. Then, a specific approach is needed to detect these events.

Another mechanism involves transposable elements (TE), one of the mechanisms leading to the creation of new miRNAs (Smalheiser and Torvik 2005). These DNA sequences, known to play a key role in genome function and evolution (Bucher et al. 2012), are able to change their position within the genome or can be duplicated in many copies. They provide functions to their hosts by several ways (Kidwell and Lisch 2001), including creation of new coding (Volff 2006) and regulatory sequences (Britten 1996; Jordan et al. 2003; van de Lagemaat et al. 2003). In mammals, many studies discovered miRNAs derived from TE (Piriyapongsa and Jordan 2007a; Borchert et al. 2006; Smalheiser and Torvik 2005), and this is probably only an incomplete list, since large scale bioinformatics approach to find new miRNAs frequently exclude TE sequences from their analysis (Bentwich et al. 2005; Li et al. 2006). TE-derived human miRNAs are generally less conserved (Piriyapongsa et al. 2007), and have been associated to interspersed among Alu (Borchert et al. 2006) and

LINE-2 (Smalheiser and Torvik 2005) transposable elements. Also, by their ubiquity and abundance, TE activity results not only in the emergence of paralogous miRNAs gene families (i.e. hsa-mir-548), but also of multiple target sites dispersed throughout the genome (Piriyapongsa et al. 2007).

Finally, it has been shown that certain miRNAs may arise through random substitutions or short indels (e.g. inverted duplications), especially when located in the right context (e.g. region that is already transcribed) and given a favourable starting point (e.g. a tRNA sequence) (Berezikov 2011), but these are relatively rare.

Overall, despite the excellent work mentioned above based on comparison of extant genomic sequences from multiple species, the accuracy and completeness of the inferred mechanisms of miRNA origination are limited. This is in part because these approaches do not have access to the ancestral sequences corresponding to the time where each miRNA is predicted to have been created.

Dissecting the evolutionary events that led to the creation of a miRNA could perhaps best be achieved if the ancestral genomic DNA sequences immediately predating its creation, as well as that immediately following it, were available. Despite recent progress in sequencing ancient DNA (Mouttham et al. 2015; Shapiro and Hofreiter 2014), samples older than 1 Myrs have never been sequenced. An alternate approach comes from the field of paleogenomics (Salse et al. 2009; Putnam et al. 2008; Chauve and Tannier 2008). Here, one infers ancient DNA sequences computationally, by comparison of a set of related extant genomes. Ancestral genome reconstruction involves several steps, including multiple sequence alignment, inference of insertions/deletions (indels), inference of substitutions and gene rearrangements (Blanchette et al. 2008). Given a multiple sequence alignment of orthologous extant sequences and a phylogenetic tree, the inferAncestors approach (Diallo et al. 2010) infers ancestral sequences at each ancestral node of the tree. Applied to mammalian genomes, this approach has been shown to be highly accurate; in particular, the Boreoeutherian ancestor, an early mammal living approximately 75 Myrs ago, was

reconstructed with estimated base-by-base accuracy of 98-99% (Blanchette et al. 2004a).

In this paper, we use the computationally reconstructed ancestral mammalian sequences to study the key evolutionary mechanisms leading to the creation of new miRNAs, specifically focusing on primate-specific miRNAs. We first estimate the period of origin of many human pre-miRNAs and then identify the set of evolutionary events that have led to their creation. We found that 53% of human miRNAs of miRbase v20 are primate specific, and we characterized nine possible mechanisms of origination. A large proportion of human miRNA genes were created *De novo* by random mutations, but also by insertion of both transposable and non-transposable distal genomic elements.

3.4 Results and discussion

3.4.1 Dating the Period of Origin of Human MiRNAs

Whole-genome alignments for a set of 35 mammals (Figure III-1A) and computationally reconstructed ancestral sequences (see Methods) were analyzed to identify the period of origin (PO) and mechanism of creation of each human miRNA. Two methods were used to infer the PO of each miRNA (see Figure III-1B and Methods). The first one, denoted PO-AR, analyzes computationally inferred ancestral pre-miRNA sequences; it is defined as the most ancient ancestral pre-miRNA sequence that is predicted to form a valid miRNA precursor by MiPred (Jiang et al. 2007) and to contain at least one region that could form a mature miRNA, as predicted by miRdup (Leclercq et al. 2013). The second approach, denoted PO-LCA, is obtained from analyzing only the extant orthologous sequences to each human pre-miRNA, applying the test above to each ortholog, and identifying the least common ancestor (LCA) of the set of species with predicted valid miRNA gene. With both MiPred and miRdup, permissive settings were used to minimize false-negatives, as the fact that the sequences considered are orthologous or ancestral to a functional miRNA in human

places a strong positive prior on their own function as miRNAs. Each of the two approaches assigns a node of the human ancestry (or the human node itself) as the putative period of origin; in reality the creation of a miRNA would have occurred on the branch leading to that node from its immediate ancestor. Based on the phylogenetic tree used for our study (Figure III-1A), we can assign the PO of a miRNA to one of 13 nodes of the phylogenetic tree, ranging from the most recent (human itself, which is labelled as ancestor 0), to the most ancient (ancestor of all mammals, Theria, labelled as ancestor 12).

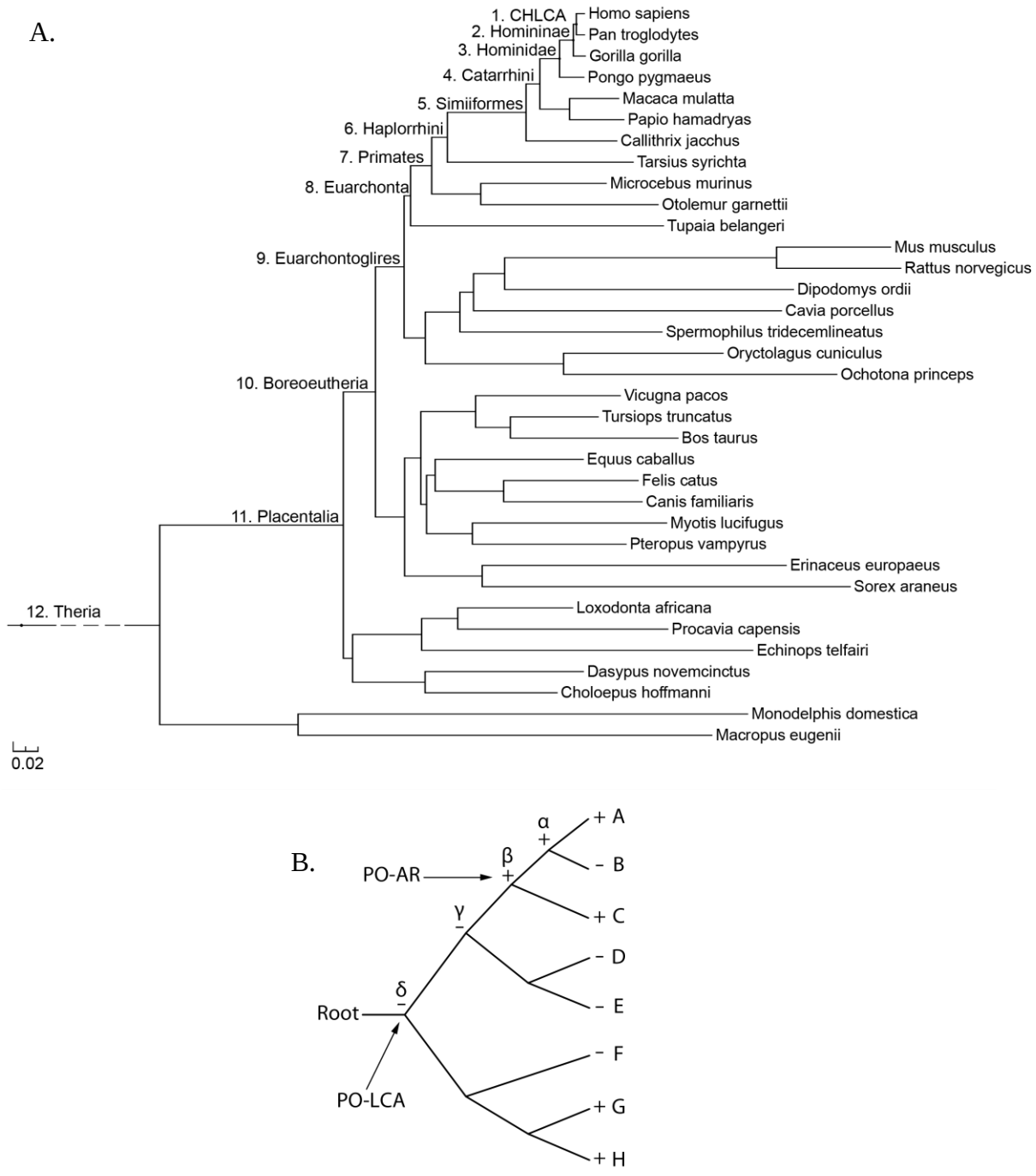


Figure III-1: A: Mammal tree from UCSC genome browser. B: Example of the estimation of the period of origin of a miRNA from the species A. The least common ancestor (PO-LCA) and ancestral reconstruction (PO-AR) are estimated based on the predicted functionality (+ for functional, - for non-functional) of the orthologous and ancestral sequences.

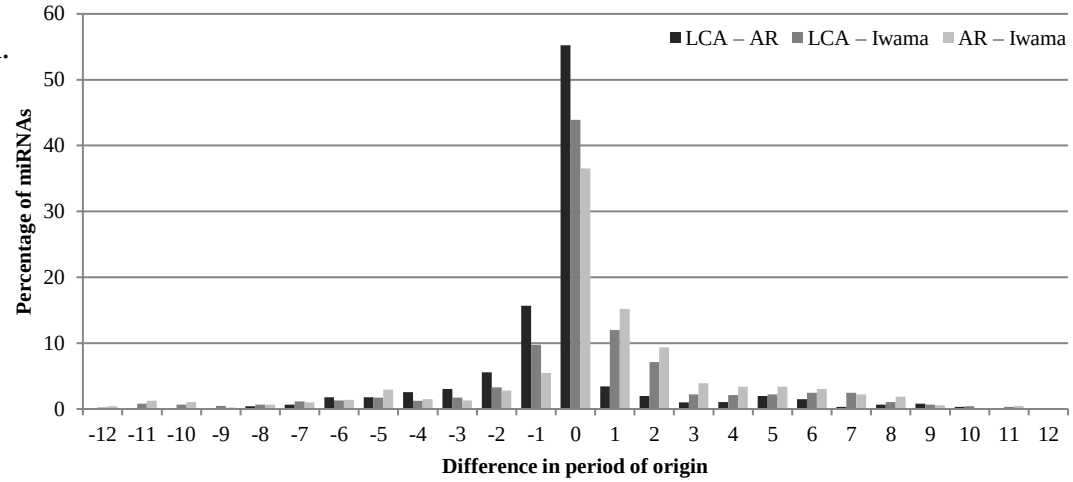
Periods of origination of miRNAs have been previously estimated by Iwama et al. (Iwama et al. 2013) using an LCA-based approach where mammalian orthologous sequences are predicted functional if (i) their seed sequence is identical to that in human, and (ii) the pre-miRNA secondary structure has a minimum folding energy (MFE) less than -13 kcal/mol. Notably, the first criterion is very strict: although mutations in the seed region are likely to have broad effects of the repertoire of target genes, such changes may be expected especially in the early life of a miRNA; our approach (miRdup) uses instead a computational prediction of the ability of Drosha to process the miRNA, irrespective of sequence conservation. Furthermore, the MFE criterion is quite relaxed: whereas 99.99% of human miRNAs (miRbase v21) satisfy this criterion, pseudo-hairpins (i.e. non-Dicer compatible) also frequently do. We replace this criterion by one based on the predictions of miPred (Jiang et al. 2007), a well-established pre-miRNA predictor, which considers not only MFE but also a number of other considerations.

Figure III-2A describes the extent to which the three approaches agree (see also Table SD III-1 A-C). PO-LCA and PO-AR estimates are identical for 55% of miRNAs (and within one for 75%), while the two LCA-based approaches (PO-LCA and PO-Iwama) also are in general agreement (within one for 65%). Unsurprisingly, because it is based on a strict criterion of seed sequence conservation, Iwama's approach tends to assign slightly more recent POs (average value of PO-Iwama is 6.18) compared to our two approaches (average PO-AR=7.6, average PO-LCA=7.4). We manually investigated some of the miRNAs that had inconsistent PO estimates. Disagreements between their method and ours generally fall into two categories: (1) miRNAs that had relatively recent mutations in the seed region but whose pre-miRNA sequence is predicted to be functional in ancestors that predate that mutation, leading to PO-Iwama < PO-AR = PO-LCA; (2) miRNAs whose seed region is conserved far back in time but whose pre-miRNA only appears to be functional in more recent ancestors, leading to PO-Iwama > PO-AR = PO-LCA. Disagreements observed between PO-AR and PO-LCA are

mainly due to incompletely sequenced genomes, incorrect alignments and low ancestral reconstruction prediction confidence score.

The proportion of miRNAs whose PO is assigned to each node by each of the three dating methods is shown in Figure III-2B. Remarkably, 5-7% of all human miRNAs are predicted to be human-specific, and 43-53% have been created on the branch leading to Simiiformes or more recently. This is significantly more than the <15% of human genomic DNA that is primate specific. Two distinct periods have a large number of creations of new pre-miRNAs: the period extending from the Haplorrhini ancestor to the Catarrhini ancestors (PO=5 & PO=4) and the period that predates the Placentals' ancestor (PO=12 & PO=11).

A.



B

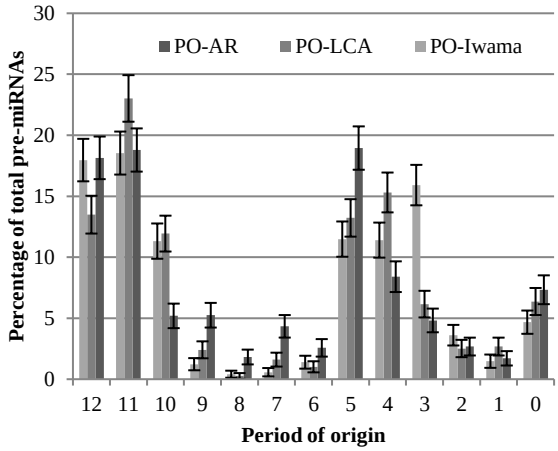


Figure III-2: (A) Difference of period of origin estimation between the three pairs of dating methods. Only the 1219 miRNAs that we could have extracted from Iwama et al. supplementary material are included in this graph. (B) Percentage of miRNAs assigned to each period of origin. Confidence interval corresponds to one standard error on the proportion p of the number of miRNAs n: $p = \sqrt{((p(1-p))/n)}$. Iwama percentages were retrieved from Iwama et al. paper. For B and C, PO-AR and PO-LCA are in percentage of 1868 human miRNAs.

Having dated the origin of each miRNA, we turn to studying the evolutionary mechanisms behind the creation of new miRNAs. We focused here on a set of 488 human miRNAs that are predicted to have been gained on the branch leading to the primate ancestor or more recently (PO-AR≤7) and whose ancestral sequences could be determined with high confidence, based on the following criteria (see also

Methods): (i) Consistent PO estimates from both PO-AR and PO-LCA (|PO-AR – PO-LCA|≤ 1; in case of disagreement, PO-AR is retained); (ii) All intermediate ancestors between the PO and human (including human itself) pass our pre-miRNA and miRNA prediction criteria; (iii) high confidence level on ancestral sequence at estimated PO. These 488 miRNAs represent 51.4% of the set of miRNAs with PO-AR≤7. This selection does not bias in a substantial manner the age distribution of miRNAs considered (See Figure SD III-1). We preferred not to analyze low confidence miRNAs, because erroneous PO estimation would possibly lead to inaccurate characterization of the mechanism of origin based to the genetic changes between PO+1 and PO. These miRNAs were predicted older by PO-AR compared to PO-LCA, due to missing or very low conservation of orthologous sequences.

3.4.2 Increased levels of selective pressure follows the period of origin.

To confirm our estimates of period of origin, we studied miRNA's and pre-miRNA's mutation rates before (PO+1) and after (PO-1) their predicted period of origin. A sudden decrease in observed mutation rate in a given genomic region hints to increase selective pressure associated to a gain of function. For each miRNA, we compared the mutation rate (substitutions and indels) between its sequence at PO and its modern version in human, and contrasted it with the mutation rate between sequences at PO+1 and PO (see Methods). We observe a significant decrease in insertion and deletion rates after the prediction PO (Figure III-3A and B), and a slight decrease in the substitution rate (Figure III-3C). The decrease in mutation rates in most notable in the seed region, which is indeed the region that is believed to be under the strongest levels negative selection due to its role in target selection. These observations support the hypothesis that PO estimates are generally accurate.

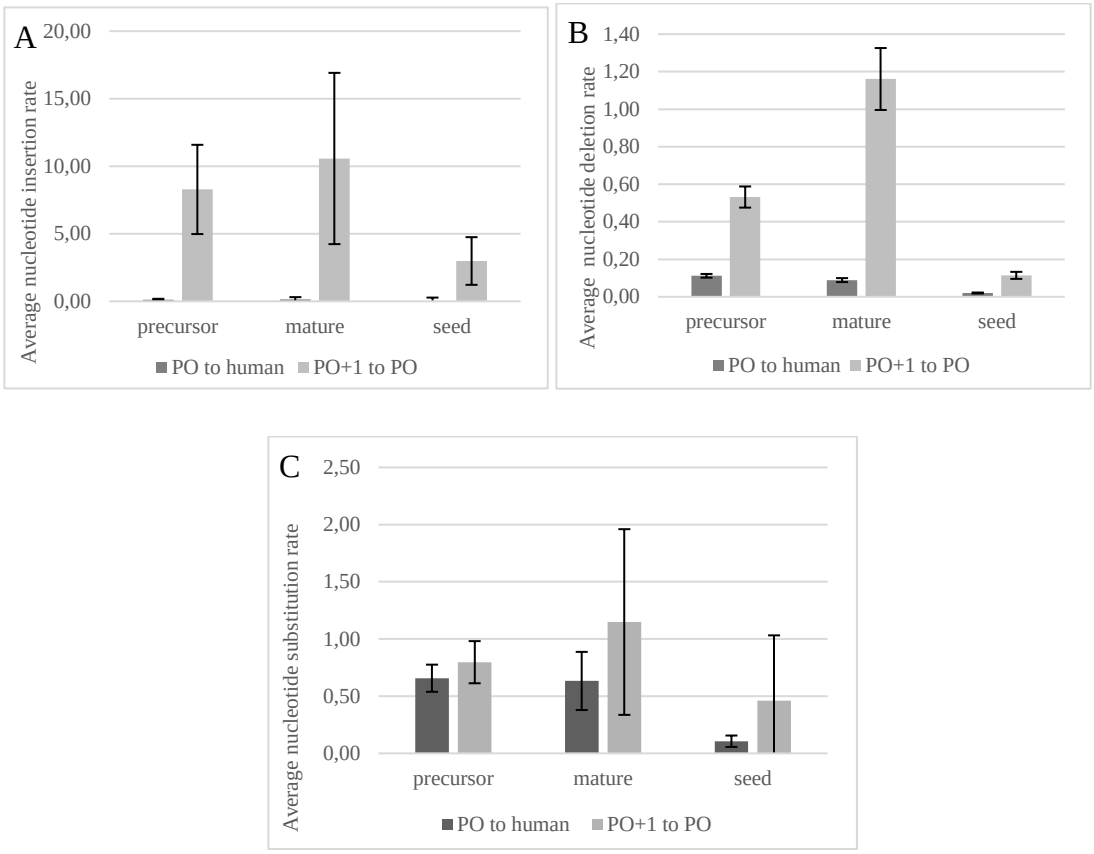


Figure III-3: Average nucleotides insertion (A), deletion (B) and substitution (C) rates between PO to human and PO+1 to PO in entire pre-miRNAs, mature miRNAs, and mature miRNAs seeds. They were calculated by dividing the number of mutations by branch length and sequence length.

3.4.3 Evolutionary mechanisms leading to new primate miRNA genes

Three main types of mechanisms of miRNAs creation are proposed (Figure III-4): (1) Full duplication: an existing functional pre-miRNA is copied to another region in the genome through a segmental or tandem duplication, or via a trans-duplication by a transposable element. (2) Insertions: a pre-miRNA-like hairpin is created by the insertion a portion of DNA that is partially complementary to a pre-existing nearby region; the inserted DNA may have its origin in cis (e.g. tandem or inverted tandem duplication), in trans (transposable element, retroposed gene, or segmental duplication), or may be of unknown origin. (3) Accumulation of local mutations (substitutions or short indels) that ends up forming a functional pre-miRNA. The

availability of accurate ancestral sequences at PO+1 and PO allows us to accurately quantify the rate at which each of these mechanisms contributed to the creation of new miRNAs. The decision tree used to determine the mechanism of creation of each miRNA (see Methods for details), together with the number of miRNAs found at each step of the decision process, are shown in Figure III-5.

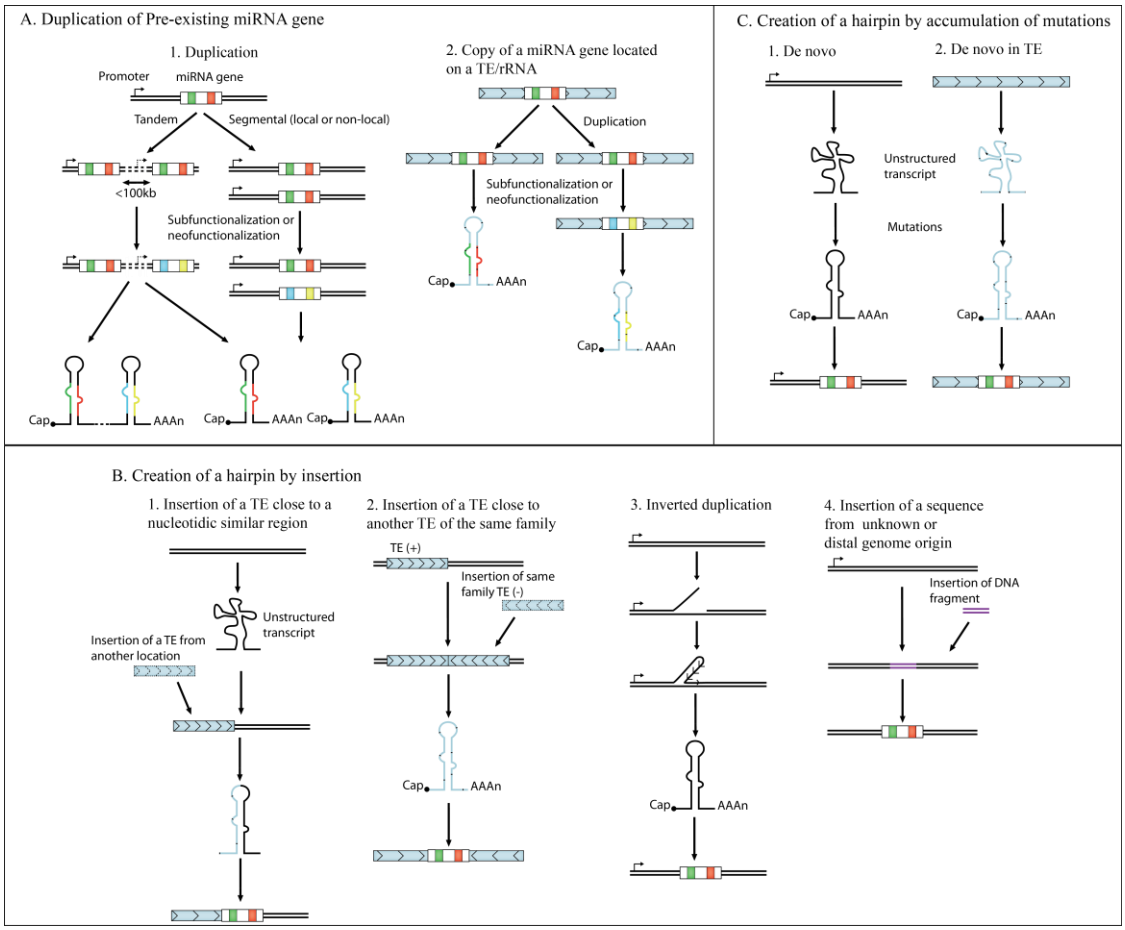


Figure III-4: Mechanisms leading to new miRNA genes **in primates. Figure inspired from Berezikov 2011. A. Duplication events. A.1:** miRNA genes **are created by duplication of an existing functional miRNA. The copy is located either on the same chromosome within 100 kb (tandem duplication) or elsewhere in the genome (segmental duplication). The copy eventually mutates to become a new miRNA gene having a new function (subfunctionalization or neofunctionalization). A.2: Transduplication event, where a functional miRNA gene is located on a TE (transposable element) or an rRNA and is duplicated. The new copy eventually evolves to a new miRNA gene. B. Insertion events. B.1: The insertion of a TE close to a similar region but inverted at nucleotide content level creates a new hairpin, leading to a new miRNA gene. B.2: A new miRNA gene is created with the insertion of a TE close to a copy of itself on the reverse strand, thereby creating a hairpin. B.3: An inverted duplication, by strand slippage or snapback DNA synthesis creates a perfect hairpin, leading to a new miRNA gene. B.4: An insertion of a sequence from unknown or distant location in the genome creates a hairpin. C. De novo events. C.1: De novo miRNA gene creation involves many mutations in a transcribed region leading slowly to a dicer-compatible hairpin. C.2: A miRNA gene is located in a TE, but was not functional at time of TE insertion, hence considered as *de novo*.**

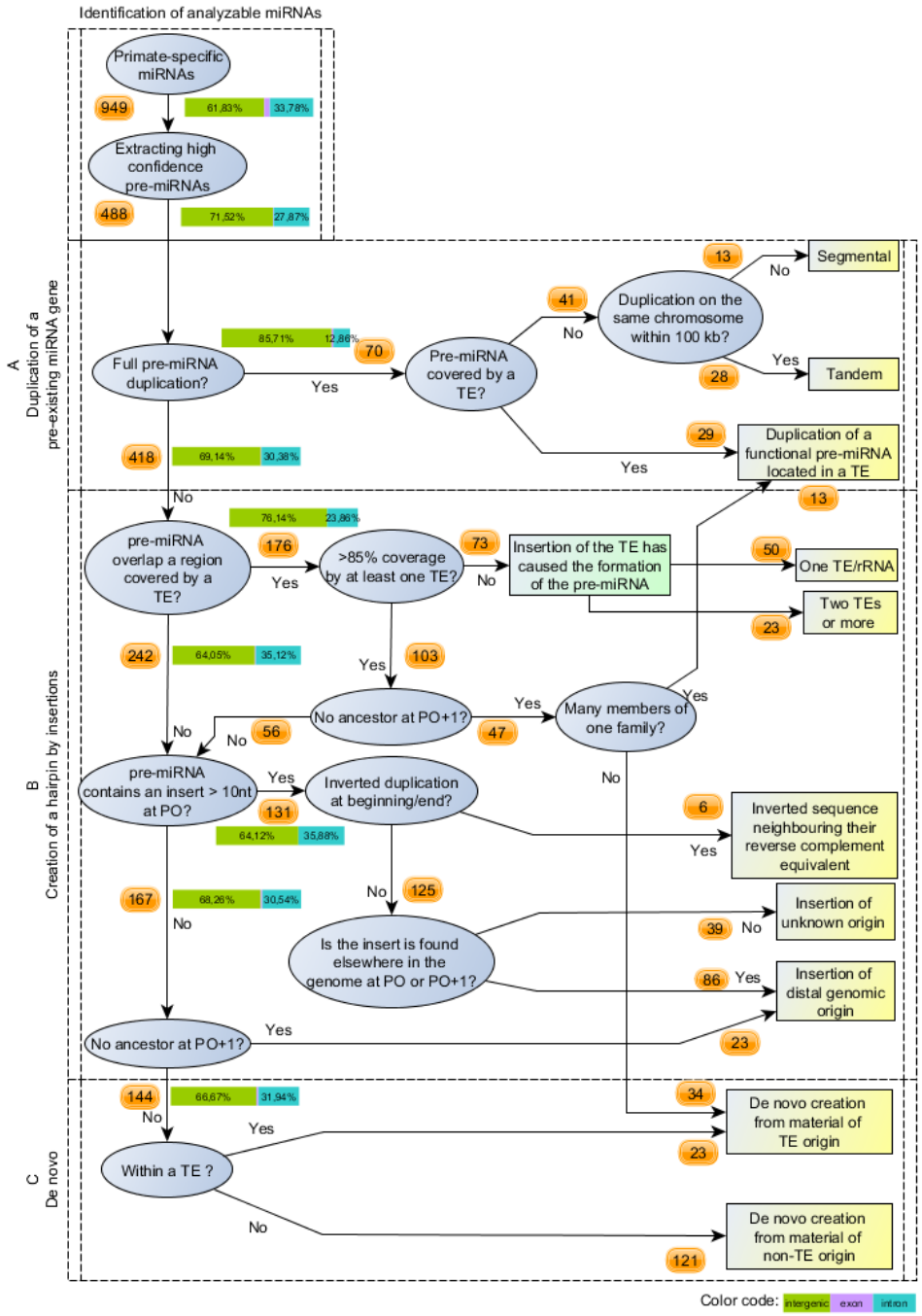


Figure III-5: Classification of human miRNA genes **by mechanisms of creation. Each number represents the number of pre-miRNAs involved at each decision node. MiRNAs involved in each mechanism of creation of this figure are listed in Table SD III-4. Details of the region type (intergenic, exonic and intronic) for each mechanism are listed in Table SD III-5.**

3.4.3.1 Duplication of pre-existing miRNAs

Duplication of a pre-existing miRNA gene, either through tandem or segmental duplication or TE-mediated trans-duplication, is the most direct route to the creation of a new miRNA. MicroRNA duplications are known to have been a major source of new miRNA genes prior to the mammalian radiation (Hertel et al. 2006). The duplicated miRNA gene eventually changes to get a new function (neofunctionalization) or retains aspects of the original function (subfunctionalization). Our results show that they are also important in more recent evolution (i.e. since the Primates ancestor), although perhaps less so than in earlier periods, having contributed to the creation of 17% (83) of the miRNA gene we analyzed (see examples in Figure III-6, full results in Table SD III-2), including 13 and 28 from segmental and tandem duplication resp., and 42 (29+13) duplicated by transposable elements, mostly involving DNA class type MADE1 family. Some miRNAs families show a large expansion because of duplication events in primates. For example, the mir-6511 family has seen its number of members grow from 0 to 6 between the primate ancestor and human. Another family mir-548, and in a lower proportion mir-3118, have also seen a large increase in number of members, but this time due to TE-mediated trans-duplication. Finally, we noted that 52 miRNA genes on 70 that had an origin from non-analyzable pool of miRNA genes.

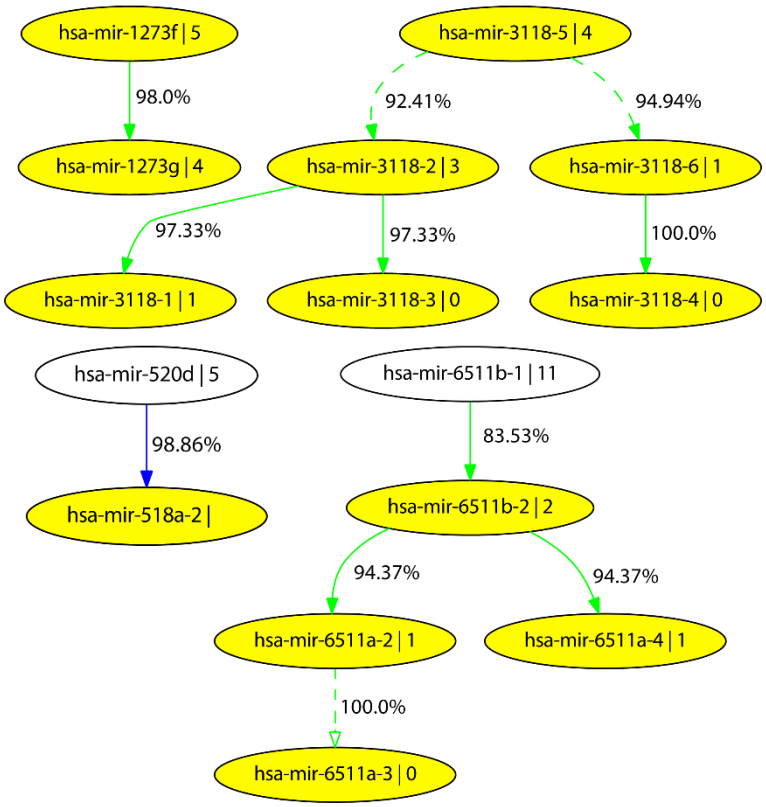


Figure III-6: Duplication event paths examples. Each bubble contains the miRNA gene and its associated PO. MiRNA genes **in yellow bubbles are present in our analyzable pool. Between two copies, an arrow, which represents the direction of the duplication, is labelled with the similarity score, the percentage identity, the shared sequence proportion, and the period when the duplication occurred. Green arrows represent a duplication occurring in a same miRNA gene family, blue arrows for a different family. Tandem duplications are represented by dashed arrows and segmental duplications by full arrows.**

3.4.3.2 Insertions leading to the creation of new pre-miRNAs

Here, we consider the 131 cases where newly inserted DNA combined (at least 10 bp long) with pre-existing DNA resulted in the creation of a new pre-miRNA. We break down our analysis based on the source of the new DNA insertion.

3.4.3.3 Insertions of transposable or repetitive elements

We consider first insertions of transposable element (TE) origin. We identified 73 miRNAs (see

Table SD III-3) whose sequence at their PO consisted up to 85% of newly inserted TE DNA. 50 involved only one TE, and 23 two TEs. A special case, mir-548h-3, has been created by the insertion of a TE MADE1 (DNA class, TcMar-Mariner family) in the middle of two other L3 TEs (LINE class, CR1 family). We observed that members of family hsa-mir-548 are mostly derived from MADE1 transposable elements, which are short miniature inverted-repeat transposable elements (Piriyapongsa and Jordan 2007b), but also a family associated with many functional roles, with high levels of nucleotide divergence and whose seeds show uneven evolutionary patterns (Liang et al. 2012). On the 72 annotated members of the hsa-mir-548 family, our pipeline classified 50 of them as being derived from TEs, including 52% identified as pre-existing miRNA genes duplicated by TE, and 20% created by the insertion of a TE.

The 103 remaining miRNA genes had no predicted ancestral sequence before their PO (No ancestor at PO+1), thus originating directly from duplication event of a TE. In this case, two scenarios are possible: either the TE was already carrying a functional miRNA gene before the copy, which should result in the presence of many human miRNAs members of one family (13 miRNA genes were in that case, classified as duplication events), either the region became a functional gene after the insertion of the TE (34 miRNA genes were in that case). In this last scenario, either the TE carried a functional hairpin and was inserted close to a promoter, either our resolution in ancestors, i.e. number of ancestral species, is too weak to determine that the miRNA gene existed in other ancestors after its PO. For miRNA genes whose section(s) of the sequence exists in other ancestors after the PO, we do not consider that the TE is the source of the miRNA gene creation (56 miRNA genes were in that case).

Finally, two miRNA genes overlapped the same rRNA 5S: miR-7641-1 (created at CHLCA, PO_1) and miR-7641-2 (hominida, PO_3). These sequences of the same family

have 83.6% similarity and are located on different chromosomes. Both have no existing ancestral sequence before their PO, implying that the rRNA has the same PO at these positions. We classified miR-7641-1 as a duplication event from miR-7641-2 despite both were overlapping the same rRNA. The reason is that about 200 copies of this rRNA exist in the human genome, and the miR-7641 family has only two members. Then, the most probable scenario is that miR-7641-2 is directly derived from an rRNA which have been duplicated with sequence content modifications after its copy, thus classified as De novo.

On the 120 miRNA genes (73+47) created by the insertion of a TE or rRNA, most were SINE, LINE and DNA types (25%, 31%, 25% resp.), created in a large proportion (61%) between Haplorrhini (PO_6) and Hominida (PO_3) periods (see Figure III-7).

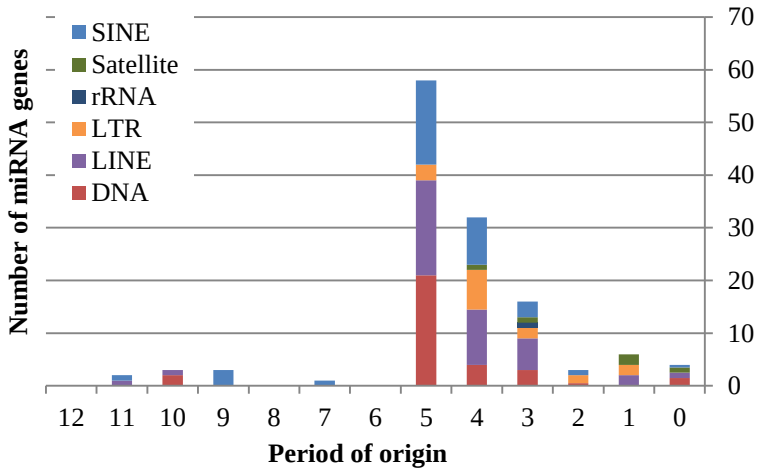


Figure III-7: Distribution of the number of miRNA genes by their period of origin created by the insertion of a repetitive/transposable element or rRNA.

3.4.3.4 Inverted duplications

In many cases, the DNA insertion that led to the creation of a miRNA could be linked to transposable elements, but were instead caused by other mechanisms. Inverted duplications, copy in tandem the reverse complement of a genomic region – a prime

mechanism for the creation of hairpins. This is often caused by snapback DNA, where a DNA can renature to form a hairpin structure after synthetization of its self-complementarity sequence (Lechner et al. 1983), or strand slippage (Petruska et al. 1998). While this mechanism has been observed for miRNAs in plants (Voinnet 2004; Fahlgren et al. 2007), surprisingly, we found 6 miRNAs whose origin can be traced back to an inverted duplication event, based on hairpin sequence analysis.

3.4.3.5 Short segmental duplications and insertions of unknown origin

131 miRNA-creating insertions could not be tied to transposable elements. Nevertheless, 33 are partially (ex: hsa-mir-5692c-2) or fully (ex: hsa-mir-606) covered by a TE, but the functionality of the miRNA were acquired because of another type of insertion. To understand their origin, we blasted the inserted sequences against the ancestral genomes at PO and PO+1. 86 returned one or more significant hits, sometimes on different chromosomes. The origin of the 39 insertions that are unaccounted for remains unclear, although most are suspected of coming from distal genomic regions, but they are quite short (30bp on average), which makes determining their origin difficult.

3.4.3.6 Full insertions of distal genomic origin

A total of 23 analyzable miRNA genes have their sequence almost perfectly conserved from their period of origin to humans. They were not detected as duplication from pre-existing miRNA genes, and we couldn't predict an ancestral sequence before PO. These genes do not overlap a TE, and Blast showed that they exist at multiple positions in the genome at PO, with highly significant hits. These miRNA genes sequences are probably passengers of other non-TE or poorly annotated TE DNA elements that were duplicated and inserted close to an existing promoter, or a promoter has been created downstream or upstream after their insertion. These two cases are not currently investigated by our pipeline, which does not take in account the possibility that a region lacking promoter can already have a nucleotide content that fold into a miRNA-like hairpin once transcribed.

3.4.3.7 De novo

The last type of mechanism we consider is one where a genomic region evolves into a functional pre-miRNA through a series of substitutions that turn a transcript that is not processed by Drosha into one that is. We identified 144 pre-miRNAs whose origin could not be ascribed to any of previously described mechanisms, including 23 covered by TE material, and for which the mutations between PO+1 and PO did not include large insertions. This process is made easier if the region in question is already transcribed for some other reason (e.g. it is in the intron of a protein-coding gene).

Berezikov et al. (2011) reported that after many mutations, it is possible that a tRNA or a snoRNA mutate to a hairpin shape and acquire miRNA-like features. Rare are the pre-miRNAs derived from these molecules. After aligning human pre-miRNAs and human tRNA/snoRNAs, only 7 had a similarity greater than 80%, all with snoRNA: miR-1248, miR-1291, miR-3651, miR-3653, miR-6516, miR-664a and miR-664b. But none of them belongs to a primate's period of origin. Only miR-4521 shares by three nucleotides a human tRNA (chr17.tRNA7-SerGCT), which is not enough to pretend that this miRNA is derived from this tRNA. Finally, we also submitted ancestral sequences of human miRNA genes to tRNA (ARAGORN (Laslett and Canback 2004)) and snoRNA predictors (SnoReport (Hertel et al. 2008)), but no positive results were returned.

3.4.4 Intragenic, intergenic, pseudogene

One may expect that genomic regions that already are transcribed for some reason (e.g. intronic region of protein-coding genes) may be more fertile grounds for the birth of new miRNAs, especially those that are not created by full duplication events, because this waives the requirement of developing a transcriptional regulatory mechanism. That is indeed the case. The proportion of miRNA gains by insertions or de novo events taking place within intronic regions is 2.4 times higher than among miRNA created by full duplications (Chi-square p-value is .00554). Nonetheless, nearly two thirds of miRNA creations by insertions or de novo occur outside of protein-coding

transcripts, which raise the question of how these regions gained at the same time the ability to be transcribed and that of being processed by Drosha.

We set three subcategories for every mechanism: origination events can be exonic, intronic or intergenic, determined by the location of the miRNA gene in RefSeq genes. Currently, about 62% of the 1868 human miRNA genes are located in intergenic regions, 34% in intronic regions, and the remaining 4% in exons. This proportion is relatively maintained in the 488 primate-specific miRNAs we analyzed (72% intergenic, 28% intronic). We kept track of these proportions along the pipeline, after the filter steps, and we observed that 85% of the miRNA genes that originated from duplication events are in intergenic regions, and only 13% in intronic regions. These proportions are quite different among miRNA genes that originated from insertions and De novo events, of which 75% and 72% resp. lie in intergenic regions while 25% and 27% resp. are intronic (See Table SD III-5). Exonic locations are rare except for miRNA genes originated from segmental duplications events. Since intragenic regions are less prone to undergo large insertions of genetic material because of positive selective pressure, this was expected. MiRNA genes created by the insertion of transposable elements are also generally located in intergenic regions (76%). The remaining 24% miRNA genes are located in intronic regions, which is, in proportion, much higher than the 4% of protein-coding regions of all human genes affected by TE insertions (Nekrutenko and Li 2001). We denoted no cases of insertion of TE that causes the creation of miRNA gene within exonic regions, which is consistent with previous studies who reported very low proportions of TEs in human CDS (Kapusta et al. 2013).

3.4.5 MiRNA functions by period of origin and mechanism of origination

We analyzed the gene ontology (biological processes) profiles of the experimentally validated target genes of miRNAs classified in each mechanism or origination and PO (Figure III-8), based on data from miRTarBase v6 (Hsu et al. 2011). We found a large heterogeneity of processes across our classification, but some interesting differences

were noted. For example, RNA metabolic process (GO:0016070), defined by the cellular chemical reactions and pathways involving RNA, is enriched among the targets of miRNAs coming from all mechanisms of origination. For periods of origin, RNA metabolic process is also represented in all PO from human to primates' ancestor, and is one of the rare significant process found enriched among the targets of miRNA that arose in human (PO_0), CHLCA (PO_1) and Hominidae (PO_2). A particularity of miRNAs that were created in Simiiformes (PO_5) and Haplorrhini (PO_6) compared to other PO is that they include miRNA target repertoire involved in chromatin organization (GO:0006325), neuron differentiation (GO:0030182) and phosphorylation (GO:0016310).

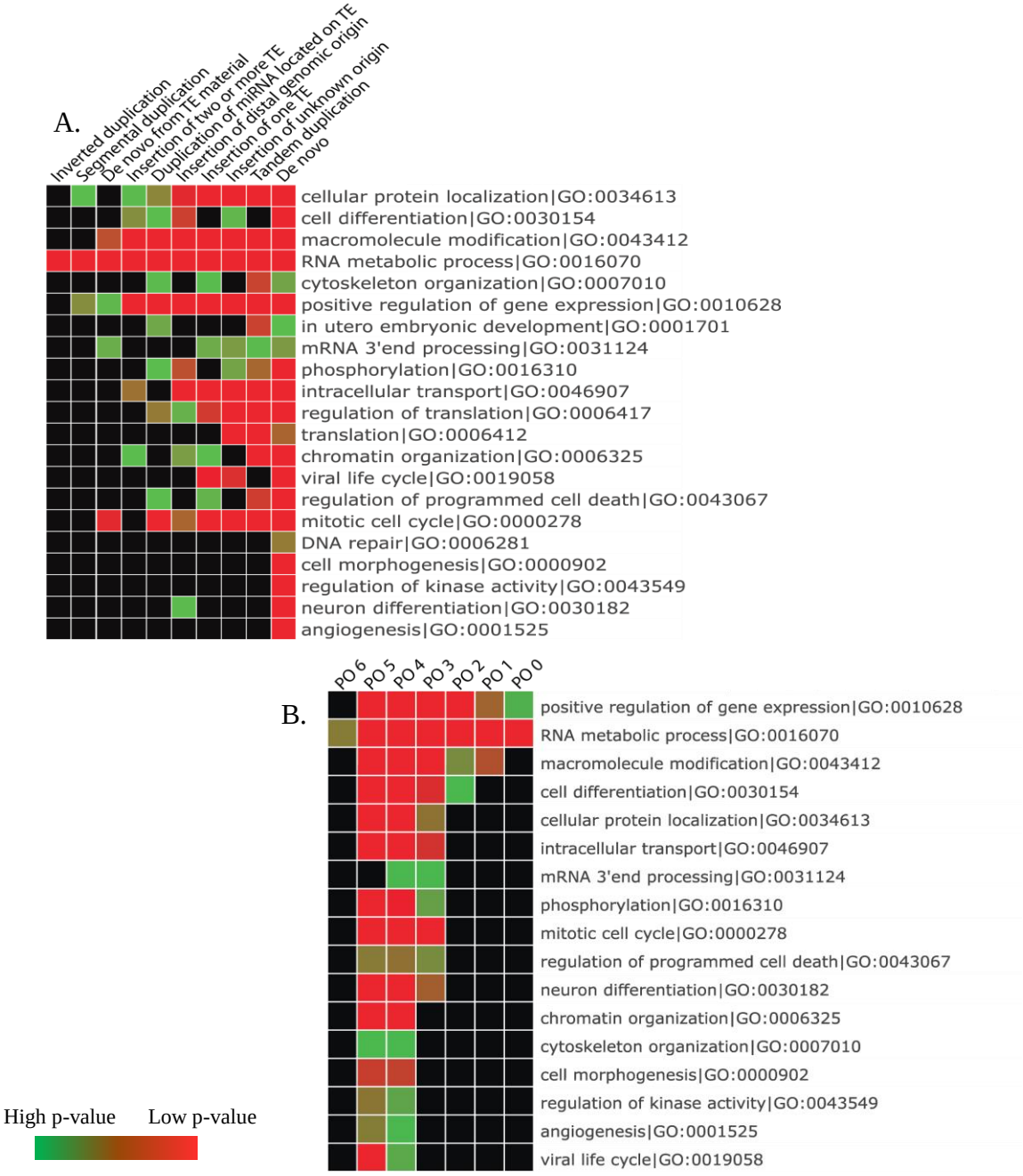


Figure III-8: Heatmap of gene ontologies' biological processes of miRNAs' experimentally validated target genes of miRTarBase v6 in (A) each mechanism of origination and (B) period of origin. P-values of ontologies (red is lowest, green highest, black is no target genes associated to given process) have been calculated with G:cocoa (Reimand et al. 2007) by comparing a non-redundant selection of enriched GO terms of

each dataset, and the heatmap has been created with GiTools v2.2 (Perez-Llamas and Lopez-Bigas 2011).

3.4.6 Comparison with other studies

Yuan et al. (Yuan et al. 2011) found 223 miRNAs originated from TEs based on strict parameters: TE-derived miRNAs are considered as such if the coverage of the repetitive element was at least 50% of the miRNA gene or 100% in one of the associated mature miRNA sequences. While this approach is reasonable to detect the overlap with a TE, it cannot tell if the insertion of the TE is intimately associated to the creation of the miRNA. On the 94 miRNA genes we have identified to have been created because of a TE element (duplicated by TE, insertion of a TE), we had 44 in common with Yuan et al.. The remaining 50 are disagreements, and although covered by a TE, according to ancestral sequences many of these miRNA genes were created after the insertion of another non-TE DNA fragment within a TE (e.g mir-3667, mir-3937). Others were classified in De novo from TE material (e.g. mir-3164, mir-588), because we believe that if a miRNA is localized in a TE they would have existed in many copies of their family members.

Piriyapongsa et al. (Piriyapongsa et al. 2007) found 55 TE-derived miRNAs. Of the 18 being in our set of analyzable miRNAs, we confirm 9 TE-derived. Other studies identified miRNAs originated from TEs, such as Smalheiser et al. (Smalheiser and Torvik 2005), but we couldn't compare our results since identified miRNAs are not in our set of analyzable miRNAs

3.5 Conclusions

Figure III-9 summarizes our classification of primate-specific miRNA genes based on their mechanism of origination. We estimated that 949 miRNA genes were created during the primate's evolution, and we were able to analyze 51% of them. Among the 488 analyzable miRNAs, three major mechanisms are identified: 30% of miRNA

genes were created *De novo* through random point mutations, 18% are the result of random insertion of DNA of non-transposable origin, and 15% are derived from transposable or repetitive elements that created the hairpin because of their insertion. We note that most of the origination events appeared between Haplorrhini ancestor to the Catarrhini ancestors (PO=5 & PO=4) periods, mostly by insertion events (Figure III-10). During these periods, 180 originated from an insertion event, 90 from duplication events, and 69 from *De novo* mechanism.

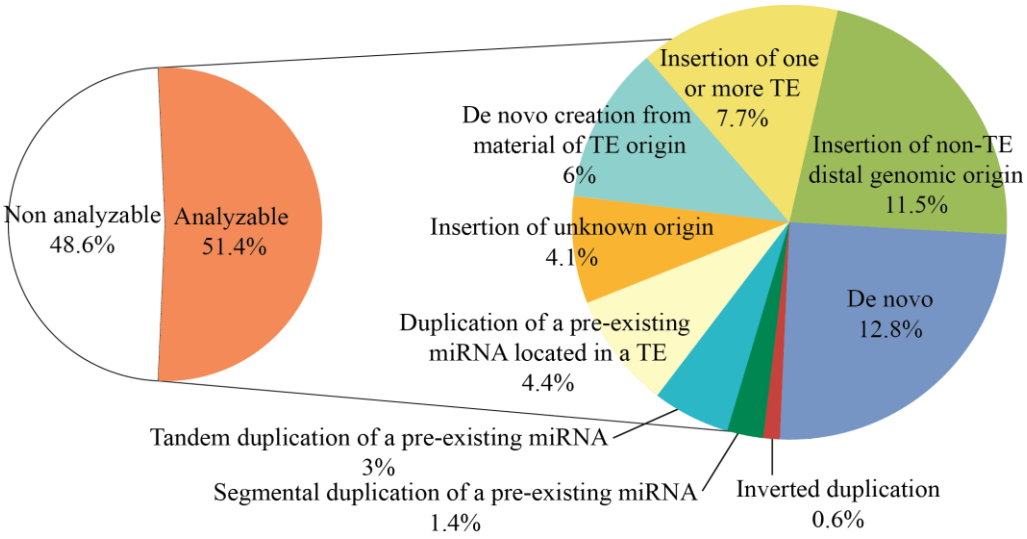


Figure III-9: Overall distribution of mechanisms leading to new primate-specific human pre-miRNAs.

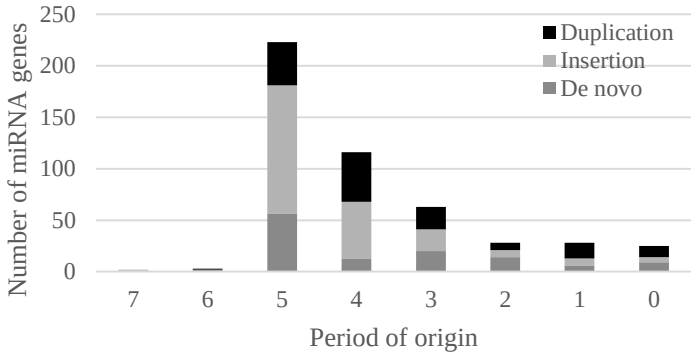


Figure III-10: Number of analyzable miRNA genes by period of origin from humans (0) to primate's ancestor (7) in the main categories of mechanisms, i.e. duplication, insertion, and De novo events.

In this study, we restricted our analyses to high confidence miRNAs to allow a more accurate estimation of the mechanisms of origination. These represent more than half of the primate's miRNAs, which provides a good perspective on the global distribution of all primates' miRNA's mechanisms of origination. Much remains unknown about the exact mechanisms of gains of miRNAs, and further investigation is needed to find the other factors that may be at play. Furthermore, compared to other studies that have characterized the mechanisms of origination of some miRNA genes, we identified many new miRNAs members in TE-originated and duplication events. We also found many some disagreements but we believe that ancestral reconstruction provides a much better precision to better identify how miRNA were created in the past. As future work, one may want to restart this study with alignments containing more species, which will improve the accuracy of ancestral reconstruction and increase the number of ancestors, so as the possible periods of origin. Also it could be interesting to analyze the regions around the new miRNA genes to find if they are the results of the promoter's presence or insertion.

3.6 Material and Methods

3.6.1 Datasets

Experimentally validated human miRNAs precursors and mature sequences coordinates were retrieved from MiRbase v20 (Griffiths-Jones et al. 2006), for a total of 1868 pre-miRNAs and 2575 mature miRNAs (many pre-miRNAs contain two mature miRNAs). Whole genome multiple alignments of 34 mammals (see Figure III-1A), referenced on human genome assembly GRCh37/hg19, was downloaded from UCSC genome browser (Kent et al. 2002; Schwartz et al. 2003; Blanchette et al. 2004b). The human complete protein-coding gene annotation used to classify genomic regions into exonic, intronic, or intergenic regions was obtained from the same source (UCSC known genes).

3.6.2 Ancestral reconstruction

Ancestral genomes were reconstructed with an improved local version of Ancestor (Diallo et al. 2010), which uses a maximum likelihood approach based on an evolutionary model that takes in account insertions, deletions and substitutions. The reconstruction is computed from whole-genome multiple alignments mentioned earlier and the phylogenetic tree from Murphy et al (Murphy et al. 2001). The result is an augmented multiple genome alignment, which includes both extant sequences and computationally inferred ancestral sequences. We defined 13 periods of origin for human pre-miRNAs (see Figure III-1A): modern human (PO_0), CHLCA (PO_1), Homininae (PO_2), Hominidae (PO_3), Catarrhini (PO_4), Simiiformes (PO_5), Haplorrhini (PO_6), Primates (PO_7), Euarchonta (PO_8), Euarchontoglires (PO_9), Boreoeutheria (PO_{10}), Placentalia (PO_{11}), and Theria (PO_{12}).

Pre-miRNAs' and mature miRNAs' ancestral sequences were retrieved from the augmented genome alignment based on the human's genomic coordinates from miRbase. Ancestral reconstruction provides a confidence score assigned to every reconstructed base (Blanchette et al. 2004a). This score varies between ancestors and

between genomic regions. Thus, in order to avoid bias observations due to low score reconstruction, we excluded reconstructed sequence having a confidence score below 90%, and all pre-miRNAs ancestral sequences containing more than 80% of gaps were discarded from our results, considering them as result from wrong alignments.

3.6.3 Inferring the period of origin of miRNAs

The period of origin of a human miRNA genes estimated by ancestral reconstruction (PO-AR) were determined as follows: First, all human miRNA genes' ancestral sequences were retrieved from ancestral reconstruction results, from human (PO_0) to the mammal common ancestor (PO_{12}). Each sequence was then submitted to (i) miPred (Jiang et al. 2007) to obtain a score describing the likelihood that the sequence would form a functional pre-miRNA, and (ii) miRdup (Leclercq et al. 2013) to determine whether the ancestral miRNA, based on homology to the mature human miRNA, is likely to be processed by dicer. Since selected pre-miRNAs are functional in human, we have prior information that increases our belief that their ancestors are also functional. Moreover, despite the fact that miPred seems to be a good tool to predict human pre-miRNAs considering its high reported accuracy (Hu et al. 2012), it rejects 42.8% of experimentally validated human pre-miRNAs from miRbase v20, including 7.3% it classifies as non-hairpins and 35.5% as pseudo-hairpins. Thus, miPred's distinction between real and pseudo-human pre-miRNAs was ignored, and both were considered positive predictions. To predict mature miRNAs in ancestral pre-miRNAs, we used miRdup (trained on mammalian miRNAs) which is the only existing software that tests the compatibility of a miRNA within a given hairpin.

To estimate the period of origin of a pre-miRNA based on least common ancestor method (PO-LCA), we first extracted extant sequences that are orthologous to the human pre-miRNA using the multiple alignment. We then tested each sequence with miPred and miRdup as described above. The PO-LCA was then obtained as the least common ancestor of all modern sequences where both the miPred and miRdup predictions were positive.

3.6.4 Classification of mechanisms of origination

Duplications: These events were identified by pairwise alignment of pre-miRNAs of lengths $L1$ and $L2$ at their PO using Stretcher (Rice et al. 2000). Only pre-miRNAs having a match with percent identity ($\frac{\#matchs}{\#matchs + \#mismatches} \times 100$) and a shared sequence ($\frac{\#matchs + \#mismatches}{\max(L1,L2)} \times 100$) greater than 80% were retained.

Tandem or segmental duplication were determined by comparing the location of duplicated pre-miRNAs. A duplication event is called tandem when both copies are within 100 kb. Segmental duplications are either local if both copies are on the same chromosome and distant of more than 100 kb, or non-local if located on different chromosomes.

To detect inverted duplications, we focused on pre-miRNAs having an insertion of more than 10 nt at PO, and rejected those whose first calculated ancestor was equal to PO. We define the first ancestor of a human sequence as the earliest ancestor species for which at least one nucleotide is predicted. Inserts localized at start or end of the pre-miRNAs were aligned to the reverse complement of the existing sequence before insertion, using Stretcher (Rice et al. 2000). If the percentage identity was greater than 80%, the insert was considered as an inverted duplication event. Otherwise, it was identified as an insertion from another source.

Source of duplicated miRNA genes: Although only miRNA genes from the analyzable pool are reported in the duplication events of Figure III-5, all the 1868 human miRNA genes were considered as a potential source of the duplications. For duplication events whose source originated from non-analyzable pool of miRNA genes, i.e. having uncertain PO, we stated the direction of the duplication using PO-LCA. Moreover, pre-miRNA B is called a duplication of a pre-miRNA A when the PO of A is greater or equal to that of B. When many sources existed for a copy, we kept the source having the highest percentage identity with its copy.

Insertions: Sources of insertions were identified by BLAST against the inferred ancestral genome at PO and PO+1, and against NCBI's genome databases.

Transposable elements: To identify pre-miRNAs derived from the transposable elements, we overlapped elements in UCSC table browser (Karolchik et al. 2004) between RepeatMasker (Smit et al. 1996) track and human pre-miRNAs.

3.6.5 Mutation rates

Mutation rate after (MRA, human to PO) and before (MRB, PO to PO+1), normalized by branch lengths and sequence lengths. A mutation can either be a substitution, insertion or deletion.

$$MRA = \frac{\left(\frac{\#mutations}{branchLength(human \rightarrow PO)}\right)}{sequenceLength}$$

$$MRB = \frac{\left(\frac{\#mutations}{branchLength(PO+1 \rightarrow PO)}\right)}{sequenceLength}$$

3.7 Funding

This work is funded in part by a fellowship to ML by the Fonds Québécois pour la Recherche sur la Nature et les Technologies, and by an NSERC Discovery grant to MB.

3.8 SUPPLEMENTARY DATA

Table SD III-1: Number of miRNA genes estimated for each period of origin between two methods, from 0 (human) to 12 (older than mammal's ancestor). A: x-axis =PO-AR, y-axis=PO-LCA. B: x =PO-AR, y=Iwama et al. C: x=PO-LCA, y=Iwama et al.

A. LCA\AR	0	1	2	3	4	5	6	7	8	9	10	11	12
0	52	2	1	5	2	1	1	0	2	0	0	0	0
1	0	13	8	3	1	3	0	0	0	0	1	0	0
2	2	0	8	9	3	5	0	0	0	0	1	0	0
3	1	0	0	21	13	14	2	3	1	5	0	1	0
4	3	0	3	4	58	67	11	14	3	11	9	6	1
5	2	0	1	0	2	114	6	20	6	13	7	6	2
6	0	0	1	1	1	4	2	1	1	1	0	1	1
7	1	0	0	0	0	4	1	4	0	2	1	7	1
8	0	0	0	0	0	0	0	0	3	1	0	0	0
9	6	1	1	1	2	4	1	2	1	2	4	3	2
10	3	1	5	0	9	16	1	4	5	14	19	32	10
11	1	0	3	2	1	7	3	4	2	7	11	194	48
12	1	0	1	0	0	1	1	1	0	2	0	5	183

B. Iwama\AR	0	1	2	3	4	5	6	7	8	9	10	11	12
0	11	2	2	10	9	7	3	3	3	0	1	6	0
1	3	4	1	2	2	3	0	1	0	1	1	0	0
2	9	4	5	1	7	8	0	3	2	0	1	4	0
3	8	2	7	18	41	62	3	12	3	17	9	10	2
4	4	1	5	4	14	54	8	12	2	11	5	11	8
5	6	1	0	5	9	61	6	10	9	11	10	8	4
6	1	0	2	0	1	3	1	2	0	0	0	6	1
7	1	0	0	0	0	0	0	0	1	0	0	4	1
8	0	0	0	1	0	0	0	1	0	1	1	1	0
9	0	0	1	1	1	3	0	0	0	0	3	3	3
10	10	1	5	1	6	23	4	5	2	8	14	40	19
11	14	1	2	3	2	9	4	3	1	8	6	140	33
12	5	1	2	0	0	7	0	1	1	1	2	22	177

C. Iwama\LCA	0	1	2	3	4	5	6	7	8	9	10	11	12
0	14	2	1	7	13	8	0	1	0	3	4	4	0
1	3	4	0	2	2	1	0	1	0	2	2	1	0
2	9	4	8	1	10	4	0	0	0	1	4	3	0
3	11	9	9	22	55	58	4	3	0	5	12	6	0
4	4	5	1	7	54	24	1	5	0	5	16	16	1
5	4	1	1	9	31	69	1	0	2	3	11	8	0
6	1	1	0	0	2	3	1	1	0	0	5	3	0
7	0	0	0	0	0	0	0	2	1	1	2	1	0
8	0	0	0	1	2	0	0	1	0	0	0	1	0
9	0	1	1	0	4	1	0	0	0	0	2	6	0
10	7	1	3	8	12	8	4	3	1	6	38	39	8
11	10	1	4	3	4	2	2	3	0	3	17	157	20
12	3	0	0	1	1	1	1	1	0	1	6	38	166

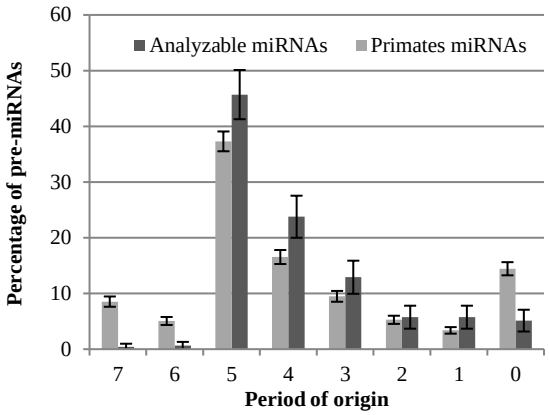


Figure SD III-1: Percentage of primates and analyzable miRNAs on a total of 949 and 488 miRNAs resp., having a PO-AR<=7.

Table SD III-2: Analyzable miRNAs originated from a duplication event.

Source	PO-LCA = PO-AR	PO-LCA of source	Copy	PO-LCA of copy	Percentage Identity	Same family between source and copy	Segmental duplication (else tandem)
hsa-mir-1184-1	x	10	hsa-mir-1184-2	4	100.0	x	x
hsa-mir-1283-2	x	5	hsa-mir-1283-1	4	100.0	x	x
hsa-mir-3118-2		3	hsa-mir-3118-1	1	97.33	x	x
hsa-mir-3118-2		3	hsa-mir-3118-3	0	97.33	x	x
hsa-mir-3118-5		4	hsa-mir-3118-2	3	92.41	x	
hsa-mir-3118-5		4	hsa-mir-3118-6	1	94.94	x	
hsa-mir-3118-6		1	hsa-mir-3118-4	0	100.0	x	x
hsa-mir-3156-2		5	hsa-mir-3156-3	4	97.47	x	
hsa-mir-3179-3		5	hsa-mir-3179-2	3	100.0	x	x
hsa-mir-3198-2	x	12	hsa-mir-3198-1	3	100.0	x	
hsa-mir-3689f	x	9	hsa-mir-3689a	2	82.28	x	x
hsa-mir-3690-2		2	hsa-mir-3690-1	0	100.0	x	
hsa-mir-4253	x	11	hsa-mir-4301	0	80.88		
hsa-mir-4444-1	x	11	hsa-mir-4444-2	0	100.0	x	
hsa-mir-512-2	x	7	hsa-mir-512-1	4	85.71	x	x
hsa-mir-515-2	x	5	hsa-mir-515-1	4	100.0	x	x
hsa-mir-516a-1	x	5	hsa-mir-516a-2	4	100.0	x	x
hsa-mir-516b-1	x	7	hsa-mir-516b-2	4	97.83	x	x
hsa-mir-516b-1	x	7	hsa-mir-518c	5	85.29		x
hsa-mir-516b-1	x	7	hsa-mir-519a-1	5	94.44		x
hsa-mir-516b-1	x	7	hsa-mir-519e	5	93.33		x
hsa-mir-516b-1	x	7	hsa-mir-520a	5	94.44		x
hsa-mir-516b-1	x	7	hsa-mir-520c	5	98.89		x
hsa-mir-516b-1	x	7	hsa-mir-523	5	97.78		x
hsa-mir-516b-1	x	7	hsa-mir-526b	5	92.22		x
hsa-mir-517c	x	5	hsa-mir-517a	4	86.73	x	x
hsa-mir-518a-1	x	5	hsa-mir-524	4	100.0		x
hsa-mir-518e	x	5	hsa-mir-522	4	100.0		x
hsa-mir-519a-1		5	hsa-mir-517b	4	81.18		x
hsa-mir-519b		5	hsa-mir-519c	4	93.1	x	x
hsa-mir-519e		5	hsa-mir-520h	4	92.22		x

hsa-mir-520c		5	hsa-mir-520f	4	100.0	x	x
hsa-mir-520d	x	5	hsa-mir-518a-2	4	98.86		x
hsa-mir-520e	x	7	hsa-mir-519a-2	5	100.0		x
hsa-mir-520e	x	7	hsa-mir-519b	5	93.1		x
hsa-mir-520e	x	7	hsa-mir-527	5	98.86		x
hsa-mir-520g	x	5	hsa-mir-519d	4	98.9		x
hsa-mir-521-1	x	5	hsa-mir-521-2	4	100.0	x	x
hsa-mir-548ad		5	hsa-mir-548f-5	4	91.86	x	
hsa-mir-548ae-2		4	hsa-mir-548ba	3	83.58	x	
hsa-mir-548an		5	hsa-mir-548d-2	1	85.57	x	
hsa-mir-548ax		5	hsa-mir-548o-2	4	87.67	x	
hsa-mir-548f-1		5	hsa-mir-548e	4	82.14	x	x
hsa-mir-548g		4	hsa-mir-548f-3	3	91.01	x	
hsa-mir-548h-2		3	hsa-mir-548aa-2	1	90.72	x	
hsa-mir-548h-4		5	hsa-mir-548ay	4	84.11	x	
hsa-mir-548h-4		5	hsa-mir-548h-2	3	87.13	x	
hsa-mir-548h-4		5	hsa-mir-548j	4	81.42	x	
hsa-mir-548l	x	5	hsa-mir-548ah	4	84.88	x	
hsa-mir-548n	x	5	hsa-mir-548g	4	80.9	x	
hsa-mir-548n	x	5	hsa-mir-548w	4	94.59	x	
hsa-mir-548o-2		4	hsa-mir-548ar	1	81.43	x	
hsa-mir-548u		5	hsa-mir-548ae-2	4	80.49	x	
hsa-mir-548v		5	hsa-mir-548a-3	4	82.47	x	x
hsa-mir-548v		5	hsa-mir-548am	3	82.5	x	
hsa-mir-548v		5	hsa-mir-548ap	4	81.63	x	
hsa-mir-548v		5	hsa-mir-548x-2	3	80.0	x	
hsa-mir-550a-1	x	5	hsa-mir-550a-3	1	97.94	x	x
hsa-mir-5692a-2		3	hsa-mir-5692a-1	1	85.51	x	
hsa-mir-570	x	3	hsa-mir-548al	2	83.51		
hsa-mir-5701-2		4	hsa-mir-5701-1	3	100.0	x	x
hsa-mir-620		3	hsa-mir-3669	0	80.0		
hsa-mir-6511a-2		1	hsa-mir-6511a-3	0	100.0	x	x
hsa-mir-6511b-1	x	11	hsa-mir-6511b-2	2	83.53	x	x
hsa-mir-6511b-2		2	hsa-mir-6511a-2	1	94.37	x	x
hsa-mir-6511b-2		2	hsa-mir-6511a-4	1	94.37	x	x
hsa-mir-6770-1	x	10	hsa-mir-6770-3	1	100.0	x	x
hsa-mir-7641-2		3	hsa-mir-7641-1	1	86.89	x	
hsa-mir-941-2		3	hsa-mir-941-4	1	100.0	x	x

Table SD III-3: MiRNA genes created by the insertion of one of more transposable elements (TE) or duplicated by a TE. Name, class and family are provided by RepeatMasker. PO is the period of origin based on PO-AR. Percentage of shared bases between the TE and the pre-miRNA are is in terms of pre-miRNA length.

Pre-miRNA	TE Name	TE class	TE family	PO	Percentage of shared bases between TE and pre-miRNA	Insertion of one of more TE	Duplication of a pre-existing miRNA located in TE
hsa-mir-1266	MIR3	SINE	MIR	5	80%	x	
hsa-mir-1273g	AluJb	SINE	Alu	4	100%		x
hsa-mir-1304	AluJo	SINE	Alu	5	73%	x	
hsa-mir-1972-1	AluSx	SINE	Alu	4	35%	x	
hsa-mir-1972-1	FLAM_A	SINE	Alu	4	34%	x	
hsa-mir-1972-2	FLAM_A	SINE	Alu	4	34%	x	
hsa-mir-1972-2	AluSx	SINE	Alu	4	35%	x	
hsa-mir-3118-1	L1PA13	LINE	L1	1	100%		x
hsa-mir-3118-2	L1PA13	LINE	L1	3	100%		x

hsa-mir-3118-3	L1PA13	LINE	L1	0	100%		x
hsa-mir-3118-4	L1PA13	LINE	L1	0	100%		x
hsa-mir-3118-6	L1PA13	LINE	L1	1	100%		x
hsa-mir-3137	Tigger3b	DNA	TcMar-Tigger	5	68%	x	
hsa-mir-3137	Tigger3c	DNA	TcMar-Tigger	5	55%	x	
hsa-mir-3149	L1ME3G	LINE	L1	5	65%	x	
hsa-mir-3166	L2a	LINE	L2	1	79%	x	
hsa-mir-3166	L2a	LINE	L2	1	34%	x	
hsa-mir-3169	MIRb	SINE	MIR	5	42%	x	
hsa-mir-3179-2	AluJo	SINE	Alu	3	1%		x
hsa-mir-3179-3	AluJo	SINE	Alu	5	1%	x	
hsa-mir-3622b	AluJo	SINE	Alu	5	3%	x	
hsa-mir-3646	MIR	SINE	MIR	5	1%	x	
hsa-mir-3664	MER46C	DNA	TcMar-Tigger	5	26%	x	
hsa-mir-3670-1	LTR16A1	LTR	ERVL	1	80%	x	
hsa-mir-3670-2	LTR16A1	LTR	ERVL	1	80%	x	
hsa-mir-3680-1	MER96	DNA	hAT-Tip100	3	100%		x
hsa-mir-3680-2	MER96	DNA	hAT-Tip100	3	100%		x
hsa-mir-378d-1	MIRb	SINE	MIR	4	44%	x	
hsa-mir-378d-2	MIRc	SINE	MIR	5	68%	x	
hsa-mir-3908	FLAM_A	SINE	Alu	3	40%	x	
hsa-mir-3908	AluSx	SINE	Alu	3	37%	x	
hsa-mir-3912	MER39	LTR	ERV1	4	9%	x	
hsa-mir-3912	L1ME3G	LINE	L1	4	41%	x	
hsa-mir-3919	L1M6	LINE	L1	3	52%	x	
hsa-mir-3920	L2a	LINE	L2	5	35%	x	
hsa-mir-4317	MIR	SINE	MIR	5	3%	x	
hsa-mir-4424	L1MA9	LINE	L1	5	43%	x	
hsa-mir-4457	L1MEc	LINE	L1	5	59%	x	
hsa-mir-4472-2	AluSz6	SINE	Alu	3	73%	x	
hsa-mir-4477a	Tigger1	DNA	TcMar-Tigger	4	100%		x
hsa-mir-4477b	Tigger1	DNA	TcMar-Tigger	4	100%		x
hsa-mir-4480	MIRb	SINE	MIR	3	70%	x	
hsa-mir-4484	MER50-int	LTR	ERV1	4	39%	x	
hsa-mir-4491	MER81	DNA	hAT-Blackjack	4	15%	x	
hsa-mir-4495	MIRc	SINE	MIR	5	52%	x	
hsa-mir-4495	MIRb	SINE	MIR	5	29%	x	
hsa-mir-4504	L1M2	LINE	L1	5	47%	x	
hsa-mir-4504	L1MA8	LINE	L1	5	38%	x	
hsa-mir-4508	MIR	SINE	MIR	5	24%	x	
hsa-mir-4512	AluSz	SINE	Alu	4	45%	x	
hsa-mir-4512	AluJb	SINE	Alu	4	51%	x	
hsa-mir-4518	MER117	DNA	hAT-Charlie	0	49%	x	
hsa-mir-4520a	MIRb	SINE	MIR	5	11%	x	
hsa-mir-4520b	MIRb	SINE	MIR	5	2%	x	
hsa-mir-4656	MIRc	SINE	MIR	5	19%	x	
hsa-mir-466	L1ME3	LINE	L1	3	11%	x	
hsa-mir-466	(CATAn	Simple_repeat	null	3	57%	x	
hsa-mir-4684	Charlie8	DNA	hAT-Charlie	5	9%	x	
hsa-mir-4703	Tigger18a	DNA	TcMar-Tigger	5	48%	x	
hsa-mir-4753	HAL1	LINE	L1	5	18%	x	
hsa-mir-4781	L3	LINE	CR1	5	7%	x	
hsa-mir-4797	AmnSINE1	SINE	Deu	4	41%	x	
hsa-mir-4797	AmnSINE1	SINE	Deu	4	42%	x	
hsa-mir-4999	L1MA9	LINE	L1	5	13%	x	
hsa-mir-5007	MSTA	LTR	ERVL-MaLR	4	2%	x	
hsa-mir-5011	Tigger3a	DNA	TcMar-Tigger	2	9%	x	
hsa-mir-5011	MER66C	LTR	ERV1	2	2%	x	
hsa-mir-5095	Charlie1a	DNA	hAT-Charlie	0	63%	x	
hsa-mir-5095	AluSq2	SINE	Alu	0	38%	x	
hsa-mir-548a-3	MLT1G1	LTR	ERVL-MaLR	4	5%		x
hsa-mir-548a-3	MADE1	DNA	TcMar-Mariner	4	81%		x

hsa-mir-548a-3	MLT1G1	LTR	ERVL-MaLR	4	12%		x
hsa-mir-548aa-1	MADE1	DNA	TcMar-Mariner	5	81%	x	
hsa-mir-548aa-2	MADE1	DNA	TcMar-Mariner	1	81%		x
hsa-mir-548ae-2	MADE1	DNA	TcMar-Mariner	4	100%		x
hsa-mir-548ag-1	MADE1	DNA	TcMar-Mariner	5	100%		x
hsa-mir-548ah	MADE1	DNA	TcMar-Mariner	4	100%		x
hsa-mir-548al	MADE1	DNA	TcMar-Mariner	2	95%		x
hsa-mir-548am	MADE1	DNA	TcMar-Mariner	3	97%		x
hsa-mir-548ap	MER50	LTR	ERV1	4	100%		x
hsa-mir-548ar	MADE1	DNA	TcMar-Mariner	1	100%		x
hsa-mir-548av	MADE1	DNA	TcMar-Mariner	5	100%		x
hsa-mir-548ay	LTR43	LTR	ERV1	4	100%		x
hsa-mir-548b	MADE1	DNA	TcMar-Mariner	5	81%	x	
hsa-mir-548ba	MADE1	DNA	TcMar-Mariner	3	100%		x
hsa-mir-548c	MADE1	DNA	TcMar-Mariner	5	81%	x	
hsa-mir-548d-1	MADE1	DNA	TcMar-Mariner	5	81%	x	
hsa-mir-548d-2	MADE1	DNA	TcMar-Mariner	1	81%		x
hsa-mir-548e	L1M5	LINE	L1	4	3%		x
hsa-mir-548e	MADE1	DNA	TcMar-Mariner	4	88%		x
hsa-mir-548e	L1M5	LINE	L1	4	8%		x
hsa-mir-548f-2	MADE1	DNA	TcMar-Mariner	5	81%	x	
hsa-mir-548f-3	L1M3	LINE	L1	3	3%		x
hsa-mir-548f-3	MADE1	DNA	TcMar-Mariner	3	94%		x
hsa-mir-548f-3	L1M3	LINE	L1	3	1%		x
hsa-mir-548f-4	L1MEd	LINE	L1	5	13%	x	
hsa-mir-548f-4	MADE1	DNA	TcMar-Mariner	5	70%	x	
hsa-mir-548f-5	MADE1	DNA	TcMar-Mariner	4	92%		x
hsa-mir-548g	MADE1	DNA	TcMar-Mariner	4	87%		x
hsa-mir-548h-1	Charlie1a	DNA	hAT-Charlie	5	14%	x	
hsa-mir-548h-1	MADE1	DNA	TcMar-Mariner	5	77%	x	
hsa-mir-548h-2	L1MB3	LINE	L1	3	6%		x
hsa-mir-548h-2	MADE1	DNA	TcMar-Mariner	3	90%		x
hsa-mir-548h-3	L3	LINE	CR1	5	19%	x	
hsa-mir-548h-3	MADE1	DNA	TcMar-Mariner	5	67%	x	
hsa-mir-548h-3	L3	LINE	CR1	5	14%	x	
hsa-mir-548h-4	MADE1	DNA	TcMar-Mariner	5	87%		x
hsa-mir-548h-5	MADE1	DNA	TcMar-Mariner	5	100%		x
hsa-mir-548j	MADE1	DNA	TcMar-Mariner	4	65%		x
hsa-mir-548k	MADE1	DNA	TcMar-Mariner	4	67%	x	
hsa-mir-548m	MADE1	DNA	TcMar-Mariner	5	90%		x
hsa-mir-548m	L1M5	LINE	L1	5	3%		x
hsa-mir-548o-2	MADE1	DNA	TcMar-Mariner	4	100%		x
hsa-mir-548u	MADE1	DNA	TcMar-Mariner	5	96%		x
hsa-mir-548w	MADE1	DNA	TcMar-Mariner	4	100%		x
hsa-mir-548x-2	MamGypLTR1a	LTR	Gypsy	3	10%		x
hsa-mir-548x-2	MADE1	DNA	TcMar-Mariner	3	79%		x
hsa-mir-548x-2	MamGypLTR1a	LTR	Gypsy	3	10%		x
hsa-mir-548z	MADE1	DNA	TcMar-Mariner	5	81%	x	
hsa-mir-549a	MIR	SINE	MIR	5	24%	x	
hsa-mir-553	MIR3	SINE	MIR	2	18%	x	
hsa-mir-553	MIR3	SINE	MIR	2	3%	x	
hsa-mir-5591	L1PB3	LINE	L1	5	35%	x	
hsa-mir-5591	L1PB3	LINE	L1	5	52%	x	
hsa-mir-5681a	MIRc	SINE	MIR	5	4%	x	
hsa-mir-5684	AluJr	SINE	Alu	4	48%	x	
hsa-mir-5684	AluJo	SINE	Alu	4	51%	x	
hsa-mir-5692a-2	SATR2	Satellite	null	3	100%		x
hsa-mir-5692b	SATR1	Satellite	null	0	100%		x
hsa-mir-5692b	SATR2	Satellite	null	0	37%		x
hsa-mir-5692c-1	SATR2	Satellite	null	1	100%		x
hsa-mir-5697	HAL1	LINE	L1	4	68%	x	
hsa-mir-5697	HAL1	LINE	L1	4	41%	x	

hsa-mir-5701-1	REP522	Satellite	telo	3	100%		x
hsa-mir-5708	AluJr	SINE	Alu	5	34%	x	
hsa-mir-5708	AluSx	SINE	Alu	5	34%	x	
hsa-mir-571	L1MA9	LINE	L1	4	32%	x	
hsa-mir-571	L1MA9	LINE	L1	4	64%	x	
hsa-mir-585	MLT1C	LTR	ERVL-MaLR	2	31%	x	
hsa-mir-587	MER115	DNA	hAT-Tip100	3	66%	x	
hsa-mir-607	MIR	SINE	MIR	5	55%	x	
hsa-mir-607	MIR	SINE	MIR	5	67%	x	
hsa-mir-634	L1ME3A	LINE	L1	3	47%	x	
hsa-mir-637	L1MC4a	LINE	L1	4	38%	x	
hsa-mir-644a	L1MB3	LINE	L1	4	62%	x	
hsa-mir-6839	LTR7C	LTR	ERV1	4	55%	x	
hsa-mir-7641-1	5S	rRNA	null	1	97%		x
hsa-mir-7849	MLT1L	LTR	ERVL-MaLR	5	17%	x	
hsa-mir-8076	MIRc	SINE	MIR	4	14%	x	
hsa-mir-8084	L1ME3Cz	LINE	L1	3	70%	x	

Table SD III-4: Mechanisms and period of origin of the 488 analyzable MiRNAs genes. Their number in each category is associated with the pipeline in Figure III-5.

Mechanism of origination	**Number of** miRNA genes	MiRNA genes **and period of origin**							
Segmental duplication of a pre-existing miRNA	13	hsa-mir-1184-2	4	hsa-mir-3690-1	0	hsa-mir-5692a-1	1	hsa-mir-6770-3	1
		hsa-mir-3156-3	4	hsa-mir-4301	0	hsa-mir-6511a-2	1		
		sa-mir-3198-1	3	hsa-mir-4444-2	0	hsa-mir-6511a-4	1		
		hsa-mir-3669	0	hsa-mir-550a-3	1	hsa-mir-6511b-2	2		
Tandem duplication of a pre-existing miRNA	28	hsa-mir-1283-1	4	hsa-mir-517b	4	hsa-mir-519d	4	hsa-mir-522	4
		hsa-mir-3689a	2	hsa-mir-518a-2	4	hsa-mir-519e	5	hsa-mir-523	5
		hsa-mir-512-1	4	hsa-mir-518c	5	hsa-mir-520a	5	hsa-mir-524	4
		hsa-mir-515-1	4	hsa-mir-519a-1	5	hsa-mir-520c	5	hsa-mir-526b	5
		hsa-mir-516a-2	4	hsa-mir-519a-2	5	hsa-mir-520f	4	hsa-mir-527	5
		hsa-mir-516b-2	4	hsa-mir-519b	5	hsa-mir-520h	4	hsa-mir-6511a-3	0
		hsa-mir-517a	4	hsa-mir-519c	4	hsa-mir-521-2	4	hsa-mir-941-4	1
Duplication of a pre-existing miRNA located in a TE	42 (29+13)	hsa-mir-1273g	4	hsa-mir-548al	2	hsa-mir-548h-2	3	hsa-mir-548ag-1	5
		hsa-mir-3118-1	1	hsa-mir-548am	3	hsa-mir-548j	4	hsa-mir-548av	5
		hsa-mir-3118-2	3	hsa-mir-548ap	4	hsa-mir-548o-2	4	hsa-mir-548h-4	5
		hsa-mir-3118-3	0	hsa-mir-548ar	1	hsa-mir-548w	4	hsa-mir-548h-5	5
		hsa-mir-3118-4	0	hsa-mir-548ay	4	hsa-mir-548x-2	3	hsa-mir-548m	5
		hsa-mir-3118-6	1	hsa-mir-548ba	3	hsa-mir-5701-1	3	hsa-mir-548u	5
		hsa-mir-3179-2	3	hsa-mir-548d-2	1	hsa-mir-7641-1	1	hsa-mir-5692a-2	3
		hsa-mir-548a-3	4	hsa-mir-548e	4	hsa-mir-3680-1	3	hsa-mir-5692b	0
		hsa-mir-548aa-2	1	hsa-mir-548f-3	3	hsa-mir-3680-2	3	hsa-mir-5692c-1	1
		hsa-mir-548ae-2	4	hsa-mir-548f-5	4	hsa-mir-4477a	4		
		hsa-mir-548ah	4	hsa-mir-548g	4	hsa-mir-4477b	4		
Insertion of one TE	50	hsa-mir-1266	5	hsa-mir-3920	5	hsa-mir-4684	5	hsa-mir-549a	5
		hsa-mir-1304	5	hsa-mir-4317	5	hsa-mir-4703	5	hsa-mir-5681a	5
		hsa-mir-3149	5	hsa-mir-4424	5	hsa-mir-4753	5	hsa-mir-585	2
		hsa-mir-3169	5	hsa-mir-4457	5	hsa-mir-4781	5	hsa-mir-587	3
		hsa-mir-3179-3	5	hsa-mir-4472-2	3	hsa-mir-4999	5	hsa-mir-634	3
		hsa-mir-3622b	5	hsa-mir-4480	3	hsa-mir-5007	4	hsa-mir-637	4
		hsa-mir-3646	5	hsa-mir-4484	4	hsa-mir-548aa-1	5	hsa-mir-644a	4
		hsa-mir-3664	5	hsa-mir-4491	4	hsa-mir-548b	5	hsa-mir-6839	4
		hsa-mir-3670-1	1	hsa-mir-4508	5	hsa-mir-548c	5	hsa-mir-7849	5
		hsa-mir-3670-2	1	hsa-mir-4518	0	hsa-mir-548d-1	5	hsa-mir-8076	4

		hsa-mir-378d-1	4	hsa-mir-4520a	5	hsa-mir-548f-2	5	hsa-mir-8084	3
		hsa-mir-378d-2	5	hsa-mir-4520b	5	hsa-mir-548k	4		
		hsa-mir-3919	3	hsa-mir-4656	5	hsa-mir-548z	5		
Insertion of two or more TE	23	hsa-mir-1972-1	4	hsa-mir-4495	5	hsa-mir-5095	0	hsa-mir-5697	4
		hsa-mir-1972-2	4	hsa-mir-4504	5	hsa-mir-548f-4	5	hsa-mir-5708	5
		hsa-mir-3137	5	hsa-mir-4512	4	hsa-mir-548h-1	5	hsa-mir-571	4
		hsa-mir-3166	1	hsa-mir-466	3	hsa-mir-553	2	hsa-mir-607	5
		hsa-mir-3908	3	hsa-mir-4797	4	hsa-mir-5591	5	hsa-mir-548h-3	5
		hsa-mir-3912	4	hsa-mir-5011	2	hsa-mir-5684	4		
Inverted duplication	6	hsa-mir-3622a	5	hsa-mir-4643	5	hsa-mir-4755	2		
		hsa-mir-4436b-2	4	hsa-mir-4714	5	hsa-mir-548ac	5		
Insertion of unknown origin	39	hsa-mir-1267	5	hsa-mir-3942	5	hsa-mir-4752	4	hsa-mir-627	5
		hsa-mir-3129	4	hsa-mir-4440	3	hsa-mir-4771-2	5	hsa-mir-642a	5
		hsa-mir-3148	5	hsa-mir-4490	5	hsa-mir-4780	5	hsa-mir-642b	5
		hsa-mir-3150a	4	hsa-mir-4493	7	hsa-mir-5000	5	hsa-mir-6512	5
		hsa-mir-3202-1	5	hsa-mir-4642	5	hsa-mir-5680	5	hsa-mir-6718	5
		hsa-mir-3910-2	5	hsa-mir-4704	5	hsa-mir-5685	2	hsa-mir-6888	3
		hsa-mir-3913-1	4	hsa-mir-4719	5	hsa-mir-5696	5	hsa-mir-7702	5
		hsa-mir-3913-2	4	hsa-mir-4720	5	hsa-mir-5700	5	hsa-mir-7850	4
		hsa-mir-3922	4	hsa-mir-4727	3	hsa-mir-576	5	hsa-mir-944	5
		hsa-mir-3926-1	4	hsa-mir-4744	5	hsa-mir-610	5		
Insertion of non-TE distal genomic origin	109 (86+23)	hsa-mir-1265	5	hsa-mir-3938	5	hsa-mir-4803	5	hsa-mir-934	5
		hsa-mir-1269a	5	hsa-mir-4436a	4	hsa-mir-4804	5	hsa-mir-941-2	3
		hsa-mir-1270-1	4	hsa-mir-4439	4	hsa-mir-5008	5	hsa-mir-1324	1
		hsa-mir-1273d	5	hsa-mir-4465	4	hsa-mir-5087	2	hsa-mir-1827	5
		hsa-mir-1273f	5	hsa-mir-4471	5	hsa-mir-5191	5	hsa-mir-3121	3
		hsa-mir-1273h	4	hsa-mir-4474	4	hsa-mir-548ad	5	hsa-mir-3180-1	4
		hsa-mir-1285-1	5	hsa-mir-4524a	1	hsa-mir-548ai	5	hsa-mir-3180-3	0
		hsa-mir-2116	4	hsa-mir-4524b	1	hsa-mir-548an	5	hsa-mir-3673	0
		hsa-mir-2681	5	hsa-mir-4529	5	hsa-mir-548aq	5	hsa-mir-3675	3
		hsa-mir-3116-1	3	hsa-mir-4536-1	5	hsa-mir-548ax	5	hsa-mir-3687	4
		hsa-mir-3116-2	3	hsa-mir-4536-2	5	hsa-mir-548f-1	5	hsa-mir-3690-2	2
		hsa-mir-3119-1	5	hsa-mir-4633	5	hsa-mir-548q	5	hsa-mir-4283-2	4
		hsa-mir-3119-2	5	hsa-mir-4636	5	hsa-mir-548s	5	hsa-mir-4313	4
		hsa-mir-3122	4	hsa-mir-4637	4	hsa-mir-548v	5	hsa-mir-4441	3
		hsa-mir-3145	5	hsa-mir-4661	4	hsa-mir-548x	5	hsa-mir-4509-2	4
		hsa-mir-3150b	4	hsa-mir-4670	5	hsa-mir-5579	5	hsa-mir-4635	4
		hsa-mir-3153	4	hsa-mir-4692	5	hsa-mir-5582	4	hsa-mir-520b	5
		hsa-mir-3156-2	5	hsa-mir-4698	5	hsa-mir-561	5	hsa-mir-5706	4
		hsa-mir-3202-2	5	hsa-mir-4716	5	hsa-mir-5687	5	hsa-mir-572	0
		hsa-mir-320e	4	hsa-mir-4729	5	hsa-mir-5690	5	hsa-mir-6080	4
		hsa-mir-3659	5	hsa-mir-4735	5	hsa-mir-5692c-2	1	hsa-mir-622	4
		hsa-mir-3667	4	hsa-mir-4737	3	hsa-mir-605	5	hsa-mir-6859-3	4
		hsa-mir-3678	3	hsa-mir-4759	5	hsa-mir-606	4	hsa-mir-8069	2
		hsa-mir-3686	5	hsa-mir-4764	5	hsa-mir-620	3	hsa-mir-8071-2	4
		hsa-mir-3910-1	5	hsa-mir-4765	5	hsa-mir-625	5	hsa-mir-941-3	3
		hsa-mir-3926-2	4	hsa-mir-4788	5	hsa-mir-629	4		
		hsa-mir-3927	5	hsa-mir-4796	5	hsa-mir-633	5		
		hsa-mir-3937	5	hsa-mir-4798	5	hsa-mir-6744	4		
De novo creation from material of TE origin	57 (34+23)	hsa-mir-1254-2	5	hsa-mir-4428	5	hsa-mir-6500	4	hsa-mir-3915	0
		hsa-mir-1255b-2	5	hsa-mir-4438	4	hsa-mir-6507	4	hsa-mir-422a	5
		hsa-mir-1256	5	hsa-mir-4445	4	hsa-mir-7641-2	3	hsa-mir-4419a	3
		hsa-mir-2115	1	hsa-mir-4448	3	hsa-mir-7975	3	hsa-mir-4425	5
		hsa-mir-3118-5	4	hsa-mir-4487	0	hsa-mir-1202	4	hsa-mir-4447	5
		hsa-mir-3133	5	hsa-mir-4525	4	hsa-mir-1255b-1	0	hsa-mir-4502	4
		hsa-mir-3134	4	hsa-mir-4666b	5	hsa-mir-1261	3	hsa-mir-548ag-2	5
		hsa-mir-3144	3	hsa-mir-5096	4	hsa-mir-1269b	2	hsa-mir-548au	5
		hsa-mir-3164	5	hsa-mir-5585	5	hsa-mir-1285-2	3	hsa-mir-548t	5
		hsa-mir-3657	5	hsa-mir-5586	5	hsa-mir-1290	5	hsa-mir-640	5

		hsa-mir-3672	5	hsa-mir-5590	5	hsa-mir-151b	5	hsa-mir-7151	3
		hsa-mir-3674	4	hsa-mir-5698	4	hsa-mir-302e	6	hsa-mir-7157	2
		hsa-mir-3683	1	hsa-mir-5701-2	4	hsa-mir-3135a	2		
		hsa-mir-3929	4	hsa-mir-579	5	hsa-mir-3163	4		
		hsa-mir-4421	4	hsa-mir-588	3	hsa-mir-3611	3		
De novo	121	hsa-mir-1182	5	hsa-mir-4256	5	hsa-mir-4503	5	hsa-mir-6077-1	0
		hsa-mir-1258	5	hsa-mir-4258	3	hsa-mir-4505	5	hsa-mir-6079	2
		hsa-mir-1262	5	hsa-mir-4261	2	hsa-mir-4509-3	4	hsa-mir-6089-1	1
		hsa-mir-1272	2	hsa-mir-4264	5	hsa-mir-4517	5	hsa-mir-6089-2	1
		hsa-mir-1276	5	hsa-mir-4265	4	hsa-mir-4521	2	hsa-mir-6126	5
		hsa-mir-1286	5	hsa-mir-4268	5	hsa-mir-4528	4	hsa-mir-614	3
		hsa-mir-147a	3	hsa-mir-4278	3	hsa-mir-4645	5	hsa-mir-624	5
		hsa-mir-1913	3	hsa-mir-4282	6	hsa-mir-4664	3	hsa-mir-641	5
		hsa-mir-2278	4	hsa-mir-4283-1	3	hsa-mir-4673	5	hsa-mir-648	5
		hsa-mir-3117	6	hsa-mir-4284	2	hsa-mir-4697	2	hsa-mir-6508	5
		hsa-mir-3128	5	hsa-mir-4289	5	hsa-mir-4705	5	hsa-mir-6509	5
		hsa-mir-3146	4	hsa-mir-4290	5	hsa-mir-4711	5	hsa-mir-6511a-1	0
		hsa-mir-3155b	5	hsa-mir-4294	5	hsa-mir-4717	0	hsa-mir-6755	5
		hsa-mir-3160-1	2	hsa-mir-4298	3	hsa-mir-4718	5	hsa-mir-6822	5
		hsa-mir-3160-2	2	hsa-mir-4305	5	hsa-mir-4733	5	hsa-mir-6874	3
		hsa-mir-3176	1	hsa-mir-4310	3	hsa-mir-4740	3	hsa-mir-7114	5
		hsa-mir-3182	3	hsa-mir-4316	2	hsa-mir-4746	4	hsa-mir-7150	7
		hsa-mir-3185	4	hsa-mir-4319	5	hsa-mir-4770	5	hsa-mir-765	3
		hsa-mir-3196	5	hsa-mir-4324	3	hsa-mir-4777	5	hsa-mir-7852	5
		hsa-mir-3199-2	5	hsa-mir-4326	2	hsa-mir-4791	5	hsa-mir-8054	5
		hsa-mir-3606	1	hsa-mir-4328	0	hsa-mir-5091	2	hsa-mir-8056	3
		hsa-mir-3609	2	hsa-mir-4329	5	hsa-mir-5092	5	hsa-mir-8057	5
		hsa-mir-3671	2	hsa-mir-4417	5	hsa-mir-5192	5	hsa-mir-8063	2
		hsa-mir-3689d-2	0	hsa-mir-4443	5	hsa-mir-5195	1	hsa-mir-8065	0
		hsa-mir-3914-1	5	hsa-mir-4453	3	hsa-mir-550b-1	5	hsa-mir-8066	3
		hsa-mir-3916	0	hsa-mir-4466	5	hsa-mir-554	5	hsa-mir-8081	4
		hsa-mir-3935	5	hsa-mir-4473	5	hsa-mir-559	4	hsa-mir-8088	3
		hsa-mir-3941	5	hsa-mir-4475	4	hsa-mir-5681b	0	hsa-mir-943	0
		hsa-mir-3945	4	hsa-mir-4478	5	hsa-mir-5688	5		
		hsa-mir-3974	1	hsa-mir-4479	3	hsa-mir-583	5		
		hsa-mir-3978	4	hsa-mir-4482	4	hsa-mir-596	3		

Table SD III-5: Proportion of miRNA for each mechanism of origination in exonic, intronic and intergenic regions. Number of miRNA in each category is associated with the pipeline in Figure III-5.

Mechanism of origination	**Number of** miRNA genes	**Proportion in intergenic regions**	**Proportion in exonic regions**	**Proportion in intronic regions**
Segmental duplication of a pre-existing miRNA	13	69,2%	7,7%	23,1%
Tandem duplication of a pre-existing miRNA	28	96,4%	0,0%	3,6%
Duplication of a pre-existing miRNA located in a TE	29	82,8%	0,0%	17,2%
	13	92,3%	0,0%	7,7%
Total and average of duplication events	*83*	*85,2%*	*1,9%*	*12,9%*

Insertion of one TE	50	74,0%	0,0%	26,0%
Insertion of two or more TE	23	72,7%	0,0%	27,3%
Inverted duplication	6	100,0%	0,0%	0,0%
Insertion of unknown origin	39	64,1%	0,0%	35,9%
Insertion of non-TE distal genomic origin	86	61,6%	0,0%	38,4%
	23	78,3%	0,0%	21,7%
Total and average of insertion events	*227*	*75,1%*	*0,0%*	*24,9%*
De novo creation from material of TE origin	34	73,5%	0,0%	26,5%
	23	78,3%	0,0%	21,7%
De novo	121	64,5%	1,7%	33,9%
Total and average of De novo events	*178*	*72,1%*	*0,6%*	*27,4%*

CHAPTER IV : PREDICTION OF HUMAN MIRNA TARGET GENES USING COMPUTATIONALLY RECONSTRUCTED ANCESTRAL MAMMALIAN SEQUENCES

4.1 Preface

This fourth chapter present MirAncesTar (microRNA Ancestral Target predictor), which proposes a new approach based on the analysis of ancestral genome sequences to improve the existing methods to predict miRNAs target genes. MirAncesTar was developed in continuity to the study on the mechanisms of origination of miRNAs, presented in CHAPTER III We realized that the inferred ancestral sequences could also be useful to help improve target genes predictions.

The problem of miRNAs target genes prediction is challenging: the average recall rate of experimentally validated target genes in humans (i.e. sensitivity) by the best known tools (e.g. miRanda (Enright et al. 2003), or TargetScan (Agarwal et al. 2015)) remain currently very low (about 15 to 25% in average of all human miRNAs having more than 200 experimentally validated targets in miRTarBase v6). Many reasons explain this situation (see Section 1.4), and a lot of work remains to be done to achieve a more acceptable level of prediction accuracy. New techniques are presented in the literature every year since the first version of miRanda in 2003 (Enright et al. 2003), most of the time improving the last microRNA target gene predictor released (at least in the hands of the author of these papers). The study presented in this chapter brings its contribution to this field, by boosting considerably the recall rate of existing target site prediction tools.

The rest of this chapter is reproduced from

> *Leclercq M, Diallo AB, Blanchette M (2016) Prediction of Human miRNA Target Genes using Computationally Reconstructed Ancestral Mammalian Sequence.* Paper submitted in February 2016 at Nucleic Acids Research.

4.2 Abstract

MicroRNAs (miRNA) are short single stranded RNA molecules derived from hairpin-forming precursors that play a crucial role as post-transcriptional regulators in eukaryotes and viruses. In the past years, many microRNA target genes (MTGs) have been identified experimentally. However, because of the high costs of experimental approaches, target genes databases remain incomplete. Although many target prediction programs have been developed in the recent years to identify MTGs *in silico*, their specificity and sensitivity remain low. Here, we propose a new approach called MirAncesTar, which uses ancestral genome reconstruction to boost the accuracy of existing MTGs prediction tools for human genome. For each miRNA and each putative human target UTR, our algorithm makes uses of existing prediction tools to identify putative target sites in the human UTR, its mammalian orthologs and inferred ancestral sequences. It then evaluates evidence in support of selective pressure to maintain target site counts (rather than sequences), accounting for the possibility of target site turnover. It finally integrates this measure with several simpler ones using a logistic regression. MirAncesTar improves the accuracy of existing MTG predictors by 26% to 157%. Source code and prediction results for human miRNAs, as well as supporting evolutionary data are available at http://cs.mcgill.ca/~blanchem/mirancestar.

4.3 Introduction

MicroRNAs (miRNAs) form a class of evolutionary conserved non-coding single-stranded RNA molecules involved in the regulation of gene expression by translational repression and mRNA destabilization (Ambros 1989; Ruvkun 2001; Swami 2010;

Kane et al. 2014). They are involved in the regulation of most animal and plant physiological processes (Osman 2012; Lawrie 2013; Teruel-Montoya et al. 2014), are implicated in many human diseases (Cooper et al. 2009; Dangwal et al. 2012; Goodall et al. 2013), and represent promising therapeutic applications (Lawrie 2013; Hammond 2015).

Unlike in plants, where the gene silencing requires a near-perfect complementarity between the miRNA and its mRNA target site, the repression of mRNA expression in animals is determined in part by the complementarity of a short region of the miRNA, called the seed. The seed is usually located between positions 2 to 7 of the miRNA, but variations exist (Bartel 2009). MiRNA target binding sites (MTBS) are generally located in the 3'UTR (3' untranslated region) of genes, but also, in a lower proportion, in their 5'UTR and open reading frame (ORF) (Lytle et al. 2007). MiRNAs produced from a single locus have the potential to silence a large number of genes (henceforth called its miRNA target genes (MTG)), and silenced genes are often targeted by more than one miRNA (Gennarino et al. 2012).

Experimental identification of miRNA target genes involves techniques such as gene expression analysis, using expression of ectopic miRNAs followed by the quantification of remaining non-degraded target mRNA on a genome-wide scale with microarrays or RNA-seq (Thomson et al. 2011), as well as approaches that directly identify interactions between mRNAs and proteins such as argonaute, including AGO2-PAR-CLIP (Farazi et al. 2014). But the number of experiments required to identify all MTGs of all miRNAs, in all tissues, conditions, and species of interest remains impractical. Therefore, computational methods to predict MTGs continue to be necessary.

Over the last few years, many tools predicting MTGs in various species have been developed. A first set of approaches, including miRanda (Enright et al. 2003) and PicTar (Krek et al. 2005), focused on identifying thermodynamically stable interaction sites between miRNAs and putative target genes. Later, various rule-based approaches,

such as PITA (Kertesz et al. 2007), or machine learning approaches, such as MirTarget2 (miRDB) (Wang 2008; Wang and El Naqa 2008) and TargetMiner (Bandyopadhyay and Mitra 2009), were proposed to integrate miRNA-mRNA duplex structural information with other types of features, such as target site accessibility, A/U content or target-site abundance, in order to improve prediction accuracy (Zheng et al. 2013).

Although these approaches have grown increasingly accurate over the past few years, and despite significant efforts, existing programs continue to produce high rates of false positives and false negatives (Zheng et al. 2013). In an effort to alleviate this problem, several programs, including mirMark (Menor et al. 2014), Diana-microT (Maragkakis et al. 2011) and TargetScan (Agarwal et al. 2015), have proposed to use inter-species sequence conservation as an indication of functional binding. MirMark considers as part of its input cross-species sequence conservation scores from PhastCons (Siepel et al. 2005), and TargetScan makes direct use of UTR sequence alignments to measure conservation on each branch of a calculated phylogenetic tree.

The underlying principle of using interspecies conservation is that functional miRNA target sites are important to the appropriate regulation of a gene's expression, so mutations that would disrupt binding are generally deleterious and over time more mutations should accumulate outside target sites than within them. However, concerns about the site conservation condition have been raised by Farh et al. (Farh et al. 2005) and Xu et al. (Xu et al. 2013), who observed that a large fraction of MTBS is not highly conserved among mammals orders. Applying strict requirements of sequence conservation thus results in an increased false-negative rate. Nevertheless, more than 60% of human protein-coding genes are under selective pressure to maintain pairing to miRNAs (Friedman et al. 2009), which explains in part why most mammalian miRNAs' 3'UTRs target sites are conserved above background levels (Xie et al. 2005).

The failure of conservation-based approaches to identify certain MTBS is partly due to an evolutionary process called binding site turnover (Venkataram and Fay 2010).

(Note that this concept is unrelated to that of miRNA turnover, which describes a change in miRNA expression due to degradation (Rogers and Chen 2013)). Because MTBS are short, random mutations can easily create new sites in the vicinity of existing ones. Since MTBS are generally not dependant on their exact position in the UTRs of a regulated gene, as long as the new site's position is in an accessible portion of the folded mRNA, the newly created site may be as potent as the previous one, thus reducing the selective pressure to maintain both. A mutation that would abrogate the old site would thus not be deleterious. The result is a turnover event, where although the target gene has continuously been targeted by the miRNA over evolutionary time, the position of the functional binding site has changed. Interspecies comparison would reveal that the sequence of neither the old nor the new site is particularly conserved, because both have been evolving neutrally for some time. This phenomenon is well characterized for transcription factor binding sites (Moses et al. 2006; Schmidt et al. 2010; Dermitzakis and Clark 2002) and taking it into consideration has been shown to improve the accuracy of binding predictions (Blanchette 2012). For miRNAs, target site turnover has been observed in cases where a target gene has multiple target sites for the same miRNA, a situation called cooperative targeting that allows MTBS to be lost and gained over time, as long as one or more remain present (Saetrom et al. 2007). Simkin et al. (Simkin et al. 2014) have recently exhibited several cases of miRNA target site turnover within primates.

In this paper, we introduce *MirAncesTar*, an approach to improve the miRNA target gene predictions made by existing tools by taking into account MTBS turnover. MirAncesTar uses computationally reconstructed ancestral mRNA sequences, rather than relying on pure conservation scores such as phastCons or PhyloP (Pollard et al. 2010; Siepel et al. 2005; Siepel and Haussler 2005). Our approach is not a predictor in itself, but rather an accuracy booster that can be applied to any existing predictor. Applied to three of the most commonly used MTBS predictors, MirAncesTar results in a large improvement in accuracy and compares favourably with three of the recent

MTG predictors making use of sequence conservation, mirMark (Menor et al. 2014), Diana-microT (Maragkakis et al. 2011), and TargetScan (Agarwal et al. 2015).

4.4 Material and Methods

4.4.1 Datasets

Human miRNAs were retrieved from miRbase v20 (Griffiths-Jones et al. 2006, 2008), for a total of 2,580 mature miRNAs. Experimentally validated miRNA targets (called known targets in this paper) were downloaded from miRTarBase version 6 (Hsu et al. 2011) which contains a total of 324,219 interactions between 2,619 miRNAs and 12,738 target genes. Of those, three subsets of miRNAs were considered: (i) M_{100} is a set of 100 miRNAs that had at least 200 known targets in the union of miRTarBase (release 5.0) and mirWalk (version 1) (Table SDIV-1) (ii) M_{396} is a set of 396 miRNAs that had at least 200 known targets in the most recent version of miRTarBase (release 6.0) (Table SDIV-2) (iii) $M_{308} \subset M_{396}$, a set of 308 miRNAs for which target predictions are available from both TargetScan and Diana-microT. The number known targets (based on miRTarBase (release 6.0)) used for the training and evaluation varies from 47,388 miRNA-targetGene pairs for M_{100} to 150,892 pairs for M_{396}.

Human 5’ and 3’ UTRs sequences of human protein-coding genes were retrieved from the UCSC genome browser (build GRCh37/hg19, RefSeq genes annotation). PhastCons conservation scores and conserved regions based on a 100-way multiple sequence alignment were also retrieved from UCSC genome browser.

4.4.2 Target gene predictors

MTGs predictors were selected based on their availability and running time. We considered five target gene predictors:

1. MiRanda (August 2010 version; (Enright et al. 2003)), which identifies putative targets by sequence alignment and ranks them based on thermodynamic stability. Default options.

2. RNAhybrid (Krüger and Rehmsmeier 2006), which determines the most stable hybridization site based on Energy parameters from Mathews et al. (Mathews et al. 1999), with length restrictions established for bulges and internal loops (Krüger and Rehmsmeier 2006). Default options, except for target length option (-m 1000000), a p-value threshold (-p 0.1) and the appropriate species selection (-s 3utr_human).
3. MirMark (version 1.0; (Menor et al. 2014)), a machine learning based method using more than 700 features describing the interactions between a miRNA and a UTR, such as target site availability, structure and sequence features, and PhastCons46way conservation data. Default options.
4. TargetScan (Agarwal et al. 2015), predicts miRNA target genes by searching for the presence of 6 to 8mer sites that match the seed region of a given miRNA and make use of species alignment to locate conserved sites. The most recent version integrates a regression model to improve TargetScan predictions. We did not run this tool ourselves but instead downloaded its predictions from targetscan.org, release 7.0, august 2015. Both conserved and non-conserved were considered.
5. Diana-microT v4 (Maragkakis et al. 2011), trained on miRbase v18, is based on binding and conservation features identified in high throughput experimental data, and calculated for each miRNA and each miRNA recognition elements responsible for the interaction with a target gene. Again, the score of a given gene was obtained by summing the scores of all predicted target sites in that gene.

For each tool, we obtained a ranked list of putative targets for each miRNA, sorted in decreasing order of the sum of confidence scores of predicted target sites.

4.4.3 Ancestral reconstruction

Ancestral genomes were reconstructed with an improved local version of Ancestor (Diallo et al. 2010), a tool which uses a maximum likelihood approach based on an

evolutionary model that takes in account insertions, deletions and substitutions. The reconstruction is computed from whole-genome multiple alignments of 46 vertebrate species, built with blastZ/Multiz pipeline (Blanchette et al. 2004b; Schwartz et al. 2003), and the phylogenetic tree reflecting their distance among each other, available from UCSC genome browser (Miller et al. 2007). 5'UTR and 3'UTR reconstructed ancestral sequences are available as supplementary data on our site.

4.4.4 Measuring evidence of selective pressure on predicted target site count

To identify targets for a given miRNA *M*, target site predictions are obtained first for each human 5' and 3'UTRs, their orthologs and ancestral sequences, using a given Single-Sequence Target Site Predictors (SSTSP). Consider Let the branch (*p,u*) of the phylogenetic tree, where *p* is the parent of *u*. We first build a null evolutionary model of the target site count, which is aiming at describing how the number of predicted target sites changes along branch *(p,u)*, assuming that the sequence under consideration is *not* a true target of *M*. In other words, we model the evolution of the count of false-positive predictions in UTRs. Let X_u denote the random variable corresponding to the number of sites at node *u*, and let $x_{g,u}$ denote the observed number of target sites predicted in the sequence at node *u* for gene *g*. Let $T_{(p,u)}(a,b) = \Pr[\, X_u = b \mid X_p = a \,]$ be the conditional probability of the sequence at *u* containing *b* sites given that the sequence at *p* contained *a* sites. $T_{(p,u)}$ is estimated on the basis that the vast majority of predicted target sites for *M* are false-positives, so that

$$T_{(p,u)}(a,b) = \frac{\sum_{g:Genes} \mathbb{I}(x_{g,p}, a) \cdot \mathbb{I}(x_{g,u}, b)}{\sum_{g:Genes} \mathbb{I}(x_{g,p}, a)},$$

where $\mathbb{I}(i,j) = 1$ if $i=j$ and 0 otherwise. Figure IV-1 illustrates some of the T conditional distributions for branches of the tree that have different lengths. Let $P_{(p,u)}(a,b) = \sum_{b' \geq b} T(a,b')$ be the p-value associated to observing *b* sites at node *u* given that there

were *a* sites at node *p*. The score of gene *g* as a putative target for miRNA *M* is obtained as

$$MirAncestarRaw(g,M) = \sum_{(p,u)\in Tree\ branches} -\log(P_{(p,u)}(x_p,x_u))$$

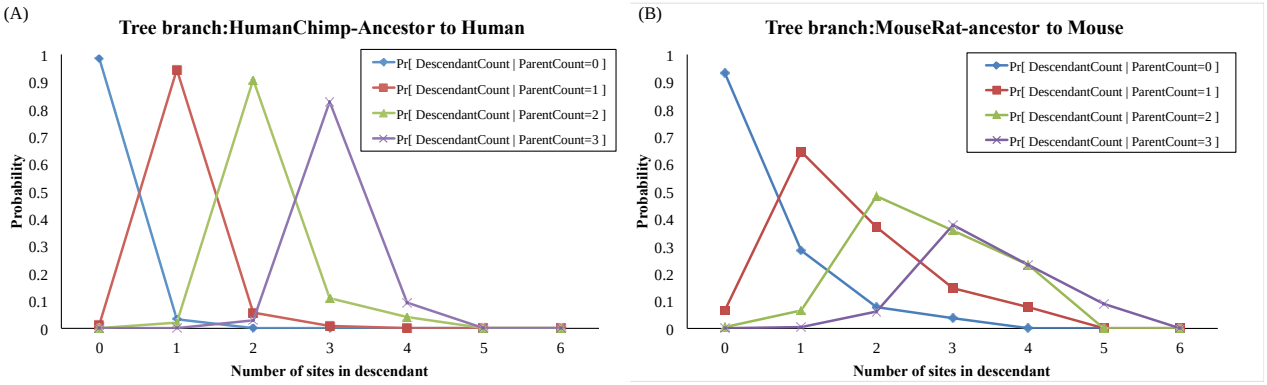


Figure IV-1: Examples of the posterior probability of the count of predicted target sites for let7a-5p, for two different branches of the phylogenetic tree: (A) the short branch leading from the human-chimp ancestor; (B) The longer branch leading from the mouse-rate ancestor to mouse.

4.4.5 Normalized conservation score

To take into account the fact that longer UTRs have a higher probability to be targeted than shorter ones, we introduce a second scoring mechanism that calculates for each branch (*p*,*u*) a p-value conditioned on the (binned) length *L(u)* of the sequence at node *p*. Specifically,

$$Tnorm_{(p,u),L}(a,b) = \frac{\sum_{g:Genes\ s.t.\ bin(L(g,u))=L}(\mathbb{I}_{x_{g,p}=a}) \cdot (\mathbb{I}_{x_{g,u}=b})}{\sum_{g:Genes\ s.t.\ bin(L(g,u))=L}(\mathbb{I}_{x_{g,p}=a})}$$

$$Pnorm_{(p,u),L}(a,b) = \sum_{b'\geq b} Tnorm_{(p,u),L}(a,b')$$

$$MirAncestarNorm(g,M) = \sum_{(p,u)\in Tree\ branches} -\log(Pnorm_{(p,u),bin(L(g,u))}(x_p,x_u))$$

The length binning function *bin(·)* is chosen so that approximately 500 genes fall within each bin.

4.4.6 Posterior probability normalized conservation score

While we found that the MirAncestarNorm score performed well, we realized that it over-penalizes genes with long UTRs, by intrinsically assuming that all genes are equally likely to be targets, irrespective of their UTR lengths. In reality, longer UTRs are generally more likely to be targets for any given miRNA. We thus introduced a last score called MirAncestarPost, which captures the posterior probability of a gene g being a target for *M*, given its length *L(g)* (in human) and its length-normalized score $MirAncestarNorm(g, M)$ (abbreviated $MAN(g, M)$ in the formula below). Let *P(g,M)* denote the event that g is a target of *M*.

$$MirAncestarPost(g, M) = \Pr[\, P(g, M) \mid L(g), MAN(g, M)]$$

$$= \frac{\Pr[L(g), MAN(g, M) \mid P(g, M)] \cdot \Pr[P(g, M)]}{\Pr[L(g), MAN(g, M) \mid P(g, M)] \cdot \Pr[P(g, M)] + \Pr[L(g), MAN(g, M) \mid \overline{P(g, M)}] \cdot \Pr[\overline{P(g, M)}]}$$

$$= \frac{\Pr[L(g) \mid P(g, M)] \cdot \Pr[MAN(g, M) \mid P(g, M)] \cdot \Pr[P(g, M)]}{\Pr[L(g) \mid P(g, M)] \cdot \Pr[MAN(g, M) \mid P(g, M)] \cdot \Pr[P(g, M)] + \Pr[L(g) \mid \overline{P(g, M)}] \cdot \Pr[MAN(g, M) \mid \overline{P(g, M)}] \cdot \Pr[\overline{P(g, M)}]}$$

where $\Pr[L(g) \mid P(g, M)]$, $\Pr[MAN(g, M) \mid P(g, M)]$, $\Pr[L(g) \mid \overline{P(g, M)}]$, and $\Pr[MAN(g, M) \mid \overline{P(g, M)}]$ are represented using multinomial distributions and estimated from the known targets genes and non-target genes (separately in each cross-validation iteration, with binning of the MirAncestarNorm score and length).

4.4.7 MirAncestar feature set and training

While the MirAncestarPost scoring approach is in itself competitive with existing SSTSPs, we are aware that it is not capturing some properties that could be useful for prediction. Thus, we elected to instead combine the three scoring schemes presented above (MirAncestarRaw, MirAncestarNorm, and MirAncestarPost) with a set of 7 other simpler measures:

1. UTRlength: The total length of the gene's UTRs in human.
2. TotalSitesCount: The total number of target sites predicted in the human gene, its orthologs, and ancestors.
3. TotalSitesCountNorm: TotalSitesCount/UTRlength.
4. HumanTotalScore-Conserved: The sum of the SSTSP scores of all predicted target sites predicted in the human sequence, limited to the highly conserved portions (defined by the PhastCons 46-way predictions).
5. HumanTotalScore-NonConserved: The sum of the SSTSP scores of all predicted target sites predicted in the human sequence, outside of the highly conserved portions.
6. HumanMaxScore-Conserved: The maximum of the SSTSP scores of all predicted target sites predicted in the human sequence, limited to the highly conserved portions.
7. HumanMaxScore-NonConserved: The maximum of the SSTSP scores of all predicted target sites predicted in the human sequence, outside of the highly conserved portions.

The 10 features are combined using a logistic regression approach trained and evaluated using 10-fold cross validation, using Weka (Hall et al. 2009). Because we work with unbalanced classes, we used a cost sensitive classifier, used to reweight training instances according to the total cost assigned to each class. This weighting method simulates stratification, avoiding downsampling the majority class and allowing taking advantage of the full available data. The cost matrix associated with the cost-sensitive classifier was set as follows: False-negatives were assigned a cost of 1, while false-positives were assigned a cost of |PositiveTrainingSet| / |NegativeTrainingSet|. The logistic regression parameters were learned based on a positive training set consisting of the set of known targets of M_{100} miRNAs with more than 200 known targets, and the negative training set was the set of non-targets for the same M_{100} miRNAs. For each SSTSP, a different set of logistic regression parameters were learned.

4.5 Results

MirAncesTar is an approach that makes use of comparative genomics data to improve the predictions of the target genes of a given microRNA by evaluating the conservation of the *count* of predicted target sites among mammalian orthologs and their ancestors. MirAncesTar exploits existing Single-Sequence Target Site Predictors (SSTSP) such as miRanda (Enright et al. 2003) to identify candidate target sites in genes of the genome under study (here, human), their orthologs (here, from 34 other mammals) and computationally reconstructed ancestral sequences. The method does not directly evaluate sequence conservation of target sites *per se*, but instead seeks evidence for selective pressure to maintain a certain number of target sites in gene's UTRs (irrespective of their position), thus allowing for target site turnover. The target site count conservation score is then combined with other simpler measures (UTR length, sum and maximum of site SSTSP scores inside and outside conserved regions, and total number of predicted sites (see Methods)), using a logistic regression predictor. Here, we report our evaluation of the accuracy of MirAncesTar compared to a variety of other existing tools, and investigate the factors that affect its performance.

4.5.1 MirAncesTar improves the accuracy of miRNA target gene prediction

For each of the 18,653 UTRs sequences of human genes annotated in RefSeq release 66 (after merging isoforms), we extracted orthologous mammalian sequences from the UCSC 46-way vertebrate whole-genome alignment (Blanchette et al. 2004b; Kent et al. 2002), which yielded a maximum of 34 aligned mammalian orthologs. Ancestral sequences for each of the 34 internal nodes in the phylogenetic tree (Figure SD IV-1) were inferred using a local version of Ancestors 1.1 (Blanchette et al. 2008; Diallo et al. 2010), which was previously estimated to be able to infer ancestral mammalian sequences with accuracy ranging from 85 to 98%, depending on the ancestral node. This produced a set of up to 69 extant or ancestral orthologous sequences per human gene, although for most genes, orthologs are missing in a small number of species

(average number of orthologs/ancestors per gene: 65.3; 0.1% of genes have no orthologs outside primates).

We trained and tested (using 10-fold cross-validation) our various predictors on experimentally identified target sites of a set of 100 well-characterized miRNAs (see Methods). These 100 miRNAs have on average 474 known targets per miRNA. For each SSTSP $P \in$ {miRanda, RNAhybrid, mirMark}, we evaluated the accuracy of MirAncesTar$_P$, the MirAncesTar predictor based on the predictions obtained with P, and compared it to P itself when applied to the human sequences alone. For each miRNA and each predictor, we obtained the ranked list of predicted targets among the RefSeq genes, sorted by the sum of confidence values (prediction score) of predicted targets. We then evaluated the proportion of all known targets captured among the top k predictions (recall), for k ranging from 1 to 1000 (Figure IV-2A-C). Although receiving-operator curves (ROC) are a more classical way to evaluate predictors (presented in Figure SD IV-2), we find that the former provides a more intuitive and practical evaluation of a predictor, by providing the answer to the question: if a researcher was to look at the top k predictions made by a given tool, what fraction of the known targets would be recovered?

Figure IV-2A compares the recall curves of miRanda and MirAncesTar$_{\text{miRanda}}$. The latter provides a notable improvement. For example, at k = 1000, MirAncesTar$_{\text{miRanda}}$ has an average recall of 26.1%, compared to 18.4% for miRanda, a relative increase of 20.7%. The recall relative increase is actually much larger when limiting our attention to a smaller number of top predictions; for example, at k = 100, MirAncesTar$_{\text{miRanda}}$ improves the recall of miRanda by 67%. The improvements in recall are even more significant for RNAhybrid (Figure IV-2B) where MirAncesTar yields a 158% increase in recall (at k = 1000). MirMark is not a true single-sequence predictor because it uses as part of its input a measure of interspecies sequence conservation (PhastCons score (Siepel et al. 2005)). As such, we were not able to use it directly to predict targets sites in orthologs and ancestors and instead modified it to

not take sequence conservation into consideration (see Methods). The resulting predictor (MirMark0) had a recall that was slightly worse than the original MirMark (Figure IV-2C), but MirAncesTar$_{MirMark0}$ nonetheless succeeded at increasing the recall value 63% above that of MirMark (at k = 1000). (Because MirMark produces better results if we calculate the recall based on the maximum of the scores of the putative sites instead of their sum, we used the former method in this case). Overall, MirAncesTar produced significant improvements over all SSTSP we considered. The best recall curve was obtained using MirAncesTar$_{miRanda}$, which outperformed the other two MirAncesTar-based predictors, by 72 to 78% at k = 1000, and even more for smaller values of k.

Although MirAncesTar performs on average better than SSTSP predictors, its accuracy varies depending on the miRNA whose targets are being predicted. Figure IV-2D presents the recall obtained by MirAncesTar$_{miRanda}$ (at k = 1000) for each miRNA, compared to that obtained with miRanda alone. MirAncesTar$_{miRanda}$ improves the recall for 93 of the 100 miRNAs considered, including 39 where the improvement was statistically significant (in red in the figure; $p \leq 0.05$; two-tailed Student t-test). In one case, the recall is more than doubled. Figure IV-2E-F show the analogous results for RNAhybrid and MirMark. Improved recall values were obtained for 99% and 98% of miRNAs resp., with 85% and 78% of these improvements being statistically significant.

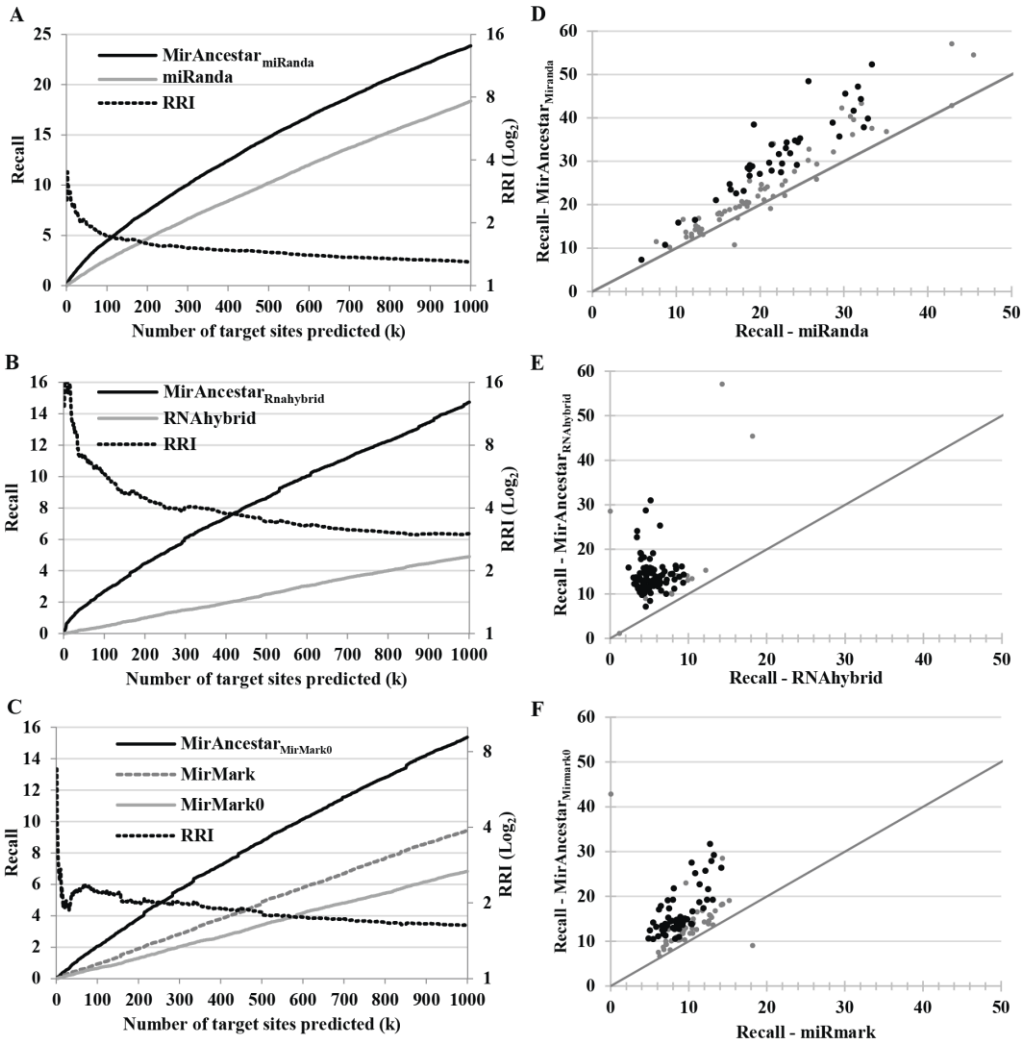


Figure IV-2. Comparison of the recall (primary y-axis) and relative recall improvement (RRI, secondary y-axis, log-scale) of single-sequence target gene predictors and their corresponding MirAncesTar predictors. (A-C) Average (over 100 miRNAs) of the recall (percentage of known targets recovered) as a function of the number of sites being predicted (k). (A) miRanda; (B) RNAhybrid; (C) mirMark with and without PhastCons. (D-F) Recall (at k = 1000 predictions), for each of the 100 miRNAs, for each SSTSP (x-axis) and its corresponding MirAncesTar predictor (y-axis). MiRNAs for which the difference between the two recall values is statistically significant (p-value<0.05 based on two-tailed Student t-test) are shown in black.

TargetScan (Agarwal et al. 2015) and Diana-microT (Maragkakis et al. 2011) are two of the most widely used miRNA target gene predictors that exploit interspecies comparisons to score putative target sites. For that reason, we could not apply them as

a SSTSP for MirAncesTar to be based off. Because both tools offers precomputed target predictions for a large set of miRNAs, we were able to expand our study to a larger set of 308 well-characterized miRNAs having at least 200 known targets and for which target gene predictions were available from both TargetScan and Diana-microT. Other SSTSPs were not used on this larger data set because of their excessive running time. To estimate the recall rate, we again listed in decreasing order the scores provided by the tools. Diana-microT is constituted of a list of genes for each miRNA, associated to a score. For TargetScan, target ranking was produced based on the sum of the context++ scores of each gene (including both conserved and non-conserved sites). Figure IV-3 show that MirAncesTar$_{Miranda}$ obtains recall values that are significantly larger than those of Diana-microT (by approximately 25 to 40%, depending on the value *k*). Recall values are comparable to those of TargetScan at *k*=1000, but approximately 10% better for $k < 400$. For a larger set of miRNAs (Figure SD IV-3), MirAncesTar reports in average a higher recall rate than TargetScan for all values of *k*.

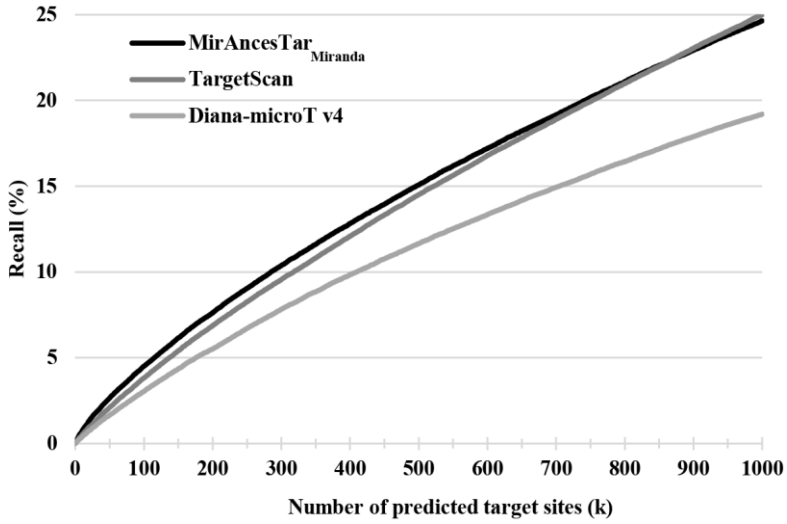


Figure IV-3: Recall obtained by MirAncesTar$_{Miranda}$, TargetScan and Diana-microT v4, averaged across 308 miRNAs.

To better understand the properties of different prediction methods, we compared the target predictions of miRanda, TargetScan, Diana-microT, and MirAncesTar$_{Miranda}$ on the same set of 308 miRNAs. Interestingly, the set of target predictions made by the three tools have only moderate overlap (Figure IV-4). This suggests that the three tools are somewhat complementary. Genes predicted as targets by all four tools have large positive predictive value (PPV; fraction of positive predictions that are currently known to be correct), at 21.6%. Those predicted by three of the tools also have high PPV, ranging from 23.9% (MirAncestar+TargetScan+Diana-microT) to only 8.7% (TargetScan+Diana-microT+miRanda). Targets predicted by a single tool had lower PPV, ranging from 4.2% (miRanda alone) to 6% (TargetScan alone). This shows that significant gains in specificity can be obtained by combining the three comparative genomics based predictors.

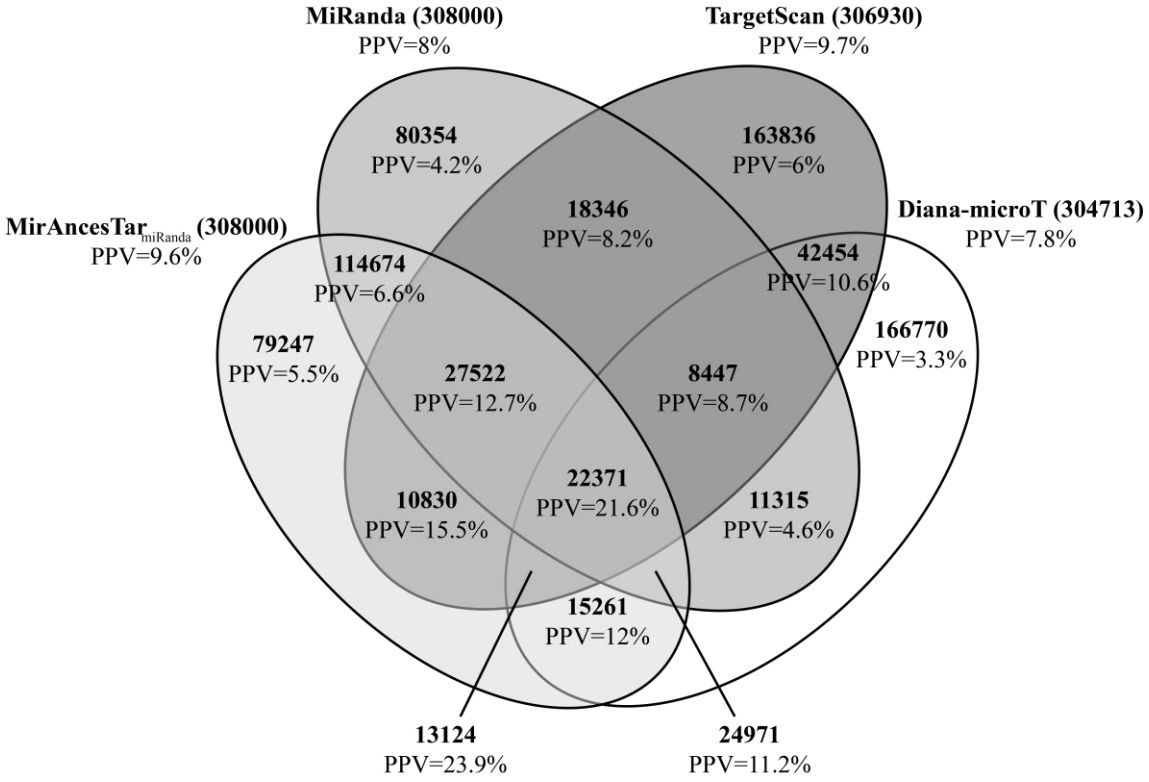


Figure IV-4: Venn diagrams of the predictions made with miRanda, MirAncesTar$_{miRanda}$, and TargetScan, Diana-microT on 308 miRNAs, with *k*=1000 for each tool and miRNA.

4.5.2 MirAncesTar exploits sequence conservation but is robust with respect to target site turnover

As seen in Figure IV-2D, the recall of $MirAncesTar_{miRanda}$ (at k=1000) varies quite widely between miRNAs, ranging from 7 to 57%. Two main reasons appear to explain this variability. The first is the ability of miRanda to correctly identify candidate target sites in human. Indeed, the correlation between the recall values of miRanda and $MirAncesTar_{miRanda}$ is quite high (R^2 = 0.84, Figure IV-2D); this is unsurprising, since $MirAncesTar_{miRanda}$ builds off miRanda. Second, the extent to which $MirAncesTar_{miRanda}$ improves the target recall (at k = 1000) compared to miRanda varies from a 2-fold increase for miR-92b-3p (from 19.2% to 38.4%) to no improvement for several miRNAs, and, in the case of let-7i-3p, miR-324-3p, 324-5p, 30b-3p, 373-3p, 30d-3p and 92a-1-5p, to a slight decrease in recall. We sought to understand the particular characteristics of a miRNA that may be associated with a gain or loss in accuracy with $MirAncesTar_{miRanda}$. We regressed the $MirAncesTar_{miRanda}$ recall improvement against a number of miRNA properties (nucleotide content, average PhastCons UTR conservation scores of known targets, total predicted target sites count, etc.). The only significant interaction identified was with the average PhastCons conservation scores of known targets (p-value = 2.7×10^{-6}), which suggests that, unsurprisingly, MirAncesTar is more effective for miRNAs whose target genes have a tendency to have more conserved UTRs. Those are often miRNAs that target transcription factors, especially those whose family is involved in regulation of embryonic development and gastrulation such as let-7d (Wong et al. 2012), let-7e (Colas et al. 2012) and mir-124 (Lee et al. 2010), which are the three miRNAs for which MirAncesTar has the highest recall values.

One of the key innovations of MirAncesTar is its ability to tolerate MTBS turnover. This is supported by the fact that the UTRs correctly predicted as targets by MirAncesTar tend to have lower conservation levels (avg. PhastCons of 0.305) than those predicted by TargetScan, Diana-microT, MirMark (respectively avg. PhastCons

score of 0.322, 0.419, and 0.507). Figure IV-5 illustrates the predicted target sites for hsa-let-7a-5p in the *SMCR8* gene, a known target of that miRNA, which obtained a high prediction score (target ranked 39th out of 18,653 genes) by MirAncesTar but was scored poorly by other conservation-based tools (target ranking by mirMark: 4,896th, TargetScan: 768th, Diana-microT: not in the top 7338 predictions available for this miRNA). Clearly, no specific target site predicted in human is conserved across all mammals. Interestingly however, there is evidence of a turnover event in rodents (mouse, rat, and kangaroo-rat), where a site that was otherwise conserved in most mammals was shifted by approximately 600 bp. Overall, the number of predicted sites in extant ancestral sequences (shown on the phylogenetic tree in the figure) is remarkably constant, which is why this target is scored well by MirAncesTar.

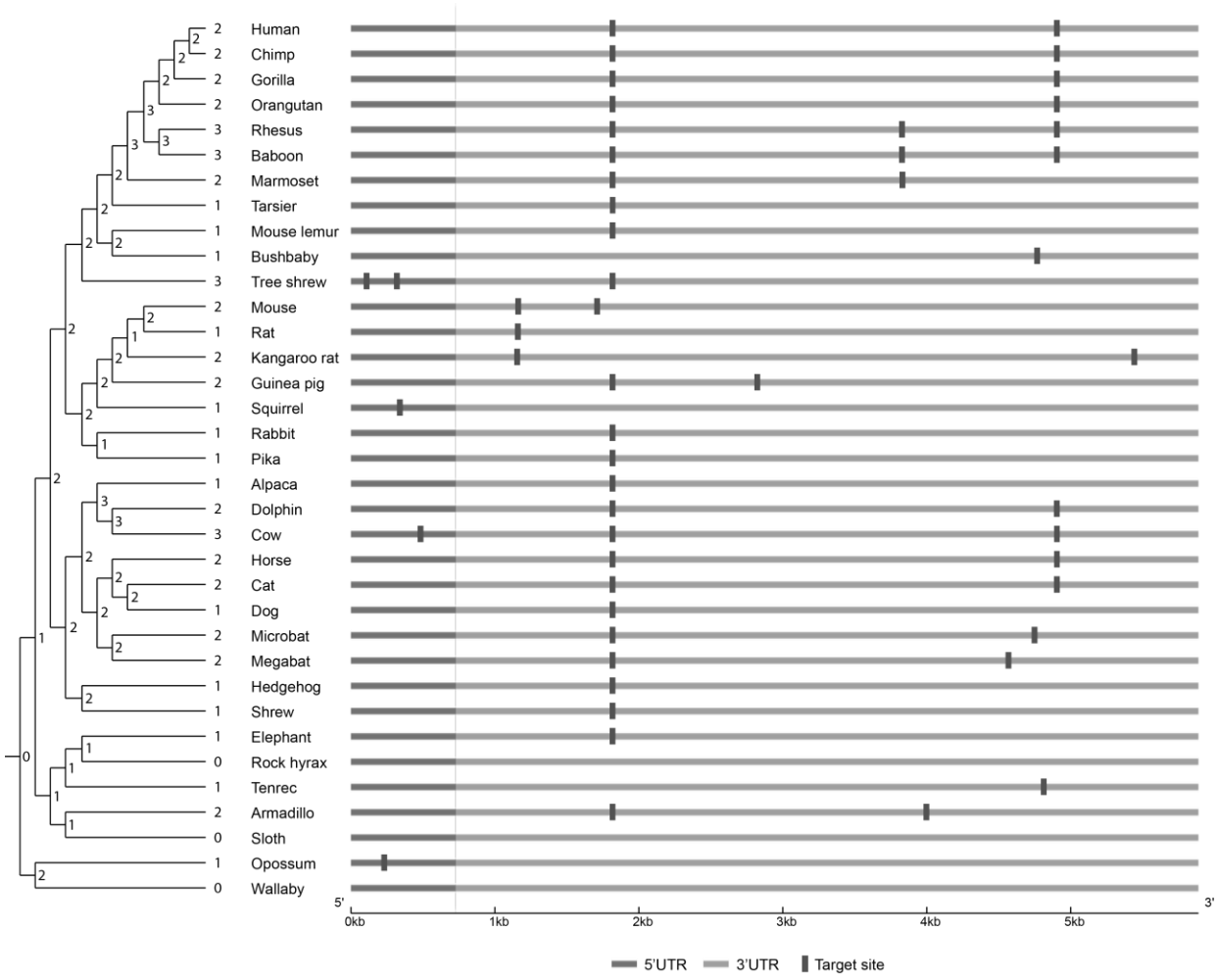


Figure IV-5: Example of putative target site turnover for hsa-let-7a-5p in the SMCR8 gene. Putative target sites predicted by miRanda in each species are marked. The number of predicted target sites in each species and each computationally reconstructed ancestral sequence is shown on the nodes of the species tree. The position of sites for non-human species is converted to that of its human orthologous position through the multiple sequence alignment.

4.5.3 Contribution of the different features used by MirAncesTar

MirAncesTar is a logistic regression predictor where each putative target is represented using ten features that capture in different ways the number of predicted target sites in the species of interest (human) and/or in its orthologs and ancestors (see Methods). It is instructive to consider how each of these features contributes to the overall accuracy of the predictor. Figure SD IV-3 shows the recall curves obtained for

each of the ten features when used individually as predictor, for SSTSP = miRanda. By far the most informative feature is MirAncesTarPost, a score that captures evidence of selective pressure to maintain the number of candidate target sites during the evolution of the putative target. In itself, it is competitive with TargetScan and outperforms the three SSTSP used in this study. Interestingly, the second most predictive feature is the number of sites predicted by miRanda *outside* highly conserved portions of the UTR (PhastCons), which ranks better than the analogous number of target sites located within such conserved regions. This counterintuitive result is caused by the fact that most validated target UTRs contain zero conserved predicted targets.

4.6 Discussion

We propose here a new algorithm that relies on ancestral sequence reconstruction to improve the predictions of miRNA predictors in human, based on the idea that, despite the fact that UTRs are generally under negative selective pressure to maintain a given set of miRNA target sites, individual target sites are often subject to turnover. MirAncesTar builds off an evolutionary model that characterizes how the number of predicted targets in a neutrally evolving sequence changes over time, and seeks to identify UTRs that depart from that null model. It uses predictions made by existing SSTSPs, executed on UTRs of mammalian species and their ancestors, to identify genes that exhibit evidence of this type of selective pressure. It then learns how best to combine this measure of selective pressure with other simpler measures of target site content in the target species and its ancestors/orthologs. MirAncesTar significantly improved the overall accuracy of the three single-sequence target site predictors it was based off (miRanda, RNAhybrid, and mirMark). For certain miRNAs, recall (at k = 1000) was more than doubled, while we found no miRNA for which recall was significantly decreased. The best overall accuracy was obtained using miRanda as SSTSP, although MirAncesTar produced its largest increase in accuracy for RNAhybrid (158% increase in recall at k = 1000). $\text{MirAncesTar}_{\text{miRanda}}$ also

outperforms existing sequence conservation based predictors Diana-microT and MirMark, and has slightly better performance than TargetScan. Notably, the accuracy gains obtained using MirAncesTar$_{miRanda}$ appear to be largely due to its ability to tolerate target site turnover. Not all miRNAs benefit equally from the application of MirAncesTar$_{miRanda}$. Those for which MirAncesTar results in the largest increase in recall are those that (i) are already well predicted by miRanda, and (ii) whose known targets tend to exhibit elevated levels of sequence conservation, such as miRNAs whose function is to regulate cell differentiation or organismal development.

An important benefit of MirAncesTar is that it can be used with any existing single-sequence target site predictor, and with the three such predictors considered here, it results in significant gains in accuracy. By decoupling the individual target site prediction task (performed by miRanda, RNAhydbrid, mirMark, or other tools) from the evaluation of selective pressure on target site count (performed by MirAncesTar), we obtain an approach that will age well because it will benefit from future improvements in single sequence target site predictors.

Although the overall recall of TargetScan and MirAncesTar are similar, the properties of predicted targets are quite different. Part of the explanation lies in how UTR length affects prediction accuracy. The recall of TargetScan is almost independent of UTR length: short targets (<500 bp) are recovered with the same recall as long ones (>5000 bp) (Figure SD IV-4A). Conversely, the recall of MirAncesTar increases with target length, from only 3% for short UTRs to more than 50% for long ones. This is due to the fact that evidence of selective pressure on target site counts is easier to detect for target genes that contain a relatively large number of predicted sites. On the contrary, the precision (positive predictive value) of MirAncesTar is largely independent of target length: in other words, a gene that is predicted to be a target by MirAncesTar has approximately 10% probability of being a known target, irrespective of its length (Figure SD IV-4B). Instead, the precision of TargetScan is length-dependent, ranging from only 6% for genes with very short UTRs to more than 15% for genes with

relatively long UTRs. The predictions made by Diana-microT show an intermediate effect.

A similar analysis is instructive to highlight the effect of UTR sequence conservation on precision and recall. Unsurprisingly, the precision of each method improves with sequence conservation (average UTR PhastCons score) (Figure SD IV-4C). However, large differences are observed in terms of recall (Figure SD IV-4D): while both TargetScan and MirAncesTar recover 20-30% of known targets irrespective of their sequence conservation, Diana-microT has recall values that range from very poor (6%) for weakly conserved UTRs to very high (>40%) for highly conserved ones.

These differences have important consequences on the interpretation of the predictions made by these tools. On the one hand, the length bias of MirAncesTar prediction, and the conservation bias of Diana-microT, can induce artificial functional enrichment (e.g. for a gene ontology enrichment analysis) among predicted targets. On the other hand, investigators interested in validating experimentally predicted targets should expect a length-dependent success rate if they base their study on TargetScan, but not so with MirAncesTar.

Several possible directions may prove fruitful to explore in order to further improve the accuracy of MirAncesTar. First, in the present version, the position of predicted sites is not taken into consideration; only the total count matters. While this conveniently allows for target site turnover, it could be that an approach that would be semi-position specific would have some benefits. One could for example consider a model where changes in target site position are allowed but penalized. Second, improvements may be obtained by considering more sophisticated machine learning predictors to replace our logistic regression classifier, or by considering additional sets of features. In particular, one may attempt to predict target genes based on the target site predictions of more than one SSTSP, although this would come at the expense of additional running time.

Finally, we note that although our focus here was on predicting target genes for human miRNAs, it should be equally powerful in other mammalian species (provided a sufficiently large number of known miRNA target sites are available for the training). MirAncesTar should also be applicable to other groups of species that are where sufficiently many closely related taxa are sequenced, such as fruit flies (Clark et al. 2007) or crucifers (Haudry et al. 2013), although the accuracy of ancestral sequence reconstruction may not be as high for these lineages.

In conclusion, this paper is a striking example of a prediction task that can be achieved more accurately through a careful analysis of not only a human sequence and its orthologs, but also of computationally reconstructed ancestral sequences. Tracing the evolution of a region across the mammalian phylogeny significantly eases the detection of compensatory events such as target site turnover, by helping resolve the timing of these events. Did the loss of a particular target site precede or follow the creation of another one nearby? The answer to this question lies in the analysis of ancestral sequences, and is crucial for detecting evidence of selective pressure. We note that this concept is quite general and could quite easily be applied to other sequence-based prediction tasks. As the number of species whose genome increases (Koepfli et al. 2015) so will the power of this family of approaches.

4.7 Acknowledgements

This work was funded in part by a NSERC Discovery grant to MB, and by a FRQNT scholarship to ML. We would like to thank the Clumeq (Supercomputer Consortium Laval UQAM McGill and Eastern Quebec) for the access to the clusters Colosse and Guillimin, and the LICEF research center for their access to the ERASME cluster.

4.8 Supplementary Data

Table SDIV-1: List of the 100 miRNAs used to train and test MirAncesTar

hsa-let-7a-3p	hsa-let-7i-5p	hsa-miR-145-5p	hsa-miR-18a-3p	hsa-miR-25-3p	hsa-miR-30e-3p	hsa-miR-744-5p
hsa-let-7a-5p	hsa-miR-100-5p	hsa-miR-146a-5p	hsa-miR-18a-5p	hsa-miR-26a-5p	hsa-miR-30e-5p	hsa-miR-7-5p
hsa-let-7b-3p	hsa-miR-101-3p	hsa-miR-148b-3p	hsa-miR-192-5p	hsa-miR-26b-5p	hsa-miR-320a	hsa-miR-877-3p
hsa-let-7b-5p	hsa-miR-103a-3p	hsa-miR-149-5p	hsa-miR-193b-3p	hsa-miR-296-3p	hsa-miR-324-3p	hsa-miR-92a-1-5p
hsa-let-7c-5p	hsa-miR-106b-5p	hsa-miR-155-3p	hsa-miR-196a-5p	hsa-miR-29a-3p	hsa-miR-324-5p	hsa-miR-92a-3p
hsa-let-7d-3p	hsa-miR-10a-5p	hsa-miR-15a-3p	hsa-miR-19b-3p	hsa-miR-29a-5p	hsa-miR-331-3p	hsa-miR-92b-3p
hsa-let-7d-5p	hsa-miR-10b-5p	hsa-miR-15a-5p	hsa-miR-200c-3p	hsa-miR-29c-3p	hsa-miR-335-5p	hsa-miR-93-3p
hsa-let-7e-3p	hsa-miR-122-5p	hsa-miR-15b-5p	hsa-miR-20a-5p	hsa-miR-30a-3p	hsa-miR-34a-3p	hsa-miR-93-5p
hsa-let-7e-5p	hsa-miR-124-3p	hsa-miR-16-5p	hsa-miR-21-3p	hsa-miR-30a-5p	hsa-miR-34a-5p	hsa-miR-9-5p
hsa-let-7f-1-3p	hsa-miR-124-5p	hsa-miR-17-3p	hsa-miR-215-5p	hsa-miR-30b-3p	hsa-miR-373-3p	hsa-miR-98-5p
hsa-let-7f-2-3p	hsa-miR-125a-5p	hsa-miR-17-5p	hsa-miR-21-5p	hsa-miR-30b-5p	hsa-miR-373-5p	
hsa-let-7f-5p	hsa-miR-125b-5p	hsa-miR-181a-5p	hsa-miR-221-3p	hsa-miR-30c-1-3p	hsa-miR-375	
hsa-let-7g-3p	hsa-miR-128-3p	hsa-miR-181b-5p	hsa-miR-222-3p	hsa-miR-30c-2-3p	hsa-miR-423-3p	
hsa-let-7g-5p	hsa-miR-130b-3p	hsa-miR-183-5p	hsa-miR-23b-3p	hsa-miR-30c-5p	hsa-miR-423-5p	
hsa-let-7i-3p	hsa-miR-132-3p	hsa-miR-186-5p	hsa-miR-24-3p	hsa-miR-30d-3p	hsa-miR-484	

Table SDIV-2: List of the 396 miRNAs having more than 200 experimentally validated target genes

hsa-let-7a-5p	hsa-miR-181b-5p	hsa-miR-30e-5p	hsa-miR-410-3p	hsa-miR-4781-3p	hsa-miR-5582-3p	hsa-miR-6799-5p
hsa-let-7b-5p	hsa-miR-181c-5p	hsa-miR-3122	hsa-miR-421	hsa-miR-4789-3p	hsa-miR-5589-5p	hsa-miR-6807-5p
hsa-let-7c-5p	hsa-miR-181d-5p	hsa-miR-3135b	hsa-miR-423-3p	hsa-miR-4789-5p	hsa-miR-5590-3p	hsa-miR-6808-5p
hsa-let-7d-5p	hsa-miR-1827	hsa-miR-3148	hsa-miR-423-5p	hsa-miR-4793-3p	hsa-miR-5692a	hsa-miR-6809-3p
hsa-let-7e-5p	hsa-miR-183-5p	hsa-miR-3163	hsa-miR-424-5p	hsa-miR-4796-3p	hsa-miR-5693	hsa-miR-6817-3p
hsa-let-7f-5p	hsa-miR-185-5p	hsa-miR-3175	hsa-miR-4252	hsa-miR-484	hsa-miR-5698	hsa-miR-6821-3p
hsa-let-7g-5p	hsa-miR-186-3p	hsa-miR-3183	hsa-miR-4257	hsa-miR-485-5p	hsa-miR-574-5p	hsa-miR-6825-5p
hsa-let-7i-5p	hsa-miR-186-5p	hsa-miR-3187-3p	hsa-miR-4279	hsa-miR-497-5p	hsa-miR-588	hsa-miR-6829-3p
hsa-miR-100-5p	hsa-miR-18a-3p	hsa-miR-3190-5p	hsa-miR-4282	hsa-miR-498	hsa-miR-590-3p	hsa-miR-6832-3p
hsa-miR-101-3p	hsa-miR-18a-5p	hsa-miR-32-5p	hsa-miR-4284	hsa-miR-5006-3p	hsa-miR-603	hsa-miR-6832-5p
hsa-miR-103a-3p	hsa-miR-190a-3p	hsa-miR-320a	hsa-miR-4287	hsa-miR-5011-5p	hsa-miR-607	hsa-miR-6833-3p
hsa-miR-106a-5p	hsa-miR-1910-3p	hsa-miR-324-3p	hsa-miR-4295	hsa-miR-503-5p	hsa-miR-6077	hsa-miR-6838-5p
hsa-miR-106b-5p	hsa-miR-192-5p	hsa-miR-324-5p	hsa-miR-4419a	hsa-miR-504-3p	hsa-miR-6086	hsa-miR-6840-3p
hsa-miR-107	hsa-miR-193b-3p	hsa-miR-329-3p	hsa-miR-4419b	hsa-miR-505-3p	hsa-miR-6127	hsa-miR-6843-3p
hsa-miR-10a-5p	hsa-miR-195-5p	hsa-miR-331-3p	hsa-miR-4430	hsa-miR-508-5p	hsa-miR-6129	hsa-miR-6845-3p
hsa-miR-10b-5p	hsa-miR-196a-5p	hsa-miR-335-3p	hsa-miR-4435	hsa-miR-5089-5p	hsa-miR-6130	hsa-miR-6848-3p
hsa-miR-122-5p	hsa-miR-197-3p	hsa-miR-335-5p	hsa-miR-4438	hsa-miR-5095	hsa-miR-6131	hsa-miR-6849-3p
hsa-miR-1224-3p	hsa-miR-1976	hsa-miR-339-5p	hsa-miR-4446-5p	hsa-miR-5096	hsa-miR-6133	hsa-miR-6851-5p
hsa-miR-1226-3p	hsa-miR-19a-3p	hsa-miR-340-5p	hsa-miR-4458	hsa-miR-512-3p	hsa-miR-6134	hsa-miR-6864-3p
hsa-miR-1228-3p	hsa-miR-19b-3p	hsa-miR-342-3p	hsa-miR-4459	hsa-miR-5193	hsa-miR-615-3p	hsa-miR-6867-3p
hsa-miR-1236-3p	hsa-miR-204-5p	hsa-miR-34a-5p	hsa-miR-4469	hsa-miR-5196-5p	hsa-miR-616-5p	hsa-miR-6867-5p
hsa-miR-124-3p	hsa-miR-20a-5p	hsa-miR-3609	hsa-miR-4478	hsa-miR-519a-3p	hsa-miR-619-5p	hsa-miR-6873-3p
hsa-miR-1247-3p	hsa-miR-20b-5p	hsa-miR-3612	hsa-miR-4500	hsa-miR-519b-3p	hsa-miR-623	hsa-miR-6875-3p
hsa-miR-125a-3p	hsa-miR-21-5p	hsa-miR-3613-3p	hsa-miR-450a-1-3p	hsa-miR-519c-3p	hsa-miR-627-3p	hsa-miR-6881-3p
hsa-miR-125a-5p	hsa-miR-211-5p	hsa-miR-362-3p	hsa-miR-4510	hsa-miR-519d-3p	hsa-miR-642a-5p	hsa-miR-6883-5p
hsa-miR-125b-5p	hsa-miR-215-5p	hsa-miR-363-3p	hsa-miR-4524a-3p	hsa-miR-520a-3p	hsa-miR-646	hsa-miR-6884-5p
hsa-miR-1260b	hsa-miR-216a-3p	hsa-miR-3652	hsa-miR-454-3p	hsa-miR-520b	hsa-miR-6499-3p	hsa-miR-6890-3p
hsa-miR-1273e	hsa-miR-218-5p	hsa-miR-3662	hsa-miR-455-3p	hsa-miR-520c-3p	hsa-miR-650	hsa-miR-6893-5p
hsa-miR-1273f	hsa-miR-221-3p	hsa-miR-3663-5p	hsa-miR-4635	hsa-miR-520d-3p	hsa-miR-6504-3p	hsa-miR-7-5p
hsa-miR-1273g-3p	hsa-miR-222-3p	hsa-miR-3666	hsa-miR-4638-5p	hsa-miR-520e	hsa-miR-6506-5p	hsa-miR-7106-5p
hsa-miR-1273h-5p	hsa-miR-223-5p	hsa-miR-3667-3p	hsa-miR-4649-3p	hsa-miR-520g-3p	hsa-miR-6511a-5p	hsa-miR-7110-3p
hsa-miR-1277-5p	hsa-miR-23a-3p	hsa-miR-367-3p	hsa-miR-4659a-3p	hsa-miR-520h	hsa-miR-6512-3p	hsa-miR-7111-3p

hsa-miR-128-3p	hsa-miR-23b-3p	hsa-miR-3672	hsa-miR-4659b-3p	hsa-miR-526b-3p	hsa-miR-6513-5p	hsa-miR-7111-5p
hsa-miR-129-5p	hsa-miR-24-3p	hsa-miR-3681-3p	hsa-miR-4667-3p	hsa-miR-526b-5p	hsa-miR-6516-5p	hsa-miR-7151-3p
hsa-miR-1304-3p	hsa-miR-2467-3p	hsa-miR-3689a-3p	hsa-miR-4668-5p	hsa-miR-532-3p	hsa-miR-660-3p	hsa-miR-7160-5p
hsa-miR-1307-3p	hsa-miR-25-3p	hsa-miR-3689b-3p	hsa-miR-4684-5p	hsa-miR-548ac	hsa-miR-661	hsa-miR-744-5p
hsa-miR-130a-3p	hsa-miR-26a-5p	hsa-miR-3689c	hsa-miR-4685-3p	hsa-miR-548ah-3p	hsa-miR-665	hsa-miR-764
hsa-miR-130b-3p	hsa-miR-26b-5p	hsa-miR-3689d	hsa-miR-4691-5p	hsa-miR-548ah-5p	hsa-miR-6720-5p	hsa-miR-765
hsa-miR-130b-5p	hsa-miR-27a-3p	hsa-miR-371a-5p	hsa-miR-4695-5p	hsa-miR-548aj-3p	hsa-miR-6727-3p	hsa-miR-766-3p
hsa-miR-132-3p	hsa-miR-27b-3p	hsa-miR-371b-5p	hsa-miR-4698	hsa-miR-548aj-5p	hsa-miR-6731-5p	hsa-miR-7703
hsa-miR-1321	hsa-miR-296-3p	hsa-miR-372-3p	hsa-miR-4701-5p	hsa-miR-548am-3p	hsa-miR-6734-3p	hsa-miR-7977
hsa-miR-142-3p	hsa-miR-29a-3p	hsa-miR-372-5p	hsa-miR-4722-3p	hsa-miR-548aq-3p	hsa-miR-6736-3p	hsa-miR-8085
hsa-miR-142-5p	hsa-miR-29b-3p	hsa-miR-373-3p	hsa-miR-4722-5p	hsa-miR-548aw	hsa-miR-6741-3p	hsa-miR-873-5p
hsa-miR-143-5p	hsa-miR-29c-3p	hsa-miR-373-5p	hsa-miR-4723-3p	hsa-miR-548az-5p	hsa-miR-6742-3p	hsa-miR-877-3p
hsa-miR-145-5p	hsa-miR-301a-3p	hsa-miR-374a-5p	hsa-miR-4728-5p	hsa-miR-548c-3p	hsa-miR-6747-3p	hsa-miR-877-5p
hsa-miR-1468-3p	hsa-miR-302a-3p	hsa-miR-374b-5p	hsa-miR-4731-5p	hsa-miR-548d-3p	hsa-miR-6749-3p	hsa-miR-887-5p
hsa-miR-148b-3p	hsa-miR-302b-3p	hsa-miR-375	hsa-miR-4739	hsa-miR-548f-5p	hsa-miR-6758-5p	hsa-miR-9-5p
hsa-miR-149-3p	hsa-miR-302c-3p	hsa-miR-377-3p	hsa-miR-4747-5p	hsa-miR-548g-5p	hsa-miR-6769b-3p	hsa-miR-92a-3p
hsa-miR-149-5p	hsa-miR-302d-3p	hsa-miR-377-5p	hsa-miR-4753-3p	hsa-miR-548h-3p	hsa-miR-6778-3p	hsa-miR-92b-3p
hsa-miR-150-5p	hsa-miR-302e	hsa-miR-378a-5p	hsa-miR-4755-3p	hsa-miR-548j-3p	hsa-miR-6779-5p	hsa-miR-93-3p
hsa-miR-153-5p	hsa-miR-30a-5p	hsa-miR-383-3p	hsa-miR-4755-5p	hsa-miR-548n	hsa-miR-6780a-5p	hsa-miR-93-5p
hsa-miR-155-5p	hsa-miR-30b-3p	hsa-miR-3913-5p	hsa-miR-4756-5p	hsa-miR-548s	hsa-miR-6785-5p	hsa-miR-939-3p
hsa-miR-15a-5p	hsa-miR-30b-5p	hsa-miR-3924	hsa-miR-4768-3p	hsa-miR-548t-5p	hsa-miR-6787-3p	hsa-miR-940
hsa-miR-15b-5p	hsa-miR-30c-1-3p	hsa-miR-3926	hsa-miR-4768-5p	hsa-miR-548x-3p	hsa-miR-6788-5p	hsa-miR-98-5p
hsa-miR-16-5p	hsa-miR-30c-2-3p	hsa-miR-3929	hsa-miR-4772-3p	hsa-miR-548x-5p	hsa-miR-6790-3p	
hsa-miR-17-5p	hsa-miR-30c-5p	hsa-miR-3934-5p	hsa-miR-4775	hsa-miR-548z	hsa-miR-6791-3p	
hsa-miR-181a-5p	hsa-miR-30d-5p	hsa-miR-3941	hsa-miR-4779	hsa-miR-5580-3p	hsa-miR-6792-3p	

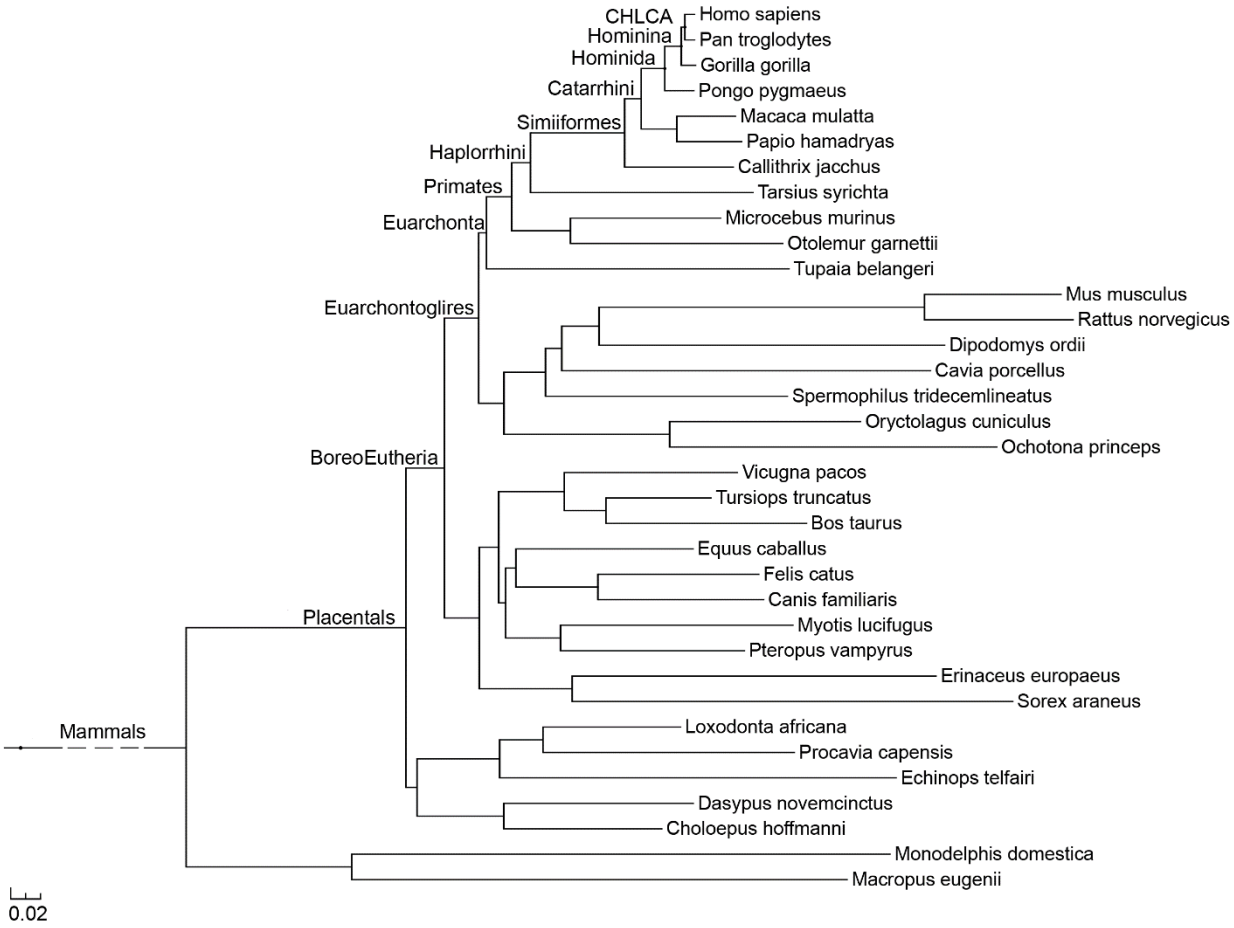


Figure SD IV-1: Mammalian species phylogenetic tree used in the study. Extracted from the UCSC genome browser.

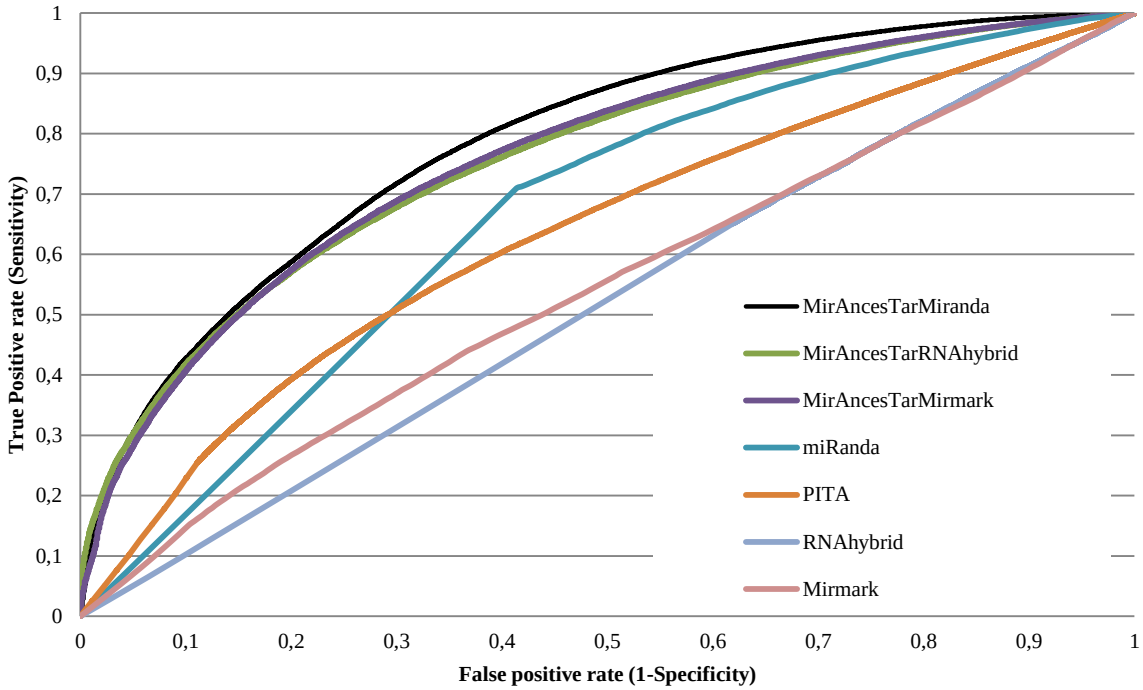


Figure SD IV-2: Receiver–operating characteristic curves of MirAncesTar classifiers.

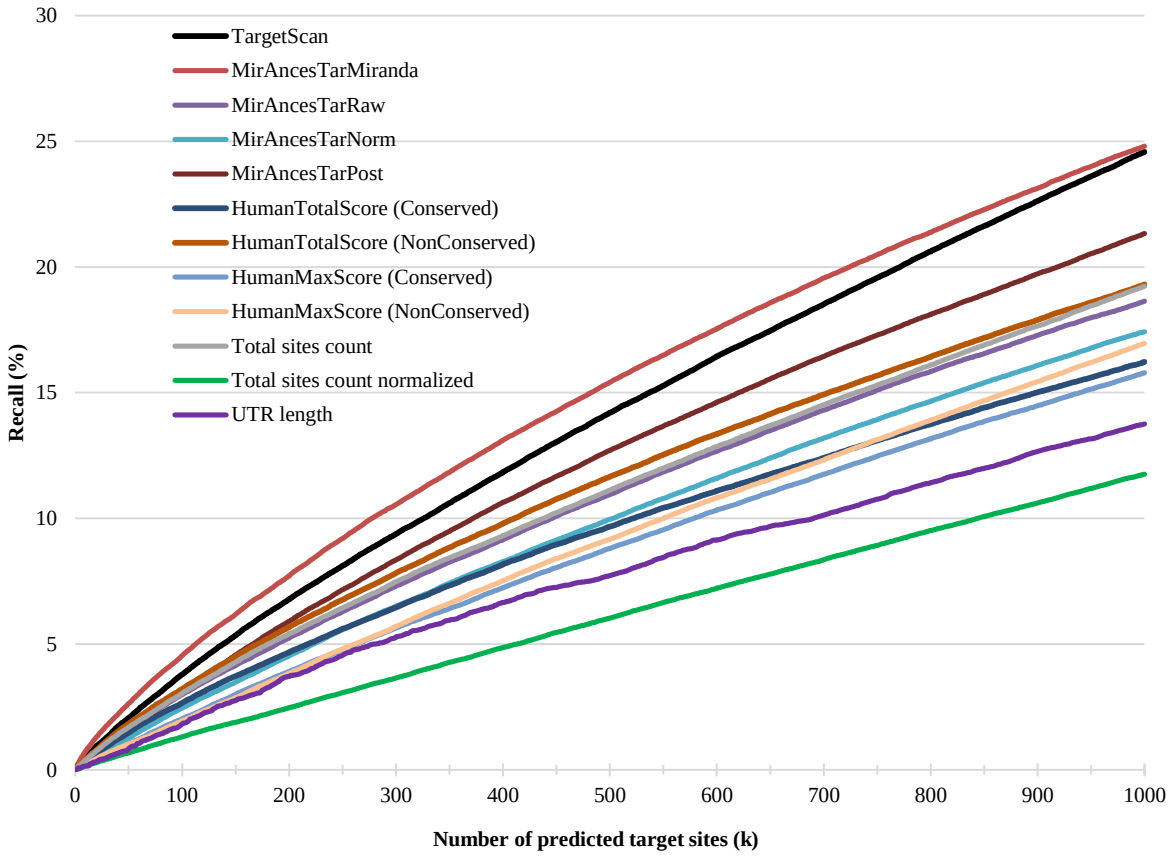


Figure SD IV-3: Recall obtained by predictors based on each of the individual features considered by MirAncesTar$_{Miranda}$, averaged across 396 miRNAs having more than 200 known targets in miRTarBase v6. Features are detailed in Methods.

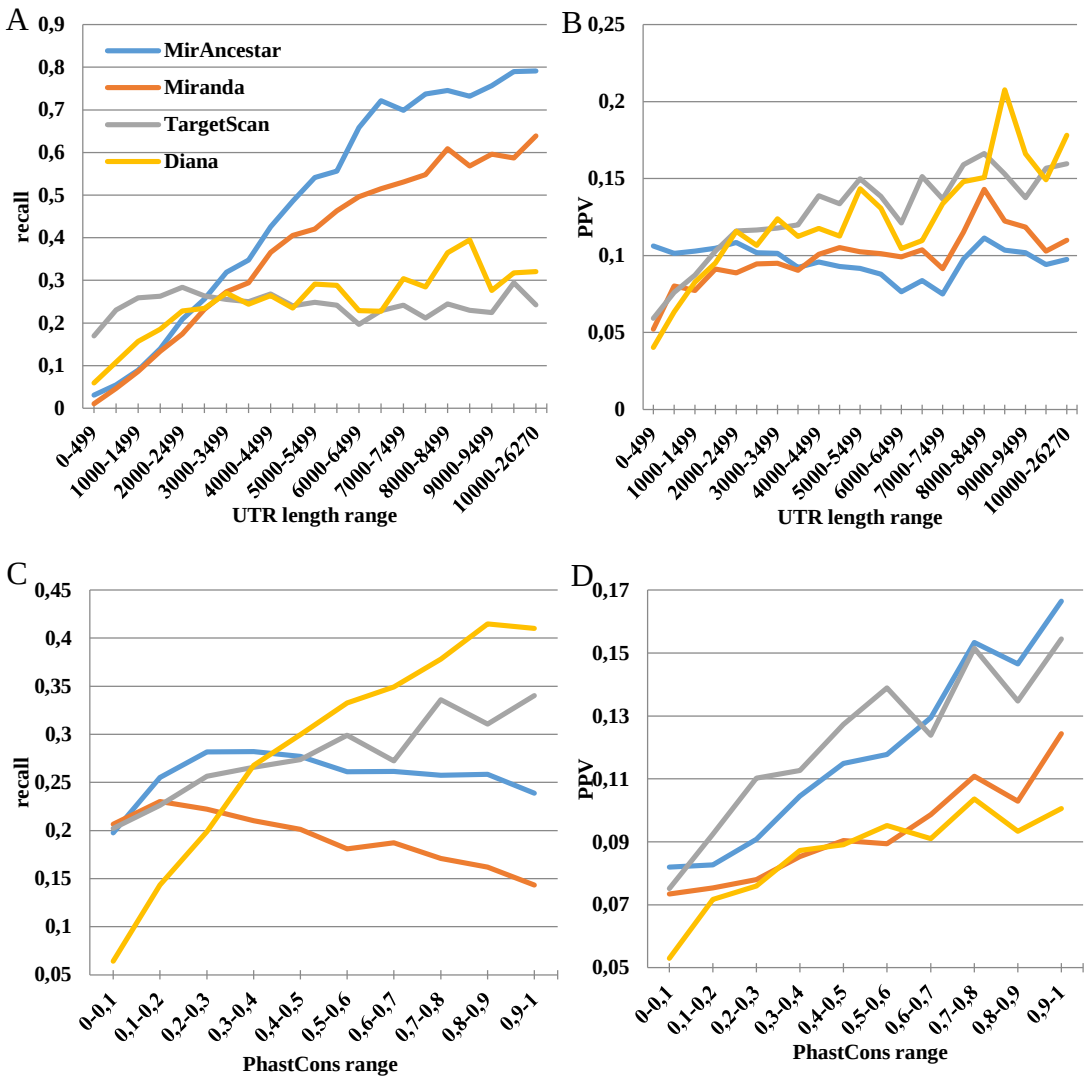


Figure SD IV-4: Precision and recall rate by UTR length and PhastCons scores for MirAncesTar, Miranda, TargetScan and Diana-microT. (A) Recall with respect to target length; (B) Precision (or PPV) with respect to target length; (C) Recall with respect to target conservation; (D) Precision (or PPV) with respect to target conservation.

CHAPTER V : CONCLUSION

5.1 Summary of Contributions

There are numerous of topics of research involving miRNAs, including biogenesis (e.g. transcription initiation, precursors folding and transport, processing of the mature miRNA), target gene silencing processes, evolution through animal and plants, differential expression analysis, implications in physiological processes and diseases, etc. This thesis focuses on the development of novel algorithms to predict the mature miRNAs sequences and their target genes, where we greatly improved the existing methods. I also focused my research on the identification of the period of origin and the mechanisms of creation of novel miRNAs in mammals, an evolution topic relatively unexplored in the literature.

In CHAPTER II , I introduce in the miRdup algorithm, a machine learning approach for the identification of the most likely position of a mature miRNA within a given miRNA precursor. Despite its importance, this is a problem that had received relatively little attention in the literature at that time. Moreover, I emphasize that the problem addressed in the paper is quite different from that solved by dozens of existing computational approaches incorrectly called miRNA predictors, which are in fact predictors of miRNA precursors, not of mature miRNAs. I believe that, in combination with existing miRNA precursor predictors, miRdup will be valuable to a large

community of users, from those interested in the de novo miRNA annotation of new genomes to those aiming to analyze short-RNA sequencing data and separate miRNAs from other short RNA species. Moreover, the program is able to support multi-loop precursors as input, and it has been designed to automatically retrain itself on all species of the most recent version of the main miRNA repository, miRbase. It contains an automatic updater and is able to train new models depending user needs. In consequence, at the opposite of almost all predictions tools, which are trained once, at publication time, miRdup will continue to improve as more and better data gets inserted in miRbase. Furthermore, beyond introducing a highly accurate predictive tool, the manuscript reports key sequence and structural features of the miRNA-miRNA* duplex that allow its recognition and shows that many of these properties differ quite significantly between clades.

CHAPTER III presents an original approach to determine the period of origin and the mechanisms of origination that led to the creation of human miRNAs in the evolution of primates. For the first time, ancestral reconstruction is exploited to unveil how many miRNAs arose in our genome. We were able to track the genetic modifications that led to become Drosha-compatible hairpins. A total of 488 primate's miRNAs have been classified into one of nine different mechanisms of origination. We found that a large proportion of miRNAs has been created by accumulation of mutations over time (De novo), insertions of non-transposable elements from distal genomic regions, and insertions of transposable elements. Our study also adds more miRNAs on those that have been already identified to result from transposable elements in other studies.

I finally present in CHAPTER IV MirAncesTar, a new tool specialized in the identification of microRNA target genes. *In silico* prediction of such genes is a challenging task, as the best existing tools have a great difficulty to assign good prediction scores to known target genes. For the first time, I propose to apply the knowledge of ancestral reconstruction to this field, which allows the tracking of targeted sites in genes along ancient genomes. MirAncesTar increases the recall of

existing miRNA target genes predictors by 26 to 157%. It implements a new algorithm that evaluates the selective pressure that maintains the predicted target site counts in evolution, boosted with machine learning. This approach allows the prediction of target sites that may be poorly conserved in terms of sequence identity due to target site turnover. Moreover, a big strength of MirAncesTar is that it can be applied to any single-sequence predictor. As these predictors become better over time, so will the accuracy of MirAncesTar. This new method will be valuable to many researchers, including those interested in improving miRNA target genes prediction tools, and those aiming to analyze miRNA-genes interaction networks. Finally, this study promotes the usage of ancestral reconstruction compared to relying on pure conservation scores, and thus provides new ideas for other genetic research fields.

5.2 Perspectives on future work

Each paper presented in the chapters II to IV has the potential to be improved. I propose in this section some directions for enhancements.

5.2.1 Mature miRNA prediction

New programs competing miRdup have already been created by other bioinformatics teams around the world and more will come, as this field of research is so dynamic. Although the dual purpose of miRdup (i.e. verification and identification of mature miRNAs), combined to its capacity to be tailored to any species and its unique auto-updating function, should make it attractive for a longer time than similar programs, it can still be improved. Inclusion of more features of interest could help increase the predictions accuracy of the machine learning predictors, This was the route taken by miRLocator (Cui et al. 2015), a recent competitor to miRdup, which was only able to improve upon miRdup on plant miRNAs. Also, in September 2015, a paper claimed to predict the position of a mature miRNA in a given precursor with a higher accuracy than miRdup. The new tool, called MatPred (Li et al. 2015a), implements interesting features that could be brought to miRdup in order to improve it. Nevertheless, the

program, which is not yet publicly available, is trained and evaluated on less than one third of human known miRNAs, and authors do not describe how they trained miRdup. A fairer comparison would have been to compare MatPred vs miRdup trained on the same set of selected human miRNAs. Finally, recently a review from Wang et al. (Wang et al. 2015) show that miRdup remained the most accurate mature miRNA predictor,,

Also, in the Summer 2015, our group released an online version of miRdup (Remita et al. 2015). This online version will make the tool more easily accessible and hopefully increase its utilization. The next step will be to propose to train a miRdup model online. Users could then choose the species on which miRdup would be trained.

5.2.1 Period of origin and mechanisms of origination of miRNAs

The proposed approach based on ancestral reconstruction to estimate the mechanisms of origination of miRNAs is something new. While we focused our research on miRNAs, the same exercise could be performed for other types of RNAs or DNA elements. Nevertheless, currently this method has limitations. We couldn't identify the origin of all human miRNAs because many predate the earliest mammalian ancestor, and many others were discarded from our analysis because we estimated that the precision of the reconstruction was insufficient. This weakness may be attenuated if we had more species genomes in the alignment used to infer ancestral states of current mammals. As the number of fully sequenced species increases, the precision of ancestral reconstructed genome will too and also our ability to accurately identify the period of origin and mechanisms of origination. Recently, a bigger alignment of 100 species has been made publicly available on the UCSC genome browser. It would be interesting to repeat this whole work on this new dataset. Also, if the understanding of mechanisms of creation of miRNA or other genetic elements increases, the same study can be performed from other alignments of various species.

Finally, this work has generated a lot of data and it would be interesting to create a website that would present all these results (e.g. ancestral states, sequence evolution of miRNA genes, blast results, etc.), to better explain the history of miRNA genes.

5.2.1 MiRNAs target gene prediction

MirAncesTar relies on ancestral genomes to boost the accuracy of existing single-sequence target site predictors. It provides very good results but requires heavy computing resources, mainly because predictions of a miRNA target gene predictor must be computed over all ancestors and orthologous species of choice. Also, if a user wants to try other predictors than those tested in our paper, he will need to code a parser to extract information from of the predictions output, then execute MirAncesTar to calculate the features that have to be submitted to a logistic regression classifier. We provide in the user manual of mirAncesTar all guidance to go through these steps. However, it would be conceivable to develop a program that automates parsing and learning tasks to make it more user-friendly. For example, for the second step, one would create an automatic parser where a user could provide or select the column IDs used by MirAncesTar. Then, it would be possible to develop a similar tool such as miRdup, which supports auto-learning, adapted to handle a much larger amount of input data.

MirAncesTar has place to improvement. More features could be calculated to increase the prediction accuracy, such as considering target site positions within the UTR, or maybe considering the possibility that a target for specific miRNA could be replaced by one for another miRNA. In other words, the target site would still exist, but for another miRNA. In that case, we would need to use more than one miRNA at a time. Also, to make MirAncesTar more visible, a web server will be developed. Ideally, we would like to create a platform offering the possibility to get the target genes of a given miRNA and vice versa. For each interaction between a miRNA and its target gene, we would provide various types of information supporting the prediction (i.e. prediction scores, experimental validation if exists, etc.) and a picture of aligned putative target

sites in every orthologs and ancestors, similar to the Figure IV-5, showing the cases under turnover.

Finally, currently, the large majority of target genes predictors, including MirAncesTar, restrain the research of targeted sites within genes' UTRs. Nevertheless, in recent years, there has been increasing evidence that miRNAs also bind in the genes' coding regions (CDS) (Hausser et al. 2013b). If these genes' regions are taken in account, new challenges will arise for *in silico* research of putative miRNAs target genes, especially because the targeted sequence length to analyze will increase, and also because since CDS are generally much more conserved than UTRs, it will lead to a higher rate of false positive predictions. Moreover, calculation times are known to be heavy in the research of target genes, thus there is place to improving algorithms that would reduce execution time.

5.2.1 Other advances

Although I have not focused my research in this thesis on pre-miRNAs prediction, we rely on these predictions for miRdup and for the identification of miRNAs' periods of origin. Indeed, miRdup is a post-processing tool that predicts and validates the position of mature miRNAs on predicted or experimentally validated pre-miRNAs, and we were able to predict the period of origin of miRNAs by predicting pre-miRNAs in ancestral genomes. Prediction of miRNA precursors is relatively accurate, but recent discoveries in the miRNA biogenesis could bring further improvements in the mature and pre-miRNA prediction field. It has been shown in a recent study that Drosha serves as a "ruler" by measuring about 11 bp from the basal ssRNA-dsRNA junction (Nguyen et al. 2015) and approximately 22 bp away from the 'apical' junction linked to the terminal loop (Ha and Kim 2014). Also, Drosha and DGCR8, respectively, recognize the basal UG and apical UGU motifs, which ensure proper orientation of the complex (Figure V-1). No existing predictor take in account the pri-miRNA. This sequence can be retrieved from the human genome, based on pre-miRNA coordinates. Starting from

that molecule instead of the pre-miRNA and using the features recently discovered could lead to a better prediction of miRNAs molecules.

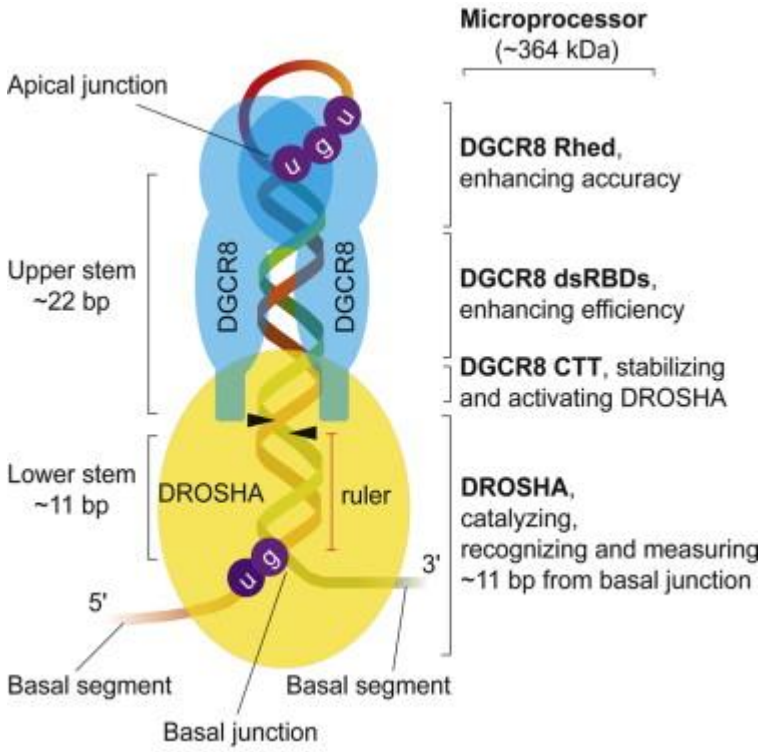


Figure V-1: Model of the functional Microprocessor on a pri-miRNA molecule (Nguyen et al. 2015).

Another work that would be useful for miRNA research is the creation of a tool that transcribes old miRNA names to the recent nomenclature. Since its creation, the main miRNA registry, miRbase, has adapted its nomenclature to novel discoveries. Thus, the comparison is often laborious when the results of old miRNA-related programs have to be tested and compared to new tools. The correspondence tool, ideally online, would offer name translation of input miRNA names between miRbase versions. The user would select the version of input miRNA names and a destination version. The implementation would consist of gathering all miRbase databases versions, and make

the correspondence between miRNA names based on their accession ID, which hasn’t changed since miRbase creation.

In conclusion, miRNAs research is a young field that has started 15 years ago. Many discoveries helped to understand their implication in living beings and revolutionized the way we think about non-coding parts of the genome (often been referred to as "junk DNA") and post-transcriptional regulation, but a lot of work remains to do. For example, we do not clearly comprehend exactly how Dicer select the mature miRNA on a pre-miRNA hairpin, or how mature miRNAs select their target sites. Once these processes will be much more understood, and they will, better prediction models will arise. Finally, I believe that a great therapeutic potential reside in miRNAs, especially in the form of cancer biomarkers. MiRNA-based diagnostics have already reached the clinic in laboratory-developed tests. Maybe one day, to cure some diseases, synthesized miRNAs will be used in next-generation drugs to artificially restore the regulation of specific genes. The miRNAs’ story is just beginning!

REFERENCES

Agarwal V, Bell GW, Nam J-W, Bartel DP. 2015. Predicting effective microRNA target sites in mammalian mRNAs. *Elife* **4**: e05005.

Agharbaoui Z, Leclercq M, Remita MA, Badawi MA, Lord E, Houde M, Danyluk J, Diallo AB, Sarhan F. 2015. An integrative approach to identify hexaploid wheat miRNAome associated with development and tolerance to abiotic stress. *BMC Genomics* **16**: 339.

Agrawal N, Dasaradhi PVN, Mohmmed A, Malhotra P, Bhatnagar RK, Mukherjee SK. 2003. RNA interference: biology, mechanism, and applications. *Microbiol Mol Biol Rev* **67**: 657–685.

Alberts B. 2002. *Molecular biology of the cell*. 4th ed. Garland Science, New York.

Alexander RP, Fang G, Rozowsky J, Snyder M, Gerstein MB. 2010. Annotating non-coding regions of the genome. *Nat Rev Genet* **11**: 559–571.

Alexiou P, Maragkakis M, Papadopoulos GL, Simmosis VA, Zhang L, Hatzigeorgiou AG. 2010. The DIANA-mirExTra web server: From gene expression data to microRNA function. *PLoS One* **5**.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215**: 403–410.

Ambros V. 1989. A hierarchy of regulatory genes controls a larva-to-adult developmental switch in C. elegans. *Cell* **57**: 49–57.

Ambros V. 2003. A uniform system for microRNA annotation. *RNA* **9**: 277–279.

Ambros V. 2001. microRNAs: tiny regulators with great potential. *Cell* **107**: 823–826.

Ambros V. 2004. The functions of animal microRNAs. *Nature* **431**: 350–355.

Ameres SL, Martinez J, Schroeder R. 2007. Molecular Basis for Target RNA Recognition and Cleavage by Human RISC. *Cell* **130**: 101–112.

An J, Lai J, Lehman ML, Nelson CC. 2013. miRDeep*: an integrated application tool for miRNA identification from RNA sequencing data. *Nucleic Acids Res* **41**: 727–37.

Andorfer CA, Necela BM, Thompson EA, Perez EA. 2011. MicroRNA signatures: Clinical biomarkers for the diagnosis and treatment of breast cancer. *Trends Mol Med* **17**: 313–319.

Bahadur B, Venkat Rajam M, Sahijram L, Krishnamurthy K V., eds. 2015. *Plant Biology and Biotechnology*. Springer India, New Delhi.

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

Barh D, Carpi A, Verma M, Gunduz M. 2014. Cancer Biomarkers: Non-Invasive Early Diagnosis and Prognosis.

Bartel DP. 2009. MicroRNAs: target recognition and regulatory functions. *Cell* **136**: 215–33.

Batuwita R, Palade V. 2009. microPred: effective classification of pre-miRNAs for human miRNA gene prediction. *Bioinformatics* **25**: 989–995.

Bentwich I, Avniel A, Karov Y, Aharonov R, Gilad S, Barad O, Barzilai A, Einat P, Einav U, Meiri E, et al. 2005. Identification of hundreds of conserved and nonconserved human microRNAs. *Nat Genet* **37**: 766–70.

Berezikov E. 2011. Evolution of microRNA diversity and regulation in animals. *Nat Rev Genet* **12**: 846–60.

Berezikov E, Cuppen E, Plasterk RH. 2006a. Approaches to microRNA discovery. *Nat Genet* **38 Suppl**: S2–7.

Berezikov E, Guryev V, Van De Belt J, Wienholds E, Plasterk RHA, Cuppen E. 2005. Phylogenetic shadowing and computational identification of human microRNA genes. *Cell* **120**: 21–24.

Berezikov E, Van Tetering G, Verheul M, Van De Belt J, Van Laake L, Vos J, Verloop R, Van De Wetering M, Guryev V, Takada S, et al. 2006b. Many novel mammalian microRNA candidates identified by extensive cloning and RAKE analysis. *Genome Res* **16**: 1289–1298.

Bernstein E, Kim SY, Carmell MA, Murchison EP, Alcorn H, Li MZ, Mills AA, Elledge SJ, Anderson K V, Hannon GJ. 2003. Dicer is essential for mouse development. *Nat Genet* **35**: 215–217.

Betel D, Wilson M, Gabow A, Marks DS, Sander C. 2008. The microRNA.org resource: targets and expression. *Nucleic Acids Res* **36**: D149–53.

Birnstiel ML, Schaufele FJ. 1988. Structure and function of minor snRNPs. *Struct Funct Major Minor Small Nucl Ribonucleoprotein Part ML Birnstiel, Ed Springer-Verlag New York Inc, New York* **155**: 182.

Blanchette M. 2012. Exploiting ancestral mammalian genomes for the prediction of human transcription factor binding sites. *BMC Bioinformatics* **13 Suppl 1**: S2.

Blanchette M, Diallo AB, Green ED, Miller W, Haussler D. 2008. Computational reconstruction of ancestral DNA sequences. *Methods Mol Biol* **422**: 171–84.

Blanchette M, Green ED, Miller W, Haussler D. 2004a. Reconstructing large regions of an ancestral mammalian genome in silico. *Genome Res* **14**: 2412–23.

Blanchette M, Kent WJ, Riemer C, Elnitski L, Smith AFA, Roskin KM, Baertsch R,

Rosenbloom K, Clawson H, Green ED, et al. 2004b. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* **14**: 708–15.

Bonnet E, He Y, Billiau K, de Peer Y, Van de Peer Y. 2010. TAPIR, a web server for the prediction of plant microRNA targets, including target mimics. *Bioinformatics* **26**: 1566.

Borchert GM, Lanier W, Davidson BL. 2006. RNA polymerase III transcribes human microRNAs. *Nat Struct Mol Biol* **13**: 1097–101.

Boualem A, Laporte P, Jovanovic M, Laffont C, Plet J, Combier JP, Niebel A, Crespi M, Frugier F. 2008. MicroRNA166 controls root and nodule development in Medicago truncatula. *Plant J* **54**: 876–887.

Breakfield NW, Corcoran DL, Petricka JJ, Shen J, Sae-Seaw J, Rubio-Somoza I, Weigel D, Ohler U, Benfey PN. 2012. High-resolution experimental and computational profiling of tissue-specific known and novel miRNAs in Arabidopsis. *Genome Res* **22**: 163–176.

Breiman L. 2001. Random forests. *Mach Learn* **45**: 5–32.

Brennecke J, Hipfner DR, Stark A, Russell RB, Cohen SM. 2003. bantam Encodes a Developmentally Regulated microRNA that Controls Cell Proliferation and Regulates the Proapoptotic Gene hid in Drosophila. *Cell* **113**: 25–36.

Brenner S, Jacob F, Meselson M. 1961. An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature* **190**: 576–581.

Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo S, McCurdy S, Foy M, Ewan M, et al. 2000. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat Biotechnol* **18**: 630–4.

Britten RJ. 1996. DNA sequence insertion and evolutionary variation in gene regulation. *Proc Natl Acad Sci U S A* **93**: 9374–7.

Brünger a T, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, et al. 1998. Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr D Biol Crystallogr* **54**: 905–21.

Bucher E, Reinders J, Mirouze M. 2012. Epigenetic control of transposon transcription and mobility in Arabidopsis. *Curr Opin Plant Biol* **15**: 503–10.

Calvin S. 2013. *XAFS for Everyone*. CRC Press.

Carrington JC, Ambros V. 2003. Role of microRNAs in plant and animal development : Developmental timing. *Science (80-)* **301**: 336–338.

Castanotto D, Rossi JJ. 2009. The promises and pitfalls of RNA-interference-based therapeutics. *Nature* **457**: 426–433.

Cessie S Le, Houwelingen JC Van, Society RS, Le Cessie S, Van Houwelingen JC. 1992. Ridge estimators in logistic regression. *Appl Stat* **41**: 191–201.

Chang CC, Lin CJ. 2001. LIBSVM: a library for support vector machines. Software. *Comput Sci Inf Eng*.

Chang T-C, Mendell JT. 2007. microRNAs in vertebrate physiology and human disease. *Annu Rev Genomics Hum Genet* **8**: 215–239.

Chauve C, Tannier E. 2008. A methodological framework for the reconstruction of contiguous regions of ancestral genomes and its application to mammalian genomes. *PLoS Comput Biol* **4**: e1000234.

Chen JF, Mandel EM, Thomson JM, Wu Q, Callis TE, Hammond SM, Conlon FL, Wang DZ. 2006. The role of microRNA-1 and microRNA-133 in skeletal muscle proliferation and differentiation. *Nat Genet* **38**: 228.

Chen JJ-L, Chen J-L, Greider CW, Greider CW. 2005. Functional analysis of the pseudoknot structure in human telomerase RNA. *Proc Natl Acad Sci U S A* **102**: 8080–5; discussion 8077–9.

Chen L, Zhang Y, Ren Y, Xu J, Zhang Z, Wang Y. 2012. Genome-wide identification of cold-responsive and new microRNAs in Populus tomentosa by high-throughput sequencing. *Biochem Biophys Res Commun* **417**: 892–896.

Chen X. 2004. A microRNA as a translational repressor of APETALA2 in Arabidopsis flower development. *Science (80-)* **303**: 2022.

Chen X, Ba Y, Ma L, Cai X, Yin Y, Wang K, Guo J, Zhang Y, Chen J, Guo X, et al. 2008. Characterization of microRNAs in serum: a novel class of biomarkers for diagnosis of cancer and other diseases. *Cell Res* **18**: 997–1006.

Chiang HR, Schoenfeld LW, Ruby JG, Auyeung VC, Spies N, Baek D, Johnston WK, Russ C, Luo S, Babiarz JE, et al. 2010. Mammalian microRNAs: Experimental evaluation of novel and previously annotated genes. *Genes Dev* **24**: 992–1009.

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**: 1084–8.

Clancy S, Shaw KM. 2008. DNA Deletion and Duplication and the Associated Genetic Disorders. *Nat Educ* **1**.

Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow T a, Kaufman TC, Kellis M, Gelbart W, Iyer VN, et al. 2007. Evolution of genes and genomes on the Drosophila phylogeny. *Nature* **450**: 203–18.

Clop A, Marcq F, Takeda H, Pirottin D, Tordoir X, Bibé B, Bouix J, Caiment F, Elsen JM, Eychenne F, et al. 2006. A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep. *Nat Genet* **38**: 813–818.

Cohen WW. 1995. Fast Effective Rule Induction. *Proc Twelfth Int Conf Mach Learn Lake Tahoe, Calif*.

Colas AR, McKeithan WL, Cunningham TJ, Bushway PJ, Garmire LX, Duester G, Subramaniam S, Mercola M. 2012. Whole-genome microRNA screening identifies let-7 and mir-18 as regulators of germ layer formation during early embryogenesis. *Genes Dev* **26**: 2567–79.

Collobert R, Bengio S. 2004. Links between perceptrons, MLPs and SVMs. *Proc twenty-first Int Conf Mach Learn* 23.

Cooper TA, Wan L, Dreyfuss G. 2009. RNA and disease. *Cell* **136**: 777–793.

Cortes C, Vapnik V. 1995. Support-Vector Networks. *Mach Learn* **20**: 273–297.

Costa FF. 2008. Non-coding RNAs, epigenetics and complexity. *Gene* **410**: 9–17.

Creemers EE, Tijsen AJ, Pinto YM. 2012. Circulating MicroRNAs: Novel biomarkers and extracellular communicators in cardiovascular disease? *Circ Res* **110**: 483–495.

Crick FH. 1966. Codon--anticodon pairing: the wobble hypothesis. *J Mol Biol* **19**: 548–555.

Cuellar TL, McManus MT. 2005. MicroRNAs and endocrine biology. *J Endocrinol* **187**: 327–332.

Cui H, Zhai J, Ma C. 2015. miRLocator: Machine Learning-Based Prediction of Mature MicroRNAs within Plant Pre-miRNA Sequences. *PLoS One* **10**: e0142753.

Cuperus JT, Fahlgren N, Carrington JC. 2011. Evolution and functional diversification of MIRNA genes. *Plant Cell* **23**: 431–442.

Dai X, Zhao PX. 2011. psRNATarget: a plant small RNA target analysis server. *Nucleic Acids Res* **39**: W155.

Dangwal S, Bang C, Thum T. 2012. Novel techniques and targets in cardiovascular microRNA research. *Cardiovasc Res* **93**: 545–554.

Dardel F, Kapas F. 2002. *Bioinformatique : genomique et post-genomique*. Editions de l'ecole polytechnique, Palaiseau [France].

Darwin C. 1859. Origin of Species. *Library (Lond)* **475**: 424.

Dash M, Liu H. 1997. Feature selection for classification. *Intell data Anal* **1**: 131–156.

Denman RB. 1993. Using RNAFOLD to predict the activity of small catalytic RNAs. *Biotechniques* **15**: 1090–1094.

Dermitzakis ET, Clark AG. 2002. Evolution of Transcription Factor Binding Sites in Mammalian Gene Regulatory Regions: Conservation and Turnover. *Mol Biol Evol* **19**: 1114–1121.

Diallo AB, Makarenkov V, Blanchette M. 2010. Ancestors 1.0: a web server for ancestral sequence reconstruction. *Bioinformatics* **26**: 130–1.

Didiano D, Hobert O. 2006. Perfect seed pairing is not a generally reliable predictor for miRNA-target interactions. *Nat Struct Mol Biol* **13**: 849–851.

Ding J, Zhou S, Guan J. 2010. MiRenSVM: towards better prediction of microRNA precursors using an ensemble SVM classifier with multi-loop features. *BMC Bioinformatics* **11**: S11.

Dirks RM, Lin M, Winfree E, Pierce NA. 2004. Paradigms for computational nucleic acid design. *Nucleic Acids Res* **32**: 1392–1403.

Doshi KJ, Cannone JJ, Cobaugh CW, Gutell RR. 2004. Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. *BMC Bioinformatics* **5**: 105.

Dumay-Odelot H, Durrieu-Gaillard S, Da Silva D, Roeder RG, Teichmann M. 2010. Cell

growth- and differentiation-dependent regulation of RNA polymerase III transcription. *Cell Cycle* **9**: 3687–3699.

Durbin R, Eddy S, Krogh A, Mitchison G. 1998. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*.

Dweep H, Sticht C, Pandey P, Gretz N. 2011. MiRWalk - Database: Prediction of possible miRNA binding sites by " walking" the genes of three genomes. *J Biomed Inform* **44**: 839–847.

Eddy SR, Durbin R. 1994. RNA sequence analysis using covariance models. *Nucleic Acids Res* **22**: 2079.

Ekimler S, Sahin K. 2014. Computational Methods for MicroRNA Target Prediction. *Genes (Basel)* **5**: 671–83.

Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS. 2003. MicroRNA targets in Drosophila. *Genome Biol* **5**: R1.

Esteller M. 2011. Non-coding RNAs in human disease. *Nat Rev Genet* **12**: 861–874.

Fahlgren N, Howell MD, Kasschau KD, Chapman EJ, Sullivan CM, Cumbie JS, Givan SA, Law TF, Grant SR, Dangl JL, et al. 2007. High-throughput sequencing of Arabidopsis microRNAs: Evidence for frequent birth and death of MIRNA genes. *PLoS One* **2**.

Fahlgren N, Jogdeo S, Kasschau KD, Sullivan CM, Chapman EJ, Laubinger S, Smith LM, Dasenko M, Givan SA, Weigel D, et al. 2010. MicroRNA gene evolution in Arabidopsis lyrata and Arabidopsis thaliana. *Plant Cell* **22**: 1074–1089.

Farazi T a, Ten Hoeve JJ, Brown M, Mihailovic A, Horlings HM, van de Vijver MJ, Tuschl T, Wessels LF. 2014. Identification of distinct miRNA target regulation between breast cancer molecular subtypes using AGO2-PAR-CLIP and patient datasets. *Genome Biol* **15**: R9.

Farh KK-H, Grimson A, Jan C, Lewis BP, Johnston WK, Lim LP, Burge CB, Bartel DP. 2005. The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. *Science* **310**: 1817–21.

Fasanaro P, Romani S, Voellenkle C, Maimone B, Capogrossi MC, Martelli F. 2012. ROD1 Is a Seedless Target Gene of Hypoxia-Induced miR-210. *PLoS One* **7**.

Fatica A, Bozzoni I. 2014. Long non-coding RNAs: new players in cell differentiation and development. *Nat Rev Genet* **15**: 7–21.

Filipowicz W, Bhattacharyya SN, Sonenberg N. 2008. Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nat Rev Genet* **9**: 102–114.

Freund Y, Schapire R. 1995. A desicion-theoretic generalization of on-line learning and an application to boosting. *Comput Learn theory* 23–37.

Friedlander MR, Chen W, Adamidi C, Maaskola J, Einspanier R, Knespel S, Rajewsky N. 2008. Discovering microRNAs from deep sequencing data using miRDeep. *Nat Biotechnol* **26**: 407–415.

Friedländer MR, Chen W, Adamidi C, Maaskola J, Einspanier R, Knespel S, Rajewsky N.

2008. Discovering microRNAs from deep sequencing data using miRDeep. *Nat Biotechnol* **26**: 407–15.

Friedlander MR, Mackowiak SD, Li N, Chen W, Rajewsky N. 2012. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res* **40**: 37–52.

Friedman RC, Farh KK-HH, Burge CB, Bartel DP. 2009. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res* **19**: 92–105.

Fujii H, Chiou TJ, Lin SI, Aung K, Zhu JK. 2005. A miRNA involved in phosphate-starvation response in Arabidopsis. *Curr Biol* **15**: 2038–2043.

Fujiwara T, Yada T. 2013. miRNA-target prediction based on transcriptional regulation. *BMC Genomics* **14 Suppl 2**: S3.

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 lsy-6 and other microRNAs. *Nat Struct Mol Biol* **18**: 1139–1146.

Gennarino VA, D'Angelo G, Dharmalingam G, Fernandez S, Russolillo G, Sanges R, Mutarelli M, Belcastro V, Ballabio A, Verde P, et al. 2012. Identification of microRNA-regulated gene networks by expression analysis of target genes. *Genome Res* **22**: 1163–72.

Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, Korbel JO, Emanuelsson O, Zhang ZD, Weissman S, Snyder M. 2007. What is a gene, post-ENCODE? History and updated definition. *Genome Res* **17**: 669–681.

Gkirtzou K, Tsamardinos I, Tsakalides P, Poirazi P. 2010. MatureBayes: A Probabilistic Algorithm for Identifying the Mature miRNA within Novel Precursors. *PLoS One* **5**: e11843.

Gontan C, Jonkers I, Gribnau J. 2011. Long Noncoding RNAs and X Chromosome Inactivation. *Prog Mol Subcell Biol* **51**: 43–64.

Goodall EF, Heath PR, Bandmann O, Kirby J, Shaw PJ. 2013. Neuronal dark matter: the emerging role of microRNAs in neurodegeneration. *Front Cell Neurosci* **7**: 178.

Goodfellow SJ, White RJ. 2007. Regulation of RNA polymerase III transcription during mammalian cell growth. *Cell Cycle* **6**: 2323–2326.

Grey F, Antoniewicz A, Allen E, Saugstad J, McShea A, Carrington JC, Nelson J. 2005. Identification and characterization of human cytomegalovirus-encoded microRNAs. *J Virol* **79**: 12095.

Griffiths-Jones S. 2004. The microRNA registry. *Nucleic Acids Res* **32**: D109.

Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ. 2006. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res* **34**: D140.

Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A. 2005. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res* **33**: D121.

Griffiths-Jones S, Saini HK, Van Dongen S, Enright AJ. 2008. miRBase: tools for microRNA genomics. *Nucleic Acids Res* **36**: D154.

Grimson A, Farh KK-H, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP. 2007. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell* **27**: 91–105.

Gros F, Hiatt H, Gilbert W, Kurland CG, Risebrough RW, Watson JD. 1961. Unstable Ribonucleic Acid Revealed by Pulse Labelling of Escherichia Coli. *Nature* **190**: 581–585.

Grosswendt S, Filipchyk A, Manzano M, Klironomos F, Schilling M, Herzog M, Gottwein E, Rajewsky N. 2014. Unambiguous Identification of miRNA: Target site interactions by different types of ligation reactions. *Mol Cell* **54**: 1042–1054.

Grundhoff A, Sullivan CS, Ganem D. 2006. A combined computational and microarray-based approach identifies novel microRNAs encoded by human gamma-herpesviruses. *RNA* **12**: 733–750.

Guan D, Zhang W, Liu G-H, Belmonte JCI. 2013. Switching cell fate, ncRNAs coming to play. *Cell Death Dis* **4**: e464.

Guan D-G, Liao J-Y, Qu Z-H, Zhang Y, Qu L-H. 2011. mirExplorer: Detecting microRNAs from genome and next generation sequencing data using the AdaBoost method with transition probability matrix and combined features. *RNA Biol* **8**: 922–934.

Guo H, Ingolia NT, Weissman JS, Bartel DP. 2010. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* **466**: 835–840.

Guy CL. 1990. Cold Accelimation and Freezing Stress Tolerance: Role of Protein Metabolism. *Annu Rev Plant Biol* **41**: 187–223.

Guyon I, Weston J, Barnhill S, Vapnik V. 2002. Gene selection for cancer classification using support vector machines. *Mach Learn* **46**: 389–422.

Ha M, Kim VN. 2014. Regulation of microRNA biogenesis. *Nat Rev Mol Cell Biol* **15**: 509–524.

Hackenberg M, Rodriguez-Ezpeleta N, Aransay AM. 2011. miRanalyzer: an update on the detection and analysis of microRNAs in high-throughput sequencing experiments. *Nucleic Acids Res* **39**: W132–8.

Hackenberg M, Sturm M, Langenberger D, Falcon-Perez JM, Aransay AM. 2009. miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments. *Nucleic Acids Res* **37**: W68.

Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, Rothballer A, Ascano M, Jungkamp AC, Munschauer M, et al. 2010. Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP. *Cell* **141**: 129–141.

Hajdin CE, Bellaousov S, Huggins W, Leonard CW, Mathews DH, Weeks KM. 2013. Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. *Proc Natl Acad Sci U S A* **110**: 5498–503.

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. 2009. The WEKA data mining software: an update. *ACM SIGKDD Explor Newsl* **11**: 10–18.

Hamilton AJ, Baulcombe DC. 1999. A species of small antisense RNA in posttranscriptional gene silencing in plants. *Science (80-)* **286**: 950.

Hammond SM. 2015. An overview of microRNAs. *Adv Drug Deliv Rev*.

Hamzeiy H, Allmer J, Yousef M. 2014. Computational Methods for MicroRNA Target Prediction. *Methods Mol Biol* **1107**: 207–21.

Hansen TB, Veno MT, Kjems J, Damgaard CK. 2014. miRdentify: high stringency miRNA predictor identifies several novel animal miRNAs. *Nucleic Acids Res* 1–11.

Harada M, Luo X, Murohara T, Yang B, Dobrev D, Nattel S. 2014. MicroRNA regulation and cardiac calcium signaling: Role in cardiac disease and therapeutic potential. *Circ Res* **114**: 689–705.

Haudry A, Platts AE, Vello E, Hoen DR, Leclercq M, Williamson RJ, Forczek E, Joly-Lopez Z, Steffen JG, Hazzouri KM, et al. 2013. An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. *Nat Genet* **45**: 891–8.

Hausser J, Syed AP, Bilen B, Zavolan M. 2013a. Analysis of CDS-located miRNA target sites suggests that they can effectively inhibit translation. *Genome Res* **23**: 604–615.

Hausser J, Syed AP, Bilen B, Zavolan M. 2013b. Analysis of CDS-located miRNA target sites suggests that they can effectively inhibit translation. *Genome Res* **23**: 604–15.

He C, Li YX, Zhang G, Gu Z, Yang R, Li J, Lu ZJ, Zhou ZH, Zhang C, Wang J. 2012. MiRmat: mature microRNA sequence prediction. *PLoS One* **7**: e51673.

Heimberg AM, Cowper-Sal-lari R, Sémon M, Donoghue PCJ, Peterson KJ. 2010. microRNAs reveal the interrelationships of hagfish, lampreys, and gnathostomes and the nature of the ancestral vertebrate. *Proc Natl Acad Sci U S A* **107**: 19379–83.

Hellweg P. 2012. *The American Heritage Dictionary, Fifth Edition: Office Edition*. Houghton Mifflin Harcourt Publishing Company.

Hendrix D, Levine M, Shi W. 2010. miRTRAP, a computational method for the systematic identification of miRNAs from high throughput sequencing data. *Genome Biol* **11**: R39.

Hertel J, Hofacker IL, Stadler PF. 2008. SnoReport: computational identification of snoRNAs with unknown targets. *Bioinformatics* **24**: 158–64.

Hertel J, Lindemeyer M, Missal K, Fried C, Tanzer A, Flamm C, Hofacker IL, Stadler PF. 2006. The expansion of the metazoan microRNA repertoire. *BMC Genomics* **7**: 25.

Hertel J, Stadler PF. 2006. Hairpins in a Haystack: Recognizing microRNA precursors in comparative genomics data. In *Bioinformatics*, Vol. 22 of.

Higashi S, Fournier C, Gautier C, Gaspin C, Sagot M-F. 2015. Mirinho: An efficient and general plant and animal pre-miRNA predictor for genomic and deep sequencing data. *BMC Bioinformatics* **16**: 179.

Hofacker IL. 2004. RNA secondary structure analysis using the Vienna RNA package. *Curr Protoc Bioinforma*.

Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P. 1994. Fast

folding and comparison of RNA secondary structures. *Monatshefte f{\"u}r Chemie/Chemical Mon* **125**: 167–188.

Hong SJ. 1997. Use of contextual information for feature ranking and discretization. **9**: 718–730.

Hsu S Da, Lin FM, Wu WY, Liang C, Huang WC, Chan WL, Tsai WT, Chen GZ, Lee CJ, Chiu CM, et al. 2011. MiRTarBase: A database curates experimentally validated microRNA-target interactions. *Nucleic Acids Res* **39**.

Hu LL, Huang Y, Wang QC, Zou Q, Jiang Y. 2012. Benchmark comparison of ab initio microRNA identification methods and software. *Genet Mol Res* **11**: 4525–38.

Huang TH, Fan B, Rothschild M, Hu ZL, Li K, Zhao SH. 2007. MiRFinder: an improved approach and software implementation for genome-wide fast microRNA precursor scans. *BMC Bioinformatics* **8**: 341.

Huntzinger E, Izaurralde E. 2011. Gene silencing by microRNAs: contributions of translational repression and mRNA decay. *Nat Rev Genet* **12**: 99–110.

Iwama H, Kato K, Imachi H, Murao K, Masaki T. 2013. Human microRNAs originated from two periods at accelerated rates in mammalian evolution. *Mol Biol Evol* **30**: 613–26.

Jacob F, Monod J. 1961. Genetic Regulatory Mechanisms in the Syntheis of Proteins. *J Mol Biol* **3**: 318–356.

Jacobsen A, Krogh A, Kauppinen S, Lindow M. 2010. miRMaid: a unified programming interface for microRNA data resources. *BMC Bioinformatics* **11**: 29.

Jiang P, Wu H, Wang W, Ma W, Sun X, Lu Z. 2007. MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Res* **35**: W339–44.

Jones-Rhoades MW. 2010. Prediction of plant miRNA genes. *Methods Mol Biol (Clifton, NJ)* **592**: 19.

Jones-Rhoades MW, Bartel DP. 2004. Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. *Mol Cell* **14**: 787–799.

Jordan IK, Rogozin IB, Glazko G V, Koonin E V. 2003. Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet* **19**: 68–72.

Kadri S, Hinman V, Benos P V. 2009. HHMMiR: efficient de novo prediction of microRNAs using hierarchical hidden Markov models. *BMC Bioinformatics* **10**: S35.

Kaikkonen MU, Lam MTY, Glass CK. 2011. Non-coding RNAs as regulators of gene expression and epigenetics. *Cardiovasc Res* **90**: 430–440.

Kane NM, Thrasher AJ, Angelini GD, Emanueli C. 2014. Concise Review: MicroRNAs as Modulators of Stem Cells and Angiogenesis. *Stem Cells* **32**: 1059–66.

Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermüller J, Hofacker IL, et al. 2007. RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316**: 1484–1488.

Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, Bourque G, Yandell M,

Feschotte C. 2013. Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS Genet* **9**: e1003470.

Karginov F V, Cheloufi S, Chong MMW, 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.

Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. 2004. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* **32**: D493–6.

Kedde M, Agami R. 2008. Interplay between microRNAs and RNA-binding proteins determines developmental processes. *Cell Cycle* **7**: 899–903.

Keiler KC. 2008. Biology of trans-translation. *Annu Rev Microbiol* **62**: 133–151.

Keiler KC, Ramadoss NS. 2011. Bifunctional transfer-messenger RNA. *Biochimie* **93**: 1993–1997.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome Res* **12**: 996–1006.

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.

Khraiwesh B, Zhu J-K, Zhu J. 2012. Role of miRNAs and siRNAs in biotic and abiotic stress responses of plants. *Biochim Biophys Acta - Gene Regul Mech* **1819**: 137–148.

Kidwell MG, Lisch DR. 2001. Perspective: transposable elements, parasitic DNA, and genome evolution. *Evolution* **55**: 1–24.

Kim J, Jung JH, Reyes JL, Kim YS, Kim SY, Chung KS, Kim JA, Lee M, Lee Y, Narry Kim V, et al. 2005. microRNA-directed cleavage of ATHB15 mRNA regulates vascular development in Arabidopsis inflorescence stems. *Plant J* **42**: 84–94.

Kiss T. 2001. Review: Small nucleolar RNA-guided post-transcriptional modification of cellular RNAs. *EMBO J* **20**: 3617.

Kleftogiannis D, Theofilatos K, Likothanassis S, Mavroudi S. 2015. YamiPred: A novel evolutionary method for predicting pre-miRNAs and selecting relevant features. *IEEE/ACM Trans Comput Biol Bioinforma* **PP**: 1–1.

Klug WS, Cummings MR. 1997. *Concepts of Genetics*. Prentice Hall.

Koepfli K, Paten B, Genome 10K Community of Scientists, O'Brien SJ. 2015. The Genome 10K Project: a way forward. *Annu Rev Anim Biosci* **3**: 57–111.

Koerner M V, Pauler FM, Huang R, Barlow DP. 2009. The function of non-coding RNAs in genomic imprinting. *Development* **136**: 1771–1783.

Krek A, Grün D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, MacMenamin P, da Piedade I, Gunsalus KC, Stoffel M, et al. 2005. Combinatorial microRNA target predictions. *Nat Genet* **37**: 495–500.

Krol J, Loedige I, Filipowicz W. 2010. The widespread regulation of microRNA biogenesis, function and decay. *Nat Rev Genet* **11**: 597–610.

Krol J, Sobczak K, Wilczynska U, Drath M, Jasinska A, Kaczynska D, Krzyzosiak WJ.

2004. Structural features of MicroRNA (miRNA) precursors and their relevance to miRNA biogenesis and small interfering RNA/Short hairpin RNA design. *J Biol Chem* **279**: 42230–42239.

Krüger J, Rehmsmeier M. 2006. RNAhybrid: microRNA target prediction easy, fast and flexible. *Nucleic Acids Res* **34**: W451–4.

Kuhn DE, Martin MM, Feldman DS, Terry A V., Nuovo GJ, Elton TS. 2008. Experimental validation of miRNA targets. *Methods* **44**: 47–54.

Kulkarni M, Ozgur S, Stoecklin G. 2010. On track with P-bodies. *Biochem Soc Trans* **38**: 242–251.

Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T. 2001. Identification of novel genes coding for small expressed RNAs. *Science (80-)* **294**: 853.

Lagos-Quintana M, Rauhut R, Meyer J, Borkhardt A, Tuschl T. 2003. New microRNAs from mouse and human. *RNA* **9**: 175–179.

Lagos-Quintana M, Rauhut R, Yalcin A, Meyer J, Lendeckel W, Tuschl T. 2002. Identification of tissue-specific microRNAs from mouse. *Curr Biol* **12**: 735–739.

Lai EC, Tomancak P, Williams RW, Rubin GM. 2003. Computational identification of Drosophila microRNA genes. *Genome Biol* **4**: R42.

Laing C, Schlick T. 2010. Computational approaches to 3D modeling of RNA. *J Phys Condens Matter* **22**: 283101.

Lal A, Navarro F, Maher CA, Maliszewski LE, Yan N, O'Day E, Chowdhury D, Dykxhoorn DM, Tsai P, Hofmann O, et al. 2009. miR-24 Inhibits Cell Proliferation by Targeting E2F2, MYC, and Other Cell-Cycle Genes via Binding to "Seedless" 3???UTR MicroRNA Recognition Elements. *Mol Cell* **35**: 610–625.

Laslett D, Canback B. 2004. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res* **32**: 11–16.

Lau NC, Lim LP, Weinstein EG, Bartel DP. 2001. An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science* **294**: 858–862.

Lawrie CH, ed. 2013. *MicroRNAs in Medicine*. John Wiley & Sons, Hoboken, NJ, USA.

Lechner RL, Engler MJ, Richardson CC. 1983. Characterization of strand displacement synthesis catalyzed by bacteriophage T7 DNA polymerase. *J Biol Chem* **258**: 11174–84.

Leclercq M, Diallo AB, Blanchette M. 2013. Computational prediction of the localization of microRNAs within their pre-miRNA. *Nucleic Acids Res* **41**: 7200–11.

Lee MR, Kim JS, Kim KS. 2010. MiR-124a is important for migratory cell fate transition during gastrulation of human embryonic stem cells. *Stem Cells* **28**: 1550–1559.

Lee RC, Ambros V. 2001. An extensive class of small RNAs in Caenorhabditis elegans. *Science* **294**: 862–864.

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**: 843–854.

Lee Y, Kim M, Han J, Yeom KH, Lee S, Baek SH, Kim VN. 2004. MicroRNA genes are

transcribed by RNA polymerase II. *EMBO J* **23**: 4051–4060.

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.

Lewis BP, Shih IH, others, Jones-Rhoades MW, Bartel DP, Burge CB. 2003. Prediction of mammalian microRNA targets. *Cell* **115**: 787–798.

Li J, Wang Y, Wang L, Feng W, Luan K, Dai X, Xu C, Meng X, Zhang Q, Liang H. 2015a. MatPred: Computational Identification of Mature MicroRNAs within Novel Pre-MicroRNAs. *Biomed Res Int* **2015**: 546763.

Li J, Zhang Y, Li D, Liu Y, Chu D, Jiang X, Hou D, Zen K, Zhang C-Y. 2015b. Small non-coding RNAs transfer through mammalian placenta and directly regulate fetal gene expression. *Protein Cell* **6**: 391–6.

Li S-C, Pan C-Y, Lin W. 2006. Bioinformatic discovery of microRNA precursors from human ESTs and introns. *BMC Genomics* **7**: 164.

Li SC, Shiau CK, Lin W. 2008. Vir-Mir db: prediction of viral microRNA candidate hairpins. *Nucleic Acids Res* **36**: D184.

Li Z-F, Zhang Y-C, Chen Y-Q. 2015c. miRNAs and lncRNAs in reproductive development. *Plant Sci* **238**: 46–52.

Liang T, Guo L, Liu C. 2012. Genome-wide analysis of mir-548 gene family reveals evolutionary and functional implications. *J Biomed Biotechnol* **2012**.

Lianidou ES, Markou A, Zavridou M. 2015. MicroRNA signatures as clinical biomarkers in lung cancer. *Curr Biomark Find* **Volume 5**: 35.

Lim LP, Lau NC, Garrett-Engele P, Grimson A, Schelter JM, Castle J, Bartel DP, Linsley PS, Johnson JM. 2005. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* **433**: 769–773.

Lindow M, Jacobsen A, Nygaard S, Mang Y, Krogh A. 2007. Intragenomic matching reveals a huge potential for miRNA-mediated regulation in plants. *PLoS Comput Biol* **3**: e238.

Liu B, Fang L, Chen J, Liu F, Wang X. 2015a. miRNA-dis: microRNA precursor identification based on distance structure status pairs. *Mol Biosyst* **11**: 1194–204.

Liu B, Fang L, Liu F, Wang X, Chou K-C. 2015b. iMiRNA-PseDPC: microRNA precursor identification with a pseudo distance-pair composition approach. *J Biomol Struct Dyn* 1–13.

Liu H, Yue D, Chen Y, Gao SJ, Huang Y. 2010. Improving performance of mammalian microRNA target prediction. *BMC Bioinformatics* **11**.

Long D, Lee R, Williams P, Chan CY, Ambros V, Ding Y. 2007. Potent effect of target structure on microRNA function. *Nat Struct Mol Biol* **14**: 287–294.

Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. 2011. ViennaRNA Package 2.0. *Algorithms Mol Biol* **6**: 26.

Lu C, Tej SS, Luo S, Haudenschild CD, Meyers BC, Green PJ. 2005. Elucidation of the

small RNA component of the transcriptome. *Science (80-)* **309**: 1567.

Lu ZJ, Yip KY, Wang G, Shou C, Hillier LW, Khurana E, Agarwal A, Auerbach R, Rozowsky J, Cheng C, et al. 2011. Prediction and characterization of noncoding RNAs in C. elegans by integrating conservation, secondary structure, and high-throughput sequencing and array data. *Genome Res* **21**: 276–285.

Lv DK, Bai X, Li Y, Ding XD, Ge Y, Cai H, Ji W, Wu N, Zhu YM. 2010. Profiling of cold-stress-responsive miRNAs in rice by microarrays. *Gene* **459**: 39–47.

Lyngso RB. 2004. Complexity of pseudoknot prediction in simple models. In *Automata, Languages And Programming: 31st International Colloquium, ICALP*, Vol. 3142 of, pp. 919–931.

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**: 9667–72.

Lyu Y, Shen Y, Li H, Chen Y, Guo L, Zhao Y, Hungate E, Shi S, Wu C-I, Tang T. 2014. New MicroRNAs in Drosophila—Birth, Death and Cycles of Adaptive Evolution ed. H.S. Malik. *PLoS Genet* **10**: e1004096.

Malecová B, Morris K V. 2010. Transcriptional gene silencing through epigenetic changes mediated by non-coding RNAs. *Curr Opin Mol Ther* **12**: 214–222.

Maragkakis M, Reczko M, Simossis VA, Alexiou P, Papadopoulos GL, Dalamagas T, Giannopoulos G, Goumas G, Koukis E, Kourtis K, et al. 2009. DIANA-microT web server: elucidating microRNA functions through target prediction. *Nucleic Acids Res* **37**: W273.

Maragkakis M, Vergoulis T, Alexiou P, Reczko M, Plomaritou K, Gousis M, Kourtis K, Koziris N, Dalamagas T, Hatzigeorgiou AG. 2011. DIANA-microT Web server upgrade supports Fly and Worm miRNA target prediction and bibliographic miRNA to disease association. *Nucleic Acids Res* **39**.

Martin MW, Grazhdankin D V, Bowring S a., Evans D a, Fedonkin M a, Kirschvink JL. 2000. Age of Neoproterozoic bilatarian body and trace fossils, White Sea, Russia: implications for metazoan evolution. *Science* **288**: 841–845.

Matera AG, Terns RM, Terns MP. 2007. Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. *Nat Rev Mol Cell Biol* **8**: 209–220.

Mathelier A, Carbone A. 2010. MIReNA: finding microRNAs with high accuracy and no learning at genome scale and from deep sequencing data. *Bioinformatics* **26**: 2226–34.

Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, Turner DH. 2004. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc Natl Acad Sci U S A* **101**: 7287–7292.

Mathews DH, Sabina J, Zuker M, Turner DH. 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol* **288**: 911–940.

McCaskill JS. 1990. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* **29**: 1105–19.

Mendes ND, Freitas AT, Sagot MF. 2009. Current tools for the identification of miRNA genes and their targets. *Nucleic Acids Res* **37**: 2419–2433.

Menor M, Ching T, Zhu X, Garmire D, Garmire LX. 2014. mirMark: a site-level and UTR-level classifier for miRNA target prediction. *Genome Biol* **15**: 500.

Metpally RPR, Nasser S, Courtright A, Carlson E, Ghaffari L, Villa S, Tembe W, Van Keuren-Jensen K. 2013. Comparison of analysis tools for miRNA high throughput sequencing using nerve crush as a model. *Front Genet* **4**.

Meza-Sosa KF, Valle-García D, Pedraza-Alva G, Pérez-Martínez L. 2012. Role of microRNAs in central nervous system development and pathology. *J Neurosci Res* **90**: 1–12.

Miller W, Rosenbloom K, Hardison RC, Hou M, Taylor J, Raney B, Burhans R, King DC, Baertsch R, Blankenberg D, et al. 2007. 28-way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome Res* **17**: 1797–808.

Miska EA, Alvarez-Saavedra E, Abbott AL, Lau NC, Hellman AB, McGonagle SM, Bartel DP, Ambros VR, Horvitz HR. 2007. Most Caenorhabditis elegans microRNAs are individually not essential for development or viability. *PLoS Genet* **3**: e215.

Moses AM, Pollard DA, Nix DA, Iyer VN, Li X-Y, Biggin MD, Eisen MB. 2006. Large-scale turnover of functional transcription factor binding sites in Drosophila. *PLoS Comput Biol* **2**: e130.

Mouttham N, Klunk J, Kuch M, Fourney R, Poinar H. 2015. Surveying the repair of ancient DNA from bones via high-throughput sequencing. *Biotechniques* **59**: 19–25.

Murchison EP, Stein P, Xuan Z, Pan H, Zhang MQ, Schultz RM, Hannon GJ. 2007. Critical roles for Dicer in the female germline. *Genes Dev* **21**: 682.

Murphy WJ, Eizirik E, O'Brien SJ, Madsen O, Scally M, Douady CJ, Teeling E, Ryder O a, Stanhope MJ, de Jong WW, et al. 2001. Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science* **294**: 2348–2351.

Nam JW, Kim J, Kim SK, Zhang BT. 2006. ProMiR II: a web server for the probabilistic prediction of clustered, nonclustered, conserved and nonconserved microRNAs. *Nucleic Acids Res* **34**: W455.

Nam JW, Shin KR, Han JJ, Lee Y, Kim VN, Zhang BT. 2005. Human microRNA prediction through a probabilistic co-learning model of sequence and structure. *Nucleic Acids Res* **33**: 3570–3581.

Nana-Sinkam SP, Croce CM. 2013. Clinical applications for microRNAs in cancer. *Clin Pharmacol Ther* **93**: 98–104.

Navarro L, Dunoyer P, Jay F, Arnold B, Dharmasiri N, Estelle M, Voinnet O, Jones JDG. 2006. A plant miRNA contributes to antibacterial resistance by repressing auxin signaling. *Science (80-)* **312**: 436.

Navarro L, Jay F, Nomura K, He SY, Voinnet O. 2008. Suppression of the microRNA pathway by bacterial effector proteins. *Science* **321**: 964–967.

Nawrocki EP, Kolbe DL, Eddy SR. 2009. Infernal 1.0: Inference of RNA alignments. *Bioinformatics* **25**: 1335–1337.

Nekrutenko A, Li W. 2001. Transposable elements are found in a large number of human protein-coding genes. *Trends Genet* **17**: 619–621.

Nelson JA. 2007. Small RNAs and large DNA viruses. *N Engl J Med* **357**: 2630–2632.

Nguyen TA, Jo MH, Choi Y-G, Park J, Kwon SC, Hohng S, Kim VN, Woo J-S. 2015. Functional Anatomy of the Human Microprocessor. *Cell* **161**: 1374–1387.

Nicolas FE. 2011. Experimental validation of microRNA targets using a luciferase reporter system. *Methods Mol Biol* **732**: 139–152.

Nissen P, Hansen J, Ban N, Moore PB, Steitz TA. 2000. The structural basis of ribosome activity in peptide bond synthesis. *Science* **289**: 920–930.

Nussinov R, Pieczenik G, Griggs JR, Kleitman DJ. 1978. Algorithms for Loop Matchings. *SIAM J Appl Math* **35**: 68–82.

Osada H, Takahashi T. 2007. MicroRNAs in biological processes and carcinogenesis. *Carcinogenesis* **28**: 2–12.

Osman A. 2012. MicroRNAs in health and disease--basic science and clinical applications. *Clin Lab* **58**: 393–402.

Osman IH, Kelly JP. 1996. *Meta-Heuristics: Theory and Applications*. Springer Science & Business Media.

Ouellet DL, Provost P. 2010. Current knowledge of MicroRNAs and noncoding RNAs in virus-infected cells. *Methods Mol Biol* **623**: 35–65.

Oulas A, Boutla A, Gkirtzou K, Reczko M, Kalantidis K, Poirazi P. 2009. Prediction of novel microRNA genes in cancer-associated genomic regionsa combined computational and experimental approach. *Nucleic Acids Res* **37**: 3276–3287.

Oulas A, Karathanasis N, Louloupi A, Pavlopoulos GA, Poirazi P, Kalantidis K, Iliopoulos I. 2015. Prediction of miRNA targets. *Methods Mol Biol* **1269**: 207–29.

Palatnik JF, Allen E, Wu X, Schommer C, Schwab R, Carrington JC, Weigel D. 2003. Control of leaf morphogenesis by microRNAs. *Nature* **425**: 257–263.

Parisien M, Major F. 2008. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature* **452**: 51–55.

Park JE, Heo I, Tian Y, Simanshu DK, Chang H, Jee D, Patel DJ, Kim VN. 2011. Dicer recognizes the 5 [prime] end of RNA for efficient and accurate processing. *Nature* **475**: 201–205.

Park JH, Shin C. 2014. MicroRNA-directed cleavage of targets: mechanism and experimental approaches. *BMB Rep* **47**: 417–23.

Pellicer J, FAY MF, LEITCH IJ. 2010. The largest eukaryotic genome of them all? *Bot J Linn Soc* **164**: 10–15.

Perez-Llamas C, Lopez-Bigas N. 2011. Gitools: Analysis and visualisation of genomic data using interactive heat-maps. *PLoS One* **6**.

Peterson SM, Thompson JA, Ufkin ML, Sathyanarayana P, Liaw L, Congdon CB. 2014. Common features of microRNA target prediction tools. *Front Genet* **5**.

Petruska J, Hartenstine MJ, Goodman MF. 1998. Analysis of strand slippage in DNA polymerase expansions of CAG/CTG triplet repeats associated with neurodegenerative disease. *J Biol Chem* **273**: 5204–10.

Pfeffer S, Sewer A, Lagos-Quintana M, Sheridan R, Sander C, Grasser FA, van Dyk LF, Ho CK, Shuman S, Chien M, et al. 2005. Identification of microRNAs of the herpesvirus family. *Nat Methods* **2**: 269–276.

Piriyapongsa J, Jordan IK. 2007a. A family of human microRNA genes from miniature inverted-repeat transposable elements. *PLoS One* **2**: e203.

Piriyapongsa J, Jordan IK. 2007b. A family of human microRNA genes from miniature inverted-repeat transposable elements. *PLoS One* **2**: e203.

Piriyapongsa J, Mariño-Ramírez L, Jordan IK. 2007. Origin and evolution of human microRNAs from transposable elements. *Genetics* **176**: 1323–37.

Plescia OJ, Palczuk NC, Cora-Figueroa E, Mukherjee A, Braun W. 1965. Production of antibodies to soluble RNA (sRNA). *Proc Natl Acad Sci U S A* **54**: 1281–5.

Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* **20**: 110–121.

Ponting CP, Oliver PL, Reik W. 2009. Evolution and Functions of Long Noncoding RNAs. *Cell* **136**: 629–641.

Porrello ER. 2013. microRNAs in cardiac development and regeneration. *Clin Sci (Lond)* **125**: 151–66.

Poy MN, Eliasson L, Krutzfeldt J, Kuwajima S, Ma X, Macdonald PE, Pfeffer S, Tuschl T, Rajewsky N, Rorsman P, et al. 2004. A pancreatic islet-specific microRNA regulates insulin secretion. *Nature* **432**: 226–230.

Putnam NH, Butts T, Ferrier DEK, Furlong RF, Hellsten U, Kawashima T, Robinson-Rechavi M, Shoguchi E, Terry A, Yu J-K, et al. 2008. The amphioxus genome and the evolution of the chordate karyotype. *Nature* **453**: 1064–71.

Qu J, Ye J, Fang R. 2007. Artificial microRNA-mediated virus resistance in plants. *J Virol* **81**: 6690–6699.

Quinlan JR. 1993. *C4.5: programs for machine learning*. Morgan kaufmann.

Quinlan JR. 1986. Induction of decision trees. *Mach Learn* **1**: 81–106.

Reczko M, Maragkakis M, Alexiou P, Grosse I, Hatzigeorgiou AG. 2012. Functional microRNA targets in protein coding sequences. *Bioinformatics* **28**: 771–6.

Redova M, Sana J, Slaby O. 2013. Circulating miRNAs as new blood-based biomarkers for solid cancers. *Future Oncol* **9**: 387–402.

Reimand J, Kull M, Peterson H, Hansen J, Vilo J. 2007. G:Profiler-a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res* **35**.

Remita MA, Lord E, Agharbaoui Z, Leclercq M, Badawi M, Makarenkov V, Sarhan F, Diallo AB. 2015. *WMP: A novel comprehensive wheat miRNA database, including*

related bioinformatics software. Cold Spring Harbor Labs Journals.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet* **16**: 276–277.

Rinn JL, Chang HY. 2012. Genome Regulation by Long Noncoding RNAs. *Annu Rev Biochem* **81**: 145–166.

Ritchie W, Flamant S, Rasko JEJ. 2009. Predicting microRNA targets and functions: traps for the unwary. *Nat Methods* **6**: 397–8.

Robnik-Sikonja M. 2004. Improving random forests. *Mach Learn Ecml 2004, Proc* **3201**: 359–370.

Robnik-Sikonja M, Kononenko I. 2003. Theoretical and empirical analysis of ReliefF and RReliefF. *Mach Learn* **53**: 23–69.

Rogers K, Chen X. 2013. Biogenesis, turnover, and mode of action of plant microRNAs. *Plant Cell* **25**: 2383–99.

Ruvkun G. 2001. Molecular biology: glimpses of a tiny RNA world. *Sci STKE* **294**: 797.

Saetrom P, Heale BSE, Snøve O, Aagaard L, Alluin J, Rossi JJ. 2007. Distance constraints between microRNA target sites dictate efficacy and cooperativity. *Nucleic Acids Res* **35**: 2333–42.

Sahoo S, Losordo DW. 2014. Exosomes and cardiac repair after myocardial infarction. *Circ Res* **114**: 333–344.

Salse J, Abrouk M, Bolot S, Guilhot N, Courcelle E, Faraut T, Waugh R, Close TJ, Messing J, Feuillet C. 2009. Reconstruction of monocotelydoneous proto-chromosomes reveals faster evolution in plants than in animals. *Proc Natl Acad Sci U S A* **106**: 14908–13.

Salser W. 1978. Globin mRNA Sequences: Analysis of Base Pairing and Evolutionary Implications. *Cold Spring Harb Symp Quant Biol* **42**: 985–1002.

Sankoff D. 1985. Simultaneous Solution of the RNA Folding, Alignment and Protosequence Problems. *SIAM J Appl Math* **45**: 810–825.

Saqib M, Zorb C, Schubert S. 2008. Silicon-mediated improvement in the salt resistance of wheat (Triticum aestivum) results from increased sodium exclusion and resistance to oxidative stress. *Funct plant Biol* **35**: 633–639.

Sarnow P, Jopling CL, Norman KL, Sch\"utz S, Wehner KA. 2006. MicroRNAs: expression, avoidance and subversion by vertebrate viruses. *Nat Rev Microbiol* **4**: 651–659.

Saxe JP, Lin H. 2011. Small noncoding RNAs in the germline. *Cold Spring Harb Perspect Biol* **3**: 1–16.

Saxena A, Carninci P. 2011. Long non-coding RNA modifies chromatin: epigenetic silencing by long non-coding RNAs. *Bioessays* **33**: 830–9.

Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S, Martinez-Jimenez CP, Mackay S, et al. 2010. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328**: 1036–40.

Schulte JH, Marschall T, Martin M, Rosenstiel P, Mestdagh P, Schlierf S, Thor T, Vandesompele J, Eggert A, Schreiber S, et al. 2010. Deep sequencing reveals differential expression of microRNAs in favorable versus unfavorable neuroblastoma. *Nucleic Acids Res* 1–10.

Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W. 2003. Human-mouse alignments with BLASTZ. *Genome Res* **13**: 103–7.

Schwarz DS, Hutvágner G, Du T, Xu Z, Aronin N, Zamore PD. 2003. Asymmetry in the assembly of the RNAi enzyme complex. *Cell* **115**: 199–208.

Sethupathy P, Corda B, Hatzigeorgiou AG. 2006. TarBase: A comprehensive database of experimentally supported animal microRNA targets. *Rna* **12**: 192.

Seto AG, Kingston RE, Lau NC. 2007. The coming of age for Piwi proteins. *Mol Cell* **26**: 603–609.

Sewer A, Paul N, Landgraf P, Aravin A, Pfeffer S, Brownstein MJ, Tuschl T, van Nimwegen E, Zavolan M. 2005. Identification of clustered microRNAs using an ab initio prediction method. *BMC Bioinformatics* **6**: 267.

Shannon CE, Weaver W, Blahut RE, Hajek B. 1949. *The mathematical theory of communication*. University of Illinois press Urbana.

Shapiro B, Hofreiter M. 2014. A paleogenomic perspective on evolution and gene function: new insights from ancient DNA. *Science (80-)* **343**: 1236573.

Shin C, Nam J-W, Farh KK-H, Chiang HR, Shkumatava A, Bartel DP. 2010. Expanding the microRNA targeting code: functional sites with centered pairing. *Mol Cell* **38**: 789–802.

Shivdasani RA. 2006. MicroRNAs: regulators of gene expression and cell differentiation. *Blood* **108**: 3646–3653.

Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**: 1034–1050.

Siepel A, Haussler D. 2005. Phylogenetic Hidden Markov Models. *Engineering* 325–351.

Simkin AT, Bailey JA, Gao F-B, Jensen JD. 2014. Inferring the evolutionary history of primate microRNA binding sites: overcoming motif counting biases. *Mol Biol Evol* msu129–.

Smalheiser NR, Torvik VI. 2005. Mammalian microRNAs derived from genomic repeats. *Trends Genet* **21**: 322–6.

Smit AFA, Hubley R, Green P. 1996. RepeatMasker Open-3.0.

Soumillon M, Meunier J, Weier M, Guschanski K, Hu H, Khaitovich P, Kaessmann H. 2013. Birth and expression evolution of mammalian microRNA genes. 34–45.

Srinivasan S, Selvan ST, Archunan G, Gulyas B, Padmanabhan P. 2013. MicroRNAs -the next generation therapeutic targets in human diseases. *Theranostics* **3**: 930–942.

Srivastava PK, Moturu TR, Pandey P, Baldwin IT, Pandey SP. 2014. A comparison of performance of plant miRNA target prediction tools and the characterization of

features for genome-wide target prediction. *BMC Genomics* **15**: 348.

Staple DW, Butcher SE. 2005. Pseudoknots: RNA structures with diverse functions. *PLoS Biol* **3**: 0956–0959.

Suh MR, Lee Y, Kim JY, Kim SK, Moon SH, Lee JY, Cha KY, Chung HM, Yoon HS, Moon SY, et al. 2004. Human embryonic stem cells express a unique set of microRNAs. *Dev Biol* **270**: 488–498.

Sunkar R, Kapoor A, Zhu JK. 2006. Posttranscriptional induction of two Cu/Zn superoxide dismutase genes in Arabidopsis is mediated by downregulation of miR398 and important for oxidative stress tolerance. *Plant Cell Online* **18**: 2051.

Sunkar R, Li YF, Jagadeeswaran G. 2012. Functions of microRNAs in plant stress responses. *Trends Plant Sci* **17**: 196–203.

Sunkar R, Zhou X, Zheng Y, Zhang W, Zhu J-KK. 2008. Identification of novel and candidate miRNAs in rice by high throughput sequencing. *BMC Plant Biol* **8**: 25.

Swami M. 2010. Small RNAS: An epigenetic silencing influence. *Nat Rev Genet* **11**: 172–173.

Szymanski M, Barciszewska MZ, Erdmann VA, Barciszewski J. 2005. A new frontier for molecular medicine: noncoding RNAs. *Biochim Biophys Acta* **1756**: 65–75.

Taft RJ, Glazov EA, Cloonan N, Simons C, Stephen S, Faulkner GJ, Lassmann T, Forrest ARR, Grimmond SM, Schroder K, et al. 2009. Tiny RNAs associated with transcription start sites in animals. *Nat Genet* **41**: 572–578.

Taft RJ, Pang KC, Mercer TR, Dinger M, Mattick JS. 2010. Non-coding RNAs: Regulators of disease. *J Pathol* **220**: 126–139.

Taylor RS, Tarver JE, Hiscock SJ, Donoghue PCJ. 2014. Evolutionary history of plant microRNAs. *Trends Plant Sci* **19**: 175–82.

Teruel-Montoya R, Kong X, Abraham S, Ma L, Kunapuli SP, Holinstat M, Shaw CA, McKenzie SE, Edelstein LC, Bray PF. 2014. MicroRNA expression differences in human hematopoietic cell lineages enable regulated transgene expression. *PLoS One* **9**: e102259.

Teune JH, Steger G. 2010. NOVOMIR: de novo prediction of microRNA-coding regions in a single plant-genome. *J Nucleic Acids* **2010**: 10.

Thieme CJ, Gramzow L, Lobbes D, TheiBen G. 2011. SplamiR: prediction of spliced miRNAs in plants. *Bioinformatics* **27**: 1215–1223.

Thomson DW, Bracken CP, Goodall GJ. 2011. Experimental strategies for microRNA target identification. *Nucleic Acids Res* **39**: 6845–53.

Tran VDT, Tempel S, Zerath B, Zehraoui F, Tahi F. 2015. miRBoost: boosting support vector machines for microRNA precursor classification. *RNA* **21**: 775–85.

Turner DH, Mathews DH. 2009. NNDB: The nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Res* **38**.

van de Lagemaat LN, Landry J-R, Mager DL, Medstrand P. 2003. Transposable elements in mammals promote regulatory variation and diversification of genes with

specialized functions. *Trends Genet* **19**: 530–6.

Van Rooij E, Purcell AL, Levin AA. 2012. Developing MicroRNA therapeutics. *Circ Res* **110**: 496–507.

Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. 1995. Serial analysis of gene expression. *Science* **270**: 484–7.

Venkataram S, Fay JC. 2010. Is transcription factor binding site turnover a sufficient explanation for cis-regulatory sequence divergence? *Genome Biol Evol* **2**: 851–8.

Vigneault F, Ter-Ovanesyan D, Alon S, Eminaga S, C Christodoulou D, Seidman JG, Eisenberg E, M Church G. 2012. High-throughput multiplex sequencing of miRNA. *Curr Protoc Hum Genet* **Chapter 11**: Unit 11.12.1–10.

Vlachos IS, Paraskevopoulou MD, Karagkouni D, Georgakilas G, Vergoulis T, Kanellos I, Anastasopoulos I-L, Maniou S, Karathanou K, Kalfakakou D, et al. 2015. DIANA-TarBase v7.0: indexing more than half a million experimentally supported miRNA:mRNA interactions. *Nucleic Acids Res* **43**: D153–9.

Voet D, Voet JG. 2010. *Biochemistry*. Wiley.

Voinnet O. 2004. Shaping small RNAs in plants by gene duplication. *Nat Genet* **36**: 1245–1246.

Volff J-N. 2006. Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. *Bioessays* **28**: 913–22.

Wang W-C, Lin F-M, Chang W-C, Lin K-Y, Huang H-D, Lin N-S. 2009. miRExpress: analyzing high-throughput sequencing data for profiling microRNA expression. *BMC Bioinformatics* **10**: 328.

Wang X. 2014. Composition of seed sequence is a major determinant of microRNA targeting patterns. *Bioinformatics* **30**: 1377–1383.

Wang X. 2008. miRDB: a microRNA target prediction and functional annotation database with a wiki interface. *Rna* **14**: 1012.

Wang X, El Naqa IM. 2008. Prediction of both conserved and nonconserved microRNA targets in animals. *Bioinformatics* **24**: 325–332.

Wang X, Zhang J, Li F, Gu J, He T, Zhang X, Li Y. 2005. MicroRNA identification based on sequence and structure alignment. *Bioinformatics* **21**: 3610–3614.

Wang Y, Dai X, Ru J, Lv D, Li J. 2015. Computational methods for the identification of mature microRNAs within their Pre-miRNA. In *2015 8th International Congress on Image and Signal Processing (CISP)*, pp. 1241–1245, IEEE.

Warthmann N, Chen H, Ossowski S, Weigel D, Herve P. 2008. Highly specific gene silencing by artificial miRNAs in rice. *PLoS One* **3**.

Washietl S, Hofacker IL, Stadler PF. 2005. Fast and reliable prediction of noncoding RNAs. *Proc Natl Acad Sci U S A* **102**: 2454–2459.

Watson JD, Crick FHC. 1953. Molecular structure of nucleic acids. *Nature* **171**: 737–738.

Wienholds E, Koudijs MJ, van Eeden FJM, Cuppen E, Plasterk RHA. 2003. The microRNA-producing enzyme Dicer1 is essential for zebrafish development. *Nat*

Genet **35**: 217–218.

Wilfred BR, Wang WX, Nelson PT. 2007. Energizing miRNA research: A review of the role of miRNAs in lipid metabolism, with a prediction that miR-103/107 regulates human metabolic pathways. *Mol Genet Metab* **91**: 209–217.

Winter J, Jung S, Keller S, Gregory RI, Diederichs S. 2009. Many roads to maturity: microRNA biogenesis pathways and their regulation. *Nat Cell Biol* **11**: 228–234.

Witkos TM, Koscianska E, Krzyzosiak WJ. 2011. Practical Aspects of microRNA Target Prediction. *Curr Mol Med* **11**: 93–109.

Wong SSY, Ritner C, Ramachandran S, Aurigui J, Pitt C, Chandra P, Ling VB, Yabut O, Bernstein HS. 2012. miR-125b promotes early germ layer specification through Lin28/let-7d and preferential differentiation of mesoderm in human embryonic stem cells. *PLoS One* **7**: e36121.

Wower J, Zwieb C. 2000. tmRDB (tmRNA database). *Nucleic Acids Res* **28**: 169–170.

Wu GZ, Huang ZP, Wang DZ. 2013a. MicroRNAs in cardiac regeneration and cardiovascular disease. *Sci China Life Sci* **56**: 907–913.

Wu J, Liu Q, Wang X, Zheng J, Wang T, You M, Sheng Sun Z, Shi Q. 2013b. mirTools 2.0 for non-coding RNA discovery, profiling, and functional annotation based on high-throughput sequencing. *RNA Biol* **10**: 1087–92.

Wu L, Belasco JG. 2008. Let Me Count the Ways: Mechanisms of Gene Regulation by miRNAs and siRNAs. *Mol Cell* **29**: 1–7.

Wu Y, Wei B, Liu H, Li T, Rayner S. 2011. MiRPara: a SVM-based software tool for prediction of most probable microRNA coding regions in genome scale sequences. *BMC Bioinformatics* **12**: 107.

Wuchty S, Fontana W, Hofacker IL, Schuster P. 1999. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers* **49**: 145–165.

Xia T, Liao Q, Jiang X, Shao Y, Xiao B, Xi Y, Guo J. 2014. Long noncoding RNA associated-competing endogenous RNAs in gastric cancer. *Sci Rep* **4**: 6088.

Xiao C, Rajewsky K. 2009. MicroRNA control in the immune system: basic principles. *Cell* **136**: 26–36.

Xiao F, Zuo Z, Cai G, Kang S, Gao X, Li T. 2009. miRecords: an integrated resource for microRNA--target interactions. *Nucleic Acids Res* **37**: D105.

Xie B, Ding Q, Han H, Wu D. 2013. MiRCancer: A microRNA-cancer association database constructed by text mining on literature. *Bioinformatics* **29**: 638–644.

Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M. 2005. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* **434**: 338–345.

Xu J, Zhang R, Shen Y, Liu G, Lu X, Wu C-I. 2013. The evolution of evolvability in microRNA target sites in vertebrates. *Genome Res* **23**: 1810–6.

Xuan P, Guo MZ, Huang YC, Li WB, Huang YF. 2011. MaturePred: Efficient Identification of MicroRNAs within Novel Plant Pre-miRNAs. *PLoS One* **6**: e27422.

Xue C, Li F, He T, Liu G-PP, Li Y, Zhang X. 2005. Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics* **6**: 310.

Yang J-SS, Phillips MD, Betel D, Mu P, Ventura A, Siepel AC, Chen KC, Lai EC. 2011. Widespread regulatory activity of vertebrate microRNA* species. *RNA* **17**: 312–326.

Yousef M, Nebozhyn M, Shatkay H, Kanterakis S, Showe LC, Showe MK. 2006. Combining multi-species genomic data for microRNA identification using a Naive Bayes classifier. *Bioinformatics* **22**: 1325–1334.

Yuan Z, Sun X, Liu H, Xie J. 2011. MicroRNA genes derived from repetitive elements and expanded by segmental duplication events in mammalian genomes. *PLoS One* **6**: e17666.

Zampetaki A, Mayr M. 2012. MicroRNAs in vascular and metabolic disease. *Circ Res* **110**: 508–522.

Zhang BH, Pan XP, Cox SB, Cobb GP, Anderson TA. 2006. Evidence that miRNAs are different from other RNAs. *Cell Mol Life Sci* **63**: 246–254.

Zhang Y. 2005. miRU: an automated plant miRNA target prediction server. *Nucleic Acids Res* **33**: W701.

Zhang ZZ, Yu J, Li D, Liu F, Zhou X, Wang T, Ling Y, Su Z. 2010. PMRD: plant microRNA database. *Nucleic Acids Res* **38**: D806–13.

Zhao D, Wang Y, Luo D, Shi X, Wang L, Xu D, Yu J, Liang Y. 2010a. PMirP: A pre-microRNA prediction method based on structure-sequence hybrid features. *Artif Intell Med* **49**: 127–132.

Zhao P, Zhang WB, Chen SJ. 2010b. Predicting secondary structural folding kinetics for nucleic acids. *Biophys J* **98**: 1617–1625.

Zhao Y, Ransom JF, Li A, Vedantham V, von Drehle M, Muth AN, Tsuchihashi T, McManus MT, Schwartz RJ, Srivastava D. 2007. Dysregulation of Cardiogenesis, Cardiac Conduction, and Cell Cycle in Mice Lacking miRNA-1-2. *Cell* **129**: 303–317.

Zheng H, Fu R, Wang J-T, Liu Q, Chen H, Jiang S-W. 2013. Advances in the Techniques for the Prediction of microRNA Targets. *Int J Mol Sci* **14**: 8179–87.

Zhou J, Foster DP, Stine RA, Ungar LH. 2006. Streamwise feature selection. *J Mach Learn Res* **7**: 1861–1885.

Zhu E, Zhao F, Xu G, Hou H, Zhou L, Li X, Sun Z, Wu J. 2010. MirTools: MicroRNA profiling and discovery based on high-throughput sequencing. *Nucleic Acids Res* **38**.

Zou Q, Mao Y, Hu L, Wu Y, Ji Z. 2014. miRClassify: An advanced web server for miRNA family classification and annotation. *Comput Biol Med* **45**: 157–60.

Zuker M, Stiegler P. 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res* **9**: 133.