Pattern recognition for non-coding RNA promoters

A dissertation submitted

by

Georgios K. Georgakilas

to

the Department of Electrical and Computer Engineering

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

University of Thessaly

Volos, Greece

September 2015

Dissertation Committee

Prof. Artemis G. Hatzigeorgiou, Supervisor

Prof. Elias Houstis

Prof. Katerina Housti

Pattern recognition for non-coding RNA promoters

The central dogma of Biology summarizes the flow of genetic information from deoxyribonucleic acid (DNA) to ribonucleic acid (RNA) to proteins. Early estimations considered that only 1% of the human genome encodes proteins while the rest constitutes "junk" DNA. During the last decade and due to the invention of novel experimental methodologies and platforms there has been an increasing number of publications that continuously remove the obscurity surrounding the central dogma of Biology. Recent estimations consider that 3% of the human genome encodes proteins, 62% transcribes functional non-coding RNA (ncRNA) elements and approximately 80% participates in at least one biochemical event. These findings suggest that RNA and especially the class of ncRNAs constitutes an integral part of every cellular process and its "elusive" role remains to be unveiled.

Traditionally, the concept of ncRNAs has been utilized as a blanket term for a wide range of molecules which have initially been categorized into subfamilies based on their size. NcRNAs shorter than 200 nucleotides (nts) are generally termed small RNAs while the rest constitute the long non-coding RNA (lncRNA) subspecies. MicroRNAs (miRNAs), which were first discovered in 1993 and are the main focus of this dissertation, are single stranded RNA (ssRNA) molecules (~22 nts) that post-transcriptionally regulate gene expression by translation suppression and/or messenger RNA (mRNA) degradation. Since the discovery of their abundant transcription in 2001 there has been an explosion of miRNA-related publications, estimated to exceed 40,000 (Sep 2015). Even though there have been substantial breakthroughs in research related to miRNA biogenesis, function and disease implication, there are still open questions regarding their expression regulation due to the rapid processing and degradation of their primary transcripts (pri-miRNAs) in the nucleus by Drosha enzyme.

Aim of my Doctoral studies was to design an algorithm and implement it into a robust computational framework that would facilitate the assembly of a genome-wide, accurate and high-resolution map of miRNA transcription start sites (TSS). This goal has been achieved by developing microTSS, an algorithm that combines Machine Learning and

Next Generation Sequencing (NGS) data in order to provide highly accurate, single nucleotide resolution predictions for miRNA gene TSSs. MicroTSS integrates RNA Sequencing data with active transcription marks derived from chromatin immunoprecipitation (ChIP) and DNase Sequencing and enables the characterization of tissue-specific miRNA TSSs. MicroTSS was validated with RNA Sequencing data derived from a *Drosha* null/conditional-null mouse model specifically designed for this purpose and generated using the conditional by inversion (COIN) methodology.

During the course of my Doctoral studies, I participated in six publications that provide robust computational methods, able to complement microTSS and facilitate every aspect of miRNA-related research. The implemented algorithms are readily applicable to a variety of cell lines or organisms and they can be utilized separately or combined, depending on the study setting. The identification of differences in miRNA expression regulation as well as the target repertoire between pathological and physiological conditions, cell types and species, could inaugurate a new era for the elucidation of miRNA expression and function, redefining their role into the wider context of biological networks and pathways.

Τριμελής Επιτροπή

Καθ. Άρτεμις Γ. Χατζηγεωργίου, Επιβλέπουσα

Καθ. Ηλίας Χούστης

Καθ. Κατερίνα Χούστη

Αναγνώριση μοτίβων για υποκινητές μη-κωδικών RNA

Το κεντρικό δόγμα της Βιολογίας συνοψίζει τη ροή της γενετικής πληροφορίας από το δεοξυριβονουκλεϊκό οξύ (DNA) προς το ριβονουκλεϊκό οξύ (RNA) προς τις πρωτεΐνες. Αρχικές εκτιμήσεις ανέφεραν πως μόλις 1% του ανθρώπινου γονιδιώματος κωδικοποιεί πρωτεΐνες ενώ το υπόλοιπο χαρακτηρίζεται ως «άχρηστο» DNA. Κατά τη διάρκεια της τελευταίας δεκαετίας, η εμφάνιση καινοτόμων πειραματικών μεθοδολογιών επέτρεψε τη συνεχόμενη αύξηση του αριθμού των μελετών που αποκαλύπτουν νέα στοιχεία και αποσαφηνίζουν σκοτεινά σημεία γύρω από το κεντρικό δόγμα της Βιολογίας. Πρόσφατες εκτιμήσεις υπολογίζουν πως 3% του ανθρώπινου γονιδιώματος κωδικοποιεί πρωτεΐνες, 62% μεταγράφει μη-κωδικά RNA ενώ το 80% συμμετέχει σε μία τουλάχιστον βιοχημική διεργασία. Τα παραπάνω ευρήματα υποδεικνύουν ότι το RNA και ειδικότερα τα μη-κωδικά RNA αποτελούν αναπόσπαστο κομμάτι της κυτταρικής λειτουργίας και ο ρόλος τους μένει να αποσαφηνιστεί.

Παραδοσιακά, η έννοια μη-κωδικό RNA χρησιμοποιείται σαν όρος ομπρέλα για πληθώρα μορίων που κατηγοριοποιούνται σε υποοικογένειες με βάση το μήκος τους. Μη-κωδικά RNA μικρότερα των 200 νουκλεοτιδίων γενικά ονομάζονται μικρά RNAs ενώ τα υπόλοιπα αποτελούν το υποείδος των μακρών μη-κωδικών RNA. Τα microRNA (miRNA), τα οποία ανακαλύφθηκαν το 1993 και στα οποία εστιάζει η παρούσα διδακτορική διατριβή, είναι μικρά μόρια RNA, μήκους περίπου 22 νουκλεοτιδίων, που ρυθμίζουν μετά-μεταφραστικά την έκφραση των γονιδίων είτε εμποδίζοντας τη σύνθεση των πρωτεϊνών ή οδηγώντας το αγγελιοφόρο RNA σε αποδόμηση. Από το 2001 που επιβεβαιώθηκε η ευρύτητα της έκφρασής τους στα κύτταρα και έπειτα, σημειώθηκε έκρηξη στον αριθμό (περισσότερες από 40,000 – Σεπτέμβρης 2015) των δημοσιεύσεων που σχετίζονται με την έρευνα των miRNAs. Αυτό οδήγησε σε σημαντικές ανακαλύψεις σχετικά με τον μηχανισμό ωρίμανσης και δράσης των miRNAs καθώς και στον τρόπο που εμπλέκονται στις ασθένειες. Παρόλα αυτά, υπάρχουν ακόμα ανοικτά ερωτήματα που σχετίζονται με τις διεργασίες και τους παράγοντες που ελέγχουν την έκφρασή τους. Αυτό οφείλεται στο ένζυμο Drosha το οποίο προκαλεί ταχύτατη αποδόμηση των πρώιμων μετάγραφων RNA, από τα οποία παράγονται τα miRNA, εμποδίζοντας την ανίχνευση των γονιδίων τους με συμβατικές πειραματικές τεχνολογίες.

Στόχος της παρούσας Διδακτορικής διατριβής αποτελεί η σχεδίαση ενός αλγορίθμου κατάλληλου για τη δημιουργία ενός ακριβούς και υψηλής ανάλυσης χάρτη θέσεων έναρξης μεταγραφής miRNA γονιδίων. Ο στόχος αυτός επετεύχθη με την υλοποίηση του microTSS, ενός αλγορίθμου που συνδυάζει Μηχανική Μάθηση και δεδομένα Αλληλούχισης Επόμενης Γενιάς και παρέχει ακριβείς και υψηλής ανάλυσης προβλέψεις θέσεων έναρξης μεταγραφής miRNA γονιδίων. Η αξιολόγηση του αλγορίθμου επετεύχθη με την αξιοποίηση ενός ζωικού μοντέλου (μυς) από το οποίο αφαιρέθηκε το γονίδιο *Drosha* επιτρέποντας την ανίχνευση των πρώιμων μετάγραφων των miRNA γονιδίων με τεχνικές αλληλούχισης RNA.

Κατά τη διάρκεια των Διδακτορικών μου σπουδών, συμμετείχα σε έξι ακόμα δημοσιεύσεις μελετών που περιγράφουν εργαλεία και υπολογιστικές τεχνικές που δρουν συμπληρωματικά στον αλγόριθμο microTSS και διευκολύνουν με πολύπλευρο τρόπο την έρευνα που σχετίζεται με τα miRNA. Όλες οι μέθοδοι είναι εφαρμόσιμοι σε πληθώρα ιστών, κυτταρικών σειρών και οργανισμών και μπορούν να αξιοποιηθούν μεμονωμένα ή συνδυαστικά ανάλογα με το πλαίσιο της εκάστοτε μελέτης. Η ανίχνευση διαφορών, ανάμεσα σε φυσιολογικές και παθολογικές καταστάσεις, που αφορούν στη ρύθμιση της έκφρασης των miRNA και των στόχων τους δύναται να εγκαινιάσει μια νέα εποχή στην Βιολογία επαναπροσδιορίζοντας το ρόλο των σημαντικών αυτών μορίων στο ευρύτερο πλαίσιο των δικτύων γονιδιακής έκφρασης.

Acknowledgments

I would like to thank the members of the dissertation committee Professor Elias Houstis, Professor Katerina Housti and especially Professor Artemis G. Hatzigeorgiou who provided me with her exceptional guidance, as a supervisor of my Doctoral studies, and the opportunity to work in the cutting-edge research field of Bioinformatics and microRNAs in a state-of-the-art and stimulating working environment.

I would like to thank my colleagues and lab members, Dimitra Karagkouni, Konstantinos Liakos and especially Dr. Ioannis Vlachos and Maria Paraskevopoulou for the exceptional and very productive collaboration that led to numerous publications in highly esteemed journals as well as all those stimulating conversations that contributed in evolving as scientists.

I would also like to thank previous lab members and colleagues, Dr. Panagiotis Alexiou, Dr. Manolis Maragkakis and Dr. Martin Reczko for their invaluable help during my first steps.

The work, presented in this dissertation, was funded by research projects 09 SYN - 13 - 1055 “MIKRORNA” and “TOM”, “ARISTEIA” Action of the “OPERATIONAL PROGRAMME EDUCATION AND LIFELONG LEARNING”, General Secretariat for Research and Technology, Ministry of Education, Greece, European Social Fund (ESF).

Dedicated to PT and my family.

Contents

List of Figures

List of Tables

Abbreviations

3'UTR	3-prime untranslated region
5'UTR	5-prime untranslated region
bp	Base pairs
CAGE	Cap analysis of gene expression
CB	Cajal bodies
CDS	Coding sequence
ceRNA	Competing endogenous RNA
ChIP-Seq	Chromatin immunoprecipitation sequencing
COIN	Conditional by inversion
DGF	Digital genomic footprint
DNA	Deoxyribonucleic acid
endo-siRNA	Endogenous small interfering RNA
FDR	False discovery rate
FP	False positive
GRO-Seq	Global run-on sequencing
HITS-CLIP	RNA isolated by crosslinking immunoprecipitation
kb	Kilo-bases
kbp	Kilo-base pairs
lincRNA	Long intergenic non-coding RNA
lncRNA	Long non-coding RNA
mESC	Mouse embryonic stem cell
miRNA	microRNA
miRNP	Micro ribonucleic protein
MRE	microRNA recognition element
mRNA	Messenger RNA
ncRNA	Non-coding RNA
NGS	Next generation sequencing
nt	Nucleotide

PAR-CLIP	Photoactivatable-ribonucleoside-enhanced crosslinking and immunoprecipitation
P-bodies	Cytoplasmic processing bodies
PE	Paired-end
piRNA	PIWI-associated interfering RNA
Pol2	RNA polymerase II
Pol3	RNA polymerase III
pre-miRNA	Precursor microRNA
pri-miRNA	Primary microRNA
qPCR	Quantitative polymerase chain reaction
RBF	Radial basis function
RISC	RNA induced silencing complex
RNA	Ribonucleic acid
RNAi	RNA interference
RNP	Ribonucleic protein
rRNA	Ribosomal RNA
RT-PCR	Real time polymerase chain reaction
scaRNP	Small cajal body -specific ribonucleic protein
SE	Single-end
snoRNA	Small nucleolar RNA
snRNA	Small nuclear RNA
snRNP	Small nuclear ribonucleic protein
SVM	Support vector machines
TF	Transcription factor
TP	True positive
tRNA	Transfer RNA

1. Introduction

1.1. The Dogma revisited

The traditional view of the central dogma of biology states that genetic information which is encoded in the form of deoxyribonucleic acid (DNA) is transcribed into individual molecular units called ribonucleic acids (RNA) that are subsequently translated to proteins (**Fig. 1**). Initially, genes with the ability to translate proteins and few classes of RNA were believed to be the functional part of genome. The rest was considered "junk" DNA meant to act as buffer against inherited or environmentally-driven mutation causing mechanisms that could lead to various pathological states such as cancer.

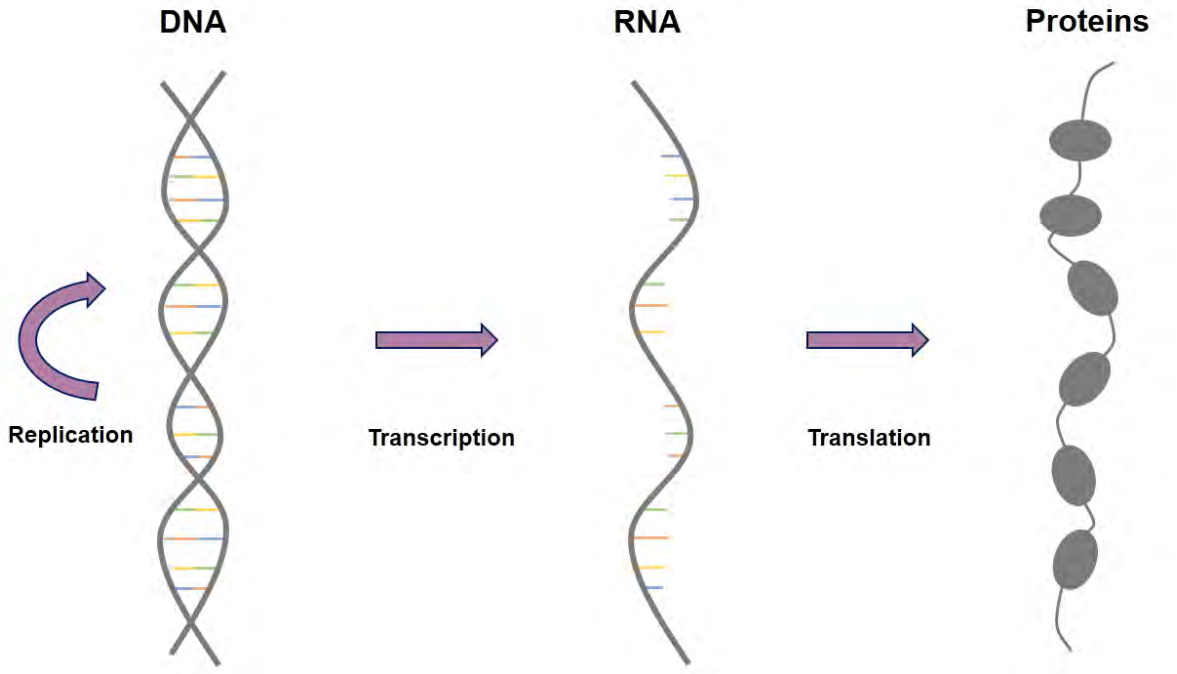


Figure 1. The central dogma of Biology. This figure has been designed for the purposes of this dissertation.

During the last decade, in the light of Next Generation Sequencing (NGS), novel experimental methodologies have removed the obscurity surrounding RNA and turned the "desert" region of DNA into a research hotpot. The ENCODE Project Consortium has been an integral part in the process of unveiling secrets of the human genome regarding its function and organization. According to the consortium's flagship publication (Consortium, 2012) approximately 3% of the human genome encodes proteins, 62% transcribes functional non-coding RNA (ncRNA) molecules and 80% participates in at least one biochemical RNA- and/or chromatin-associated event. These findings suggest

that RNA and especially the family of ncRNAs constitutes an integral part of every cellular process and its "elusive" role remains to be unveiled.

1.2. Non-coding RNAs

During the last several decades numerous publications have revealed a plethora of regulatory RNAs of multiple shapes and sizes, forcing a paradigm shift on the historical notion regarding the roles of RNA in gene expression regulation and various cellular processes in general. Traditionally, the concept of ncRNAs has been utilized as a blanket term for a wide range of molecules which have initially been classified based on their size. Those that are shorter than 200 nucleotides (nts) are generally termed small RNAs. Molecules longer than 200 nts constitute the long non-coding RNA (lncRNA) family. An overview of the most important ncRNA subspecies and their associated function is presented in **Table 1**.

Table 1. Overview of non-coding RNA subfamilies. This table has been designed for the purposes of this dissertation.

Non-coding RNA subfamilies and their function	
Category	**Associated function**
miRNAs	Post-transcriptional gene expression regulation
piRNAs	Silencing of transposable elements during germ line development
siRNAs	Post-transcriptional gene silencing, similar to miRNAs
snRNAs	Post-transcriptional modification of RNAs
snoRNAs	Guidance of chemical modifications of other RNAs
rRNAs	Formation of ribosomal units
tRNAs	Facilitates protein synthesis by carrying amino-acids to ribosomal units
lncRNAs	Epigenetic gene expression regulation, scaffolds, decoys, sponges, transporters

1.2.1. Ribosomal RNAs

One of the first ncRNA subfamily that has been discovered includes ribosomal RNAs (rRNAs) which constitute the predominant component of the ribosomal unit. Depending on the cell type, rRNA synthesis accounts for the majority of transcriptional activity due to increased demand for ribosome and protein production. Structural cleavages and

nucleotide modifications occur as the ribosomal protein counterparts are incorporated leading to the maturation of subunits (Planta & Mager, 1998).

1.2.2. Transfer RNAs

Another ncRNA subspecies that plays an important role in protein synthesis includes transfer RNAs (tRNAs) which are responsible for carrying amino acids to the ribosomal units. Like rRNAs, tRNAs are highly transcribed leading to the production of, in example, approximately 3 million molecules in yeast per generation (Waldron & Lacroute, 1975) compared to roughly 60 thousand messenger RNAs (mRNAs) (Ares, Grate, & Pauling, 1999).

1.2.3. Small nuclear and nucleolar RNAs

One of the most conserved class of ncRNAs includes small nucleolar RNAs (snoRNAs) whose size varies between ~80 to several hundred nts. Their role is to guide by base pairing the 2′-O-ribose methylation and pseudouridylation of specific rRNA nucleotides (Kiss, 2002; Reichow, Hamma, Ferre-D'Amare, & Varani, 2007) while some of them are required for pre-rRNA endonucleolytic processing. There is increasing evidence that the target repertoire of snoRNAs is not limited to rRNAs (Matera, Terns, & Terns, 2007). Yet another important subfamily of ncRNAs includes small nuclear RNAs (snRNAs) which are metabolically stable molecules of 60-450 nucleotides and reside in the nucleus in the form of small nuclear ribonucleic proteins (snRNPs) (Stanek et al., 2008). These ribonucleic protein complexes are mainly responsible for the removal of pre-mRNA introns leading in the maturation of coding transcripts. A recently discovered group of snRNAs, termed small Cajal body (CB)-specific ribonucleic proteins (scaRNPs), accumulate in Cajal bodies and direct 2′-O-ribose methylation and pseudouridylation of specific nucleotides of spliceosomal snRNAs (Darzacq et al., 2002).

1.2.4. Endogenous small interfering and PIWI-associated RNAs

There are three distinct subfamilies of ncRNAs that are highly associated to gene expression regulation. Endogenous small interfering RNA (endo-siRNA) precursors are derived from repetitive sequences of the genome, antisense pairs or long stem-loop structures which are processed by Dicer resulting to the mature product (~21 nts in length) which is subsequently loaded mainly on AGO2. Endo-siRNAs have been shown to function as post-transcriptional regulators that target RNAs (Czech et al., 2008; Kawamura et al., 2008; Okamura et al., 2008). On the other hand, PIWI-associated RNAs (piRNAs), are 24-31 nts in length and can be derived from genomic regions enriched in retrotransposons through the “ping-pong” mechanism (Siomi, Sato, Pezic, & Aravin, 2011) and mainly from intergenic regions depleted from repetitive elements (Lau et al., 2006). piRNAs are associated with Piwi-subfamily proteins and their biogenesis does not depend on Dicer (Vagin et al., 2006). They are highly abundant in germ cells and they are

involved in transposon silencing through heterochromatin formation and RNA destabilization.

1.3. The discovery of microRNAs

MicroRNAs (miRNAs), which are the main focus of this dissertation, are single stranded RNA molecules (~22 nts in length) that post-transcriptionally regulate gene expression by translation suppression and/or mRNA degradation. The first miRNAs were discovered during developmental progress experimentation in *C. elegans* (R. C. Lee, Feinbaum, & Ambros, 1993). According to this study, *lin-4* gene produces an approximately 22 nt long ncRNA molecule that binds to 3-prime untranslated region (3'UTR) of *lin-14* mRNA inhibiting the translation process. More than six years later, another miRNA, named let-7, has been found to repress *lin-41* expression in *C. elegans* by targeting its 3'UTR (Reinhart et al., 2000). Let-7 is conserved in numerous species and at that time this observation suggested similar regulatory RNAs should exist in other organisms as well (Pasquinelli et al., 2000). In 2001, it became evident that miRNAs are highly abundant in a wide variety of organisms (Ambros, 2001; Lau, Lim, Weinstein, & Bartel, 2001; R. C. Lee & Ambros, 2001). At the same time period miRNAs were associated with a newly proposed mechanism of gene expression regulation, named RNA interference (RNAi) (Fire et al., 1998), allowing them to become a biological research hotspot. The interest in the field continuously mounts ever since.

1.3.1. The growing number of annotated microRNAs

Since 2001, there has been an ever increasing interest in miRNAs which has been followed by an explosion of newly identified miRNAs in a plethora of organisms ranging from single-cell organisms, plants, animals and even viruses. In 2002, the "microRNA Registry" (Griffiths-Jones, 2004) has been established in order to provide guidelines for miRNA annotation and create a repository of the identified miRNA sequences.

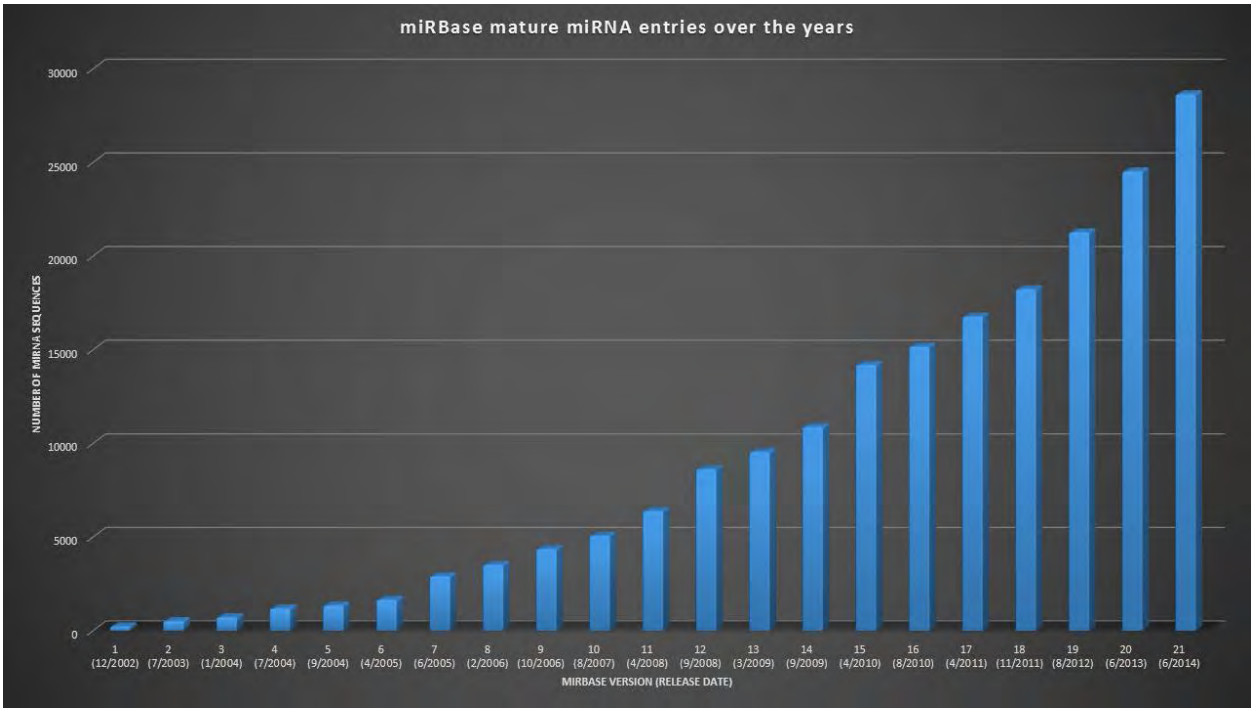


Figure 2. Cumulative number of mature miRNA sequences registered in miRBase repository since the initial launch in 2002. This figure has been designed for the purposes of this dissertation.

Later on, the microRNA Registry became miRBase (Griffiths-Jones, Grocock, van Dongen, Bateman, & Enright, 2006). The initial version of miRBase (2002) included 218 miRNA sequences in 5 species while the latest release (Kozomara & Griffiths-Jones, 2014) contains more than 28,000 sequences in 223 species (**Fig. 2**). The enormous number of annotated miRNAs and the overwhelming rate of their discovery suggest that we have only experienced the tip of the iceberg in miRNA-related research. The aforementioned hypothesis is supported by recent studies (Londin et al., 2015) which have unveiled thousands of novel primate- and tissue-specific miRNA sequences.

1.3.2. Biogenesis of microRNAs

The transcription of the majority of mammalian miRNAs is driven by RNA Polymerase II (Pol2) (Y. Lee et al., 2004) resulting in the formation of capped, polyadenylated and in many cases spliced transcripts, named primary-microRNAs (pri-miRNAs). Over the years, it has become apparent that miRNA genes share the same mechanisms of transcription and post-transcriptional processing with protein-coding genes. There have been studies (Borchert, Lanier, & Davidson, 2006) showing that certain miRNAs could derive from Alu repeat elements, however, these transcripts are processed by RNA Polymerase III (Pol3). More than half of mammalian miRNAs are encoded in close proximity to other miRNA loci. These clustered miRNAs are considered to be derived from a single polycistronic transcription unit (Georgakilas et al., 2014; Y. Lee, Jeon, Lee,

Kim, & Kim, 2002). In general, miRNAs can be divided into two categories depending on their location relative to protein-coding genes. Intragenic miRNAs are located in either the intronic or exonic part of protein-coding genes. On the other hand, intergenic miRNAs are encoded in individual miRNA genes located in the genomic space between coding loci (**Fig. 3**). The majority of intragenic miRNAs share the same promoter with the host gene, however, is some cases they have their own regulatory loci residing in upstream intronic regions. Transcription is an integral part in the mechanism of miRNA biogenesis regulation. A plethora of Pol2-associated transcription factors are involved in the regulation of miRNA genes. In example, MyoD1 induces the transcription of miR-1 and miR-133 during myogenesis (J. F. Chen et al., 2006; Rao, Kumar, Farkhondeh, Baskerville, & Lodish, 2006). Other miRNAs are regulated by tumor suppressive or oncogenic transcription factors such as p53 which enables the transcription of miR-34 family (L. He, He, Lowe, & Hannon, 2007). Oncogenic protein MyC represses numerous miRNAs involved in cell cycle and apoptosis (T. C. Chang et al., 2008).

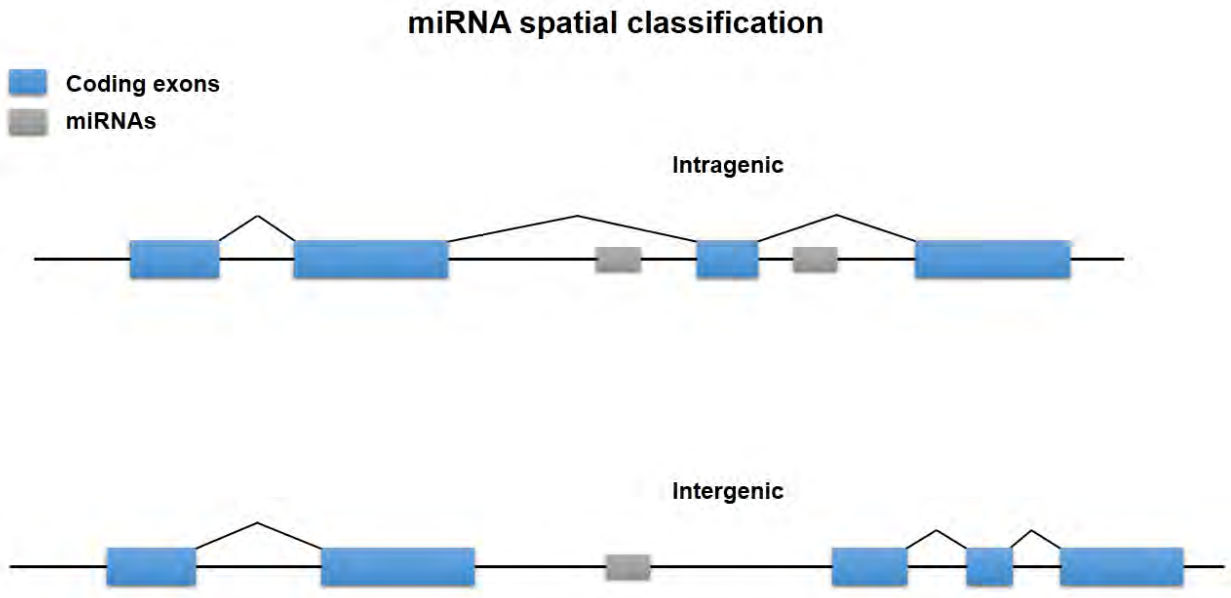


Figure 3. Classification of miRNAs based on their localization in respect to protein-coding genes. This figure has been designed for the purposes of this dissertation.

The initial products of Pol2-mediated miRNA gene transcription are pri-miRNAs whose length varies between a few hundred to several hundred kilobases (kb). Pri-miRNAs contain hairpin-like structures (**Fig. 4**), named precursor-microRNAs (pre-miRNAs), which can be identified by the nuclear RNase III type protein Drosha (Y. Lee et al., 2003).

This step of miRNA biogenesis pathway is localized in the nucleus and requires another protein, named DGCR8, in order for a large dimer known as Microprocessor complex to be formed (Denli, Tops, Plasterk, Ketting, & Hannon, 2004; Gregory et al., 2004; Landthaler, Yalcin, & Tuschl, 2004). Typically, pre-miRNAs include the mature miRNA and its complementary sequence, the stem which is approximately 33 nts in length and flanking single stranded RNAs (ssRNAs). Drosha is able to identify and cleave the substrate ~11 bp away from the ssRNA-stem junction. This process is catalyzed by the interaction between DGCR8, the stem loop and the ssRNA segments (Han et al., 2006; Zeng & Cullen, 2005). There is increasing evidence that pri-miRNA processing may be a co-transcriptional process. In cases where the hairpin structure is located inside exonic region of protein-coding genes, the cleavage by the Microprocessor complex is able to induce reduced protein production. In addition, Drosha is also able to identify mRNAs that contain long hairpins. An intriguing example is the deregulation of DGCR8 protein synthesis caused by the identification and cleavage of hairpin structures by Drosha (Han et al., 2009).

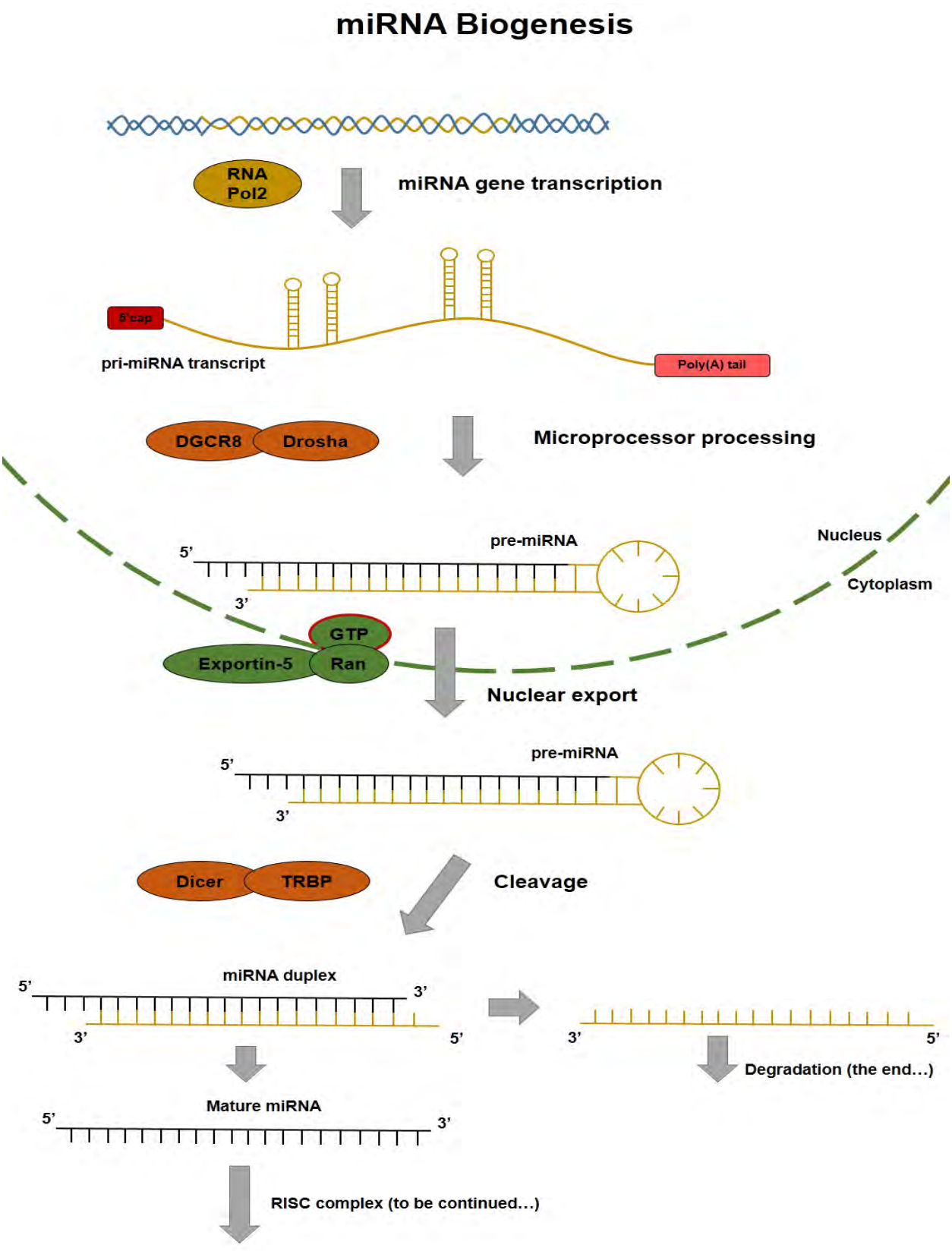


Figure 4. Overview of miRNA biogenesis. This figure has been designed for the purposes of this dissertation.

Pre-miRNAs, released from primary transcripts in the nucleus, are subsequently exported to the cytoplasm by a member of the nuclear transport receptor family of proteins, named EXP5. This process is facilitated by the recognition of more than 14 nts of pre-miRNA's stem and a short 3' overhang of 1 to 8 nts in length, by EXP5 (V. N. Kim, 2004). Upon reaching cytoplasm, RNase III type protein Dicer cleaves ~22 nts away from the stem base of the hairpin structure releasing miRNA duplexes of approximately 18-23 nts in length (Hutvagner et al., 2001). One strand of the produced RNA duplex will be subsequently loaded to the RNA-induced silencing complex (RISC) complex while the other strand is typically degraded. In some cases, some pre-miRNAs produce mature sequences from both strands that survive and are functional in comparable frequencies (Khvorova, Reynolds, & Jayasena, 2003).

1.3.3. Function of mature microRNAs

In the final step of miRNA biogenesis pathway, the mature products are loaded onto complexes, either referred to as ribonucleic proteins (RNPs) or RISCs, following Dicer processing and removal of the stem loop from the hairpin structures (**Fig. 5**). The major component of RISC complex is the Argonaute family of proteins (Peters & Meister, 2007; Tolia & Joshua-Tor, 2007) which utilizes the loaded mature miRNA sequence and its complementarity to miRNA recognition elements (MREs) located on mRNAs as a guide in order to regulate gene expression. Mammals contain four AGO proteins, AGO1-4. AGO2 is the only member of Argonautes that has an RNaseH-like PIWI domain. miRNPs often include other proteins, beside AGO1-4, which most likely act as microRNA ribonucleic protein (miRNP) assembly or as regulatory factors catalyzing the repressive miRNP functions (Peters & Meister, 2007).

In plants, miRNAs are paired to MREs with nearly perfect complementarity and cause mRNA cleavage by a mechanism that resembles mammalian RNAi (Jones-Rhoades, Bartel, & Bartel, 2006). A similar mechanism is sometimes utilized by vertebrate and viral miRNAs. However, in most instances, metazoan miRNAs form imperfect pairs with their target regions, following a set of rules derived by experimental and bioinformatics analyses (Brennecke, Stark, Russell, & Cohen, 2005; Doench & Sharp, 2004; Grimson et al., 2007; Lewis, Burge, & Bartel, 2005; Nielsen et al., 2007). These studies combined reporter genes assays and miRNA overexpression experiments and revealed that the major determinant of target specificity is the perfect base pairing between the 5' ends of miRNAs, nucleotides 2-7 (or -8) in particular, and 3'UTRs (**Fig. 6**). miRNAs can also act as post-transcriptional regulator by binding on MREs located on 5-prime untranslated regions (5'UTRs) and coding sequence (CDS) regions of mRNAs (Hafner et al., 2010; Kloosterman, Wienholds, Ketting, & Plasterk, 2004; Lytle, Yario, & Steitz, 2007). However, there are many reports suggesting regulation of sites without perfect seed

complementarity (Betel, Koppal, Agius, Sander, & Leslie, 2010; Brennecke et al., 2005; Didiano & Hobert, 2006; Lal et al., 2009; Vella, Choi, Lin, Reinert, & Slack, 2004).

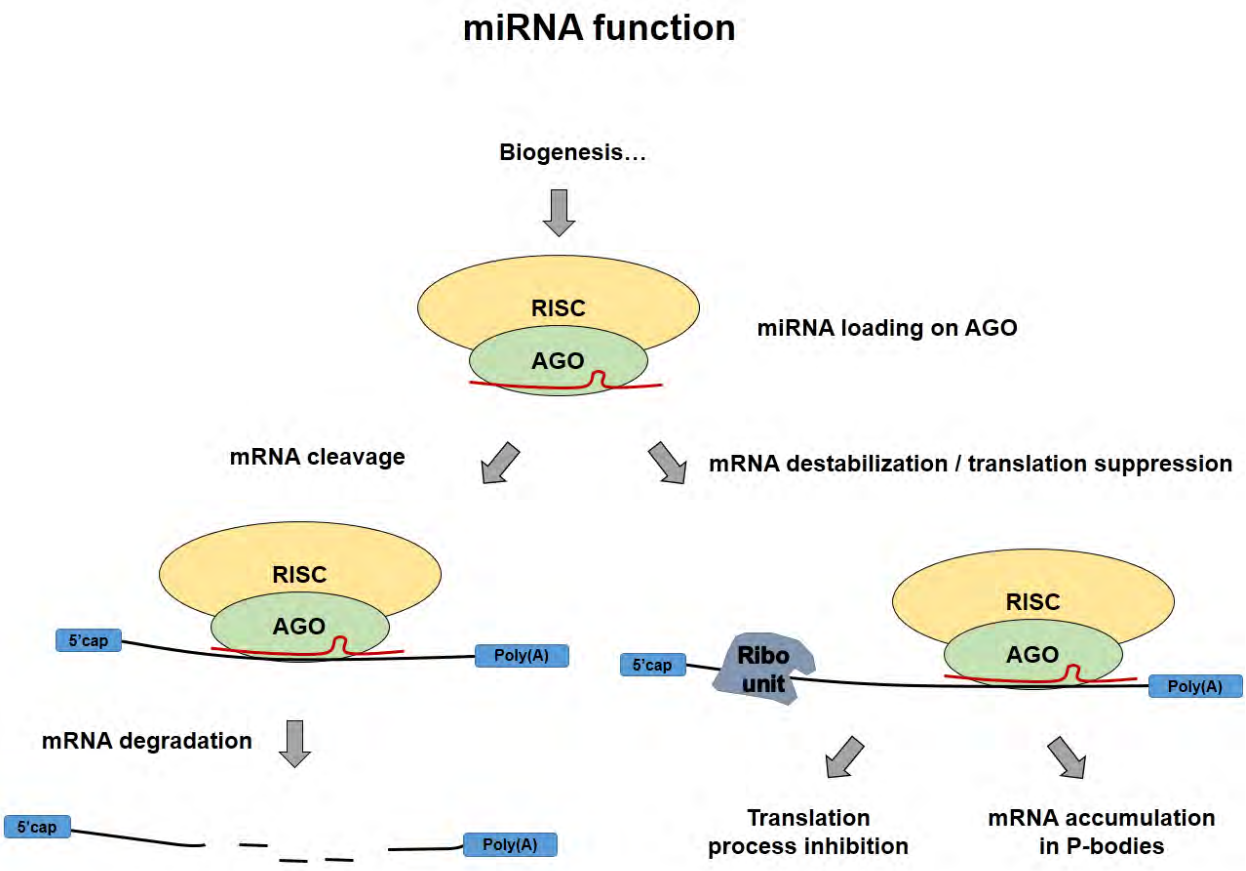


Figure 5. Overview of miRNA function through the RNAi machinery. This figure has been designed for the purposes of this dissertation.

Most experimental effort has been focused on verifying seed-based interactions, however, cases of non-seed based interactions have also been demonstrated, although far less frequently (Bartel, 2009; Chi, Hannon, & Darnell, 2012; Grimson et al., 2007). The most frequent cases appear to be "seed-like" with mismatches or wobbles in positions 5, 6, and 7, and "G-bulge" sites where the mRNA nucleotide that would normally pair with position 6 of the miRNA is bulged out of the interaction (Chi et al., 2012). The development of experimental protocols such as RNA isolated by crosslinking immunoprecipitation (HITS-CLIP) and photoactivatable-ribonucleoside-enhanced crosslinking and immunoprecipitation (PAR-CLIP), unveiled the localization of AGO binding sites on the transcriptome (Chi, Zang, Mele, & Darnell, 2009; Hafner et al., 2010).

A sizable portion of the CLIP-derived AGO binding sites did not contain seed matches. However, it was unclear whether this type of targeting was caused by miRNA independent mechanisms or by non-canonical miRNA-target interactions. In a recent study (Loeb et al., 2012), the authors attempted to address this issue by combining genetic, biochemical and computational approaches. The analysis included MREs derived from differential HITS-CLIP and mRNA expression changes isolated from wild-type or miR-155 deficient mice. The results confirmed that exact complementarity between the seed region and the MREs was present in the majority of miR-155-associated binding sites. On the other hand, perfect seed complementarity was absent in ~40% of binding regions. These non-canonical pairs were strongly enriched in inexact seed matches and contain a mismatch in the seed at a single nucleotide position. These non-canonical miRNA binding sites regulate gene expression with lower potency than canonical sites.

Recent studies unveiled the increased evolutionary conservation of the central region of miRNAs (Grimson et al., 2007). This observation initiated a search for "centered site"-mediated miRNA activity, showing that 11 nucleotides of perfect complementarity starting at position 3, 4 or 5 could inhibit mRNA translation (Shin et al., 2010). However, these centered sites were observed only occasionally within the human miRNA targetome, with frequency similar to 3-prime supplementary and 3-prime complementary sites that account for less than 10% of interactions (Bartel, 2009) when combined together. Imperfect centered sites could also occur more frequently, according to a recent study (Martin et al., 2014). The authors employed the biotin pull-down approach (Cloonan et al., 2011) in order to identify the direct targets of miR-10a and miR-10b, which share identical seed sites, but differ by a single nucleotide in the center region. The results demonstrated that imperfect-centered sites occur frequently, a finding that may explain the evolutionary conservation of the central region of miRNAs. Examples of different types of binding categories between miRNAs and MREs are presented in **Figure 6**.

miRNA-mediated gene expression regulation can be caused by either direct mRNA cleavage and degradation or mRNA destabilization and translation suppression. Initial evidence regarding mRNA degradation unveiled that extensive pairing complementarity directs AGO-catalyzed mRNA cleavage (Hutvagner & Zamore, 2002; J. J. Song, Smith, Hannon, & Joshua-Tor, 2004; Yekta, Shih, & Bartel, 2004). Additional studies (Bagga et al., 2005) have continued to shed light on this type of regulation by revealing that let-7 promotes degradation of lin-41 target mRNA in *C. elegans* and does not require perfect base-pairing. This kind of mRNA decay is also referred to as "slicer" activity. One additional requirement for miRNA-mediated mRNA cleavage is AGO2 protein to be present inside the RISC complex due to its catalytic RNaseH domain.

An additional layer of miRNP-associated gene expression regulation involves the RNA degradation machinery. Eukaryotic cells contain two well-conserved pathways for the degradation of mRNA, both of which require initial removal of the 3-prime polyadenylated (3′poly(A)) tail in a process known as deadenylation (Parker & Song, 2004). Usually, deadenylation is followed by 3′-to-5′ exonucleolytic degradation by the exosome.

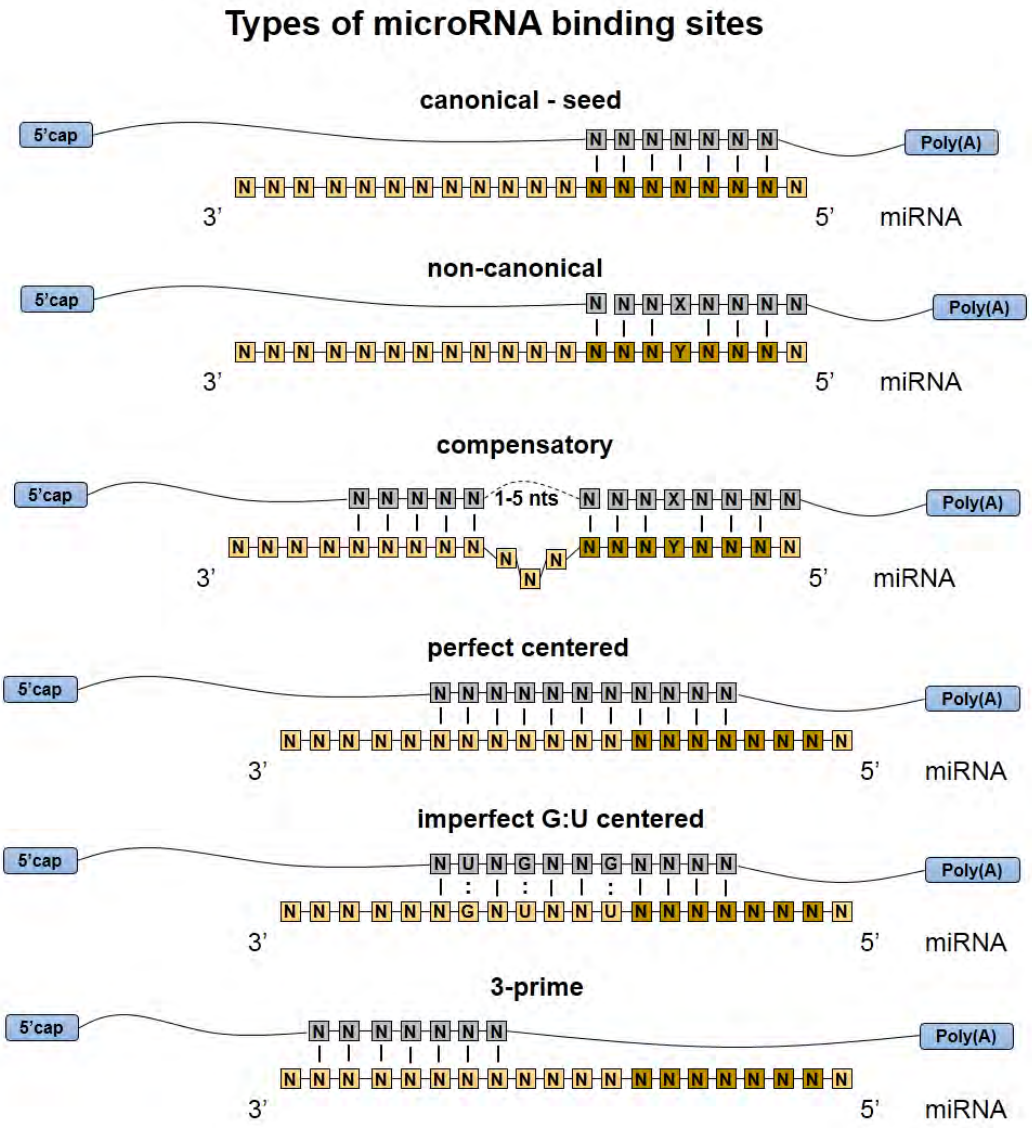


Figure 6. Examples of different types of miRNA binding sites. This figure has been designed for the purposes of this dissertation.

Alternatively, after deadenylation, mRNAs can be decapped by the Dcp1/2 decapping enzymes and degraded by the 5′-to-3′ exoribonuclease, Xrn1p. Initial evidence that

miRNAs may mark mRNAs for decapping were presented in studies comparing the subcellular localization of Argonaute proteins and various components of the decapping machinery. In a range of eukaryotic organisms the enzymes and several activators associated with the decapping process are concentrated in specific cytoplasmic foci known as cytoplasmic processing bodies (P-bodies), which can be sites of mRNA decapping and degradation (Cougot, Babajko, & Seraphin, 2004; Sheth & Parker, 2003). Other studies have shown that all four versions of the mammalian AGO proteins are concentrated in P-bodies and can co-immunoprecipitate with the decapping enzyme (Jakymiw et al., 2005; Liu, Valencia-Sanchez, Hannon, & Parker, 2005; Pillai et al., 2005; Sen & Blau, 2005). miRNA targets also accumulate within P-bodies in a miRNA-dependent manner (Liu, Rivas, et al., 2005). According to real time polymerase chain reaction (RT–PCR) analysis, the majority of a specific mRNA repressed by let-7, is found in a complex containing P-bodies (Liu, Rivas, et al., 2005). Based on these results, a hypothesis emerged suggesting that miRNAs target mRNAs to P-bodies, reducing their levels by decapping and 5′-to-3′ degradation.

The final layer of miRNA-mediated gene silencing includes translation suppression. Translation initiation process is based on a series of key steps (Kapp & Lorsch, 2004). 5′cap is recognized by the cap-binding protein known as eIF-4E which is part of the eIF-4F initiation complex. This complex recruits another complex that contains eIF3, the 40S ribosomal subunit, and a ternary complex of eIF2, GTP as well as the initiator tRNA. The 40S subunit subsequently scans 5′UTR until an AUG start codon is recognized, enabling the 60S subunit to join and begin the elongation phase of translation. First, translation initiation can be inhibited by affecting the ability of the mRNA to complete a step in the initiation process (Richter & Sonenberg, 2005). Alternatively, translation initiation can also be repressed by a competition between P-body mRNP and translation initiation associated complex, suggesting a model where cytoplasmic mRNAs are in equilibrium between translation complexes and P-body mRNPs (Brengues, Teixeira, & Parker, 2005; Coller & Parker, 2005). Moreover, mRNA-specific repression complexes might be involved into this general competition. In the initial publication (R. C. Lee et al., 1993) where miRNAs were first introduced it was shown that the lin-4 miRNA reduced the amount of lin-14 protein while the abundance of lin-14 mRNA remained unchanged. Although recent observations suggest that the lin-4 might also affect mRNA levels (Sheth & Parker, 2003), there are now multiple examples where silencing by a miRNA is observed with either no change in the mRNA level, or with a significantly smaller decrease in mRNA levels than the one observed for the corresponding protein (Brennecke, Hipfner, Stark, Russell, & Cohen, 2003; X. Chen, 2004; Cimmino et al., 2005; Poy et al., 2004). These results suggest two possibly overlapping miRNA-RISC mechanisms that repress translation. In the first one, one of the RISC components inhibits a certain translation initiation factor and forces the target mRNA to exit the ribosomal

units and accumulate in P-bodies. Alternatively, or in parallel, RISC might contain or recruit proteins that induce the formation of mRNPs that can accumulate within P-bodies and be excluded from the translation machinery. It should be noted that P-bodies are dynamic structures and mRNAs are able to cycle in and out of these structures. As a result, the RISC-mediated translation repression could be a kinetic effect regulating the P-bodies entry and exit rate of mRNAs.

1.3.4. Interplay between physiological/pathological conditions and miRNA function

Since their first discovery in 1993, miRNAs have been vigorously researched for their implication in various physiological and pathological states in a plethora of organisms. Ever since and due to intense research, there has been an explosion of miRNA-related publications which according to PubMed-derived statistics are estimated to exceed 40,000 (Sep, 2015).

Numerous studies have reported a tight interplay between miRNA expression and important mechanisms responsible for the development of various species. Small alterations in miRNA biogenesis pathway, in example the inhibition of critical components of pri- and pre-miRNA processing machineries, result in global inhibition of miRNA expression leading to lethality in early embryonic stages (Bernstein et al., 2003; Y. Wang, Medvid, Melton, Jaenisch, & Blelloch, 2007). The role of miRNAs in differentiation is not limited to embryonic development. In example, conditional *Dicer*-knockout mice models in myogenic tissues result in abnormal morphology of muscle fibers (O'Rourke et al., 2007). The normal development of the haematopoietic lineage has been found to depend on proper expression of certain miRNAs regulated by cell type specific transcription factors (TFs). GATA1 activates the transcription of miR-451 and miR-23 which are responsible for the final differentiation of erythroid progenitors (Dore et al., 2008; Zhu et al., 2013). Neural stem cell differentiation in mouse is controlled by several self-reinforcing feedback loops that regulate the expression of neuronal-related miRNA. REST-SCP1 protein complex is known to silence neuronal genes in non-neuronal cells by repressing neuronal-specific miR-9 and miR-124. However, miR-124 feeds back to the REST-SCP1 complex by targeting SCP1, thus suppressing its activity during neuronal differentiation (Visvanathan, Lee, Lee, Lee, & Lee, 2007).

Since the early stages of miRNA research, cancer has been the most prominent amongst human diseases with a clear role for miRNA regulation. Initial evidence involved the observation of lower miR-15 and miR-16 abundance in 65% of B-cell chronic lymphoma leukemia patients (Calin et al., 2002). Subsequent expression profiling studies further established the connection between aberrant miRNA expression patterns of miR-125b, miR-145, miR-21 and miR-155 and the increased risk of breast cancer (Iorio et al., 2005). Up-regulation of miR-155 and down-regulation of let-7a have been correlated with low

survival rates of lung cancer patients (Yanaihara et al., 2006). Additional studies highlight the imbalance between cell death and proliferation during the development of various types of cancer such as hepatocellular carcinoma (Shimizu et al., 2010), breast cancer (Kong et al., 2010) and adenocarcinoma (Cho, Chow, & Au, 2011). Depending on the target repertoire of malignancy-related miRNAs, they can be divided in two categories, tumor suppressors and tumor promoters. The first category refers to miRNAs that target genes whose expression is aberrantly increased and promote cancer by regulating relevant pathways. Usually these miRNAs are down-regulated in pathological conditions. The second category includes miRNAs that are usually up-regulated in pathological conditions and target genes that exhibit low expression and their role is to disrupt cancer-related pathways. In example, miR-15 and miR-16 are considered tumor suppressors since they target anti-apoptotic gene *BCL2* and promote cell death in cancers (Cimmino et al., 2005). On the other hand, miR-21 has been found to directly act as tumor promoter in breast cancer (Iorio et al., 2005) and glioblastomas (Chan, Krichevsky, & Kosik, 2005). The role of miRNAs in malignancies extends in every hallmark of cancer including invasion and metastasis. Analyses of miRNA expression profiling have associated the continuously declining expression of miR-145 with gradual progression of primary gastric cancer and secondary metastasis (Gao et al., 2013) while the up-regulation of miR-210 has been associated with invasive transition of breast cancer (Volinia et al., 2012).

The type of pathological conditions associated to miRNAs are not limited to cancers. Many immune-related diseases such as fatty liver disease, systemic lupus erythematosus, type I/II diabetes and multiple sclerosis have shown established connections with specific miRNAs. Numerous mature miRNAs were identified as signatures by analyzing the expression profiles of healthy controls and relapsing multiple sclerosis patients (Keller et al., 2009). In two independent studies including hundreds of systemic lupus erythematosus patients and healthy controls, the decreased miR-146a expression was strongly correlated with increased risk for the disease among European and Asian populations (Lofgren et al., 2012; Luo et al., 2011). Similar studies have also identified miRNAs related to type II diabetes such as miR-144, miR-146a, miR-150 and miR-182 (Karolina et al., 2011). In addition, over-expression of miR-200a, miR-200b and miR-429 and down-regulation of miR-122, miR-451 and miR-27 was connected to diet-mediated nonalcoholic fatty liver disease in rats (Alisi et al., 2011). The pathogenesis of neuronal degeneration such as Parkinson and Alzheimer still remains poorly understood. There is a progressively growing number of studies attempting to shed light on the implication of miRNAs in such diseases. In example, expression deregulation of miR-133b might contribute to the progression of Parkinson's disease, since the miR-133b-*PIXT3* feedback loop is essential for maintaining dopaminergic neurons in the brain (J. Kim et al., 2007). Similarly, miR-29a, miR-29b-1 and miR-9 have been found to be significantly down-

regulated in Alzheimer's disease patients (Hebert et al., 2008), resulting in aberrant over-expression of their target BACE1, a critical protein for the disease's pathogenesis (Willem et al., 2006).

Beside eukaryotic miRNAs, viral-encoded miRNAs have been discovered in multiple DNA viruses as well. The first cases of viral-encoded miRNAs have been derived from a Burkitt's lumphoma cell line which was infected by Epstein-Barr virus (Pfeffer et al., 2004). Since then, bioinformatics and cloning approaches we utilized in order to identify viral miRNAs in polyoma virus (Sullivan, Grundhoff, Tevethia, Pipas, & Ganem, 2005), adenovirus (Andersson et al., 2005) and several subtypes of the herpes viruses (Cai et al., 2005). The function of viral miRNAs mainly focuses on targeting host genes that assist the cell to enter the apoptotic cycle. However, other DNA viruses such as papillomaviruses and poxviruses do not encode any miRNAs. Since most miRNAs are generated from endonucleolytic processing of longer transcripts, the common conception was that RNA viruses will not encode miRNAs in order to avoid unproductive cleavage of their genome or mRNAs. Initial studies did not identify any viral miRNAs in a wide range of RNA viruses (Skalsky & Cullen, 2010). Later on, it was discovered that bovine leukemia retrovirus with an RNA genome, encodes a conserved cluster of miRNAs that are transcribed by RNA polymerase III (Kincaid, Burke, & Sullivan, 2012). Another recent example is HIV-1 which has been found to encode miR-H3 by combining computational prediction and deep sequencing. Overexpression of miR-H3 increases viral production and artificially induced mutations in miR-H3 sequence significantly impair the viral replication of wildtype HIV-1 viruses, suggesting that this is a replication-enhancing miRNA.

There is no doubt that the number of publications that attempt to elucidate the role of miRNAs in disease will keep increasing. Such studies are destined to go hand-in-hand with research dedicated to fill the puzzle of biological pathways and regulatory networks. During the last decades, miRNAs have been successfully incorporated in networks regulating gene expression, however, in the last few years lncRNAs have been identified as an additional layer of regulation that exhibits a strong interplay between miRNAs and mRNAs.

1.4. microRNA target prediction algorithms

Despite the fact that there has been a significant increase of experimentally validated miRNA binding sites (Vergoulis et al., 2012; Vlachos, Paraskevopoulou, et al., 2015) knowledge regarding the majority of miRNA target genes remains elusive. Therefore, *in silico* target prediction methodologies are still considered the only rapid and cost-free source of putative miRNA target identification. During the last decade, several miRNA target prediction algorithms have been established. The majority of these *in silico*

techniques utilize the alignment between the miRNA seed region to the mRNA sequence of candidate target genes as the main prediction feature. Their predictive power is enhanced by measuring the evolutionary conservation of the binding region, identifying the accessible regions of the mRNA, characterizing the nucleotide composition of the region surrounding the binding site and taking into account the location of the binding sites within the mRNA.

1.4.1. DIANA-microT-CDS

DIANA-microT-CDS (Reczko, Maragkakis, Alexiou, Grosse, & Hatzigeorgiou, 2012) is the latest version of microT algorithm which is capable of predicting miRNA targets in both 3′UTR and CDS regions of protein-coding genes. The algorithm was trained on positive and negative sets of MREs derived from publically available PAR-CLIP datasets (Hafner et al., 2010). The analysis of the experimentally validated MREs was performed independently for CDS and 3′UTR enabling the identification of region-specific binding features. Separate prediction models are trained for 3′UTR and CDS regions which are subsequently combined in order to produce a final score characterizing the interaction strength and quality. DIANA-microT-CDS was the first and until recently the only algorithm capable of identifying protein-coding genes that are targeted in the CDS and not in the 3′UTR.

The algorithm is available through the latest version of DIANA-microT Web Server (Paraskevopoulou, Georgakilas, Kostoulas, Vlachos, et al., 2013). This major update is able to detect more than 11 million interactions between 3,876 miRNAs and 64,750 protein-coding genes in *H. sapiens*, *M. musculus*, *D. melanogaster* and *C. elegans*. Furthermore, it has been upgraded to miRBase v18 (Kozomara & Griffiths-Jones, 2011) and Ensembl v69 (Flicek et al., 2012). DIANA-microT Web Server v5 hosts numerous integrated analyses in the form of ready-made advanced pipelines, covering a wide range of inquiries regarding predicted or validated miRNA:gene interactions and their impact on metabolic and signalling pathways. These pipelines can be used to analyze user data derived from small scale and high-throughput experiments directly from the DIANA-microT Web Server interface, without the necessity to install or implement any kind of software.

1.4.2. ElMMo

ElMMo (Gaidatzis, van Nimwegen, Hausser, & Zavolan, 2007) is a miRNA target prediction algorithm which is based on Bayesian theory. Every binding site of all available miRNAs are predicted in a distinct set of organisms including flies, worms, fish and mammals. For each miRNA, the homologous binding sites are utilized in order to model their evolution in a set of related species. The algorithm explicitly infers the phylogenetic distribution of functional binding sites, independently for each miRNA and allows to identify species- and clade-specific miRNA binding. ElMMo serves as the basis

for the association of miRNAs and specific biochemical pathways by analyzing miRNA targets and their association to KEGG database (Kanehisa, Goto, Sato, Furumichi, & Tanabe, 2012).

1.4.3. miRanda

miRanda (B. John et al., 2004) is a target prediction technique that utilizes a two-step approach. Initially, dynamic programming is applied in order to identify local alignments between miRNA and 3'UTR sequences that correspond to double-stranded antiparallel duplexes. Mismatches and different wobble pairs are weighted accordingly depending on their naturally occurred frequency, as derived from experimental procedures. In addition, complementarity scores at the first eleven positions are weighted by a different scaling factor than the rest of the binding site in order to reflect the experimentally observed asymmetry of the binding region complementarity. The resulting set of binding sites is subjected to a conservation analysis which includes the identification of homologous sites in human, mouse, rat, fugu and zebrafish. The set of MREs is subsequently filtered depending on the level of conservation.

1.4.4. Pictar

Pictar (Lall et al., 2006) target prediction algorithm distinguishes the identified miRNA binding sites in two categories. The first one includes MREs that exhibit perfect complementarity between the miRNA seed and the 3′UTR binding region. The second category includes MREs whose perfect complementarity is interrupted by at most one nucleotide bulge, mismatch, or G:U wobble. In any case, the algorithm requires that the binding stability of the putative interaction exceeds a specified threshold. The algorithm subsequently labels highly conserved binding sites sites as "anchors" and removes 3′UTRs that do not contain enough anchors. The likelihood of 3′UTRs being targeted by miRNAs in a combinatorial manner is accessed by applying a Hidden Markov model. These scores are computed for a set of species and combined to compute the final score.

1.4.5. PITA

PITA (Kertesz, Iovino, Unnerstall, Gaul, & Segal, 2007) is a target prediction algorithm that follows the typical seed parameter settings of length 6 to 8 bases, beginning at position 2 of the mature miRNA sequence. This setup does not allow any mismatches or loops, except from a single G:U wobble in 7- or 8-mers. The interaction is described by an energy score which represents the difference between the energy gained by the miRNA binding on the target sequence and the energy required to make the binding regions accessible. In order to combine the score of multiple binding sites for a single miRNA on the same mRNA into a total interaction score, PITA computes the statistical weight of all configurations in which exactly one of the sites is bound by the miRNA sequence.

1.4.6. RNA22

The RNA22 (Miranda et al., 2006) target prediction algorithm incorporates the identification of redundant patterns in mature miRNA sequences. The statistical significance of these patterns is approximated by a second-order Markov chain. The algorithm subsequently identifies the reverse complement of all miRNA patterns within 3′UTR regions. Regions exhibiting increased accumulation of reverse complement hits are characterized as "target islands". The association between miRNAs and target islands is accomplished by measuring the strength of the pairing which is calculated based on the free energy and the number of nucleotides involved.

1.4.7. TargetScan

The latest version of TargetScan (Garcia et al., 2011) algorithm is based on several observations regarding the sequence surrounding experimentally derived binding sites such as the ones that occur between LSY6 mRNA and miR-23. Therefore it considers seed-pairing-stability and target-site abundance as two independent variables when performing multiple linear regression. The remaining parameters are described by the context score which models the relative contributions of previously identified targeting features such as site type, site number, site location, local AU content and 3′-supplementary pairing.

1.4.8. TargetSpy

TargetSpy (Sturm, Hackenberg, Langenberger, & Frishman, 2010) is an in silico target prediction methodology that does not require the presence of seed match. The algorithm considers a wide range of sequence and structural characteristics such as the general extent of miRNA:mRNA binding, G:U base pairing, bulge-related features of duplexes, position specificity, GC content and accessibility. The model does not rely on evolutionary conservation, which allows the detection of species-specific interactions.

1.4.9. TargetS

TargetS (Xu, San Lucas, Wang, & Liu, 2014) algorithm is capable of predicting miRNA binding sites located along entire gene sequences permitting the identification of targets beyond 3′UTR. The alogrithm is based on both canonical and non-canonical seed pairing but it does not rely on evolutionary conservation. It additionally incorporates the stability between miRNA:mRNA bindings as well as the remaining free energy of the interaction without considering any context-related sequence features.

1.4.10. Comparison of target prediction algorithms

In the most recent review (Alexiou, Maragkakis, Papadopoulos, Reczko, & Hatzigeorgiou, 2009) of target prediction methods, the performance of early versions of DIANA-microT, PITA, Pictar, Targetscan, ElMMo and RNA22 were evaluated. This was

accomplished by utilizing measured changes of protein levels after over- or under-expression of specific miRNAs (Baek et al., 2008; Selbach et al., 2008). The evaluation process showed that programs relying on the evolutionary conservation of the seed or an extension of the seed region perform better, exhibiting ~50% precision and 6 to 12% sensitivity. In addition, all possible union and intersection combinations of the aforementioned programs were calculated in order to assess the performance of merged prediction sets. It was observed that in most cases, an accurate algorithm was better than a combination of predictions.

Since then, some of the algorithms (i.e. DIANA-microT, TargetScan) have been updated to new and improved versions, others were never upgraded (i.e. ElMMo, PITA, Pictar) and in some cases, novel methodologies emerged (i.e. TargetSpy, TargetS). In the publication of the latest version of DIANA-microT (Reczko et al., 2012), the algorithm was evaluated against the measured changes of protein levels that were observed in pSILAC experiments (Selbach et al., 2008) and its performance was compared to TargetScan v5 (Friedman, Farh, Burge, & Bartel, 2009), PicTar (Lall et al., 2006), RNA22 (Miranda et al., 2006), miRanda (B. John et al., 2004), DIANA-microT-v3 (Maragkakis et al., 2009) and a seed measure, whose prediction score is defined through the number of miRNA seed matches on 3′UTR of protein-coding genes. Sensitivity and precision were measured at different prediction score thresholds and showed that DIANA-microT-CDS program exhibits the highest sensitivity at any level of specificity.

During the last few years there has been an increased accumulation of experimentally verified miRNA:mRNA interactions (Vergoulis et al., 2012; Vlachos, Paraskevopoulou, et al., 2015). Such repositories include interactions validated with both low- and high-throughput experimental techniques and should be utilized for fine-tuning target prediction methodologies as well as performing extensive evaluation assays of their performance.

1.5. Repositories of experimentally validated microRNA targets

Bioinformatics algorithms and tools are playing a significant role in miRNA target identification. Such algorithms, which have been presented in previous sections of this dissertation, attempt to tackle the problem computationally. Some targets can be confidently predicted with currently available techniques. However, precision and sensitivity of state-of-the-art algorithms were estimated as ~50% and 12%, respectively, when tested against proteomics supported miRNA targets (Reczko et al., 2012), highlighting the necessity for mass experimental miRNA target validation.

miRNA targets can be experimentally verified with gene-specific, as well as high-throughput techniques. Specific techniques include reporter gene assays, assessment of

miRNA and target mRNA co-expression, in example northern blotting or qPCR, and estimation of miRNA effect on target protein such as ELISA, western blotting and immunohistochemistry (Kuhn et al., 2008). High-throughput techniques can be a simple extension of an existing gene-specific technique in a high-throughput setting, for example the utilization of microarray screening instead of qPCR. They can also involve novel relevant methodologies, such as RNA-Seq, HITS-CLIP, PAR-CLIP, biotin tagging of miRNAs, parallel analysis of RNA ends (PARE) and various proteomics approaches such as SILAC.

As the relevant literature and the number of experiments increase with a super linear rate, databases that curate and collect experimentally verified miRNA targets have gradually emerged. Their aim is to face this challenge by providing a significant increase of available miRNA targets derived from all contemporary experimental techniques (gene specific and high-throughput).

1.5.1. miR2Disease

miR2Disease (Jiang et al., 2009) was first released in 2009. It is a manually curated database that aims to provide information regarding miRNA-related pathologies. The database includes 809 miRNA:gene interactions for *H. sapiens*, coupled with related disease information derived from relevant literature. The 3,273 miRNA disease-related entries consist the strongest point of the database (last updated 14-Mar-2011). The user can search by miRNA, target gene or disease name. Further details include method of validation, relation with the pathology, manuscript information and links to target predicting algorithms.

1.5.2. MirnaMAP

MirnaMAP (S. D. Hsu et al., 2008) was first released in 2006. It contains data derived from an outdated version of TarBase (346 targets) and by manual curation (29 targets). MirnaMAP has not been updated since 2008 and contains a limited amount of experimentally validated targets for *H. sapiens*. The largest amount of mirRNAMAP entries is based on predicted interactions for 2,464 miRNAs in 12 species. MirnaMAP provides a wealth of available data for each database entry, including miRNA and gene information, bead-array miRNA tissue expression profile, qPCR tissue expression profile, predicted target genes, as well as relevant literature.

1.5.3. MiRecords

MiRecords (Xiao et al., 2009) was first released in 2009. It contains manually curated and predicted miRNA targets. The validated targets component of the database contains 2,286 interactions between 548 miRNAs and 1,579 target genes in nine species (last update 25-Nov-2010). At the time of writing of this dissertation, the official web site was inaccessible. The largest number of those interactions is derived from gene-specific

experiments. The database provides miRNA, gene and target site-related information, as well as links to miRBase and RefSeq. miRNA:gene interactions are supported with data regarding manuscript information, experimental method used for validation, as well as a selected passage from the manuscript stating the experimental result. However, the user does not have the ability to filter results based on any of the available predicted or validated component fields. The miRecords interface also enables the user to insert new interactions.

1.5.4. miRSel

miRSel (Naeem, Kuffner, Csaba, & Zimmer, 2010) database was first released in 2010. It contains miRNA interaction data derived solely from text mining of MedLine abstracts. The text mining algorithm manages to extract miRNA:gene associations with 65% precision, 90% recall, based on a test performed on 89 selected sentences, derived from 50 PubMed abstracts. MiRSel contains 3,690 miRNA:gene text mined associations. By applying less stringent criteria, the user can have access to approximately 8,000 pairs, which are deemed as less reliable by the developers. In miRSel, the user can also search for miRNAs related to specific MedLine articles that contain a subset of desired terms or that are related to Gene Ontology entries. Links to external databases such as miRBase and Entrez Gene are provided for each entry. Information regarding the experimental method used for miRNA target validation is not available. Data derived from other curated miRNA interaction databases such as TarBase v5, miR2Disease and miRecords have also been integrated.

1.5.5. miRTarBase

miRTarBase (S. D. Hsu et al., 2014) was first released in 2011. The latest version, which was released in 2014, includes manually curated data for 51,460 experimentally verified interactions between 17,520 genes and 1,232 miRNA in 14 species (last update 27-May-2015). It provides information related to the miRNA, the target gene and the target site. In many cases, where the articles do not explicitly present target site information, miRTarBase can provide predicted regions by using a computational target prediction algorithm. Information regarding available experimental findings supporting the interaction is also included. The user-interface provides links to external data sources such as NCBI Entrez, UCSC Genome Browser, miRBase, BioGPS, iHOP and HGNC. Optionally the user can submit data for non-indexed interactions.

1.5.6. miRWalk

miRWalk (Dweep, Sticht, Pandey, & Gretz, 2011) was first released in 2011. It provides experimentally supported miRNA targets identified solely from text-mined abstracts available in MedLine. The latest version of the database incorporates in silico predicted as well as experimentally derived interactions in a plethora of species. The text mining

approach of the authors enabled them to also collect data for disease targets, organs, cell lines and pathways.

1.5.7. StarBase

StarBase (Yang et al., 2011) was first released in 2011. It is a platform focused on the analysis of high-throughput CLIP-Seq (HITS-CLIP and PAR-CLIP) and degradome sequencing (Degradome-Seq and PARE) data. The latest version (Li, Liu, Zhou, Qu, & Yang, 2014) of the database has been designed for decoding pan-cancer and interaction networks of lncRNAs, miRNAs, ceRNAs, RNA-binding proteins and mRNAs from large-scale CLIP-Seq (HITS-CLIP, PAR-CLIP, iCLIP, CLASH) data and tumor samples (14 cancer types and more than 6,000 samples). StarBase was also developed for deciphering protein-RNA and miRNA-target interactions, such as protein:lncRNA, protein:sncRNA, protein:mRNA, protein:pseudogene, miRNA:lncRNA, miRNA:mRNA, miRNA:circRNA, miRNA:pseudogene, miRNA:sncRNA interactions and ceRNA networks from 108 CLIP-Seq datasets derived from 37 studies. StarBase provides miRFunction and ceRNAFunction web tools to predict the function of ncRNAs (miRNAs, lncRNAs, pseudogenes) and protein-coding genes from the miRNA-mediated (ceRNA) regulatory networks.

1.5.8. DIANA-TarBase

The fifth version of DIANA Lab's TarBase (Papadopoulos, Reczko, Simossis, Sethupathy, & Hatzigeorgiou, 2009) was released in 2009 and included 1,300 experimentally supported targets from eight species that were manually curated from relevant literature. The transition from TarBase v5 to TarBase v6 (Vergoulis et al., 2012) included a 50-fold target increase (65,814 miRNA:gene interactions), coupled with a significant extension of specific research-oriented features. TarBase v6 accommodated a significant number of outcomes procured from state-of-the-art high-throughput studies. Importantly, the database hosted data derived from 3 CLIP-Seq and 12 Degradome-Seq studies, which at that time, was a 87.5%-fold increase compared to the eight studies supporting StarBase.

DIANA-TarBase v7 (Vlachos, Paraskevopoulou, et al., 2015) provides for the first time hundreds of thousands of high-quality manually curated experimentally validated miRNA:gene interactions, enhanced with detailed meta-data. The database enables users to easily identify positive or negative experimental results, the utilized experimental methodology, experimental conditions including cell/tissue type and treatment. The new interface also provides advanced information ranging from the binding site location, as identified experimentally as well as *in silico*, to the primer sequences used for cloning experiments. More than half a million miRNA:gene interactions have been curated from published experiments on 356 different cell types in 24 species, corresponding to 9- to 250-fold more entries than any other relevant database.

1.6. Integration of microRNAs in biological pathways

The characterization of miRNA function still remains an open challenge. In silico miRNA target prediction algorithms have been proven invaluable tools for the elucidation of miRNA function. Currently available state-of-the-art implementations can identify miRNA:gene interactions in 3′UTR as well as CDS regions, using complex physical models and/or machine learning approaches (Garcia et al., 2011; Reczko et al., 2012). However, even the most advanced methods still require experimental validation, since they exhibit a high number of false positive results. To this end, numerous low yield and high throughput wet lab techniques have been developed, that can be used to validate, explore and/or complement predicted results (Vlachos, Paraskevopoulou, et al., 2015).

These approaches have revealed the complex functional roles of miRNAs. Each miRNA can control up to dozens of genes, while multiple miRNAs have been also shown to collaborate in targeting extensive cellular processes and molecular pathways. The high number of miRNAs, in example miRBase v21 includes more than 2,500 human miRNAs, poses a significant bottleneck to the elucidation of their functional impact. Multiple targets have to be taken into account, which can be present in numerous pathways. The complexity of the problem increases when assessing the combinatorial effect of multiple miRNAs.

A series of functional analysis web servers and packages have been developed, in order to assist in the assessment of the functional impact of miRNAs on biological processes and pathways. Some of the most commonly used applications are presented in the following sections. StarBase has been excluded from the list since it has been extensively described in previous sections.

1.6.1. CORNA

CORNA (X. Wu & Watson, 2009) is a software package developed in R Statistical Language that allows scientists to analyze lists of genes that are targeted by miRNAs. The software is able to utilize existing methods such as hypergeometric, Fisher's exact and chi-square tests in order to identify significant miRNA:gene relationships in gene lists, and to test for significant associations between miRNAs and pathways or GO terms. CORNA includes plotting functions for visualizing quantitative data associated with miRNA targets.

1.6.2. miRTar

miRTar (J. B. Hsu et al., 2011) is a web based application which adopts various analyzing scenarios to identify putative miRNA:gene interactions in order to elucidate the biological functions of miRNAs and their implication in biological pathways. The

algorithm utilizes already established computational target prediction methods in order to consider various analyzing scenarios (1 miRNA:1 gene, 1:N, N:1, N:M, all miRNAs:N genes, and N miRNAs: genes involved in a pathway) to easily identify the regulatory relationships between important miRNAs and their targets, in 3'UTR, 5'UTR and coding regions. Subsequently, miRTar analyzes and highlights groups of miRNA-regulated genes that participate in particular KEGG pathways in order to elucidate the biological roles of miRNAs in biological pathways. The web server can also provide further information for elucidating miRNA regulation, such as the effect of alternative splicing.

1.6.3. miTalos

miTALOS (Kowarsch, Preusse, Marr, & Theis, 2011) is an interactive tool that integrates tissue and pathway filters to restrict the functional analysis. MiTALOS performs an enrichment and proximity analysis of predicted target genes in signaling pathways to infer miRNA-pathway associations. As the enrichment analysis focuses on the whole signaling pathway as a set of genes without taking its topology into account, subcascade-specific relations between miRNAs and pathways are ignored. In order to also account for such interactions, miTALOS performs simultaneous analysis of multiple miRNAs or even predefined genomic miRNA clusters. In addition, target genes and miRNAs are linked to external databases to offer additional information. Finally, graphical visualization of the miRNA targets in a given pathway allows functional insights into miRNA-dependent regulation of signaling pathways.

1.6.4. DIANA-miRPath

During the course of my Doctoral studies two versions of DIANA-miRPath (Vlachos et al., 2012; Vlachos, Zagganas, et al., 2015) have been developed in order to accomplish the integration of miRNAs in biological pathways. DIANA-miRPath is an online software suite dedicated to the assessment of miRNA regulatory roles and the identification of controlled pathways. The latest version of the web server renders possible the functional annotation of one or more miRNAs using standard (hypergeometric distributions), unbiased empirical distributions and/or meta-analysis statistics. DIANA-miRPath v3 database and functionality have been significantly extended to support all analyses for KEGG molecular pathways (Kanehisa et al., 2014), as well as multiple slices of Gene Ontology (Ashburner et al., 2000) in seven species including *H. sapiens*, *M. musculus*, *R. norvegicus*, *D. melanogaster*, *C. elegans*, *G. gallus* and *D. rerio*). Importantly, more than 600,000 experimentally supported miRNA targets from DIANA-TarBase v7 have been incorporated into the new schema. Users of DIANA-miRPath v3 can harness this wealth of information and substitute or combine the available in silico predicted targets from DIANA-microT-CDS and/or TargetScan v6.2 with high quality experimentally supported interactions. A unique feature of DIANA-miRPath v3 is its redesigned Reverse Search module, which enables users to identify and visualize miRNAs significantly

controlling selected pathways or belonging to specific GO categories based on in silico or experimental data.

1.7. Long non-coding RNAs as a novel layer of gene expression regulation

During the last decade, high-throughput sequencing technologies have emerged enabling the detection of novel coding and non-coding transcripts with unprecedented accuracy and sensitivity. The most convenient way to categorize the vast number of reported lncRNAs is to classify them according to genomic context and especially protein-coding genes. In this manner, lncRNAs can be grouped in several broad but mutually nonexclusive categories which do not correlate with their function or evolutionary origin.

1.7.1. Intergenic long non-coding RNAs

Intergenic lncRNAs (lincRNAs) are transcribed by individual genomic regions that do not overlap coding loci (Cabili et al., 2011; Guttman et al., 2009). Evidence of their existence initially occurred by analyzing signatures of active transcription such as H3K4me3 and H3K36me3 for promoters and gene bodies respectively. Many of the characterized lincRNAs are transcribed by Pol2, they undergo processing from the splicing machinery and are polyadenylated. They exhibit alternative isoforms, however, less frequently than protein-coding genes due to their limited number of exons. Typically they have an average length of 1kb, though there have been recent reports of lincRNAs whose length exceeds 300 kb (Georgakilas et al., 2014). The most notable ones include XIST (Brockdorff et al., 1992), H19 (Brannan, Dees, Ingram, & Tilghman, 1990), HOTAIR (Rinn et al., 2007) and MALAT1 (Ji et al., 2003).

1.7.2. Antisense long non-coding RNAs

Recently, there is increased evidence that abundant transcription occurs opposite the sense strand of annotated transcription units. More than 70% of sense transcripts have reported antisense counterparts (Faghihi & Wahlestedt, 2009; Y. He, Vogelstein, Velculescu, Papadopoulos, & Kinzler, 2008; Katayama et al., 2005). The overlap between sense-antisense pairs might be complete, however, natural antisense transcripts are most frequently enriched around the 5′ or 3′ end of the sense transcript. Many imprinted regions contain coding/noncoding sense-antisense pairs, such as *KCNQ1/KCNQLOT1* (Kanduri, Thakur, & Pandey, 2006) and *IGF2R/AIR* (Lyle et al., 2000).

1.7.3. Sense long non-coding RNAs

Small ncRNAs such as snoRNAs and miRNAs have long been known to reside inside introns of protein coding genes. Recently, large-scale transcriptomic and computational

analyses have reported numerous lncRNAs encoded within introns of annotated genes (Louro, Smirnova, & Verjovski-Almeida, 2009). The majority of these transcripts, also referred to as sense lncRNAs, exhibit differential expression patterns, stimuli response, deregulated expression in cancer, but only a few have been studied in detail (Guil et al., 2012).

1.7.4. Divergent, promoter- and enhancer-associated long non-coding RNAs

Short transcripts ranging from 20 bp to 2.5 kb have been recently found to be abundantly produced from the vicinity of active transcription start sites in both sense and antisense directions, corresponding to pausing-derived Pol2 peaks (Buratowski, 2008; Core, Waterfall, & Lis, 2008; Y. He et al., 2008; Preker et al., 2008; Seila et al., 2008). The shortest of these, often referred to as transcription start site-associated RNAs or divergent RNAs, might either be degradation products, processed from longer upstream antisense RNAs or promoter upstream transcripts. In most cases, these heterogeneous transcripts are capped and polyadenylated and exhibit low abundance and rapid degradation by exosomes. Until recently, it was not clear whether these transcription start site associated RNAs are transcriptional by-products from nucleosome-free regions surrounding promoters or their transcription assists in maintaining the chromatin open. Recently, a subset called promoter-associated short RNAs was found to interact with epigenetic factors such as Polycomb proteins (Kanhere et al., 2010) while divergent RNAs were found to host abundantly expressed miRNA precursors (Georgakilas et al., 2014). Another class of genomic regulatory elements, the enhancers, was also found to produce bidirectional transcripts up to 2 kb, which tend to remain unprocessed and to lack a known biological function (T. K. Kim et al., 2010).

1.7.5. Pseudogenes

Pseudogenes are considered remnants of genes that have lost their coding potential due to nonsense, frameshift, and other mutations (Pink et al., 2011). Many pseudogenes are products of tandem gene duplication or of mRNAs being carried along during retrotransposition, both of which create extra gene copies that are no longer under selective pressure. Most pseudogenes exhibit neutral conservation rates and no expression. However, recent estimations suggest that 2-20% of pseudogenes are transcribed while some of them present high levels of sequence conservation. Only a few rare examples have been shown to translate proteins. Certain pseudogenes were found to often regulate gene expression and especially their ancestral coding genes by post-transcriptional and epigenetic mechanisms. In example, there is a hypothesis that *XIST* evolved by the pseudogenization of the protein-coding gene Lnx3 and the integration of various transposon-derived repeat elements (Duret, Chureau, Samain, Weissenbach, & Avner, 2006).

1.7.6. Function of long non-coding RNAs

Although the mechanistic details for the function of the vast majority of annotated lncRNAs have not yet been unveiled, there are few cases that have shed light on how lncRNAs carry out their biological roles. There is increasing evidence that lncRNAs play a major role in epigenetic mechanisms by acting as recruiters, tethers and scaffolds. Large-scale studies of RNA-protein interactions have shown that chromatin-modifying complexes, such as PRC2, interact with a large number of lncRNAs (Guil et al., 2012; Kanhere et al., 2010; Khalil et al., 2009). The mechanisms that mediate the recruitment of Polycomb complexes to specific genomic loci have not yet been unveiled in mammals, especially when consensus binding sequences are absent. However, recent observations regarding the interaction between lncRNAs and Polycomb proteins suggest that Polycomb recruitment may be ncRNA-mediated. *HOTAIR* is a member of *HOXC* cluster and was found to repress the transcription of *HOXD in-trans* by interacting with PRC2 (Rinn et al., 2007). H3K9 methyltransferase G9a is another epigenetic complex which interacts with the imprinted lncRNA Air (Nagano et al., 2008). In some cases, lncRNAs act as scaffolds enabling the assembly of numerous protein complexes and facilitating the coordination of multiple layers of chromatin modifications. In example, *KCNQ1OT1* is hypothesized to recruit both PRC2 and G9a to the promoter of *KCNQ1* (Pandey et al., 2008). LncRNAs can also act by recruiting factors associated with gene activation. *MISTRAL* and *HOTTIP* are two lncRNAs that belong to the *HOXA* cluster and their role is to recruit the MLL complex *in-cis* (Bertani, Sauer, Bolotin, & Sauer, 2011). In addition, lncRNAs are able to influence epigenetic regulation by modulating DNA methylation at CpG dinucleotides, which is a crucial step in the stability of genes' repression process (Law & Jacobsen, 2010).

Numerous studies have unveiled the role of lncRNAs in transcription by acting as decoys, co-regulators and Pol2 inhibitors. *PANDA* lncRNA acts as decoy in order to remove NF-YA away from its pro-apoptotic target genes (Hung et al., 2011). In a similar fashion, other lncRNAs compete for TF binding. In example, *GAS5* recognizes the DNA-binding domain of nuclear glucocorticoid receptors and inhibits their contact with glucocorticoid response elements on the genome (Kino, Hurt, Ichijo, Nader, & Chrousos, 2010). LncRNAs can also directly interfere with Pol2 activity. One case is the inhibition of the transcription of dihydrofolate reductase's major transcript (Schnell, Dyson, & Wright, 2004). The minor promoter of *DHFR* transcribes a lncRNA molecule which inhibits the assembly of the transcription pre-initiation complex at the gene's major promoter. This mechanism most likely functions through the direct binding of the general transcription factor TFIIB on the promoter-derived lncRNA. There is also a possibility that DNA:RNA triplex formation occurs at the major promoter (Martianov, Ramadass, Serra Barros, Chow, & Akoulitchev, 2007).

The most prominent role of lncRNAs is their implication in RNAi-mediated gene expression regulation. In example, the antisense transcript of Alzheimer's-associated *BACE*, known as *BACE-AS*, increases the stability of *BACE* mRNA, by acting as "sponge" for miR-485-5p (Faghihi et al., 2010). Many mammalian pseudogenes such as *PTENP1* and *KRASP1* (Poliseno et al., 2010), as well as other lncRNAs (Cesana et al., 2011) harbor miRNA-binding sites in their 3'UTRs and might therefore function as sponges to remove miRNAs away from their intended targets. This is a phenomenon that was initially discovered in plants (Franco-Zorrilla et al., 2007) and was hypothesized to be part of a genome-wide regulatory network comprising miRNA pseudo-targets (Seitz, 2009) called competing endogenous RNAs (ceRNAs) (Salmena, Poliseno, Tay, Kats, & Pandolfi, 2011). The conceptual idea behind ceRNAs is that if the expression level of one member of this network changes I would affect the amount of miRNAs binding on it. This would affect the overall accessible pool of miRNAs shared with other members of the network, leading to subsequent changes in the transcript levels of the network. In example, *LINC-MD1*, whose expression is developmentally-dependent, has been reported to act as a sponge and influence the mRNA levels of miRNA-targeted muscle differentiation genes (Cesana et al., 2011). Recent studies have revealed that miRNA:lncRNA interactions are wide-spread and common in both human and mouse species (Paraskevopoulou, Georgakilas, Kostoulas, Reczko, et al., 2013). The task now is to determine whether this actually represents a new layer of post-transcriptional regulation directed by precise and signal-responsive changes in a ceRNA's expression level or if this is simply an inevitable consequence of several mRNAs and ncRNAs being regulated by the same pool of miRNAs.

1.8. microRNA targets on long non-coding RNAs

The recent shift of the research community's attention towards lncRNAs resulted in the continuously growing number of experiments in order to study their physiological/pathological implications. Consequently, and due to the rapid rate in the annotation of transcriptional units, databases indexing lncRNA properties and function gradually become essential tools to this process. The interplay between miRNAs and lncRNAs has been extensively discussed in previous section. The proposal of ceRNA networks has introduced a novel layer of gene expression regulation. The identification of the underlying links between lncRNA and miRNA families will provide new insights in molecular biology.

During the past few years, several databases emerged in order to compensate for the need of functional characterization of lncRNAs. These databases focus on the annotation of miRNA:lncRNA interactions either by utilizing target prediction algorithms or by analyzing experimentally derived AGO binding sites on the transcriptome.

1.8.1. lnCeDB

lnCeDB (Das, Ghosal, Sen, & Chakrabarti, 2014) provides a catalogue of human lncRNAs that can potentially interfere in ceRNA networks. The authors have utilized AGO binding sites that are available in StarBase in order to identify experimentally derived MREs. The association of regions enriched in AGO binding with the mature miRNA sequences has been accomplished with TargetScan. Additional miRNA:lncRNA interactions were derived from miRCode database. An in-house developed algorithm based on seed-matching has been utilized in order to identify miRNA binding sites on the remaining lncRNAs.

In order to establish the probability of an lncRNA:mRNA pair to function in a ceRNA context two approaches have been utilized. The ceRNA score is calculated from the ratio of the number of shared MREs between the pair with the total number of MREs of the individual candidate gene. Alternatively, the p-value for each ceRNA pair is determined with the hypergeometric test by utilizing the number of shared miRNAs between the ceRNA pair against the number of miRNAs interacting with the individual RNAs.

1.8.2. LncRNADisease

In the last few years, there is increased evidence showing that lncRNAs are associated with a variety of biological processes. LncRNADisease (G. Chen et al., 2013) is a repository that attempts to associate lncRNAs and disease by utilizing computation and experimental methodologies. The database aims to uveil the role of lncRNAs in diseases and to identify candidate molecules for disease diagnosis, treatment and prognosis. LncRNADisease includes 480 entries of experimentally supported lncRNA-disease associations corresponding to 166 diseases. The database also hosts 478 interactions between lncRNAs and proteins, RNAs, miRNAs and DNA.

1.8.3. LncRNABase

LncRNABase is a module of StarBase v2 (Li et al., 2014) that hosts the interactions between various types of ncRNAs. These interactions have been derived by systematic analysis of 108 CLIP-Seq datasets generated by 37 independent studies. In order to identify genome-wide interactions between miRNA and lncRNAs, the authors utilized conserved MREs predicted by TargetScan, miRanda, Pictar, PITA and RNA22 algorithms which were subsequently intersected with the AGO CLIP clusters resulting in CLIP supported sites.

1.8.4. miRCode

In a recent study, Jeggari *et al.* (Jeggari, Marks, & Larsson, 2012) have identified putative miRNA-binding sites across all annotated human transcripts of GENCODE v11 release, which included 10,419 lncRNAs. The authors provided access to these in silico predicted

miRNA sites through a web interface named "miRCode". miRCode supports seed-related information; genomic location, binding type, percentage of evolutionary conservation across primates/non-primate mammals/non-mammalian vertebrates as well as possible overlaps with repeat sequences. The implemented prediction pipeline has been based on a seed complementarity algorithm and on TargetScan v6 miRNA seed family nomenclature. However, miRCode includes limited MRE-related information and only on a small fraction of the publicly available annotated human lncRNAs. miRNA target predictions for other species as well as experimentally verified binding sites on lncRNAs are not supported.

1.8.5. DIANA-LncBase

DIANA-LncBase (Paraskevopoulou, Georgakilas, Kostoulas, Reczko, et al., 2013) repository hosts transcriptome-wide experimentally verified and computationally predicted MREs on human and mouse lncRNAs. The database can be accessed by an intuitive and user-friendly web interface which provides two distinct modules. The experimental module hosts validated miRNA:lncRNA interactions while the computational module provides access to a plethora of information related to the *in silico* predicted pairs.

DIANA-LncBase provides a comprehensive collection of more than 5,000 miRNA:lncRNA interactions interactions supported by experimental data for both human and mouse species. miRNA targets of large collections of mouse lncRNA transcripts have not yet been extensively studied and there are only few miRNA:lncRNA interactions reported in the available literature. The analysis performed includes all available lncRNA data resources in human and mouse and the identification of experimentally verified miRNA targets with the use of high-throughput PAR-CLIP (Hafner et al., 2010) and HITS-CLIP (Chi et al., 2009) experiments.

The in silico analysis performed includes an integration of most of the available lncRNA resources and state-of-the-art computational target predictions (Reczko et al., 2012) which resulted in more than 10 million miRNA:lncRNA interactions between 56,097 lncRNAs and 3,078 miRNAs in human and mouse. DIANA-LncBase (predicted module) hosts detailed information for each miRNA:lncRNA pair, such as external links, graphic plots of transcripts' genomic location, representation of the binding sites, lncRNA tissue expression as well as MREs conservation and prediction scores.

1.9. Next Generation Sequencing

The first generation of sequencing methodologies was introduced during the 1970s when Sanger *et al.* (Sanger, Nicklen, & Coulson, 1977) and Maxam and Gilbert (Maxam & Gilbert, 1977) developed termination and fragmentation methods respectively, in order

to sequence DNA. This was a decisive moment in research since these sequencing techniques transformed Biology by providing the tools to decipher complete genes and later on entire genomes. Sanger sequencing became the dominant DNA sequencing method for the next 30 years, finally enabling the completion of the first human genome sequence in 2004 (International Human Genome Sequencing, 2004).

Table 2. Maximum throughput of well-established NGS platforms in Giga-bases. This table has been designed for the purposes of this dissertation.

Throughput of NGS solutions		
	Gigabytes per run	
Solution	**Initial version**	**Latest version**
Ion Torrent	0.3	10
PacBio	0.012	0.5
SOLiD	3	320
454	0.02	0.7
Illumina	1	1800

The Human Genome Project required enormous amounts of resources and especially time, highlighting the need for the development of higher throughput, faster and significantly cheaper technologies. In the same year (2004) the National Human Genome Research Institute (NHGRI) initiated a funding program aiming to reduce the cost of human genome sequencing to less than 1,000 USD in the following decade. This stimulated the development and commercialization of next-generation sequencing technologies which were introduced in 2005. The enormous numbers of reads (**Table 2**) generated by NGS enabled the sequencing of entire genomes at unprecedented running time (**Table 3**).

Table 3. Time needed to complete the sequencing of a bacterial genome on different NGS platforms. This table has been designed for the purposes of this dissertation.

Time required for sequencing bacterial genome	
	Hours
Solution	
Ion Torrent	3
PacBio	3
SOLiD	336
454	23
Illumina	240

Numerous sequencing platforms have emerged since 2005, however, only five are considered well-established in the NGS market; 454, Illumina, SOLiD, Ion Torrent and PacBio. Each platform has its own unique characteristics and aims at different target groups. In example, for quantitative studies, Illumina and SOLiD platforms were more suitable than 454 and PacBio due to their higher throughput. On the other hand, 454 and PacBio platforms are preferred for genome assembly studies, since the length of the produced reads is significantly higher when compared to the competition (**Table 4**).

Table 4. Maximum read length produced by the most prominent NGS platforms. This table has been designed for the purposes of this dissertation.

Read length by NGS solutions		
	Number of nucleotides per read	
Solution	**Initial version**	**Latest version**
Ion Torrent	200	400
PacBio	4,000	20,000
SOLiD	35	75

454	110	1,000
Illumina	35	300

As the sequencing technologies evolved, an increasing number of sample preparations methods and a plethora of different protocols have emerged enabling researchers to study biological systems at unprecedented speed and resolution. In example, advances in throughput and cost reduction allowed genomic DNA sequencing to be applied on a population scale. This included the first large-scale human genetic variation study, named 1000 Genomes Project (Genomes Project et al., 2010), which was followed by even larger projects involving the sequencing of thousands of genomes (Genome, 2009). Such projects are able to revolutionize our understanding of the relationship between genomic variation and phenotype. The ever evolving field of sequencing techniques has provided a breakthrough in the knowledge regarding the transcriptome landscape of eukaryotes. The first strand specific RNA-Sequencing protocols emerged in 2009 (Z. Wang, Gerstein, & Snyder, 2009) and enabled the identification of novel antisense regulatory transcripts that exhibit important biological functions. Nowadays, transcriptome analysis can be performed at single cell level providing a detailed view of transcription dynamics. Such studies have revealed that there can be substantial transcriptional heterogeneity among seemingly identical cells (Shalek et al., 2014). NGS methodologies have also been developed to study protein-DNA interactions in genome-wide scale with chromatin immunoprecipitation sequencing, known as ChIP-Seq (Johnson, Mortazavi, Myers, & Wold, 2007). This protocol has been extended ever since in order to facilitate studies for the identification of protein-RNA interactions; CLIP-Seq (Sanford et al., 2009), iCLIP (Konig et al., 2010), PAR-CLIP (Hafner et al., 2010) and HITS-CLIP (Chi et al., 2009); RNA-DNA interactions; CHART (Simon et al., 2011) and CHiRP (Chu, Qu, Zhong, Artandi, & Chang, 2011); DNA-DNA interactions; ChIA-PET(Dekker, Marti-Renom, & Mirny, 2013); and open chromatin domains; DNase-Seq (L. Song & Crawford, 2010).

The following sections will focus on NGS techniques that have been utilized in order to develop the algorithms and computational techniques presented in this dissertation.

1.9.1. RNA-Sequencing

One of the first and most important NGS methodology is RNA-Seq (Z. Wang et al., 2009). During the last decade RNA-Seq has been used in countless studies in order to measure gene expression, discover and annotate complete transcriptomes and characterize alternative splicing sites and polyadenylation. Depending on the experimental framework and the biological question that needs to be answered, there are many variations of the RNA-Seq protocol which focus on capture different types of RNA, such

as polyadenylated, non-polyadenylated, ribosomal and different classes of small RNAs. A simplified overview of the protocol is presented in **Figure 7**. Since the goal of RNA-Seq is to characterize the transcriptome the first step involves isolating and purifying cellular RNAs. This typically involves disrupting cells in the presence of detergents and chaotropic agents. Subsequently, RNA can be recovered from the total cell lysate and undergo the selection step which includes size selection in order to distinguish small from large RNA molecules, poly(A) selection in order to distinguish between adenylated and non-adenylated RNAs etc. The next step includes RNA fragmentation which depends on the NGS platform that will be utilized in the sequencing phase. The most common fragmentation methods are metal ion, heat, enzymatic-induced and sonication.

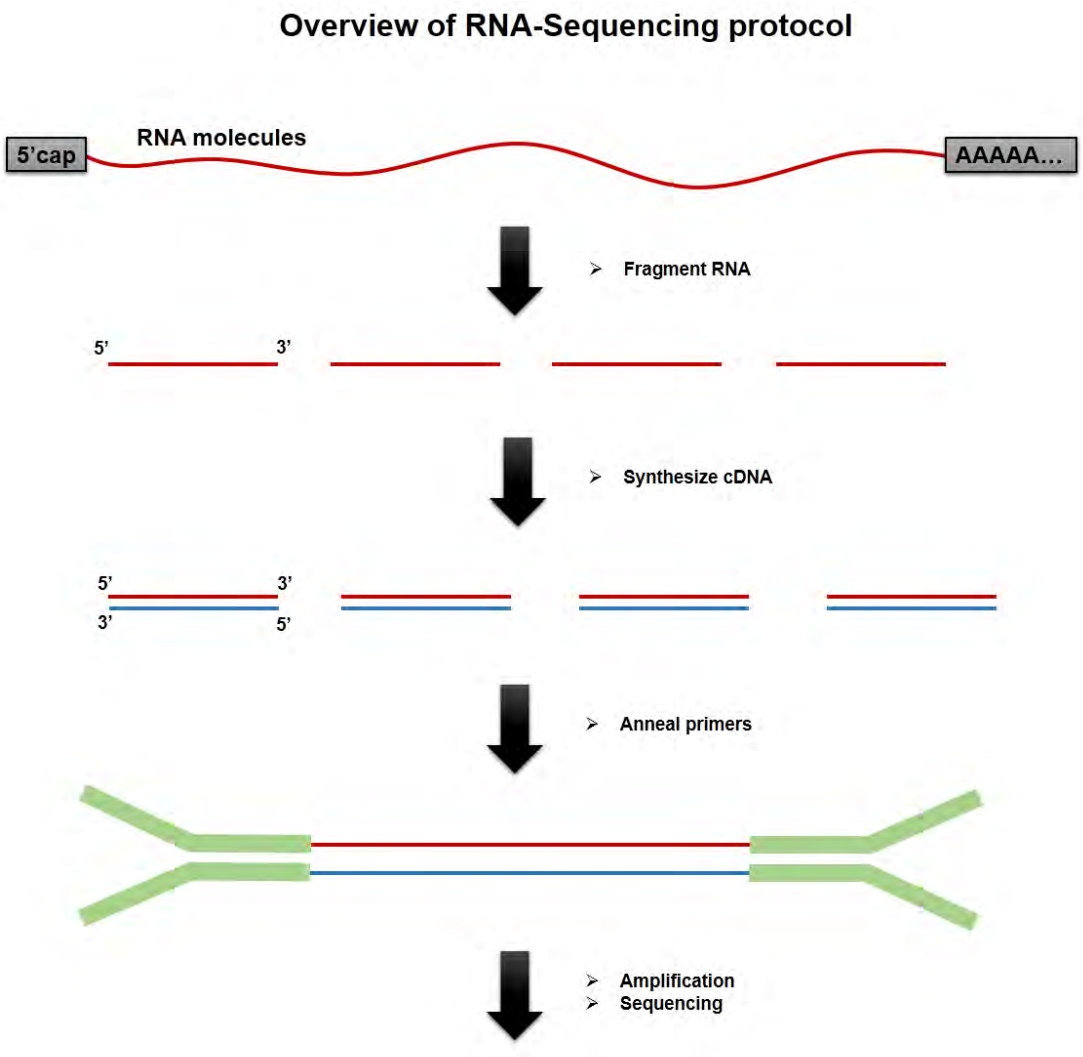


Figure 7. Simplified overview of RNA-Seq protocol. This figure has been designed for the purposes of this dissertation.

After fragmentation the RNA must be converted to double stranded complementary DNA and ligated with special sequences, commonly referred to as "adapters", that are platform-specific. Adapter sequences are a very important part of the process since they can be detected by the sequencing instrument and start the sequencing process. The library preparation is finished with the amplification step and the reads are forwarded to the sequencing platform.

1.9.2. ChIP-Sequencing

Since the completion of the sequencing of the human genome, a number of functional approaches have been taken to understand how the genome functions. Gene expression is a dynamic and complex process and is regulated by multi-protein transcriptional machinery including protein-DNA and protein-protein interactions. The proteins include TFs, histones, enhancers, suppressors and others. Each TF can regulate the expression of many genes binding near the transcription start site and can play important roles in defining the physiological state of a cell. ChIP-Seq (Johnson et al., 2007) has become a widely used method for determining the in vivo binding sites of a TF and locations of chromatin modifications.

A simplified overview of ChIP-Seq protocol is presented in **Figure 8**. The initial cell population is cross-linked and then homogenized in order to eliminate cytoplasmic proteins and significantly reduce the number of possible cross-reactive proteins. Nuclei subsequently undergo lysis and the chromatin is sheared by sonication. Antibodies raised against a particular TF or DNA-binding protein of interest are used to immunoprecipitate specific DNA-protein complexes.

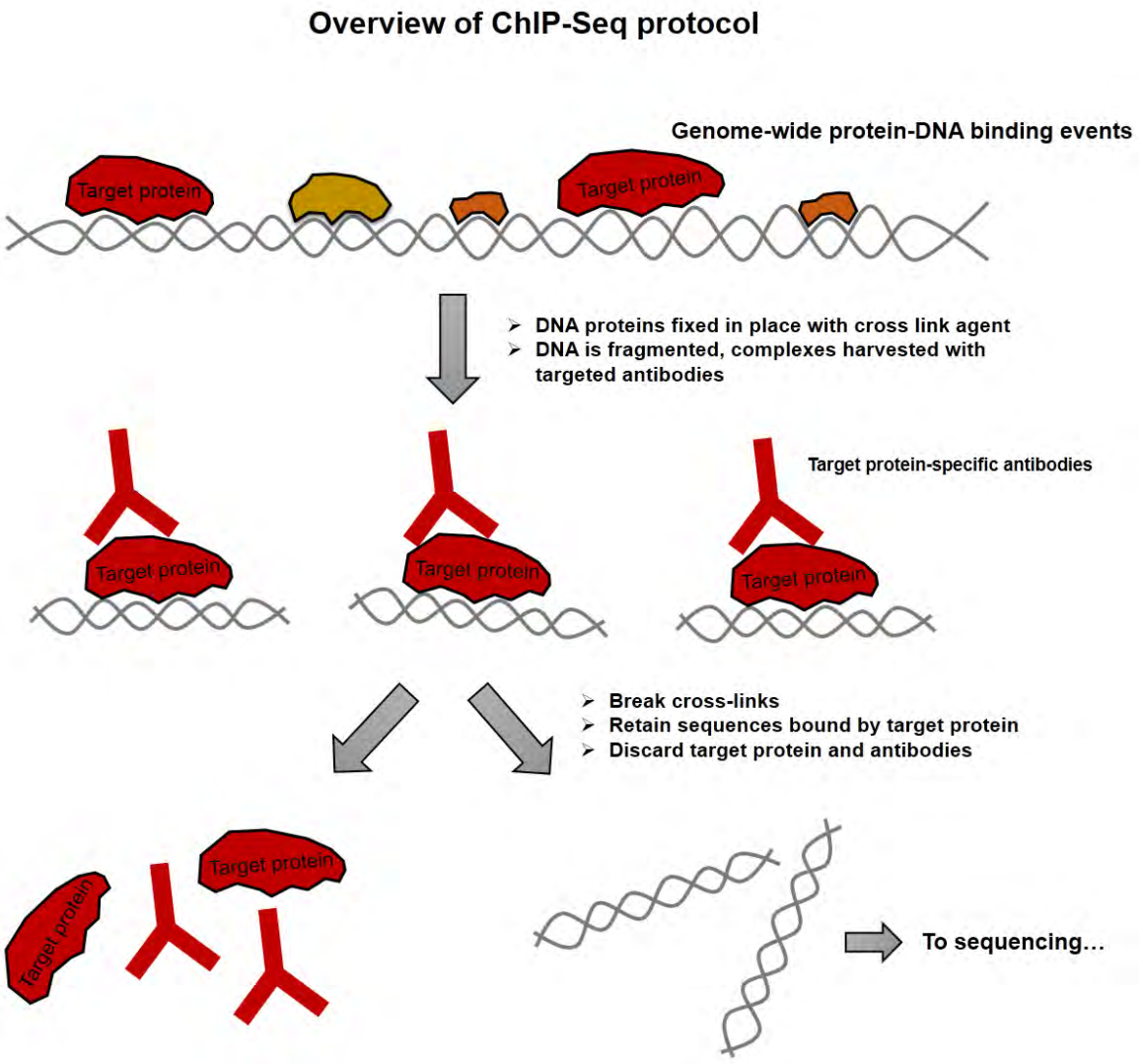


Figure 8. Simplified pipeline of ChIP-Seq protocol. This figure has been designed for the purposes of this dissertation.

ChIP DNA is then separated from proteins by reverse cross-linking, followed by RNase and proteinase K digestion. Purified ChIP DNA is then prepared for sequencing by ligating adapters and amplify reads with PCR for a limited number of cycles.

1.9.3. HITS/PAR-CLIP

Most RBPs recognize short, degenerate RNA motifs, and therefore they might often bind at several sites on most RNAs. Thus, it is important to define the full landscape of interactions between RBPs and various types of RNA. CLIP is a state-of-the-art technology that enables the identification of such landscapes and relies on the principle that precise and stringent mapping of binding sites is achieved by preserving the in vivo-protein-RNA interactions by irradiation of living cells or tissue with ultraviolet C light.

The UVC light induces the formation of covalent crosslinks only at sites of direct contact between proteins and RNA (**Fig. 9**). On cell lysis, the protein-RNA complex is immunoprecipitated with and antibody that is specific for the protein of interest. The co-purified RNA molecules are reverse-transcribed and amplified with the aid of 5′ and 3′ adaptors.

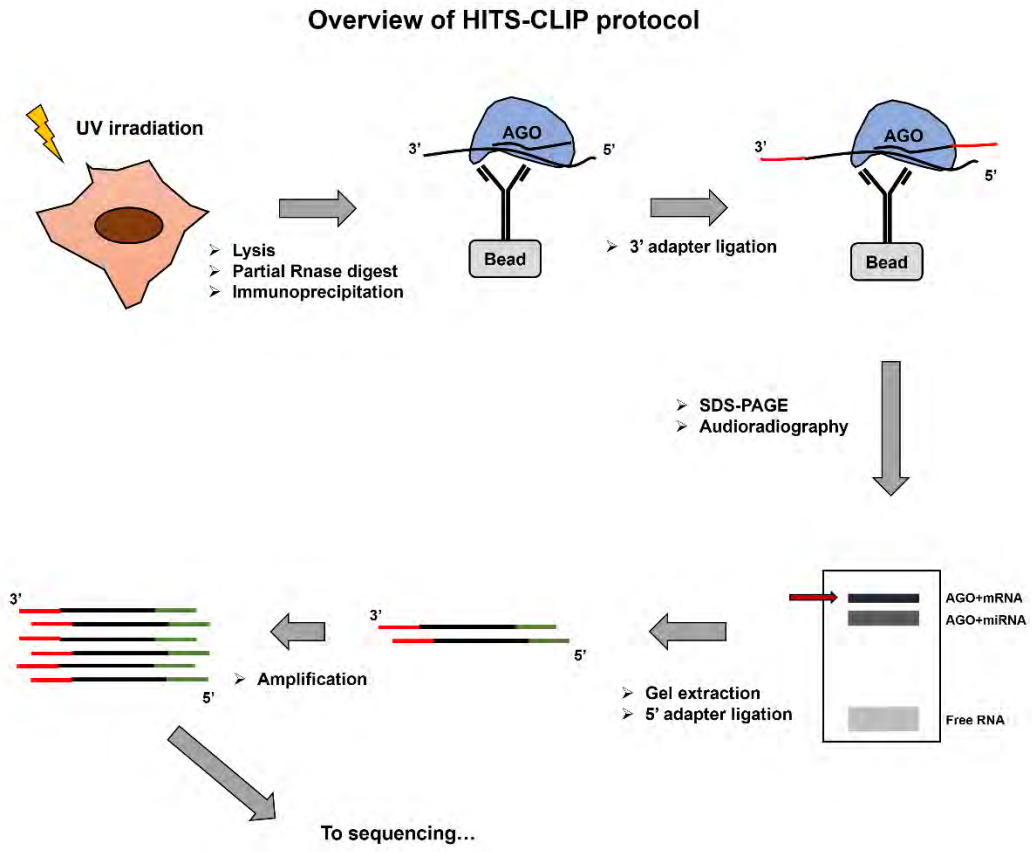


Figure 9. Concept of HITS-CLIP protocol. This figure has been designed for the purposes of this dissertation.

The reads that have been ligated with adapters are subjected to PCR amplification and then forwarded to the sequencing instrument.

In the traditional CLIP protocol (Chi et al., 2009), the resolution of binding site detection mostly corresponds to the length of the fragmented RNAs. Subsequent studies (Granneman, Kudla, Petfalski, & Tollervey, 2009) showed that crosslink-induced point mutations and deletions can be used to identify the crosslink sites of RBPs within snoRNAs. On the other hand, in the photoactivatable ribonucleoside-enhanced CLIP approach (Hafner et al., 2010), nucleotide analogues such as 4-thiouridine or 6-

thioguanosine are used which can be efficiently crosslinked with ultraviolet A light. The nucleotide analogues are readily taken up by cells and become incorporated into newly synthesized transcripts. Importantly, they lead to a base transition at the crosslink site during reverse transcription. Therefore, mutation analysis of the resulting cDNA sequences can be used to pinpoint crosslink sites at nucleotide resolution.

An alternative method for achieving nucleotide resolution is known as individual nucleotide resolution CLIP (Konig et al., 2010). This method is based on the concept that reverse transcription can stop at nucleotides that are cross-linked to the peptides that remain after proteinase K digestion. Sequencing of the truncated cDNAs provides direct identification of the cross-link position, which is located one nucleotide upstream of the truncation site.

1.9.4. DNase-Sequencing

Traditionally, open chromatin has been identified by the hypersensitivity of genomic sites to nuclease treatment with MNase and the non-specific double-strand endonuclease DNase I. The identification of hypersensitive sites has initially been based to Southern blotting and involved laborious and time-consuming steps that limit the applicability of the method to a narrow extent of the genome. The advent of NGS gave rise to DNase-Seq allowing the genome-wide identification of hypersensitive sites with unparalleled specificity, throughput and sensitivity in a single reaction. In recent times the drop of sequencing costs and the increased quality of the data have made DNase-Seq the "golden standard" for probing chromatin accessibility. During a typical (**Fig. 10**) DNase-Seq experiment (S. John et al., 2013) DNA from nuclei is digested with limiting DNase I concentrations.

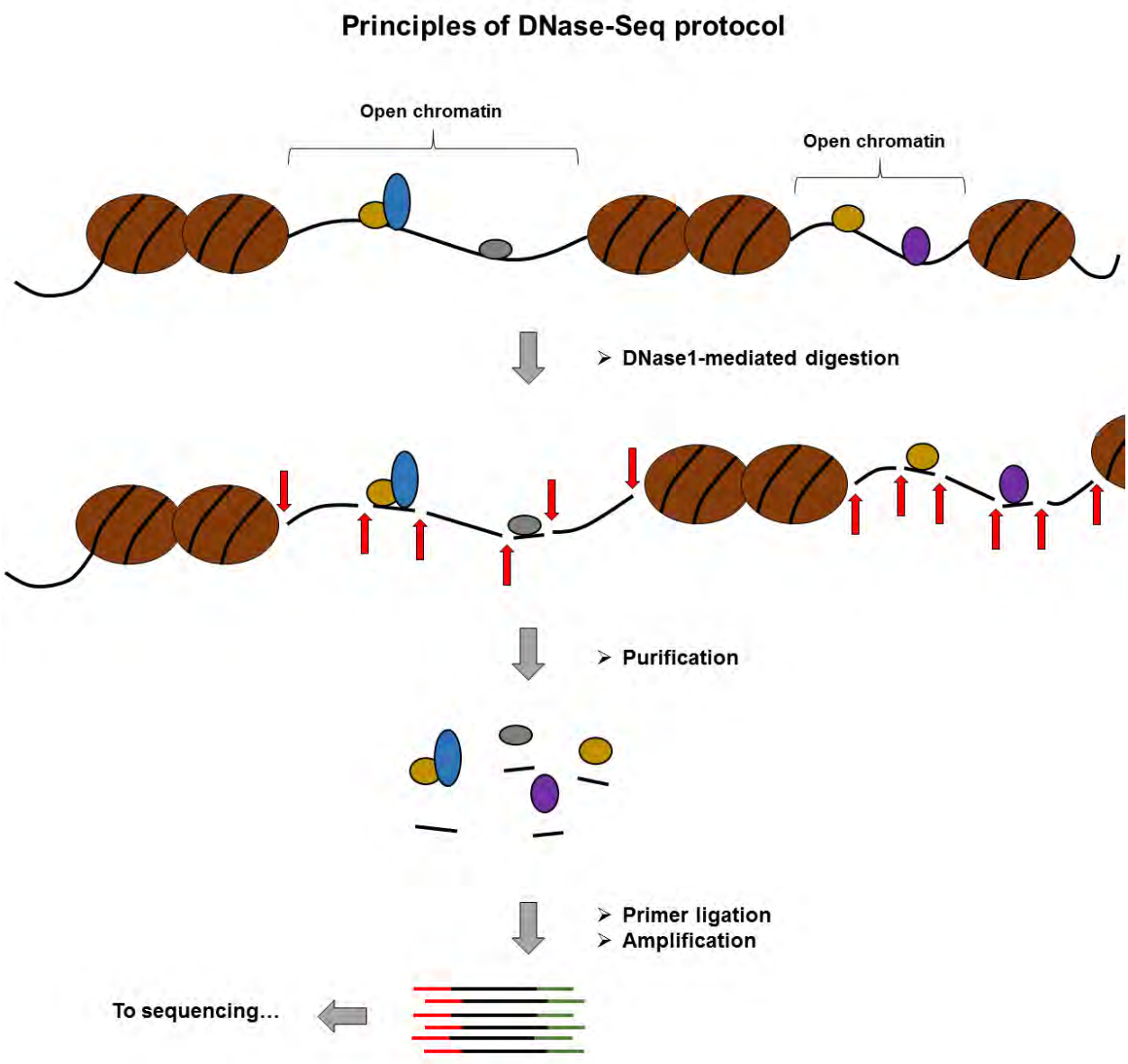


Figure 10. Simplistic overview of DNase-Seq protocol. This figure has been designed for the purposes of this dissertation.

Optimal digestions are purified with size selection of fragments smaller than 500 bp and are submitted to sequencing after library construction.

1.9.5. GRO-Sequencing

Global run-on-sequencing (GRO-Seq) technique (Core et al., 2008) has been developed in order to enable the mapping and quantification of transcriptionally engaged polymerase density in genome-wide scale. GRO-Seq provides a snapshot of genome-wide transcription and allow us to directly evaluate promoter-proximal pausing on all genes.

In order to specifically isolate Nuclear Run-On RNAs, a ribonucleotide analog 5-bromouridine 5′-triphosphate (BrUTP) is added to nascent RNA during the Run-On step (**Fig. 11**). The Nuclear Run-On RNAs are subsequently chemically hydrolyzed into short fragments (~100 bases) to facilitate high-resolution mapping of the polymerase origin at the time of assay.

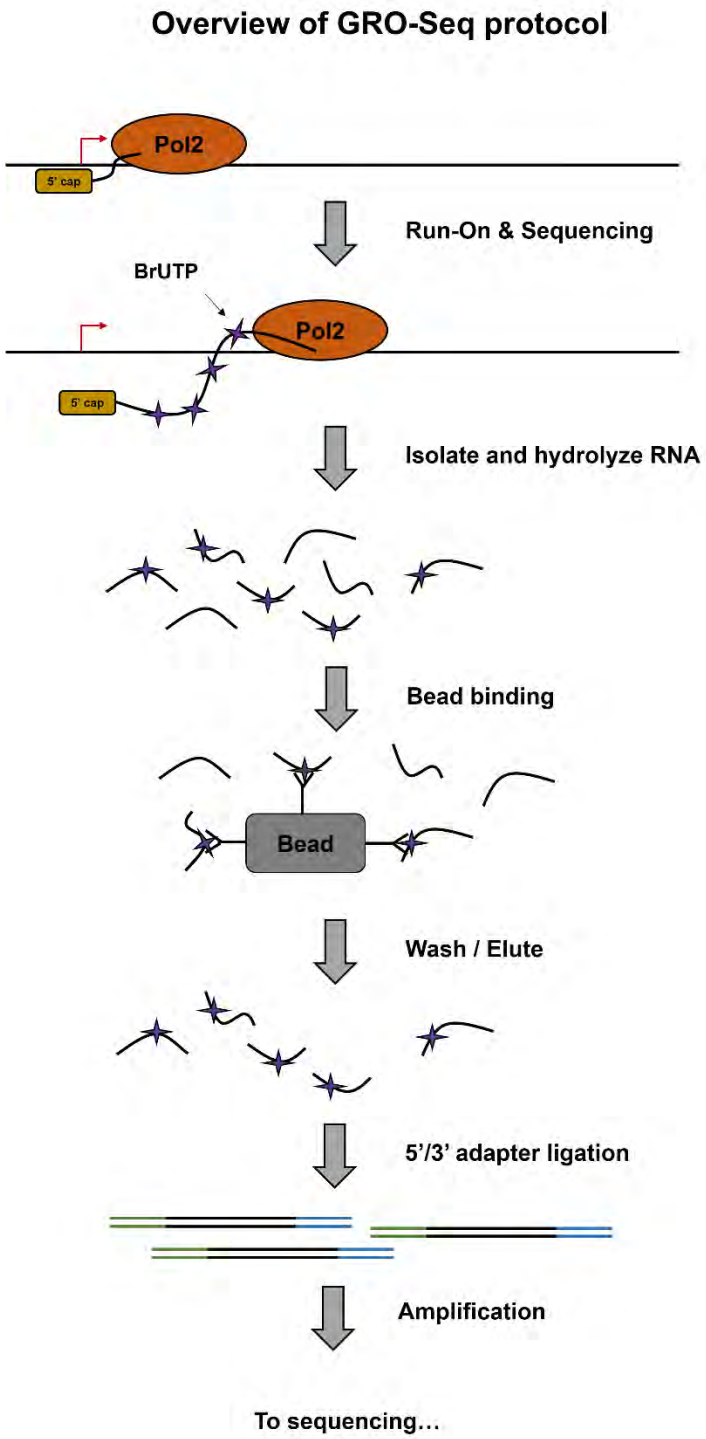


Figure 11. Overview of GRO-Seq protocol. This figure has been designed for the purposes of this dissertation.

BrU-containing Nuclear Run-On RNAs are immunopurified with an antibody that is specific for this nucleotide analog. Adapters for both 5′ and 3′ ends are ligated, the RNA

fragments are subjected to reverse transcription and amplified. The origin and the orientation of the RNAs and therefore the associated transcriptionally engaged polymerases can be documented genome-wide by mapping the reads to the reference genome.

1.10. Machine Learning

The scientific field of Machine Learning is associated with the development and evaluation of algorithms that are able to perform pattern recognition, classification and prediction by utilizing prior knowledge of existing data in order to improve the derived models (Mitchell, 1997).

Some of the most important discoveries in Biology (among other scientific fields) have been facilitated by utilizing Machine Learning methodologies. On the other hand, the Machine Learning field itself has been enriched in techniques explicitly developed to answer open questions in Biology. The most notable example is the perceptron (Rosenblatt, 1958) that was the first attempt to model neuronal behavior and resulted in the creation of artificial neural networks. The perceptron has also been utilized in order to recognize translation initiation sites in the genome of Escherichia coli (Stormo, Schneider, Gold, & Ehrenfeuch, 1982). Machine learning has also been useful for making sense out of large genomic data sets and facilitating the annotation of a wide variety of genomic sequence elements. In example, the recognition of transcription start sites in a genome sequences (Chien et al., 2011; Georgakilas et al., 2014; Marsico et al., 2013; Megraw, Pereira, Jensen, Ohler, & Hatzigeorgiou, 2009). In a similar fashion, Machine Learning models can also be trained to recognize splice sites (Degroeve, De Baets, Van de Peer, & Rouze, 2002), promoters (Corcoran et al., 2009; Marson et al., 2008; Ozsolak et al., 2008), enhancers (Heintzman et al., 2007) or nucleosome protected (Segal et al., 2006) genomic regions. In general, the combination of different models that each recognize specific types of genomic elements and their genomic context can facilitate the developments of Machine Learning models capable of annotating entire genes and other important regulatory regions of the genome. During the last decade, the flood of NGS data sets such as RNA-Seq, DNase-Seq and ChIP-Seq has allowed the integration of experimental techniques and Machine Learning methodologies resulting in the identification of new classes of genomic elements and the annotation of the genome in an unsupervised way (Ernst & Kellis, 2012; Guttman et al., 2009; Roadmap Epigenomics et al., 2015).

The field of Machine Learning resembles the one of Statistical Learning since the majority of the aforementioned questions can be answered by utilizing statistical methodologies. Even though the boundaries between these two fields are sometimes not clear, it is a fact that Machine Learning originates from the artificial intelligence community. Scientists of

Machine Learning field focus in the analysis of large heterogeneous data, while important statistical concepts do not exist in the field's literature. The flexibility of Machine Learning algorithms has grown hand-in-hand with frameworks for evaluating their reliability, and hopefully they will enable the discovery of hidden information in the ever increasing volume and complexity of biological data.

In the following sections, the concept of categorizing Machine Learning algorithms in Supervised, Un-supervised and Semi-supervised methods is presented depending on the utilized data as well as a brief introduction of the most notable techniques' functionality.

1.10.1. Supervised, Unsupervised and Semi-supervised Learning

Machine Learning algorithms can be categorized in two main families, named Supervised and Un-supervised Learning methodologies (**Fig. 12**). In Supervised Learning, the models are trained on already labelled instances and subsequently utilized to predict the labels of previously unknown data. With the structure of the train data already available, the goal of Supervised Learning is to accurately predict the structure of new instances based on the available features. On the other hand, Un-supervised Learning algorithms do not require prior knowledge of the under-study data properties. They are mainly utilized in order to discover the structure of unlabeled instances in datasets of uncharted scientific fields.

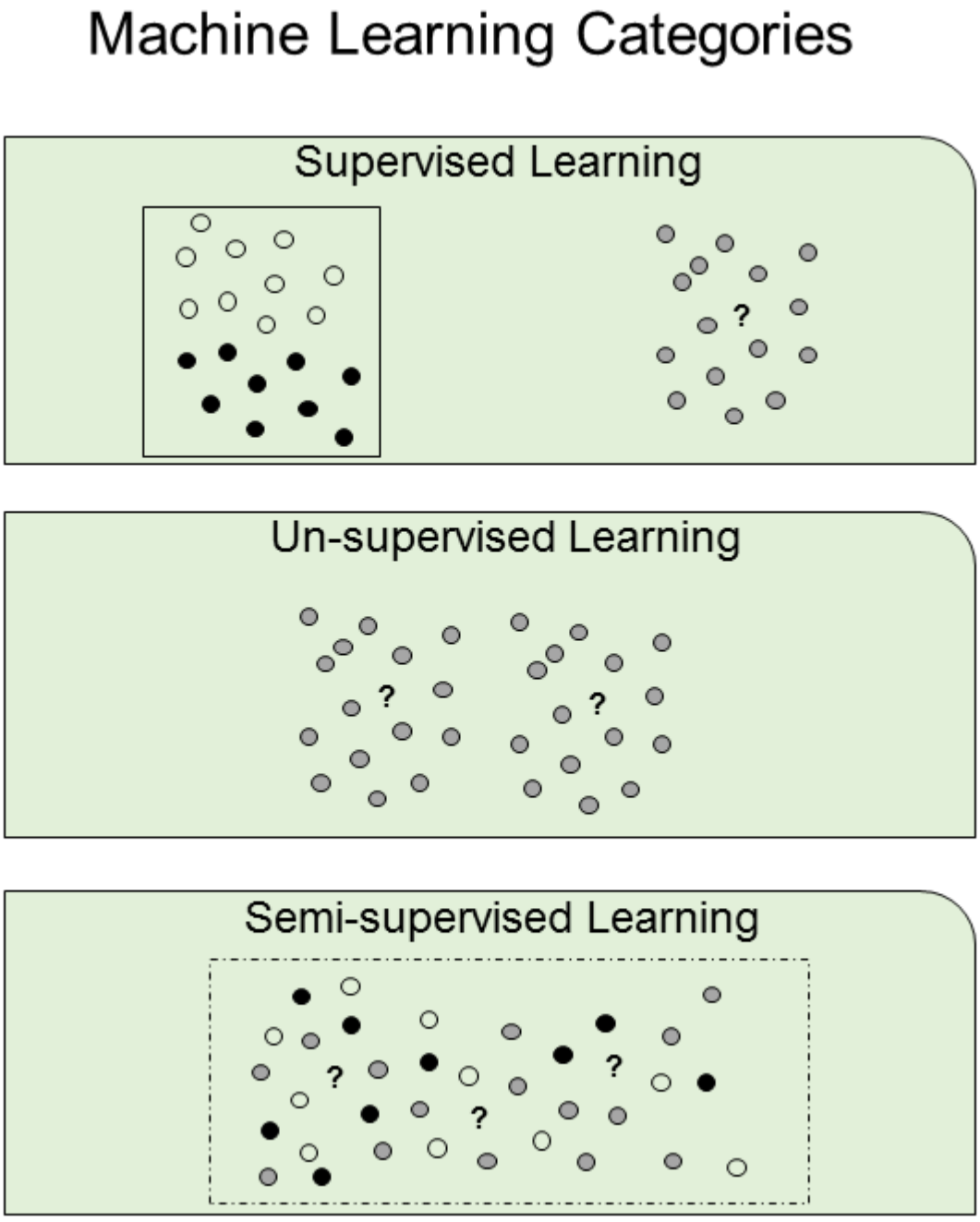


Figure 12. Visualization of training sets in different Machine Learning algorithmic categories. This figure has been designed for the purposes of this dissertation.

In example, algorithms that provide solutions in the problem of miRNA TSS identification (Georgakilas et al., 2014) utilize Supervised Machine Learning approaches. MicroTSS extracts features related to known TSSs of protein coding genes in order to train Support Vector Machine models and subsequently utilize them in order to predict miRNA TSSs in previously unseen genomic regions. On the other hand, labelled training

sets are not always available resulting in the need of Un-supervised Learning. The most recent example (Roadmap Epigenomics et al., 2015), considers the interpretation of a heterogeneous epigenomic data set collection in order to discover novel regulatory regions in the human genome.

There are cases where the properties of the under-study data cannot fit neither Supervised nor Un-supervised framework. In Semi-supervised Learning, which is a mixture of the two previously described approaches, the algorithm utilizes a collection of instances, but only a subset of them have associated labels. In example, gene-finding algorithms are typically trained in a Semi-supervised fashion, where the input is a set of annotated genes and a whole-genome sequence without labels. In this case, an initial model is trained on the basis of the labelled subset alone. The model is subsequently utilized to scan the sequence and assign labels throughout the genome, which can further be used to refine the model. The process is repeated until zero new gene discovery is reached.

The description of the enormous variety of Machine Learning algorithms far exceeds the purpose of the current dissertation. The following sections summarize the properties of the most notable and widely used core methodologies without delving into different versions of each technique.

1.10.2. Regression

Regression and General Linear Models could be characterized as Supervised methods and are tightly connected with the very foundations of mathematical thought, but their current form was made possible during the 18th century when the theory of algebraic invariants emerged. Direct result of this theory was the development of correlational methods and Linear Regression models during the 19th century. Both of these methods serve as the foundation of General Linear Models which enables the description of relationships among dependent and independent variables in a simplified mathematical equation. Regression models can enable the estimation of the dependent variable values from the observed values of the independent variables.

The most basic form of Regression is Linear Regression (**Fig. 13**) which is used to describe the linear relationship between a dependent continuous variable and one or more independent continuous, binary or categorical variables. Usually, a scatter plot of the two variables should be utilized as the initial judgment of their putative linear relationship. In the majority of real-life scenarios, the dependent variable cannot be explained by a single independent variable, therefore a multivariate linear regression model is needed in order to describe the effects of multiple variables on the dependent variable.

One way in which the General Linear Model differs from the Multiple Regression Model is in terms of the number of dependent variables that can be analyzed. The General Linear

Model goes a step beyond the multivariate regression model by allowing for linear transformations or linear combinations of multiple dependent variables. This extension gives the general linear model important advantages over the multiple and the so-called multivariate regression models, both of which are inherently univariate (single dependent variable) methods.

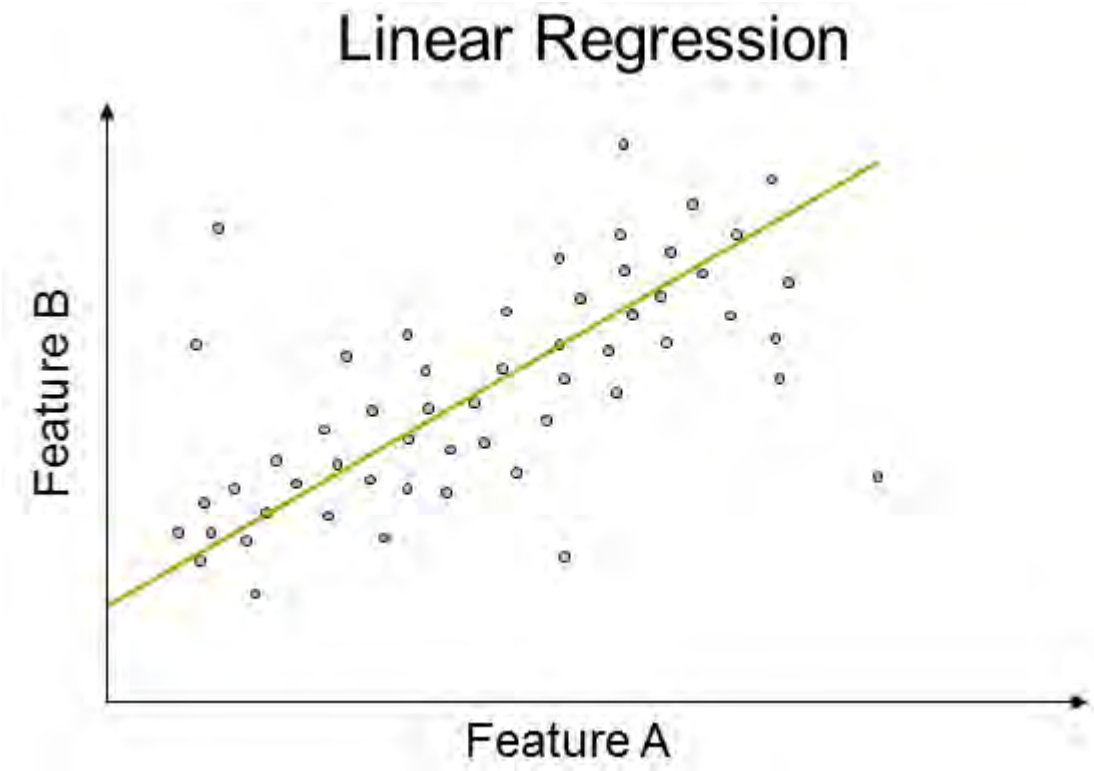


Figure 13. Linear Regression example. This figure has been designed for the purposes of this dissertation.

In many cases, we want to explore the relationship between dependent and independent variables but the dependent variable is not continuous. There are many situations where we have binary outcomes and the independent variables can be either continuous or discrete. In such cases, the solution is Logistic Regression, which belongs to the family of General Linear Models. The main mathematical concept behind Logistic Regression is the natural logarithm of an odds ratio. Considering a simple case with one continuous independent variable A and a dependent binary variable B, the visualization of such data would resemble two parallel lines (**Fig. 14**). Such lines cannot be described with typical least squares regression equations. A solution would be to create categories for the independent variable and calculate the mean of the dependent variable. These categories will appear linear in the middle but curved at the ends. Logistic Regression applies the natural logarithm of an odds ratio transformation B from A.

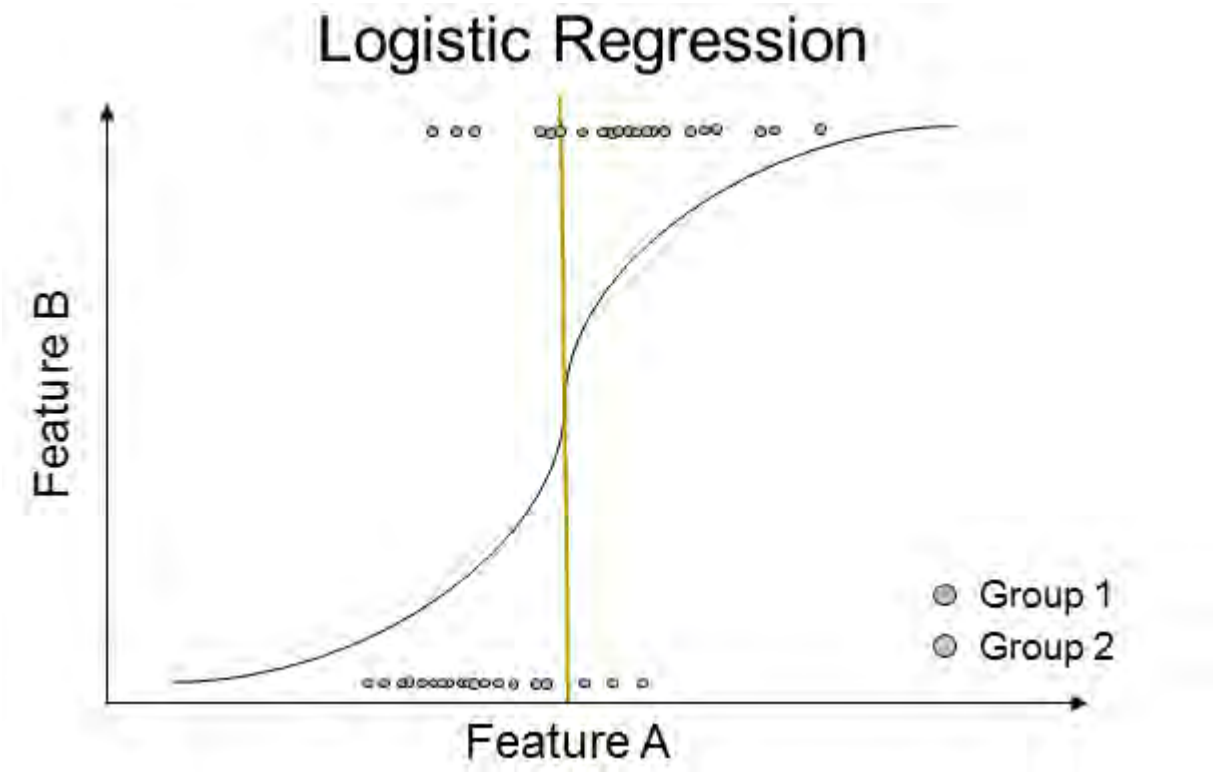


Figure 14. Example of Logistic Regression. This figure has been designed for the purposes of this dissertation.

1.10.3. Decision Trees

One could describe Decision Trees as Supervised models built to work sequentially by combining tests that compare a single numeric value against a threshold or a set of putative values.

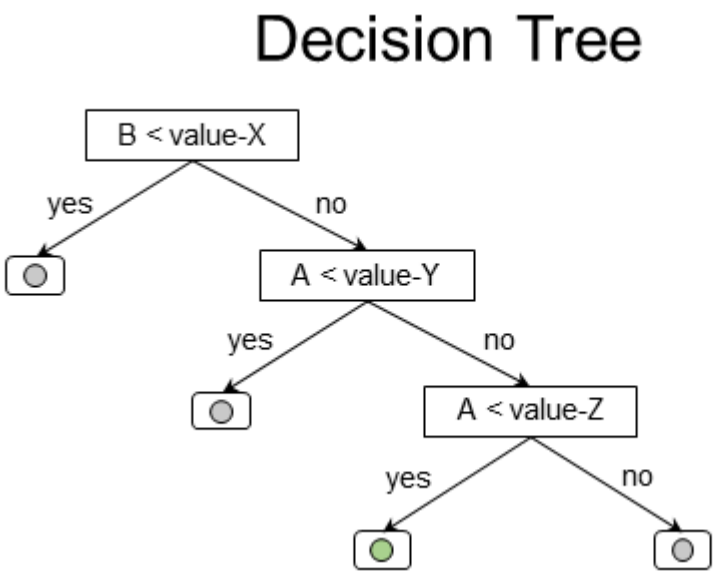


Figure 15. Example of a Decision Tree's rules. This figure has been designed for the purposes of this dissertation.

A Decision Tree classifies instances (**Fig. 15,16**) by incorporating a number of questions related to the characteristics associated with each instance. Every node represents such a question and each child node represents the outcome of its parent node question. It is only natural that these questions form a structure that resemble a tree. An instance is therefore classified into a category by traversing the path from the root to a leaf depending on the answers at each intermediate node. Every leaf is associated with a class. In certain variations of the algorithm, the leafs correspond to probability distributions that estimate the conditional probability that an instance belongs to a certain class.

Decision Trees have an advantage over models such as Artificial Neural Networks or Support Vector Machines because they combine simple questions about the data in an understandable way.

Figure 16. Example of a Decision Tree's plot. This figure has been designed for the purposes of this dissertation.

Decision Trees are considered flexible due to their ability to work on instances with a combination of continuous and categorical features and instances with missing features and they inherently support multiple class classification problems.

1.10.4. Artificial Neural Networks

Many years after 1943, when the first Artificial Neuron was introduced by McCulloch and Pits, the field of Artificial Neural Networks has emerged as a Machine Learning research hotspot when during 1980's, scientists recognized real potential in Neural Networks.

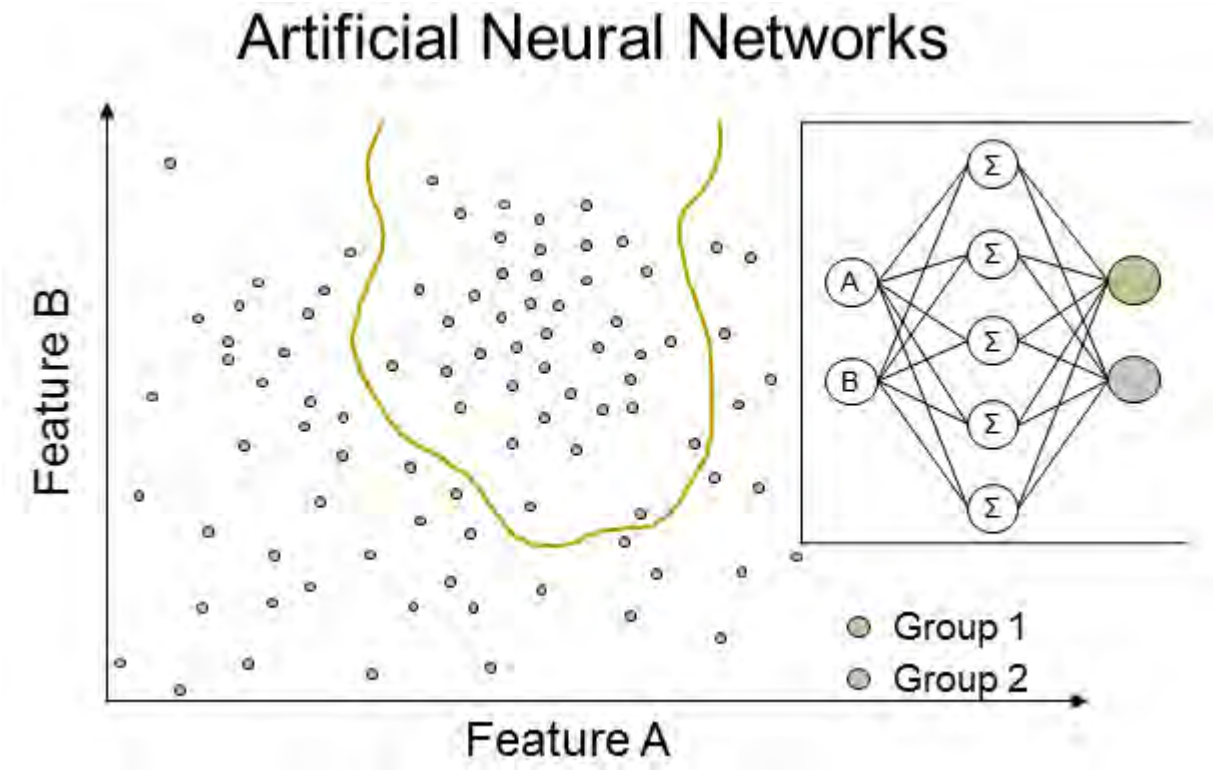


Figure 17. Example of Artificial Neural Network. This figure has been designed for the purposes of this dissertation.

The Neural Network was inspired by the way that the human brain processes information. Such a network consists of numerous interconnected processing nodes that learn by example (**Fig. 17**), the same way as a network of biological neurons works. Neural Networks are considered Supervised Learning algorithms since they rely on already labelled instances in order to calculate the weights of each intermediate processing node. Neural Networks are typically utilized in pattern recognition due to their generalization and accurate response to unexpected patterns ability.

The training phase of a Neural Network is based on observing labeled instances through iteratively adjusting each processing node weight. The goal of the iteration process is to progressively improve the model in such a way that the error rate is minimized. Neural Networks have their own limitations since it is a common phenomenon to fall into local minimum during the training process. Since the early 1990's Evolutionary Algorithms and later on Genetic Algorithms were utilized in in order to optimize the network design, pre-process the input data and assemble the Neural Network.

1.10.5. Support Vector Machines

The field of Support Vector Machines (SVM) originates from the work of Vapnik and Chervonenkis in 1974 and since then it has been increasingly gaining popularity due to their promising performance and elegant features. Support Vector Machines are

Supervised Learning models that utilize a fundamentally geometric idea in order to classify instances. This idea could be summarized in an attempt to project the instances to a higher dimensional space where the two classes can be separated by the construction of a hyperplane (**Fig. 18**). This corresponds to constructing such a hyperplane that the first side contains examples from the first class and the second side contains the instances of the second class. The possibility exists where an infinite number of separating hyperplanes exist. In such a case, the algorithm decides which hyperplane to keep by considering the maximization of the margin between the two classes. Machine Learning algorithms are typically benchmarked on the basis of their generalization error of the error rate when applied on unlabeled instances.

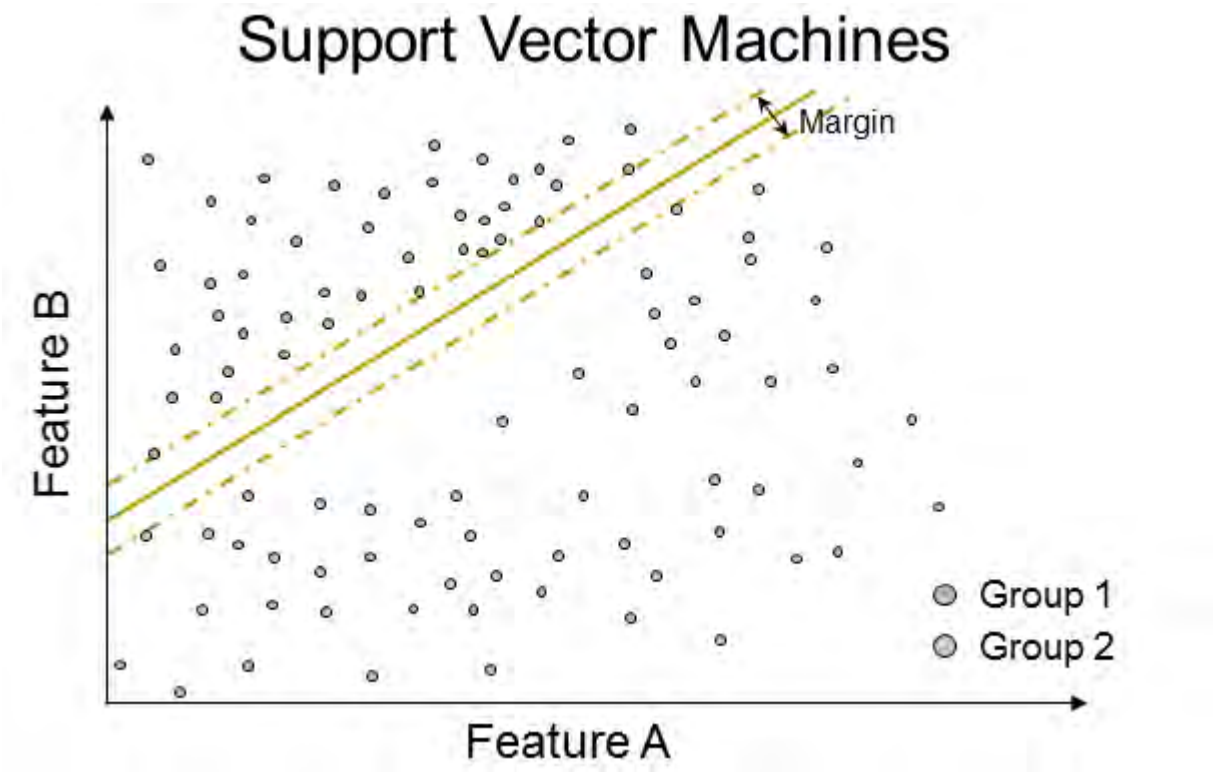


Figure 18. Support Vector Machines example. This figure has been designed for the purposes of this dissertation.

Support Vector Machines have become popular and are utilized in many scientific fields due to their ability to learn independently of the feature space dimensionality. This is summarized in the fact that the complexity of the input data is measured not by the number of features but from the margin that separates them, allowing them to generalize independently of the number of feature dimensions.

1.10.6. Clustering

During the early 1960's the work of Sokal and Sneath triggered the research on the development of clustering techniques. Clustering algorithms comprise the majority and most renown Un-supervised Learning techniques and as such they are utilized in discovering structure on unlabeled data. The are three types of Clustering approaches: Hierarchical, Data Partitioning and Data Grouping. The most notable clustering algorithm is "k-means", which has been proposed by many scientists in different forms in the past, is based on the sum-of-squares criterion and belong in the Data Partitioning family (**Fig. 19**).

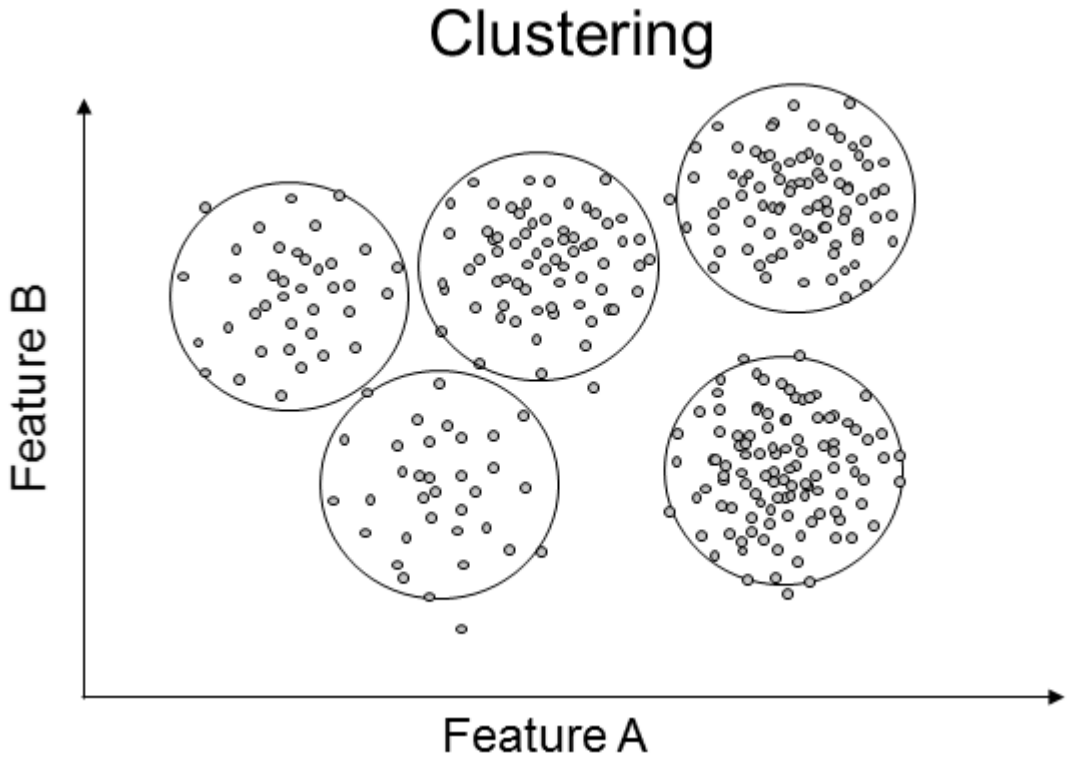


Figure 19. Example of k-means clustering, where k=5. This figure has been designed for the purposes of this dissertation.

Cluster analysis typically arranges instances in groups solely based on the information found within the data and their in between relationship. Clustering algorithms attempt to create such groups in a way that the members of each group have more similarities with one another than with the members of other groups. Increased similarity between members of a group can be translated in increased distance between groups. In order to achieve, the algorithm initiates the process by randomly (or defined by the user) selecting one point for each cluster. Each instance is assigned to the closest point typically based on a distance metric such as Euclidean distance and the group of instances connected

with each point comprises the cluster. This process is repeated until no changes in the clusters are observed.

2. Characterization of microRNA transcription regulation

The main objective of my Doctoral studies was the accurate and genome-wide characterization of miRNA gene transcription start sites and regulatory regions. Even though significant progress has been achieved for the identification of miRNA function, information regarding miRNA transcription regulation still remains significantly limited. Such knowledge will enable the genome-wide identification of miRNA expression regulators, including transcription factors (TFs), other non-coding RNAs and epigenetic modifiers; providing significant breakthroughs in understanding the mechanisms underlying miRNA expression in development and disease.

During the past few years, *in silico* miRNA promoter recognition methods have been elaborated, as a means to address the increased difficulty of high throughput miRNA promoter identification. Initial approaches (Saini, Enright, & Griffiths-Jones, 2008; Saini, Griffiths-Jones, & Enright, 2007; Zhou, Ruan, Wang, & Zhang, 2007) utilized DNA sequence features such as over-represented k-mers, transcription factor weight matrices and CpG content extracted from well annotated promoters of protein-coding genes, which were subsequently applied to identify promoters proximal to miRNA loci. These techniques provided the first indications of miRNA transcription start site positions on a genome-wide scale. However, they exhibit high false-positive rates and require vigorous filtering and validation of the provided results.

Megraw *et al* (Megraw et al., 2009) proposed S-Peaker, a model for "single-peaked TSS" identification based solely on known transcription factors and their respective regions of positional enrichment. In this work, Cap Analysis of Gene Expression (CAGE) data have been utilized in order to derive training and test sets and categorize promoters into single-peak and multi-peak TSSs based on the width of CAGE peaks. S-Peaker provides a probabilistic score for each nucleotide in the search space upstream of miRNAs. This score reflects the nucleotide's likelihood of being a TSS. S-Peaker supports multiple predictions per miRNA that include clusters of similarly scored nucleotides, forming peaks. Depending on the probability threshold, the width of these peaks may vary from tens up to hundreds of nucleotides.

Other studies (Barski et al., 2009; Corcoran et al., 2009; Ozsolak et al., 2008) utilize experimental data from active transcription marks (i.e. H3K4me3, Pol2 and nucleosome positioning) derived from high-throughput techniques such as ChIP-Seq. The methodology introduced by Marson *et al* (Marson et al., 2008) relies on H3K4me3 ChIP-Seq data. The algorithm considers regions enriched in H3K4me3 signals as putative

promoters. An empirically derived scoring system has been deployed to score each candidate region. Positive scores were given to enriched sites if they were either the start of a known gene or expressed sequence tag (EST) spanning the miRNA. Additional positive scores were given to enriched sites within 5 kb of the miRNA. Negative scores were assigned based on the number of intervening H3K4me3 sites and in the case where the enriched region could be assigned to a gene or EST not overlapping the miRNA.

The main disadvantage of techniques utilizing active transcription marks and sequence characteristics is the underlying low resolution and thus non-informative broad predictions. Deep sequencing data from epigenetic modifications and TF binding motifs are indicative of broad promoter regions and are unable to support high-resolution TSS identification (**Fig. 12**).

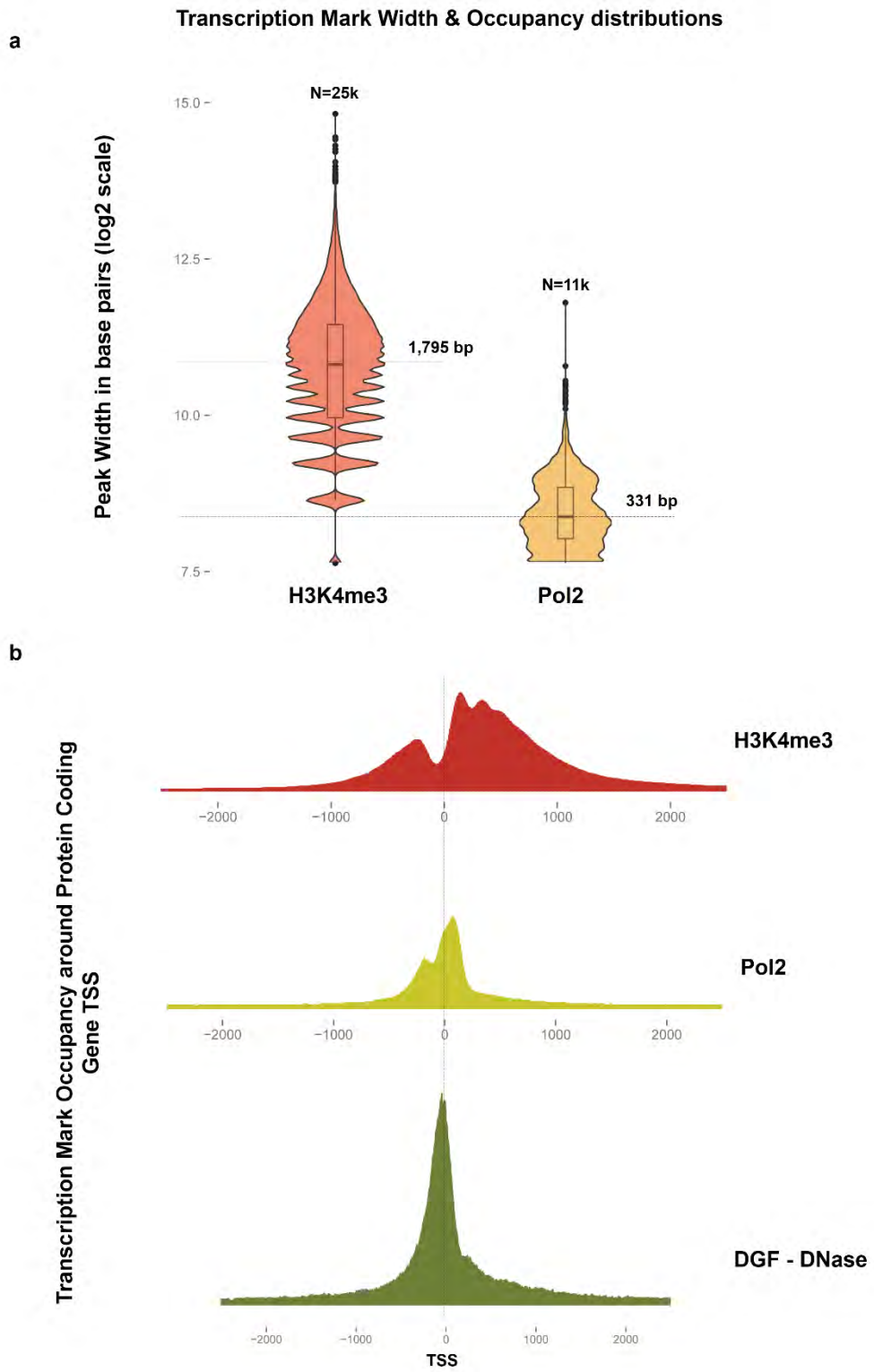


Figure 20. Transcription marks utilized by microTSS. a) Comparison between H3K4me3 and Pol2 peak width. b) H3K4me3, Pol2 and DGF coverage distribution around protein-coding genes. This image has been taken from the relevant publication of microTSS (Georgakilas et al., 2014).

miRStart (Chien et al., 2011) is a computational approach that integrates CAGE with TSS-Seq and H3K4me3 ChIP-Seq datasets. The algorithm utilizes these data in order to extract a signature profile around the TSS of protein coding genes, which is subsequently considered as the basis for training a Support Vector Machine (SVM) model. The SVM model identifies putative promoter regions upstream of mature miRNAs. miRStart filters each candidate promoter based on the distance from the corresponding miRNA and the number of overlapping ESTs or protein coding exons.

PROmiRNA (Marsico et al., 2013) is one of the latest and most advanced available algorithms. PROmiRNA utilizes CAGE data from all available tissues in FANTOM 4 database and combines them with sequence features for the characterization of miRNA promoters. It especially emphasizes in intronic miRNAs. The algorithm considers loci upstream of precursor miRNAs enriched in CAGE signals as putative promoters. Each candidate as well as randomly selected intergenic and intronic regions serve as positive and negative examples for training a probabilistic model which additionally incorporates CpG content, conservation, TATA box affinity and mature miRNA proximity.

TSS identification algorithms utilizing NGS data can be further divided into two distinct categories based on the scope of their predictions: a) generalized algorithms and b) experiment-specific. The first group comprises algorithms integrating data derived from multiple cell lines (e.g. PROmiRNA) or DNA motif analysis (e.g. S-Peaker), providing multiple predictions per miRNA that correspond to different promoters, potentially active in different tissues, cell lines and conditions. These algorithms can suggest in a single run different putative miRNA TSS locations but cannot identify those active in a specific experiment (e.g. cell line, treatment or tissue), since they are agnostic to its conditions. The second group (e.g. Marson *et al*) utilizes NGS data from a specific experiment and provides a "snapshot" of the currently active promoters in the investigated tissue or cell line. Such *in silico* methodologies enable experimentalists to focus only on those promoters that are active in the cell type or condition of interest and use their results as a stepping stone for building tissue specific regulatory networks or to identify interventions. On the other hand, these methodologies require separate runs using data from different experiments, in order to map promoters active in different conditions.

A common characteristic for existing studies in both categories is the absence of a rigid high-throughput experimental framework for validating their predictions. Well-established techniques such as 5′ RACE and RT-PCR coupled with promoter cloning are frequently utilized in the scope of miRNA promoter validation. These protocols are time consuming and low-throughput since they support single promoter validation per experiment. Most available algorithms utilized indirect means of validation (e.g. existence of Pol2 ChIP-Seq signals near the prediction site) and/or direct testing of selected 1-2 promoters as proof of concept.

Table 5. Detailed information regarding the analysis of raw RNA/GRO/ChIP-Seq data and the number of DGF TF binding sites utilized during the development of microTSS algorithm. This table has been taken from the relevant publication of microTSS (Georgakilas et al., 2014).

RNA/GRO/ChIP/DNase-Seq data utilized in the study			
GEO accession	**Uniquely mapped reads**	**Total reads**	**Specifications**
WT mESCs RNA-Seq			
GSM973235 Rep 1	180M	236M	PE, 101bp
GSM973235 Rep 2	250M	328M	PE, 101bp
WT hESCs RNA-Seq			
CSHL-H1EScs-Nucleus-PolyA+ Rep 2	174M	208M	PE, 76bp
WT human IMR90 RNA-Seq			
GSM981249	218M	N/A	N/A
Drosha -/- mESCs RNA-Seq			
Replicate 1 - GSM1342579	8.2M	11.4M	SE, 32bp
Replicate 2 - GSM1342580	8M	10.9M	SE, 32bp
Replicate 3 - GSM1342581	10.9M	15M	SE, 32bp
Drosha +/+ mESCs RNA-Seq			
Replicate 1 - GSM1342576	7.4M	10.3M	SE, 32bp
Replicate 2 - GSM1342577	2.7M	3.81M	SE, 32bp
Replicate 3 - GSM1342578	8.9M	12.2M	SE, 32bp
WT mESCs small RNA-Seq			
GSM886478	14.4M	23.9M	SE, 50bp
WT H3K4me3 mESCs ChIP-Seq			
GSM723017	18.9M	23M	SE, 36bp
WT H3K4me3 hESCs ChIP-Seq			
GSM605315	10M	20.4M	SE, 36bp
WT H3K4me3 human IMR90 ChIP-Seq			
GSM1055816*	13.7M	N/A	N/A

WT Pol II mESCs ChIP-Seq			
GSM723019	13.6M	18M	SE, 36bp
WT Pol II human IMR90 ChIP-Seq			
GSM935513*	20.6M	N/A	N/A
WT Pol II hESCs ChIP-Seq			
GSM803366	13.2M	40M	SE, 36bp
WT mESCs GRO-Seq			
GSE27037 (mESC Rep1| Rep2| Rep3)	14.1M	17.7M	SE, 36|36|35bp
WT hESCs GRO-Seq			
GSM1006728 (hESC Rep1|Rep2|Rep3)	241M	331M	SE, 101|50|40bp
WT human IMR90 GRO-Seq			
GSM1055806*	33.5M	N/A	N/A
WT mESCs DNase-Seq			
GSE40869	623K DGFs		
WT hESCs DNase-Seq			
GSE32970	1.2M DGFs		
WT human IMR90 DNase-Seq			
GSM1008586	970K DGFs		

Until recently, most RNA-Seq studies provided limited sequencing depth and were not sensitive enough to capture the elusive pri-miRNA transcripts, due to increased cost and/or technical limitations. Recent improvements in deep sequencing enabled the creation of datasets comprising more than 200 million reads per sequenced sample. Such data are already available from extensive consortia and collaborations (e.g. ENCODE Consortium). The detailed analysis of such RNA-Seq datasets derived from 2 mouse embryonic stem cell (mESC) replicates comprising more than 430 million uniquely mapped reads (**Table 5**) revealed that pri-miRNA transcripts can be detected in datasets of high sequencing depth (**Fig. 13**).

We therefore hypothesized that the *in silico* examination of such datasets, by utilizing machine learning algorithms empowered with multiple signatures of active transcription marks, could provide accurate and high-resolution miRNA TSS identification.

Importantly, extensive experimental validation of the *in silico* identified miRNA promoters was considered essential for the determination of the implementation's accuracy and performance, as well as for comparison with previously elaborated methodologies.

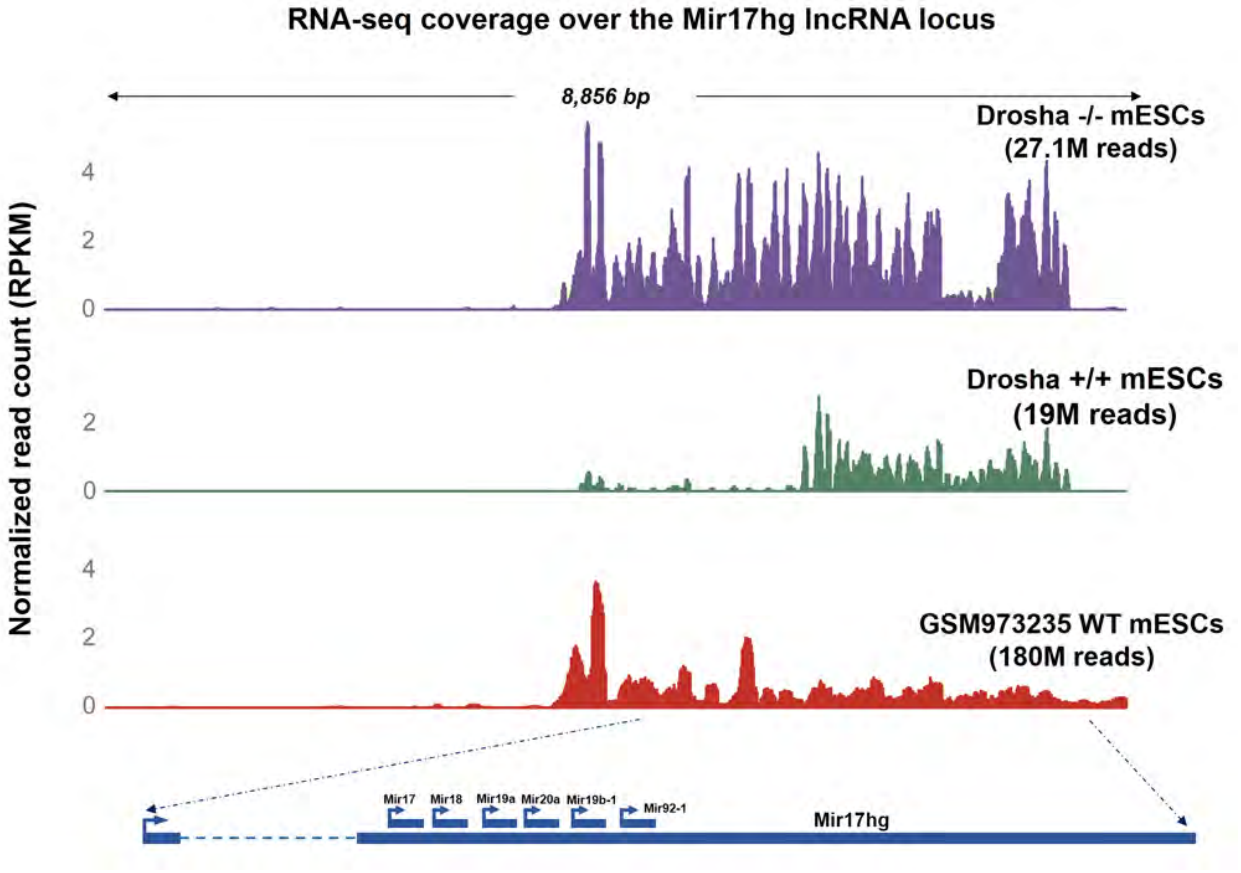


Figure 21. Comparison of RNA-Seq coverage between Drosha -/- and wild-type mouse ESCs. The example depicts Mir17hg locus transcribing a cluster of 6 precursor miRNAs. Purple color represents the coverage of Drosha -/- mouse ESCs (~27M uniquely mapped SE reads), while green color is utilized for Drosha +/+ ESCs (~19M uniquely mapped SE reads). The "normal-depth" Drosha +/+ dataset depicts the effect of Drosha processing, which is the main reason for the current lack of pri-miRNA TSS characterization. Currently annotated Mir17hg TSS is close to the start site of Drosha +/+ Mir17hg expression. Red color represents the coverage of the deeply sequenced RNA-Seq dataset (~250M uniquely mapped PE reads) from wild-type mouse ESCs derived from ENCODE project. This figure illustrates the ability of Drosha -/- and deeply sequenced RNA-Seq datasets to capture the elusive pri-miRNA expression. In addition, it shows that the TSS of Mir17hg is clearly upstream from its currently annotated position. This image has been taken from the relevant publication of microTSS (Georgakilas et al., 2014).

To this end, we implemented an experimental, as well as a computational framework for high-throughput miRNA TSS identification. The former consists of a *Drosha* null/conditional-null (*Drosha*$^{LacZ/e4COIN}$) mouse model that has been generated using the novel conditional by inversion (COIN) methodology (Economides et al., 2013). Whole transcriptome sequencing from mESCs derived from *Drosha*$^{LacZ/e4COIN}$ resulted to an

extensive set of experimentally identified miRNA TSSs. This experimentally derived dataset was kept as an independent test set, and was utilized for the thorough evaluation of the computational methods.

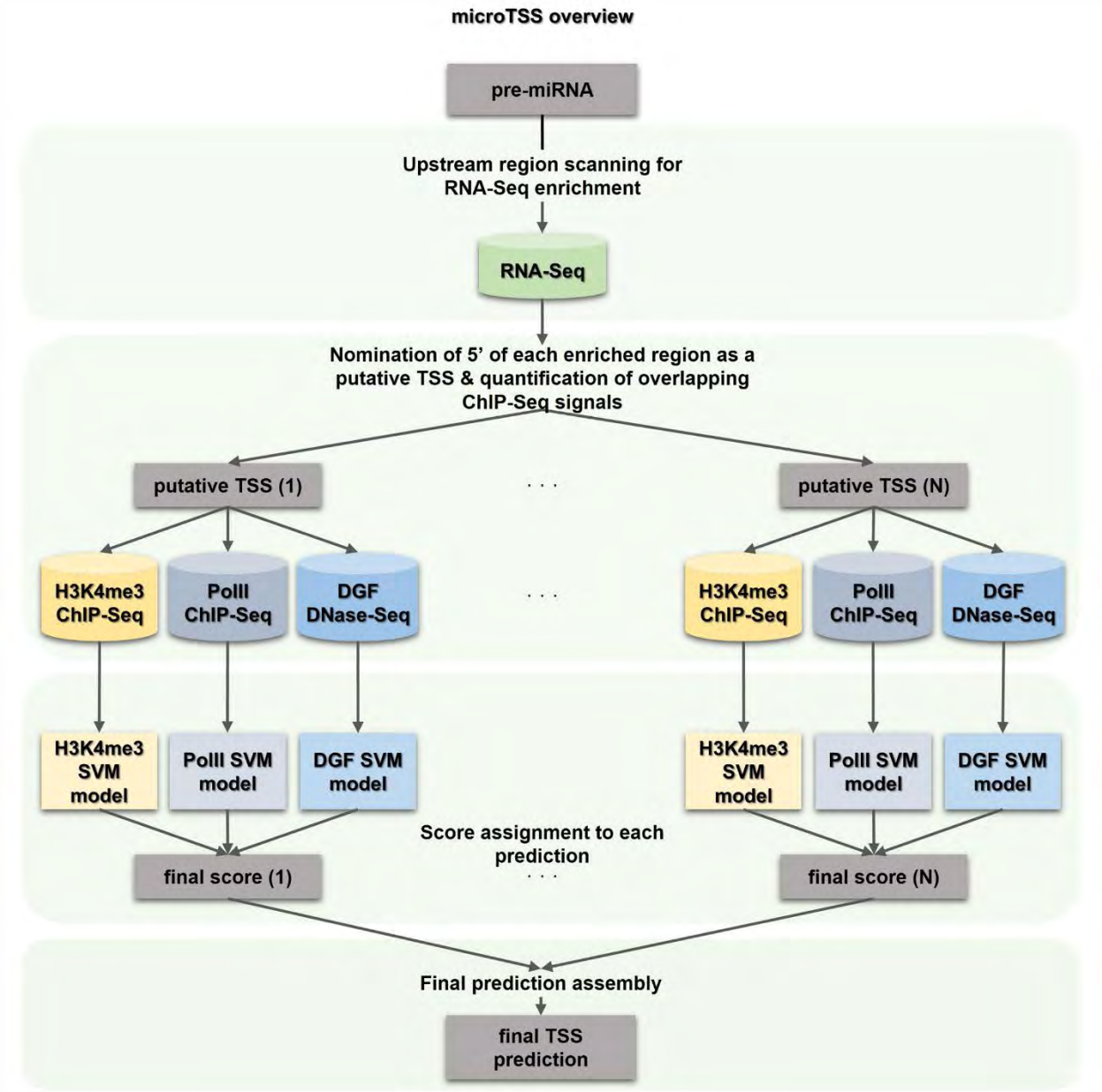


Figure 22. Overview of microTSS algorithm. For each precursor, microTSS utilizes a sliding window initialized at the pre-miRNA genomic location and identifies upstream regions enriched in RNA-Seq signal. The 5' end of each identified enriched locus is treated as a TSS candidate. The area surrounding each candidate is divided into bins of fixed/predefined size and different for each transcription marker (H3K4me3, Pol II and DNase-derived TF footprints). Each bin is assigned a score which represents the number of overlapping ChIP-Seq reads and TF footprints. Three separately trained SVM models utilize the scored bins as features and emit probabilistic estimates (one for each transcription mark) which are subsequently combined to a final score. This image has been taken from the relevant publication of microTSS (Georgakilas et al., 2014).

The latter, microTSS (Georgakilas et al., 2014), is an *in silico* approach that focuses on the identification of intergenic miRNA TSSs and relies on deeply sequenced RNA-Seq data. The algorithm integrates RNA-Seq data by creating "islands" of transcription (i.e. regions with increased RNA-Seq coverage) upstream of intergenic pre-miRNAs. The 5′ end of each identified expressed region is treated as a putative TSS (**Fig. 14**). This step is the backbone of the algorithm since it provides TSS candidates with single nucleotide resolution. A combination of three independent SVM models is subsequently utilized to score each candidate TSS and derive the final predictions. These SVM models have been trained on H3K4me3 and Pol2 occupancy around protein coding TSSs, as well as on the existence of open chromatin domains, as identified by DNase-Seq (**Fig. 15**). microTSS is finally tested against TSSs identified using *Drosha* null/conditional-null mESCs, as well as TSSs detected using deeply sequenced global run-on sequencing (GRO-Seq) data in human IMR90 and ES cells.

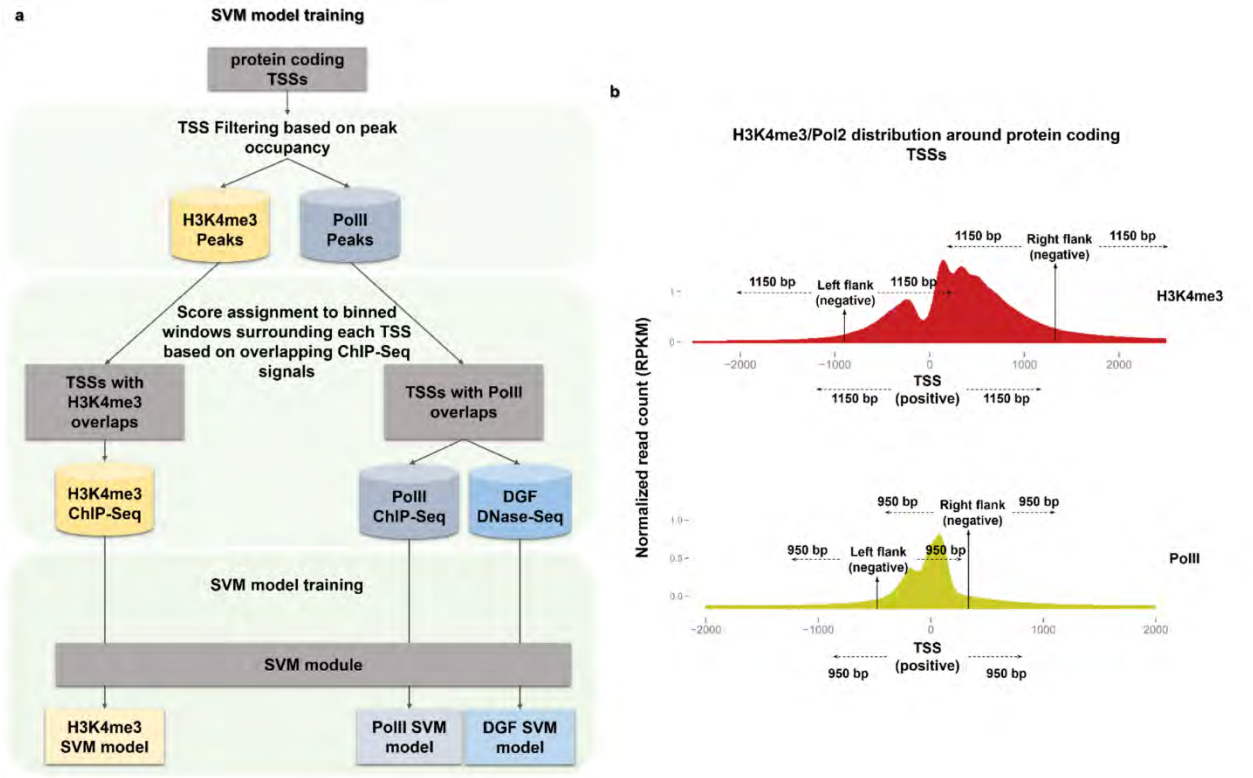


Figure 23. SVM training pipeline and H3K4me3/PolII occupancy around the TSSs of protein coding genes. a) The initial set of protein coding TSSs is divided into two subsets based on H3K4me3 or Pol II occupancy. The region surrounding each TSS is divided into bins and each bin is assigned a score, which is the number of overlapping ChIP-Seq reads or TF footprints. Subsequently, the scored bins are utilized as features in order to develop three separately trained SVMs, modeling the distribution of each transcription mark around protein coding TSSs. b) In order to train the SVM models, the annotated TSSs were selected as positive instances and the flanking regions of each active transcription mark as negatives. In addition, two randomly selected intergenic spots are selected as negatives, resulting in a 1:4 positive to negative ratio. The area (+/- 1,150 bp and +/- 950 bp for

H3K4me3 and Pol II, respectively) surrounding each instance is divided in similarly scored bins of 100 nts. Both Polymerase II and DGF models share the same training set, while the region (+/- 2050 bp) surrounding each DGF instance is divided in bins of 200 bps (not shown). This image has been taken from the relevant publication of microTSS (Georgakilas et al., 2014).

2.1. Methods

2.1.1. Drosha-null and Drosha-wild-type data generation

Drosha-null mESCs were generated by treating *Drosha*$^{ex4COIN/LacZ}$; *Gt(ROSA)26Sor*$^{CreERt2/+}$ cells with Tamoxifen (500ng/ml) to activate the *CreERt2* recombinase, and clones with inverted COIN module were identified, one of which (LD12) was used in this study. LD12 exhibits abrogation of *Drosha* expression and absence of a mature microRNA miR-293, and concomitant accumulation of its precursor pri-miR-293, indicating lack of *Drosha* functionality (Economides et al., 2013). Mouse ES cells of WT or *Drosha*-null genotype were cultured on gelatinized plates free of feeder cells. Total RNA was extracted by miRNeasy Mini Kit (Qiagen). 2 μg of total RNA was converted to poly(A)+ RNA using oligo-dT coated magnetic beads (Invitrogen). Poly(A)+ RNA was converted to strand specific Illumina sequencing libraries with 8 bp barcodes using Epicenter ScriptSeq V1 RNA-Seq library preparation kit (Epicenter, Illumina Inc). RNA-Seq libraries were hybridized to a single-end flow cell and individual fragments were clonally amplified by bridge amplification on the Illumina cBot. Upon completion of clustering, the flow cell was loaded on the HiSeq 2000 (Illumina Inc, USA) and sequenced using Illumina's SBS chemistry. Samples were run for 33 bp sequencing reads as well as 9 bp index reads. Base call (.bcl) files for each cycle of sequencing were generated by Illumina Real Time Analysis (RTA) software and de-multiplexed to FASTQ files, which were used for analysis in this study.

2.1.2. RNA-Seq and GRO-Seq analysis

Apart from the generated *Drosha* -/- and +/+ RNA-Seq datasets, mESC RNA-Seq data have been derived from the ENCODE consortium repository (GEO accessions GSE49847 and GSM758574). GRO-Seq data were obtained from the studies of Min *et al* (Min et al., 2011), Sigova *et al* (Sigova et al., 2013) and Jin *et al* (Jin et al., 2013). Quality control has been performed using FastQC (www.bioinformatics.babraham.ac.uk/projects/fastqc). Contaminants were detected and removed utilizing a combination of an in-house developed algorithm and already available tools such as minion (Davis, van Dongen, Abreu-Goodger, Bartonicek, & Enright, 2013) and trimgalore (www.bioinformatics.babraham.ac.uk/projects/trim_galore). Following pre-processing, GSNAP spliced aligner (T. D. Wu & Nacu, 2010) was utilized to map the reads against the reference genomes (GRCm38/mm10 and GRCh37/hg19 genome assemblies). GSNAP has been appropriately parameterized in order to detect novel and known splice junctions. The analysis resulted in ~849M uniquely mapped paired-end (PE) reads (WT

RNA-Seq), ~27M uniquely mapped single-end (SE) reads (*Drosha*-null mESCs RNA-Seq) and ~288M uniquely mapped SE reads (wild-type GRO-Seq). GRO-Seq data were aligned against the genome using Bowtie v1 (Langmead, Trapnell, Pop, & Salzberg, 2009). Reads aligned to more than one genomic location have been discarded from subsequent analyses (**Table 5**). Differential expression analysis was performed using EDGER (Robinson, McCarthy, & Smyth, 2010).

2.1.3. Small RNA-Seq analysis

ESC small RNA-Seq data were derived from the study of Chang *et al* (G. Chang et al., 2014). Following pre-processing, adapter trimmed reads were aligned against known human mature-miRNA sequences (miRBase v20) using Bowtie v1. Unaligned reads were subsequently mapped against known pre-miRNAs. Reads mapped on pre-miRNAs not clearly overlapping a mature miRNA sequence were discarded. Alignments on identical mature miRNAs deriving from distinct pre-miRNAs were collapsed. Identification of the miRNA expression was finally estimated on mature miRNA level by combining both alignment results.

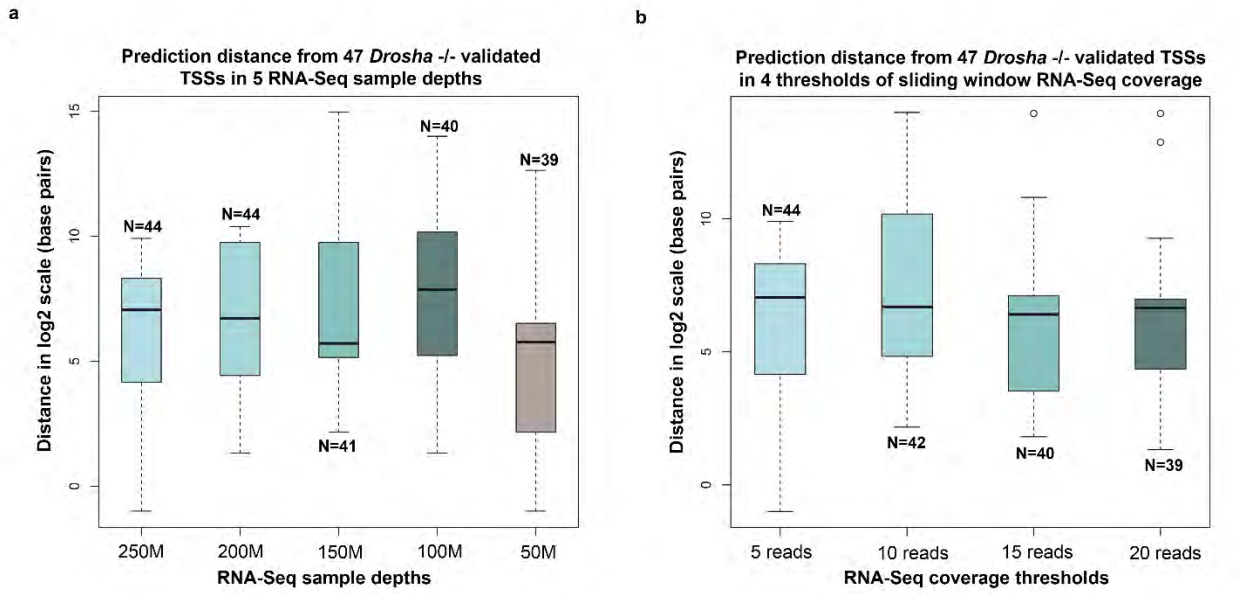
2.1.4. ChIP-Seq and DNase-Seq analysis

ESC raw H3K4me3 and Pol2 ChIP-Seq data have been derived from the published collection of Shen *et al* (Shen et al., 2012), Derrien *et al* (Derrien et al., 2012) and Jin *et al* (Jin et al., 2013). Quality control and contaminant removal was performed using the same tools and techniques as for RNA-Seq and GRO-Seq data. Bowtie v1 has been utilized in order to align the reads to the reference genome (GRCm38/mm10 and GRCh37/hg19 genome assemblies). The analysis resulted in ~42M and ~47M uniquely mapped H3K4me3 and Pol2 reads respectively. SICER (Zang et al., 2009) and Macs2 (Zhang et al., 2008) have been used in order to identify enriched regions in H3K4me3 and Pol2 signals. Digital Genomic Footprinting (DGF) data produced by DNase-Seq have been derived from the ENCODE consortium repository (GEO accessions GSE40869, GSE32970 and GSM1008586) and the migration from mm9 to mm10 has been accomplished using liftover tool provided by UCSC (**Table 5**). The integration of these datasets has facilitated the training and optimization processes of the SVM models.

2.1.5. Description of the algorithm

microTSS is composed of two distinct modules (**Fig. 14**). Initially, the algorithm identifies regions enriched with RNA-Seq reads upstream of intergenic pre-miRNAs. This is accomplished by utilizing a sliding window initialized at the pre-miRNA genomic location, covering a user-defined distance. Each window is assigned a score which represents the number of overlapping RNA-Seq reads. The applied window size, sliding step and score threshold are also parameterized. The suggested parameters are 30 nts as

default length for the sliding window and the relevant score threshold at 5 overlapping RNA-Seq reads (**Fig. 16**), regardless of sequencing depth.



c

Number of microTSS predictions on expressed pre-miRNAs, in 5 RNA-Seq sample depths

Expressed precursors	Intragenic*	Intergenic	250M	200M	150M	100M	50M
N=186	52	134	57	55	52	49	35
		Total Predictions	118	110	97	82	59

d

Number of microTSS predictions on expressed pre-miRNAs, in 4 thresholds of sliding window RNA-Seq coverage

Expressed precursors	Intragenic*	Intergenic	Sliding window threshold=5	Sliding window threshold=10	Sliding window threshold=15	Sliding window threshold=20
N=186	52	134	57	49	42	39
		Total Predictions	118	85	68	59

Figure 24. microTSS performance following perturbations of key parameters in mESCs. a) Random subsampling of the original WT RNA-Seq sample derived from ENCODE (GSM973235 Rep 2) has been accomplished with samtools. The algorithm has been applied on each subsample with the rest parameters fixed, in order to assess the importance of sequencing depth in miRNA TSS identification. b) The algorithm has been applied on the original WT RNA-Seq sample, with different thresholds for the sliding window RNA-Seq coverage. Evaluation of the algorithm based on prediction distance from the Drosha -/- RNA-Seq validated TSSs is shown in the boxplot. The algorithm's sensitivity in predicting TSSs of expressed pre-miRNAs is shown in the tables. c), d) Detected expressed pre-miRNAs in various sequencing depth and RNA-Seq coverage thresholds. This image has been taken from the relevant publication of microTSS (Georgakilas et al., 2014).

These are microTSS recommended default values that result in maximum sensitivity without compromising the algorithm's accuracy. microTSS filters out windows according to the threshold score and merges the remaining ones based on a user-defined distance, enabling the identification of genomic loci enriched in RNA-Seq reads. Assessing the performance of the algorithm for a wide range of this parameter values, we observed that a robust selection, in terms of sensitivity and precision, is 200 nts. The length of the scanning region upstream of each pre-miRNA has been set to 400 kb. This value has been selected based on previous studies that have identified TSSs located more than 100 kb away from their corresponding precursors and in some cases even 150 kb.

The 5' ends of the identified RNA-Seq enriched loci serve as putative TSSs. Subsequently, microTSS combines three Support Vector Machines (SVM) models in order to score each putative TSS and to filter out false positives. The SVM models have been trained on H3K4me3 and Pol2 ChIP-Seq as well as digital genomic footprint (DGF) of transcription factor binding occupancy on a set of annotated protein coding genes (**Fig. 15**). Each candidate TSS position is assigned three different windows of varying size, depending on the corresponding SVM model. The H3K4me3 window length is +/- 1,150 bp around the candidate TSS, while the Pol2 and DGF are +/- 950 bp and +/- 2,050 bp respectively. The H3K4me3 and Pol2 windows are divided in bins of 100nts, while DGF are divided in bins of 200nts. Each bin is assigned a specific score, which is the number of overlapping ChIP-Seq reads or TF footprints. The scores for all bins are subsequently forwarded to the SVM models as features, which in turn estimate the probability for the candidate position to actually include a bona fide TSS. The final score of each candidate TSS is the sum of the three probabilities. Cases exhibiting a final score below a threshold are filtered out. From the remaining candidates, microTSS reports the one corresponding to the highest final score.

2.1.6. Support Vector Machines model training

The promoters of miRNA genes have been shown to present similar characteristics with protein coding genes, since their transcription is regulated by Pol2. H3K4me3, Pol2 and TFs are considered key elements in the initiation of gene transcription. H3K4me3 has been found to occupy the promoters of actively transcribed genes or genes poised for transcription. TFs are required for recruiting the transcription machinery which is driven by Pol2. Due to the observed underlying hierarchy in promoter occupancy, in many cases TSSs of protein coding genes have been found to correlate only with H3K4me3 peaks, others have been shown to be occupied by H3K4me3 and TFs, while the majority is controlled by all three transcription marks.

In order to properly capture the information residing in each proposed active transcription mark, three distinct SMV models have been trained on a set of annotated

protein coding TSSs derived from Ensembl v74 (Flicek et al., 2013), utilizing ChIP-Seq data against H3K4me3 and Pol2 as well as DGF TF binding sites (**Fig. 15**).

The training procedure has been accomplished using LIBSVM v3 (C. Chang, Lin, C., 2011), which provides probability estimations instead of performing binary classification. Radial Basis Function (RBF) has been chosen over other kernels since it performed better in cross-validation tests. ChIP-Seq signals corresponding to TSSs of multiple genes with an in-between distance smaller than 10 kb have been filtered out. The finalized set of protein coding genes comprises 10,929 entries. This group of genes has been subsequently divided into two sets with a ratio of 4 to 1: 8,740 TSSs were utilized for training and 2,189 for testing the SVM models. SICER and Macs2 have been applied to identify genomic locations (peaks) enriched in H3K4me3 and Pol2 respectively, enabling the development of a robust predictive model. Peaks exhibiting a false discovery rate (FDR) higher than 0.05 have been filtered out. Out of the 8,740 protein coding genes in the training set, 4,504 have been found to overlap with H3K4me3 peaks and 1,623 with Pol2 peaks (**Table 6**). In order to train each model, the center of each peak served as a positive instance while the leftmost and rightmost positions were treated as negatives. In addition to the flanking positions of the peak, two randomly selected intergenic spots are selected as negatives, resulting in a 1:4 positive to negative ratio (**Fig. 15**). In order to develop a robust DGF model, its training set should consist of promoters with fully recruited TF machineries. This could only be the case for promoters occupied by Pol2. Thus, the set of protein coding genes utilized for training the Pol2 model has also served as training set for the DGF SVM model.

Table 6. microTSS performance on training (10-fold CV) and test set comprised of protein-coding genes. This table has been taken from the relevant publication of microTSS (Georgakilas et al., 2014).

Performance of microTSS on protein coding genes				
	Training Set			**Test Set**
	H3K4me3 10-CV (4,504 positives & 18,016 negatives)	**Pol2 10-CV (1,623 positives & 6,492 negatives)**	**DGF 10-CV (1,623 positives & 6,492 negatives)**	**Merging of Models (2,189 positives & 8,756 negatives)**
Accuracy	99%	98%	98%	99.5%
Precision	95%	95%	93%	98.2%
Specificity	98%	98%	98%	99.5%
Sensitivity	99%	96%	96%	99.7%

Ten-fold cross-validation has been performed on the training data in order to estimate the performance of each model, achieving 98% accuracy for the DGF model, 98% for the Pol2 model and 99% for the H3K4me3 model (**Table 6**). The protein coding test set was utilized in order to evaluate the performance of the final combined model, as well as to estimate its generalization ability and to avoid over-fitting. Even by applying a loose threshold on the final score of each prediction (as explained in the previous section) the algorithm can predict TSSs of the unknown test genes with 99.5% Accuracy, 98.2% Precision, 99.5% Specificity and 99.7% Sensitivity (**Table 6**).

microTSS was initially developed to combine H3K4me3, Pol2 and TF occupancy within a single model. However, this approach resulted in H3K4me3 consistently overshadowing/masking the other marks' properties. H3K4me3 consistently occupies transcription start sites but its binding region tends to be very wide. Pol2 and DGFs on the other hand, occupy fewer TSSs than H3K4me3 but in a significantly narrower region. **Figure 12** demonstrates the size of the binding region of each transcription mark, suggesting that all three features are informative and equally important. The score of each model acts as additive value/evidence strengthening the likelihood of each candidate TSS and removing the majority of false positives. The distribution of the score provided by each individual model remains unaffected by the expression level and is similar for both protein coding and miRNA genes.

2.1.7. Precursor miRNA spatial classification and conservation

Human and mouse pre-miRNAs have been divided into six categories depending on their genomic location relative to protein coding genes (**Fig. 17**). Precursors residing inside protein coding exons/introns have been classified as "exonic"/"intronic".

Table 7. Pre-miRNA classification in respect to protein-coding genes. The annotation has been derived from miRBase v20. This table has been taken from the relevant publication of microTSS (Georgakilas et al., 2014).

pre-miRNA classification		
Category	**Mouse pre-miRNAs (1,181)**	**Human pre-miRNAs (1,870)**
Intronic	551 (~46.6%)	808 (~43.2%)
Exonic	94 (~7.8%)	108 (~5.8%)
Antisense	96 (~8.1%)	175 (~9.3%)
Divergent	13 (~1.1%)	43 (~2.3%)
Intergenic	371 (~31.5%)	670 (~35.8%)

Read-through	56 (~4.7%)	66 (~3.6%)

miRNAs located in the opposite strand of protein coding loci were classified as antisense. Pre-miRNAs located in the immediate (less than 4,000 bp) upstream/downstream sense region of protein coding genes have been labeled as read-through. RNA-Seq signal profile at these loci suggests common transcription regulation for both coding and non-coding genes. On the other hand, miRNAs located in the upstream antisense region (less than 2,000 bp) of coding loci were classified as divergent. The remaining precursors were characterized as intergenic (**Table 7**).

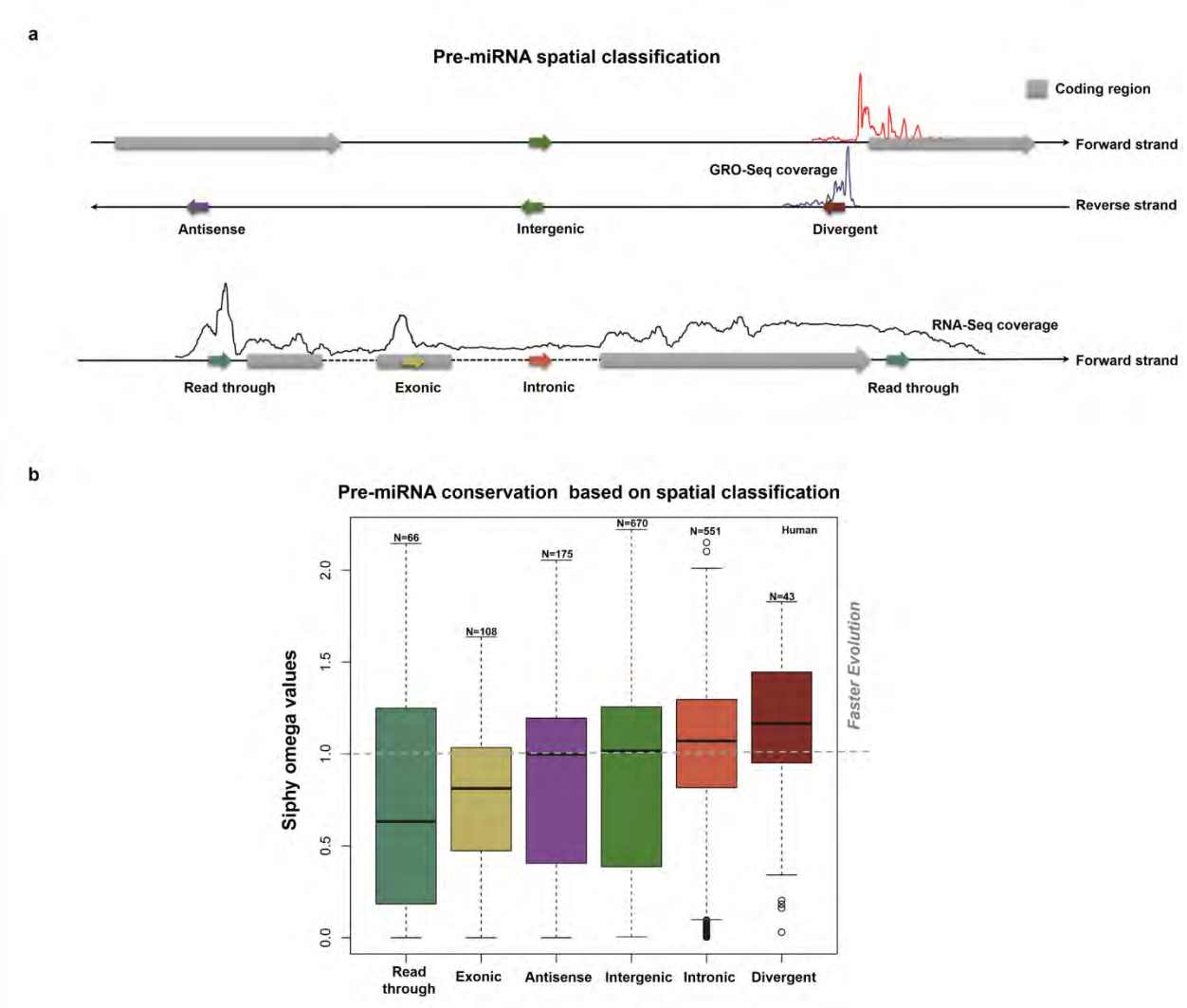


Figure 25. Precursor miRNA spatial classification and conservation. a) miRNA categories are based on their location relative to protein coding genes. b) Evolution rate for each spatial class as calculated by SiPhy. Divergent precursors have been found to be the least conserved group of miRNAs. This image has been taken from the relevant publication of microTSS (Georgakilas et al., 2014).

In order to identify the evolutionary rate of each category, multiple alignment files between 21 mammals in MAF format have been downloaded from UCSC repository. SiPhy (Garber et al., 2009) has been utilized to calculate the local rate of substitutions compared to a neutral phylogenetic tree model, which is depicted in the estimated omega values. Higher omega scores are associated to less conserved regions and precursors surpassing the cut-off value 1.0, as determined by SiPhy, are considered rapidly evolving sequences. Due to the limited amount of identified divergent miRNAs in mouse, statistical analysis on the conservation results has been performed only for human precursors.

2.2. Results

2.2.1. *Drosha* null/conditional-null mouse model

Drosha$^{LacZ/e4COI}$ mouse model was generated to enable the identification of full-length pri-miRNA transcripts, not processed by *Drosha* enzyme in the nucleus (Economides et al., 2013). The conditional-null allele of *Drosha* phenocopies the null allele both in mESCs and in mice, upon conversion to the null state with Cre. Lack of *Drosha* enzyme expression results in an abundance of unprocessed, full-length pri-miRNA transcripts that can be readily identified. Whole transcriptome sequencing of *Drosha*-null mESCs resulted in the identification of 22 high-quality intergenic miRNA gene TSSs, incorporating 47 pre-miRNAs. The validated miRNA TSSs were utilized to assess the accuracy of the implemented microTSS algorithm.

2.2.2. Comparison between *Drosha*-null and *Drosha*-wild-type

Drosha-null samples exhibited significantly increased coverage of pri-miRNA regions compared to WT (**Table 8**). Differential expression analysis was performed on the set of verified pri-miRNAs following removal of the hairpin pre-miRNA region, in order to identify differences in coverage of the pri-miRNA portion which is normally cleaved within the nucleus.

Table 8. Differential expression analysis between the Drosha +/+ and Drosha -/- samples for 21 experimentally derived pri-miRNAs. The table provides reads per kilobase per million uniquely mapped reads, normalized expression (RPKM), log2 fold change (log2fc) and false discovery rate levels (FDR). This image has been taken from the relevant publication of microTSS (Georgakilas et al., 2014).

Differential Expression analysis of identified pri-miRNA regions between Drosha -/- and +/+ mice				
pri-miRNA (associated lncRNA)	**Drosha +/+ RPKM**	**Drosha -/- RPKM**	**log2FC**	**FDR**

pri-mir-101a (E130102H24Rik)	6.18	35.69	2.42	4.14E-21
pri-mir-1199	9.61	8.50	-0.15	0.834
pri-mir-130a	8.31	11.44	0.57	0.0512
pri-mir-142	0.33	1.75	2.61	2.31E-04
pri-mir-16-1/15a (Dleu2)	16.28	13.60	-0.50	0.34
pri-mir-183/182/96	2.41	83.57	5.11	3.67E-97
pri-mir-1839 (2900076A07Rik)	16.09	7.26	-1.16	2.83E-07
pri-mir-191/425	0.48	5.06	3.05	2.66E-14
pri-mir-1949 (Snhg4)	59.74	29.58	-1.05	1.67E-07
pri-mir-196a-1 (Gm53)	9.86	19.51	1.0	2.07E-06
pri-mir-20a/17/19b-1/18a/92a-1/19a (Mir17hg)	29.72	104.64	1.66	1.62E-14
pri-mir-20b/363/92a-2/19b-2/106a/18b (Kis2)	2.90	9.58	1.47	2.28E-08
pri-mir-22 (Mir22hg)	5.51	18.63	1.91	9.35E-13
pri-mir-290a/294/292/291a/291b/295/293 (D7Ertd143e)	12.32	433.40	5.0	2.61E-122
pri-mir-296/298 (Nespas)	23.15	89.14	2.04	3.33E-20
pri-mir-3069 (Snhg10)	3.09	1.89	-0.72	1.18E-01
pri-mir-322/503/351 (C430049B03Rik)	0.91	11.24	3.30	1.39E-19
pri-mir-345	1.06	1.01	0.066	1.0
pri-mir-6516 (2810008D09Rik)	21.55	21.54	0.044	1.0
pri-mir-675 (H19)	2.46	56.69	4.67	1.02E-23

pri-let-7d/7f-1/7a-1	6.96	11.51	0.66	5.36E-03

Fourteen (63.6%) pri-miRNA regions were significantly up-regulated in *Drosha*-null samples, while only 2 (9.1%) were down-regulated. The majority of the experimentally derived pri-miRNA transcripts (14 out of 22) partially (**Fig. 18**) or fully (**Fig. 19**) overlap with previously annotated long non-coding RNA genes (lncRNAs), suggesting incomplete annotation and/or multiple functionality.

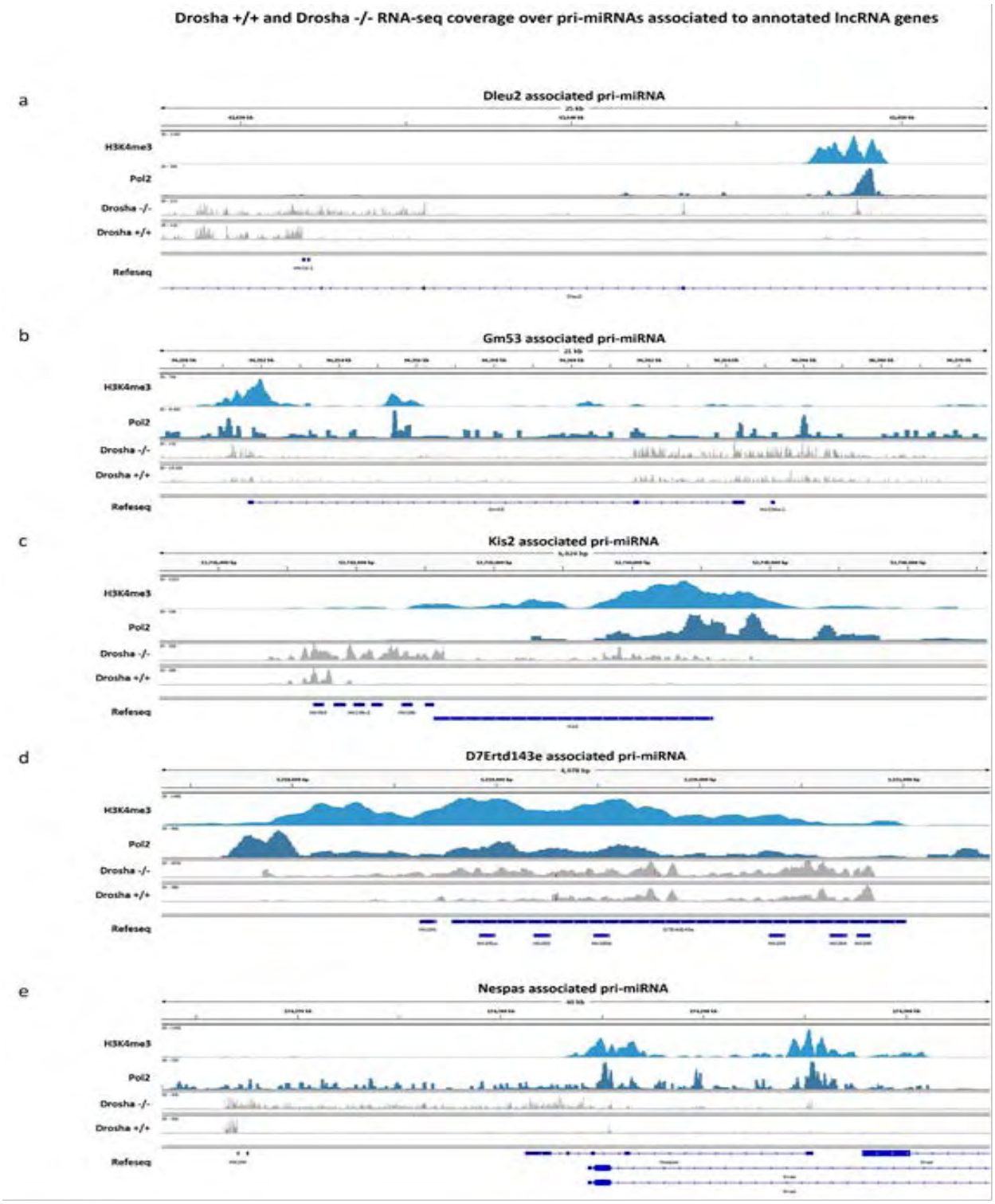


Figure 26. Drosha +/+ and Drosha -/- RNA-seq coverage over pri-miRNAs that partially or fully overlap with annotated lncRNAs. Expression and DE significance values for each pri-miRNA are listed in Supplementary Table 5. a) pri-mir-16-1/15a has been found to be regulated by its own promoter which is located inside the Dleu2 lncRNA gene body. b) mmu-mir-196a-1 is located less than 1kb downstream to Gm53 lncRNA 3' end. The analysis of the Drosha -/- RNA-seq data, however, suggests that pri-mir-196a-1 is part of the Gm53 locus suggesting multiple functionality. c) mmu-mir-196a-1, mmu-mir-20b/363/92a-2/19b-2/106a/18b cluster is also located immediately downstream (and partially overlaps) to Kis2 lncRNA 3' end. The results suggest that pri-mir-20b/363/92a-2/19b-2/106a/18b is transcribed from the Kis2 locus. d) D7Ertd143e lncRNA locus is part of the pri-miRNA

transcribing mmu-mir-290a/294/292/291a/291b/295/293 cluster, which transcription start is located less than 1kb upstream compared to the existing annotation. e) The RNA-seq data analysis suggests that Nespas lncRNA is transcribed from an imprinted locus that also acts as a pri-miRNA. This image has been taken from the relevant publication of microTSS (Georgakilas et al., 2014).

Both down-regulated transcripts overlap with such loci and their expression could be also connected to lncRNA transcript functions. For instance, one of the down-regulated pri-miRNA regions overlaps with *Snhg4* (small nucleolar RNA host gene 4), which also hosts a known snoRNA transcript.

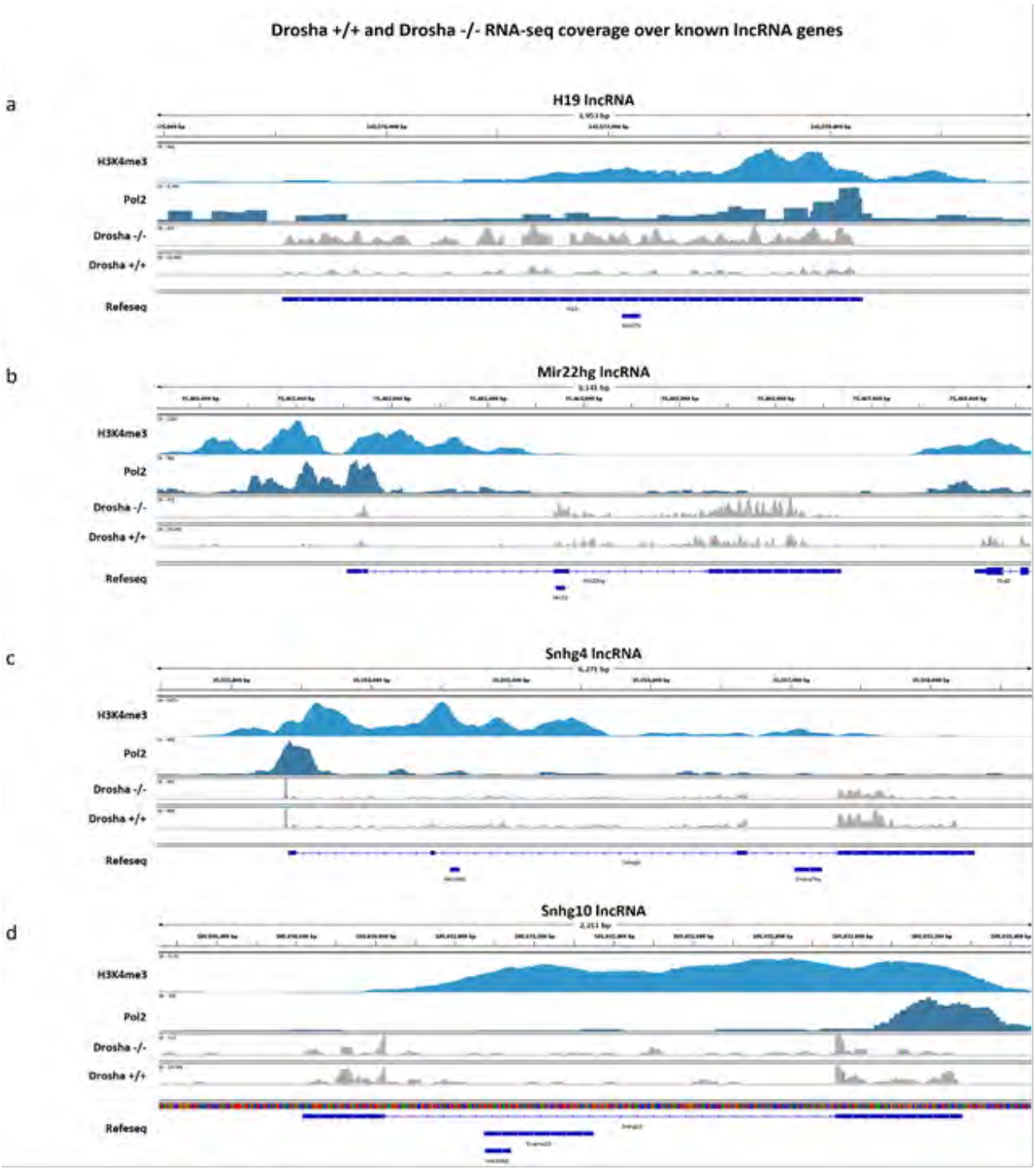


Figure 27. An example of Drosha +/+ and Drosha -/- RNA-seq coverage over annotated lncRNA suggesting multiple functionality. Expression and DE significance values for each pri-miRNA can be found in Supplementary Table 5. H19 a) and Mir22hg b) have been found to be up-regulated in Drosha -/- samples. Snhg4 c) expression levels have been deemed down-regulated while Snhg10 d) expression has been identified as unchanged. The down-regulated miRNA expression levels in the Drosha -/- model could be also connected to other functions of the lncRNA transcripts. This image has been taken from the relevant publication of microTSS (Georgakilas et al., 2014).

2.2.3. Comparison between microTSS and previous methods

In order to construct an extensive validation set of miRNA TSSs in human, GRO-Seq datasets, derived from human IMR90 and ES cell samples (Jin et al., 2013; Sigova et al., 2013), were analyzed (**Table 5**). In contrast to Pol2 ChIP-Seq, GRO-Seq data are strand-specific. They map and quantify only transcriptionally engaged Pol2 (Core et al., 2008). GRO-Seq density sharply peaks near the TSS in sense and anti-sense directions (**Fig. 25**). A sliding window was applied on the region upstream of pre-miRNAs resulting in the identification of loci enriched in GRO-Seq signal. Regions correlated with H3K4me3 and Pol2 ChIP-Seq derived peaks have been marked as TSSs. Precursors presenting no overlap with enriched regions have been filtered out. This pipeline resulted to the identification of TSSs for 72 pre-miRNAs in human ES and 81 pre-miRNAs in IMR90 cells. Human ESC GRO-Seq signal around pri-miRNA TSSs is depicted in **Figure 20c**. These human miRNA TSSs served as two additional independent test sets.

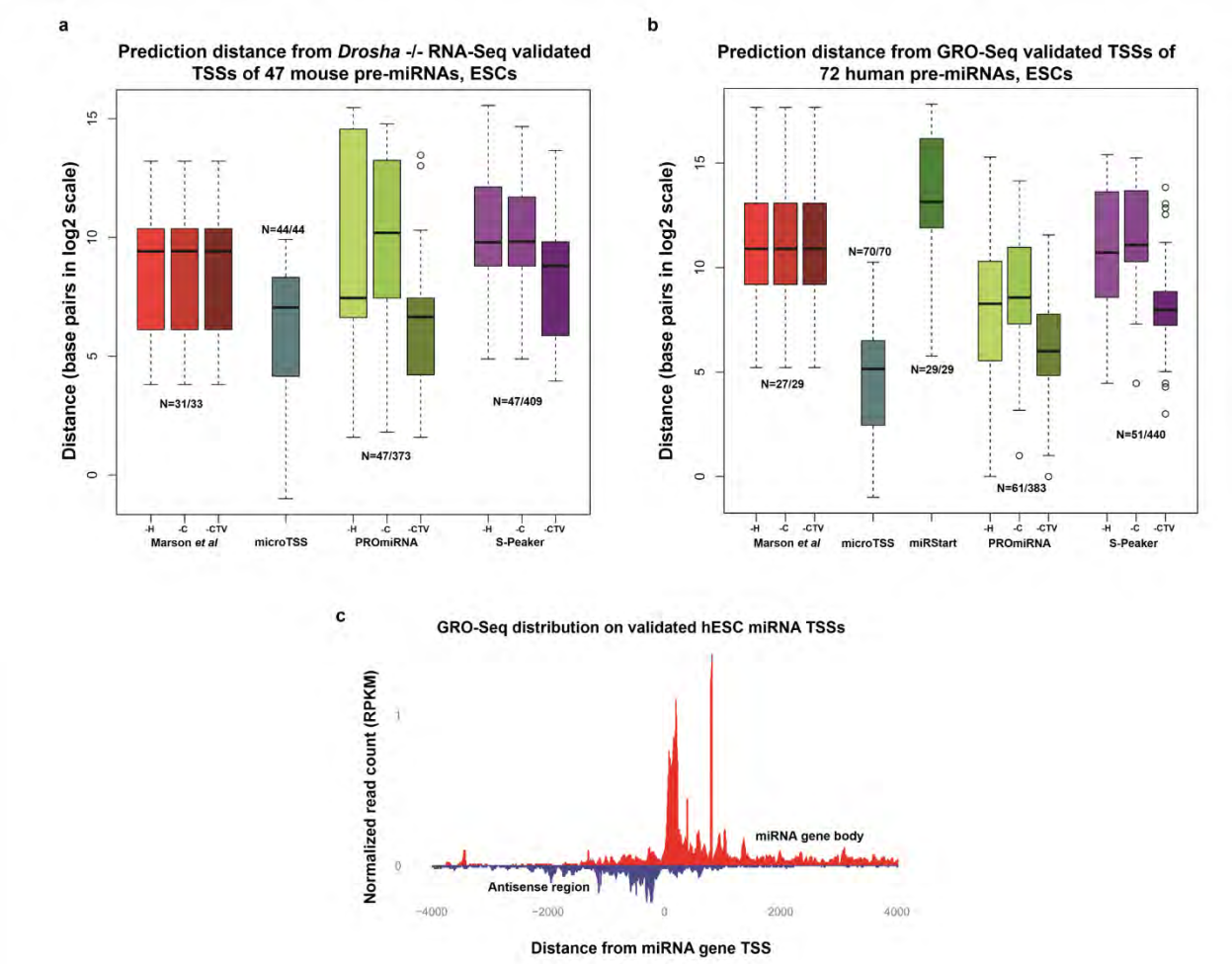


Figure 28. Comparing prediction distance from validated TSSs. PROmiRNA, S-Peaker and Marson et al support multiple predictions per miRNA. The total amount of predicted TSSs is given in X/Y notation to provide a sense of precision for each algorithm. X represents the number of supported miRNAs and Y the total amount of predictions for the supported miRNAs. a) Comparison between the algorithms in terms of prediction distance from Drosha-null validated miRNA TSSs. Distance has been transformed in log2 scale. b) The same comparison methodology based on 72 GRO-Seq derived TSSs in hESCs. c) Signal distribution around the GRO-Seq validated miRNA gene transcription start site in hESCs. This image has been taken from the relevant publication of microTSS (Georgakilas et al., 2014).

By applying microTSS on deeply sequenced NGS data derived from the ENCODE consortium we have identified 70 intergenic miRNA gene TSSs, corresponding to 118 miRNA precursors in mESCs. In hESCs we have identified 63 TSSs corresponding to 86 pre-miRNAs and in IMR90 cells 50 TSSs associated to 82 precursors.

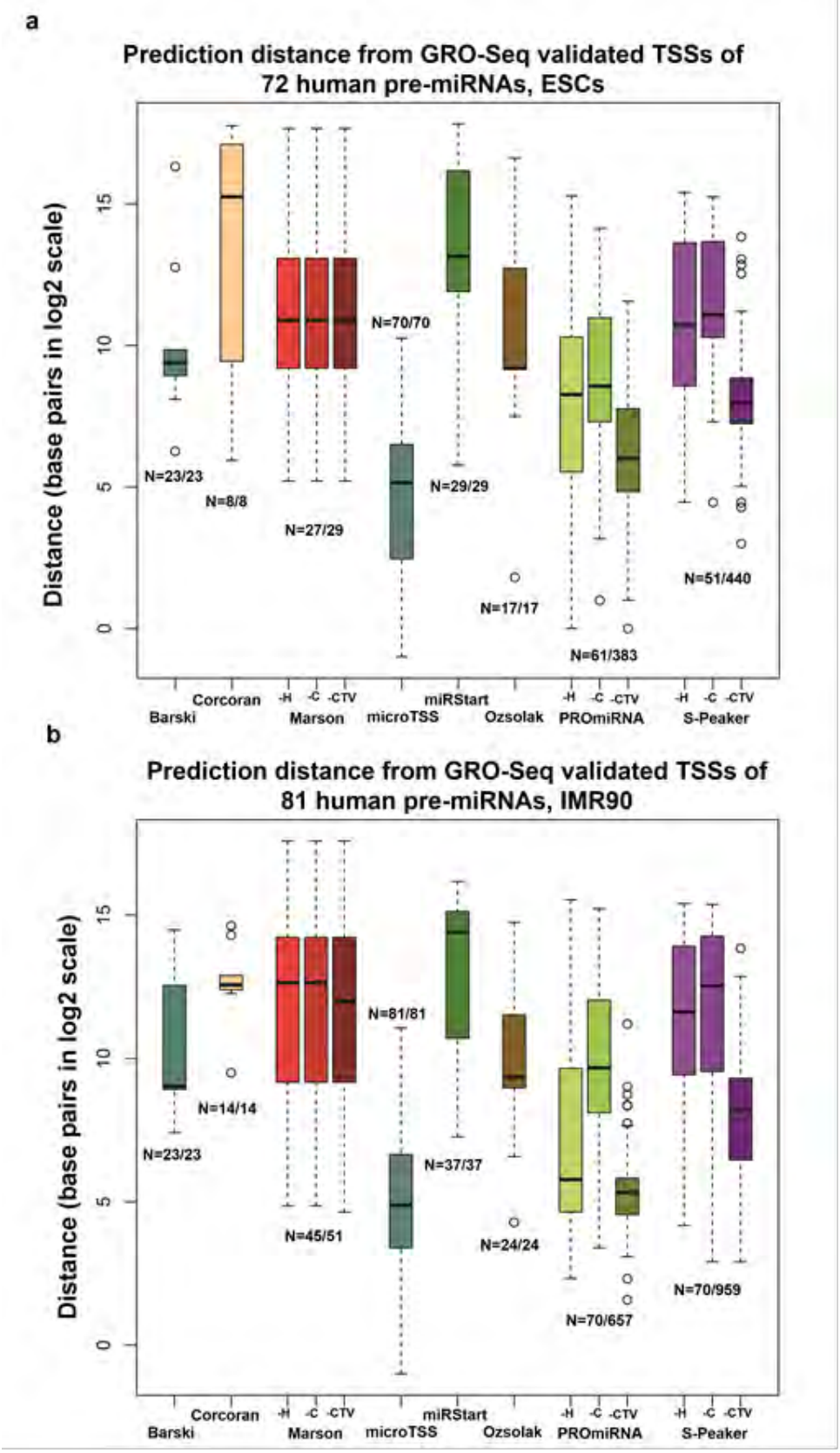


Figure 29. Algorithms' performance in terms of prediction distance from validated TSSs. Distances in y-axis are log2 transformed. The number of supported miRNAs and the total numbers of predictions are shown in N=X/Y notation. X denotes the number of supported miRNAs out of the set of validated precursors. Y denotes the number of total predictions for the supported miRNAs. Marson et al, PROmiRNA and S-Peaker provide multiple predictions per miRNA. For these three algorithms, -H, -C and -CTV correspond to the highest scored prediction, closest to precursor and closest to validated TSS respectively. a) The comparison between the algorithms has been achieved with GRO-Seq validated TSSs of 72 miRNA precursors in human ESCs. b) Additional evaluation of the algorithms' performance has been based on GRO-Seq derived TSSs of 81 pre-miRNAs in human IMR90 cells. This image has been taken from the relevant publication of microTSS (Georgakilas et al., 2014).

From the existing miRNA promoter recognition techniques, only the algorithms introduced by Marson *et al* (Marson et al., 2008), PROmiRNA (Marsico et al., 2013) and S-Peaker (Megraw et al., 2009) support predictions in mouse genome. Since source codes for miRStart (Chien et al., 2011) and Marson *et al* (Marson et al., 2008) algorithms are not available, we have utilized their precompiled predictions. Additionally, we took into account the fact that these algorithms are based on outdated miRBase versions, comprising fewer miRNAs than miRBase v20, which is utilized by microTSS, PROmiRNA and S-Peaker. Therefore, the prediction set of these algorithms has been reduced to the annotation utilized for their implementation.

Some of the algorithms that we have identified offer multiple TSS predictions per miRNA, while others offer a single prediction. In order to perform a robust comparison and account for the fundamental differences between the algorithms in both categories we established two distinct evaluation pipelines.

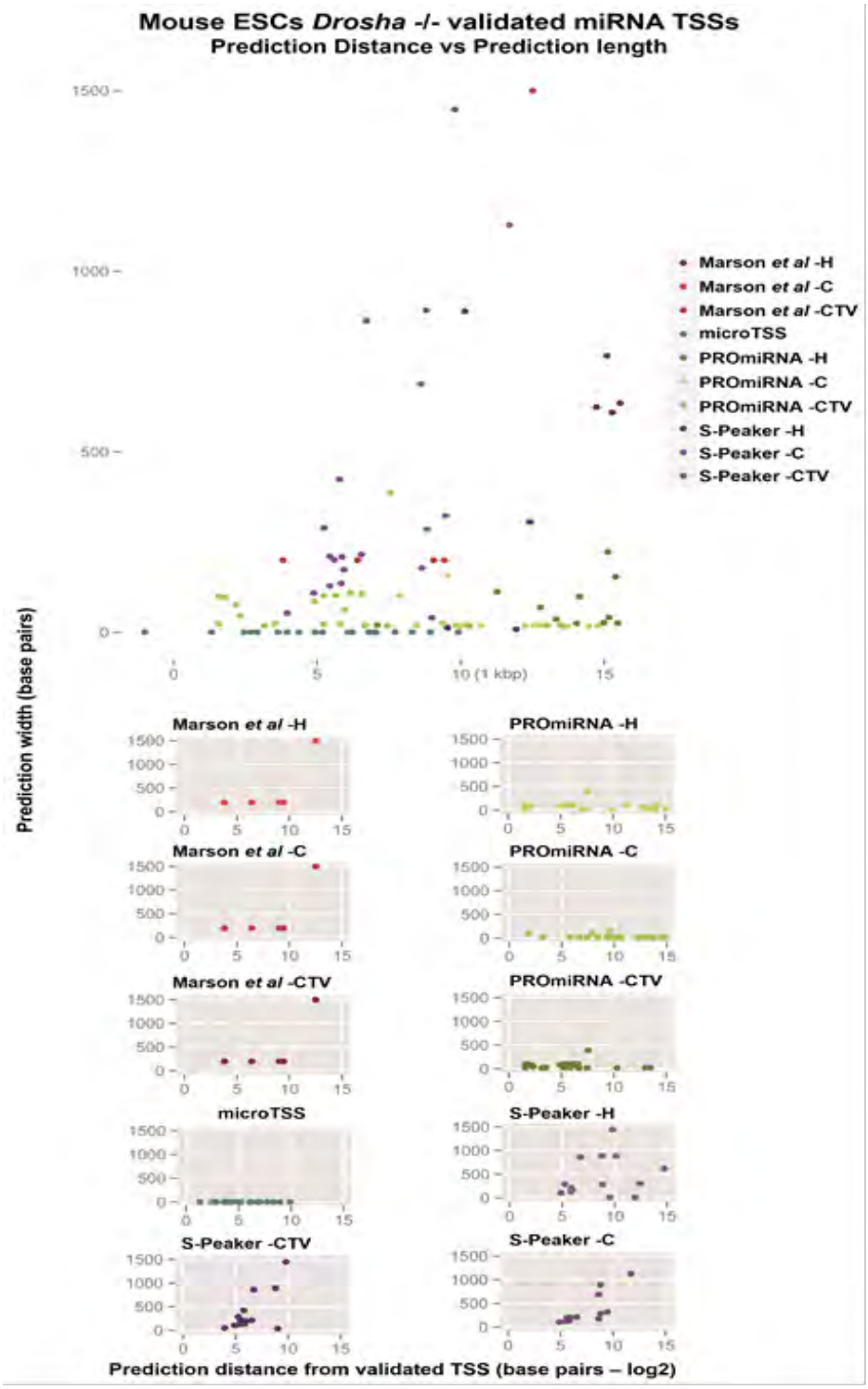


Figure 30. Prediction distance vs width for Drosha -/- validated TSSs of 47 mouse miRNA precursors. X-axis is limited to 15 (log2 scale). The y-axis is limited to 1.5 kb. This image has been taken from the relevant publication of microTSS (Georgakilas et al., 2014).

In the first approach we have selected one prediction per miRNA for each method. For the algorithms in the first category this corresponds to the standard set of supported predictions. On the other hand, for the methods in the second category, three distinct subsets of predictions have been created. The first (denoted with the extension -H) comprises the highest scored TSSs in the region upstream of miRNAs, while the second includes the closest predictions to each precursor (denoted with the extension -C). The last subset contains the closest predictions to the experimentally verified TSS (denoted with the extension –CTV). It should be noted that the last set (-CTV) requires a priori knowledge of the true TSS in order to be defined and can be applied from the user if all predictions per miRNA are taken into account.

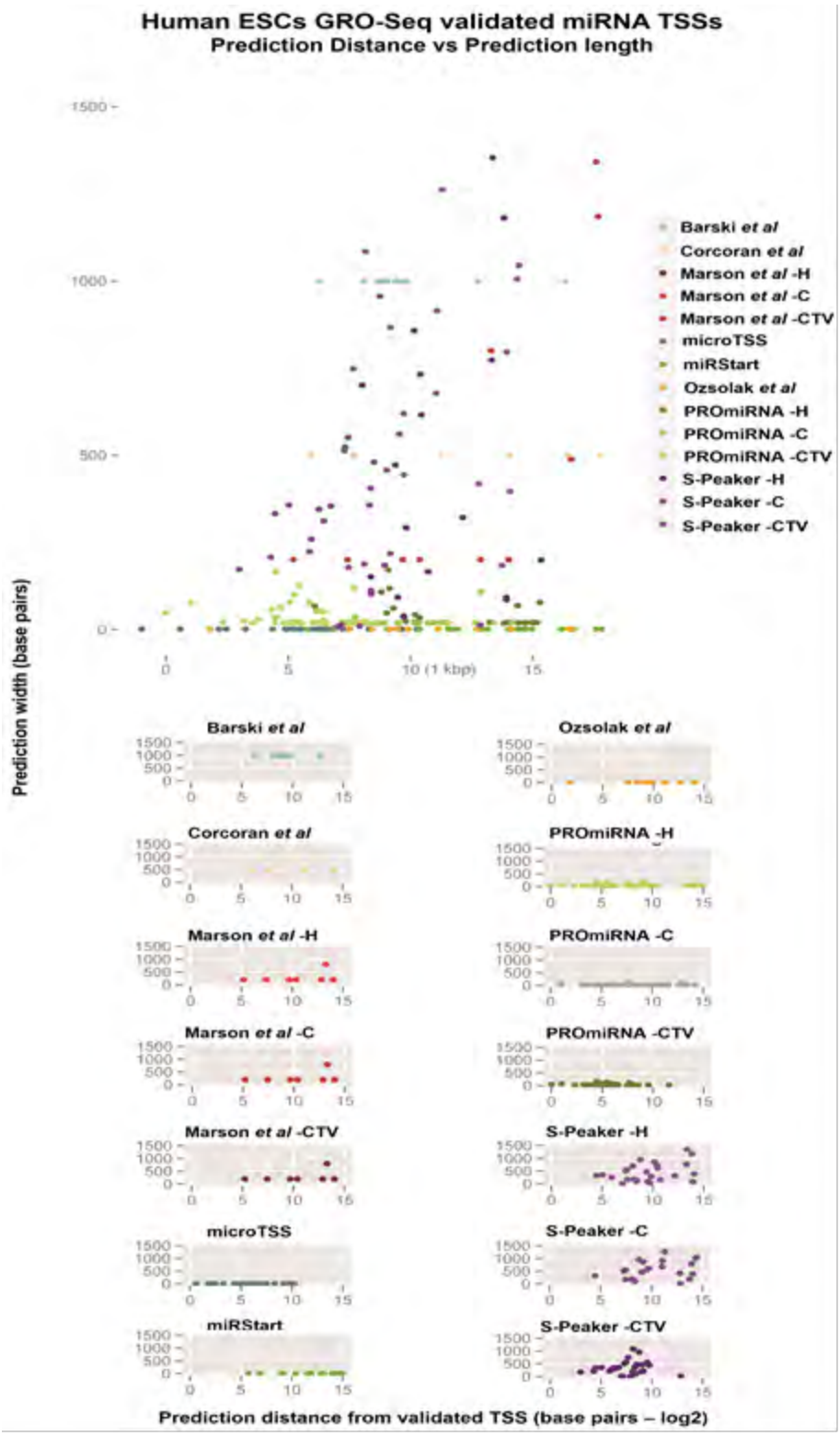


Figure 31. Comparing prediction distance vs width for GRO-Seq validated TSSs of 72 human ESCs miRNA precursors. Prediction distance in x-axis is limited to 15 (log2 scale). The y-axis is limited to 1.5 kbp. This image has been taken from the relevant publication of microTSS (Georgakilas et al., 2014).

The distance of all predictions relative to the corresponding validated TSSs has been calculated and the number of all predictions is also noted (**Fig. 20a,b**, **Fig. 21-24**), including descriptive and inferential statistics (**Tables 9-11**).

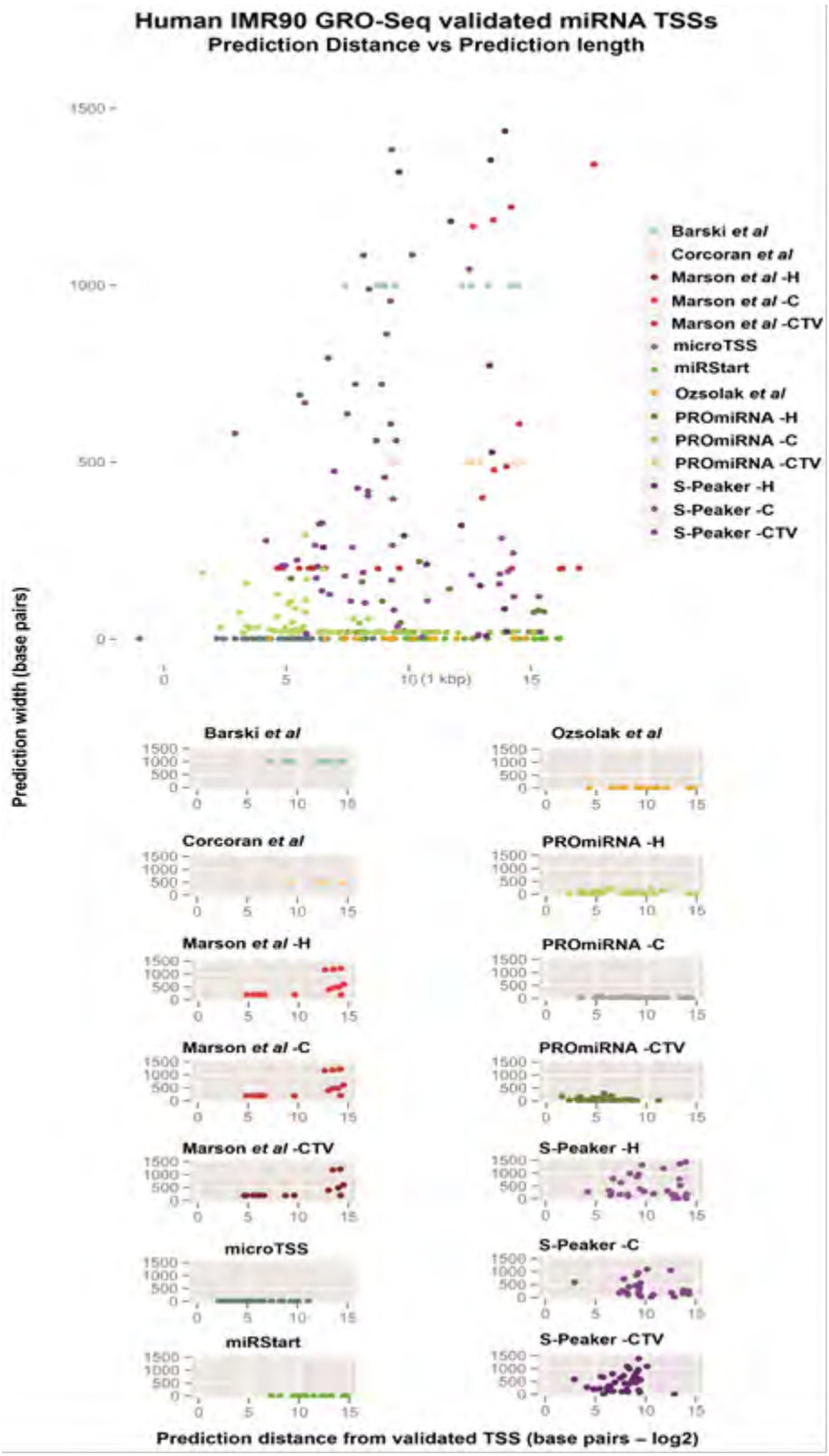


Figure 32. Comparing prediction distance vs width for GRO-Seq validated TSSs of 81 human IMR90 miRNA precursors. Prediction distance in x-axis is limited to 15 (log2 scale). The y-axis is limited to 1.5 kbp. This image has been taken from the relevant publication of microTSS (Georgakilas et al., 2014).

It can be observed that microTSS performs significantly better than all the other programs of the same category exhibiting median distance, between the predicted and validated TSS, smaller than 35nts in human and 130nts in mouse.

Table 9. Algorithms' performance in terms of prediction distance from Drosha -/- RNA-Seq validated miRNA TSSs in mouse. Marson et al utilizes an older miRBase version, resulting in the smallest sample size. PROmiRNA-H, Marson et al –H and S-Peaker-H refers to to the highest-score prediction. PROmiRNA-C, S-Peaker-C and Marson et al –C corresponds to each precursor's closest predicted TSS. PROmiRNA-CTV, S-Peaker-CTV and Marson et al –CTV corresponds to each precursor's predicted TSS closest to the experimentally verified TSS. The second part of the table includes statistical significance levels of post-hoc pairwise statistical comparisons (FDR corrected). Statistically significant differences are marked with blue. This table has been taken from the relevant publication of microTSS (Georgakilas et al., 2014).

Prediction distance statistics on Drosha -/- RNA-Seq derived TSSs of 47 pre-miRNAs (bp) in mESCs										
	Marson *et al* (N=31)			**microTSS (N=44)**	**PROmiRNA (N=47)**			**S-Peaker (N=47)**		
	-C	**-H**	**-CTV**		**-C**	**-H**	**-CTV**	**-C**	**-H**	**-CTV**
Median	683	683	683	128	1,168	174	100	897	885	442
IQR	1,245	1,245	1,245	299	9,500	23,889	156	2,884	4,009	839

Statistical significance									
	Marson *et al*			**PROmiRNA**			**S-Peaker**		
	-C	**-CTV**	**-H**	**-C**	-CTV	**-H**	-C	**-H**	-CTV
Marson *et al* –CTV	1	-	-	-	-	-	-	-	-
Marson *et al* –H	1	1	-	-	-	-	-	-	-
PROmiRNA -C	0.088	0.088	0.088	-	-	-	-	-	-
PROmiRNA -CTV	0.0329	0.0329	0.0329	0.00011	-	-	-	-	-
PROmiRNA -H	0.64	0.64	0.64	0.75	0.01	-	-	-	-
S-Peaker -C	0.21	0.21	0.21	0.53	0.00045	0.8	-	-	-
S-Peaker -H	0.164	0.164	0.164	0.8	0.0001	0.61	0.96	-	-
S-Peaker –CTV	0.58	0.58	0.58	0.00345	0.088	0.164	0.0095	0.01503	-
microTSS	0.0035	0.0035	0.0035	0.000017	0.65	0.00618	0.0000083	0.000017	0.088

MicroTSS outperforms the –C and –H sets of the programs in the second category (where one prediction per miRNA has been selected) and is comparable to the –CTV set where the closest predictions to the validated TSSs (out of several predictions for each miRNA) has been used. In the second evaluation pipeline, the predictions provided by the algorithms have been utilized in order to measure their sensitivity and precision. To this end, we have applied a threshold of 1,000 bp on the prediction distance from validated TSSs. Predictions located closer than 1 kbp from the validated TSS are considered True Positives (TP) and the rest are treated as False Positives (FP). Precision has been calculated as the number of TPs divided by the number of total predictions (TPs + FPs).

Table 10. Comparing prediction distance between microTSS and the available algorithms on GRO-Seq validated miRNA TSSs in human ESCs. The differences in each algorithm's sample size originate from older miRBase versions and/or from the fact that different methodologies utilize different search space upstream of pre-miRNAs. PROmiRNA-H, Marson et al –H and S-Peaker-H refers to the highest-score prediction. PROmiRNA-C, S-Peaker-C and Marson et al –C corresponds to each precursor's closest predicted TSS. PROmiRNA-CTV, S-Peaker-CTV and Marson et al –CTV corresponds to each precursor's predicted TSS closest to the experimentally verified TSS. The second part of the table includes statistical significance levels of post-hoc pairwise comparisons (FDR corrected). Statistically significant differences are marked with blue. This table has been taken from the relevant publication of microTSS (Georgakilas et al., 2014).

Prediction distance (bp) statistics on GRO-Seq derived TSSs of 72 pre-miRNAs in hESCs														
	Barski *et al* (N=21)	**Corcoran *et al* (N=8)**	**Marson *et al* (N=27)**			**microTSS (N=70)**	**miRStart (N=29)**	**Ozsolak *et al* (N=17)**	**PROmiRNA (N=61)**			**S-Peaker (N=51)**		
			-C	**-H**	**-CTV**				**-C**	**-H**	**-CTV**	**-C**	**-H**	**-CTV**
Median	670	38,967	1,910	1,910	1,910	35	9,026	584	377	306	64	2,164	1,675	249
IQR	435	110,478	7,608	7,608	7,608	84	69,910	6,209	1,848	1,131	188	11,791	12,036	308

Statistical significance													
	Barski *et al*	**Corcoran *et al***	**Marson *et al***			**miRStart**	**Ozsolak *et al***	**PROmiRNA**			**S-Peaker**		
			-C	**-CTV**	**-H**			**-C**	-CTV	**-H**	-C	**-H**	-CTV
Corcoran *et al*	0.09	-	-	-	-	-	-	-	-	-	-	-	-
Marson *et al* -C	0.06	0.20	-	-	-	-	-	-	-	-	-	-	-
Marson *et al* –CTV	0.06	0.20	1	-	-	-	-	-	-	-	-	-	-
Marson *et al* -H	0.06	0.20	1	1	-	-	-	-	-	-	-	-	-
miRStart	0.0003	0.79	0.02	0.02	0.02	-	-	-	-	-	-	-	-
Ozsolak *et al*	0.39	0.25	0.46	0.46	0.46	0.02	-	-	-	-	-	-	-
PROmiRNA -C	0.26	0.02	0.01	0.01	0.01	6.5E-08	0.11	-	-	-	-	-	-
PROmiRNA -CTV	2.3E-08	0.0007	2.5E-09	2.5E-09	2.5E-09	1.8E-11	3.6E-06	0.0000005	-	-	-	-	-

PROmiRNA -H	0.12	0.02	0.005	0.005	0.005	7.1E-06	0.08	0.78	0.00008	-	-	-	-
S-Peaker -C	0.007	0.13	0.31	0.31	0.31	0.01	0.28	0.00001	2.6E-14	0.0003	-	-	-
S-Peaker -H	0.14	0.07	0.78	0.78	0.78	0.00232	0.75	0.01	1.8E-11	0.007	0.19	-	-
S-Peaker -CTV	0.0009	0.01	0.00002	0.00002	0.00002	1.8E-08	0.001	0.17	0.000007	0.36	1.10E-08	1.20E-05	-
microTSS	7E-09	0.0001	1E-10	1E-10	1E-10	2E-12	3.9E-07	2.7E-11	0.01	2.00E-08	<2E-16	1.30E-14	2.40E-10

Sensitivity is defined as the number of TPs divided by the number of positives (supported miRNAs from the validation set). Marson *et al.* achieves 54% and 64.5% in mESCs, 15.2% and 40.7% in hESCs, 18.5% and 29.4% in IMR90 sensitivity and precision respectively. miRStart on the other hand, achieves 5.5%/4.9% and 13.7%/10.8% sensitivity and precision in hES/IMR90 cells. microTSS significantly outperforms the algorithms of the same category by exhibiting 93.6% and 100% in mESCs, 94.4% and 97.1% in hESCs, 91.3% and 91.3% in IMR90 sensitivity and precision respectively. The algorithms of the second category that provide multiple TSS predictions per miRNA possibly active in different cell types/tissues (i.e. PROmiRNA and S-Peaker) have been excluded from this evaluation pipeline since the evaluation sets consist only from promoters specifically active in the investigated cell lines.

Table 11. Comparing prediction distance between microTSS and the available algorithms on GRO-Seq validated miRNA TSSs in human IMR90 cells. The differences in each algorithm's sample size originate from older miRBase versions and/or from the fact that different methodologies utilize different search space upstream of pre-miRNAs. PROmiRNA-H, Marson et al –H and S-Peaker-H refers to the highest-score prediction. PROmiRNA-C, S-Peaker-C and Marson et al –C corresponds to each precursor's closest predicted TSS. PROmiRNA-CTV, S-Peaker-CTV and Marson et al –CTV corresponds to each precursor's predicted TSS closest to the experimentally verified TSS. The second part of the table includes statistical significance levels of post-hoc pairwise comparisons (FDR corrected). Statistically significant differences are marked with blue. This table has been taken from the relevant publication of microTSS (Georgakilas et al., 2014).

Prediction distance (bp) statistics on GRO-Seq derived TSSs of 81 pre-miRNAs in IMR90 cells														
	Barski *et al* (N=23)	**Corcoran *et al* (N=14)**	**Marson *et al* (N=45)**			**microTSS (N=81)**	**miRStart (N=37)**	**Ozsolak *et al* (N=24)**	**PROmiRNA (N=71)**			**S-Peaker (N=71)**		
			-C	**-H**	**-CTV**				**-C**	**-H**	**-CTV**	**-C**	**-H**	**-CTV**
Median	522	6,038	4,039	4,039	4,039	29	21,618	643	814	54	40	5,873	3,125	290
IQR	5,512	2,223	18,506	18,506	18,506	90	33,947	2,020	3,562	778	33	18,137	13,979	543
Statistical significance														
			Marson *et al*			**miRStart**		**PROmiRNA**				**S-Peaker**		

	Barski *et al*	**Corcoran *et al***	**-C**	**-CTV**	**-H**		**Ozsolak *et al***	**-C**	-CTV	**-H**	-C	**-H**	-CTV
Corcoran *et al*	0.01	-	-	-	-	-	-	-	-	-	-	-	-
Marson *et al* -C	0.07	0.62	-	-	-	-	-	-	-	-	-	-	-
Marson *et al* –CTV	0.19	0.27	0.78	-	-	-	-	-	-	-	-	-	-
Marson *et al* -H	0.07	0.62	1	0.78	-	-	-	-	-	-	-	-	-
miRStart	0.0002	0.20	0.09	0.04	0.09	-	-	-	-	-	-	-	-
Ozsolak *et al*	0.98	0.0036	0.07	0.14	0.07	0.00011	-	-	-	-	-	-	-
PROmiRNA -C	0.54	0.001	0.008	0.03	0.008	1.10E-06	0.78	-	-	-	-	-	-
PROmiRNA -CTV	4.80E-11	1.50E-08	9.60E-14	5.70E-13	9.60E-14	1.60E-15	1.00E-09	1.90E-15	-	-	-	-	-
PROmiRNA -H	9.10E-05	4.70E-06	3.70E-08	2.20E-07	3.70E-08	7.40E-11	0.0004	4.20E-06	0.007	-	-	-	-
S-Peaker -C	0.02	0.54	1	0.66	1	0.009	0.01	0.0003	< 2e-16	1.40E-11	-	-	-
S-Peaker -H	0.17	0.16	0.42	0.67	0.42	0.0009	0.10	0.016	< 2e-16	7.70E-10	0.28	-	-
S-Peaker –CTV--	9.00E-05	1.10E-07	1.70E-07	2.60E-06	1.70E-07	7.20E-13	0.001	0.0001	8.00E-12	0.013	1.40E-13	7.40E-11	-
microTSS	2.30E-09	1.50E-08	9.60E-14	5.70E-13	9.60E-14	1.90E-15	2.30E-08	7.50E-14	0.65	0.0014	< 2e-16	< 2e-16	1.10E-09

These results depict the fundamental differences between the methodologies of the two categories. Algorithms such as PROmiRNA and S-Peaker provide high quality predictions close to the validated TSS (-CTV results) but are often lost within numerous predictions since in most cases they are not highly scored. This results to increased False Positive Rate due to the high number of predictions per miRNA decreasing prediction precision. On the other hand, microTSS addresses this issue by utilizing expression data from the investigated cell line or tissue, providing single predictions per miRNA close to the validated TSS and provides a "snapshot" of the currently active promoters. The unique combination of high precision and sensitivity provided by microTSS enables the study of miRNA regulation and their complete integration in cell line/tissue specific regulatory networks.

2.2.4. Effects of sequencing depth and RNA-Seq coverage threshold on microTSS performance

Sequencing depth and the algorithm's sliding window threshold of RNA-Seq coverage are key parameters in microTSS performance. In order to assess their effects on the algorithm's outcome, we have performed two distinct tests. In the first test (**Fig. 16a**), random subsampling has been applied on the WT mESC RNA-Seq data (GSM973235) resulting in four subsets of 20%, 40%, 60% and 80% of the initial dataset's depth (2 X 125M uniquely mapped, strand-specific, paired-end reads). The performance of microTSS on each subset has been evaluated using the set of *Drosha* -/- validated TSSs. The analysis suggests that even at lower sequencing depths (e.g. 2 X 25M uniquely mapped reads), microTSS is able to accurately identify TSSs corresponding to the most abundant pri-miRNAs and expressed precursors, i.e. miRNA transcripts with low degradation rate. Gradual increments in the sequencing depth enable microTSS to capture pri-miRNAs and precursors of lower abundance and expression rate, respectively.

In the second test (**Fig. 16b**), microTSS has been applied on the same WT mESCs RNA-Seq dataset (GSM973235) by utilizing four different thresholds for the RNA-Seq coverage. The threshold of 5 reads, which is the default, is able to identify TSSs of pri-miRNAs with high degradation rate without compromising the prediction accuracy. The algorithm is less sensitive, at the same levels of precision, as the threshold increases.

2.2.5. Polycistronic pri-miRNAs and coverage of annotated lncRNAs

The analysis of microTSS predictions revealed that 37.1% of TSSs in mESCs (26 out of 70), 19% in human ESCs (12 out of 63) and 30% (15 out of 50) IMR90 cells are associated with multiple pre-miRNAs. 40.6% of miRNAs in hESCs (35 out of 86), 57.3% in IMR90 cells (47 out of 82) and 62% in mECSs (74 out of 118) are derived from polycistronic miRNA gene clusters. Moreover, 28% of TSSs in mESCs (20 out of 70), 25.3% in hESCs (16 out of 63) and 44% in IMR90 (22 out of 50) correspond to pri-miRNAs that partially or fully overlap with already annotated lncRNA genes. For example, our findings regarding mouse pri-mir-675 are in agreement with previous studies (Kallen et al., 2013; Monnier et al., 2013) showing that it fully overlaps with H19 lncRNA gene, which has been found to control several genes within the imprinted gene network. H19 recruits MBD1 and forms a lncRNA:protein complex which interacts with histone lysine methyltransferases and represses its target genes (Monnier et al., 2013). A recent study has also revealed that H19 hosts both canonical and non-canonical binding sites for the let-7 family, thus acting as a molecular sponge (Kallen et al., 2013). Another example of incomplete annotation is Mir17hg which has been classified as a small RNA host transcript (Clark et al., 2012). The analysis of microTSS predictions shows that Mir17hg is a polycistronic miRNA gene cluster, hosting 6 precursors (mir-20a, mir-17, mir-19b-1, mir-18a, mir-92a-1 and mir19a)

whose identified TSS is located several hundred base pairs upstream of the current annotation.

The analysis of small RNA-Seq data in mESCs, presented in the **Supplementary Data** of microTSS publication (Georgakilas et al., 2014), revealed different patterns of pre-miRNA expression in polycistronic miRNA genes. There are cases where all members of the same cluster share similar expression levels. MiR-365-1 and miR-193b are transcribed from the same pri-miRNA exhibiting very low reads per kilo-base per million mapped reads (RPKM) values. In other cases, co-clustered miRNAs present significantly different expression levels. *D7ertd143e* polycistronic miRNA locus hosts miR-292, miR-291a, miR-295, miR-293 and miR-294 located in the top 64 expressed pre-miRNAs in mESCs, while the remaining two precursors of the cluster, miR-290a and miR-291b exhibit significantly lower expression levels. These results are in agreement with previous studies suggesting that there are post-transcriptional mechanisms responsible for blocking individual members of polycistronic miRNA genes from the maturation process. In a recent study (Chawla & Sokol, 2014), adenosine deaminases acting on RNAs (*ADARs*) were shown to alter the structural conformation of let-7 polycistronic pri-miRNA transcript, resulting in limited Drosha processing for individual members of the cluster and enhanced processing for others.

2.2.6. Divergent antisense pri-miRNAs identified with GRO-Seq

Several recent studies have shown that the majority of mammalian promoters initiate transcription on both sense and antisense directions, a phenomenon known as divergent transcription (Seila et al., 2008). Divergent transcription generates upstream antisense RNAs near the 5′ end of genes that are typically short (50–2,000 nucleotides) and in many cases unstable (X. Wu & Sharp, 2013). These results suggest that the common phenomenon of divergent transcription of active promoters may help promoter regions to maintain a state poised for subsequent regulation and has been proposed as a model for new gene formation. In mouse and human ESCs divergent transcription from promoter and enhancer regions of protein-coding genes is the major source of intergenic transcription.

Table 12. Precursor miRNAs derived from upstream antisense pri-miRNAs as identified by analyzing GRO-Seq datasets in mESCs. Increasing Siphy score corresponds to fast evolving, thus less conserved sequences. This table has been taken from the relevant publication of microTSS (Georgakilas et al., 2014).

Mouse divergent pre-miRNAs		
Precursor	**Protein Coding**	**Siphy Score**
mmu-mir-3569	Kmt2b	1.0065
mmu-mir-320	Polr3d	0.2279

mmu-mir-7666	Micu3	1.9528
mmu-mir-1934	Mpdu1	1.5571
mmu-mir-8102	Pip4k2b	0.7794
mmu-mir-219c	Ring1	0.3400
mmu-mir-92b	Muc1	0.1542
mmu-mir-5627	Dnajb9	1.2490
mmu-mir-5135	Plekhg3	1.2044
mmu-mir-345	Slc25a29	0.4189
mmu-mir-5615-1	Map2k2	0.9442

The analysis of microTSS predictions based on their distance from protein coding genes revealed a significant number of precursors residing very close to coding loci. We subsequently performed spatial classification of all pre-miRNAs in miRBase identifying 13 (1.1%) putative divergent miRNAs in mouse and 43 (2.3%) in human, based on the distance to their corresponding protein coding gene. In order to validate that these pri-miRNAs are indeed transcribed divergently upstream from active protein coding gene promoters, we analyzed mouse and human ESC GRO-Seq data. Eleven out of 13 (84.6%) mouse divergent miRNAs TSSs (**Table 12**) and 26 out of 43 (60.4%) human (**Table 13**), exhibit divergent GRO-Seq signals 2-3kb upstream of the closest protein coding gene, fully overlapping with expressed regions of these miRNA precursors (**Fig. 25**). Six out of 11 (54.5%) mouse and 11 out of 26 (42.3%) human GRO-Seq verified divergent pri-miRNAs have also been identified using microTSS algorithm and deep ESC RNA-Seq data, further supporting our initial hypothesis. miRNA precursors from such loci are significantly less conserved, consistent with the recently proposed model of new gene formation (X. Wu & Sharp, 2013). Relevant graphs and descriptive as well as inferential statistics are presented in **Figure 17**, **Table 14**.

Table 13. GRO-Seq analysis in hESCs revealed 26 precursor miRNAs derived from upstream antisense pri-miRNAs. Increasing Siphy score corresponds to fast evolving, thus less conserved sequences. This table has been taken from the relevant publication of microTSS (Georgakilas et al., 2014).

Human divergent pre-miRNAs		
Precursor	**Protein Coding**	**Siphy Score**
hsa-mir-1289-1	Cep250	1.0414
hsa-mir-345	Slc25a29	0.3925

hsa-mir-3928	Rnf185	1.1882
hsa-mir-3188	Jund	1.2983
hsa-mir-4754	Rps5	1.4129
hsa-mir-320a	Polr3d	0.1828
hsa-mir-4470	Asph	0.6532
hsa-mir-378d-2	Pdp1	1.1392
hsa-mir-3124	Sh3bp5l	1.7301
hsa-mir-548al	Pgm2l1	0.8040
hsa-mir-3166	Rab38	1.1970
hsa-mir-3939	Fgfr1op	1.7221
hsa-mir-4522	Wsb1	1.8283
hsa-mir-4733	Nf1	0.9516
hsa-mir-4727	Cwc25	0.8872
hsa-mir-3678	Grb2	1.0872
hsa-mir-1538	Nfat5	0.3414
hsa-mir-4795	Chmp2b	0.9491
hsa-mir-5188	Ubc	1.4889
hsa-mir-5091	Bod1l1	1.5594
hsa-mir-5696	Rnf149	1.1647
hsa-mir-1302-11	Wash1	1.1395
hsa-mir-548aw	Tsc1	1.6339
hsa-mir-4482	Gsto2	1.0264
hsa-mir-3912	Npm1	0.9785
hsa-mir-4638	Trim41	1.7165

The analysis of precursor miRNAs in mESCs, presented in the Supplementary Material in (Georgakilas et al., 2014) revealed that out of the eleven GRO-Seq validated divergent pre-miRNAs, only mir-320 was highly expressed in the small RNA-Seq sample. Mir-1934, mir219c and mir-345 have been found to exhibit very low expression levels and the rest have not been detected at all. These four precursors correspond to only 8 out of 24 mature divergent miRNA candidates.

Mir-320 and mir-345 are highly conserved and divergently transcribed from *Polr3d* and *Slc25a29*, respectively (**Tables 11** and **12**). Out of the expressed divergent precursors in this cell line only these two miRNAs were identified to interact with coding genes in TarBase (Vergoulis et al., 2012; Vlachos, Paraskevopoulou, et al., 2015), an extensive database of experimentally supported miRNA:gene interactions. In fact, mir-320 targets the same gene (*Hspb6*), among many others, in both species as well as its adjacent gene (*Polr3d*) in human. Mir-345 has been found in the same database to interact only with 3 genes in human. At the same time the in silico analysis of divergently transcribed miRNAs with microT –CDS (Paraskevopoulou, Georgakilas, Kostoulas, Vlachos, et al., 2013) provide a significant number of targets for all miRNAs.

Table 14. Siphy omega values are utilized to measure evolutionary ratings for spatially classified miRNA precursors. The second part of the table includes statistical significance levels of post-hoc pairwise comparisons (FDR corrected). Statistically significant differences are marked with blue. This table has been taken from the relevant publication of microTSS (Georgakilas et al., 2014).

Human pre-miRNA conservation (Siphy omega value) based on spatial classification					
Category		**Median**		**IQR**	
Read-through (N=66)		0.63		1.06	
Antisense (N=175)		0.99		0.79	
Divergent (N=43)		1.16		0.49	
Intergenic (N=670)		1.01		0.87	
Intronic (N=808)		1.07		0.48	
Exonic (N=108)		0.81		0.56	
Statistical significance					
	Read-through	**Exonic**	**Antisense**	**Intergenic**	**Intronic**
Exonic	0.65	-	-	-	-
Antisense	0.15	0.033	-	-	-
Intergenic	0.0496	0.0029	0.38	-	-
Intronic	0.0038	< 10-6	0.006	0.006	-
Divergent	0.005	< 10-4	0.006	0.012	0.083

The small RNA seq analysis further revealed that out of the 99 mature miRNAs located in expressed intergenic transcripts, close to 20% (20 mature miRNAs) reside in

divergently expressed loci; while in the small-RNA-Seq dataset, they correspond to a ~2% fraction (8 out of 411 expressed mature intergenic miRNAs). This is an indication that these divergently transcribed miRNAs are more often repressed on a later stage of miRNA biogenesis, exhibiting a significantly smaller transcription vs expression ratio, as identified with GRO-Seq and small-RNA-Seq, respectively ($p<10^{-12}$).

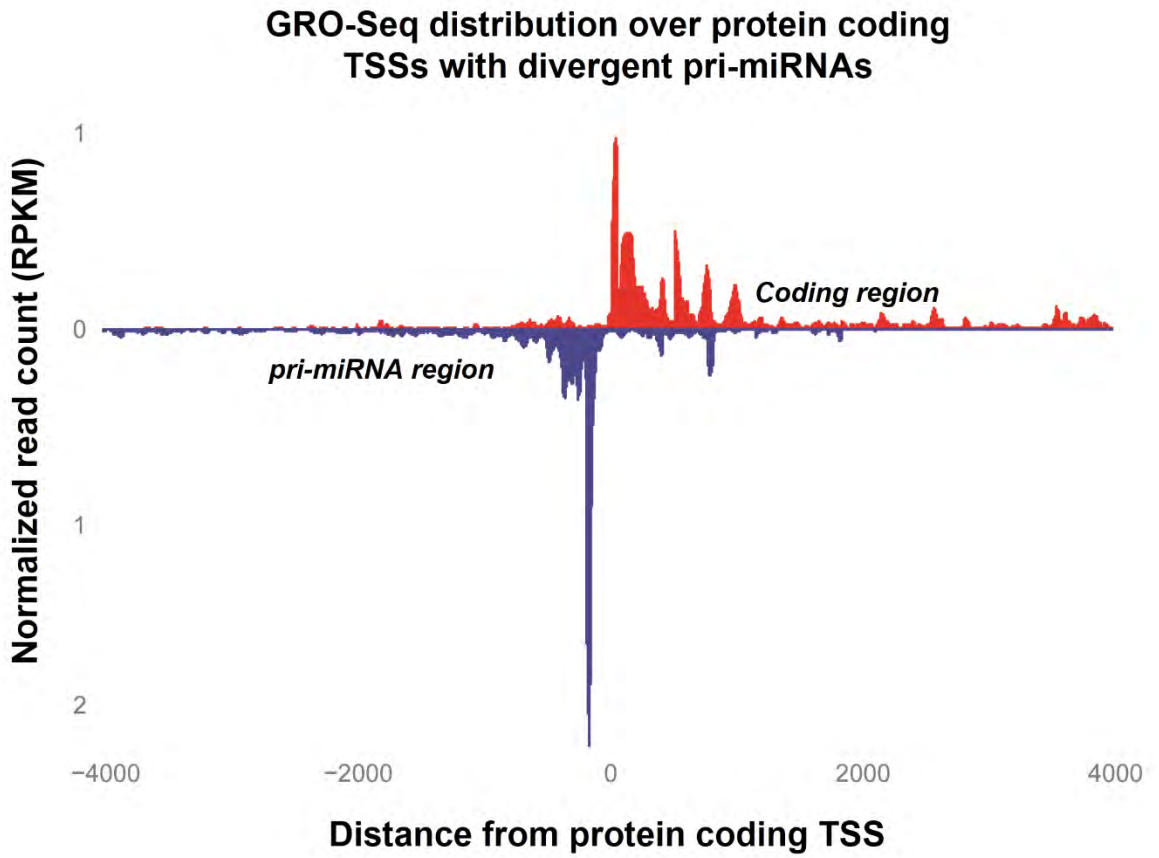


Figure 33. GRO-Seq distribution around protein coding TSSs with divergent pri-miRNAs supporting the hypothesis that divergent transcription might play an additional role in the cell by generating mature miRNAs. All identified precursor miRNAs are transcribed by the pri-miRNA region that exhibits a clear divergent transcription profile, since it fully overlaps with the GRO-Seq signal which dissipates 2 kb upstream of coding TSSs. This image has been taken from the relevant publication of microTSS (Georgakilas et al., 2014).

We have identified more miRNAs expressed using the small-RNA-Seq dataset than trancribed, as identified by the GRO-Seq data. Even though both datasets have comparable numbers of reads, in our opinion the underlying cause is small-RNA-Seq's strict size fractionation, which restricts the available sequencing depth to miRNA-specific RNA lengths.

3. Conclusions - Discussion

Since the discovery of the abundant transcription of miRNAs in 2001 (Lau et al., 2001; R. C. Lee & Ambros, 2001) there has been an explosion of miRNA-related publications which are estimated to exceed 38,000 (May 2015). These studies can be divided in three distinct categories which depending on the research framework can sometimes overlap; studies focusing on i) miRNA biogenesis, ii) miRNA function and iii) miRNA implications in physiological and pathological conditions.

Initial studies were focused in characterizing the mechanisms involved in miRNA biogenesis and function. The miRNA biogenesis pathway has been extensively characterized (Y. Lee et al., 2003; Y. Lee et al., 2004), and the mechanism that mediates miRNA-associated gene expression regulation has been vigorously studied unveiling multiple types of miRNA:mRNA binding (Brennecke et al., 2005; Doench & Sharp, 2004; Jones-Rhoades et al., 2006). Due to lack of experimental as well as computational frameworks, studies focusing on the identification of miRNA gene transcription regulation have only started to emerge six years after the discovery of the abundant transcription of miRNAs in 2001, however they have been unsuccessful in providing an accurate and high resolution map of pri-miRNAs. More contemporary studies have shifted the scientific interest towards the mechanisms that implicate miRNAs in various types of physiological and pathological conditions by integrating them in biological pathways.

Even though each aforementioned category includes includes numerous publications there are still open questions in every aspect of miRNA-related research. In example, there is a great gap in the knowledge regarding miRNA gene transcription regulation due to the fact that existing studies either provide extremely low resolution or totally inaccurate solutions to the problem of miRNA transcription start site identification. On the other hand, the function of miRNAs has been extensively characterized and a substantial part of the miRNA targetome has been mapped. However, this wealth of information is scattered among thousands of manuscripts creating the need for the development of database solutions that will serve as miRNA:target interaction repositories. The final frontier is to utilize the miRNA expression regulation as well as the miRNA targetome in order to complete the mosaic of gene expression regulatory networks.

During the course of the Doctoral studies, numerous strudies have been published that provide robust computational methods for answering the majority of miRNA-related open questions. The work presenting microTSS algorithm (Georgakilas et al., 2014) is the first available study implementing a *Drosha* conditional allele animal model for the study of unprocessed pri-miRNA transcripts. The utilized mouse enabled for the first time a high-throughput pri-miRNA transcript identification using conventional RNA-

Sequencing. Recent advances in the field of Next Generation Sequencing resulted in a concurrent cost reduction and quality increase of derived data. As demonstrated in the same study, detection of intergenic pri-miRNAs is now achievable with the use of deeply sequenced transcriptomic RNA-Seq, ChIP-Seq and DNase-Seq experiments. Such data can be analyzed using microTSS, in order to provide accurate and high-resolution miRNA TSS predictions. The novelty of the algorithm resides in its ability to integrate tissue specific deeply sequenced RNA-Seq data, resulting in single nucleotide TSS predictions. It is able to detect TSSs currently active in cell lines or conditions of interest. The analysis of microTSS predictions in mES, hES and IMR90 cells showed that a significant number of pri-miRNAs overlap partially or completely with previously annotated lncRNAs suggesting incomplete annotation of certain non-coding loci and/or multiple functionality. microTSS is a resource able to facilitate the annotation of pri-miRNAs and non-coding transcripts in general, as well as to support targeted functional studies of lncRNAs. microTSS results have also revealed novel dicistronic and polycistronic miRNA transcripts. The analysis of small RNA-Seq data in mESCs has additionally depicted variable expression levels between co-clustered precursors. There are cases where specific miRNAs exhibit zero or low expression as compared to other members of the same cluster. Such observations have also been reported in previous studies (Chawla & Sokol, 2014), suggesting post-transcriptional mechanisms able to block the maturation process of individual members derived from polycistronic miRNA genes. The analysis of GRO-Seq data unveiled a significant number of divergent pri-miRNAs upstream of protein coding gene promoters. The significantly smaller degree of conservation in these precursor sequences directly supports the proposed hypothesis (X. Wu & Sharp, 2013) that divergent transcription is a model of new gene formation. The analysis of long and small RNA-Seq data in mouse indicates that the maturation process of miRNAs located in divergent transcripts is repressed more often than expected on a later stage. Even though the small RNA dataset is deeply sequenced, we cannot exclude the possibility that these miRNAs are expressed below its detection limit. However, even in this case there is a strong indication that their rate of maturation is either blocked or actively regulated. It can also be connected to the aberrant divergent transcription observed in ES cells, serving as a means of cell protection from redundant miRNA transcription. Only a few of the miRNAs located in divergent transcripts had experimentally validated targets but all were predicted to have a significant number of *in silico* identified interactions. It could be possible that processing of divergent miRNA transcripts is more difficult to be regulated since it is not independent from the transcription of adjacent protein coding genes. A recently discovered mechanism (Chawla & Sokol, 2014) enables cells to distinguish miRNAs located in the same polycistronic transcript by blocking others and preferentially allowing only their maturation. The existence of this additional layer of post-transcriptional miRNA biogenesis regulation might be the only way to enable the preferential tissue or cell line

specific expression/repression of divergent miRNAs (and other monocistronic pri-miRNAs). Future experiments in multiple tissues would provide valuable information towards the evaluation of this hypothesis.

The transition from TarBase v5 (Papadopoulos et al., 2009) to TarBase v6 (Vergoulis et al., 2012) included a 50-fold target increase, coupled with a significant extension of specific research-oriented features. The latest version of DIANA-TarBase (Vlachos, Paraskevopoulou, et al., 2015) has been completely redesigned in every aspect. The utilized database has been significantly extended with richer meta-data and detailed information for each interaction, while the interface now supports advanced real-time querying and result filtering. During the past few years NGS methodologies have revolutionized almost every aspect of biological research. Novel NGS-based high-throughput miRNA target identification techniques have enabled the identification of thousands of interactions present in specific cell types or experimental conditions. By analyzing more than 150 raw NGS data sets and extracting interactions from hundreds of meticulously curated articles, DIANA-TarBase v7 is the first relevant database to break the barrier of 100 000 entries by indexing more than half a million interactions in 24 species, 9–250 times more than any other manually curated database. This wealth of information can be utilized for exploratory studies, enforcing or even at cases substituting in silico predicted interactions.

As we learn more about miRNA:gene interactions, the in silico analysis tools and applications mature and grow, in order to support more demanding research scenarios. The latest version of DIANA-miRPath (Vlachos, Zagganas, et al., 2015) combines leading state of the art target prediction algorithms (Garcia et al., 2011; Maragkakis et al., 2009; Paraskevopoulou, Georgakilas, Kostoulas, Vlachos, et al., 2013; Reczko et al., 2012), with the most extensive manually curated miRNA:gene interaction dataset to date (Vlachos, Paraskevopoulou, et al., 2015), in order to chart the miRNA target search space. The new user interface enables extensive parameterization and tailor-made analyses, with selection options spanning from prediction thresholds to settings deep under the hood of the statistics engine. The latter has been significantly redesigned, in order to support novel statistics methodologies that are based on empirical distributions and do not suffer from the biases observed in standard approaches. DIANA-miRPath v3 is designed to accommodate diverse research needs that require accurate functional characterization of one or more microRNAs. The incorporation of KEGG pathways and multiple gene ontologies, as well as numerous links to DIANA and external tools or databases, meta-data, SNP analysis modules, clustering algorithms and advanced visualizations, render DIANA-miRPath v3 as an one-stop hub for miRNA research projects. Unique features such as “Reverse Search Module” enables miRNA functional analysis tools to be also utilized as a first exploratory research step, as well as a companion along a research endeavor.

The implemented computational methods are readily applicable to a variety of cell lines or organisms. These methodologies can be utilized separately or combined, depending on the study setting, availability of datasets and genome annotation of the examined organism. The identification of differences in miRNA expression regulation as well as target repertoire between pathological and physiological conditions, cell types and species, could inaugurate a new era for the elucidation of miRNA expression and function redefining their role into the wider context of biological pathways.

4. Publications

During the course of my Doctoral studies, a record of seven publications has been achieved which is presented below in chronological order.

1. Vlachos IS, Zagganas K, Paraskevopoulou MD, **Georgakilas G**, Karagkouni D, Vergoulis T, Dalamagas T, Hatzigeorgiou AG. DIANA-miRPath v3.0: Deciphering microRNA function with experimental support. Nucleic Acids Research (**8.8 Impact Factor, 2015**). 2015.

2. **Georgakilas G**, Vlachos IS, Paraskevopoulou MD, Yang P, Zhang Y, Economides AN, Hatzigeorgiou AG. microTSS: accurate microRNA transcription start site identification reveals a significant number of divergent pri-miRNAs. Nature Communications (**10.7 Impact Factor, 2015**). 2014.

3. Vlachos IS and Paraskevopoulou MD, Karagkouni D, **Georgakilas G**, Vergoulis T, Kanellos I, Anastasopoulos IL, Maniou S, Karathanou K, Kalfakakou D, Dalamagas T, Hatzigeorgiou AG. DIANA-TarBase v7.0: Indexing more than half a million experimentally supported miRNA:mRNA interactions. Nucleic Acids Research (**8.8 Impact Factor, 2015**). 2014.

4. Vergoulis T, Kanellos I, Kostoulas N, **Georgakilas G**, Sellis T, Hatzigeorgiou AG, Dalamagas T. mirPub: a database for searching microRNA publications. BMC Bioinformatics (**2.6 Impact Factor, 2015**). 2014.

5. Paraskevopoulou MD and **Georgakilas G**, Kostoulas N, Vlachos IS, Vergoulis T, Reczko M, Filippidis C, Dalamagas T, Hatzigeorgiou AG. DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows. Nucleic Acids Research (**8.8 Impact Factor, 2015**). 2013. (**joint first authorship**)

6. Paraskevopoulou MD and **Georgakilas G**, Kostoulas N, Reczko M, Maragkakis M, Dalamagas TM, Hatzigeorgiou AG. DIANA-LncBase: experimentally verified and computationally predicted microRNA targets on long non-coding RNAs. Nucleic Acids Research (**8.8 Impact Factor, 2015**). 2013. (**joint first authorship**)

7. Vlachos IS and Kostoulas N, Vergoulis T, **Georgakilas G**, Reczko M, Maragkakis M, Paraskevopoulou MD, Prionidis K, Dalamagas T, Hatzigeorgiou AG. DIANA miRPath v.2.0: investigating the combinatorial effect of microRNAs in pathways. Nucleic Acids Research (**8.8 Impact Factor, 2015**). 2012.

5. References

Alexiou, P., Maragkakis, M., Papadopoulos, G. L., Reczko, M., & Hatzigeorgiou, A. G. (2009). Lost in translation: an assessment and perspective for computational microRNA target identification. *Bioinformatics, 25*(23), 3049-3055. doi: 10.1093/bioinformatics/btp565

Alisi, A., Da Sacco, L., Bruscalupi, G., Piemonte, F., Panera, N., De Vito, R., . . . Nobili, V. (2011). Mirnome analysis reveals novel molecular determinants in the pathogenesis of diet-induced nonalcoholic fatty liver disease. *Lab Invest, 91*(2), 283-293. doi: 10.1038/labinvest.2010.166

Ambros, V. (2001). microRNAs: tiny regulators with great potential. *Cell, 107*(7), 823-826.

Andersson, M. G., Haasnoot, P. C., Xu, N., Berenjian, S., Berkhout, B., & Akusjarvi, G. (2005). Suppression of RNA interference by adenovirus virus-associated RNA. *J Virol, 79*(15), 9556-9565. doi: 10.1128/JVI.79.15.9556-9565.2005

Ares, M., Jr., Grate, L., & Pauling, M. H. (1999). A handful of intron-containing genes produces the lion's share of yeast mRNA. *RNA, 5*(9), 1138-1139.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., . . . Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet, 25*(1), 25-29. doi: 10.1038/75556

Baek, D., Villen, J., Shin, C., Camargo, F. D., Gygi, S. P., & Bartel, D. P. (2008). The impact of microRNAs on protein output. *Nature, 455*(7209), 64-71. doi: 10.1038/nature07242

Bagga, S., Bracht, J., Hunter, S., Massirer, K., Holtz, J., Eachus, R., & Pasquinelli, A. E. (2005). Regulation by let-7 and lin-4 miRNAs results in target mRNA degradation. *Cell, 122*(4), 553-563. doi: 10.1016/j.cell.2005.07.031

Barski, A., Jothi, R., Cuddapah, S., Cui, K., Roh, T. Y., Schones, D. E., & Zhao, K. (2009). Chromatin poises miRNA- and protein-coding genes for expression. *Genome Res, 19*(10), 1742-1751. doi: 10.1101/gr.090951.109

Bartel, D. P. (2009). MicroRNAs: target recognition and regulatory functions. *Cell, 136*(2), 215-233. doi: 10.1016/j.cell.2009.01.002

Bernstein, E., Kim, S. Y., Carmell, M. A., Murchison, E. P., Alcorn, H., Li, M. Z., . . . Hannon, G. J. (2003). Dicer is essential for mouse development. *Nat Genet, 35*(3), 215-217. doi: 10.1038/ng1253

Bertani, S., Sauer, S., Bolotin, E., & Sauer, F. (2011). The noncoding RNA Mistral activates Hoxa6 and Hoxa7 expression and stem cell differentiation by recruiting MLL1 to chromatin. *Mol Cell, 43*(6), 1040-1046. doi: 10.1016/j.molcel.2011.08.019

Betel, D., Koppal, A., Agius, P., Sander, C., & Leslie, C. (2010). Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biol, 11*(8), R90. doi: 10.1186/gb-2010-11-8-r90

Borchert, G. M., Lanier, W., & Davidson, B. L. (2006). RNA polymerase III transcribes human microRNAs. *Nat Struct Mol Biol, 13*(12), 1097-1101. doi: 10.1038/nsmb1167

Brannan, C. I., Dees, E. C., Ingram, R. S., & Tilghman, S. M. (1990). The product of the H19 gene may function as an RNA. *Mol Cell Biol, 10*(1), 28-36.

Brengues, M., Teixeira, D., & Parker, R. (2005). Movement of eukaryotic mRNAs between polysomes and cytoplasmic processing bodies. *Science, 310*(5747), 486-489. doi: 10.1126/science.1115791

Brennecke, J., Hipfner, D. R., Stark, A., Russell, R. B., & Cohen, S. M. (2003). bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in Drosophila. *Cell, 113*(1), 25-36.

Brennecke, J., Stark, A., Russell, R. B., & Cohen, S. M. (2005). Principles of microRNA-target recognition. *PLoS Biol, 3*(3), e85. doi: 10.1371/journal.pbio.0030085

Brockdorff, N., Ashworth, A., Kay, G. F., McCabe, V. M., Norris, D. P., Cooper, P. J., . . . Rastan, S. (1992). The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. *Cell, 71*(3), 515-526.

Buratowski, S. (2008). Transcription. Gene expression--where to start? *Science, 322*(5909), 1804-1805. doi: 10.1126/science.1168805

Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., & Rinn, J. L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev, 25*(18), 1915-1927. doi: 10.1101/gad.17446611

Cai, X., Lu, S., Zhang, Z., Gonzalez, C. M., Damania, B., & Cullen, B. R. (2005). Kaposi's sarcoma-associated herpesvirus expresses an array of viral microRNAs in latently infected cells. *Proc Natl Acad Sci U S A, 102*(15), 5570-5575. doi: 10.1073/pnas.0408192102

Calin, G. A., Dumitru, C. D., Shimizu, M., Bichi, R., Zupo, S., Noch, E., . . . Croce, C. M. (2002). Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. *Proc Natl Acad Sci U S A, 99*(24), 15524-15529. doi: 10.1073/pnas.242606799

Cesana, M., Cacchiarelli, D., Legnini, I., Santini, T., Sthandier, O., Chinappi, M., . . . Bozzoni, I. (2011). A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. *Cell, 147*(2), 358-369. doi: 10.1016/j.cell.2011.09.028

Chan, J. A., Krichevsky, A. M., & Kosik, K. S. (2005). MicroRNA-21 is an antiapoptotic factor in human glioblastoma cells. *Cancer Res, 65*(14), 6029-6033. doi: 10.1158/0008-5472.CAN-05-0137

Chang, C., Lin, C. (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology, 2*(27:21), 27:27.

Chang, G., Gao, S., Hou, X., Xu, Z., Liu, Y., Kang, L., . . . Tian, J. (2014). High-throughput sequencing reveals the disruption of methylation of imprinted gene in induced pluripotent stem cells. *Cell Res, 24*(3), 293-306. doi: 10.1038/cr.2013.173

Chang, T. C., Yu, D., Lee, Y. S., Wentzel, E. A., Arking, D. E., West, K. M., . . . Mendell, J. T. (2008). Widespread microRNA repression by Myc contributes to tumorigenesis. *Nat Genet, 40*(1), 43-50. doi: 10.1038/ng.2007.30

Chawla, G., & Sokol, N. S. (2014). ADAR mediates differential expression of polycistronic microRNAs. *Nucleic Acids Res, 42*(8), 5245-5255. doi: 10.1093/nar/gku145

Chen, G., Wang, Z., Wang, D., Qiu, C., Liu, M., Chen, X., . . . Cui, Q. (2013). LncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Res, 41*(Database issue), D983-986. doi: 10.1093/nar/gks1099

Chen, J. F., Mandel, E. M., Thomson, J. M., Wu, Q., Callis, T. E., Hammond, S. M., . . . Wang, D. Z. (2006). The role of microRNA-1 and microRNA-133 in skeletal muscle proliferation and differentiation. *Nat Genet, 38*(2), 228-233. doi: 10.1038/ng1725

Chen, X. (2004). A microRNA as a translational repressor of APETALA2 in Arabidopsis flower development. *Science, 303*(5666), 2022-2025. doi: 10.1126/science.1088060

Chi, S. W., Hannon, G. J., & Darnell, R. B. (2012). An alternative mode of microRNA target recognition. *Nat Struct Mol Biol, 19*(3), 321-327. doi: 10.1038/nsmb.2230

Chi, S. W., Zang, J. B., Mele, A., & Darnell, R. B. (2009). Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature, 460*(7254), 479-486. doi: 10.1038/nature08170

Chien, C. H., Sun, Y. M., Chang, W. C., Chiang-Hsieh, P. Y., Lee, T. Y., Tsai, W. C., . . . Huang, H. D. (2011). Identifying transcriptional start sites of human microRNAs based on high-throughput sequencing data. *Nucleic Acids Res, 39*(21), 9345-9356. doi: 10.1093/nar/gkr604

Cho, W. C., Chow, A. S., & Au, J. S. (2011). MiR-145 inhibits cell proliferation of human lung adenocarcinoma by targeting EGFR and NUDT1. *RNA Biol, 8*(1), 125-131.

Chu, C., Qu, K., Zhong, F. L., Artandi, S. E., & Chang, H. Y. (2011). Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. *Mol Cell, 44*(4), 667-678. doi: 10.1016/j.molcel.2011.08.027

Cimmino, A., Calin, G. A., Fabbri, M., Iorio, M. V., Ferracin, M., Shimizu, M., . . . Croce, C. M. (2005). miR-15 and miR-16 induce apoptosis by targeting BCL2. *Proc Natl Acad Sci U S A, 102*(39), 13944-13949. doi: 10.1073/pnas.0506654102

Clark, M. B., Johnston, R. L., Inostroza-Ponta, M., Fox, A. H., Fortini, E., Moscato, P., . . . Mattick, J. S. (2012). Genome-wide analysis of long noncoding RNA stability. *Genome Res, 22*(5), 885-898. doi: 10.1101/gr.131037.111

Cloonan, N., Wani, S., Xu, Q., Gu, J., Lea, K., Heater, S., . . . Grimmond, S. M. (2011). MicroRNAs and their isomiRs function cooperatively to target common biological pathways. *Genome Biol, 12*(12), R126. doi: 10.1186/gb-2011-12-12-r126

Coller, J., & Parker, R. (2005). General translational repression by activators of mRNA decapping. *Cell, 122*(6), 875-886. doi: 10.1016/j.cell.2005.07.012

Consortium, Encode Project. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature, 489*(7414), 57-74. doi: 10.1038/nature11247

Corcoran, D. L., Pandit, K. V., Gordon, B., Bhattacharjee, A., Kaminski, N., & Benos, P. V. (2009). Features of mammalian microRNA promoters emerge from polymerase II chromatin immunoprecipitation data. *PLoS One, 4*(4), e5279. doi: 10.1371/journal.pone.0005279

Core, L. J., Waterfall, J. J., & Lis, J. T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science, 322*(5909), 1845-1848. doi: 10.1126/science.1162228

Cougot, N., Babajko, S., & Seraphin, B. (2004). Cytoplasmic foci are sites of mRNA decay in human cells. *J Cell Biol, 165*(1), 31-40. doi: 10.1083/jcb.200309008

Czech, B., Malone, C. D., Zhou, R., Stark, A., Schlingeheyde, C., Dus, M., . . . Brennecke, J. (2008). An endogenous small interfering RNA pathway in Drosophila. *Nature, 453*(7196), 798-802. doi: 10.1038/nature07007

Darzacq, X., Jady, B. E., Verheggen, C., Kiss, A. M., Bertrand, E., & Kiss, T. (2002). Cajal body-specific small nuclear RNAs: a novel class of 2'-O-methylation and pseudouridylation guide RNAs. *EMBO J, 21*(11), 2746-2756. doi: 10.1093/emboj/21.11.2746

Das, S., Ghosal, S., Sen, R., & Chakrabarti, J. (2014). lnCeDB: database of human long noncoding RNA acting as competing endogenous RNA. *PLoS One, 9*(6), e98965. doi: 10.1371/journal.pone.0098965

Davis, M. P., van Dongen, S., Abreu-Goodger, C., Bartonicek, N., & Enright, A. J. (2013). Kraken: a set of tools for quality control and analysis of high-throughput sequence data. *Methods, 63*(1), 41-49. doi: 10.1016/j.ymeth.2013.06.027

Degroeve, S., De Baets, B., Van de Peer, Y., & Rouze, P. (2002). Feature subset selection for splice site prediction. *Bioinformatics, 18 Suppl 2*, S75-83.

Dekker, J., Marti-Renom, M. A., & Mirny, L. A. (2013). Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat Rev Genet, 14*(6), 390-403. doi: 10.1038/nrg3454

Denli, A. M., Tops, B. B., Plasterk, R. H., Ketting, R. F., & Hannon, G. J. (2004). Processing of primary microRNAs by the Microprocessor complex. *Nature, 432*(7014), 231-235. doi: 10.1038/nature03049

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., . . . Guigo, R. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res, 22*(9), 1775-1789. doi: 10.1101/gr.132159.111

Didiano, D., & Hobert, O. (2006). Perfect seed pairing is not a generally reliable predictor for miRNA-target interactions. *Nat Struct Mol Biol, 13*(9), 849-851. doi: 10.1038/nsmb1138

Doench, J. G., & Sharp, P. A. (2004). Specificity of microRNA target selection in translational repression. *Genes Dev, 18*(5), 504-511. doi: 10.1101/gad.1184404

Dore, L. C., Amigo, J. D., Dos Santos, C. O., Zhang, Z., Gai, X., Tobias, J. W., . . . Weiss, M. J. (2008). A GATA-1-regulated microRNA locus essential for erythropoiesis. *Proc Natl Acad Sci U S A, 105*(9), 3333-3338. doi: 10.1073/pnas.0712312105

Duret, L., Chureau, C., Samain, S., Weissenbach, J., & Avner, P. (2006). The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. *Science, 312*(5780), 1653-1655. doi: 10.1126/science.1126316

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*(5), 839-847. doi: 10.1016/j.jbi.2011.05.002

Economides, A. N., Frendewey, D., Yang, P., Dominguez, M. G., Dore, A. T., Lobov, I. B., . . . Yancopoulos, G. D. (2013). Conditionals by inversion provide a universal method for the generation of conditional alleles. *Proc Natl Acad Sci U S A, 110*(34), E3179-3188. doi: 10.1073/pnas.1217812110

Ernst, J., & Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. *Nat Methods, 9*(3), 215-216. doi: 10.1038/nmeth.1906

Faghihi, M. A., & Wahlestedt, C. (2009). Regulatory roles of natural antisense transcripts. *Nat Rev Mol Cell Biol, 10*(9), 637-643. doi: 10.1038/nrm2738

Faghihi, M. A., Zhang, M., Huang, J., Modarresi, F., Van der Brug, M. P., Nalls, M. A., . . . Wahlestedt, C. (2010). Evidence for natural antisense transcript-mediated inhibition of microRNA function. *Genome Biol, 11*(5), R56. doi: 10.1186/gb-2010-11-5-r56

Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., & Mello, C. C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature, 391*(6669), 806-811. doi: 10.1038/35888

Flicek, P., Ahmed, I., Amode, M. R., Barrell, D., Beal, K., Brent, S., . . . Searle, S. M. (2013). Ensembl 2013. *Nucleic Acids Res, 41*(Database issue), D48-55. doi: 10.1093/nar/gks1236

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., . . . Searle, S. M. (2012). Ensembl 2012. *Nucleic Acids Res, 40*(Database issue), D84-90. doi: 10.1093/nar/gkr991

Franco-Zorrilla, J. M., Valli, A., Todesco, M., Mateos, I., Puga, M. I., Rubio-Somoza, I., . . . Paz-Ares, J. (2007). Target mimicry provides a new mechanism for regulation of microRNA activity. *Nat Genet, 39*(8), 1033-1037. doi: 10.1038/ng2079

Friedman, R. C., Farh, K. K., Burge, C. B., & Bartel, D. P. (2009). Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res, 19*(1), 92-105. doi: 10.1101/gr.082701.108

Gaidatzis, D., van Nimwegen, E., Hausser, J., & Zavolan, M. (2007). Inference of miRNA targets using evolutionary conservation and pathway analysis. *BMC Bioinformatics, 8*, 69. doi: 10.1186/1471-2105-8-69

Gao, P., Xing, A. Y., Zhou, G. Y., Zhang, T. G., Zhang, J. P., Gao, C., . . . Shi, D. B. (2013). The molecular mechanism of microRNA-145 to suppress invasion-metastasis cascade in gastric cancer. *Oncogene, 32*(4), 491-501. doi: 10.1038/onc.2012.61

Garber, M., Guttman, M., Clamp, M., Zody, M. C., Friedman, N., & Xie, X. (2009). Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics, 25*(12), i54-62. doi: 10.1093/bioinformatics/btp190

Garcia, D. M., Baek, D., Shin, C., Bell, G. W., Grimson, A., & Bartel, D. P. (2011). Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. *Nat Struct Mol Biol, 18*(10), 1139-1146. doi: 10.1038/nsmb.2115

Genome, K. Community of Scientists. (2009). Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. *J Hered, 100*(6), 659-674. doi: 10.1093/jhered/esp086

Genomes Project, Consortium, Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., . . . McVean, G. A. (2010). A map of human genome variation from population-scale sequencing. *Nature, 467*(7319), 1061-1073. doi: 10.1038/nature09534

Georgakilas, G., Vlachos, I. S., Paraskevopoulou, M. D., Yang, P., Zhang, Y., Economides, A. N., & Hatzigeorgiou, A. G. (2014). microTSS: accurate microRNA transcription start site identification reveals a significant number of divergent pri-miRNAs. *Nat Commun, 5*, 5700. doi: 10.1038/ncomms6700

Granneman, S., Kudla, G., Petfalski, E., & Tollervey, D. (2009). Identification of protein binding sites on U3 snoRNA and pre-rRNA by UV cross-linking and high-throughput analysis of cDNAs. *Proc Natl Acad Sci U S A, 106*(24), 9613-9618. doi: 10.1073/pnas.0901997106

Gregory, R. I., Yan, K. P., Amuthan, G., Chendrimada, T., Doratotaj, B., Cooch, N., & Shiekhattar, R. (2004). The Microprocessor complex mediates the genesis of microRNAs. *Nature, 432*(7014), 235-240. doi: 10.1038/nature03120

Griffiths-Jones, S. (2004). The microRNA Registry. *Nucleic Acids Res, 32*(Database issue), D109-111. doi: 10.1093/nar/gkh023

Griffiths-Jones, S., Grocock, R. J., van Dongen, S., Bateman, A., & Enright, A. J. (2006). miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res, 34*(Database issue), D140-144. doi: 10.1093/nar/gkj112

Grimson, A., Farh, K. K., Johnston, W. K., Garrett-Engele, P., Lim, L. P., & Bartel, D. P. (2007). MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell, 27*(1), 91-105. doi: 10.1016/j.molcel.2007.06.017

Guil, S., Soler, M., Portela, A., Carrere, J., Fonalleras, E., Gomez, A., . . . Esteller, M. (2012). Intronic RNAs mediate EZH2 regulation of epigenetic targets. *Nat Struct Mol Biol, 19*(7), 664-670. doi: 10.1038/nsmb.2315

Guttman, M., Amit, I., Garber, M., French, C., Lin, M. F., Feldser, D., . . . Lander, E. S. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature, 458*(7235), 223-227. doi: 10.1038/nature07672

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., . . . Tuschl, T. (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell, 141*(1), 129-141. doi: 10.1016/j.cell.2010.03.009

Han, J., Lee, Y., Yeom, K. H., Nam, J. W., Heo, I., Rhee, J. K., . . . Kim, V. N. (2006). Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. *Cell, 125*(5), 887-901. doi: 10.1016/j.cell.2006.03.043

Han, J., Pedersen, J. S., Kwon, S. C., Belair, C. D., Kim, Y. K., Yeom, K. H., . . . Kim, V. N. (2009). Posttranscriptional crossregulation between Drosha and DGCR8. *Cell, 136*(1), 75-84. doi: 10.1016/j.cell.2008.10.053

He, L., He, X., Lowe, S. W., & Hannon, G. J. (2007). microRNAs join the p53 network--another piece in the tumour-suppression puzzle. *Nat Rev Cancer, 7*(11), 819-822. doi: 10.1038/nrc2232

He, Y., Vogelstein, B., Velculescu, V. E., Papadopoulos, N., & Kinzler, K. W. (2008). The antisense transcriptomes of human cells. *Science, 322*(5909), 1855-1857. doi: 10.1126/science.1163853

Hebert, S. S., Horre, K., Nicolai, L., Papadopoulou, A. S., Mandemakers, W., Silahtaroglu, A. N., . . . De Strooper, B. (2008). Loss of microRNA cluster miR-29a/b-1 in sporadic Alzheimer's

disease correlates with increased BACE1/beta-secretase expression. *Proc Natl Acad Sci U S A, 105*(17), 6415-6420. doi: 10.1073/pnas.0710263105

Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., . . . Ren, B. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet, 39*(3), 311-318. doi: 10.1038/ng1966

Hsu, J. B., Chiu, C. M., Hsu, S. D., Huang, W. Y., Chien, C. H., Lee, T. Y., & Huang, H. D. (2011). miRTar: an integrated system for identifying miRNA-target interactions in human. *BMC Bioinformatics, 12*, 300. doi: 10.1186/1471-2105-12-300

Hsu, S. D., Chu, C. H., Tsou, A. P., Chen, S. J., Chen, H. C., Hsu, P. W., . . . Huang, H. D. (2008). miRNAMap 2.0: genomic maps of microRNAs in metazoan genomes. *Nucleic Acids Res, 36*(Database issue), D165-169. doi: 10.1093/nar/gkm1012

Hsu, S. D., Tseng, Y. T., Shrestha, S., Lin, Y. L., Khaleel, A., Chou, C. H., . . . Huang, H. D. (2014). miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. *Nucleic Acids Res, 42*(Database issue), D78-85. doi: 10.1093/nar/gkt1266

Hung, T., Wang, Y., Lin, M. F., Koegel, A. K., Kotake, Y., Grant, G. D., . . . Chang, H. Y. (2011). Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters. *Nat Genet, 43*(7), 621-629. doi: 10.1038/ng.848

Hutvagner, G., McLachlan, J., Pasquinelli, A. E., Balint, E., Tuschl, T., & Zamore, P. D. (2001). A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. *Science, 293*(5531), 834-838. doi: 10.1126/science.1062961

Hutvagner, G., & Zamore, P. D. (2002). A microRNA in a multiple-turnover RNAi enzyme complex. *Science, 297*(5589), 2056-2060. doi: 10.1126/science.1073827

International Human Genome Sequencing, Consortium. (2004). Finishing the euchromatic sequence of the human genome. *Nature, 431*(7011), 931-945. doi: 10.1038/nature03001

Iorio, M. V., Ferracin, M., Liu, C. G., Veronese, A., Spizzo, R., Sabbioni, S., . . . Croce, C. M. (2005). MicroRNA gene expression deregulation in human breast cancer. *Cancer Res, 65*(16), 7065-7070. doi: 10.1158/0008-5472.CAN-05-1783

Jakymiw, A., Lian, S., Eystathioy, T., Li, S., Satoh, M., Hamel, J. C., . . . Chan, E. K. (2005). Disruption of GW bodies impairs mammalian RNA interference. *Nat Cell Biol, 7*(12), 1267-1274. doi: 10.1038/ncb1334

Jeggari, A., Marks, D. S., & Larsson, E. (2012). miRcode: a map of putative microRNA target sites in the long non-coding transcriptome. *Bioinformatics, 28*(15), 2062-2063. doi: 10.1093/bioinformatics/bts344

Ji, P., Diederichs, S., Wang, W., Boing, S., Metzger, R., Schneider, P. M., . . . Muller-Tidow, C. (2003). MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene, 22*(39), 8031-8041. doi: 10.1038/sj.onc.1206928

Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., . . . Liu, Y. (2009). miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res, 37*(Database issue), D98-104. doi: 10.1093/nar/gkn714

Jin, F., Li, Y., Dixon, J. R., Selvaraj, S., Ye, Z., Lee, A. Y., . . . Ren, B. (2013). A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature, 503*(7475), 290-294. doi: 10.1038/nature12644

John, B., Enright, A. J., Aravin, A., Tuschl, T., Sander, C., & Marks, D. S. (2004). Human MicroRNA targets. *PLoS Biol, 2*(11), e363. doi: 10.1371/journal.pbio.0020363

John, S., Sabo, P. J., Canfield, T. K., Lee, K., Vong, S., Weaver, M., . . . Stamatoyannopoulos, J. A. (2013). Genome-scale mapping of DNase I hypersensitivity. *Curr Protoc Mol Biol, Chapter 27*, Unit 21 27. doi: 10.1002/0471142727.mb2127s103

Johnson, D. S., Mortazavi, A., Myers, R. M., & Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science, 316*(5830), 1497-1502. doi: 10.1126/science.1141319

Jones-Rhoades, M. W., Bartel, D. P., & Bartel, B. (2006). MicroRNAS and their regulatory roles in plants. *Annu Rev Plant Biol, 57*, 19-53. doi: 10.1146/annurev.arplant.57.032905.105218

Kallen, A. N., Zhou, X. B., Xu, J., Qiao, C., Ma, J., Yan, L., . . . Huang, Y. (2013). The imprinted H19 lncRNA antagonizes let-7 microRNAs. *Mol Cell, 52*(1), 101-112. doi: 10.1016/j.molcel.2013.08.027

Kanduri, C., Thakur, N., & Pandey, R. R. (2006). The length of the transcript encoded from the Kcnq1ot1 antisense promoter determines the degree of silencing. *EMBO J, 25*(10), 2096-2106. doi: 10.1038/sj.emboj.7601090

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., & Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res, 40*(Database issue), D109-114. doi: 10.1093/nar/gkr988

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2014). Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res, 42*(Database issue), D199-205. doi: 10.1093/nar/gkt1076

Kanhere, A., Viiri, K., Araujo, C. C., Rasaiyaah, J., Bouwman, R. D., Whyte, W. A., . . . Jenner, R. G. (2010). Short RNAs are transcribed from repressed polycomb target genes and interact with polycomb repressive complex-2. *Mol Cell, 38*(5), 675-688. doi: 10.1016/j.molcel.2010.03.019

Kapp, L. D., & Lorsch, J. R. (2004). The molecular mechanics of eukaryotic translation. *Annu Rev Biochem, 73*, 657-704. doi: 10.1146/annurev.biochem.73.030403.080419

Karolina, D. S., Armugam, A., Tavintharan, S., Wong, M. T., Lim, S. C., Sum, C. F., & Jeyaseelan, K. (2011). MicroRNA 144 impairs insulin signaling by inhibiting the expression of insulin receptor substrate 1 in type 2 diabetes mellitus. *PLoS One, 6*(8), e22839. doi: 10.1371/journal.pone.0022839

Katayama, S., Tomaru, Y., Kasukawa, T., Waki, K., Nakanishi, M., Nakamura, M., . . . Consortium, Fantom. (2005). Antisense transcription in the mammalian transcriptome. *Science, 309*(5740), 1564-1566. doi: 10.1126/science.1112009

Kawamura, Y., Saito, K., Kin, T., Ono, Y., Asai, K., Sunohara, T., . . . Siomi, H. (2008). Drosophila endogenous small RNAs bind to Argonaute 2 in somatic cells. *Nature, 453*(7196), 793-797. doi: 10.1038/nature06938

Keller, A., Leidinger, P., Lange, J., Borries, A., Schroers, H., Scheffler, M., . . . Meese, E. (2009). Multiple sclerosis: microRNA expression profiles accurately differentiate patients with relapsing-remitting disease from healthy controls. *PLoS One, 4*(10), e7440. doi: 10.1371/journal.pone.0007440

Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U., & Segal, E. (2007). The role of site accessibility in microRNA target recognition. *Nat Genet, 39*(10), 1278-1284. doi: 10.1038/ng2135

Khalil, A. M., Guttman, M., Huarte, M., Garber, M., Raj, A., Rivea Morales, D., . . . Rinn, J. L. (2009). Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc Natl Acad Sci U S A, 106*(28), 11667-11672. doi: 10.1073/pnas.0904715106

Khvorova, A., Reynolds, A., & Jayasena, S. D. (2003). Functional siRNAs and miRNAs exhibit strand bias. *Cell, 115*(2), 209-216.

Kim, J., Inoue, K., Ishii, J., Vanti, W. B., Voronov, S. V., Murchison, E., . . . Abeliovich, A. (2007). A MicroRNA feedback circuit in midbrain dopamine neurons. *Science, 317*(5842), 1220-1224. doi: 10.1126/science.1140481

Kim, T. K., Hemberg, M., Gray, J. M., Costa, A. M., Bear, D. M., Wu, J., . . . Greenberg, M. E. (2010). Widespread transcription at neuronal activity-regulated enhancers. *Nature, 465*(7295), 182-187. doi: 10.1038/nature09033

Kim, V. N. (2004). MicroRNA precursors in motion: exportin-5 mediates their nuclear export. *Trends Cell Biol, 14*(4), 156-159.

Kincaid, R. P., Burke, J. M., & Sullivan, C. S. (2012). RNA virus microRNA that mimics a B-cell oncomiR. *Proc Natl Acad Sci U S A, 109*(8), 3077-3082. doi: 10.1073/pnas.1116107109

Kino, T., Hurt, D. E., Ichijo, T., Nader, N., & Chrousos, G. P. (2010). Noncoding RNA gas5 is a growth arrest- and starvation-associated repressor of the glucocorticoid receptor. *Sci Signal, 3*(107), ra8. doi: 10.1126/scisignal.2000568

Kiss, T. (2002). Small nucleolar RNAs: an abundant group of noncoding RNAs with diverse cellular functions. *Cell, 109*(2), 145-148.

Kloosterman, W. P., Wienholds, E., Ketting, R. F., & Plasterk, R. H. (2004). Substrate requirements for let-7 function in the developing zebrafish embryo. *Nucleic Acids Res, 32*(21), 6284-6291. doi: 10.1093/nar/gkh968

Kong, W., He, L., Coppola, M., Guo, J., Esposito, N. N., Coppola, D., & Cheng, J. Q. (2010). MicroRNA-155 regulates cell survival, growth, and chemosensitivity by targeting FOXO3a in breast cancer. *J Biol Chem, 285*(23), 17869-17879. doi: 10.1074/jbc.M110.101055

Konig, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., . . . Ule, J. (2010). iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat Struct Mol Biol, 17*(7), 909-915. doi: 10.1038/nsmb.1838

Kowarsch, A., Preusse, M., Marr, C., & Theis, F. J. (2011). miTALOS: analyzing the tissue-specific regulation of signaling pathways by human and mouse microRNAs. *RNA, 17*(5), 809-819. doi: 10.1261/rna.2474511

Kozomara, A., & Griffiths-Jones, S. (2011). miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res, 39*(Database issue), D152-157. doi: 10.1093/nar/gkq1027

Kozomara, A., & Griffiths-Jones, S. (2014). miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res, 42*(Database issue), D68-73. doi: 10.1093/nar/gkt1181

Kuhn, D. E., Martin, M. M., Feldman, D. S., Terry, A. V., Jr., Nuovo, G. J., & Elton, T. S. (2008). Experimental validation of miRNA targets. *Methods, 44*(1), 47-54. doi: 10.1016/j.ymeth.2007.09.005

Lal, A., Navarro, F., Maher, C. A., Maliszewski, L. E., Yan, N., O'Day, E., . . . Lieberman, J. (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*(5), 610-625. doi: 10.1016/j.molcel.2009.08.020

Lall, S., Grun, D., Krek, A., Chen, K., Wang, Y. L., Dewey, C. N., . . . Rajewsky, N. (2006). A genome-wide map of conserved microRNA targets in C. elegans. *Curr Biol, 16*(5), 460-471. doi: 10.1016/j.cub.2006.01.050

Landthaler, M., Yalcin, A., & Tuschl, T. (2004). The human DiGeorge syndrome critical region gene 8 and Its D. melanogaster homolog are required for miRNA biogenesis. *Curr Biol, 14*(23), 2162-2167. doi: 10.1016/j.cub.2004.11.001

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol, 10*(3), R25. doi: 10.1186/gb-2009-10-3-r25

Lau, N. C., Lim, L. P., Weinstein, E. G., & Bartel, D. P. (2001). An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science, 294*(5543), 858-862. doi: 10.1126/science.1065062

Lau, N. C., Seto, A. G., Kim, J., Kuramochi-Miyagawa, S., Nakano, T., Bartel, D. P., & Kingston, R. E. (2006). Characterization of the piRNA complex from rat testes. *Science, 313*(5785), 363-367. doi: 10.1126/science.1130164

Law, J. A., & Jacobsen, S. E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat Rev Genet, 11*(3), 204-220. doi: 10.1038/nrg2719

Lee, R. C., & Ambros, V. (2001). An extensive class of small RNAs in Caenorhabditis elegans. *Science, 294*(5543), 862-864. doi: 10.1126/science.1065329

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

Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., . . . Kim, V. N. (2003). The nuclear RNase III Drosha initiates microRNA processing. *Nature, 425*(6956), 415-419. doi: 10.1038/nature01957

Lee, Y., Jeon, K., Lee, J. T., Kim, S., & Kim, V. N. (2002). MicroRNA maturation: stepwise processing and subcellular localization. *EMBO J, 21*(17), 4663-4670.

Lee, Y., Kim, M., Han, J., Yeom, K. H., Lee, S., Baek, S. H., & Kim, V. N. (2004). MicroRNA genes are transcribed by RNA polymerase II. *EMBO J, 23*(20), 4051-4060. doi: 10.1038/sj.emboj.7600385

Lewis, B. P., Burge, C. B., & Bartel, D. P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell, 120*(1), 15-20. doi: 10.1016/j.cell.2004.12.035

Li, J. H., Liu, S., Zhou, H., Qu, L. H., & Yang, J. H. (2014). starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res, 42*(Database issue), D92-97. doi: 10.1093/nar/gkt1248

Liu, J., Rivas, F. V., Wohlschlegel, J., Yates, J. R., 3rd, Parker, R., & Hannon, G. J. (2005). A role for the P-body component GW182 in microRNA function. *Nat Cell Biol, 7*(12), 1261-1266. doi: 10.1038/ncb1333

Liu, J., Valencia-Sanchez, M. A., Hannon, G. J., & Parker, R. (2005). MicroRNA-dependent localization of targeted mRNAs to mammalian P-bodies. *Nat Cell Biol, 7*(7), 719-723. doi: 10.1038/ncb1274

Loeb, G. B., Khan, A. A., Canner, D., Hiatt, J. B., Shendure, J., Darnell, R. B., . . . Rudensky, A. Y. (2012). Transcriptome-wide miR-155 binding map reveals widespread noncanonical microRNA targeting. *Mol Cell, 48*(5), 760-770. doi: 10.1016/j.molcel.2012.10.002

Lofgren, S. E., Frostegard, J., Truedsson, L., Pons-Estel, B. A., D'Alfonso, S., Witte, T., . . . Alarcon-Riquelme, M. E. (2012). Genetic association of miRNA-146a with systemic lupus erythematosus in Europeans through decreased expression of the gene. *Genes Immun, 13*(3), 268-274. doi: 10.1038/gene.2011.84

Londin, E., Loher, P., Telonis, A. G., Quann, K., Clark, P., Jing, Y., . . . Rigoutsos, I. (2015). Analysis of 13 cell types reveals evidence for the expression of numerous novel primate- and tissue-specific microRNAs. *Proc Natl Acad Sci U S A, 112*(10), E1106-1115. doi: 10.1073/pnas.1420955112

Louro, R., Smirnova, A. S., & Verjovski-Almeida, S. (2009). Long intronic noncoding RNA transcription: expression noise or expression choice? *Genomics, 93*(4), 291-298. doi: 10.1016/j.ygeno.2008.11.009

Luo, X., Yang, W., Ye, D. Q., Cui, H., Zhang, Y., Hirankarn, N., . . . Shen, N. (2011). A functional variant in microRNA-146a promoter modulates its expression and confers disease risk for systemic lupus erythematosus. *PLoS Genet, 7*(6), e1002128. doi: 10.1371/journal.pgen.1002128

Lyle, R., Watanabe, D., te Vruchte, D., Lerchner, W., Smrzka, O. W., Wutz, A., . . . Barlow, D. P. (2000). The imprinted antisense RNA at the Igf2r locus overlaps but does not imprint Mas1. *Nat Genet, 25*(1), 19-21. doi: 10.1038/75546

Lytle, J. R., Yario, T. A., & Steitz, J. A. (2007). Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5' UTR as in the 3' UTR. *Proc Natl Acad Sci U S A, 104*(23), 9667-9672. doi: 10.1073/pnas.0703820104

Maragkakis, M., Alexiou, P., Papadopoulos, G. L., Reczko, M., Dalamagas, T., Giannopoulos, G., . . . Hatzigeorgiou, A. G. (2009). Accurate microRNA target prediction correlates with protein repression levels. *BMC Bioinformatics, 10*, 295. doi: 10.1186/1471-2105-10-295

Marsico, A., Huska, M. R., Lasserre, J., Hu, H., Vucicevic, D., Musahl, A., . . . Vingron, M. (2013). PROmiRNA: a new miRNA promoter recognition method uncovers the complex regulation of intronic miRNAs. *Genome Biol, 14*(8), R84. doi: 10.1186/gb-2013-14-8-r84

Marson, A., Levine, S. S., Cole, M. F., Frampton, G. M., Brambrink, T., Johnstone, S., . . . Young, R. A. (2008). Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell, 134*(3), 521-533. doi: 10.1016/j.cell.2008.07.020

Martianov, I., Ramadass, A., Serra Barros, A., Chow, N., & Akoulitchev, A. (2007). Repression of the human dihydrofolate reductase gene by a non-coding interfering transcript. *Nature, 445*(7128), 666-670. doi: 10.1038/nature05519

Martin, H. C., Wani, S., Steptoe, A. L., Krishnan, K., Nones, K., Nourbakhsh, E., . . . Cloonan, N. (2014). Imperfect centered miRNA binding sites are common and can mediate repression of target mRNAs. *Genome Biol, 15*(3), R51. doi: 10.1186/gb-2014-15-3-r51

Matera, A. G., Terns, R. M., & Terns, M. P. (2007). Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. *Nat Rev Mol Cell Biol, 8*(3), 209-220. doi: 10.1038/nrm2124

Maxam, A. M., & Gilbert, W. (1977). A new method for sequencing DNA. *Proc Natl Acad Sci U S A, 74*(2), 560-564.

Megraw, M., Pereira, F., Jensen, S. T., Ohler, U., & Hatzigeorgiou, A. G. (2009). A transcription factor affinity-based code for mammalian transcription initiation. *Genome Res, 19*(4), 644-656. doi: 10.1101/gr.085449.108

Min, I. M., Waterfall, J. J., Core, L. J., Munroe, R. J., Schimenti, J., & Lis, J. T. (2011). Regulating RNA polymerase pausing and transcription elongation in embryonic stem cells. *Genes Dev, 25*(7), 742-754. doi: 10.1101/gad.2005511

Miranda, K. C., Huynh, T., Tay, Y., Ang, Y. S., Tam, W. L., Thomson, A. M., . . . Rigoutsos, I. (2006). A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell, 126*(6), 1203-1217. doi: 10.1016/j.cell.2006.07.031

Mitchell, T. (1997). Machine Learning. *McGraw-Hill*.

Monnier, P., Martinet, C., Pontis, J., Stancheva, I., Ait-Si-Ali, S., & Dandolo, L. (2013). H19 lncRNA controls gene expression of the Imprinted Gene Network by recruiting MBD1. *Proc Natl Acad Sci U S A, 110*(51), 20693-20698. doi: 10.1073/pnas.1310201110

Naeem, H., Kuffner, R., Csaba, G., & Zimmer, R. (2010). miRSel: automated extraction of associations between microRNAs and genes from the biomedical literature. *BMC Bioinformatics, 11*, 135. doi: 10.1186/1471-2105-11-135

Nagano, T., Mitchell, J. A., Sanz, L. A., Pauler, F. M., Ferguson-Smith, A. C., Feil, R., & Fraser, P. (2008). The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. *Science, 322*(5908), 1717-1720. doi: 10.1126/science.1163802

Nielsen, C. B., Shomron, N., Sandberg, R., Hornstein, E., Kitzman, J., & Burge, C. B. (2007). Determinants of targeting by endogenous and exogenous microRNAs and siRNAs. *RNA, 13*(11), 1894-1910. doi: 10.1261/rna.768207

O'Rourke, J. R., Georges, S. A., Seay, H. R., Tapscott, S. J., McManus, M. T., Goldhamer, D. J., . . . Harfe, B. D. (2007). Essential role for Dicer during skeletal muscle development. *Dev Biol, 311*(2), 359-368. doi: 10.1016/j.ydbio.2007.08.032

Okamura, K., Chung, W. J., Ruby, J. G., Guo, H., Bartel, D. P., & Lai, E. C. (2008). The Drosophila hairpin RNA pathway generates endogenous short interfering RNAs. *Nature, 453*(7196), 803-806. doi: 10.1038/nature07015

Ozsolak, F., Poling, L. L., Wang, Z., Liu, H., Liu, X. S., Roeder, R. G., . . . Fisher, D. E. (2008). Chromatin structure analyses identify miRNA promoters. *Genes Dev, 22*(22), 3172-3183. doi: 10.1101/gad.1706508

Pandey, R. R., Mondal, T., Mohammad, F., Enroth, S., Redrup, L., Komorowski, J., . . . Kanduri, C. (2008). Kcnq1ot1 antisense noncoding RNA mediates lineage-specific transcriptional silencing through chromatin-level regulation. *Mol Cell, 32*(2), 232-246. doi: 10.1016/j.molcel.2008.08.022

Papadopoulos, G. L., Reczko, M., Simossis, V. A., Sethupathy, P., & Hatzigeorgiou, A. G. (2009). The database of experimentally supported targets: a functional update of TarBase. *Nucleic Acids Res, 37*(Database issue), D155-158. doi: 10.1093/nar/gkn809

Paraskevopoulou, M. D., Georgakilas, G., Kostoulas, N., Reczko, M., Maragkakis, M., Dalamagas, T. M., & Hatzigeorgiou, A. G. (2013). DIANA-LncBase: experimentally verified and computationally predicted microRNA targets on long non-coding RNAs. *Nucleic Acids Res, 41*(Database issue), D239-245. doi: 10.1093/nar/gks1246

Paraskevopoulou, M. D., Georgakilas, G., Kostoulas, N., Vlachos, I. S., Vergoulis, T., Reczko, M., . . . Hatzigeorgiou, A. G. (2013). DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows. *Nucleic Acids Res, 41*(Web Server issue), W169-173. doi: 10.1093/nar/gkt393

Parker, R., & Song, H. (2004). The enzymes and control of eukaryotic mRNA turnover. *Nat Struct Mol Biol, 11*(2), 121-127. doi: 10.1038/nsmb724

Pasquinelli, A. E., Reinhart, B. J., Slack, F., Martindale, M. Q., Kuroda, M. I., Maller, B., . . . Ruvkun, G. (2000). Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature, 408*(6808), 86-89. doi: 10.1038/35040556

Peters, L., & Meister, G. (2007). Argonaute proteins: mediators of RNA silencing. *Mol Cell, 26*(5), 611-623. doi: 10.1016/j.molcel.2007.05.001

Pfeffer, S., Zavolan, M., Grasser, F. A., Chien, M., Russo, J. J., Ju, J., . . . Tuschl, T. (2004). Identification of virus-encoded microRNAs. *Science, 304*(5671), 734-736. doi: 10.1126/science.1096781

Pillai, R. S., Bhattacharyya, S. N., Artus, C. G., Zoller, T., Cougot, N., Basyuk, E., . . . Filipowicz, W. (2005). Inhibition of translational initiation by Let-7 MicroRNA in human cells. *Science, 309*(5740), 1573-1576. doi: 10.1126/science.1115079

Pink, R. C., Wicks, K., Caley, D. P., Punch, E. K., Jacobs, L., & Carter, D. R. (2011). Pseudogenes: pseudo-functional or key regulators in health and disease? *RNA, 17*(5), 792-798. doi: 10.1261/rna.2658311

Planta, R. J., & Mager, W. H. (1998). The list of cytoplasmic ribosomal proteins of Saccharomyces cerevisiae. *Yeast, 14*(5), 471-477. doi: 10.1002/(SICI)1097-0061(19980330)14:5<471::AID-YEA241>3.0.CO;2-U

Poliseno, L., Salmena, L., Zhang, J., Carver, B., Haveman, W. J., & Pandolfi, P. P. (2010). A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature, 465*(7301), 1033-1038. doi: 10.1038/nature09144

Poy, M. N., Eliasson, L., Krutzfeldt, J., Kuwajima, S., Ma, X., Macdonald, P. E., . . . Stoffel, M. (2004). A pancreatic islet-specific microRNA regulates insulin secretion. *Nature, 432*(7014), 226-230. doi: 10.1038/nature03076

Preker, P., Nielsen, J., Kammler, S., Lykke-Andersen, S., Christensen, M. S., Mapendano, C. K., . . . Jensen, T. H. (2008). RNA exosome depletion reveals transcription upstream of active human promoters. *Science, 322*(5909), 1851-1854. doi: 10.1126/science.1164096

Rao, P. K., Kumar, R. M., Farkhondeh, M., Baskerville, S., & Lodish, H. F. (2006). Myogenic factors that regulate expression of muscle-specific microRNAs. *Proc Natl Acad Sci U S A, 103*(23), 8721-8726. doi: 10.1073/pnas.0602831103

Reczko, M., Maragkakis, M., Alexiou, P., Grosse, I., & Hatzigeorgiou, A. G. (2012). Functional microRNA targets in protein coding sequences. *Bioinformatics, 28*(6), 771-776. doi: 10.1093/bioinformatics/bts043

Reichow, S. L., Hamma, T., Ferre-D'Amare, A. R., & Varani, G. (2007). The structure and function of small nucleolar ribonucleoproteins. *Nucleic Acids Res, 35*(5), 1452-1464. doi: 10.1093/nar/gkl1172

Reinhart, B. J., Slack, F. J., Basson, M., Pasquinelli, A. E., Bettinger, J. C., Rougvie, A. E., . . . Ruvkun, G. (2000). The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature, 403*(6772), 901-906. doi: 10.1038/35002607

Richter, J. D., & Sonenberg, N. (2005). Regulation of cap-dependent translation by eIF4E inhibitory proteins. *Nature, 433*(7025), 477-480. doi: 10.1038/nature03205

Rinn, J. L., Kertesz, M., Wang, J. K., Squazzo, S. L., Xu, X., Brugmann, S. A., . . . Chang, H. Y. (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell, 129*(7), 1311-1323. doi: 10.1016/j.cell.2007.05.022

Roadmap Epigenomics, Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., . . . Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. *Nature, 518*(7539), 317-330. doi: 10.1038/nature14248

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics, 26*(1), 139-140. doi: 10.1093/bioinformatics/btp616

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychol Rev*(65), 386-408.

Saini, H. K., Enright, A. J., & Griffiths-Jones, S. (2008). Annotation of mammalian primary microRNAs. *BMC Genomics, 9*, 564. doi: 10.1186/1471-2164-9-564

Saini, H. K., Griffiths-Jones, S., & Enright, A. J. (2007). Genomic analysis of human microRNA transcripts. *Proc Natl Acad Sci U S A, 104*(45), 17719-17724. doi: 10.1073/pnas.0703890104

Salmena, L., Poliseno, L., Tay, Y., Kats, L., & Pandolfi, P. P. (2011). A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell, 146*(3), 353-358. doi: 10.1016/j.cell.2011.07.014

Sanford, J. R., Wang, X., Mort, M., Vanduyn, N., Cooper, D. N., Mooney, S. D., . . . Liu, Y. (2009). Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts. *Genome Res, 19*(3), 381-394. doi: 10.1101/gr.082503.108

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A, 74*(12), 5463-5467.

Schnell, J. R., Dyson, H. J., & Wright, P. E. (2004). Structure, dynamics, and catalytic function of dihydrofolate reductase. *Annu Rev Biophys Biomol Struct, 33*, 119-140. doi: 10.1146/annurev.biophys.33.110502.133613

Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thastrom, A., Field, Y., Moore, I. K., . . . Widom, J. (2006). A genomic code for nucleosome positioning. *Nature, 442*(7104), 772-778. doi: 10.1038/nature04979

Seila, A. C., Calabrese, J. M., Levine, S. S., Yeo, G. W., Rahl, P. B., Flynn, R. A., . . . Sharp, P. A. (2008). Divergent transcription from active promoters. *Science, 322*(5909), 1849-1851. doi: 10.1126/science.1162253

Seitz, H. (2009). Redefining microRNA targets. *Curr Biol, 19*(10), 870-873. doi: 10.1016/j.cub.2009.03.059

Selbach, M., Schwanhausser, B., Thierfelder, N., Fang, Z., Khanin, R., & Rajewsky, N. (2008). Widespread changes in protein synthesis induced by microRNAs. *Nature, 455*(7209), 58-63. doi: 10.1038/nature07228

Sen, G. L., & Blau, H. M. (2005). Argonaute 2/RISC resides in sites of mammalian mRNA decay known as cytoplasmic bodies. *Nat Cell Biol, 7*(6), 633-636. doi: 10.1038/ncb1265

Shalek, A. K., Satija, R., Shuga, J., Trombetta, J. J., Gennert, D., Lu, D., . . . Regev, A. (2014). Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature, 510*(7505), 363-369. doi: 10.1038/nature13437

Shen, Y., Yue, F., McCleary, D. F., Ye, Z., Edsall, L., Kuan, S., . . . Ren, B. (2012). A map of the cis-regulatory sequences in the mouse genome. *Nature, 488*(7409), 116-120. doi: 10.1038/nature11243

Sheth, U., & Parker, R. (2003). Decapping and decay of messenger RNA occur in cytoplasmic processing bodies. *Science, 300*(5620), 805-808. doi: 10.1126/science.1082320

Shimizu, S., Takehara, T., Hikita, H., Kodama, T., Miyagi, T., Hosui, A., . . . Hayashi, N. (2010). The let-7 family of microRNAs inhibits Bcl-xL expression and potentiates sorafenib-induced apoptosis in human hepatocellular carcinoma. *J Hepatol, 52*(5), 698-704. doi: 10.1016/j.jhep.2009.12.024

Shin, C., Nam, J. W., Farh, K. K., Chiang, H. R., Shkumatava, A., & Bartel, D. P. (2010). Expanding the microRNA targeting code: functional sites with centered pairing. *Mol Cell, 38*(6), 789-802. doi: 10.1016/j.molcel.2010.06.005

Sigova, A. A., Mullen, A. C., Molinie, B., Gupta, S., Orlando, D. A., Guenther, M. G., . . . Young, R. A. (2013). Divergent transcription of long noncoding RNA/mRNA gene pairs in

embryonic stem cells. *Proc Natl Acad Sci U S A, 110*(8), 2876-2881. doi: 10.1073/pnas.1221904110

Simon, M. D., Wang, C. I., Kharchenko, P. V., West, J. A., Chapman, B. A., Alekseyenko, A. A., . . . Kingston, R. E. (2011). The genomic binding sites of a noncoding RNA. *Proc Natl Acad Sci U S A, 108*(51), 20497-20502. doi: 10.1073/pnas.1113536108

Siomi, M. C., Sato, K., Pezic, D., & Aravin, A. A. (2011). PIWI-interacting small RNAs: the vanguard of genome defence. *Nat Rev Mol Cell Biol, 12*(4), 246-258. doi: 10.1038/nrm3089

Skalsky, R. L., & Cullen, B. R. (2010). Viruses, microRNAs, and host interactions. *Annu Rev Microbiol, 64*, 123-141. doi: 10.1146/annurev.micro.112408.134243

Song, J. J., Smith, S. K., Hannon, G. J., & Joshua-Tor, L. (2004). Crystal structure of Argonaute and its implications for RISC slicer activity. *Science, 305*(5689), 1434-1437. doi: 10.1126/science.1102514

Song, L., & Crawford, G. E. (2010). DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb Protoc, 2010*(2), pdb prot5384. doi: 10.1101/pdb.prot5384

Stanek, D., Pridalova-Hnilicova, J., Novotny, I., Huranova, M., Blazikova, M., Wen, X., . . . Neugebauer, K. M. (2008). Spliceosomal small nuclear ribonucleoprotein particles repeatedly cycle through Cajal bodies. *Mol Biol Cell, 19*(6), 2534-2543. doi: 10.1091/mbc.E07-12-1259

Stormo, G.D., Schneider, T.D., Gold, I., & Ehrenfeuch, A. (1982). Use of the perceptron algorithm to distinguish translation initiation sites in E. coli. *Nucleic Acids Res, 10*, 2997-3011.

Sturm, M., Hackenberg, M., Langenberger, D., & Frishman, D. (2010). TargetSpy: a supervised machine learning approach for microRNA target prediction. *BMC Bioinformatics, 11*, 292. doi: 10.1186/1471-2105-11-292

Sullivan, C. S., Grundhoff, A. T., Tevethia, S., Pipas, J. M., & Ganem, D. (2005). SV40-encoded microRNAs regulate viral gene expression and reduce susceptibility to cytotoxic T cells. *Nature, 435*(7042), 682-686. doi: 10.1038/nature03576

Tolia, N. H., & Joshua-Tor, L. (2007). Slicer and the argonautes. *Nat Chem Biol, 3*(1), 36-43. doi: 10.1038/nchembio848

Vagin, V. V., Sigova, A., Li, C., Seitz, H., Gvozdev, V., & Zamore, P. D. (2006). A distinct small RNA pathway silences selfish genetic elements in the germline. *Science, 313*(5785), 320-324. doi: 10.1126/science.1129333

Vella, M. C., Choi, E. Y., Lin, S. Y., Reinert, K., & Slack, F. J. (2004). The C. elegans microRNA let-7 binds to imperfect let-7 complementary sites from the lin-41 3'UTR. *Genes Dev, 18*(2), 132-137. doi: 10.1101/gad.1165404

Vergoulis, T., Vlachos, I. S., Alexiou, P., Georgakilas, G., Maragkakis, M., Reczko, M., . . . Hatzigeorgiou, A. G. (2012). TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support. *Nucleic Acids Res, 40*(Database issue), D222-229. doi: 10.1093/nar/gkr1161

Visvanathan, J., Lee, S., Lee, B., Lee, J. W., & Lee, S. K. (2007). The microRNA miR-124 antagonizes the anti-neural REST/SCP1 pathway during embryonic CNS development. *Genes Dev, 21*(7), 744-749. doi: 10.1101/gad.1519107

Vlachos, I. S., Kostoulas, N., Vergoulis, T., Georgakilas, G., Reczko, M., Maragkakis, M., . . . Hatzigeorgiou, A. G. (2012). DIANA miRPath v.2.0: investigating the combinatorial effect of microRNAs in pathways. *Nucleic Acids Res, 40*(Web Server issue), W498-504. doi: 10.1093/nar/gks494

Vlachos, I. S., Paraskevopoulou, M. D., Karagkouni, D., Georgakilas, G., Vergoulis, T., Kanellos, I., . . . Hatzigeorgiou, A. G. (2015). DIANA-TarBase v7.0: indexing more than half a million experimentally supported miRNA:mRNA interactions. *Nucleic Acids Res, 43*(Database issue), D153-159. doi: 10.1093/nar/gku1215

Vlachos, I. S., Zagganas, K., Paraskevopoulou, M. D., Georgakilas, G., Karagkouni, D., Vergoulis, T., . . . Hatzigeorgiou, A. G. (2015). DIANA-miRPath v3.0: deciphering microRNA function with experimental support. *Nucleic Acids Res*. doi: 10.1093/nar/gkv403

Volinia, S., Galasso, M., Sana, M. E., Wise, T. F., Palatini, J., Huebner, K., & Croce, C. M. (2012). Breast cancer signatures for invasiveness and prognosis defined by deep sequencing of microRNA. *Proc Natl Acad Sci U S A, 109*(8), 3024-3029. doi: 10.1073/pnas.1200010109

Waldron, C., & Lacroute, F. (1975). Effect of growth rate on the amounts of ribosomal and transfer ribonucleic acids in yeast. *J Bacteriol, 122*(3), 855-865.

Wang, Y., Medvid, R., Melton, C., Jaenisch, R., & Blelloch, R. (2007). DGCR8 is essential for microRNA biogenesis and silencing of embryonic stem cell self-renewal. *Nat Genet, 39*(3), 380-385. doi: 10.1038/ng1969

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet, 10*(1), 57-63. doi: 10.1038/nrg2484

Willem, M., Garratt, A. N., Novak, B., Citron, M., Kaufmann, S., Rittger, A., . . . Haass, C. (2006). Control of peripheral nerve myelination by the beta-secretase BACE1. *Science, 314*(5799), 664-666. doi: 10.1126/science.1132341

Wu, T. D., & Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics, 26*(7), 873-881. doi: 10.1093/bioinformatics/btq057

Wu, X., & Sharp, P. A. (2013). Divergent transcription: a driving force for new gene origination? *Cell, 155*(5), 990-996. doi: 10.1016/j.cell.2013.10.048

Wu, X., & Watson, M. (2009). CORNA: testing gene lists for regulation by microRNAs. *Bioinformatics, 25*(6), 832-833. doi: 10.1093/bioinformatics/btp059

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*(Database issue), D105-110. doi: 10.1093/nar/gkn851

Xu, W., San Lucas, A., Wang, Z., & Liu, Y. (2014). Identifying microRNA targets in different gene regions. *BMC Bioinformatics, 15 Suppl 7*, S4. doi: 10.1186/1471-2105-15-S7-S4

Yanaihara, N., Caplen, N., Bowman, E., Seike, M., Kumamoto, K., Yi, M., . . . Harris, C. C. (2006). Unique microRNA molecular profiles in lung cancer diagnosis and prognosis. *Cancer Cell, 9*(3), 189-198. doi: 10.1016/j.ccr.2006.01.025

Yang, J. H., Li, J. H., Shao, P., Zhou, H., Chen, Y. Q., & Qu, L. H. (2011). starBase: a database for exploring microRNA-mRNA interaction maps from Argonaute CLIP-Seq and Degradome-Seq data. *Nucleic Acids Res, 39*(Database issue), D202-209. doi: 10.1093/nar/gkq1056

Yekta, S., Shih, I. H., & Bartel, D. P. (2004). MicroRNA-directed cleavage of HOXB8 mRNA. *Science, 304*(5670), 594-596. doi: 10.1126/science.1097434

Zang, C., Schones, D. E., Zeng, C., Cui, K., Zhao, K., & Peng, W. (2009). A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics, 25*(15), 1952-1958. doi: 10.1093/bioinformatics/btp340

Zeng, Y., & Cullen, B. R. (2005). Efficient processing of primary microRNA hairpins by Drosha requires flanking nonstructured RNA sequences. *J Biol Chem, 280*(30), 27595-27603. doi: 10.1074/jbc.M504714200

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., . . . Liu, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). *Genome Biol, 9*(9), R137. doi: 10.1186/gb-2008-9-9-r137

Zhou, X., Ruan, J., Wang, G., & Zhang, W. (2007). Characterization and identification of microRNA core promoters in four model species. *PLoS Comput Biol, 3*(3), e37. doi: 10.1371/journal.pcbi.0030037

Zhu, Y., Wang, D., Wang, F., Li, T., Dong, L., Liu, H., . . . Yu, J. (2013). A comprehensive analysis of GATA-1-regulated miRNAs reveals miR-23a to be a positive modulator of erythropoiesis. *Nucleic Acids Res, 41*(7), 4129-4143. doi: 10.1093/nar/gkt093