Full characterization of the small RNA transcriptome using novel computational methods for high-throughput sequencing data: study of miRNA variability in eukaryote organisms

Author: Lorena Pantano Rubiño

TESI DOCTORAL UPF / YEAR 2011

THESIS DIRECTORS

Dra. Eulàlia Martí

Genes and Disease Programme, Center for Genomic Regulation, CRG-UPF

Dr. Xavier Estivill

Genes and Disease Programme, Center for Genomic Regulation, CRG-UPF

A todas esas personas que luchan por lo que creen
To those people who fight for what they believe in

A mi hermana, siempre he estado a tu lado
To my sister, I have been always with you

Acknowledgment

First, sorry for any mis-spelling, bad translation to english and grammatical errors, because it is very difficult to write this section in english when the feelings are what you want to transmit. But I think that this section is equally or more important that the entire thesis, so it deserves to be in english, so any person can understand it. This section is always the last part to write, but for me it is the most important, because without those people who have been with me during these years, probably, I had not been able to finish it. I want to thank to one of my supervisors, **Xavier Estivill**, to give me an opportunity to grow as a scientist in one of the best center in Spain, and in the South of Europe. Thanks to give me a second opportunity when others had not done it. I will continue with my other supervisor, **Eulàlia Martí**, for everything she is. I can say many great things about her, but this section would be very long. She appeared in my life when I needed, she knew to listen to me, although I only was a student, she knew to understood me, some time this is the more difficult part, and she decided to helped me. I can say that she saved my PhD, and if I produce new scientific knowledges in the future, it will be because of her. She is my boss, my guide, my work-mate, my friend. I have to thank to my family, to my **father**, for teaching me something very value, and thanks to him I can say that I am very consequent with my acts, good and bad ones. Thanks to my **mother**, because she taught me a very difficult thing, emotional intelligence. She taught me to understand different points of view, to tolerate them, to think about them, and understand that with time everything is ok. She taught me to be strong, and fighter, to get everything with perseverance. She taught me to see what things are important in life, that is not to be rich and famous, so I can be very happy sitting in front of my computer with a cup of coffee, doing research, and then going with my friends to have dinner. I have to mention to my brothers, **Carlos and Francisco**, to always care of me, and show me that there are different ways to react and analyze the life. Because some times you have to know that there are things that you can not change, and you should accept and continue, and because some times, you have to fight for what you consider fair. Thanks to my sister (her memories), **María**, to give me strength in the worst moments, you will be always with me. Now, I want to do a parenthesis, and thanks to a other people who maybe are unknown by the majority of people. Thanks **Sergey Brin** and **Larry Page**, because they decided

one day that there were a better way to share information, to find what you are looking for; so they founded Google. Despite that any founder of a company does this for money, I think that the idea comes from a necessity. They have got that people find what they need, and thanks to that I can say that the 80 % of my knowledges come from them, and for that I have to mention it. Beside, I must thank to all the community that put its knowledges on the cloud, and share them with the rest of the world. With this, you can see that the human being, despite all bad things that we can do when we have power, also we can be generous and give the opportunity to people like me to finish a Master and PhD in Bioinformatics, coming from a Biochemistry degree. So, thanks to the **world** to show the more beautiful thing of the human being. I can not forget my work and master friends. **Albert** it is always a pleasure to discuss with you how to change science to be a better world. Thanks **Iñaki** to have always something to tell me about the last advance in technology. Always I will remember those master years, that they did us stronger. Thanks **Eneritz**, to support me in the bad moments, for all the alternative movies you introduced me, for the coffees at the terrace and to be beside me. Here I have to mention **Luci**, she always has a smile for everyone, and transmit a good mood although people do not want. Thanks **Elena, Eli**, to listen to me when I needed in complicate moments. Thanks **Mónica** to see always the positive part of everything. Finally, thanks **Sergi**, to be you. To continue fighting, to be always there, in good and bad moments. You are really good, and smart, the only thing you need it is a bit of encouragement like Eulàlia gave me. Also I want to mention to other part of my life, nothing related to science, but it has been very important during my life, and the last year. I want to thanks to the CB. Prat, my basketball team, to give me great moments in games, to make me forgetting bad moments during my thesis, and make me smiling with stupid things. **Rosa**, **Stefy**, **Puchi**, **Ana**, **Jas**, **Irune** y **Salom**, thanks to all of you. Thanks **Nuria** to be there like a sister from the beginning of the walk. Thanks coach, **Alex**, to trust me, and to teach me to fight until die for my ideals. Thanks **Alex-Jasmino** to support us, this requires a lot of female psychology!. Thanks **Lidia** to make those great meals with all the vitamins I needed. I want to finish with a special person, who makes me feel happy, who gave me strength in each of my weak moments, and thanks to her I survived this year. Thanks **Judith**, to be simply like you are, and trusted me at some point. Thanks for each moments you made me laugh when the only thing I wanted it was cry. When you know a person like you, everything is wroth. It is worth to continue trusting people, because you give hope.

Probably, I can not write all my gratitude to you here, but you know that my best way to talk to you is when I stare you. And now, thanks to **Bicho**, my little cat, to be always there when you noticed me tired and sad. Finally, thanks to all the other people I have not mentioned here, who gave good and bad moments, because everybody teaches something, that makes you stronger and a better person.

Agradecimientos

Siempre este apartado es el último en escribir, aunque realmente debería ser el primero, más que nada porque una tesis es el fruto del trabajo de todas las personas que te ayudan directamente y todas las que están junto a ti durante ese tiempo. Pero como sucede siempre, el tiempo pasa y dejas para el final este tipo de cosas porque dejamos de ver lo importantes que realmente son. Quiero empezar por agradecer a unos de mis dos directores de tesis, **Xavier Estivill**, por darme la oportunidad de haber podido madurar científicamente en unos de los mejores centro de España, y sur de Europa. Gracias por esa segunda oportunidad, cuando tal vez otros no la hubieran dado. Siguiendo en la linea, quiero agradecer a mi otra directora de tesis, **Eulàlia Martí**, por todo lo que ha sido. Puedo decir muchas cosas grandiosas de ella, pero no acabaría nunca. Se cruzó en mi vida cuando más lo necesitaba, supo escucharme a pesar de que solo era una estudiante, supo comprenderme, que a veces es lo más difícil, y sobre todo decidió apoyarme. Puedo decir abiertamente que salvó mi doctorado, y si en un futuro aporto nuevos conocimientos a la ciencia, será porque en el momento oportuno Eulália quiso ser mi directora. Es mi jefa, es mi guía, es mi compañera de trabajo, es mi amiga. Tengo que agradecer a mi familia, a mi **padre** por enseñarme algo muy valioso hoy en día, y es que gracias a él puedo decir que soy consecuente con todos mi actos, los buenos y los malos. Gracias a mi **madre**, porque me ha enseñado lo más difícil de enseñar, inteligencia emocional. Me ha enseñado como comprender diferentes puntos de vista, a tolerarlos, a reflexionar sobre cada una de las formas de ver las cosas, y comprender que solo el tiempo hace que todo acabe en su sitio. Me ha enseñado a ser fuerte, y luchadora, que con consistencia y tiempo todo se consigue. Me ha enseñado a valorar las cosas, y con ello poder ser feliz sin tener ni que ser rica ni famosa. Tengo que mencionar a mis hermanos, **Carlos y Francisco**, por cuidarme desde siempre, y mostrarme diferentes formas de reaccionar y analizar las cosas. Porque a veces hace falta saber que hay cosas que no se pueden cambiar y hay que seguir adelante, y porque a veces hay que luchar por lo que uno considera injusto. Gracias a mi hermana (al recuerdo de ella), **Maria**, por darme fuerzas en los momentos que más lo necesitaba. Ahora haré un pequeño paréntesis para poder agradecer a otro colectivo que no conozco directamente pero creo que se merecen ser mencionado, en cada tesis. Quiero agradecer a **Sergey Brin** y **Larry Page**, tal vez no muchos sepan quienes son, pero ellos decidieron un día que había una

forma mejor de compartir la información, de buscar lo que uno quiere encontrar, e inventaron **Google**. A pesar de que todo fundador de una empresa, crea una empresa por dinero, creo siempre que la idea viene de encontrar una necesidad. Y ellos han conseguido que las personas encuentren lo que necesitan. Y gracias a eso, puedo decir libremente que el 80 % de mis conocimientos de hoy en día viene del uso de ese buscador, con lo cual no puedo obviarlo en esta parte. Además de ellos, debo agradecer a todo una comunidad que ha querido compartir esos conocimientos para que cualquier persona del mundo pueda usarlos, porque aquí es cuando se ve que las personas, el ser humano, a parte de todo lo malo que puede ser cuando tiene poder, puede ser generoso y puede hacer que una persona como yo, habiendo me formado en Bioquímica, pueda superar un master y un doctorado en Bioinformática. Así que, gracias **Mundo** por mostrar la parte más bella del ser humano. No puedo olvidarme de mis compañeros de master y trabajo. **Albert**, siempre es un placer discutir contigo formas utópicas de convertir la ciencia. Gracias **Iñaki** por tener siempre algo que contarme sobre lo último en tecnología. Siempre recordare esos dos años de master, que mi hicieron cruzarme con vosotros, y que nos hicieron más fuertes. Gracias **Eneritz** por aguantar cada uno de mis momentos bajos, por todas esas películas diferentes, por todos esos cafes en la terraza, por estar junto a mi. Aquí tengo que mencionar a **Luci**, que siempre tiene una sonrisa, no importe el momento, y con ello te contagia de buen humor aunque no quieras. Gracias **Elena, Eli**, por apoyarme en muchos momentos complicados, por simplemente estar ahí para escucharme. Gracias **Mónica**, por hacer ver siempre lo positivo de todo. Por último, gracias **Sergi**, por ser tu. Por no dejar de luchar contra una tendencia, por permanecer siempre ahí, en los buenos y malos momentos. Eres bueno, eres listo, solo te hace falta un empujón, ese empujón que Eulàlia me dio a mi, y lo verás todo diferente. También quiero mencionar a otra parte de mi vida, que poco tiene que ver con ciencia, pero que ha sido muy importante durante mi vida y durante este último año. Quiero agradecer a mi equipo, CB Prat, por emocionarme en pista, por hacer que me olvidara del estrés del acabar una tesis, y hacerme reír con tonterías. **Rosa**, **Stefy**, **Puchi**, **Ana**, **Jas**, **Irune** y **Salom**, gracias a todas. Gracias **Nuria** por estar a mi lado como una hermana desde que llegue a esta ciudad. Gracias coach, **Alex**, por confiar en mi, y sobre todo enseñarme que hay que luchar hasta la muerte con tus ideales. Gracias **Alex-Jasmino** por aguantarnos, eso requiere mucha psicología femenina!. Gracias **Lidia** por esas comidas con tantas vitaminas .Quiero terminar con una persona muy especial,

que me ha hecho sentir bien conmigo misma, me ha dado fuerzas en cada uno de mis momentos de debilidad, y gracias a ella he podido sobrevivir a este año. Gracias, **Judith**, por simplemente ser como eres y confiar en mi en un momento dado. Gracias por cada momento que me has hecho reír cuando solo quería llorar. Cuando conoces a alguien como tu, todo merece la pena. Merece la pena seguir confiando en las personas, porque tu das esperanzas. Tal vez no pueda poner por escrito todo lo que te agradezco, pero sabes que mi mejor forma de hablarte es con la mirada. Y ahora si, gracias **Bicho**, mi gato, por estar siempre en mi regazo cada vez que me notabas cansada. Gracias a todas las demás personas que no menciono aquí, los que me han dado buenos y malos momentos, porque de todo se aprende en esta vida, y te hace más fuerte y mejor persona.

Abstract

In this thesis we have developed a user-friendly tool, SeqBuster, for the analysis of small RNA (sRNA) data generated by next generation sequencing strategies, with special emphasis on deep characterization of miRNA variants (isomiRs). We tested the tool using public datasets, revealing an unexpected amount of isomiRs in the total miRNA profile in different species. In addition, we detected all known classes of non-miRNA sRNAs and new sRNAs with a still unassigned function. Furthermore, we studied the implication of miRNAs and isomiRs in human brain development and aging and in Huntington disease, concluding that miRNAs/isomiRs may contribute to central nervous system physiological and pathological conditions. Overall, our results have uncovered a new layer of complexity in miRNAs, with probable consequences in mRNA mediated gene expression regulation underlying different biological functions. Furthermore SeqBuster may be extremely useful to identify sRNA sequences with a putative regulation role in selective biological processes.

Resumen

En esta tesis hemos desarrollado una herramienta, SeqBuster, para el análisis de datos de RNA (sRNA) de pequeño tamaño generados por las nuevas tecnologías de secuenciación, con especial énfasis en la caracterización de variantes de los miRNAs. Aplicamos la herramienta a datos públicos de secuenciación, lo que reveló una inesperada abundancia de isomiRs en diferentes especies. Ademas, detectamos todas las clases conocidas de otros sRNAs y de nuevos sRNAs con funciones desconocidas. También estudiamos la implicación de los miRNAs e isomiRs en el desarrollo y envejecimiento del cerebro humano, y en la enfermedad de Huntington. Nuestros resultados resaltan una posible importancia de la plasticidad de secuencia de los miRNAs, con probables consecuencias en la regulación de la expresión génica, subyacente a varias funciones biológicas. Por último, SeqBuster, podría ser extremadamente útil para identificar nuevos sRNAs con una posible función en determinados procesos biológicos.

Preface

Gene regulation is the process used by cells to control how the information on the genes is turned into proteins. From the beginning of the gene regulation characterization, the level of complexity, affecting the overall process, has increased enormously due to the discovery of novel mechanisms. One of the latest layers described to be involved in gene regulation is gene silencing by small non-coding RNAs (sRNA). sRNAs are RNA molecules 18-36 nucleotides long that target genes by sequence complementary, regulating the gene product formation. One of the best-known classes of sRNAs is the microRNA (miRNA) family (22 nt long). Lee *et al*, in 1993, discovered the first miRNA during a study in *C. elegans* development. Nowadays, this family has been found in all eukaryotic cells, except fungi, algae, and marine plants. Beside this class, other sRNAs has been described in recent years: small interfering RNAs (siRNAs) and (piRNAs). siRNAs come from double-strand RNA molecules and are involved in the RNA interference pathway originally discovered in plants. piRNAs (26-31 nt long) form RNA-protein complexes through interactions with piwi proteins that have been linked to transcriptional gene silencing of retrotransposons, particularly those in spermatogenesis. Recently, novel classes of sRNAs have emerged as a consequence of the advent of large scale sequencing technology that offers a full coverage detection of RNA and DNA molecules in cells. In addition to these novel sRNAs, these powerful strategies lead to the discovery of miRNA sequence variants, therefore increasing the complexity of non-coding transcriptome. However, very specialized tools are required for the processing and analysis of the data coming from high-throughput sequencing. As a consequence, in this thesis, we developed a complete set of tools for the characterization of the sRNA transcriptome, from miRNAs to the discovery of novel classes, in individual samples and case/control studies. By using these novel computational methods, we aimed to further study the miRNA variability and their functional implications in different species and tissues.

Contents

Figures Index xxii

Tables Index xxii

1. INTRODUCTION 1

1.1. Small RNAs . . . 3
1.1.1. MiRNAs . . . 4
1.1.2. piRNAs . . . 16
1.1.3. siRNAs . . . 20
1.1.4. Other small RNAs . . . 23
1.2. DNA sequencing . . . 27
1.2.1. Sanger method . . . 29
1.2.2. Massively Parallel Signature Sequencing (MPSS) . . . 29
1.2.3. Polony sequencing . . . 30
1.2.4. 454 / Life Sciences: Pyrosequencing method . . . 30
1.2.5. Solexa / Illumina: Reverse termination method . . . 32
1.2.6. ABI SOLiD: Sequencing by ligation method . . . 32
1.2.7. Helicos: sequency-by-synthesis . . . 33
1.2.8. Future perspective . . . 35
1.3. Application of sequencing . . . 38
1.4. Analysis tools . . . 42
1.4.1. Analysis dependent on reference genome . . . 43
1.4.2. *ab initio* analysis . . . 47
1.4.3. Deep sequencing impact in science . . . 51
1.5. Analyzing sRNA with sequencing . . . 53
1.5.1. MiRNA data evolution . . . 55

1.5.2. Other sRNAs data evolution 56
1.5.3. Tools for sRNA analysis 56
1.5.4. MiRNA analysis in diseases 59

2. RESULTS **65**

2.1. SeqBuster, a bioinformatic tool for the processing and analysis of small RNAs datasets, reveals ubiquitous miRNA modifications in human embryonic cells. 67
2.2. A non-biased framework for the annotation and classification of the non-miRNA small RNA transcriptome 101
2.3. A myriad of miRNA variants in control and Huntington's disease brain regions detected by massively parallel sequencing. 117
2.4. miRNA variants (IsomiRS) are functionally linked to biological processes in distinct species 164
2.4.1. IsomiRs in evolution 164
2.4.2. IsomiR biogenesis and expression 165
2.4.3. Expression pattern of the different sequences in each miRNA gene 167
2.4.4. General characterization 168
2.4.5. Conservation across tissues in human 172
2.4.6. Analyzing the functional role of isomiRs 175
2.4.7. Studying the miRNA-mRNA duplex structure . . . 178
2.4.8. Methods 181

3. DISCUSSION **199**

4. CONCLUSIONS **219**

List of Figures

1.1. MiRNA biogenesis 7
1.2. Type of miRNA target sites 10
1.3. Mechanisms of miRNA-mediated gene silencing in animals 13
1.4. The piRNA Ping-Pong Model 19
1.5. Roche-454 and Illumina-Solexa sequencing technologies 31
1.6. AB SOLiD sequencing technologies 34
1.7. Pacific Biosciences sequencing technology 36
1.8. Types of GEO datasets 39
1.9. Tools dedicated to each technology 43
1.10. Tools for each analysis 44
1.11. Algorithms for mapping 46
1.12. Assembly concepts 50
1.13. Impact in science 52
1.14. Flowchart of typical data-handling steps for small RNA . 54

2.1. SeqBuster and SeqCluster integration 108
2.2. usRNA definition 111
2.3. Output scheme. 113
2.4. General results of SeqCluster extension 115
2.5. IsomiR detection in different species 165
2.6. IsomiR at the 5-end and at the 3'-end 166
2.7. Expression correlation between isomiRs and reference miRNAs 169
2.8. Expression pattern of isomiRs/reference miRNAs 170
2.9. IsomiRs contribution to miRNA gene 171
2.10. IsomiRs relevance in each miRNA gene 173

2.11. Correlation between 5'-end and 3'-end variations in isomiRs 174
2.12. MiRNA expression changes in brain during life span . . 176
2.13. Overlapped genes between isomiR and reference miRNA targets 177
2.14. Gene and miRNA sequences expression profile 180

List of Tables

1.1. Technology benchmarking 28
1.2. miRNA tools 57

2.1. Comparison of mapping tools 109

1 Introduction

1.1. Small RNAs

It is widely recognized that eukaryote organisms utilize a wide range of regulation steps in the control of gene expression, at the epigenetic, transcriptional and post-transcriptional levels. These include DNA methylation, chromatin structure, transcription factors, mRNA splicing, and mechanisms of proteins localization, modification and degradation, among others. In this context, it is worth to mention that the majority of the genome is transcribed and that the biological complexity generally emerges from non-protein-coding region [187, 34]. This unanticipated level of complexity, detected thanks to the resolution of high-throughput sequencing technology has been named as 'pervasive' transcription. This term refers to the fact that the transcripts are not restricted to well-defined functional genes [125], but intergenic regions. Furthermore, these transcripts have been described as important regulators of gene expression, creating a complex network between elements in the same layer of the biology dogma: the RNA molecules [65, 140, 125]. The different types of RNAs are classified according to size and function (reviewd by Jacquier et al, see box 1). Small non-coding RNAs are functional RNA molecules smaller than 200 bases that are not translated into protein. Most of the ncRNAs identified in genomic transcriptome studies have not been studied and have yet to be ascribed any function. This category has been divided in different types according to size and function: involved in protein synthesis (tRNA,rRNA), involved in post-transcriptional modifications(small nuclear RNAs, small nucleolar RNAs) and regulatory RNAs (small interfering RNAs of 16-33 nt long) [188]. From here, the term small RNA (sRNA) will refer as RNAs of 16-33 nucleotides long. Inside the small interfering RNAs, microRNAs (miRNAs) have been well characterized. Piwi-RNAs (piRNAs) and endo-small interference RNAs(siRNAs) have been recently described, and other families remain still unclassified nowadays. During the next sections the different classes will be further addressed.

Box 1. Non-coding RNA classes adapted from Jacquier *et al*, 2009

TUFs A generic name for transcripts of unknown function.

Small RNAs (sRNAs) sRNAs are defined as any ncRNAs <200 nucleotides.

Small interfering RNA (miRNA, siRNA and piRNA) sRNAs are defined as any ncRNAs 18-36 nt long involved in a regulatory function.

Long RNAs (lncRNAs) lncRNAs are defined as any ncRNAs >200 nucleotides.

Long interspersed ncRNAs (lincRNAs) They derive from non-coding genomic regions that have transcription-dependent chromatin modifications over a distance of at least 5 kb.

Promoter-associated sRNAs (PASRs), promoter-associated lncRNAs (PALRs) and terminator-associated sRNAs (TASRs) PASRs are <200 nucleotides long; PALRs are >200 nucleotides long.

Transcription start site-associated RNAs (TSSa-RNAs) Small RNAs described in several mouse and human cell types. 20-90 nucleotides long.

Global run-on sequencing (GRO-seq) tags RNA tags generated from the human IMR90 cell line by GRO-seq, a methodology that reveals nascent transcripts. This methodology does not provide direct indications on the size of the RNA being transcribed.

Transcription-initiation RNAs (tiRNAs) Tiny RNAs (modal size of 18 nucleotides) that were identified from human cells, chicken embryos and several Drosophila melanogaster tissues by RNA-seq of gel-purified sRNA fractions.

Promoter upstream transcripts (PROMPTs) These unstable human transcripts are stabilized by the depletion of exosome factors in human HeLa cells. They are found on both strands, upstream of promoters.

Cryptic unstable transcripts (CUTs) Budding yeast unstable transcripts that are defined as RNAs that can be identified when nuclear exosome factors are mutated. They are principally found associated with promoters on both strands. 200-600 long.

1.1.1. MiRNAs

MiRNAs are regulatory sRNAs 22 nts in length that are bound by the miRNP protein complex [162, 107]. MiRNAs guide the complex to target

sites in the 3'-UTRs or, rarely, the coding sequence of mRNAs, causing mRNA degradation or translation inhibition [276, 201, 278, 19, 27, 94]. Thus, miRNAs reduce and/or buffer the expression of protein coding genes. All metazoan animals investigated have miRNA genes, ranging in number from 40 (sea anemone) to 700 (humans) [98, 97]. MiRNA genes appear to be constantly gained throughout evolution, thus some are deeply conserved and some are species-specific [207, 161, 153, 157]. Many miRNAs target hundreds of mRNAs, and it is estimated that between 30% and 60% of all metazoan protein coding genes are regulated by miRNAs in one or more cellular contexts [151, 88]. While miRNAs have been shown to be involved in most biological pathways or processes that are studied, they appear to be especially important in differentiation and in defining cell identity [172, 146, 46]. Consistent with this, many miRNAs appear to be expressed in distinct patterns in tissues in the metazoan body [155]. There are many examples of individual miRNAs that have strong impacts on development and phenotype, including the role of lin-4 in nematode embryogenesis [52], the implication of miR-430 in purging maternal transcripts from the zebrafish embryo [94] and even an example where a point mutation generates a miR-1 target site in the 3'-UTR of the myostatin mRNA, causing muscular hypertrophy of Texel sheep[92].

Biogenesis

Approximately 50% of mammalian miRNA loci are found in close proximity to other miRNAs. These clustered miRNAs are transcribed from a single polycistronic transcription unit (TU) [163], which can be generated from non-coding or portein-coding regions. Over 40% are located in the intronic region of non-coding transcripts, whereas 10% in the exonic region. Other 40% of miRNA loci are in protein-coding region, specifically in introns, and some 'mixed' miRNA genes can be assigned to intronic or exonic group depending on the alternative splicing pattern. The transcription of most miRNA genes is mediated by RNA polymerase II [164, 48], although a minor group associated to Alu

elements can be transcribed by PolII [38].

Drosha cleavage and nuclear export

Most miRNAs are transcribed by RNA polymerase II as long primary transcripts (pri-miRNAs) that are capped and polyadenylated and may be several kilobases in length [164, 48]. Each pri-miRNA contains one or more hairpin structures forming clusters that are recognized and cleaved by the Microprocessor complex while the transcript is still in the nucleus [163] (see figure 1.1). This complex consists of the Drosha endonuclease and the DGCR8 dsRNA binding protein (or Pasha), which is necessary for recognizing the hairpin structure [69, 83]. The usual hairpin structure consists in a stem of 33 bp, a terminal loop and flanking ssRNA segments. DGCR8 interacts with pri-miRNAs through the ssRNA segments and the stem, assisting Drosha to cleave the substrate 11 bp away from the ssRNA-dsRNA junction [108, 288]. This cleavage will determine one of the extremes of the miRNA. After the hairpin (also called the precursor miRNA or pre-miRNA) has been released from the pri-miRNA, it is exported to the cytosol by the Exportin-5 nuclear export protein [180, 36], previously assigned to the tRNA transport [36, 49].

Dicer cleavage

In the cytosol, the miRNA precursor hairpin is further recognized and cleaved by the endonuclease Dicer in complex with the TRBP dsRNA binding protein [33, 99, 122, 137, 148]. Before Dicer cleavage, the pre-miRNA hairpin is 70 nucleotides (nts) long and consists in a terminal loop flanked by two arms that form a stem. The stem does not contain bifurcations, but typically 20% of the nucleotides in the stem are not base paired and form bulges. The entire hairpin is energetically stable compared with other non-coding RNAs of comparable length, like rRNAs and tRNAs [37]. After the Dicer cleavage, three products are released: The loop and the two strands of the stem [228]. The loop is typically of length 10-40 nts long and is presumably rapidly degraded

by exonuclease action. The two strands of the stem, both 22 nts in length, remain bound to each other. Due to the endonuclease action, the two strands are offset thus that the duplex has 3'-overhangs two nucleotides in length in both ends of the duplex.

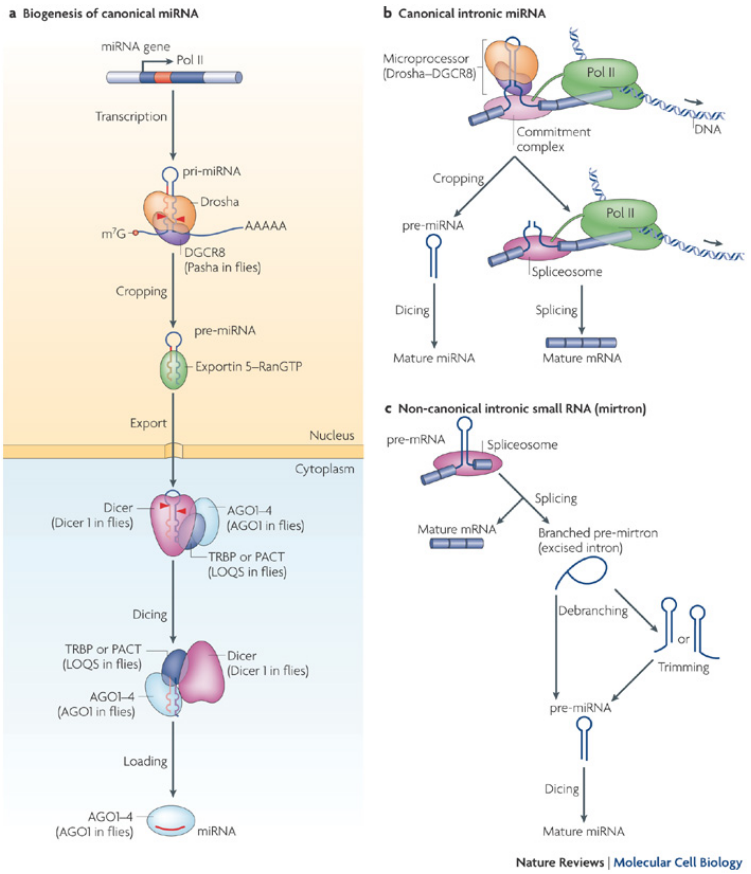


Figure 1.1: MiRNA biogenesis (Kim *et al*, 2009). a) Canonical microRNA (miRNA) genes are transcribed by RNA polymerase II to generate the primary transcripts (pri-miRNAs), which will be cropped by the Drosha generating 65 nucleotide (nt) pre-miRNAs, then recognized by the nuclear export factor exportin 5. RNase III Dicer catalyses the second processing (dicing) step to produce miRNA duplexes. Dicer, TRBP, and Argonaute mediate the processing of pre-miRNA and the assembly of the RISC in humans. b) Canonical intronic miRNAs are processed co-transcriptionally before splicing. The pre-miRNA enters the miRNA pathway. c) Non-canonical intronic sRNAs are produced from spliced introns and debranching, bypassing the Drosha-processing step.

IsomiRs generation

The advent of sequencing has permitted to elucidate a huge landscape of new miRNAs, to increase the knowledge of the biogenesis and to discover putative post-transcriptional editing processes in miRNAs ignored until now. These processes generate mainly, variation of the current miRNAs annotated in miRBase in the 3' and 5' terminus and in minor frequencies, nucleotide substitution along the miRNA length [77, 124, 205, 185]. The variations are mainly generated by a shift of Drosha and Dicer in the cleavage site, but also by nucleotide additions at the 3'-end [177], resulting new sequences different from the annotated miRNA and named isomiRs by Morin et al, 2008. IsomiRs have been well established along different species in metazoa [219, 179, 30, 102] and deeply described for first time in human stem cells and human brain samples [205, 185]. Moreover, isomiRs have been probed not to be caused by RNA degradation during sample preparation for next generation sequencing [160]. Some studies have tried to explain the miRNA diversity by structural bases of precursors but with not clear results [252]. The functionality of adenylation or uridynilation at the 3'end (3'addition isomiRs) has been related with alterations in the miRNA-3'-UTR stability [44]. Furthermore, isomiRs have been detected deregulated in *D. melanogaster* development elucidating a putative function in important process [85].

Incorporation into the miRNP effector complex

The duplex in then unwound, and typically one of the strands is selectively bound to the Argonaute protein in the miRNP (miRNA-containing ribonucleo-protein particles) effector complex while the other strand is degraded by the activity of an RNA helicase. The strand that is less tightly base paired in the 5'-end is more often incorporated into the effector complex [139, 238]. By definition, the strand that is more often incorporated is referred to as the 'mature' miRNA, while the strand that is more often degraded is the 'star' miRNA (sometimes these are referred to as the guide and passenger strands,respectively).

In practice, the distinction between the mature and star strands is blurry. For instance, the ratios of incorporated mature versus star strands can change during development in a given organism [199], and the ratios can change over evolutionary time, causing a reversal of the dominant strand [230]. Further, there is strong evidence that many miRNAs have mature and star sequences that are incorporated into the effector complex in comparable abundances and are both functional [199, 283].

Alternative routes into the miRNA biogenesis

Many miRNAs are derived from the introns of protein coding genes and may be co-transcribed with host genes [225]. However, the expression of these miRNAs do not always correlate with the expression of the host genes [230], suggesting that the miRNAs are themselves post-transcriptionally regulated. Recent studies show that some short (70 nts) introns can undergo Dicer processing and enter the miRNA pathway without previous Drosha processing by resolving the splicing process and releasing the introns which serve as miRNA precursors (mirtrons) [229, 200, 29]. Usually, these precursors need to be trimmed by exonucleases to remove extended tails at either the 5’ and 3’-end in order to become a substrate for nuclear export. Dicer-independent miRNA biogenesis that require AGO catalysis has been identified in mouse, wherein miR-451 is processed by Drosha, but its maturation does not require Dicer. Instead, the precursor becomes loaded into AGO and is cleaved to generate an intermediate 3’-end, which is then further trimmed and uridylated [55, 284, 59].

Target specificity

The miRNA-mRNA duplex structure

Once the miRNA is incorporated into the miRNP effector complex, it can direct the complex to target sites in the 3’-UTRs of mRNAs to degrade the mRNA or inhibit its translation.However, the target sequences

inserted into either coding or 5'-UTR sequences are also functional [147]. Unlike plant miRNAs, which are thought to bind to mRNA target sites with almost full complementarity [223], animal miRNAs have only partial complementarity to their target sites. It was noticed early that especially the 5'-end of miRNAs is important for the binding [154], in particular nucleotides 2-7 or 2-8 from the 5'-end, sometimes referred to as the 'seed' or the 'nucleus' region [11]. Other structures of the 3'-end of miRNAs are also involved in the duplex stability, giving a supplementary site in the case the seed is fully paired or giving a compensatory site when the 5'-end has mismatches or bulges [22] (see figure 1.2). It is true that the latter contributions to target recognition only have been proposed to affect the thermostability of the duplex rather than sensitivity.

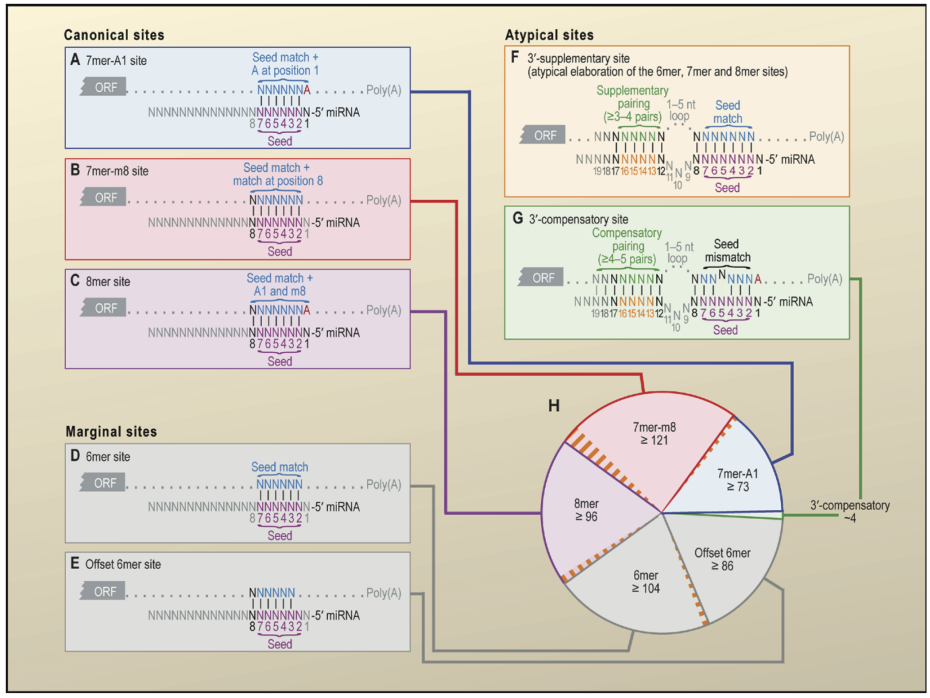


Figure 1.2: Type of miRNA target sites from Bartel *et al*, 2009. (A-C) canonical, 7-9 nt seed matched seed. (D-E) Marginal, 6 nt sites matching seed region. (F-G) Sites with productive 3' pairing. (H) Number of preferentially conserved mammalian sites matching a typical conserved miRNA. The orange-hatched subsectiors indicate the fraction of conserved sites with preferentially conserved 3' supplementari pairing.

The fate of the target mRNA is decided by the extent of base-pairing to the miRNA. It was believed that a miRNA directs destruction of the target mRNA if it has perfect or near-perfect complementarity to the target [123]. However, multiple evidences of mRNA repression instead of full degradation, with perfect target interaction indicate that the previous assumption is not correct. In these cases, an additional role has been proposed for the mRNA slicing process, related with the generation of sRNAs as it happens in the piRNAs biogenesis [42]. Nevertheless, it has been well described that central mismatches prevent slicing, which is consistent with the structural model [251]. Other factors that might influence in the way mRNA will be treated, are the type of RISC/AGO proteins, and even the protein composition of the target mRNA-protein particle [190]. Moreover, a recent study in human cells showed that transcript promoter determines the mechanism to be used for its silencing mechanism [149]. Finally, it seems clear that the presence of multiple, partially complementary sites in the target mRNA will predominately direct the inhibition of protein accumulation without strongly affecting mRNA levels [21].

Mechanisms of miRNA-mediated silincing

It is well know that miRNAs are involved in mRNA destabilization by two mechanisms: mRNA degradation and translational repression [120].

Translation repression

A fact that is relevant is that mRNAs are competent for translation if they posses a 5'-cap structure and a 3'-poly(A) tail, interacting with the factor 4G (eIF4G) which gives rise to circular mRNA that is efficiently translated and protected from degradation [70]. Another important protein in this process is the Cytoplasmic poly(A)-binding protein (PABPC), associated to the poly(A) tail of mRNAs. There is increasing evidence to suggest that animal miRNAs interfere with the function of the eIF4F complex and PABPC during mRNA stabilization. What remains unclear is how and when the interaction of miRNA-

mRNA occurs. Current studies support two main mechanisms: post-initiation repression or repression at the initiation stage. Supporting the post-initiation repression, experiments with *C. elegans* showed that lin-14 and lin-28 mRNA where detected in polysomes, suggesting that repression occurred after translation had been initiated [201, 240, 184, 196, 209]. Nottrott et al [196] proposed that the nascent polypeptide chain might be degraded co-translationally, while Petersen et al. [209] proposed that miRNAs cause ribosomes to dissociate prematurely. In contrast with this theory, other results suggested that miRNAs inhibit cap-dependent translation since mRNAs translated through cap-independent mechanism (IRES) were unsusceptible to repression by miRNAs [213]. These results argued that the silencing machinery targets the cap structure or interferes with the eIF4F complex, becoming the predominant mechanism when repression occurs.

Target degradation

Evidences of this process come from transcriptome studies showing that the abundance of miRNA targets inversely correlates with the level of miRNA [241, 102, 172, 84]. Although miRNAs can direct endonucleolytic cleavage of fully complementary targets [286], they rarely do so in animal cells, in which the vast majority of targets are partially complementary. In those cases the miRNA-mRNA duplex goes to the mRNA decay pathway, being the mRNA first deadenylated and then decapped by the enzyme DCP2 [278, 94]. *In vivo*, decapped mRNAs are ultimately degraded by the major cytoplasmatic 5-3 exonuclease XRN1 (see figure 1.3). Recent studies have reported that deadenylated mRNAs are not further degraded providing an alternative route for silencing by repression, where the mRNAs remain in cells after this step and not before [212]. However, this a debate that still exists nowadays regarding the order of events and it would need further investigation to elucidate the correct pathways of both functional roles, repression and degradation.

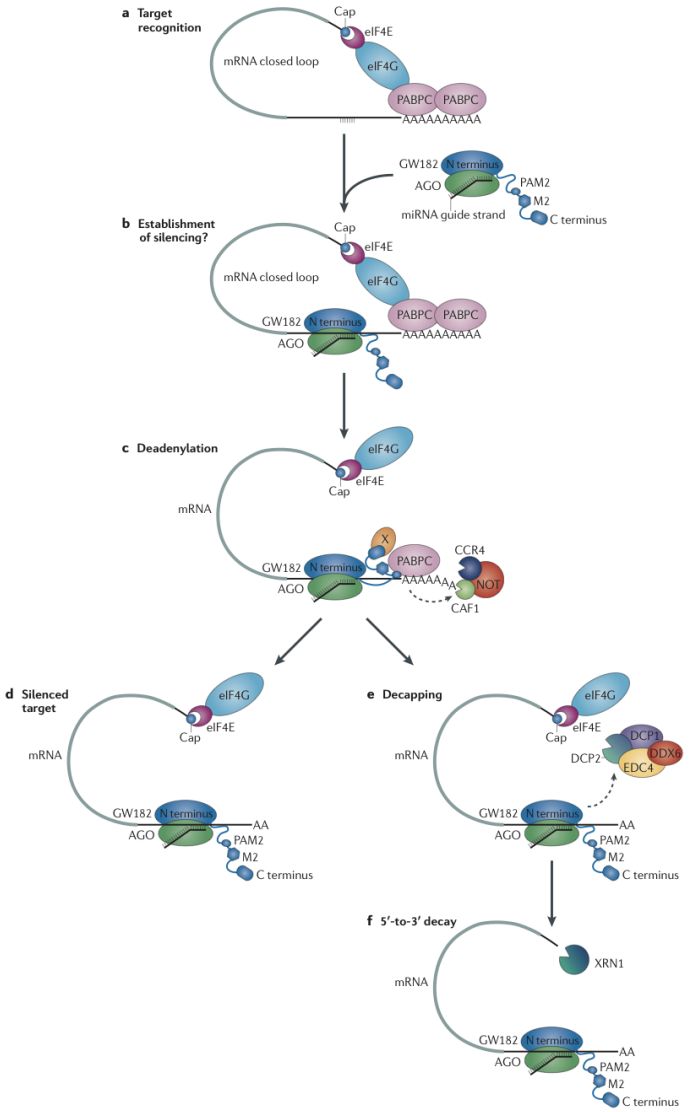


Figure 1.3: Mechanisms of miRNA-mediated gene silencing in animals (Huntzinger *et al*, 2011). a) A closed loop conformation between PABPC (poly-A binding protein) and 3' poly(A) tail and translational initiation factor (eIF4G) and cap-binding protein. b) The miRNAs loaded into AGO and coupled to GW182, a trinucleotide-repeat-containing protein, bound to the mRNA. c) Two binding site of GW182 interact with PABPC and this complex directs the deadenylation. d) Depending on the specific target, or cell type, the process goes to translational repression. e,f) In cell culture, deadenylated mRNA are decapped and rapidly degraded.

Functions

A description of miRNA biogenesis and mechanism of regulation does not confer what functions miRNAs have at the level of the cell, organism or evolution. Given that metazoans typically have hundreds of miRNA genes that together regulate 30-60% of all protein coding genes, and given the range of regulatory mechanisms available, it is difficult to make generalizations. However, a number of themes emerge from the literature, which are next summarized.

miRNAs as switches

There are some examples where miRNAs work to clear cells of transcripts from earlier development programs, enforcing a clean switch from one developmental stage to the next. In *C. elegans* the heterochronic gene lin-14 encodes a protein that is needed for the completion of the first larval stage (L1). However, unless the LIN-14 protein is depleted when the larva enters the second larval stage (L2), the first stage will be re-iterated [232]. The first miRNA described in any worm, lin-4, begins getting transcribed in the L1 to L2 transition and inhibits translation of the lin-14 mRNA by binding to seven target sites in the 3'-UTR [276, 162]. The switch function is clear: before the transition, lin-4 miRNA is absent and the LIN-14 protein is present; after the transition the reverse scenario occurs. In zebrafish, miR-430 begins getting transcribed as the zygote transits from maternal to zygotic transcription. The miRNA accelerates the degradation of hundreds of maternal transcripts [94]. This can be considered as a switch function, since the effect of miR-430 is to reduce target expression to zero. Zebrafish Dicer mutants have several defects during gastrulation and brain morphogenesis. Interestingly, injection of mature miR-430 rescues these brain defects [94]. In mice, microRNA-203 appears to signal the switch from proliferation to differentiation as skin develops in embryonic mice, rapidly upregulated to become the most abundant miRNA in the suprabasal layers of the epidermis [287].

miRNAs as expression regulatory tuners

That miRNA targets are only slightly downregulated have been supported by recent high-throughput proteomic studies [241, 18].This also holds for many target sites that are conserved, and therefore likely under positive selection. A possible explanation for this observation is that miRNAs may work as an extra layer of post-transcriptional regulation, fine-tuning the output from the transcriptional machinery. Mouse immunology can serve as a proof that fine-tuning of protein output can have a strong phenotypic effect. In mouse lymphocytes, miR-150 modulates the expression of c-Myb, which promotes B cell survival [280]. Ectopic expression of miR-150 has subtle effects on the levels of c-Myb protein (30% reduction). This modest reduction, however, has a dramatic impact on the number of B cells in the mouse (more than four-fold reduction). Overall, miRNAs may also impact the transcriptome through numerous 'soft' effects.

miRNAs as buffers

The buffering function refers to miRNAs that reduce the variance rather than the mean of gene expression. This function could theoretically help to make the expression of protein coding genes more robust and stable against stochastic fluctuations in transcription and translation efficiency and also against environmental influences. Such buffering could increase the connection between genotype and phenotype, and therefore increase heritability [116, 279, 211]. The miRNA buffering function finds theoretical support from network models, in which many miRNAs are predicted to interact with transcription factors and target genes in regulatory networks that would stabilize gene expression [116]. However, there is yet little solid evidence to support that miRNAs act as buffers [279]. One problem is that laboratory experiments are designed to minimize environmental influences that miRNAs should stabilize. Thus, experiments that simulate the stressful environment

of nature may reveal more differences between wild-type animals and Dicer knockout animals.

1.1.2. piRNAs

Piwi-interacting RNA (piRNA) is the largest class of sRNA molecules that is expressed in animal cells [242]. piRNA forms RNA-protein complexes through interactions with Piwi proteins which was shown to be essential for self-renewal of germeline stem cells [63, 64, 256]. These piRNA complexes have been linked to transcriptional gene silencing of retrotransposons and other genetic elements in germ line cells, particularly those in spermatogenesis. They are distinct from miRNA in size (26-31 nt rather than 21-24 nt), lack of sequence conservation, and increased complexity. It remains unclear how piRNAs are generated, but potential biogenesis pathways have been suggested, that clearly differ from those generating miRNAs and siRNAs (see below), although rasiRNAs (repeat associated siRNA) are a piRNA subspecies [145].

Proposed piRNA structure

piRNAs have been identified in both vertebrates and invertebrates, and although biogenesis and function mechanisms do vary somewhat between species, piRNAs have no clear secondary structure motifs. The length of a piRNA is, by definition, between 26 and 31 nucleotides, and the presence of a 5' uridine is common to piRNAs in both vertebrates and invertebrates. piRNAs in *C. elegans* have a 5' monophosphate and a 3'-modification that acts to block either the 2' or 3' oxygen by HEN1 methyltransferase [228], and this has also been confirmed to exist in fly [264], zebrafish [117], mice [142] and rats [117]. It is thought that there are many hundreds of thousands of different piRNA species found in mammals [68]. Thus far, over 50,000 unique piRNA sequences have been discovered in mice and more than 13,000 in *D. melanogaster* [173].

Location

piRNAs are found in clusters throughout the genome; these clusters may contain as few as ten or up to many thousands of piRNAs and can vary in size from one to one hundred kb [198]. While the clustering of piRNAs is highly conserved across species, the sequences are not [183]. D. melanogaster and vertebrate piRNAs have been located in areas lacking any protein coding genes [40], while in *C. elegans* have been identified amid protein coding genes [228]. In mammals, piRNAs are found only within the testes [117], with an estimated one million copies per cell in spermatocytes and spermatids [13]. In invertebrates, piRNAs have been detected in both the male and female germlines, but in no other cell types [117, 68]. At the cellular level, piRNAs have been found within both nuclei and cytoplasm, suggesting that piRNA pathways may function in both of these areas [145] and, therefore, may have multiple effects [231].

Biogenesis

The biogenesis of piRNAs is not yet fully understood, although possible mechanisms have been proposed. piRNAs show a significant strand bias, that is, they are derived from one strand of DNA only, and this may indicate that they are the product of long single stranded precursor molecules [242]. Two possible pathways have been proposed to be the mechanisms of biogenesis: First, a primary processing pathway is suggested to be the only pathway used to produce piRNAs in the pachytene stage in mice which are only associated to MILI and MIWI proteins, and related also with piRNAs derived from '*flamenco*' repeat family in fly. In this mechanism, piRNA precursors are transcribed resulting in piRNAs with marked strand asymmetry, as if they are processed from one or a few huge transcripts [14, 39]. And secondly, also a 'Ping Pong' mechanism is proposed wherein primary piRNAs recognize their complementary targets and cause the recruitment of Piwi proteins. This results in the cleavage of the transcript at a point ten nucleotides from the 5'-end of the primary piRNA, producing

the secondary piRNA [39] (see figure 1.4). These secondary piRNAs are targeted toward sequences that possess an adenine at the tenth position [14]. Since the piRNA involved in the ping pong cycle directs its attack on transposon transcripts, the ping pong cycle acts only at the level of transcription [183]. One or both of these mechanisms may be acting in different species, *C. elegans*, for instance does have piRNAs, but does not appear to use the ping pong mechanism at all [68]. A significant number of piRNAs identified in zebrafish and *D. melanogaster* contain adenine at their tenth position [145], and this has been interpreted as possible evidence of a conserved biosynthetic mechanism across species [81]. piRNAs are expressed through unique pathways, which are dissimilar to the expression pathways utilized by other sRNAs [117, 68]. However, evidence suggests that PIWI proteins must be present to stabilize piRNAs and, thus, facilitate their accumulation [68].

Function

The wide variation in piRNA sequences and PIWI function over species contributes to the difficulty in establishing the functionality of piRNAs [269]. However, like other sRNAs, piRNAs are thought to be involved in gene silencing, specifically the silencing of transposons. The majority of piRNAs are antisense to transposon sequences [183], suggesting that transposons are the piRNA target. In mammals it appears that the activity of piRNAs in transposon silencing is most important during the development of the embryo [14], and in both *C. elegans* and humans, piRNAs are necessary for spermatogenesis [269]. A recent work reported a widespread expression of a limited set of piRNAs in the hippocampus with a suggested role in spine morphogenesis [159]. piRNA has a role in RNA silencing via the formation of an RNA-induced silencing complex (RISC). piRNAs interact with Piwi proteins that are part of a family of Argonautes proteins. These are active in the testes of mammals and are required for germ-cell and stem-cell development in invertebrates. Three Piwi

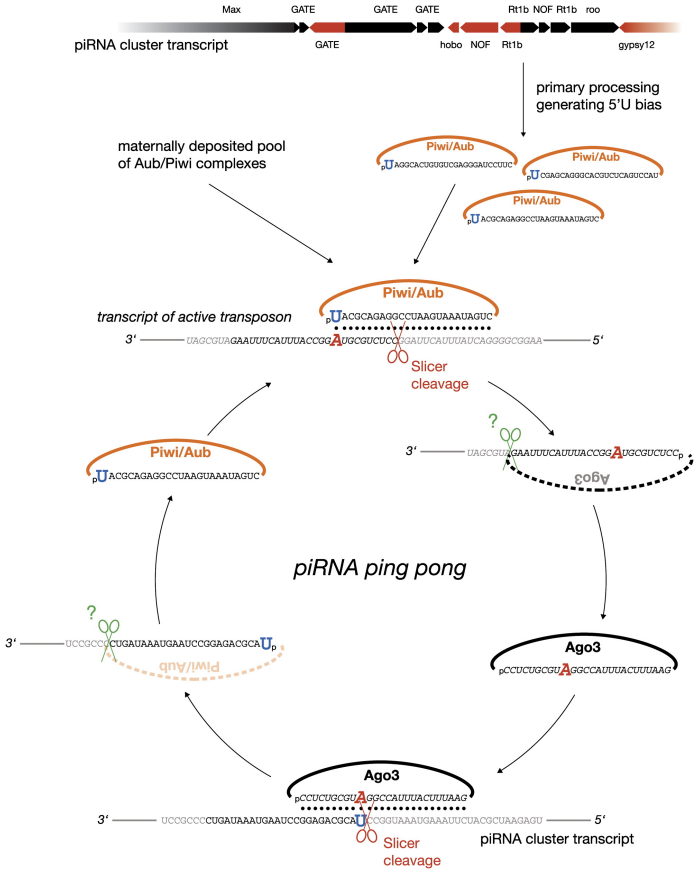


Figure 1.4: Amplification loop consisting of Piwi/Aub complexes, Ago3 complexes, piRNA cluster transcripts, and transcripts of active transposons (Brennecked *et al*, 2007). Nucleotide cleavage events are shown as scissors. Potential sources of primary piRNAs are piRNA cluster transcripts and maternally inherited piRNA complexes with a nt bias at position 1 of Uracil. The complex of Piwi/Aub and piRNA targets the transcript of active transposon and cleavages it at a point ten nucleotides from the 5'-end of the primary piRNA, producing the secondary piRNA, which is loaded into Ago3. This secondary piRNA pairs to a piRNA cluster transcript and cleavage at the Adenine nucleotide, generating a primary piRNA which will be load into Piwi/Aub proteins closing the cycle.

subfamily proteins - MIWI, MIWI2 and MILI - have been found to be essential for spermatogenesis in mice. piRNAs direct the Piwi proteins to their transposon targets. A decrease or absence of PIWI protein expression is correlated with an increased expression of transposons [14]. Transposons have a high potential to cause deleterious effect on their host, and, in fact, mutations in piRNA pathways are found to reduce fertility in *D. melanogaster* [40]. However, piRNA pathway mutations in mice do not demonstrate reduced fertility; this may indicate redundancies to the piRNA system [145]. Furthermore, it is thought that piRNA and endogenous small interfering RNA (endo-siRNA) may have comparable and even redundant functionality in transposon control in mammalian oocytes [183]. piRNAs appear to have an impact on particular methyltransferases that perform the methylations which are required to recognize and silence transposons [14], but this relationship is not well understood. piRNAs can be transmitted maternally [117], and based on research in *D. melanogaster*, piRNAs may be involved in maternally derived epigenetic effects. The activity of specific piRNAs in the epigenetic process also requires interactions between Piwi proteins and heterochromatin formation proteins, as well as other factors [173].

1.1.3. siRNAs

Small interfering RNA (siRNA), is a class of double-stranded RNA moleucles, 21 nucleotides in length, involved in the RNA interference pathway originally discovered in plants [106]. These RNAs are derived from transposon transcripts, sense-antisene pairs and long stem-loop structures [140], first described in plants and worms with a transposon silencing function. However, it is now well described how this type of sRNA is also generated in flies and mammals, and are involved in a wide range of pathways, such us, defense against viruses infection or gene regulation.

Structure

siRNAs have a well-defined structure: a short (usually 21-nt) double strand RNA (dsRNA) with 2-nt 3'-overhangs on each end being the result of processing by dicer. Each strand has a 5' phosphate group and a 3' hydroxyl (-OH) group.

Biogenesis

The biogenesis of siRNAs can be separated in two groups: RNA dependent RNA polymerases (RdRPs) pathways in worms and plants, and dsRNA derived siRNA in flies and mammals due to the lack of these proteins:

RdBP-dependent siRNAs

Worms and plants generate a huge landscape of endo-siRNAs with the help of RdRPs triggering dsRNAs from single-stranded RNAs. While plants require Dicer activity for the biogenesis of siRNAs, in worm, the processing is totally independent of Dicer. [65]. In both cases, the biogenesis process is considerably complex including several steps and factors. In *C. elegans*, primary siRNAs are products of long dsRNAs through the action of DCR-1 [99, 137, 255]. Then, siRNAs associate to RDE-1 (worm Argonaute protein) to guide it to the target transcripts recruiting an RdRP, which uses the target as a template for the synthesis of the secondary siRNAs, possesing a triphospate at the 5'-end [204, 117]. In plants, RdRPs convert ssRNAs precursors to dsRNA, which will be processed into siRNAs, classified in three major subclasses according to the Dicer family and AGO complex involved in their pathways: *trans-acting siRNAs* (ta-siRNAs), *natural antisense transcript-derived siRNAs* (nat-siRNAs), *heterochromatic siRNAs* (hc-siRNAs) (see Box 2) [65].

Box 2. Type of endo-siRNAs in plants

ta-siRNAs The process begins with miRNA-mediated cleavage of TAS1-3 non-coding RNA by miR390-AGO or miR1730-AGO1. After that, a RNA-dependent RNA polymerase (RDR6) synthesizes dsRNA using the cleavage site as entry point. Finally, 21-nt siRNA duplex are processed by Dicer-Like (DCL4) protein.

nat-siRNAs Generated using dsRNA transcripts from bidirectional transcription under biotic/abioteic stress. The proteins involved are DCL2 and DCL1, producing 24-nt or 21-nt siRNAs, respectively.

hc-siRNAs The precursors, in this case, are transposons and repeat elements processed by DCL3 and RDR2 proteins, triggering to 24-nt siRNAs.

siRNAs derived from dsRNAs

The generation of siRNAs from exogenous dsRNA is currently best understood in *D. melanogaster*, wherein the RNase III protein Dicer 2, with the help of the dsRBD co-factor and R2D2 [109, 291], cleaves exo-dsRNA sequently producing siRNA duplexes [164, 175]. However, siRNAs are also originated from endogenous dsRNA (endo-siRNA) [93, 199, 135, 58], involving a double-stranded RNA binding protein (R2D2, dsRBD) in their biogenesis [291, 93, 135, 58, 66]. This source can be sense-antisense transcripts pairs derived from transposons with natural mismatches and bulges. One example is endo-siRNAs targeting transposons, which originate from hybridization of transposon mRNAs with piRNAs cluster transcripts. Furthermore, single-stranded, but self-hybridizing, transcripts with long stem-loop structure also serve as precursors. In the same way, endo-siRNAs have been described in mammals, concretely in mouse oocytes and embryonic stem cells [262, 273, 16] using dsRNA as sources. Depending on the type of precursors, siRNAs can be *trans*-endo-siRNA, when the sense-antisense pair comes from different loci, and *cis*-endo-siRNA if the dsRNA is transcribed by convergent transcription of the same locus. An example of trasn-endo-siRNA are pseudogenes that anneal to their cognate functional transcripts generating siRNA with a role in the gene regulation [273, 262].

Function

siRNAs were first discovered as part of the post-transcriptional gene silencing in plants [106]. Since that, many studies are trying to elucidate the mechanism of action of siRNAs, enumerating a wide range of functional roles. In Arabidopsis, it was described a siRNA transposon silencing mechanism with the aim to control the mutagenic processes that are very high in this genome due to the high amount of repetitive elements [54, 105, 176, 218]. This mechanism has been also reported in *C. elegans* [138, 245, 257], fly [58] and hypothesized in mouse oocytes [140], and even with further consequence, in ciliates, where the transposons are completely removed from the genome [74]. In addition, Flies use siRNAs mechanism to defend against viruses that produce dsRNA during infection [265, 270]. The activity of siRNAs goes further, and affects the chromatin structure, modifying heterochromatin structure in pericentromeric regions of *S. pombe* yeast [222, 268]. This type of function has been also supported in the organization of centromeres of *C. elegans* [60, 101, 266]. Recently, plant siRNAs from introns has been associated to DNA methylation on their host genes [56]. Finally, siRNAs can also be exogenously (artificially) introduced into cells by various transfection methods to drive specific knockdown of a gene of interest. Essentially, any gene for which the sequence is known can be targeted based on sequence complementarity with an appropriately tailored siRNA. This has made siRNAs an important tool for discovering gene function and drug target validation studies in the post-genomic era [91].

1.1.4. Other small RNAs

Although the previous classes are the best characterized, there are a set of novel sRNAs that are being investigated nowadays, and have appeared as a consequence of the high resolution and capacity of new sequencing technologies (see high-throughput sequencing section for technical details). DeepBase is a database storing this information using 185 public datasets of different species and cell types [284]. As

the functional role remains unclear, these novel sRNAs are classified according their position in the genome and putative functions. We mention here according to the knowledge about them, in descending order: non-coding RNA (nasRNA), tiRNAs, splicing site derived sRNAs (spli-RNA), (gene termini associated human sRNA) tasiRNAs, tRNA derived sRNAs [9].

nasRNA or ncRNA-associated small RNAs

Small nucleolar RNAs (snoRNA) of 60-300 nucleotides long serve as guide for modification of selected ribosomal RNA nucelotides [144, 17]. Two main classes of snoRNAs have been described: C/D snoRNAs, which bind C/D snoRNP protein fribillarin, involved in methylation, and H/ACA snoRNAs, binding to RNP with pseudouridylation activities. Some of the features shared with miRNAs are: a) the high conservation through evolution [195], although the existence of species- and lineage specific snoRNAs and miRNAs have been also probed, and b) the procendence of snoRNAs from transposable elements [181, 274]. This has been also described for some miRNAs families [247], suggesting a widespread generation of these sRNAs by retroposition of existing RNAs, using long interspersed nuclear elements. In the last years, several studies have detected the biogenesis of sRNAs using these snoRNAs [239, 134, 259, 80] as precursors, with miRNA like functions. Indeed, a few miRNAs families are inside some of these snoRNAs, like let-7, mir-28, mir-16, mir-31 and mir-27b, supporting the hypothesis of a common functional role and evolution [239, 202]. An non-miRNA example is a sRNAs derived from snoRNA MBII-52 are involved in the regulation of serotonin alternative splicing [143].

tiRNA or transcription initiation derive sRNA

After a genome-wide profiling study in human embryonic stem cells, a population of sRNAs proceeding from almost all 5’ terminus of genes were reported. This analysis revealed that the majority of genes start their transcription, but after 70 nucleotide the process is stopped or

paused [130]. In a posterior study, to further investigate the previous results, sRNAs from human, chicken and drosophila were sequenced with Illumina and Roche FLX platform [258], detecting a peak of sRNAs, 18 nt in length, clustered downstream of transcription start sites and highly correlated to expressed genes and G+C-rich sequences. Experiments with Dicer-2 knockout indicated that their biogenesis is related to miRNA pathway. It was suggested that tiRNAs may be a product of backtracking given by RNA Pol II, which arrests +20 to +32 from TSS of certain promoters and then backtracks after encountering a nucleosome. However, a more important role may be assigned in gene regulation, wherein tiRNAs may act as a constitutive switch to counteract and silence possible anti-sense non-coding RNAs to turn on a gene [260]. As a final function, tiRNAs have been related to epigenetic marks since they have been localized in the nucleus, and strongly correlated to chromatin modification and gene transcription [261].

spli-RNA or splice site derived sRNA

A novel set of sRNAs were discovered in the nucleus of very distant metazoans, from worm to human, but not in yeast or plants [261] mapping precisely to the splice donor site of exon/intron boundaries. No evidence was found to connect the biogenesis with miRNAs or siRNAs pathways. Usually, spli-RNAs are correlated to constitutive splice sites, but a subset has been detected at alternative first exons.

tasi-RNAs or gene termini associated human RNAs

Kapranov et al [130] improved the methodology to prepare libraries for deep sequencing, avoiding biases introduced in the ligation and amplification steps. As a result, a novel sRNA family was discovered localized within 50 bp and antisense to the 3'-untranslated regions (UTRs) of annotated transcripts. It is suggested that these RNAs could potentially be produced by copying of polyadenylated mRNAs using an endogenous enzymatic activity similar to an RNA-dependent RNA polymerase starting from the poly(A) tail, supporting a novel RNA

copying mechanism and in concordance with non-expected antisense transcripts found in the CAGE analysis (cap-analysis of gene expression) [51].

tRNA derived sRNA

The fact that tRNAs lead to the generation of sRNAs arose in 2001, finding this class highly expressed in mouse [121]. Then they were detected in *G. lamblia* [170], and tripanosoma [89] as consequence of nutritional stress. Finally in human, two types were described, one localized in nucleus due to the action of RNaseZ and RNaseP, and the second localized in the cytoplasm with Dicer intervention [282, 110]. These sRNAs have preference for AGO3-4, and functional assays have shown a correlation to siRNA and miRNA silencing activities.

1.2. DNA sequencing

DNA sequencing refers to the methodology applied to determine the order of the nucleotides bases in a DNA molecule. The knowledge of DNA sequences has become indispensable for basic biological research, and numerous applied fields such as diagnosis, biotechnology, forensic biology and biological systematics. The advent of DNA sequencing has significantly accelerated biological research and discovery. With the huge evolution in this technology over history, the past years have witnessed of the deciphering of thousands of genomes from different species, including hundreds of human genomes.

As a result of a competition running in the 70s, after being recognized for the scientific community that sequencing would be the best way to investigate the life code, english and american teams developed in parallel two technologies [236, 189]. The english team, lead by Sanger F, developed the technology that was chosen for sequencing. A little more than 20 years later, a bioluminescent sequencing-by-synthesis approach was invented, called pyrosequencing [226]. Nowadays, the sequencing technology has done a huge jump to a wide genome vision generating millions of sequences in a single experiment. Well-known examples of such DNA sequencing methods include 454 pyrosequencing (introduced in 2005 and generating millions of 200-400bp reads in 2009), the Solexa system (introduced in 2006, generating hundreds of millions of 50-100bp reads in 2009) and the SOLiD system (introduced in 2007, generating billions of 50bp reads in 2009) (see table 1.1). These methods have reduced the cost from $0.01/base in 2004 to nearly $0.0001/base in 2006 and increased the sequencing capacity from 1,000,000 bases/machine/day in 2004 to more than 5,000,000,000 bases/run/day in 2010 [211]. The platforms differ in the technology used as well as in the statistics performance, and are briefly described in the following sections.

Table 1.1: Technology benchmarking

Platform	NGS chemistry	single reads (pb)	pair reads (pb)	Output (Gb)
Illumina HiSeq2	RTs	35	100	30-200
Illumina HiSeq1	RTs	35	100	13-100
Illumina GAII	RTs	35	150	25-95
Roche 454	PS	400	400	0.4
Roche junior	PS	400	400	0.04
ABI SOLiD 5500xl	SBL	75	60	20-30
ABI SOLiD 5500	PS	75	60	10-15

The benchmarking shows the difference between technologies in term of: chemistry (RT:reversible terminator; PS:pyrosequencing; SBL:sequencing by ligation), single reads length (pb), pair reads length (pb) and data produced (Gb).

1.2.1. Sanger method

Sanger method is based on the use of dideoxynucleotides in addition of the normal nucleotides, which substitute a OH group by an hydrogen in the 3' carbon [236]. With this modification the addition of further nucleotide in the reaction is backed due to the impossibility to form the phosphodiester bond with the next nucleotide. After the primer addition, which is labeled to be detected later in a gel, the sample is divided in four tubes, containing one of the four dideoxynucleotides (ddG, ddT, ddA, ddC) and the normal nucleotides. As the DNA is synthesized, nucleotides are added on to the growing chain by the DNA polymerase. However, on occasions a dideoxynucleotide is incorporated into the chain in place of a normal nucleotide, which results in a chain-terminating event. As a result, in each tube, a mixture of products is found with different sizes corresponding with the termination of the replication due to the addition of the ddNTP. For instance, in the tube with the ddGTP, all the products will end with ddGTP, indicating that at that position (the product size), in the source DNA there is a G nucleotide. With the information of all tubes, the source DNA sequence can be detected. With the advance of the technology, this method was automatized and improved with the labeling of the ddNTP with different fluorochromes, allowing the simplification and increasing the efficiency. In this content, Aplied Biosystem was the first company to develop the automatic machine in 1988 with a cost of $0.75/base.

1.2.2. Massively Parallel Signature Sequencing (MPSS)

The first of the 'next-generation' sequencing technologies, MPSS was developed in 1990s at Lynx Therapeutics, a company founded in 1992. MPSS was a bead-based method that used a complex approach of adapter ligation followed by adapter decoding, reading the sequence in increments of four nucleotides giving up to 1million of sequences per experiment [220]. Because the technology was so complex, MPSS was only performed in-house by Lynx Therapeutics and no machines were

sold. Later, they merged with Solexa which lead to the development of sequencing-by-synthesis, a more simple approach with numerous advantages, MPSS became obsolete.

1.2.3. Polony sequencing

It was developed in George Church's lab at Harvard, and was among the first next-generation sequencing systems used to sequence a full genome in 2005. It combined an *in vitro* paired-tag library with emulsion PCR, an automated microscope, and a ligation-based sequencing chemistry. It sequenced an E. coli genome at an accuracy of > 99.9999% and a cost of approximately 1/10th with respect that of Sanger sequencing. The Polony sequencing strategy was incorporated into the Applied Biosystems SOLiD platform [243].

1.2.4. 454 / Life Sciences: Pyrosequencing method

454 Life Sciences was founded originally as 454 Corporation and late in March 2007 was purchased by Roche Diagnostic 2007 [2, 28]. This technique, called pyrosequencing, is based on sequencing-by-synthesis [141]. The system relies on fragmentation of genomic DNA (300-800 base pairs) to be, later, ligated to short adaptors (see figure 1.5-A). These adapter-ligated fragments are captured by beads which are emulsified in a water-in-oil mixture with the amplification reagents, being each bead covered by millions of identical copies of the captured ligation product. Then, DNA-bound beads are placed into a fiber optic chip with a mix of enzymes such as DNA polymerase, ATP sulfurylase, and luciferase. Every time a given nucleotide is incorporated, light of a given wave length is emitted being the signal strength proportional to the number of the same nucleotides added. The light emissions are detected and translated into nucleotide sequences. This service is offered by the Genome Sequencer FLX System producing up to 1M reads of 400 bases length. A scaled version for individual labs, GS Junior, was released in 2010 generating up to 100K reads per run.

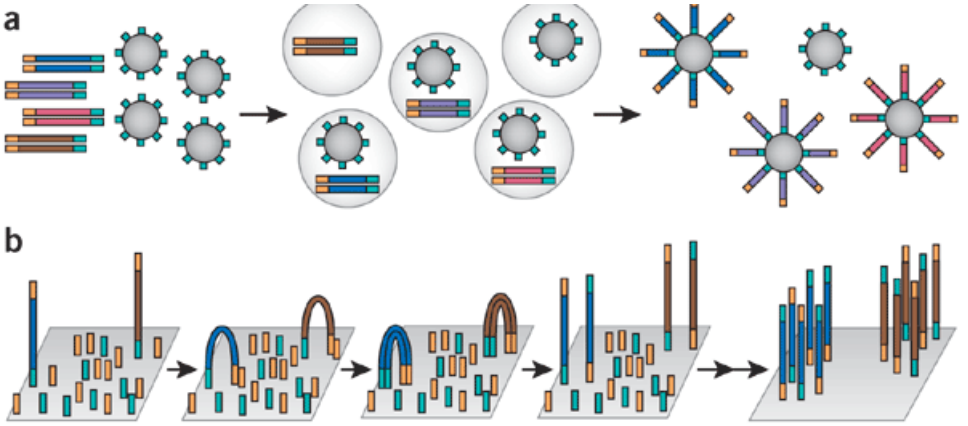


Figure 1.5: Next-generation sequencing technologies (Shendure *et al*,2005). (a) The 454 and the Polonator platforms rely on emulsion PCR to amplify clonal sequencing features. An in vitro-constructed adaptor-flanked shotgun library (shown as gold and turquoise adaptors flanking unique inserts) is PCR amplified in a water-in-oil emulsion. One of the PCR primers is tethered to the surface (5'-attached) of micron-scale beads. Bead compartments have 0/1 molecule. PCR amplicons are captured to the surface of the bead. (b) The Solexa technology relies on bridge PCR to amplify the template library. An in vitro-constructed adaptor-flanked shotgun library is PCR amplified, but both primers densely coat the surface of a solid substrate, attached at their 5'-ends. Amplification products originating from the template library remain locally tethered near the point of origin.

1.2.5. Solexa / Illumina: Reverse termination method

The first Solexa sequencer, the Genome Analyzer, was launched in 2006 and gave scientists the power to sequence 1G of bases in a single run. Solexa was acquired by Illumina in 2007 [73, 5]. In the Illumina's sequencing by synthesis (SBS) technology, first adapters are ligated to the fragmented DNA or cDNA, or to sRNAs (see figure 1.5-B). The ligation products are attached to the surface of a flow cell, to which PCR enzymes and nucleotides are added. Aside from the ligation products the flow cell is also covered by a dense lawn of primers that are complementary in sequence to the adapters. The adapters will bind to these, making each ligation product form a bridge over which amplification occurs. After numerous rounds of bridge amplification, the flow cell will be covered by millions of clusters, each containing, thousands of copies of one ligation product. Finally, enzymes and fluorescent labeled nucleotides are added and sequencing by synthesis takes place in each cluster on the flow cell. A laser excites the nucleotides that are incorporated in each cycle in each cluster, and the light emissions are translated into nucleotide sequences. Through refinements and optimization, the last generation of Illumina SBS technology based instruments generates 200G of bases per run, up to 25 Gb per day and reads with length of 75 nt. Nowadays, Illumina offers four different service according to the data production: MiSeq (up to 6.8M reads), Genome Analyzer (640G reads). HiSeq1000 (500G reads) and HiSeq2000 (2 billion reads).

1.2.6. ABI SOLiD: Sequencing by ligation method

SOLiD (Sequencing by Oligonucleotide Ligation and Detection) is a next-generation sequencing technology developed by Life Technologies and has been commercially available since 2008 [3, 217]. Adapters are ligated to the fragmented DNA or cDNA, or to sRNAs. Similar to the 454 platform, the ligation products are bound to beads and emulsion PCR reaction takes place in microreactors, resulting in beads covered by millions of copies of the same ligation product. The resulting

copies are then covalently bound to a glass slide, such that identical copies from one bead locate to one cluster on the slide. Then primers complementary to the adapter sequence are added and extended with di-base probes that compete for ligation to the primer. The di-base(two base at the same time) probes are fluorescently labeled and indicate the sequence of di-nucleotides of each cluster of identical ligation products on the glass slide. Accuracy is improved by implementing a two-base encoding system that leads to interrogation of each base twice. A sequencing run takes 6-10 days and the output is high, approximately 3-6 Gbp per run given a read length of 25-35 bases per clonally amplified bead. Further, the di-base color encoding makes it necessary to have dedicated computational tools for most downstream analysis (see figure 1.6).

1.2.7. Helicos: sequency-by-synthesis

The technology used was named True Single Molecule Sequencing (tSMS) [4], which enables the simultaneous sequencing of large numbers of strands of single DNA or RNA molecules by using a proprietary form of sequencing-by-synthesis in which labeled DNA bases are sequentially added to the nucleic acid templates captured on a flow cell. The most important characteristic of this technology is that the amplification step is removed from the workflow, decreasing the number of artifacts generated by this procedure. Billions of single DNA molecules are captured in two flow cells and serve as template for the sequencing-by-synthesis process. Addition of one of the labeled nucleotides create the nascent complementary DNA on all the templates. After a washing step, all the free nucleotides are removed, and by imaging, the positions with the nucleotide addition are recorded. In a further step, the fluorescent group is removed by cleavage, blocking the incorporation of new nucleotides. This step is then repeated with the other three bases. This results in multiple four-base cycles, generating complementary strands greater than 25 bases in length.

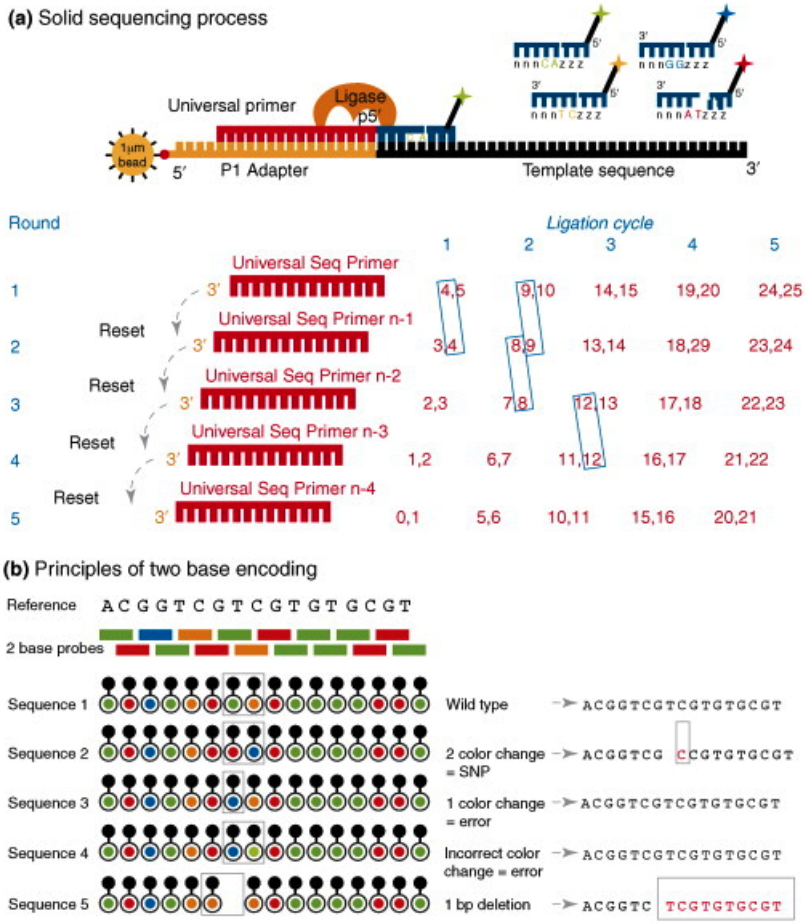


Figure 1.6: AB SOLiD sequencing technology (Mardis *et al*, 2008). (a) AB SOLiD sequencing by ligation first anneals a universal sequencing primer then goes through subsequent ligation of the appropriate labeled 8mer, followed by detection at each cycle. (b) Two base encoding of the AB SOLiD data greatly facilitates the discrimination of base calling errors from true polymorphisms or indel events.

1.2.8. Future perspective

Although sequencing progress in the last few years has resulted in a significant reduction of sequencing costs, it is still too early and too expensive to use these platforms to routinely sequence human genomes at a larger scale. In October 2006, the X Prize Foundation established an initiative to promote the development of full genome sequencing technologies, called the Archon X Prize, intending to award $10 million to the first team that can build a device and use it to sequence 100 human genomes within 10 days or less, with an accuracy of no more than one error in every 100,000 bases sequenced, with sequences accurately covering at least 98% of the genome, and at a recurring cost of no more than $10,000 (US) per genome . The competition is still running [6].

Pacific Biosciences

The technology behind Pacific Biosciences is called single molecule real time sequencing (SMRT) [8]. A proof-of-concept study showed read lengths of single DNA fragments over 1500 bases in 3000 parallel reactions. The heart of the technology is so called zero-mode waveguides (ZMW) [79] which essentially consists in nanometer scale wells with a diameter of 70 nm where a single DNA polymerase is immobilized (see figure 1.7). Nucleotides, fluorescently labeled at the terminal phosphate, are incorporated by the polymerase, exposing its base-specific fluorophore for a few milliseconds which is enough for detection. Benefits are long read lengths of thousands of bases in one stretch and high speed (10 bases per second and molecule), and the lack of amplification step. It is still at the proof-of-concept stage and no commercial instrument is ready.

Visigen Biotechnology

The platform consists of an engineered polymerase and modified nucleotides for single-molecule detection [7]. An immobilized polymerase

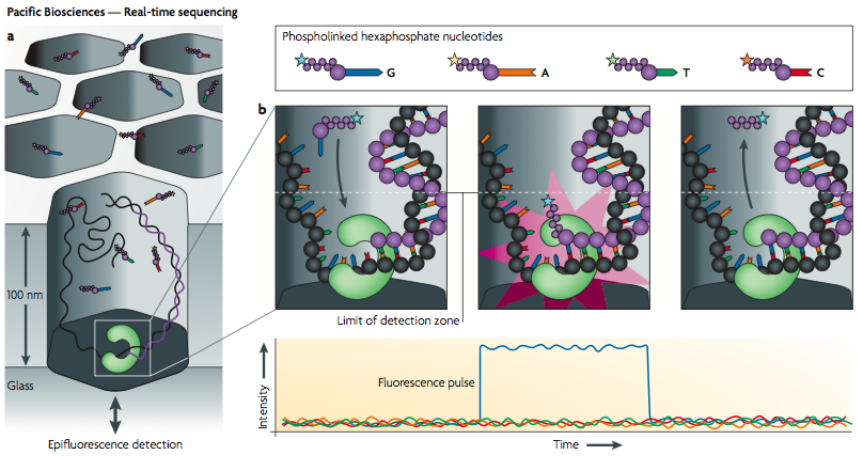


Figure 1.7: Pacific Biosciences' four-colour real-time sequencing method (Metzker *et al*, 2010). (a) The zero-mode waveguide (ZMW) design reduces the observation volume, therefore reducing the number of stray fluorescently labelled molecules that enter in the detection layer for a given period. (b) The residence time of phospholinked nucleotides in the active site is governed by the rate of catalysis (millisecond scale). This corresponds to a recorded fluorescence pulse, because only the bound, dye-labelled nucleotide occupies the ZMW detection zone. The released, dye-labelled pentaphosphate by-product quickly diffuses away, dropping the fluorescence signal to background levels. Translocation of the template marks the interphase period before binding and incorporation of the next incoming phospholinked nucleotide.

on a surface, modified with a fluorescence resonance energy transfer (FRET) donor incorporates nucleotides modified with different acceptors, allowing base-specific and real time detection of incorporation events. A theoretical throughput of 1 million bases per instrument second has been given, although no proof-of-concept study has been presented.

In advance

The future sequencing approaches are based on the physical recognition of nucleic bases. One alternative is nano pores, where the aim is to sequence a DNA strand that is pulled electrophoretically through a synthetic or natural pore, only 1.5 nm wide, measuring changes in conductivity [211], or microscopy-based techniques, such as AFM or electron microscopy that are used to identify the positions of individual nucleotides within long DNA fragments (>5,000 bp) by nucleotide labeling with heavier elements (e.g. halogens) for visual detection and recording [281].

1.3. Application of sequencing

Deep sequencing has numerous applications, not restricted to de novo genome sequencing, but also directed to solve different layers of compexity such as structural variations, SNP and transcriptome among others (see figure 1.8). When the analysis is designed to compare the data with a known genome ('re-sequencing'), the technology more used for this kind of strategies is the Illumina platform, due to the low cost and the amount of data produced by this technology, much better than its competitors. The unique handicap it would be the read size, however most possible 'words' of length >25 or 30 (illumina read size) only occur at most once even in relatively large genomes.

'*de novo* genome sequencing'. Novel genomes can be deeply sequenced, but given that the sequencing reads produced by deep sequencing are shorter than those produced by Sanger sequencing, assembly remains a challenge. The strategy to generate a de novo genome is the 454 technology, and even Sanger, to avoid problems generated by repeats region in the assembly process. When the genome is high complexity (no repeats), Solexa is also used.

'genome re-sequencing'. This is a term for the re-sequencing of genomes that have previously been sequenced and assembled. The Life Sciences/Roche company recently sequenced the genome of Dr. James Watson for less than 1.5 million dollars using only the 454 deep sequencing platform [275]. For small and low complex genomes, like microbes, Solexa technology is the chosen by the community for the low-cost and amount data produced. A deeper characterization could be extracted from DNA sequencing data, as SNP discovery, InDels (small insertions or deletions) or structural variants (insertions, deletions, copy number variants, inversions or translocations) when comparing to the reference genome.

'PEM' (paired-end mapping), refers to the sequencing of short sequences at the 5'- and 3'-ends of a DNA fragment of different size, which will determine the sensitivity of the analysis and also the library preparation. The first time, the method was used to sequence 5'/3'-ends of 3 kb fragment of two hapmap project individuals, detecting an unexpected high number of variation in the genome [150] by comparing the real distance between the 5'/3' tags mapped to a reference genome and the expected, in this case 3kb. A wide range of structural variants can be detected, from insertions (smaller than the fragment size), and deletions, to inversions and translocations. The method has evolved to sequence the 5'/3'-end of fragments of 200-400 pb, increasing the resolution and power of the structural variants detection. Nowadays, we can differentiate between mate-pairs or pair-ends strategies, depending of the origin of the 5'/3'-ends. Mate-pair protocol involves fragmentation, circularization, pull down the ends junction and sequencing, usually with a fragment size from 3kb and 40 kbs. In paired-end, the protocol is only based on fragmentation generating reads of 100-400 nucleotides, and directly sequencing the extremes.

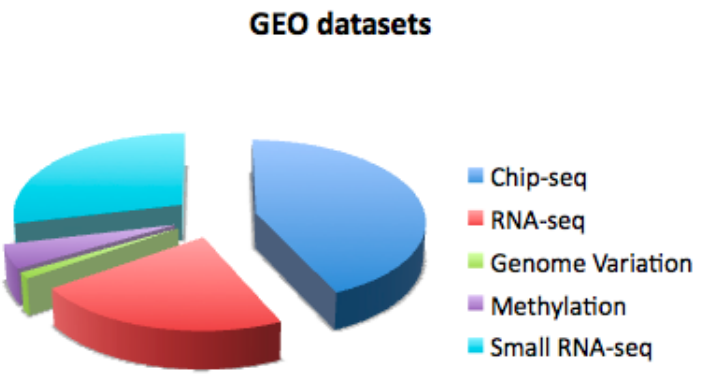


Figure 1.8: Different types of sequencing experiments. The majority of the datasets corresponds to chip-seq and RNA-seq. Source: GEO database.

'BS-Seq' (bisulfite sequencing). This is a method to selectively sequence the parts of the genome that are DNA methylated. Using a chemical reaction of sodium bisulfite, all unmethylated Cytosine nucleotides are substituted by Uraciles generating Tymine nucleotides when sequencing. In this way, comparing DNA treated and no-treated, the pattern of C methylated nucleotide are obtained. It is one way in which deep sequencing can survey epigenetic information [214].

'RE-Seq' is used in different ways for the detection of chromatin accessibility and DNA methylation using DNase I or methylated restriction enzymes, respectively. The digested DNA is then sequenced in order to generate tags of their target sequences for posterior annotation by deep sequencing and mapping. This will lead to the production of a genome map of the corresponding target enzyme.

'CHIP-seq'. This method uses immunoprecipitation to pull-down transcription factors or any enzyme and subsequent sequencing of the DNA that they bind to (reviewed in [206]). This strategy generates a huge map of possibilities in the study of motif-TF association. Furthermore, this technology is becoming the gold standard platform in epigenomics to detect DNA methylation (MeDIP-Seq) or chromatin modification by the use of different anti-bodies with high affinity to methylation or modified histones at different residues [32].

'RNA-seq'. Deep sequencing of mRNAs generates several levels of information [234]. First, the number of times a mRNA is sequenced correlates well with transcript abundances as estimated from qPCR [193]. Compared with arrays, this 'digital gene expression' is unbiased since it does not depend on pre-spotted probes on an array. Second, when exon-exon junctions are sequenced, information on splice variants and even gene fusion events is also yielded [131, 182]. Third, sequence information such as SNPs or RNA editing can also be obtained (reviewed in [271]). Frequently, in mRNA analysis the 3' polyadenylated (poly(A)) tail is targeted in order to ensure that coding RNA is separated

from noncoding RNA. This can be accomplished simply with poly (T) oligos covalently attached to a given substrate. Also, since ribosomal RNA represents over 90% of the RNA within a given cell, studies have shown that its removal via probe hybridization increases the capacity to retrieve data from the remaining portion of the transcriptome. Another key consideration concerning library construction is whether or not to prepare strand-specific libraries, as has been done in some studies [61]. These libraries have the advantage of yielding information about the orientation of transcripts, which is valuable for transcriptome annotation, especially for regions with overlapping transcription from opposite directions. In addition this strategy permits non-coding RNA gene discovery, which would have gone unnoticed unless the library preparation includes other steps to avoid it.

'CLIP-seq'. Similar to CHIP-seq, and also called RIP-Seq or HITS-CLIP, but the method uses pull-down of RNA binding proteins cross-linked to RNA [277] allowing advent in miRNA-mRNA duplex in AGO [57, 293], studying new components in splicing events [235] or describing alternative RNA processing [171]. A modified method, only directed to identifie the binding sites of cellular RNA-binding proteins (RBPs) was developed in 2010, and defined as PAR-CLIP [104].

'GRO-seq', generates cDNA tags extended from nascent transcripts synthesized in vitro from isolated human nuclei allowing the mapping of elongating RNA polymerase II [62].

'SAGE', developed in 1995, was the first technique used to analyze transcriptome in an unbiased way. With a tipe II restriction enzyme, small cDNA tags are generated to be sequenced posteriorly [267]. Currently the method has evolved to SuperSAGE, wherein tags are longer, increasing specificity, and the transcriptome coverage is high due to the advance in sequencing [186].

'CAGE' is a technique used to produce a snapshot of the 5'-end of the messenger RNA population. The small fragments (usually 20-21 nucleotides long) from the very beginning of mRNAs (5'-ends of capped transcripts) are extracted, reverse-transcribed to DNA, PCR amplified and sequenced [244].

'sRNA-seq'. Deep sequencing allows the sequencing of millions of sRNAs in a sample [191]. The library construction only requires simple steps, like RNA extraction, sRNA selection and sequencing after adapter ligation. This has enabled the discovery of sRNAs that were previously below the detection limits, such us, siRNAs or sRNA populations that have a high degree of sequence diversity, like the piRNAs.

1.4. Analysis tools

The introduction of next-generation sequencing technologies has produced a huge impact upon genomics and functional genomics. Indeed these methods are rapidly supplanting the conventional Sanger strategy [236] that has been the principal method of sequencing DNA since its inception in the late 1970s. Currently available next-generation sequencers rely on a variety of different chemistries to generate data and produce reads of differing lengths, but all are massively parallel in nature and present new challenges in terms of bioinformatics support required to maximize their experimental potential.

Given the rate of development in this field that follows the Moore's law distribution (exponential growing with time) only the most used tools will be highlighted for each sequencing application. In general, two major groups can be named: reference genome-dependent analysis and *ab initio* analysis (from scratch).

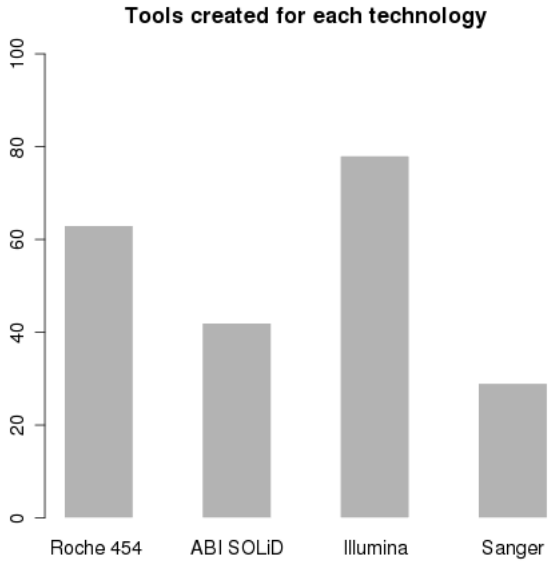


Figure 1.9: Number of tools developed to accept files from each technology. Source: http://seqswers.com/wiki:Special:BrowseData.

1.4.1. Analysis dependent on reference genome

The availability of reference sequences or genomes and transcriptomes are the base to approach importnt biological questions using high-throughput sequencing data. These include structural variation or novel alternative splicing events, and the characterization of sRNA transcriptome, among others. In this contest, 90% of the cases are using the Illumina platform due to the low cost offered and the high amount of data produced. At the beginning of this revolution, the need of a tool for mapping billions of reads in a memory and in time efficient way was the first challenge to abroad. In addition, mapping information (reads location within the reference genome) is required for the subsequent analysis. These steps are very dependent on the analysis goal. Excepting de novo sequencing, the rest of the applications mentioned in previous sections, require this information to begin with the data inspection (see figure 1.10 for a view of the number

of tools developed for each analysis).

Mapping reads is a distinct demonstration of sequence alignment, the oldest bioinformatics problem. Classical methods, like Smith-Waterman dynamic [249], indexing of longer k-mers (sequences of 7-24 nt) as BLAT [136], or combinations of the two (BLAST) [10] are not appropriate to the alignment of very large set of short sequences to a reference genome [263]. As a result, many methods are based on the similar principles and algorithms, but differ in the 'programming tricks' or the heuristics used to increase speed at the price of minimal loss of accuracy. Research in this field is growing weekly with new, or modified mapping current tools [23, 1]. Thus, I will concentrate only on a description of the general principles underlying the most successful algorithms, and very brief descriptions of a few of them.

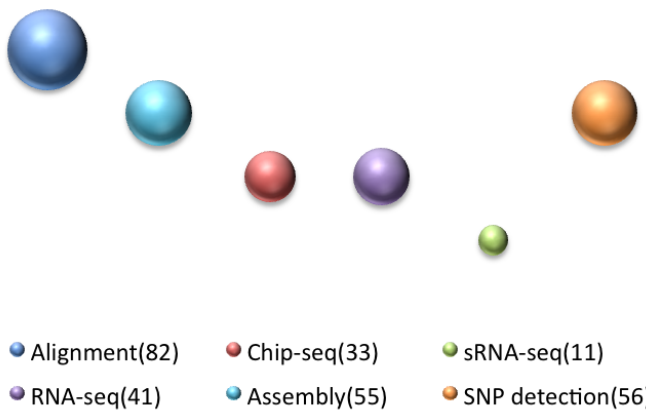


Figure 1.10: Visualization of the number of tools developed for each analysis. Source: http://seqswers.com/wiki:Special:BrowseData.

The majority of short read mapping tools underlie the principle of creating index of positions for all distinct k-mers (fragments of k nucleotides long) either the sequence reads, or the reference sequences, reviewed in [115]. The choice of which dataset is indexed can have significant implication in efficiency, meaning more memory requirements, and more amount of data. In small reference sequences,

the more efficient process is the indexation of the genome, and subsequent matching of the reads. The reason why most of the tools are choosing indexing the data reads, and matching in the reference sequences, is the easy scalability of the process: if the available memory is not enough, then the reads can be split into subsets and each subset to be indexed separately or in parallel in several computers. While the computation time is increased, tools with this strategy can be run in personal computers, or in a Bioinfromatics group limited to the standard hardware capacity. The most fundamental differences between available algorithms are whether the reference or the reads are indexed, and the method applied for that purpose. Another critical point to take into consideration for tool selection is the reporting of only unique best matches or of all matches. As mentioned before, other heuristic approaches have been implemented to speed up the analysis, such as the quality score information. For instance, reads with low quality can be removed, or only mismatches in less reliable nucleotides are allowed. Furthermore, due to the quality decrease to the 3'-end of the reads, some tools permit more mismatches at the end, or permit the trimming of the end of the sequences to find a better alignment. Also, specific strategies are needed for mapping spliced transcript sequences to genome sequences in case of working with RNA data. The performance of the different methods can be measured according to different parameters: time required, memory occupation, disk space and in the case of heuristic tools, the actual number of reads that have been assigned correctly to their original position on the genome. The choice of a given method depends on how many tags need to be mapped, the sensitivity of the alignment, and the specifications of the computing equipment available.

The first program developed to map massively sequences was ELAND, a commercial aligner for the Illumina platform, provided free for research groups that acquire the sequencer. The algorithm is based on the k-mers strategy allowing mismatches and still nowadays, is the fastest

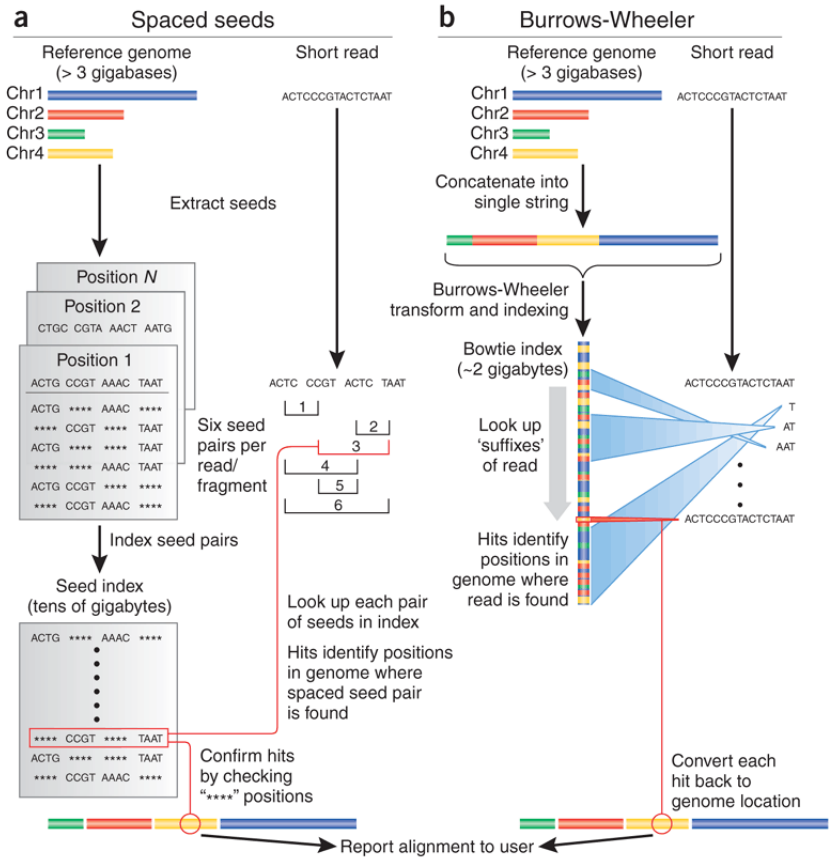


Figure 1.11: Algorithm for mapping (Trapnell and Salzberg, 2009). (a) Algorithms based on spaced-seed index the reads as follows: each position in the reference is cut into equal-sized pieces, called 'seeds' and these seeds are paired and stored in a lookup table. Each read is also cut up according to this scheme, and pairs of seeds are used as keys to look up matching positions in the reference. (b) Algorithms based on the Burrows-Wheeler transform store a memory-efficient representation of the reference genome. Reads are aligned character by character from right to left against the transformed string. With each new character, the algorithm updates an interval (indicated by blue 'beams') in the transformed string. When all characters in the read have been processed, alignments are represented by any positions within the interval.

and less memory-greedy tool, however it is not the more sensitive and specific. After that, research started to develop better algorithms to improve the deficiencies of ELAND. Some examples are SeqMap [127] which allows insertions and deletions; ZOOM [173], that introduced the term spaced seed for the indexing and is faster than ELAND; and finally, SOAP [168] converting reads and genome to numbers using 2-bits-per-base encoding (see figure 1.11). Other programs added quality information to the alignment algorithm, like MAQ [166] which is very fast but not always guarantees the best match for a read, and RMAP [248], wherein low-quality reads are removed, and nucleotides below a quality threshold induce always match. In the posterior years, this research area is suffering a boom of new tools with modifications to improve the time consuming of the process or the performance. PASS [50] or MOM [75], although the quality is better, it sacrifices the memory requirement, being up to 10 GB for the human genome. Novel algorithms are being integrated to decrease this barrier, like SOAP2 [169] or Bowtie [156] which employ a Burrows-Wheeler [45] index, reporting a memory requirement of only 1.3 GB for the human genome (see figure 1.11), but still not covering the best quality scenario since, if the best match is inexact, may be not called. Nevertheless, the differences observed between algorithms illustrate that, particularly when errors are introduced, a substantial number of artifactual placements are generated (mostly due to the presence of sequencing errors) and that the different heuristics used by diverse algorithms can find different imperfectly matching map positions [115].

1.4.2. *ab initio* analysis

The reconstruction of a genome only using information offered by the data produce by the sequencer is defined, mathematically, as a class of problem ('NP-hard') for which no efficient computational solution is known [90]. Scenically, the overlaping step is the most computational time-intensive component in a de novo assembly pipeline. This step generates bigger fragments called contigs by partial-shifted

overlapping between reads. This means that, the prefix of one read overlaps with the suffix of another , see figure 1.12. Because of this, short reads have a significant impact in the complexity of this step, leading into efficiency issues. The complexity worsens if there is no mate-pairs information and repeat regions are abundant and large in the genome [215]. Due to these challenges, short-read sequencing technologies (Solexa, SOLiD) have primarily been used in re-sequencing applications. De novo assembly of these data has largely been restricted to bacterial genomes, though an assembly of an entire human genome from Solexa reads was recently reported [246]. This approach is, therefore, better suited for long reads produced by Sanger method or 454 technology [15, 210, 221] or even for combinations of data from multiple sequencing technologies.

Greedy

Greedy algorithms represent the simplest, most intuitive, solution to the assembly problem. Individual reads are joined together into contigs in an iterative fashion, starting with the reads that overlap best, and ending once no more reads or contigs can be joined. The parameters in this case are the length and the quality of the overlapping, which limit the accuracy of the algorithm. For instance, processing always the best quality first, may generate misassemble repeats. Software, such as, phrap, TIGR, ATLAS, PCAP, CAP3 and Phusion [111, 119, 192] have implemented this type of algorithm to assembly Sanger data. However recent software (SSAKE, VCAKE and SHARCGS) [272, 126, 71], use a different strategy, wherein a read is chosen to start a contig, being extended by overlapping on its 3'-end until no more extensions are found. The process is repeated in the 5'-end using the reverse complement of the contig. With this strategy, more quality or high coverage reads are first considered.

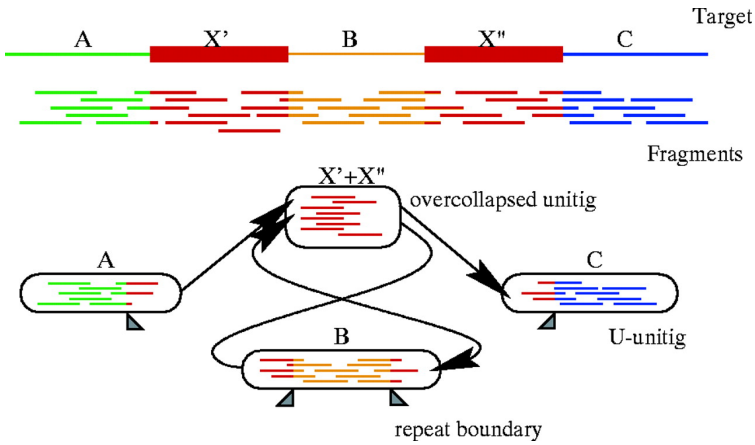
Overlap-layout-consensus or OLC

This strategy breaks down the assembly into three distinct steps in order to enable a global analysis of the relationships between the reads and is arguably the most successful assembly strategy. The first step is the same as for the greedy approach: the reads are compared to each other to construct a list of pair-wise overlaps. In a second step, this information is used to construct an overlap graph containing each contig as a node, and an edge connecting two nodes (if an overlap is identified between the corresponding reads). The last step is to build a single path that traverses each node in the overlap graph exactly once, corresponding to a reconstruction of the genome in the layout step. This module enables several analyses that are not possible with the previous algorithms. For instance, a common conflict is to resolve the boundary between a repeat and the genomic regions adjacent to the copies of this repeat throughout the genome (see figure 1.12). Also, this process can be constructed with the information of mate-pairs (paired reads that are separated by a certain distance), used by Arachne [25]. For the new data generated by the massively parallel sequencers, OLC approach has been adapted in different tools, as newbler for 454 sequencing data, Edena [113] with high performance and quality.

Eulerian path

This Eulerian path approach starts by breaking up the set of reads into their k-mer spectrum and uses the coverage information to resolving repeat events. The resulting k-mer spectrum is then used to construct a 'de Bruijn graph'. This graph contains as nodes the k-1 length prefixes and suffixes of the original k-mers, and two nodes being linked by an edge if there is an overlap of k-2 mers. The path generated is called Eulerian path, giving the name to the algorithm. Intuitively, the Eulerian approach offers several advantages over the OLC strategy. First of all, pairwise overlaps between reads are never explicitly computed, hence the overlap step, which is a time-consuming process, is avoided. Furthermore, efficient algorithms exist for finding a Eulerian path in

a graph in contrast to the OCL approach. Ultimately, the task of an assembler is to find just one of the possible paths, corresponding to the correct reconstruction of the genome. The Eulerian strategy has been proposed as an alternative to OLC for the assembly of Sanger data and was implemented in the Euler series of assemblers, Velvet and ALLPATHS [289, 47]. However, it was not widely adopted, in part because of the high sensitivity to sequencing errors and the loss of information due to chopping up the reads into a set of k-mers, increasing the problem of solving the assembly in case of short repeats.



(a) Impact in science

Figure 1.12: Assembly concepts, adapted from Myers *et al*, 2000. a) Consider the hypothetical genome consisting of three unique stretches A, B, and C with two nearly identical, interspersed copies, X? and X", of a repeat element X. This results in the four contig and overlaps shown. The contig X? +X" is overcollapsed, and the U-contig for regions A, B, and C have repeat boundaries indicating the tail portions that project into X. b) A scaffold is a collection of ordered contigs with approximately known distances between them. Our contigs are built from U-contig that form a scaffold via bundles and then have a series of rocks, stones, and pebbles filled into the gaps between them.

Scaffolding

None of the assembly strategies described above can completely reconstruct a genome from read data alone. The output of most assemblers consist of an often large collection of independent contigs

(see figure 1.12). Other sources of information can be used to determine the relative placement of these contigs along a genome in a process called scaffolding. Most commonly, scaffolding relies on mate-pair information. Two contigs can be inferred to be adjacent in the genome if one end of a mate-pair is assembled within the first contig, and the other end is assembled within the second contig. Scaffolding information can also be obtained from whole-genome mapping data. In brief, optical mapping can determine the approximate location of restriction enzyme cuts along a genome, thereby generating an ordered list of restriction fragment lengths along the genome. Such information can be used to identify the location of assembled contigs within the genome (integrated in SOMA tool), resulting in an assembly with a 80-90% of coverage of a entire bacterial genome [194]. All modern assemblers, irrespective of the underlying assembly paradigm, contain a scaffolding module, although, stand-alone scaffolders are also available, such as Bambus [216] allowing mate-pair information to be added to virtually any assembler. Note that both mate-pair and mapping-based scaffolding approaches have difficulties scaffolding short contigs and may, therefore, be difficult to apply to fragmented assemblies generated from short-read sequencing data.

1.4.3. Deep sequencing impact in science

The previously unimaginable scale and economy of these methods, coupled with their enthusiastic uptake by the scientific community and the potential for further improvements in accuracy and read length, suggest that these technologies are destined to make a huge and ongoing impact upon genomic and post-genomic biology (see figure 1.13). As an example of what this advent has influenced, the word 'sequencing' has increased its ratio of occurrence from 22.3 times/year (before 2008) to 198.48 times/year (after 2008) in one of the most relevant journals: 'Science'. International consortiums are choosing deep sequencing as the main strategy to investigate different aspects of biology. Focusing on humans, 3 relevant projects have been

started from the advent of sequencing: the first one was launched in 2008 with the aim to fully sequence 1000 individuals. Coming 2 years later the ICGC project, focused on 25000 cancer genomes, and coupled with the UK10K human genomes project directed by Wellcome Trust Institute. All of them are the very beginning of a new era of 'personal genomics' services whereby complete genome sequences of individuals will contribute to improve diagnostics and therapeutic development while requiring novel tools for data management and to ensure data privacy.

Figure 1.13: Visualization of the sequencing relevance in the popular seqanswer forum (http://seqswers.com/wiki/Special:BrowseData). The graphic has been created with the online resource wordle (http://www.wordle.net/). Words such as 'sequencing', 'sequence' and 'reads' are well represented among the tools developed until now. The concepts more visible are directly related to the development of tools meant to solve the more difficult aspects of sequencing data analysis: 'alignment', 'assembly', 'mapping' ,'resequencing', 'data', 'RNA-seq' and so on.

Furthermore, the management and analysis of next-generation sequencing data requires (and indeed has already driven) the development of informatic tools able to assemble, map, store, share and interpret huge quantities of nucleotide sequence data. However, all of the current tools are directed by command-line interaction and

are inaccessible for wet-bench researchers. There is undoubtedly a need for a more intuitive, graphic user interface instruments to render the power of these new technologies available to a wider audience within the scientific community. All of these considerations will further enhance the symbiotic relationship between modern biology and computational sciences, and ensure long and productive careers for talented and committed bioinformaticians.

1.5. Analyzing sRNA with sequencing

From the beginning, sequencing has been the method of choice for miRNA discovery, wherein researchers have used conventional cloning and Sanger sequencing. The methodology to identify new sRNAs does not vary, in general terms, either using Sanger or deep sequencing platforms. The main steps in which a typical analysis of deep sequencing data is performed are the following (see figure 1.14):

1. Retrieve the sRNA sequences from the product generated by the sequencer. Depending on the procedure, adaptor at 5' and/or 3'-ends should be removed by matching the adaptor sequence to the data allowing in some cases a custom number of mismatches assigned to error sequencing.

2. The resulting sequences are then mapped back to the reference genome to identify the loci they are transcribed from.

3. In the majority of the analyses, all sequences annotated as rRNA, tRNA, TE and so on are systematically removed of posterior analysis. In the recent years, and parallel to our work, some tools have been developed for the identification of this kind of sequences, and usually, home-made pipelines are used to detect other sRNAs that are no miRNAs.

4. If miRNAs are the focus of the analysis, the strategy stops here, however if the final objective are other sRNAs, and extra step

is included to exclude all the sequences product of degradation. Typically, a window of 60-70 nt around each sequence is examined to determine whether the sequence is the representation of a unique spot on the loci, or a continuos and overlapping sequences are detected as a consequence of degradation.

5. At the end, for non-miRNA sRNAs, a further analysis is performed to characterize some common features of the population detected. This step has been applied to identify all other sRNAs and piRNAs described in the previous sections.

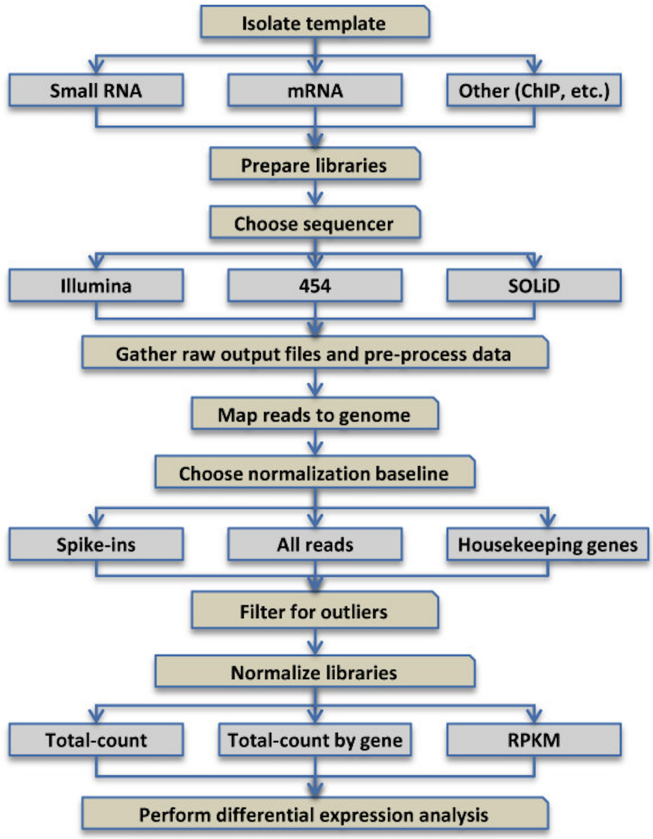


Figure 1.14: Flowchart of typical data-handling steps for small RNA (sRNA) libraries among others (McCormick *et al.*, 2010). Flowchart depicting the steps involved in creating, processing and normalizing next-generation sequencing libraries.

1.5.1. MiRNA data evolution

The first systematic studies to identify sRNAs were directed to the discovery of novel miRNAs [161, 153, 157]. To identify the putative miRNAs, a biogenesis process is expected where the sequence has been cleaved out of either arms of a putative miRNA precursor hairpin. Therefore, the flanking genomic sequence of a miRNA candidate should be able to form a hairpin when folded with an RNA structure prediction algorithm. If the candidate miRNA star sequence is also detected, it is seen as confounding evidence. This strategy enabled to the collection of enough data to create species-specific database, promoting a new branch in the miRNAs study: comparative analysis. In this analysis, the purpose is more the detection of a known miRNAs than the discovery of novel forms. In this context, miRBase the first public database, was released in 2004 [97], wherein all sequences detected as miRNA were uploaded and maintained by recollecting new publications in the field. Until the last year, the database recorded information such as, miRNA and precursor sequences, position on the genome, and validated and predicted targets. As a direct consequence of the database creation, microarray were designed to determine profiles of miRNAs expression in high-throughput approaches in different species and tissues. However, after the deep sequencing emergence, this method has been the gold standard technology for the miRNA detection. To this end, Illumina platform has been the most widely used, since the amount of data produced is bigger and cheaper (900€/20 mill reads) than the rest of the current technologies. The main advantage of over arrays is the detection of known and unknown sRNAs in an un-biased procedure. Over to 350 data series have been uploaded to GEO database [78] from 2008 until now, a 95% of the data generated by arrays in 10 years. In addition, 10 additional species have been analyzed with the new technology thanks to de novo analysis offered by sequencing. As a collateral effect of this data boom, miRBase has integrated a novel module in its database, consisting in, the visualization of all sequences mapped on a miRNA gene and detected by this technology. The main

reason to do that, is because this new and unexpected information disconcerts, in some way, the well-build concept of miRNAs and their biogenesis, assuming that only one sequence was generated from the precursor with functional relevance.

1.5.2. Other sRNAs data evolution

Before sequencing advent, only two databases were developed with a similar aim than miRBase: siRNAdb and piRNABank storing siRNAs and piRNAs small classes, with data obtained by cloning and Sanger sequencing [53, 233]. This information has been grown very fast in the last 3 years, usually, loading the raw data to GEO database [78], more than merging everything in a unique source. Only one big database has tried to group and to classify sRNAs outside miRNA family, offering for the first time a very complete overview of this type of information: DeepBase [284]. Currently this database has collected over 200 different libraries from plant to mammalian species, and has classified in different types according to their location and published discoveries: nasRNA,pasRNA,easRNA and rasRNA when the sRNAs derive from non-coding RNA, protein coding RNA, exon RNA and repeats respectively. In addition, and international web portal: www.ncRNA.org is collecting bioinformatics tools and databases specialized for functional RNAs. This site was funded by New Energy and Industrial Technology Development Organization (NEDO) from Japan, and includes a genome browser with all public databases and individual experiments related to non-coding RNA.

1.5.3. Tools for sRNA analysis

MiRNA detection tools

At the beginning of this thesis no tools had been developed for miRNAs analysis using deep sequencing data. Parallel to this work, some other tools have been emerged enumerate next (see table 1.2):

miRanalyzer, a web server tool for the analysis of deep-sequencing experiments for sRNAs. The web server tool requires a simple input file containing a list of unique reads and its copy numbers (expression levels). Using these data, miRanalyzer 1) detects all known microRNA sequences annotated in miRBase, 2) finds all perfect matches against other libraries of transcribed sequences and 3) predicts new microRNAs [103].

MiRNAkey is a software package designed to be used as a base-station for the analysis of miRNA deep sequencing data. The package implements common steps taken in the analysis of such data, as well as adds data statistics and multiple mapping levels, generating a novel platform for the analysis of miRNA expression. Through the use of a simple graphical interface, the user can determine the analysis steps. The tabular and graphical output contains detailed reports on the sequence reads and provides an accurate picture of the differentially expressed miRNAs in paired samples [227].

MirTools is a web server developed to allow researchers to perform

Table 1.2: miRNA tools

Name	type	isomir[a]	prediction[b]	others sRNAs[c]	expression[d]
miRkey	comand-line	-	+	-	miRNA
miRAnalyzer	web server	-	+	discart	-
mirTools	web server	-	+	discart	miRNA
mirExpress	comand-line	-	-	-	miRNA
DSAP	web server	+	-	discart	miRNA
miR-E	comand-line	-	-	simple	-

a:isomiRs detection(+); b:miRNA prediction(+); c: others sRNAs detection (yes='+', no='-', detected to be removed='discart'); d: expression analysis (no='-', only for miRNAs='miRNA').

a comprehensive characterization of the sRNA transcriptome. It can: 1) Align the large-scale short reads to the reference genome and exhibit the length distribution; 2) Classify of the large-scale short reads into known categories, such as known miRNAs, non-coding RNA, genomic repeats or coding sequences; 3) Provide detailed annotation information of known miRNAs, such as miRNA/miRNA*, absolute/relative reads count and the most abundant tag; 4) Discovery of the novel miRNAs that have not been characterized before from the large-scale short reads; 5) Identify the differentially expressed miRNAs between samples according to different count strategies, such as total read tag counts and the most abundant tag of specific miRNA [292].

DSAP uses a tab-delimited file as an input format, which holds the unique sequence reads (tags) and their corresponding number of copies generated by the Solexa sequencing platform. The input data will go through four analysis steps in DSAP: 1) cleanup: removal of adaptors and poly-A/T/C/G/N nucleotides; 2) clustering: grouping of cleaned sequence tags into unique sequence clusters; 3) non-coding RNA (ncRNA) matching: sequence homology mapping against a transcribed sequence library from the ncRNA database Rfam; and 4) known miRNA matching: detection of known miRNAs in miRBase based on sequence homology. The expression levels corresponding to matched ncRNAs and miRNAs are summarized in multi-color clickable bar charts linked to external databases [82].

miRExpress for extracting miRNA expression profiles from sequencing reads obtained by second-generation sequencing technology. A stand-alone software package is implemented for generating miRNA expression profiles from high-throughput sequencing of RNA without the need for sequenced genomes. The software is also a database-supported, efficient and flexible tool for investigating miRNA regulation [271].

Mir-E, a group of perl scripts to generate an expression matrix for all know non-coding RNAs detected in the input data, to report transcripts per million (correct for sequence depth) and square root transformed expression levels (variance stabilization) and to create Bed and Wig files for data visualization in the UCSC browser [43].

1.5.4. MiRNA analysis in diseases

After the discovery of the functional role of miRNAs in cells, it was a wide-spread reaction the analysis of these molecules in patients suffering some diseases [56, 26, 208]. The advent of the next generation sequencing facilitated these studies, generating a excellent miRNA profile in a shot. Since then, the more common study was the identification of miRNA families deregulation in different cancer types [253, 224, 290, 197]. Several studies have demonstrated that miRNAs are highly expressed in central nervous system (CNS), being a modulators of both CNS development and plasticity [174]. Complete loss of miRNA expression in the brain leads to neurodegeneration in several animal models [112]. A link between miRNA pathways and neurological diseases, including neurodegenerative disorders such as Alzheimer's disease, Huntington's disease and Parkinson's disease is becoming increasingly evident [158]. However, there is a lack of sequencing data in neurological diseases, maybe due to the difficulty in obtaining affected tissue from patients.

Hypothesis and objectives

Small silencing RNA is a family of non-coding RNAs (ncRNAs) 18-30 nt long, involved in gene silencing by association with the Argonaute family of proteins [93, 140, 65]. The advent in sequencing has permitted to detect novel miRNAs, to increase the knowledge of miRNA biogenesis, and to discover previously unknown putative post-transcriptional editing processes. These post-transcriptional mechanisms remain largely uncharacterized. These editing processes would generate variations of the current annotated miRNAs sequence in the miRBase repository at the 3' and 5' terminus and, in minor frequencies, nucleotide substitution along the miRNA length [77, 124, 205, 185, 191].

The progress in high-throughput sequencing technologies has also largely contributed to reveal other non-miRNA sRNAs including novel non-canonical sRNAs (derived from long non-coding RNA), repeat elements and coding regions. Since the functional role remains unclear, these novel sRNAs are classified according to their position and putative functions. Nowadays, five major groups have been detected: a) tiRNA located at the transcription initiation site [130, 258, 260], b)spli-RNA detected at splicing site of transcripts [258], c) tasi-RNA associated to gene termini [51], sRNA derived from tRNA [170, 89, 282, 110] and d) sRNA from non-coding RNA regions first described in [144, 17].

When the thesis started, no tools existed offering a user-friendly and customizable analysis of sRNA data generated from next generation sequencing technologies. The first released tools were web server dependent, providing a limited set of sRNA species for the analysis [103, 82]. Additionally, isomiRs are not well characterized in any of them, hampering a deep analysis in this kind of study. This problem remains in posterior published pipelines [271, 43, 292, 227]. Moreover, other sequences that are neither miRNA nor predicted as such, are ignored and treated as RNA degradation products, mainly because they are mapping to transcriptome datasets or repeat elements.

Based on all this, and with the aim to gain insight into the miRNA biogenesis, post-transcriptional edition mechanism, miRNA variability function and non-miRNA sRNA characterization, the following objectives were defined:

1. **Development of a bioinformatics pipeline to analyze high-throughput sequencing data from the sRNA fraction. The main goal was to build a user-friendly tool capable to deeply characterize, in an unbiased way, the small sRNA content of any sample. The following characteristics were pursued for the deepest characterization purposes:**

    1.1 To detect all miRNAs and reveal the characteristics of the different types of miRNA sequence variations.

    1.2 To integrate specific analyses such as isomiR quantitative relevance and sRNA differential expression data.

    1.3 To detect sRNAs that are unknown or non-predicted as miRNAs. This module will deal with cross-mapping events due to ambiguous reads and may distinguish sequences derived from degradation products.

2. **Study of the importance of miRNA variability in biological processes:**

    2.1 To characterize miRNA variability in different species, therefore approaching the possible relevance of isomiRs across evolution.

    2.2 To profile miRNAs and IsomiRs in normal brain development and aging and to inspect the putative impact of isomiRs as regulators of age-related genes. Specifically, evaluating the impact of isomiRs in the structural stability of the miRNA-mRNA duplexes.

3. **To profile miRNAs and isomiRs in Huntington's disease and decipher the putative relevance of miRNAs and isomiRs in gene deregulation linked to neurodegenerative processes.**

2 Results

The results section of this thesis is divided in four parts, two of them are currently published in international indexed scientific journals, and the remaining is under review/in preparation. The two first studies are focused on the development of bioinformatics tools for the analysis of sRNA data coming from next generation sequencing, separating miRNAs analysis from that of other sRNAs. The third work is centered on investigating the role of miRNA families in Huntington disease, and finally the last part addresses isomiRs behavior and characterization in different species during evolution and try to elucidate their putative functional role.

2.1. SeqBuster, a bioinformatic tool for the processing and analysis of small RNAs datasets, reveals ubiquitous miRNA modifications in human embryonic cells.

This work is focused on the development of a user-friendly tool offering a complete analysis and characterization of miRNAs from data generated by next generation high-thouhput sequencing of the sRNA fraction. The tool was integrated inside the Java platform allowing an easy development of a point- and-click interface. External resources for the data structure are MySQL platform and R language for the statistical methods and visual representation. SeqBuster is the first tool designed to inspect isomiR population in order to elucidate their putative functional role. To facilitate the latter, differential expression analysis were implemented to highlight relevant sequences with a presumed function to be experimentally verified in subsequent functional analyses. The complete set of analyses in SeqBuster were tested using public data generated from the study of human stem cells before and after differentiation [191].

The results of this study led to the publication of the following article:

SeqBuster, a bioinformatic tool for the processing and analysis of small RNAs datasets, reveals ubiquitous miRNA modifications in human embryonic cells.

Pantano L, Estivill X, Marti E.

Nucleic Acids Res. 2010 Mar;38(5):e34. Epub 2009 Dec 11.

Supplementary material

SUPPLEMENTARY MATERIAL

Figure S1: SeqBuster identifies different types of isomiRs. Histogram displaying the proportion of miRNAs with different types of isomiRs in hESC (left bars) and EB (right bars) small RNA libraries. The types of miRNA variability include trimming 5'-trimming and 3'-trimming variants, 3'-addition variants and nt-substitution variants. For every type of variability, the upper part of the graph shows the proportion of miRNAs presenting one (white), two (grey) or more than two (black) isomiRs. The abundance of the isomiR with respect to the corresponding reference miRNA is mirrored in the lower graph. The 5 brown color intensities from dark to light indicate the frequency of the isomiR with respect to the reference miRNA: >80%; 60%-80%; 40%-60%; 20%-40% and <20%.

Figure S2: Percentage of miRNAs with 5'-trimming variants (A) and the 3'-trimming variants (B) in hESC (left bars) and EB (right bars) samples, according to the nucleotide position involved in the trimming variant. Pre-miRNA slicing occurring at different positions upstream or downstream of the reference miRNA extremes (5'- and 3'-) is indicated by -3 to -1 and +1 to +3, respectively. In the upper bars the color pattern indicates the nucleotides involved in the trimming variants at every position: A (red), U (blue), C (green) and orange (G). Significant bias toward a nucleotide involved in the trimming variant is shown with a white asterisk. In every histogram the lower bars show the proportion of the isomiR with respect the reference miRNA at different positions as described in Figure 1.

Figure S3: Nucleotide substitution histogram showing the percentage of miRNAs with nucleotide modifications in hESC (left bars) and EB (right bars), at several positions of the miRNA (A). The upper bars show the type of nucleotide in the reference miRNA that presents a modification. The type of nucleotides present in the isomiRs is mirrored in the lower bars. Significant bias towards a specific nucleotide in the reference miRNA sequence is labeled with a white asterisk. The overall nucleotide substitution pattern is represented in (B). Nucleotides in the rows are those corresponding to the

reference miRNA and nucleotides in the columns are those found in the isomiR. Notice that the majority of miRNAs, 84 in hESC and 89 in EB, present a G to U substitution.

Figure S4: Analysis of the sequencing capacity with SeqBuster. In each sample, the total number of reads is ordered according to decreasing frequency. The logarithm of the counts is represented for hESC (black), and for EB (grey) samples.

Figure S5: Biological functions affected by hESC- or EB-enriched isomiRs in comparison with the corresponding reference miRNAs. The number of miRNAs and isomiRs enriched in each sample is indicated in bold. The total number of genes considered in the identification of affected biological functions is shown in a square. The isomiRs targets highlight some new top biological functions (bold).

Table S1: Detection of isomiRs and new miRNAs detected by SeqBuster. Examples of some isomiRs (highlighted in grey) with a high frequency presenting the adapter (highlighted in bold) and that were not detected in the original analysis Examples of sequences that were detected in the original analysis (highlighted in cursive), but were not annotated as miRNAs

Table S2: Invariable miRNAs. Some miRNAs appear only in hESC, and the number of counts is not displayed in EB, meaning that in EB these miRNAs present variants. A similar observation is detected for invariant miRNAs present only in EB. Notice that some miRNA lack isomiRs in both libraries (highlighted in a grey box).

Table S3: Examples of hESC and EB IsomiRs presenting a similar pattern of nt-substitution.

Table S4: Differently expressed miRNAs between libraries. A ratio ≥ 1.5 and ≤ 0.5 was considered for up- and down-regulated miRNAs, respectively. Corrected p-values are obtained applying the Hochberg and Benajmini

method on the p-value assigned by the Z-test. miRNAs with a total frequency (hESC + EB) > 50 are highlighted in bold.

Table S5: Correlation between the expression pattern of isomiRs differently expressed in hESC and EB libraries and the expression pattern of the corresponding reference miRNAs. Nt-substitution and 5'-trimming- variants affecting the seed region are tagged as 5'-variants. Addition variants and trimming variants affecting the 3'-terminus are tagged as 3'-variants.

Table S6: Seed region IsomiRs differently expressed in hESC and EB detected by SeqBuster. The seed region (2-8 nt) is shown for the IsomiRs, and the corresponding miRNAs. The modified nucleotide in the nt-substitution variants is highlighted in bold. A ratio ≥ 1.5 and ≤ 0.5 was considered for up- and down-regulated isomiRs, respectively. Corrected p-values are obtained applying the Hochberg and Benajmini method on the p-value assigned by the Z-test.

Figure S1

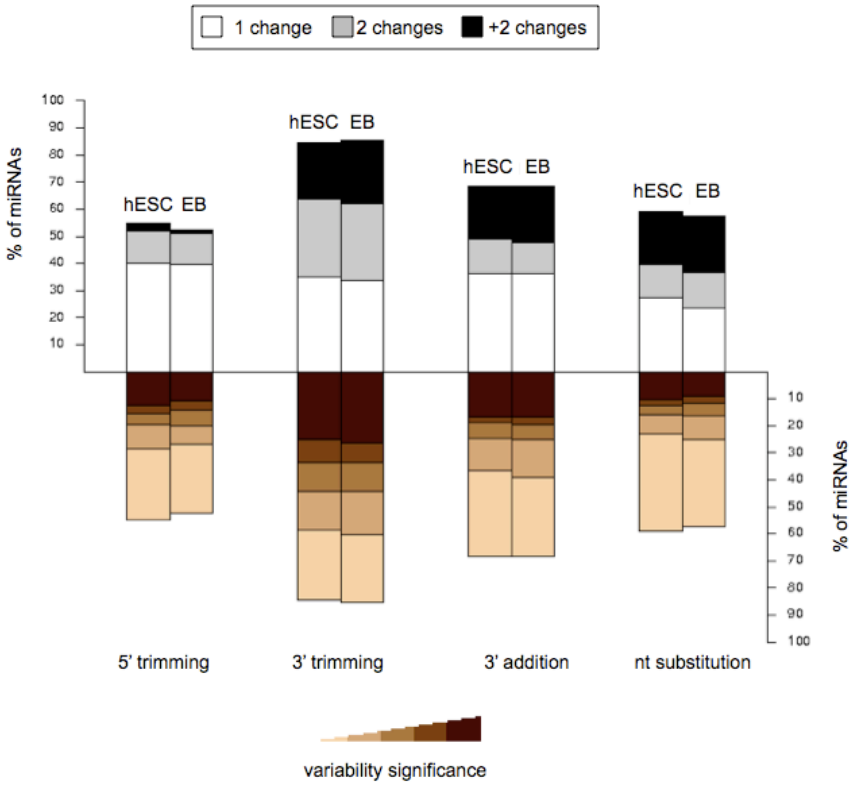

1 change
2 changes
+2 changes
% of miRNAs
hESC EB
hESC EB
hESC EB
hESC EB
5' trimming
3' trimming
3' addition
nt substitution
variability significance

Figure S2

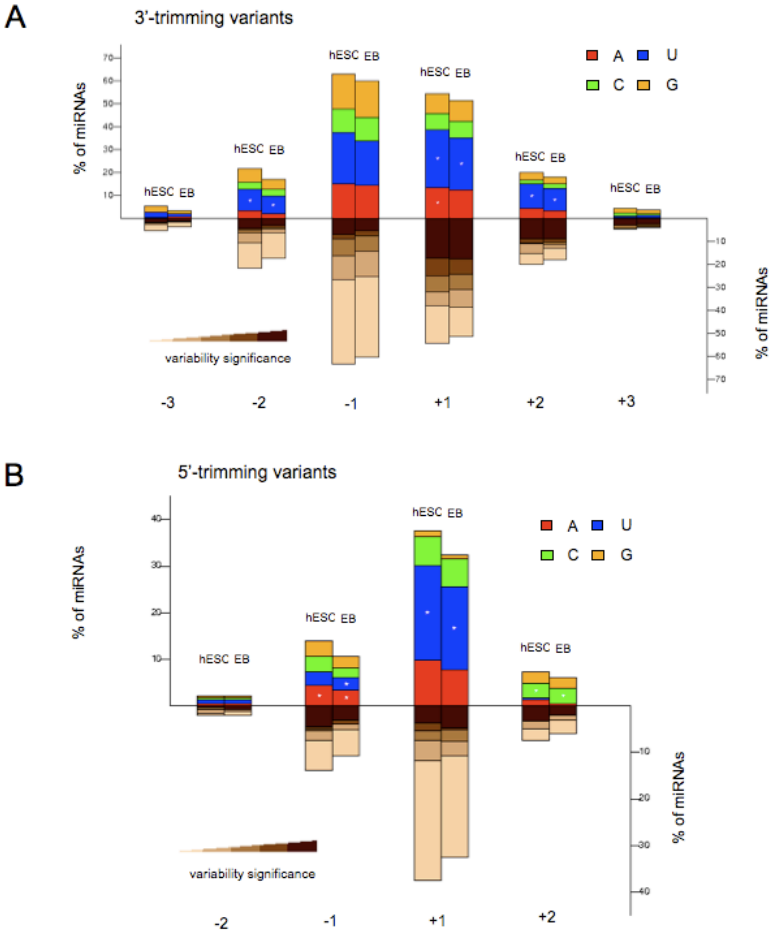

A
3'-trimming variants
% of miRNAs
hESC EB
A U
C G
variability significance
-3
-2
-1
+1
+2
+3
B
5'-trimming variants
% of miRNAs
hESC EB
A U
C G
variability significance
-2
-1
+1
+2

Figure S3

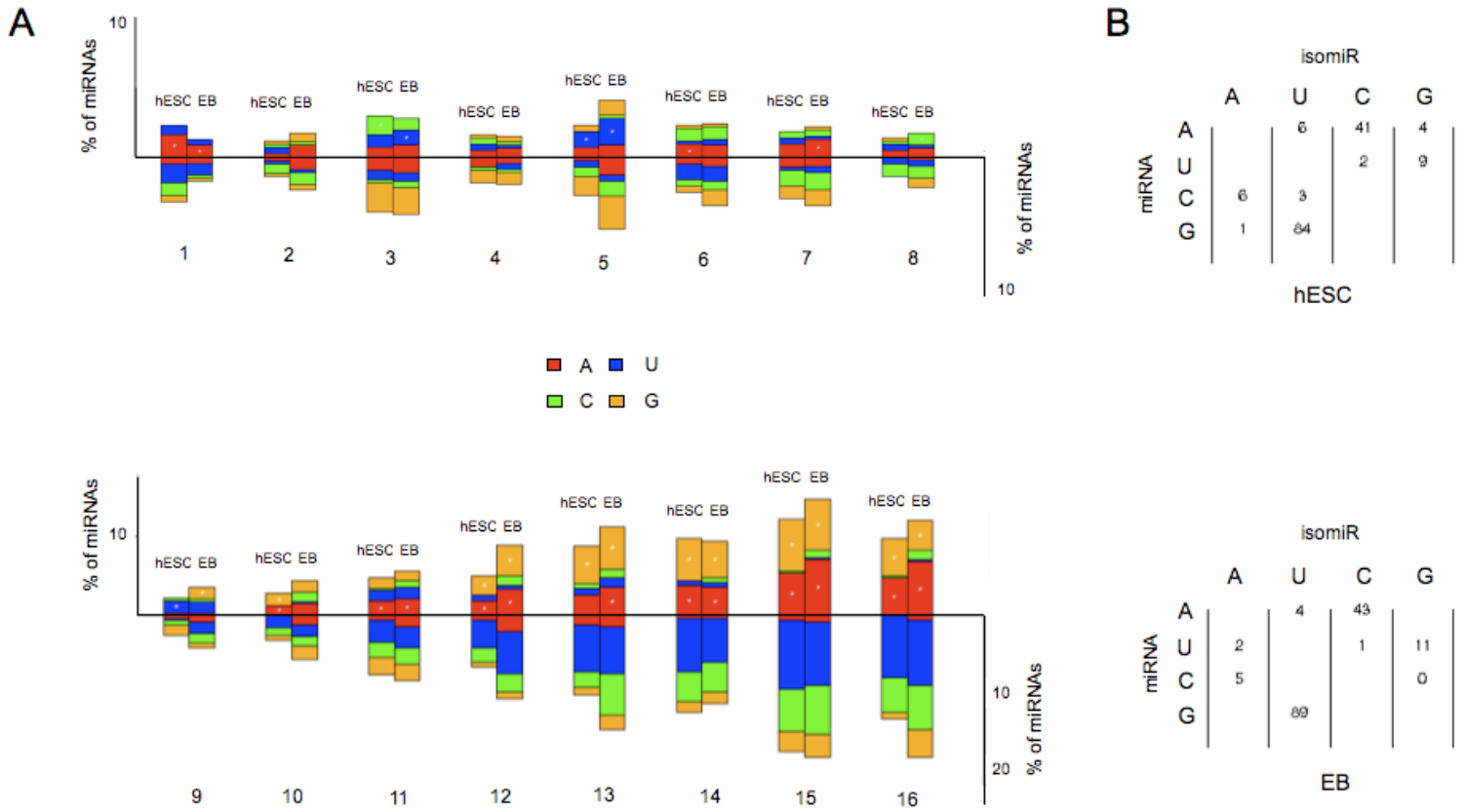

A
B
% of miRNAs
hESC EB
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
A
U
C
G
isomiR
miRNA
hESC
EB

Figure S4

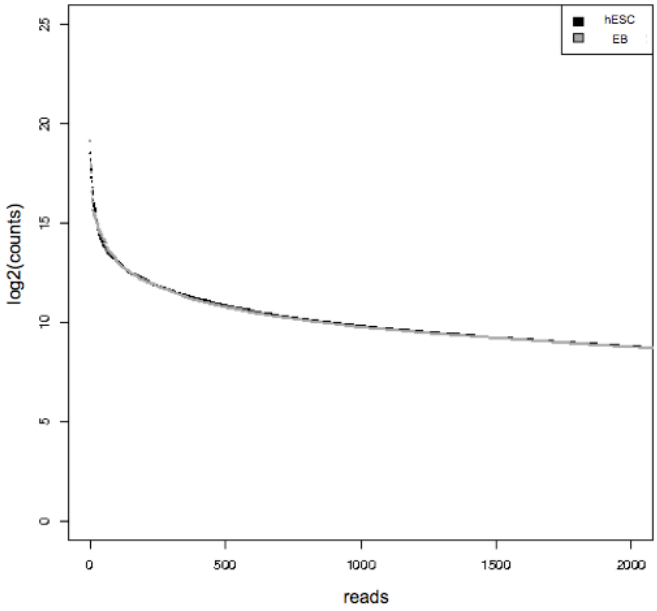


Table S1: Detection of isomiRs and new miRNAs detected by SeqBuste

miRNA	Sequence	hESC counts
hsa-miR-21	TAGCTTATCAGACTGATGTTGACA**TCGTATGCCGTC**	6779
hsa-miR-103	AGCAGCATTGTACAGGGCTATGAT**TCGTATGCCGTC**	829
hsa-miR-1208	CCGAGTTCTACCTTCCGCCTGTC**TCGTATGCCGTCT**	422
hsa-miR-92a	TATTGCACTTGTCCCGGCCTGTA**TCGTATGCCGTCT**	1206
hsa-miR-25	CATTGCACTTGTCTCGGTCTTA**TCGTATGCCGTCTT**	629
hsa-miR-422a	*ACTGGACTTAGAGTCAGAAGGC*	9
hsa-miR-513c	*TTCTCAAGGAGGTGTCGTTTAT*	4
hsa-miR-585	*TGGACGTATCTGTATGCTAGGG*	24
hsa-miR-621	*CTGCTCTGATGAAAT*	226
hsa-miR-557	*GCGGGTGGGCCTTTT*	76

Table S2: Invariable miRNAs.

miRNA	Counts	
	hESC	EB
hsa-miR-876-5p	6	13
hsa-miR-495		10
hsa-miR-1912		9
hsa-miR-517c		9
hsa-miR-376c		8
hsa-miR-433	13	7
hsa-miR-1284	3	6
hsa-miR-324-5p	3	6
hsa-miR-383		6
hsa-miR-655		6
hsa-miR-590-5p	5	5
hsa-miR-1197		5
hsa-miR-1252	4	4
hsa-miR-346	2	4
hsa-let-7d		4
hsa-miR-299-3p		4
hsa-miR-376a		4
hsa-miR-489		4
hsa-miR-513c		4
hsa-miR-628-5p	7	3
hsa-miR-328	6	3
hsa-miR-1255b	5	3
hsa-miR-33b	2	3
hsa-miR-548i	2	3
hsa-miR-1263		3
hsa-miR-154		3
hsa-miR-548p		3
hsa-miR-548o	7	2
hsa-miR-491-3p	6	2
hsa-miR-521	6	2
hsa-miR-1253	4	2
hsa-miR-1288	3	2
hsa-miR-1256	2	2
hsa-miR-1267	2	2
hsa-miR-1258		2
hsa-miR-1265		2
hsa-miR-1292		2
hsa-miR-486-3p		2
hsa-miR-493		2
hsa-miR-561		2
hsa-miR-582-5p		2
hsa-miR-592		2
hsa-miR-1277	33	
hsa-miR-429	13	
hsa-miR-532-3p	13	
hsa-miR-1294	11	
hsa-miR-618	9	
hsa-miR-509-3-5p	8	
hsa-miR-1295	4	
hsa-miR-944	4	
hsa-miR-34c-3p	3	
hsa-miR-548b-3p	3	
hsa-miR-597	3	
hsa-miR-1289	2	
hsa-miR-1299	2	
hsa-miR-1304	2	
hsa-miR-150	2	
hsa-miR-219-1-3p	2	
hsa-miR-409-5p	2	
hsa-miR-548m	2	
hsa-miR-885-5p	2	
hsa-miR-889	2	

Table S3: Examples of hESC and EB IsomiRs presenting a similar pattern of nt-substitution.

miRNA	nt-substitution	hESC	EB
hsa-miR-1246	G->A pos 12	711	426
hsa-miR-1274b	G->A pos 12	21	23
hsa-miR-222	G->A pos 12	139	25
hsa-miR-448	G->A pos 12	21	26
hsa-miR-452	G->A pos 12	3	3
hsa-miR-455-5p	G->A pos 12	2	4
hsa-miR-30a	T->G pos 3	12	26
hsa-miR-30d	T->G pos 3	18	33
hsa-miR-30e	T->G pos 3	6	9
hsa-miR-1246	A->C pos 7	73	29
hsa-miR-30e	A->C pos 7	2	4

Table S4: Differently expressed miRNAs between libraries.

miRNA	hESC	EB	ratio	*p*	*p*-corrected
hsa-miR-25	**28299**	**17858**	UP	0	0
hsa-let-7a	**16988**	**3998**	UP	0	0
hsa-miR-221	**16617**	**8658**	UP	0	0
hsa-miR-302b	**14971**	**7649**	UP	0	0
hsa-miR-423-5p	**13876**	**7795**	UP	0	0
hsa-miR-1	**10436**	**5618**	UP	0	0
hsa-miR-302d	**9814**	**5579**	UP	0	0
hsa-miR-302a	**7479**	**4275**	UP	0	0
hsa-miR-1323	**5970**	**2813**	UP	0	0
hsa-miR-320a	**5120**	**2485**	UP	0	0
hsa-miR-191	**4075**	**2448**	UP	0	0
hsa-miR-744	**3630**	**1065**	UP	0	0
hsa-miR-1298	**1380**	**623**	UP	0	0
hsa-miR-423-3p	**1240**	**537**	UP	0	0
hsa-miR-331-3p	**912**	**272**	UP	0	0
hsa-miR-302c	**793**	**319**	UP	0	0
hsa-let-7c	**501**	**42**	UP	0	0
hsa-miR-22	**490**	**268**	UP	0	0
hsa-miR-205	**486**	**189**	UP	0	0
hsa-miR-9	**438**	**253**	UP	0	0
hsa-let-7g	**427**	**143**	UP	0	0
hsa-miR-330-3p	**390**	**256**	UP	0	0
hsa-miR-184	**390**	**41**	UP	0	0
hsa-miR-151-3p	**358**	**202**	UP	0	0
hsa-miR-193b	**338**	**152**	UP	0	0
hsa-miR-518b	**267**	**123**	UP	0	0
hsa-miR-532-5p	**262**	**160**	UP	0	0
hsa-miR-28-3p	**240**	**114**	UP	0	0
hsa-miR-222	**233**	**114**	UP	0	0
hsa-miR-1308	**233**	**23**	UP	0	0
hsa-miR-1278	**218**	**144**	UP	0	0
hsa-miR-486-5p	**208**	**55**	UP	0	0
hsa-miR-504	**189**	**76**	UP	0	0
hsa-miR-129-5p	**186**	**61**	UP	0	0
hsa-miR-197	**173**	**73**	UP	0	0
hsa-miR-664	**141**	**70**	UP	0	0
hsa-miR-1255a	**136**	**90**	UP	0.02	0.01
hsa-let-7i	**130**	**23**	UP	0	0
hsa-miR-941	**129**	**45**	UP	0	0
hsa-miR-1307	**121**	**22**	UP	0	0
hsa-miR-1270	**120**	**47**	UP	0	0
hsa-let-7b	**116**	**11**	UP	0	0
hsa-miR-187	**110**	**32**	UP	0	0
hsa-miR-520g	**100**	**25**	UP	0	0
hsa-miR-548j	**97**	**64**	UP	0.06	0.03
hsa-let-7d	**93**	**5**	UP	0	0
hsa-miR-498	**85**	**51**	UP	0.03	0.01
hsa-miR-95	**77**	**47**	UP	0.04	0.02
hsa-miR-574-3p	**77**	**18**	UP	0	0
hsa-miR-484	**74**	**21**	UP	0	0
hsa-miR-720	**72**	**30**	UP	0	0
hsa-miR-100	**71**	**38**	UP	0.01	0.01
hsa-miR-1266	**67**	**10**	UP	0	0
hsa-miR-31	**62**	**19**	UP	0	0
hsa-miR-518e	**58**	**25**	UP	0	0
hsa-miR-515-5p	**52**	**30**	UP	0.07	0.04
hsa-miR-424	**46**	**17**	UP	0	0
hsa-miR-526b	**45**	**21**	UP	0.02	0.01
hsa-miR-155	**41**	**12**	UP	0	0
hsa-miR-652	**40**	**22**	UP	0.09	0.05
hsa-miR-1301	32	7	UP	0	0
hsa-miR-548h	30	15	UP	0.1	0.05

hsa-miR-525-5p	28	8	UP	0.01	0	
hsa-miR-211	28	2	UP	0	0	
hsa-miR-181d	27	12	UP	0.07	0.03	
hsa-miR-877	26	0	UP	0	0	
hsa-miR-1254	22	4	UP	0	0	
hsa-miR-1303	21	5	UP	0.01	0	
hsa-miR-522	19	5	UP	0.02	0.01	
hsa-miR-1305	16	0	UP	0	0	
hsa-miR-1296	15	5	UP	0.08	0.04	
hsa-miR-524-5p	14	0	UP	0	0	
hsa-miR-671-3p	10	2	UP	0.07	0.03	
hsa-miR-1265	10	2	UP	0.07	0.03	
hsa-miR-1258	10	2	UP	0.07	0.03	
hsa-miR-874	9	0	UP	0.01	0	
hsa-miR-1261	7	0	UP	0.03	0.01	
hsa-miR-18b	6	0	UP	0.05	0.02	
hsa-miR-133a	6	0	UP	0.05	0.02	
hsa-miR-1291	6	0	UP	0.05	0.02	
hsa-miR-1260	6	0	UP	0.05	0.02	
hsa-miR-98	5	0	UP	0.07	0.04	
hsa-miR-940	5	0	UP	0.07	0.04	
hsa-miR-524-3p	5	0	UP	0.07	0.04	
hsa-miR-1293	5	0	UP	0.07	0.04	
hsa-miR-363	**3852**	**10160**	DOWN	0	0	
hsa-miR-130a	**2908**	**5834**	DOWN	0	0	
hsa-miR-340	**2805**	**8711**	DOWN	0	0	
hsa-miR-372	**1618**	**15679**	DOWN	0	0	
hsa-miR-204	**799**	**1693**	DOWN	0	0	
hsa-let-7e	**704**	**1822**	DOWN	0	0	
hsa-miR-421	**614**	**1362**	DOWN	0	0	
hsa-miR-122	**536**	**2751**	DOWN	0	0	
hsa-miR-10a	**481**	**1190**	DOWN	0	0	
hsa-miR-708	**403**	**903**	DOWN	0	0	
hsa-miR-20b	**387**	**1065**	DOWN	0	0	
hsa-miR-152	**380**	**1725**	DOWN	0	0	
hsa-miR-143	**304**	**628**	DOWN	0	0	
hsa-miR-106a	**247**	**637**	DOWN	0	0	
hsa-miR-27b	**179**	**396**	DOWN	0	0	
hsa-miR-371-5p	**179**	**387**	DOWN	0	0	
hsa-miR-30d	**92**	**238**	DOWN	0	0	
hsa-miR-145	**74**	**177**	DOWN	0	0	
hsa-miR-29c	**68**	**141**	DOWN	0	0	
hsa-miR-96	**66**	**146**	DOWN	0	0	
hsa-miR-660	**63**	**210**	DOWN	0	0	
hsa-miR-1246	**52**	**131**	DOWN	0	0	
hsa-miR-373	**46**	**605**	DOWN	0	0	
hsa-miR-210	**43**	**285**	DOWN	0	0	
hsa-miR-1277	**41**	**84**	DOWN	0	0	mirar cutoff
hsa-miR-1269	**38**	**117**	DOWN	0	0	
hsa-miR-26b	**36**	**97**	DOWN	0	0	
hsa-miR-518a-3p	**33**	**68**	DOWN	0	0	
hsa-miR-454	**27**	**101**	DOWN	0	0	
hsa-miR-127-3p	**25**	**223**	DOWN	0	0	
hsa-miR-146b-5p	**25**	**56**	DOWN	0	0	
hsa-miR-889	**2**	**232**	DOWN	0	0	
hsa-miR-1287	**19**	**44**	DOWN	0	0	
hsa-miR-362-5p	**15**	**67**	DOWN	0	0	
hsa-miR-371-3p	**12**	**151**	DOWN	0	0	
hsa-miR-134	**10**	**67**	DOWN	0	0	
hsa-miR-219-2-3p	**12**	**50**	DOWN	0	0	
hsa-miR-452	16	33	DOWN	0.03	0.01	
hsa-miR-618	11	33	DOWN	0	0	
hsa-miR-1259	10	24	DOWN	0.03	0.01	
hsa-miR-1262	7	52	DOWN	0	0	

hsa-miR-641	6	15	DOWN	0.09	0.05
hsa-miR-523	6	15	DOWN	0.09	0.05
hsa-miR-1285	6	15	DOWN	0.09	0.05
hsa-miR-126	5	15	DOWN	0.05	0.02
hsa-miR-499-5p	4	30	DOWN	0	0
hsa-miR-518f	14	35	DOWN	0.01	0
hsa-miR-219-1-3p	2	40	DOWN	0	0
hsa-miR-1247	2	38	DOWN	0	0
hsa-miR-99a	2	27	DOWN	0	0
hsa-miR-23b	2	17	DOWN	0	0
hsa-miR-1276	2	11	DOWN	0.03	0.01
hsa-miR-329	0	47	DOWN	0	0
hsa-miR-487b	0	39	DOWN	0	0
hsa-miR-379	0	29	DOWN	0	0
hsa-miR-199b-5p	0	29	DOWN	0	0
hsa-miR-369-3p	0	28	DOWN	0	0
hsa-miR-218	0	25	DOWN	0	0
hsa-miR-410	0	24	DOWN	0	0
hsa-miR-543	0	23	DOWN	0	0
hsa-miR-382	0	23	DOWN	0	0
hsa-miR-217	0	21	DOWN	0	0
hsa-miR-323-3p	0	19	DOWN	0	0
hsa-miR-485-3p	0	17	DOWN	0	0
hsa-miR-196b	0	15	DOWN	0	0
hsa-miR-432	0	13	DOWN	0	0
hsa-miR-495	0	12	DOWN	0	0
hsa-miR-199a-5p	0	11	DOWN	0	0
hsa-miR-409-3p	0	10	DOWN	0	0
hsa-miR-376c	0	10	DOWN	0	0
hsa-miR-655	0	7	DOWN	0.02	0.01
hsa-miR-518c	0	7	DOWN	0.02	0.01
hsa-miR-383	0	7	DOWN	0.02	0.01
hsa-miR-196a	0	7	DOWN	0.02	0.01
hsa-miR-1251	0	7	DOWN	0.02	0.01
hsa-miR-1245	0	7	DOWN	0.02	0.01
hsa-miR-301b	0	6	DOWN	0.03	0.02
hsa-miR-301a	0	6	DOWN	0.03	0.02
hsa-miR-1197	0	6	DOWN	0.03	0.02
hsa-miR-548b-5p	0	5	DOWN	0.06	0.03
hsa-miR-513c	0	5	DOWN	0.06	0.03
hsa-miR-489	0	5	DOWN	0.06	0.03
hsa-miR-376a	0	5	DOWN	0.06	0.03
hsa-miR-299-3p	0	5	DOWN	0.06	0.03
hsa-miR-194	0	5	DOWN	0.06	0.03
hsa-miR-154	0	4	DOWN	0.1	0.05
hsa-miR-132	0	4	DOWN	0.1	0.05
hsa-miR-125a-3p	0	4	DOWN	0.1	0.05
hsa-miR-10b	0	4	DOWN	0.1	0.05

Table S5: Correlation between the expression pattern of isomiRs differently expressed in hESC and EB libraries and the expression pattern of the corresponding reference miRNAs.

		number of isomiRs differently expressed			
		UP	DOWN		
Type of isomiR	3’	108	11	UP	**Pattern of differential expression in the correspondent reference miRNA**
	5’	29	6		
	5’ + 3’	44	19		
	3’	24	98	DOWN	
	5’	2	20		
	5’ + 3’	98	70		

Table S6: Seed region IsomiRs differently expressed in hESC and EB detected by SeqBuster.

miRNA		isomiR	hESC	EB	hESC/EB	p	p-corrected
hsa-miR-1246	isomiR	**-**AAUGGAUU	6407	2511	UP	0	0
	Ref	AAUGGAUU					
hsa-miR-378	isomiR	-CUGGACU	530	0	UP	0	0
	Ref	ACUGGACU					
hsa-miR-423-3p	isomiR	**A**UGAGGGGC	450	193	UP	0	0
	Ref	-UGAGGGGC					
hsa-miR-1261	isomiR	-UGGAUAA	407	83	UP	0	0
	Ref	AUGGAUAA					
hsa-miR-140-3p	isomiR	**- -** CCACAG	286	159	UP	0	0
	Ref	UACCACAG					
hsa-miR-1274b	isomiR	-CCCUGUU	262	0	UP	0	0
	Ref	UCCCUGUU					
hsa-miR-222	isomiR	**-** GCUACAU	184	87	UP	0	0
	Ref	AGCUACAU					
hsa-miR-330-3p	isomiR	**-** CAAAGCA	131	0	UP	0	0
	Ref	GCAAAGCA					
hsa-miR-1290	isomiR	**-** GGAUUUU	108	70	UP	0.04	0.02
	Ref	UGGAUUUU					
hsa-miR-296-3p	isomiR	**-** AGGGUUG	102	6	UP	0	0
	Ref	GAGGGUUG					
hsa-miR-222	isomiR	**U**GCUACAU	90	34	UP	0	0
	Ref	AGCUACAU					
hsa-miR-520f	isomiR	**C**AAGUGCUU	68	32	UP	0	0
	Ref	-AAGUGCUU					
hsa-miR-520g	isomiR	**-** CAAAGUG	66	22	UP	0	0
	Ref	ACAAAGUG					
hsa-miR-1246	isomiR	AAUGG**G**UU	66	0	UP	0	0
	Ref	AAUGGAUU					
hsa-miR-1307	isomiR	**-** CUCGGCG	63	10	UP	0	0
	Ref	ACUCGGCG					
hsa-miR-1246	isomiR	AAUGG**C**UU	62	35	UP	0.04	0.02
	Ref	AAUGGAUU					
hsa-miR-1290	isomiR	**U**UGGAUUUU	61	12	UP	0	0
	Ref	-UGGAUUUU					
hsa-miR-1246	isomiR	**-** AUGGGUUU	48	15	UP	0	0
	Ref	AAUGGAUU					
hsa-miR-363	isomiR	**-** AUUGCAC	1248	3938	DOWN	0	0
	Ref	AAUUGCAC					
hsa-miR-372	isomiR	**-** AAGUGCU	1068	11044	DOWN	0	0
	Ref	AAAGUGCU					
hsa-miR-192	isomiR	-UGACCUA	376	745	DOWN	0	0
	Ref	CUGACCUA					
hsa-miR-1246	isomiR	**- -** UGGAUU	77	449	DOWN	0	0
	Ref	AAUGGAUU					
hsa-miR-30e	isomiR	**-** GUAAACA	37	78	DOWN	0	0
	Ref	UGUAAACA					
hsa-miR-518f	isomiR	**-** AAAGCGCU	33	85	DOWN	0	0
	Ref	GAAAGCGCU					
hsa-miR-502-3p	isomiR	**-** AUGCACC	29	58	DOWN	0	0
	Ref	AAUGCACC					
hsa-miR-585	isomiR	UGG**A**CGUA	22	44	DOWN	0.01	0
	Ref	UGGGCGUA					
hsa-miR-30d	isomiR	**-** GUAAACA	21	53	DOWN	0	0
	Ref	UGUAAACA					
hsa-miR-371-3p	isomiR	**-** AGUGCCG	14	83	DOWN	0	0
	Ref	AAGUGCCG					
hsa-miR-371-3p	isomiR	**- -** GUGCCG	12	153	DOWN	0	0
	Ref	AAGUGCCG					

2.2. A non-biased framework for the annotation and classification of the non-miRNA small RNA transcriptome

We developed an extension for SeqBuster to integrate the analysis of other sRNAs that are not miRNAs. In the last two years, novel classes of sRNAs have emerged due to the advent of sequencing technology. Nowadays, the published tools are only focused on the detection and prediction of miRNA families ignoring the rest of the data that constitute more than 30% of the total sRNAs in some cases. For that, we extended the previous bioinformatic analysis tool with the goal of offering a deep characterization of the non-miRNA classes of sRNAs, therefore highlighting putative novel types.

The results of this study are under review on the international journal of *Bioinformatics*:

A non-biased framework for the annotation and classification of the non-miRNA small RNA transcriptome
Pantano L, Estivill X, Marti E.
Bioinformatics. Under Review.

BIOINFORMATICS

Vol. 00 no. 00 2011
Pages 1–2

A non-biased framework for the annotation and classification of the non-miRNA small RNA transcriptome

Pantano, L [1], Estivill X [2,*] and Marti E [2*]

[1]1Genetic Causes of Disease Group, Genes and Disease Program, Centre for Genomic Regulation (CRG), UPF, Barcelona

[2]Centro de Investigacin Biomedica en Red en Epidemiologa y Salud Publica (CIBERESP)

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Recent progress in high-throughput sequencing technologies has largely contributed to reveal a highly complex landscape of small non-coding RNAs (sRNAs), including novel non-canonical sRNAs derived from long non-coding RNA, repeated elements, transcription start sites and splicing site regions among others. The published frameworks for sRNA data analysis are focused on miRNA detection and prediction, ignoring further information in the data set. As a consequence, tools for the identification and classification of the sRNAs not belonging to miRNA family are currently lacking.

Results: Here, we present, SeqCluster, an extension of the current available SeqBuster tool to identify and analyze at different levels the sRNAs not annotated or predicted as miRNAs. This new module deals with sequences mapping onto multiple locations and permits a highly versatile and user-friendly interaction with the data in order to easily classify sRNA sequences with a putative functional importance. We were able to detect all known classes of sRNAs described to date using SeqCluster with different sRNA datasets.

Availability: tool and video-tutorial are available at http://estivill_lab.crg.es/seqbuster.

Contact: eulalia.marti@crg.es,xavier.estivill@crg.es

1 INTRODUCTION

Small silencing RNAs are the best-known class of non-coding RNAs (ncRNAs) of 1830 nt in length, involved in gene silencing by association to the Argonaute family of proteins (Czech and Hannon, 2011). Recent progress in high-throughput sequencing technologies has largely contributed to elucidate the remarkable landscape of sRNAs, revealing new species of sRNAs with unknown functions. These novel sRNAs are classified according to their position in the genome and putative functions. Nowadays, five major groups of sRNAs have been detected: a) tiRNAs, located at the transcription initiation site of coding genes (Taft *et al.*, 2010); b) spli-RNAs, detected at splicing site of transcripts (Taft *et al.*, 2010); c) tasi-RNAs, associated to gene termini; d) sRNAs derived from tRNA (Haussecker *et al.*, 2010) and e) sRNAs from non-coding RNA regions (Ono *et al.*, 2011). No tools are prepared to cover a complete analysis of data coming from sRNA sequencing, and the existing ones are only for miRNA characterization and prediction. A major challenging problem using high-throughput sequencing data is annotation when the sequences map onto multiple locations. The current frameworks resolve this situation with heuristic assumptions, including non-consistent data removal or providing random annotations. This produces biased results that hamper the discovery and classification of novel sRNAs. Here, we present SeqCluster, a tool for the characterization of the non-miRNA sRNA transcriptome. SeqCluster is presented as an extension of SeqBuster, a pipeline for the characterization of miRNAs (Pantano *et al.*, 2010) and constitutes the first framework giving a complete unbiased classification of non-miRNAs data of any specie. SeqCluster permits a user a friendly interaction with the data at any level in order to easily classify and annotate small RNA sequences with a putative functional importance.

*to whom correspondence should be addressed

2 METHODS

SeqCluster, an extension of the miRNA-analysis tool SeqBuster, has been developed to analyze any kind of sRNA detected by large-scale sequencing technologies (see implementation in supplementary methods). The new framework integrates three specific processes: 1) raw data processing and miRNA detection, 2) clustering, and 3) classification (Figure 1). In the first process the adapter is trimmed from the raw sequences, and subsequently sequences are mapped onto miRNA and miRNA precursor databases. To avoid the dependency on an external tool for mapping against miRBase dataset (Griffiths-Jones, 2004), a custom algorithm based on seed (fragments of 8 nt) indexation has been integrated in java (see miRNA detection in supplementary methods). Predicted miRNAs using an external tool may be loaded to SeqCluster to avoid the incorporation of these sequences to the study of the non-miRNA sRNA transcriptome. The second step defines unit-sRNAs (usRNAs) taking into account two filters: 1) sequence similarity and, 2) genome location. In the first filter, all sequences with 100% identity (no mismatches allowed) and more than 80% of overlapping are considered as putative unit small RNA (pre-usRNA). In the second filter, all pre-usRNAs are mapped onto a custom genome using megablast (Altschul *et al.*, 1990). Otherwise, annotated data from any other mapping tool may be directly uploaded onto SeqCluster. This filter only affects ambiguously overlapped pre-usRNA and are used to make the decision on whether or not the two overlapping pre-usRNAs should be considered a single cluster of sequences (usRNA). In the rest of the cases (unique sequences or unambiguously mapped clustered sequences), pre-usRNAs are directly called as usRNAs. The premise to merge ambiguous overlapped pre-usRNA into ambiguous usRNA is that all overlapped pre-usRNAs have to share all same regions.

When the latter does not occur due to more complex situations, pre-usRNAs enter into an extra module integrating a recursive algorithm (see decision algorithm in supplementary methods). Once usRNAs are defined and located on the genome, all of them are classified according to the genome context that will cover as many classes as the users define, being the more common types: non-coding RNAs, transposable elements (TE), and genes (see usRNA classification in supplementary methods) .

3 OUTPUT

The framework generates a main MySQL table for each sample where rows are usRNAs and columns show the following information: unique identifier, number of sequences, number of locations, coordinates and finally one column for each class, according to the annotation step -repeat, ncRNA- (see output scheme in supplementary methods). For a user-friendly view, BED files are generated to be uploaded to UCSC (Kent *et al.*, 2002). Furthermore, SeqCluster permits differential expression analysis between two samples or two groups of samples in different biological contexts. Datasets involving time series experiments may be also analyzed .

4 RESULTS

We have applied SeqCluster extension to analyze small RNA datasets of human brain samples sequenced by illumina 1G in our previous work (Martí *et al.*, 2010) and other public dataset from different species (see SeqCluster application to real datasets in supplementary methods). First we detected miRNA sequences using SeqBuster with default parameters and the miRBase resource (Release 15, (Griffiths-Jones, 2004)). We applied SeqCluster extension to the rest of the data, resulting, in a total of 8335, 12366, 15614, 44265 y 82985 sequences annotated as usRNAs in human brain, human stem cells, mouse, fly and worm, respectively.

5 CONCLUSIONS

Differing from the current sRNA analysis tools, the main advantages of SeqCluster framework are: 1) the classification and annotation of the data are not restricted to specific databases, offering the possibility to perform this analysis with any custom database; 2) the implementation of filters integrated to solve, in a non-biased way, the problem of ambiguous sRNAs mapping; 3) the opportunity to distinguish presumed RNA degradation products from putative functional sRNAs and; 4) the possibility to inspect highly expressed sequences that have not been successfully classified allowing the extraction of complex sRNAs for further analysis. Our results validate SeqCluster as a tool to detect and classify all types of sRNAs in different species, including the most recently discovered classes of still unknown function (Taft *et al.*, 2010).

ACKNOWLEDGEMENT

Funding: Spanish Ministry of Health Fondo de Investigaciones Sanitarias (PI081367) and Instituto de Salud Carlos III (CIBERESP); he Spanish Ministry of Science and Innovation (SAF2008-00357)the Sixth Framework Programme of the European Commission through the SIROCCO integrated project LSHG-CT-2006-037900 and the Spanish Ministry of Science and Innovation (SAF2008-00357). E.M. is partially supported by the Spanish Ministry of Health; L.P. is recipient of a fellowship from the Spanish Ministry of Science and Innovation MICINN.

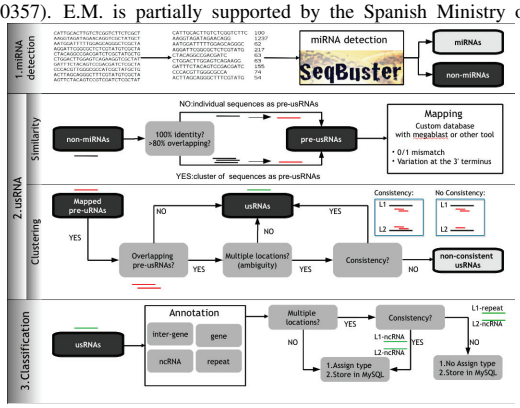


Fig. 1. SeqCluster extension framework scheme,The framework integrates three specific processes: 1) raw data processing and miRNA detection, 2) clustering and 3) classification. In the first process the adapter is trimmed from the raw sequences, and subsequently sequences are mapped onto miRNA and miRNA precursor databases. After that, all sequences not annotated as miRNAs or miRNA precursors are clustered in clustering step. This step is performed according two filters: 1) sequence similarity and, 2) genome location. In the classification step usRNAs are classified according to the genome context that will cover as many classes and sub-classes as the users define.

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, **215**(3), 403–410.

Czech, B. and Hannon, G. J. (2011). Small rna sorting: matchmaking for argonautes. *Nat Rev Genet*, **12**(1), 19–31.

Griffiths-Jones, S. (2004). The microrna registry. *Nucleic Acids Res*, **32**(Database issue), 109–111.

Haussecker, D., Huang, Y., Lau, A., Parameswaran, P., Fire, A. Z., and Kay, M. A. (2010). Human trna-derived small rnas in the global regulation of rna silencing. *RNA*, **16**(4), 673–695.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, D. (2002). The human genome browser at ucsc. *Genome Res*, **12**(6), 996–1006.

Martí, E., Pantano, L., Bañez-Coronel, M., Llorens, F., Miñones-Moyano, E., Porta, S., Sumoy, L., Ferrer, I., and Estivill, X. (2010). A myriad of mirna variants in control and huntington's disease brain regions detected by massively parallel sequencing. *Nucleic Acids Res*, **38**(20), 7219–7235.

Ono, M., Scott, M. S., Yamada, K., Avolio, F., Barton, G. J., and Lamond, A. I. (2011). Identification of human mirna precursors that resemble box c/d snornas. *Nucleic Acids Res*.

Pantano, L., Estivill, X., and Martí, E. (2010). Seqbuster, a bioinformatic tool for the processing and analysis of small rnas datasets, reveals ubiquitous mirna modifications in human embryonic cells. *Nucleic Acids Res*, **38**(5).

Taft, R. J., Simons, C., Nahkuri, S., Oey, H., Korbie, D. J., Mercer, T. R., Holst, J., Ritchie, W., Wong, J. J., Rasko, J. E., Rokhsar, D. S., Degnan, B. M., and Mattick, J. S. (2010). Nuclear-localized tiny rnas are associated with transcription initiation and splice sites in metazoans. *Nat Struct Mol Biol*, **17**(8), 1030–1034.

Supplementary material

Implementation

All the steps in the pipeline are automatized and implemented in SeqBuster 2.0 developed in JAVA platform. The new SeqBuster 2.0 has a main panel showing 5 different sections aimed at: 1) create new projects or load an existent one; 2) upload raw data or trimmed sequences; 3) run miRNA analyses; 4) run non-miRNA analyses and 5) to manage data to reset, remove or browse your processed data. Moreover, a real-time guide has been integrated offering help for each option and step. Help options display a short guide on how to perform a simple analysis of your data (Supplementary figure 1A). Specifically the SeqCluster extension panel that appears after clicking on the 'non-miRNA detection' button, is divided in 3 sections (Supplementary figure 1B): 1) load new genomes and annotation files; 2) set parameters for the analysis, such as the sample to be selected or the genome to use; and 3) select which steps to do for sRNA detection (or repeat in case of re-analysis). Fasta files containing the custom genome may be loaded to the tool. Moreover, any track (genes, repeats, mRNA,etc.) corresponding with a loaded genome can be integrated into the extension, permitting a complete custom analysis of the non-miRNAs sRNA transcriptome. In the case of UCSC gene tracks the extension extracts different types of information, including the transcription start site, splicing sites, exons, introns and promoters External sources are needed to have a complete functionality: blast repository is required for the mapping step, MySQL for the storage information and R statistical packages (standard, RmySQL and RXML) for posterior analysis. The time consumed is directly related to the second process in the framework where sequences are mapped onto a genome. The size of the genome and the number of cores used here are determinant for the total time that one sample requires to be completely analyzed. Each sample was processed in 4 hours, using a workstation (hp xw9300) with 8Gb of RAM memory and 4 cores, consuming 3 hours the mapping step.

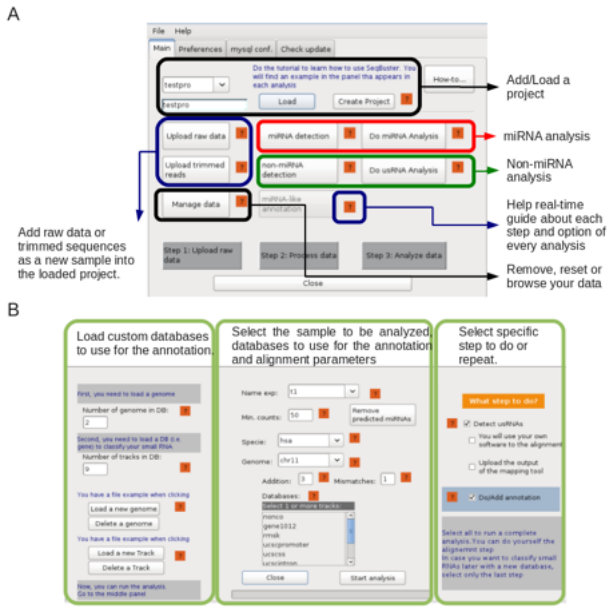


Figure 2.1: SeqBuster and SeqCluster integration. SeqBuster 2.0 and SeqCluster extension interface is developed using JAVA platform (1.6 version). A) SeqBuster panel shows 5 different sections for: 1) creating new projects (black up box); 2) uploading raw data or trimmed sequences (blue box), 3) miRNA analysis (red box), 4) non-miRNA analysis (green box); and 5) managing data to reset, remove or browse your processed data (black down box). Moreover, a real-time guide has been integrated offering help for each option and step of each analysis. B) SeqCluster extension panel is divided in 3 sections for: 1) loading new genomes and annotation files (left box); 2) setting parameters for the analysis, such as sample to be selected, genome to use ... (middle box); and 3) selection of steps to do for the non-miRNA detection (or to repeat in case of re-analysis) showed in the right box.

miRNA detection

Each fragment of 8 nt in length (seed) that appears on read sequences will be saved in a hash table structure to speed up the searching process. Then, the miRNA precursor sequences are read using windows of 8 nucleotides in order to find the previous stored seeds. When a seed is spotted at a precursor sequence, the read sequence that contains such seed and the current precursor region are fully compared to decide whether this position can be recorded as a hit of the given

Table 2.1: Comparison of mapping tools

Program	Num sequences	Time	Annotated	Pre-indexation
Blast	250000	65 min	17550	YES
ZOOM	250000	6 seg	16351	NO
SeqBuster	250000	15 seg	21350	NO

Comparison of mapping tools. Blast was used with the following parameters: word size = 7 and threshold identity= 85%. Zoom was run allowing 3 mismatches. The last column refers to the needed of database index creation previously to the mapping step.

read sequence. A hit will be considered if the read sequence is found on the precursor allowing 0/1 mismatch and up to 3 nt trimmed at the 3'-end of the sequence without matching the precursor sequence. The output stores the best hits for each read sequence with additional information about variation of the sequence when compared to the annotated miRNA. The algorithm takes 1 minute to map 200.000 sequences onto miRBase database, retrieving 99% of concordance with the blast tool (Supplementary table 1). A command-line version of this module has been released for users only interested in this step. This module needs a fasta file with the sequences to be mapped, the precursor sequences in fasta format and the position of the miRNAs on the precursors (files named hairpin.fa and miRNA.str in the miRBase repository, respectively). After that and as an optional step, the user may load data from any custom miRNA prediction pipeline to remove these sequences from the analysis. The unique information needed is the sequences predicted as putative miRNAs.

Decision algorithm

When two pre-usRNAs map on two different locations and these pre-

usRNAs only share one of the two loci, they go through a recursive algorithm to try to solve the inconsistency. The algorithm uses the percentage of overlap between pre-usRNAs as the main parameter to solve the inconsistency and select the more realistic scenario. This module will retrieve the minimal number of usRNAs made of the maximum number of pre-usRNAs with a consistent genome location using as the main parameter the percentage of overlap between pre-usRNAs. A score will be assigned indicating the reliability of the usRNA according to the number of cycles needed to solve inconsistency, giving 1 to the best score and increasing the value proportionally to the complexity of solving the cross-mapping event. For instance, in the latter case, where two pre-usRNAs have two genome locations and only one shared by both, the algorithm will assign the common location as the unique genome hit removing the remaining locations (see Supplementary figure 2). When pre-usRNAs are not solved by the previous modules, they are ignored for downstream analyses due to an inconsistent common origin.

usRNA Classification

The challenge in this step of the analysis is to avoid the exclusion of us-RNAs showing multiple locations onto the genome. This is, for instance, an important problem in miRNAs detection since many currently known miRNAs (45 human miRNAs) share genome location with TEs. This produces multiple locations in the annotation step and consequently they are not detected using a normal framework. For this reason, SeqCluster extension has integrated a final step to classify those usRNAs having multiple genome locations. In this step the consistency context is studied, meaning that if all the locations share the same context, that usRNA is directly classified. For instance, if all the locations of an usRNA map onto a ncRNA, that usRNA is labeled as ncRNA-usRNA. On the contrary, if the consistency is not observed and the usRNAs map onto several types of databases, usRNAs remain

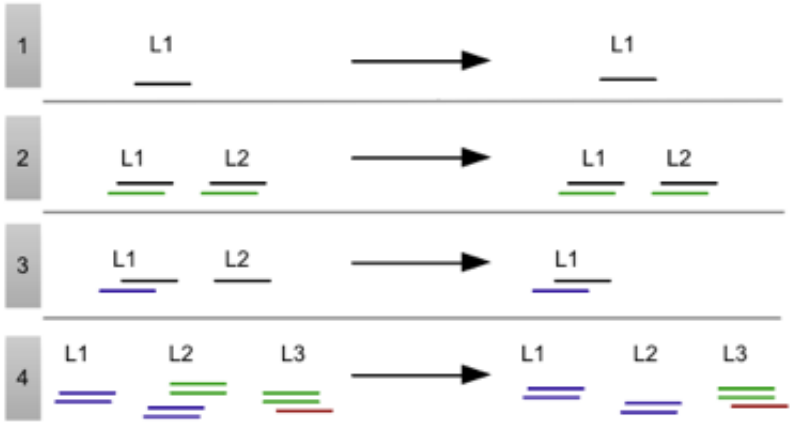


Figure 2.2: usRNA definition. Pre-usRNAs are represented by horizontal lines of different colors indicating different pre-usRNAs. Loci are indicated by the label ÒLÓ +ÓNumberÓ. 1) One pre-usRNA mapping into one location will be defined as one consistent unambiguous usRNA. 2) Two pre-usRNAs mapping in two locations, and sharing these two locations will be merged and defined as one consistent, ambiguous usRNA. 3) Two pre-usRNAs mapping in two locations, and sharing one location will be solved recursively until achieve the more realistic situation. In this case, the two pre-usRNAs will be merged into one consistent unambiguous usRNA. 4) A more complex case of the previous situation showing inconsistency. In a consistent scenario the three pre-usRNAs should map on the same locations, however, the short sequences provoke wrong mapping that may be corrected using nearest sequences information. To solve this, the overlap between pre-usRNAs on the same region is taken into account. Here, green group merges with the red group due to a higher overlap with red sequence than with purples sequence. As a consequence, one of the locations (L2) will be removed from the green pre-usRNA information since the more realistic scenario is that red and green pre-usRNAs go together. At the end, this group of pre-usRNAs will be defined as two consistent usRNAs, one mapping on L3 and the other on L1 and L2.

unclassified but not removed from the database, thus permitting the visualization and analysis of these specific cases. Furthermore, each class of genome context may be divided in several groups to extend the classification to more specific subtypes and etc provide a complete view of the genome context for each usRNAs. For instance, transposon elements may be divided into: LINE (long interspersed repetitive DNA), SINE (short intersperded repetitive DNA) or LTR (long terminal repeat).

Output scheme

The framework generates a MySQL table showing, for each usRNA, a unique identifier, number of reads, number of locations, coordinates and the identification according to the annotation step (Supplementary figure 3). SeqCluster performs a general analysis, offering standard outputs using the previous table with the following graphic and plain-text files: distribution of the different types of usRNAs, length distribution for each type of usRNAs, position in the putative precursors of each usRNAs type, and fasta files divided by types of usRNAs. Additionally, a BED file compatible with genome browsers will be generated with all the information to allow a user-friendly visualization and better comprehension. Furthermore, SeqCluster permits differential expression analysis between two samples or two groups of samples in different biological contexts to highlight those sRNAs with possible relevant functions. When the number of samples in each are higher than 10, a t-test and one-way-ANOVA (analysis of variance) are integrated to calculate the expression difference. Otherwise, a custom analysis is applied, considering only usRNAs that are not differentially expressed between intra-group samples. All samples of each group are compared between them, in pairs, calculated for each usRNA a ratio and p-value. If a 75% of the comparisons of a given usRNA do not indicate a significant differential expression, that usRNA is kept for the comparison with the other group. UsRNAs have to pass those filters in both groups to be included in the analysis. Datasets involving time series can be also analyzed using specific statistical methods to determine which sRNAs

correlate with a selective time point according to (Somel et al., 2010).

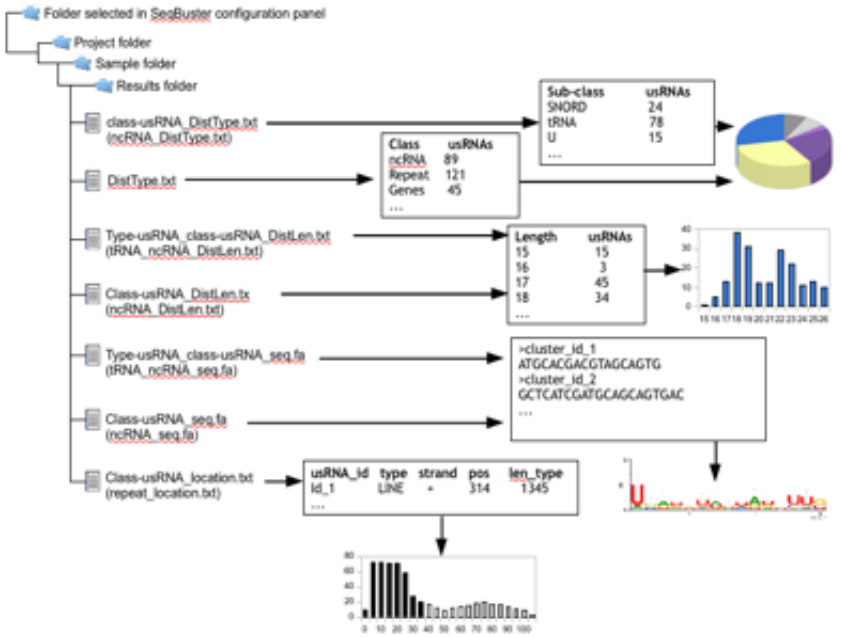


Figure 2.3: Output scheme. All files generated by SeqBuster will be stored in a folder that the user configured the first time the program is launched. For each project one folder will be created and for each sample added to the project another folder inside the project folder will appear. All the results will be stored in the sample folder, in the result section. For the non-miRNA sRNA analysis the following files will appear containing: proportion of usRNAs classified in each class/ sub-class with the suffix 'DistType.txt', length distribution of the usRNAs for each class/sub-class with the suffix 'DistLen.txt', usRNAs sequences for each class/sub-class with the suffix 'seq.fa', and finally the location of each usRNAs in their corresponding source molecule when they are classified to some group with the suffix 'location.txt'.

SeqCluster application to real datasets

We have applied SeqCluster extension to human brain samples sequenced by illumina 1G in our previous work (Marti et al.,2010). These corresponded to the frontal cortex (FC) and the striatum (ST). Furthermore, we also used published data from different species sequenced through the same technology to avoid methodology biases. The data

selected were human stem cells (SC) (Morin et al.,2008), mouse cell lines (MB) (Tam et al.,2008), *D. melanogaster* embryos (DM) (Chung et al., 2008) and *C. elegans* cells (CE) (Batista et al., 2008). The usRNAs were classified according to genome content: nasRNAs when mapping to non-coding, rasRNAs when mapping on repeat elements, gasRNAs when mapping on genes, iasRNAs when mapping on intergenic regions and nusRNAs if they were not classified (Supplementary figure 4). In mammals, nasRNAs were highly represented with a 30% of the total data, followed by rasRNAs in human samples but not in mouse, where the proportion of this type is lower. In mouse, the second most abundant class was represented by gasRNAs (40%). In *D. melanogaster* and *C. elegans*, nasRNAs were poorly represented (6%), being the more important class rasRNAs for *D. melanogaster* (75%) and iasRNAs for *C. elegans* (50%). Deeper analyses were run by SeqCluster, retreiving the more abundant classes in nasRNAs. In this case, tRNAs (78 uRNAs) and SNORD (45 uRNAs) mapping on the sense transcript were the more abundant, suggesting that those molecules could generate sRNAs, as previously described described (Kiss, 2002; Bachellerie et al., 2002; Niwa and Slack, 2007; Luo and Li, 2007; Weber, 2006; Smalheiser and Torvik, 2005; Scott et al., 2009; Kawaji et al., 2008; Taft et al., 2009a; Ender et al., 2008; Ono et al., 2011). The LINE class was the more represented type (53 uRNAs) among the rasRNAs, being 36 complementary to this TE sequences and 17 lying on sense transcripts. Other minor classes of usRNAs have been found with particular features. Among them, we highlight usRNAs that lay onto exon/exon gene junctions as reported previously (Affymetrix ENCODE Transcriptome Project ,2009). Other types of usRNAs were annotated at the splicing site region of the genes, also described in previous work (Taft et al., 2010).

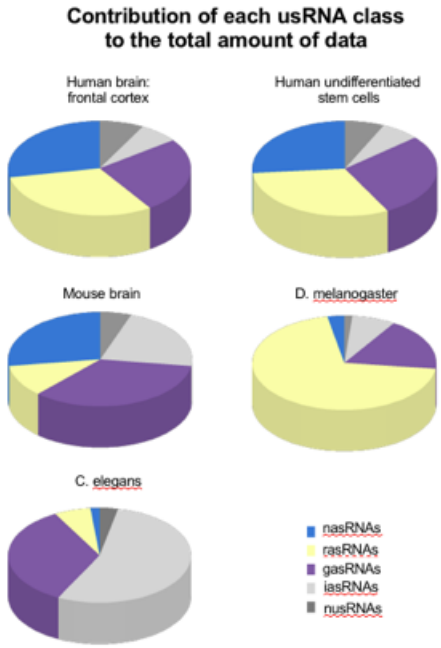


Figure 2.4: General results of SeqCluster extension. The proportion of each usRNA classes in each sample is represented. The classes used where gasRNA (genomic region), nasRNA (non-coding RNA), rasRNA (repeat elemetns), iasRNAs (intergenic regions) and nusRNAs (unclassified). The samples showed are 1) frontal cortex from human, 2) undifferentiated stem cells from human, 3) cell lines NIH3T3 from mouse, 4) adult female head from D. melanogaster and 5) larves from *C. elegans*. (See methods fro further information). Striatum from human brain showed a equal distribution than the frontal cortex sample.

2.3. A myriad of miRNA variants in control and Huntington's disease brain regions detected by massively parallel sequencing.

We have applied the potentiality of SeqBuster analysis to real data generated in our group belonging to patients with Huntington Disease and unaffected control individuals. HuntingtonÕs disease (HD) is a neurodegenerative genetic disorder that affects muscle coordination and leads to cognitive decline and dementia. In this study, we characterized the miRNA population in two affected brain regions: the frontal cortex and the striatum,by sequencing and microarray technologies, allowing a complete analysis of miRNA fraction. SeqBuster was used for the characterization of miRNAs genes in these samples, and miRNA expression deregulation obtained through SeqBuster analysis was validated in additional samples by microarray analysis. Our results corroborated the power of SeqBuster when used for deciphering the causes and/or consequences of deregulation of the sRNA population in relevant neurodegenerative diseases.

The results of this study led to the publication of the following article:

A myriad of miRNA variants in control and Huntington's disease brain regions detected by massively parallel sequencing.
Marti E, **Pantano L**, Banez-Coronel M, Llorens F, Minones-Moyano E, Porta S, Sumoy L, Ferrer I, Estivill X.
Nucleic Acids Res. 2010 Nov 1;38(20):7219-35. Epub 2010 Jun 30.

Supplementary material

LEGENDS OF SUPPLEMENTARY FIGURES

Figure S1. miRNA size distribution in C-FC, HD-FC, C-ST and HD-ST brain samples.

Figure S2. miRNA distribution in C-FC, HD-FC, C-ST and HD-ST brain samples, according to abundance range

Figure S3. (A) Histogram displaying the percentage of miRNAs with different types of isomiRs in human cells (left panel) and mouse brain (right panel), small RNA libraries obtained from immunoprecipitated Ago-1 and/or Ago-2 (GEO public database http://www.ncbi.nlm.nih.gov/geo/ based on the papers by Ender et al., Mol Cell 2008; 32(4):519-28 and Schwamborn et al., Cell 2009; 136 (5):913-25). (B) Examples of isomiRs commonly found in Ago-IPs and in the sequenced brain samples; (+) and (-) indicate the presence and absence of the indicated isomiR in the samples.

Figure S4. Sequencing performance in every sample. In each sample, the total number of reads was ordered according to decreasing frequency. The logarithm of the count number is represented for C-FC, HD-FC, C-ST and HD-ST brain samples. The p-values below are obtained using the non-parametric Willcoxon test for paired frequency distributions,

Figure S5. Huntington disease canonical pathway as described in the Ingenuity Pathway Analysis. The molecules identified as putative targets of the downregulated miRNAs/isomiRs are highlighted in grey.

LEGENDS OF SUPPLEMENTARY TABLES

Table S1. Characteristics of the human brain samples analyzed in the different approaches.

Table S2. Abundance of the human specific or evolutionary conserved miRNAs with respect to the total miRNA reads.

Table S3. Novel predicted miRNAs using the microPred algorithm. Only sequences commonly found in all the samples are included.

Table S4. List of miRNAs that do not present isomiRs.

Table S5. Examples of variants that are fare more abundant (> 80 %) than the corresponding reference miRNA.

Table S6. List of nucleotide substitution variants present in all the samples

Table S7. Nucleotide substitution variants that overlap with single nucleotide polymorphisms (SNP).

Table S8. miRNAs up-regulated in HD-FC and/or HD-ST.

Table S9. miRNAs down-regulated in HD-FC and/or HD-ST.

Table S10. Expression pattern of miRNAs whose expression has been shown to be altered in HD samples and/or HD mouse models.

Table S11. miRNAs deregulated in different neurodegenerative disorders.

Table S12. Correlation between the expression pattern of the isomiRs and the corresponding reference miRNAs. Only differently expressed sequences reaching statistical significance were considered ($p<0,05$, Hochberg and Benjamini correction). 3'-IsomiRs include 3'-trimming and 3'-addition variants; 5'-IsomiRs include 5'-trimming variants and nucleotide-substitution variants affecting the seed region. The isomiRs that present a discordant expression pattern with respect to the reference miRNA are highlighted in red.

Table S13. Top transcription factors with a possible role in miRNA expression deregulation.

Table S14. Predicted targets of the HD-downregulated miRNAs and seed

region IsomiRs involved in HD canonical pathway (p = 5,83E-03) as described in the IPA in comparison with the HD gene expression deregulation reported by Hodges et al. (2006).

Table S15A. Number of overlapping genes in HD deregulated transcriptome (Hodges et al., 2006) and the experimentally validated targets for the HD-deregulated miRNAs.

Table S15B. Identification of the overlapping genes in HD deregulated transcriptome (Hodges et al., 2006) and the experimentally validated targets for the HD-deregulated miRNAs ID of overlapping genes.

Table S16. Top biological functions and canonical pathways for predicted targets of HD-deregulated seed-region isomiRs using the IPA tool.

Figure S1

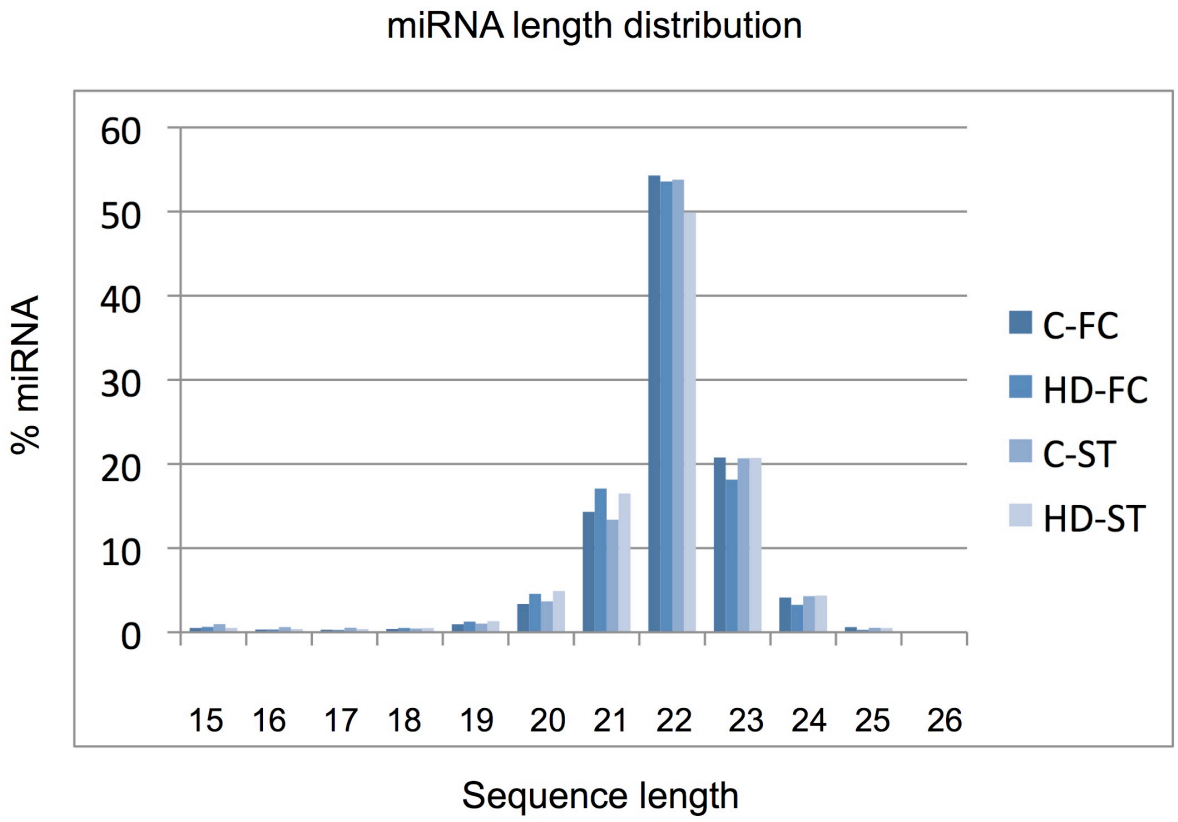

miRNA length distribution
% miRNA
60
50
40
30
20
10
0
15
16
17
18
19
20
21
22
23
24
25
26
Sequence length
C-FC
HD-FC
C-ST
HD-ST

Figure S2

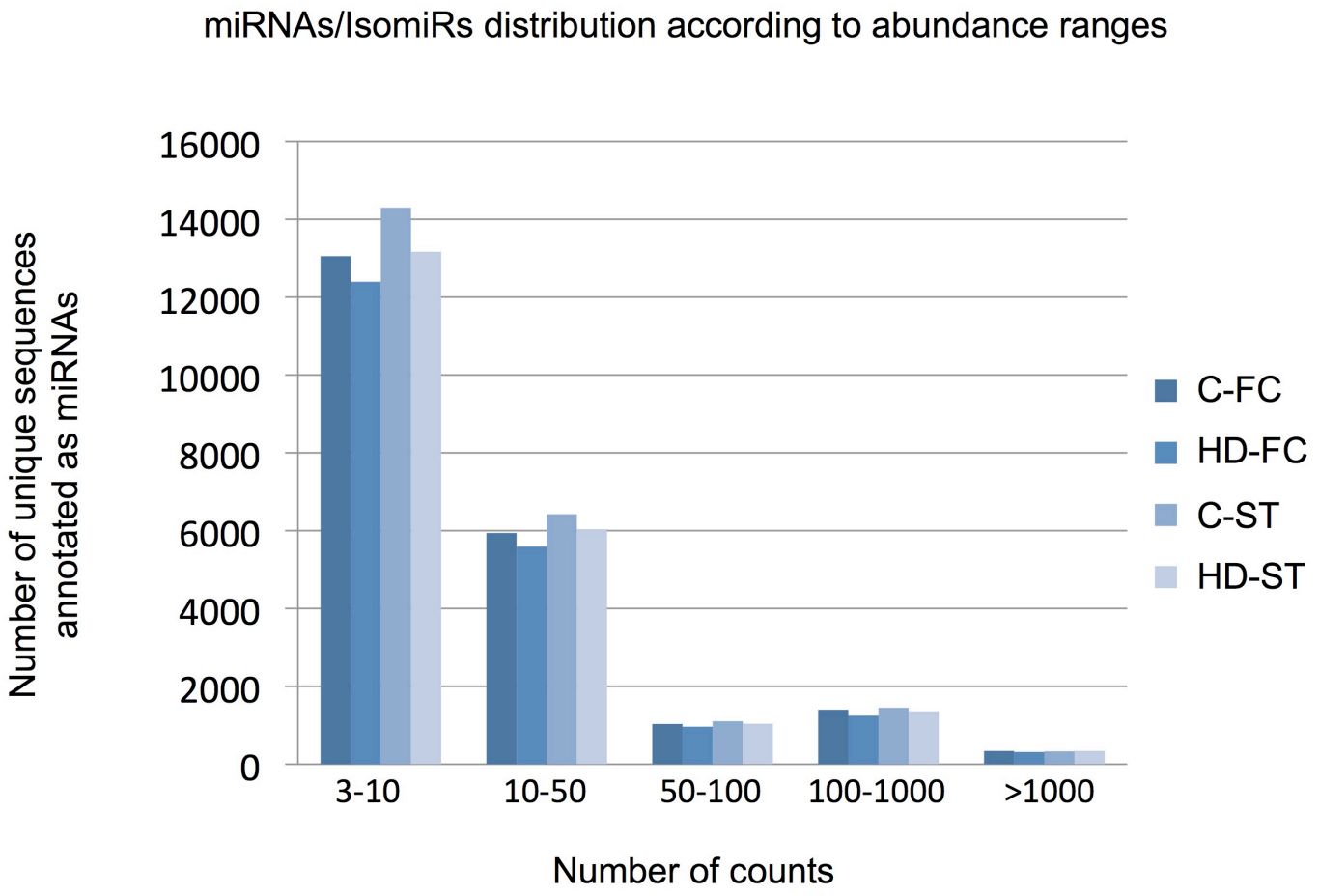
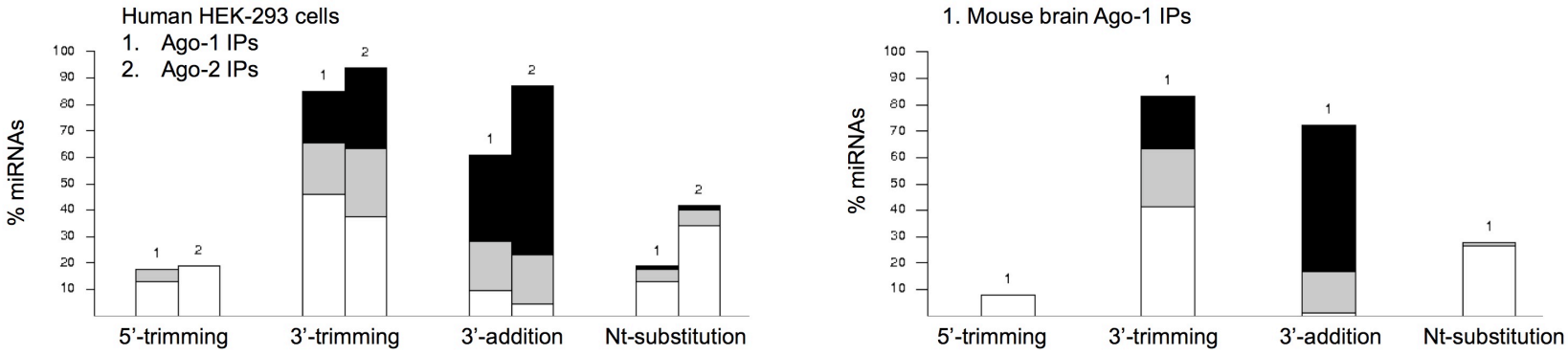

miRNAs/IsomiRs distribution according to abundance ranges
Number of unique sequences annotated as miRNAs
16000
14000
12000
10000
8000
6000
4000
2000
0
3-10
10-50
50-100
100-1000
>1000
Number of counts
C-FC
HD-FC
C-ST
HD-ST

Figure S3

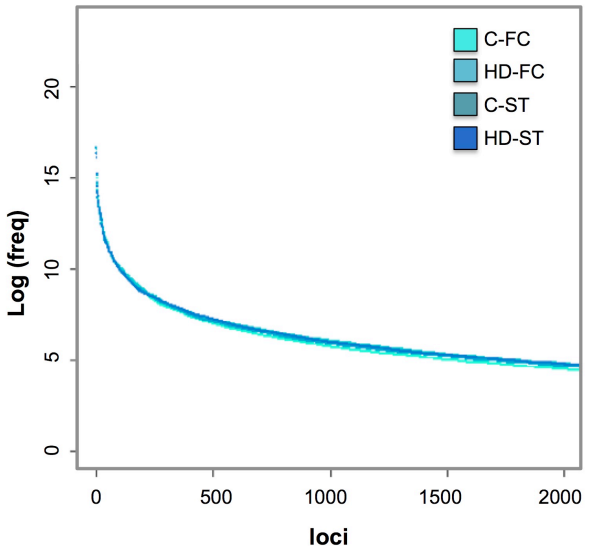
A. Proportion of miRNAs with different types of isomiRs in Ago-IP small-RNA datasets



B. Examples of IsomiRs commonly found in Ago-Ips and in the sequenced human brain samples

miRNA	Type of isomiR	sequence	human Ago1-Ips	human Ago2-Ips	mouse-brain-Ago1-Ips	FC-C	FC-HD	ST-C	ST-HD
mir-103	3'-trimming	AGCAGCAUUGUACAGGGCUAU	+	+	-	+	+	+	+
miR-106b	3'-trimming, 3'-addition	UAAAGUGCUGACAGUGCAGAAU	+	+	+	+	+	+	-
miR-17	3'-trimming	CAAAGUGCUUACAGUGCAGGU	+	+	+	+	+	+	+
mir-126	5'-trimming	CGUACCGUGAGUAAUAAUGCG	-	+	+	+	+	+	+
miR-92a	3'-addition	UAUUGCACUUGUCCCGGCCUGUAAU	+	+	-	+	+	+	+
miR-17	5'-trimming, 3'-trimming	UCAAAGUGCUUACAGUGCAGGU	+	+	-	+	+	+	+
mIR-92a	5'-trimming	AUUGCACUUGUCCCGGCCUGU	+	+	-	+	+	+	+

Figure S4



Samples	C-FC	HD-FC	C-ST	HD-ST
C-FC	1	0,812	0,904	0,846
HD-FC	0,812	1	0,733	0,667
C-ST	0,904	0,733	1	0,933
HD-ST	0,846	0,667	0,933	1

P values (Willcoxon test)

Figure S5

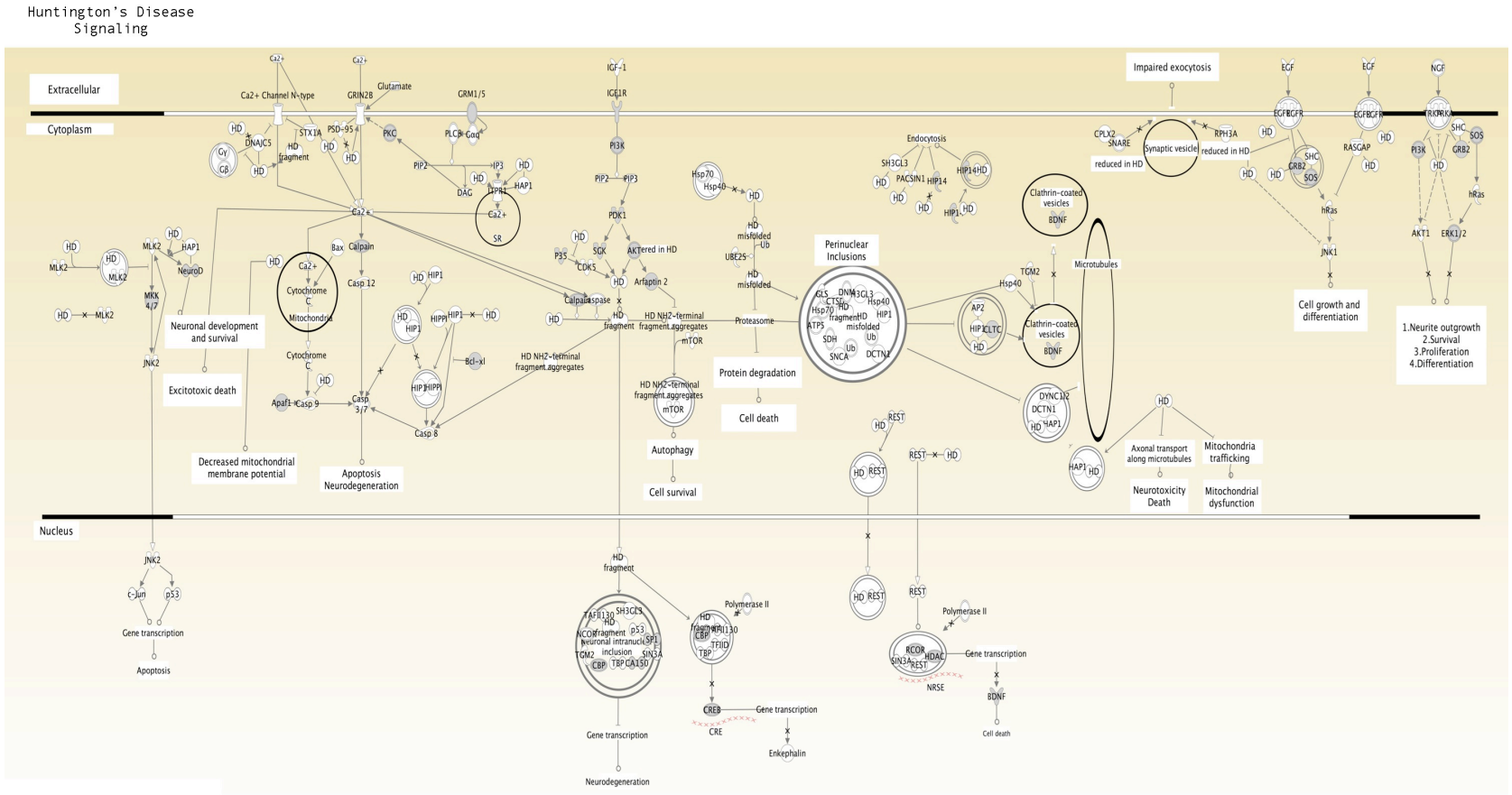

Huntington's Disease Signaling
Extracellular
Cytoplasm
Nucleus
Ca2+ Channel N-type
GRIN2B
Glutamate
GRM1/5
IGF-1
IGF1R
PI3K
PIP2
PIP3
PDK1
AKT
PLCγ
Gαq
IP3
DAG
PKC
Ca2+
SR
Calpain
Caspase
Bax
Cytochrome C
Mitochondria
Casp 12
Casp 9
Apaf-1
Casp 3/7
Casp 8
Bcl-xl
HIP1
HIPPI
MLK2
MKK 4/7
JNK2
NeuroD
HAP1
DNAJC5
STX1A
PSD-95
HD fragment
Neuronal development and survival
Excitotoxic death
Decreased mitochondrial membrane potential
Apoptosis Neurodegeneration
HD NH2-terminal fragment aggregates
mTOR
Autophagy
Cell survival
Arfaptin 2
P35
CDK5
SGK
Hsp70
Hsp40
UBE2S
Ub
Proteasome
Protein degradation
Cell death
Perinuclear Inclusions
DNM3CL3
CTSD
SDH
SNCA
DCTN1
ATP5
Endocytosis
SH3GL3
PACSIN1
HIP14
HIP1
AP2
HIP1
CLTC
Clathrin-coated vesicles
BDNF
TGM2
Microtubules
DYNC1I2
DCTN1
HAP1
Impaired exocytosis
CPLX2
SNARE
Synaptic vesicle
RPH3A
reduced in HD
EGF
EGFR
RASGAP
SHC
GRB2
SOS
hRas
JNK1
Cell growth and differentiation
NGF
TRKA
PI3K
AKT1
ERK1/2
1.Neurite outgrowth
2.Survival
3.Proliferation
4.Differentiation
Axonal transport along microtubules
Mitochondria trafficking
Neurotoxicity Death
Mitochondrial dysfunction
REST
RCOR
SIN3A
HDAC
NRSE
BDNF
Cell death
Polymerase II
TAFII130
SH3GL3
NCOR
p53
SP1
SIN3A
TGM2
TBP
CA150
CBP
TFIID
CREB
CRE
Gene transcription
Enkephalin
Neurodegeneration
c-Jun
p53
Apoptosis

Table S1: Characteristics of the brain samples analyzed in the different approaches

	AGE	SEX	GENOTYPE (1)	DISEASE STAGE (2)	FRONTAL CORTEX	CAUDATE
CONTROLS						
A07-42	53	M	-	-		qPCR
A07-44	39	M	-	-	Arrays	Arrays & qPCR
A05-5	58	M	-	-	Arrays	Arrays & qPCR
A05-18	85	F	-	-	Seq & Arrays	Seq & Arrays & qPCR
A-05-38	79	F	-	-	Seq & Arrays	Seq & Arrays & qPCR
HD PATIENTS						
BK387	28	F	62	4	Arrays	Arrays & qPCR
BK413	71	M	41	4	Arrays	Arrays & qPCR
BK518:	65	F	42	4	Seq & Arrays	Seq & Arrays & qPCR
BK 801	59	M	44	4	Seq & Arrays	Seq & Arrays & qPCR
BK 909	60	M	43	4		qPCR

(1) Includes the largest allelic CAG repeat in the HTT gene for each HD patient. (2) Disease stage according to the Vonsattel classification (J. Neuropathol Exp. Neurol., (1985) 44(6): 559-77; J. Neuropathol Exp. Neurol., (1998) 57(5): 369-84). The samples used in the sequencing study (Seq), the Agilent miRNa arrays (Array) and/or qPCR using TaqMan assays (qPCR) are indicated in the last two columns.

Table S2. Abundance of the human specific or evolutionary conserved miRNAs with respect to the total miRNA reads.

Human specific miRNAs	Freq. C-FC	Freq. HD-FC	Freq. C-ST	Freq. HD-ST
Total miRNA freq.	5035324	4019202	4558525	4374587
hsa-miR-1180	258	284	195	277
hsa-miR-1228	-	-	3	-
hsa-miR-1229	-	4	3	5
hsa-miR-1231	10	8	3	4
hsa-miR-1252	-	4	-	-
hsa-miR-1260	163	121	410	219
hsa-miR-1261	56	36	23	53
hsa-miR-1268	25	6	4	-
hsa-miR-1270	23	6	19	7
hsa-miR-1277	75	87	59	115
hsa-miR-1304	-	-	3	-
hsa-miR-1308	425	1207	1272	2058
hsa-miR-1826	3	6	-	-
hsa-miR-1908	39	21	23	26
hsa-miR-1911	-	-	3	-
hsa-miR-1912	-	-	32	21
hsa-miR-1974	846	1673	998	1567
hsa-miR-1975	147	77	180	168
hsa-miR-1976	-	-	-	3
hsa-miR-1977	-	-	3	3
hsa-miR-1978	80	131	50	87
hsa-miR-1979	349	318	395	526
hsa-miR-2110	492	283	312	221
hsa-miR-585	-	4	-	-
hsa-miR-629	40	24	46	49
hsa-miR-941	203	140	165	149
% (total miRNA reads)	0.06	0.11	0.09	0.13
Conserved miRNAs (37 different species*)				
hsa-let-7a	931597	562218	926233	755056
hsa-let-7b	410192	301072	506537	573261
hsa-let-7c	304352	120969	248063	160923
hsa-let-7d	54915	37267	46820	34688
hsa-let-7e	89217	34680	59914	35208
hsa-let-7f	884390	738614	718224	620114
hsa-let-7g	174895	195883	129265	166726
hsa-let-7i	109527	92375	79719	114616
hsa-miR-124	91155	63366	45211	17791
hsa-miR-133a	163	217	68	112
hsa-miR-133b	20	3	-	-
hsa-miR-92a	18138	24310	24622	34132
hsa-miR-92b	19644	14331	36319	27337
% (total miRNA reads)	61.3	54.4	61.9	58.1

miRNAs showing a frequency ≥3 are listed

***Species shearing conserved miRNAs**

Anopheles gambiae
Apis mellifera
Bombyx mori
Bos taurus
Branchiostoma floridae
Canis familiaris
Ciona intestinalis
Ciona savignyi
Danio rerio
Drosophila ananassae
Drosophila erecta
Drosophila grimshawi
Drosophila melanogaster
Drosophila mojavensis
Drosophila persimilis
Drosophila pseudoobscura
Drosophila sechellia
Drosophila simulans
Drosophila virilis
Drosophila willistoni
Drosophila yakuba
Fugu rubripes
Gallus gallus
Homo sapiens
Lottia gigantea
Macaca mulatta
Monodelphis domestica
Mus musculus
Ornithorhynchus anatinus
Pan troglodytes
Rattus norvegicus
Saccoglossus kowalevskii
Schmidtea mediterranea
Strongylocentrotus purpuratus
Tetraodon nigroviridis
Tribolium castaneum
Xenopus tropicalis

Table S3. Novel predicted miRNAs commonly found in all the brain samples.

Sequence (1)	Length (2)	Freq C-FC (3)	Freq HD-FC (3)	Freq C-ST (3)	Freq HD-ST (3)	Number of seq (4)	Number of loci	Chromosome	Host genes of sequences mapping onto introns	Host genes of sequences mapping onto	Sequences mapping onto "mRNA" database	Sequences mapping onto "ncRNA"	Sequences mapping onto "repeat"	Host genes of intronic small-RNAs expressed	Host genes of intronic small-RNAs expressed
GGCUGGUCCGAAGGUAGUGAGUUAUC	21	3796	3367	7089	4986	6	2	chr7,chr11,	,ARNTL,	,	,	,	scRNA,	ARNTL	ARNTL
UCGACCGGACCUCGACCGGCUCGU	21	2606	2505	1107	1857	6	1	chr10,	USMG5,	,	,	,	,	USMG5,	USMG5,
UAA**ACCAUUCGUAGACGACCUGCUUC**	23	1948	499	891	372	10	1	chr9,		,	,	,	others,		
C**UUCUCACUACUGCACUUGACUAGUCU**	26	1907	1271	2726	2462	14	24	chr7,chr14,chr1	,RALA,SNX4,SORD,FAI	,PAQR5,	,	,	scRNA,	SORD	SORD
AAACCAUUCGUAGACGACCUGCUUC	25	1846	352	759	249	5	1	chr9,		,	,	,	others,		
GGUUCCAUAGUGUAAUGGU	24	1834	1638	661	782	8	3	chr8,chr11,chr3	B3GAT3	NEFL,	,	tRNA,	,		
GUUAAGAUGGCAGAGCCCGGUAAUC	22	1510	2187	412	702	6	1	chr4,		,	,	tRNA,	,		
GCCCGGAUAGCUCAGUCGGUAGAGAC	22	1370	1703	1438	2094	7	6	chr1,chr11,chr6,	chr17,	DGCR8,	,	tRNA,	,		
C**GGCUGGUCCGAGUGCAGUGGUGUUU**A	25	1227	1578	1349	3741	4	1	chr7,		,	,	,	scRNA,		
C**CCCCCACUGCUAAAUUUGACUGGCUU**	26	921	539	1123	726	8	10	chr7,chr20,chr8	MSN,ART1,WIPF2,C18	,	,	,	,scRNA,	MSN, ART1,TRP	MSN, ART1,TRPM
GCUAAACCUAGCCCCAAACCCACU	21	800	798	758	849	6	1	chr7,		,	,	,	,		
AAC**UUAACUUGACCGCUCUGACCA**	21	746	1064	516	546	10	1	chr7,		,	,	,	,		
CUGAGGUAGUAGGUUGUAUAGUUA	23	733	486	800	785	4	1	chr19,		,	,	,	,		
GUGGCACACGUAGGGCAACU	19	652	518	201	129	3	2	chr1,chr11,		,	,	,RNU11,	,LINE,		
GAAAAGUCAUGGAGGCCAUGGGGUU	26	599	331	236	408	4	1	chr1,		,	,	tRNA,	,		
GGUCCCAUGGUGUAAUGGUUAG	19	531	3378	613	1227	2	3	chr6,chr17,		,	,	tRNA,	,		
GGCCGCGUGGCCUAAUGGA	19	513	4192	447	1496	1	4	chr6,chr16,chr15,		,	,	tRNA,	,		
AAAGCAGGAUUCAGACUACAAUAU	24	466	261	461	276	7	1	chr3,		,ABCC10,C17orf68	SNORA4,	SNORA63,	,		
GUCUCUGUGGCGCAAUCGGU	20	463	1692	325	971	2	19	chr19,chr1,chr1	LOC728855,LOC7288	,	,NDUFS7,PSMB3,	tRNA,	,		
GGCCGGUUAGCUCAGUUGGU	19	431	399	1328	819	2	20	chr14,chr6,chr1	,CXorf57,ABCC10,C17	,	,C17orf59,	tRNA,	,LTR	ABCC10, NRXN1	ABCC10, NRXN2
CU**CCUACUUGGAUAACUGUGGUAAU**	23	405	242	148	143	4	1	chr21,		,	,	,	others,		
GGU**UCGUUUCCGGCUCGAAGGACCA**	22	347	307	202	407	5	1	chr2,		,	AGBL5,	tRNA,	,		
GACUCUUAG**CGGUGGAUCACUCGGCUCGUGCG**	23	318	493	287	521	4	1	chrY,		,	,	,	,		
GAUCGAUGAUGACUUCCAUAUAUA	19	310	434	271	292	6	5	chr15,		,	,	SNORD116-14,S	,		
U**GCCCCCAUGUCUAACAACAUGGCU**	24	272	233	387	302	3	1	chr8,	CPA6,	,	,	,	,	CPA6	
AGGAGAUGGAAUAGGAGCUUGCU	23	260	135	132	73	3	3	chr20,chr14,chr11,		NEFL,	,	,	others,		
UCCCUGGUGGUCUAGUGGUUAGGAU	19	259	630	377	886	7	14	chr1,chr6,chr3,(	PPP1R12B,PLCH1,	,	,	tRNA,	,	PPP1R12B,PLCH	PPP1R12B,PLCH1
GUUUCCGUAGUGUAGCGGUUAUCA	22	255	724	304	478	3	1	chr19,		,	,	tRNA,	,		
CACCUUGCGCUACUCAGGUCUGC	23	227	228	268	203	2	1	chr22,	OSBP2,	,	,	,	,	low expression,	low expression, C
GGUUCCAUAGUGUAAUGGU	21	193	455	125	181	3	3	chr8,chr11,chr3	NEFL,B3GAT3,	,	,	tRNA,	,	NEFL, B3GAT3	NEFL, B3GAT3
GGUUCCAUGGUGUAAUGGUUAGUGGU	22	192	1160	138	444	3	10	chr8,chr11,chr6	PDHX,	,	,RAB11A,HES7,	tRNA,	,	PDHX (brain)	PDHX (brain)
GUUUCCGUAGUGUAGUGGUUAUCAC	24	173	157	133	277	2	19	chr1,chr11,chr6,	chr3,chrX,chr2,chr5,	,	,MYNN,TRIM41,HMC	tRNA,	,		
GGCUCCGUGGCGCAAUGGAU	19	168	3011	207	1185	2	1	chr1,		,	,	tRNA,	,		
GCGUUGGUGGUAUAGUGGUGAG	22	165	405	183	313	3	14	chr7,chr14,chr1	,RCOR3,	,	,	tRNA,	,SINE	RCOR3	RCOR3
CUCUCGGCGCCCCCUCGAUGCUCU	24	161	158	261	319	6	1	chr7,		,	,	,	others,		
GCG**CCGCUGGUGUAGUGGUAUCAUGCAAG**	26	155	204	72	197	3	2	chr16,chr2,		,	,	,	,		
GUCUACGGCCAUACCACCCUGAA	23	142	93	70	62	3	85	chr7,chr20,chr1	,KIAA1467,HINT3,GA	,	,DUSP5P,	,DUSP5P,GRM8,	,others,	VAMP4, KIAA14	VAMP4, KIAA1467
GGUUCCAUAGUGUAAUGGU	20	138	237	153	290	5	3	chr8,chr11,chr3	NEFL,B3GAT3,	NEFL,	,	tRNA,	,	NEFL, B3GAT3	NEFL, B3GAT4
UUCUGGGCUGUAGUGCGCUAUGCCGAU	26	132	129	48	178	4	9	chr14,chr6,chr3	,GREB1,E2F3,PTPRD,A	,	,	,	LTR,others,	GREB1, E2F3, P	GREB1, E2F3, PTI
AU**GGAUUUUUGGAAAUAGGAGAUGGAA**	26	127	186	171	360	5	2	chr11,chr9,		,	,	,	others,		
AGCGGUCGGCGUCCCCCAACUUC	23	126	543	56	89	3	1	chr16,		,	,	,	others,		
CUUGACUGAAGCUGAUGAUGAGUUUG	23	122	160	89	131	5	1	chr20,	SNRPB,	,	,	SNORD119,	,	SNRPB,	SNRPB,
CAUUCUGAAAGAACGUGUGGAAAAC	24	117	68	88	97	2	1	chr5,	HSPA9,	,	,	SNORD63,	,	HSPA9,	HSPA9,
CUAUCUGAGAGAUGGUGAUGACAUUU	25	109	95	119	131	3	1	chr9,	RPL7A,	,	SNORD36B,	SNORD24,	,	RPL7A,	RPL7A,
UCGCGUGAUGACAUUCUCCGGA	22	106	260	67	559	1	1	chr16,	RPL13	HINT3,C15orf52,	,	SNORD68,	,	RPL13	RPL13
UGUCCUGAAGAGAGGUGAUGACUUAUGAC	25	104	91	70	37	3	17	chr15,			,	SNORD115-5,SI	,		
GAUCGAUGAUGAGUCCCCCAUAAAAAA	26	101	173	221	248	3	4	chr15,			,SNORD116-10,	SNORD116-9,SI	,		
GAGCCUCGGUUGGCCUCGGAUAG	23	97	78	55	39	2	2	chr22,chr9,	MAPK1		,	,	others,	MAPK1	MAPK1
AU**AGCGAGACCCCGUCUCUA**	18	92	352	105	289	3	94	chr7,chr20,chr1	FBXL2,DDX21,DDX27,KIF9,RAB3GAP2,N		DUFAF2,C8orf79,HNF	BAX	others,SINE,	NDUFAF2,GPR1	NDUFAF2,GPR18
GAGGAAUGAUGACAAGAAAAGGCCGA	22	92	223	148	346	3	1	chr2,			,	SNORD89,	,		
CAAGAACGAAAGUCGGAGG	19	89	327	83	138	1	1	chr2,			,	NCRNA00164,	others,		
AGUUGGUCCGAGUGUUGUGGGUUAU	24	81	57	87	125	2	2	chr7,chr12,			,	,	scRNA,		
U**UGGUCGUGGUUGUAGUCCGUGCGA**	23	75	169	63	308	4	1	chr5,			,	C5orf36,	,		
ACUCCAUGAUGAACACAAAAU	19	67	272	159	409	3	1	chr11,			SNORD25,	SNORD27,	,		
AAGUUUCUCUGAACGUGUAGAGCA	24	66	49	64	62	2	10	chr3,chr6,chr17	BAI3		,	U3-3,U3-2,U3-2	others,	BAI3	BAI3
UACCUGGUUGAUCCUGCCAGUAGC	24	62	199	95	142	3	2	chr21,chr9,			,	,	,others,		
AGUCGGUAGAGCAUGAGACU	20	58	58	51	145	1	8	chr19,chr1,chr6,	chr16,chr5,		,	tRNA,	,		
CGGGAGCGGGGGACCACCAGG	21	57	99	76	152	2	9	chr14,chr8,chr3	PLXND1,PTPRD,ANKRD28,SLC25A43,		,	,	LTR,others,	PLXND1, PTPRD	PLXND1, PTPRD,
AU**GUUAUGAUGAUGGGCGAAAU**	21	54	109	95	225	3	1	chr11,			SNORA8,	SNORD6,	,		
UUCGCGGGCGAAGGCAAAGUCG	22	54	24	40	42	2	1	chr1,			,	,	,		
GCGUUGGUGGUAUAGUGGUGAGCA	23	53	251	61	140	2	13	chr7,chr14,chr1	TSC1,MGAT5B,C19orf18,DPY19L1,		,	,	SINE,	MGAT5B, DPY19	MGAT5B, DPY19L
GGCUUUGGUGACUCUAGAUAAC	22	53	131	42	63	2	1	chr21,			,	,	others,		
CCACCUCCCCUGCAAACGUCCA	22	52	78	44	59	2	1	chr22,	DGCR8,		,	MIR1306,	,	DGCR8,	DGCR8,
CAACAAGUCACAGCCGGCCUCA	22	50	24	15	10	1	1	chr19,			,	,	,		
GCAUUGGUGGUUCAGUGGUAGAA	23	49	196	88	144	2	14	chr1,chr6,chr17	,VAC14,ABCB10,COLQ,		,HSPA6,	tRNA,	,	VAC14, ABCB10	VAC14, ABCB10,
CGGAUAGCUCAGUCGGUAGAGCAUCA	26	47	29	12	17	1	6	chr1,chr6,chr11,	chr17,		,	tRNA,	,		
UUUCGAUGGUAGUCGCCGUGCCU	21	41	82	62	94	3	1	chr21,			,	,	others,		
UGUGGGAAACUCGACUGCAUA	20	40	26	26	27	2	13	chr14,chr1,chr6	,PHACTR1,		,	,	others,	,PHACTR1,	,PHACTR1,
AU**GUGAACAGCAGUUGAACAU**	20	36	146	25	78	2	1	chr2,			,	,	others,		
GAUGCUGGUGGUUGGCACUCCUGGUG	25	35	46	73	38	2	1	chr11,	MALAT1		,	MALAT1,	,		
GGUAAAAUGGCUGAGUGAAG	20	33	423	15	41	1	3	chr14,chr1,chr2	AKAP6		,	,	,	AKAP6	AKAP6
UGGUUAGGAUUCGGCGCUCUCA	22	32	35	44	115	1	10	chr1,chr6,chr3,chr2,			,	tRNA,	,		

ACCUUGGCUCUAGACUGCUUACUG	24	31	11	12	13	1	1	chr17,		,	MIR212,	,			
GUAGUAGGUUGUGUGGUUU	18	29	54	29	58	2	1	chr22,		,	LOC400931,	,			
CAGUUGAACAUGGGUCAGUCGGUCC	24	29	50	23	53	2	1	chr17,		,	,	others,			
UGUCCUCUAGGGCCUGCAGUCU	22	25	29	16	25	1	1	chr22,	TOM1	,	TOM1,	LINE,	TOM1	TOM1	
ACGAAAGUCGGAGGUUCGAAG	21	23	196	40	219	1	1	chr2,		,	NCRNA00164,	others,			
GCCGUGAUCGUAUAGUGGU	19	22	78	31	55	1	11	chr1,chr11,chr6	SHF,DOCK10	,	tRNA,	,	DOCK10	DOCK10	
AGUGUGUCAGGAUGGCCGAGUGGUCU	26	21	28	25	279	1	1	chr6,		,	tRNA,	,			
UCGGGUGCGAGAGGUCCCGG	20	19	124	14	42	1	3	chr1,chr17,chr16,		,	tRNA,	,			
UCUGGCUGUUGUGGUGUGCAA	21	19	19	14	14	1	1	chr17,	DDX5	,	,	,	DDX5	DDX5	
CUGCUAAAUUUGACUGGCUU	20	11	11	18	19	1	31	chr7,chr20,chr1	,KIAA0146,TRPM6,MSN,FOXM1,CASC4,KI	,	,	scRNA,DNA,	TRPM6, KIAA14	TRPM6, KIAA140	
AAGCAGGGUCGGGCCUGGUU	20	11	18	14	15	1	68	chr7,chr20,chr2	RAI14,MED12L,KIAA1467,C1orf14,GTF2II	,	,DUSP5P,	others,	RAI14, NT5DC1	RAI14, NT5DC1,M	
GGGGGAUUAGCUCAAAUGGUAGA	23	10	10	15	29	1	2	chr8,chr2,	,SIAE,	AGBL5,	tRNA,	,			
GUUAGUGAUGAUUUUAAAA	19	10	87	95	129	1	1	chr1,		SNORD78,	SNORD79,	,			

(1) The different representatives of the putative miRNA (the trimming variants) have been collapsed in a single sequence, where the most abundant one is highlighted in bold In the "Sequence" column. (2) The "Lenght" column idicates that of the more abundant sequence of the set of representatives with the same core sequence. (3) Frequency in the different samples corresponds to the sum of the frequencies of all the representatives of the same core sequence. (4) The column "number of seq" indicates the total of different representatives of the same core sequence that vary in lenght. The "Host genes of intronic small-RNAs expressed in the FC and ST" is based on GNF Gene Expression Atlas 2 (PNAS 2004;101(16):6062-7). This contains two replicates each of 79 human tissues run over Affymetrix microarrays. The presence of a gene indicates that this gene was detected in the array.

Table S4. List of miRNAs that do not present isomiRs.

miRNA	Freq. C-FC	Freq. HD-FC	Freq. C-ST	Freq. HD-ST
hsa-miR-1197	21	27	22	-
hsa-miR-656	16	-	-	-
hsa-miR-592	10	-	-	-
hsa-miR-1256	8	-	-	-
hsa-miR-1270	7	-	-	3
hsa-miR-1231	4	-	6	7
hsa-miR-1276	4	-	-	-
hsa-miR-1323	4	6	3	3
hsa-miR-548o	4	-	-	-
hsa-miR-96	4	-	6	-
hsa-miR-1251	3	-	-	-
hsa-miR-196b	3	-	-	-
hsa-miR-372	3	-	-	-
hsa-miR-517a	3	-	-	-
hsa-miR-654-5p	3	-	-	-
hsa-miR-1255b	-	-	-	3
hsa-miR-1258	-	-	3	5
hsa-miR-1268	-	4	-	6
hsa-miR-508-3p	-	-	3	-
hsa-miR-516a-5p	-	-	4	-
hsa-miR-618	-	-	4	-
hsa-miR-758	-	-	-	17
hsa-miR-1229	-	3	-	-
hsa-miR-1266	-	-	-	6
hsa-miR-1267	-	-	-	4
hsa-miR-1269	-	-	-	5
hsa-miR-1283	-	6	3	-
hsa-miR-1284	-	-	-	17
hsa-miR-1295	-	-	-	7
hsa-miR-182	-	-	7	-
hsa-miR-188-3p	-	-	3	-
hsa-miR-296-5p	-	28	12	-
hsa-miR-299-3p	-	12	-	-
hsa-miR-431	-	4	3	3
hsa-miR-449a	-	-	-	5
hsa-miR-450a	-	-	-	3
hsa-miR-489	-	-	8	-
hsa-miR-491-3p	-	-	-	5
hsa-miR-509-3-5p	-	7	3	-
hsa-miR-516b	-	3	-	3
hsa-miR-521	-	-	11	28
hsa-miR-548e	-	-	-	3
hsa-miR-551a	-	-	3	-
hsa-miR-570	-	-	-	4
hsa-miR-642	-	-	27	-
hsa-miR-643	-	-	5	-
hsa-miR-651	-	3	-	-
hsa-miR-671-3p	-	-	12	-
hsa-miR-886-3p	-	-	3	-

Invariable miRNAs in bold are commonly found in at least three different samples. Green and orange colours highlight invariable miRNAs in FC or ST samples, respectively. MiRNAs showing a frequency ≥3 are listed.

Table S5. Examples of variants that are far more abundant (>80%) than the corresponding reference miRNA.

miRNA	Type of variant	Freq C-FC	Freq HD-FC	Freq C-ST	Freq HD-ST
Hsa-miR-1974	5'-trimming (-1)	273	552	332	469
	Reference	27	71	31	63
Hsa-miR-1979	5'-trimming (+1)	143	182	160	269
	Reference	3	5	7	5
Hsa-miR-324-3p	5'-trimming (-2)	50	87	50	65
	Reference	2	-	2	3
Hsa-miR-1291	5'-trimming (-1)	63	62	102	64
	Reference	-	-	-	-
Hsa-miR-1827	nt-substitution 6 G-->C	132	81	51	161
	Reference	-	-	-	-
Hsa-miR-320c	nt-substitution 11 G-->U	94	55	50	70
	Reference	-	-	2	4
Hsa-miR-1260	nt-substitution 9 G-->U	165	129	415	227
	Reference	-	-	-	-
Hsa-miR-376a	nt-substitution 6 G-->A	130	138	111	82
	Reference	16	16	16	9

The upstream (-1, -2) or downstream (+1) positions at which the 5'-trimming vaiants (5'-trimming) start with respect to the reference miRNA first position are indicated. For the nucleotide substitution variants (nt-substitution), the position and the type of nucleotide change are specified. IsomiRs showing a frequency ≥10 are listed

Table S6. List of nucleotide substitution variants present in all the samples.

miRNA	nt-subtitution pattern	Freq C-FC	Freq HD-FC	Freq C-ST	Freq HD-ST
Hsa-miR-215	1 A-->C	20	26	23	9
Hsa-miR-338-5p	1 A-->C	5	5	9	3
Hsa-miR-338-5p	3 C-->A	9	5	13	5
Hsa-miR-130a	4 U-->G	14	5	5	3
Hsa-miR-1827	4 G-->A	6	11	7	7
Hsa-miR-379	5 A-->G	140	133	20	47
Hsa-miR-411	5 A-->G	400	315	202	204
Hsa-miR-193a-5p	5 U-->G	6	5	8	8
Hsa-miR-1827	5 G-->A	8	4	3	3
Hsa-miR-1308	5 G-->U	325	1016	117	1774
Hsa-miR-584	5 G-->U	9	17	17	11
Hsa-miR-376a	6 A-->G	130	138	111	82
Hsa-miR-376b	6 A-->G	66	71	62	72
Hsa-miR-376c	6 A-->G	304	297	229	176
Hsa-miR-1827	6 C-->A	14	27	8	11
Hsa-miR-320c	6 C-->A	28	12	10	10
Hsa-miR-1827	6 C-->U	14	21	16	10
Hsa-miR-320c	6 C-->U	19	12	13	20
Hsa-miR-320c	6 C-->G	136	75	43	155
Hsa-miR-320d	6 C-->G	19	11	9	11
Hsa-miR-584	7 U-->G	13	13	10	11
Hsa-miR-195	7 C-->A	6	6	9	8
Hsa-miR-328	7 C-->A	6	6	3	4
Hsa-miR-499-5p	7 C-->A	9	9	12	6
Hsa-miR-124	8 C-->A	248	315	325	121
Hsa-miR-369-3p	8 C-->A	6	3	6	5
Hsa-miR-874	8 G-->U	14	11	10	7
Hsa-miR-1827	9 U-->C	6	5	28	3
Hsa-miR-1260	9 U-->G	165	129	415	227
Hsa-miR-1827	9 U-->G	8	8	16	3
Hsa-miR-338-5p	9 C-->A	3	14	32	21
Hsa-miR-340	9 C-->A	144	187	241	150
Hsa-miR-935	9 C-->A	16	22	7	6
Hsa-miR-338-5p	9 C-->U	7	6	12	7
Hsa-miR-1827	10 A-->G	69	20	21	24
Hsa-miR-377	10 A-->G	21	8	9	5
Hsa-miR-422a	10 A-->G	7	5	5	7
Hsa-miR-124	10 C-->A	968	460	376	129
Hsa-miR-338-5p	10 C-->A	18	16	33	30
Hsa-miR-410	10 C-->A	6	4	6	6
Hsa-miR-338-5p	10 C-->U	9	11	7	16
Hsa-miR-584	10 G-->U	8	19	15	12
Hsa-miR-98	11 A-->G	168	155	163	220
Hsa-miR-107	11 U-->G	236	96	90	42
Hsa-miR-129-5p	11 U-->G	63	20	29	17
Hsa-miR-320a	11 U-->G	432	261	430	312
Hsa-miR-320b	11 U-->G	20	4	59	12
Hsa-miR-320c	11 U-->G	94	55	50	70
Hsa-miR-320d	11 U-->G	26	19	36	16
Hsa-miR-584	11 C-->A	11	41	44	32
Hsa-miR-584	11 C-->G	12	20	33	20
Hsa-miR-1827	12 A-->G	23	8	21	14
Hsa-miR-1261	12 U-->A	52	28	29	57
Hsa-miR-223	12 C-->A	5	11	7	5
Hsa-miR-374a	12 C-->A	8	6	8	6
Hsa-miR-374b	12 C-->A	4	5	14	7
Hsa-miR-485-3p	12 C-->A	32	9	27	10
Hsa-miR-584	12 C-->A	24	44	34	25
Hsa-miR-584	12 C-->U	10	13	20	13

Hsa-miR-584	12 C-->G	13	12	15	10
Hsa-miR-941	12 G-->A	8	4	7	4
Hsa-miR-1290	13 U-->G	33	104	127	497
Hsa-miR-210	13 C-->A	5	10	9	8
Hsa-miR-338-5p	13 G-->U	2	4	7	6
Hsa-miR-449b	14 U-->C	7	3	24	5
Hsa-miR-338-5p	14 U-->G	5	3	6	10
Hsa-miR-532-5p	15 A-->G	10	4	9	4
Hsa-miR-584	17 A-->C	15	38	23	13
Hsa-miR-628-3p	18 U-->G	14	17	3	9
Hsa-miR-584	18 C-->U	6	4	9	5
Hsa-miR-124	19 C-->U	287	168	108	42
Hsa-miR-628-3P	19 C-->U	18	24	28	24
Hsa-miR-1	20 U-->G	207	100	64	41
Hsa-miR-361-5p	20 U-->G	10	10	10	15
Hsa-miR-320b	20 C-->A	8	15	14	20
Hsa-miR-33a	20 C-->A	48	114	67	116
Hsa-miR-129-3p	20 C-->U	757	1117	210	189
Hsa-miR-584	20 G-->A	11	18	7	28
Hsa-miR-584	20 G-->U	18	42	30	34
Hsa-miR-320b	21 A-->U	7	10	51	10
Hsa-miR-130b	21 A-->G	35	31	24	14
Hsa-miR-329	21 U-->A	17	13	7	10
Hsa-miR-654-3p	21 U-->A	16	14	11	9
Hsa-miR-329	21 U-->C	12	8	3	4
Hsa-miR-126	21 C-->A	11	13	8	9
Hsa-miR-382	21 C-->A	20	9	11	7
Hsa-miR-503	21 C-->A	14	9	7	14
Hsa-miR-598	21 C-->A	710	544	834	263
Hsa-miR-935	21 C-->A	15	23	7	3
Hsa-miR-95	21 C-->A	14	19	28	18
Hsa-miR-138	22 C-->A	124	100	159	55
Hsa-miR-339-3P	22 C-->A	11	4	11	6

IsomiRs showing a frequency ≥3 in all samples are listed.

Table S7. Nucleotide substitution variants that overlap with single nucleotide polymorphisms (SNP).

Nucleotide substitution variants coincidental with a SNP in Control FC							
sequence	**freq**	**miRNA**	**5'-trimmed**	**3'-trimmed**	**3'-addition**	**nt-substitution**	**SNP**
UGAUUGUCUAAACGCAAUUAU	12	hsa-miR-219-5p	0	(up)CU	AU	9C-->U	rs2395322
UGAUUGUCUAAACGCAAUUCU	61	hsa-miR-219-5p	0	0	0	9C-->U	rs2395322
UGAUUGUCUAAACGCAAUUCUUG	13	hsa-miR-219-5p	0	(down)UG	0	9C-->U	rs2395322
Nucleotide substitution variants coincidental with a SNP in HD- FC							
sequence	**freq**	**miRNA**	**5'-trimmed**	**3'-trimmed**	**3'-addition**	**nt-substitution**	**SNP**
UGAUUGUCUAAACGCAAUUC	18	hsa-miR-219-5p	0	(up)U	0	9C-->U	rs2395322
UGAUUGUCUAAACGCAAUUAU	18	hsa-miR-219-5p	0	(up)CU	AU	9C-->U	rs2395322
UGAUUGUCUAAACGCAAUUCUU	22	hsa-miR-219-5p	0	(down)U	0	9C-->U	rs2395322
UGAUUGUCUAAACGCAAUUCUUG	35	hsa-miR-219-5p	0	(down)UG	0	9C-->U	rs2395322
UGAUUGUCUAAACGCAAUUCU	119	hsa-miR-219-5p	0	0	0	9C-->U	rs2395322
Nucleotide substitution variants coincidental with a SNP in Control ST							
sequnce	**freq**	**miRNA**	**5'-trimmed**	**3'-trimmed**	**3'-addition**	**nt-substitution**	**SNP**
UGAUUGUCUAAACGCAAUUAU	12	hsa-miR-219-5p	0	(up)CU	AU	9C-->U	rs2395322
UGAUUGUCUAAACGCAAUUCUU	15	hsa-miR-219-5p	0	(down)U	0	9C-->U	rs2395322
UGAUUGUCUAAACGCAAUUCUUG	28	hsa-miR-219-5p	0	(down)UG	0	9C-->U	rs2395322
UGAUUGUCUAAACGCAAUUCU	73	hsa-miR-219-5p	0	0	0	9C-->U	rs2395322
Nucleotide substitution variants coincidental with a SNP in HD-ST							
sequence	**freq**	**miRNA**	**5'-trimmed**	**3'-trimmed**	**3'-addition**	**nt-substitution**	**SNP**
UGAUUGUCUAAACGCAAUUUU	10	hsa-miR-219-5p	0	(up)CU	UU	9C--->U	rs2395322
UGAUUGUCUAAACGCAAUU	12	hsa-miR-219-5p	0	(up)CU	0	9C--->U	rs2395322
UGAUUGUCUAAACGCAAUUGU	15	hsa-miR-219-5p	0	(up)CU	GU	9C--->U	rs2395322
UGAUUGUCUAAACGCAAUUAU	17	hsa-miR-219-5p	0	(up)CU	AU	9C--->U	rs2395322
UGAUUGUCUAAACGCAAUUC	21	hsa-miR-219-5p	0	(up)U	0	9C--->U	rs2395322
UGAUUGUCUAAACGCAAUUCUUG	35	hsa-miR-219-5p	0	(down)UG	0	9C--->U	rs2395322
UGAUUGUCUAAACGCAAUUCUU	38	hsa-miR-219-5p	0	(down)U	0	9C--->U	rs2395322
UGAUUGUCUAAACGCAAUUCU	178	hsa-miR-219-5p	0	0	0	9C--->U	rs2395322

For trimming variants, (up) and (down) followed by the indictaed nucleotides show the short sequence involved in the trimming variants, upstream and downstram of the corresponding reference miRNA 3'-terminus, respectivelly. The number and type of nucleotided added to the 3'-terminus of the mature miRNA are specified in the 3'-addition column. The position and pattern of the nucleotide change with respect to the reference miRNA is shown in the nt-substitution column (Reference nucleotide-->variant nucleotide). IsomiRs showing a frequency ≥10 are listed

Table S8. miRNAs up-regulated in HD-FC and/or HD-ST

miRNAs commonly up-regulated in HD-FC and HD-ST >1,2, p<0,05 (Z-test, Benjamini and Hochberg correction)						
miRNA locus	**norm freq HD-FC**	**norm freq C-FC**	**Fold change HD-FC *vs* C-FC**	**norm freq HD-ST**	**norm freq C-ST**	**Fold change HD-ST *vs* C-ST**
hsa-let-7g	40302	30033	1.34	30910	22933	1.35
hsa-miR-100	1262	716	1.76	1491	953	1.56
hsa-miR-101	8526	4964	1.72	7524	3786	1.99
hsa-miR-106b	181	106	1.71	214	161	1.33
hsa-miR-1250	67	42	1.6	74	41	1.8
hsa-miR-126	258	205	1.26	291	137	2.12
hsa-miR-126*	315	161	1.96	186	89	2.09
hsa-miR-1308	273	83	3.29	404	265	1.52
hsa-miR-143	2800	1215	2.3	1609	918	1.75
hsa-miR-145	562	130	4.32	581	172	3.38
hsa-miR-146a	173	28	6.18	88	45	1.96
hsa-miR-148b	185	115	1.61	191	94	2.03
hsa-miR-151-3p	571	333	1.71	506	422	1.2
hsa-miR-151-5p	811	512	1.58	900	567	1.59
hsa-miR-15a	400	172	2.33	418	317	1.32
hsa-miR-15b	262	73	3.59	269	139	1.94
hsa-miR-16	1665	764	2.18	1659	1186	1.4
hsa-miR-17	174	61	2.85	289	162	1.78
hsa-miR-181a	50540	33516	1.51	90549	50595	1.79
hsa-miR-181a*	435	45	9.67	495	63	7.86
hsa-miR-193b	85	37	2.3	109	63	1.73
hsa-miR-197	135	69	1.96	143	105	1.36
hsa-miR-1974	377	171	2.2	309	223	1.39
hsa-miR-199b-3p	959	308	3.11	292	110	2.65
hsa-miR-19b	72	29	2.48	96	60	1.6
hsa-miR-204	209	90	2.32	350	78	4.49
hsa-miR-20a	128	45	2.84	175	135	1.3
hsa-miR-219-2-3p	6795	2963	2.29	6319	4730	1.34
hsa-miR-219-5p	14821	5355	2.77	19752	8505	2.32
hsa-miR-22	1308	861	1.52	1302	1071	1.22
hsa-miR-23a	825	498	1.66	929	692	1.34
hsa-miR-27b	3294	1992	1.65	3914	2145	1.82
hsa-miR-29c*	104	58	1.79	102	12	8.5
hsa-miR-30a	6463	5397	1.2	7410	5246	1.41
hsa-miR-30b	243	125	1.94	344	231	1.49
hsa-miR-30c	851	545	1.56	1037	807	1.29
hsa-miR-30e	1759	1448	1.21	2007	1557	1.29
hsa-miR-338-3p	2294	991	2.31	2904	1860	1.56
hsa-miR-33a	1472	835	1.76	1260	671	1.88
hsa-miR-33b	314	114	2.75	478	204	2.34
hsa-miR-363	368	278	1.32	424	282	1.5
hsa-miR-374a*	163	94	1.73	121	9	13.44
hsa-miR-451	668	169	3.95	579	132	4.39
hsa-miR-484	69	39	1.77	78	52	1.5
hsa-miR-486-5p	129	38	3.39	109	28	3.89
hsa-miR-574-3p	257	76	3.38	309	221	1.4
hsa-miR-664	78	33	2.36	67	38	1.76
hsa-miR-887	49	26	1.88	66	40	1.65
hsa-miR-92a	4313	2692	1.6	5174	3856	1.34
hsa-miR-93	521	254	2.05	551	408	1.35
hsa-miR-99b	875	611	1.43	1114	575	1.94
miRNAs exclusivelly up-regulated in HD-FC >1,2 p<0,05 (Z-test, Benjamini and Hochberg correction)						
hsa-let-7b*	71	20	3.55			
hsa-miR-129-3p	926	603	1.54			
hsa-miR-130b	57	31	1.84			
hsa-miR-132	262	208	1.26			
hsa-miR-132*	65	1	65			
hsa-miR-136*	69	2	34.5			
hsa-miR-139-5p	1221	834	1.46			
hsa-miR-140-3p	13937	11120	1.25			
hsa-miR-142-5p	84	47	1.79			
hsa-miR-143*	169	102	1.66			
hsa-miR-152	832	687	1.21			
hsa-miR-17*	51	26	1.96			
hsa-miR-181a-2*	54	3	18			

hsa-miR-186	626	510	1.23
hsa-miR-193a-5p	106	68	1.56
hsa-miR-195	217	168	1.29
hsa-miR-21	6698	4116	1.63
hsa-miR-210	134	80	1.68
hsa-miR-212	78	49	1.59
hsa-miR-218	59	34	1.74
hsa-miR-223	174	123	1.41
hsa-miR-23b	1888	847	2.23
hsa-miR-24	4316	3047	1.42
hsa-miR-25	2336	1720	1.36
hsa-miR-26a	21489	11980	1.79
hsa-miR-26b	2752	1470	1.87
hsa-miR-299-5p	43	23	1.87
hsa-miR-29c	3579	2701	1.33
hsa-miR-324-3p	56	30	1.87
hsa-miR-328	139	81	1.72
hsa-miR-331-3p	52	31	1.68
hsa-miR-338-5p	361	179	2.02
hsa-miR-339-5p	43	19	2.26
hsa-miR-340	5356	3663	1.46
hsa-miR-342-3p	850	706	1.2
hsa-miR-345	134	94	1.43
hsa-miR-34c-5p	749	617	1.21
hsa-miR-361-5p	267	159	1.68
hsa-miR-374a	182	96	1.9
hsa-miR-374b	193	119	1.62
hsa-miR-424	36	16	2.25
hsa-miR-425	154	74	2.08
hsa-miR-425*	132	92	1.43
hsa-miR-499-5p	362	215	1.68
hsa-miR-577	97	39	2.49
hsa-miR-584	1433	748	1.92
hsa-miR-625*	58	20	2.9
hsa-miR-652	61	34	1.79
hsa-miR-144*	47	11	4.27
hsa-miR-935	386	315	1.23

miRNAs exclusively up-regulated in HD-ST >1,2 p<0,05 (Z-test, Benjamini and Hochberg correction)			
hsa-let-7i	20400	14112	1.45
hsa-miR-1290	97	25	3.88
hsa-miR-181b	20180	14365	1.4
hsa-miR-181c*	39	24	1.62
hsa-miR-193a-3p	50	25	2
hsa-miR-211	44	11	4
hsa-miR-22*	93	5	18.6
hsa-miR-29b	1575	1156	1.36
hsa-miR-30a*	515	358	1.44
hsa-miR-30e*	532	276	1.93
hsa-miR-34b*	50	5	10
hsa-miR-365	39	19	2.05
hsa-miR-378	649	491	1.32
hsa-miR-7	166	114	1.46
hsa-miR-9	20419	13793	1.48
hsa-miR-9*	813	102	7.97
hsa-miR-98	814	630	1.29
hsa-miR-99a	1590	1122	1.42

The reference miRNA and all the variants mapping on the same miRNA locus are considered in this differenctial expression analysis. Only miRNAs showing a norm freq (norm freq C-FC + norm freq HD-FC) and norm freq (C-ST + freq HD-ST) above 50 are considered.

Table S9. miRNAs down-regulated in HD-FC and/or HD-ST.

miRNAs commonly down-regulated in HD-FC and HD-ST <-1,2; p<0,05 (Z-test, Benjamini and Hochberg correction)						
locus	**norm freq HD-FC**	**norm freq C· FC**	**Fold change HD-FC vs C-FC**	**norm freq HD-ST**	**norm freq C-ST**	**Fold change HD-ST vs C-ST**
hsa-let-7a	111880	158214	0.71	129630	167392	0.77
hsa-let-7c	23569	50689	0.46	27067	43121	0.63
hsa-let-7d	7348	9070	0.81	5824	8042	0.72
hsa-let-7e	5875	13144	0.45	5125	9391	0.55
hsa-miR-103	36651	45207	0.81	21613	30021	0.72
hsa-miR-107	2283	3940	0.58	663	1661	0.4
hsa-miR-1224-5p	73	140	0.52	18	44	0.41
hsa-miR-124	13534	16021	0.84	3430	8883	0.39
hsa-miR-127-3p	2008	2688	0.75	1239	1529	0.81
hsa-miR-128	6241	17293	0.36	2700	4493	0.6
hsa-miR-1301	154	281	0.55	115	230	0.5
hsa-miR-1307	363	477	0.76	300	369	0.81
hsa-miR-139-3p	236	686	0.34	57	375	0.15
hsa-miR-181d	1295	2677	0.48	823	1104	0.75
hsa-miR-193b*	35	60	0.58	36	53	0.68
hsa-miR-199a-3p	102	510	0.2	159	469	0.34
hsa-miR-221	3956	5206	0.76	2191	3095	0.71
hsa-miR-222	1133	1897	0.6	759	1714	0.44
hsa-miR-323-3p	210	458	0.46	151	285	0.53
hsa-miR-330-3p	2352	2807	0.84	1794	2466	0.73
hsa-miR-369-5p	50	88	0.57	29	52	0.56
hsa-miR-382	218	337	0.65	128	190	0.67
hsa-miR-383	250	414	0.6	116	273	0.42
hsa-miR-409-5p	46	102	0.45	22	44	0.5
hsa-miR-423-5p	1321	1926	0.69	1568	2005	0.78
hsa-miR-432	461	977	0.47	323	641	0.5
hsa-miR-433	561	1196	0.47	281	622	0.45
hsa-miR-485-3p	213	484	0.44	124	288	0.43
hsa-miR-485-5p	352	798	0.44	86	284	0.3
hsa-miR-92b*	41	64	0.64	62	107	0.58
hsa-miR-495	401	834	0.48	263	486	0.54
hsa-miR-543	150	468	0.32	95	250	0.38
hsa-miR-598	1969	2575	0.76	1138	2043	0.56
hsa-miR-708	277	369	0.75	231	292	0.79
hsa-miR-760	68	120	0.56	52	119	0.44
hsa-miR-95	307	418	0.73	205	291	0.7
miRNAs exclusivelly down-regulated in HD-FC <-1,2; p<0,05 (Z-test, Benjamini and Hochberg correction)						
hsa-miR-1185	39	76	0.51			
hsa-miR-122	11	45	0.24			
hsa-miR-124*	59	98	0.6			
hsa-miR-129-5p	579	1173	0.49			
hsa-miR-134	208	292	0.71			
hsa-miR-181c*	72	119	0.61			
hsa-miR-221*	110	164	0.67			
hsa-miR-27a	544	687	0.79			
hsa-miR-30a*	613	809	0.76			
hsa-miR-323-5p	19	38	0.5			
hsa-miR-369-3p	63	113	0.56			
hsa-miR-409-3p	32	58	0.55			
hsa-miR-431*	23	49	0.47			
hsa-miR-664*	55	93	0.59			
hsa-miR-7	385	718	0.54			
hsa-miR-744	1194	2255	0.53			
hsa-miR-873	95	155	0.61			
hsa-miR-885-3p	42	78	0.54			
hsa-miR-98	683	875	0.78			
hsa-miR-885-3p	42	78	0.54			

hsa-miR-488*	39	60	0.65			
hsa-miR-154*	29	43	0.67			
miRNAs exclusivelly down-regulated in HD-ST <-1,2; p<0,05 (Z-test, (Z-test, Benjamini and Hochberg correction)						
hsa-miR-1				1040	1403	0.74
hsa-miR-125a-5p				1124	1337	0.84
hsa-miR-1260				44	85	0.52
hsa-miR-1298				17	46	0.37
hsa-miR-132				139	188	0.74
hsa-miR-137				38	82	0.46
hsa-miR-138				756	1755	0.43
hsa-miR-139-5p				701	1021	0.69
hsa-miR-140-3p				9796	11643	0.84
hsa-miR-152				675	825	0.82
hsa-miR-184				23	52	0.44
hsa-miR-185				2565	3478	0.74
hsa-miR-191				1584	2123	0.75
hsa-miR-218				16	37	0.43
hsa-miR-29c				2344	2784	0.84
hsa-miR-320a				3824	5392	0.71
hsa-miR-320b				34	71	0.48
hsa-miR-320d				21	75	0.28
hsa-miR-342-5p				18	41	0.44
hsa-miR-34c-5p				2758	4361	0.63
hsa-miR-374b				174	236	0.74
hsa-miR-375				39	71	0.55
hsa-miR-448				58	215	0.27
hsa-miR-628-3p				38	64	0.59
hsa-miR-628-5p				19	40	0.48
hsa-miR-935				105	183	0.57

The reference miRNA and all the variants mapping on the same miRNA locus are considered in this differenctial expression analysis. Only miRNAs showing a norm freq (norm freq C-FC + norm freq HD-FC) and norm freq (norm freq C-ST + norm freq HD-ST) above 50 are considered.

Table S10. Expression pattern of miRNAs whose expression is altered in HD samples and/or HD mouse models.

miRNA locus	miR/isomiR	Expression pattern HD-FC *vs* C-FC	Expression pattern HD-ST vs C-ST	UP/DOWN IN HD PATIENTS	UP/DOWN IN A HD MOUSE MODEL
hsa-miR-124	Ref	-	DOWN	UNCHANGED (1) DOWN (2)	DOWN (1)
hsa-miR-124	NA	DOWN	DOWN		
hsa-miR-124	s:10A-->C	-	DOWN		
hsa-miR-124	s:19U-->C	DOWN	DOWN		
hsa-miR-124	s:8A-->C	-	DOWN		
hsa-miR-124	tr5:qU	DOWN	DOWN		
hsa-miR-29a	NA	UP	UP	UP (1) UNCHANGED (2)	DOWN (1)
hsa-miR-29a	Ref	UP	DOWN		
hsa-miR-29b	NA	UP	UP	DOWN (2)	UNCHANGED (1)
hsa-miR-29b	Ref	UP	UP		
hsa-miR-9	NA	UP	UP	UP (1) DOWN (2)	UNCHANGED (1)
hsa-miR-9	Ref	DOWN	UP		
hsa-miR-132	NA	UNCHANGED	UNCHANGED	DOWN (1) UP (2)	DOWN (1)
hsa-miR-132	Ref	UP	UNCHANGED		
hsa-miR-212	NA	UP	UNCHANGED	UNCHANGED (2)	-
hsa-miR-139-5p	NA	UP	DOWN	UNCHANGED (2)	-
hsa-miR-139-5p	Ref	UNCHANGED	DOWN		
hsa-miR-139-5p	tr5:(up)GG	UP	UNCHANGED		
hsa-miR-139-3p	tr5:(up)U	DOWN	DOWN		
hsa-miR-330-3p	NA	UNCHANGED	DOWN	UP (1)	ND (1)
hsa-miR-330-3p	Ref	UNCHANGED	DOWN		
hsa-miR-330-3p	tr5:(up)G	DOWN	DOWN		
hsa-miR-330-3p	tr5:(up)GC	DOWN			
hsa-miR-218	NA	UNCHANGED	DOWN	UNCHANGED (2)	-
hsa-miR-17	NA	UP	UP	DOWN (HD1) / UP (HD2) (2)	-
hsa-miR-17	Ref	UP	UNCHANGED		
hsa-miR-22	NA	UP	UNCHANGED	UP (HD1) / DOWN (HD2) (2)	-
hsa-miR-22	Ref	UP	UP		
hsa-miR-222	NA	DOWN	DOWN	DOWN (2)	-
hsa-miR-222	Ref	UNCHANGED	DOWN		
hsa-miR-485-3p	NA	DOWN	DOWN	DOWN (HD2) (2)	-
hsa-miR-485-3p	Ref	DOWN	DOWN		
hsa-miR-485-3p	tr5:(down)A	DOWN	UNCHANGED		
hsa-miR-485-5p	NA	DOWN	DOWN		
hsa-miR-485-5p	Ref	DOWN	DOWN		
hsa-miR-486-5p	NA	UP	UP	UP (2)	-
hsa-miR-486-5p	Ref	UP	UP		

(1), Jhonson et al., (2008) Neurobiol of Disease 29: 438-445; (2), Packer et al., (2008) J. Neuroscience 28 (53): 14341-14346; (-), Non reported in Jhonson et al., (2008); Ref, Reference miRNA; tr5, 5'-trimming variants, (up) and (down) followed by the indictaed nucleotides shows the short sequence involved in the trimming variants, upstream and downstram of the corresponding reference miRNA terminus, respectivelly; s, shows a nucleotide subtitution variant, the position of the nucleotide and the pattern of the modification is shown next; NA, Considers, for differential expression analysis, all the sequences mapping onto a specific miRNA locus different from the reference miRNA or the trimming variants, when these are shown in a separate row

Table-S11. miRNAs deregulated in different neurdegenerative diseases

microRNA	Neurodegenerative disease	Expression pattern	Reference	HD-FC (p<0,05)	HD-ST (p<0,05)
miR-93	AD	DOWN	1	UP	UP
miR-92	AD	UP	5	UP	UP
miR-9	AD	UP;DOWN	4; 1 and 5	-	UP
miR-511	AD	UP	1	-	-
miR-425	AD	DOWN	5	UP	
miR-423	AD	UP	5	DOWN	DOWN
miR-422a	AD	UP	5	-	-
miR-381	AD	UP	5	-	-
miR-370	PrD	UP	8	-	-
miR-363	AD	DOWN	1	UP	UP
miR-34a	AD	UP	5	-	-
miR-342-3p	PrD	UP	8	UP	-
miR-342-3p	PrD	UP	9	UP	-
miR-339-5p	PrD	UP	8	UP	-
miR-338-3p	PrD	DOWN	8	UP	UP
miR-337-3p	PrD	DOWN	8	-	-
miR-328	AD	DOWN	2	UP	-
miR-320	AD	UP	1	-	DOWN
miR-320	PrD	UP	8	-	DOWN
miR-30e-5p	AD	UP	5	UP	UP
miR-30c	AD	UP/DOWN (area dependent)	5	UP	UP
miR-29a/b-1	AD	DOWN	1	-	UP
miR-298	AD	DOWN	1, 2	-	-
miR-27b	AD	UP	5	UP	UP
miR-27a	AD	UP	5	DOWN	-
miR-26b	AD	DOWN	1	UP	-
miR-26a	AD	UP/DOWN (area dependent)	5	UP	-
miR-22	AD	DOWN	1	UP	UP
miR-212	AD	DOWN	5	UP	-
miR-210	AD	DOWN	1	UP	-
miR-203	PrD	UP	8	-	-
miR-200c	AD	UP/DOWN (area dependent)	5	-	-
miR-19b	AD	DOWN	1	UP	UP
miR-197	AD	UP	1	UP	UP
miR-191	PrD	UP		-	UP
miR-19	SA- 1	DOWN	7	UP	UP
miR-181c	AD	DOWN	1	-	-
miR-181a-1*	PrD	UP	8	-	-
miR-15a	AD	DOWN	1	UP	UP
miR-148a	AD	UP	5	-	-
miR-146a	AD	UP	3	UP	UP
miR-146a	PrD	UP	8	UP	UP
miR-145	AD	UP	5	UP	UP
miR-139-5p	PrD	UP	8	UP	DOWN
miR-133b	PD	DOWN	6	-	-
miR-132	AD	DOWN	5	UP	DOWN
miR-130	SA- 2	DOWN	7	UP	
miR-128a	AD	UP	4	DOWN	DOWN
miR-128	PrD	UP	8	DOWN	DOWN
miR-125b	AD	UP	5	-	-
miR-107	AD	DOWN		DOWN	DOWN
miR-106b	AD	DOWN	1	UP	UP
miR-101	AD	DOWN	1	UP	UP
miR-101	SA- 3	DOWN	7	UP	UP
miR-100	AD	UP	5	UP	UP
let-7i	AD	DOWN	1	-	UP
let-7b	PrD	UP	8	-	-

AD; Alzheimer disease; SA-1 spinocerebellar ataxia 1; PD, PArkinson's disease; PrD, Prion disease. (1) Hébert et al., Proc Natl Acad Sci U S A. 2008 Apr 29;105(17):6415-20; (2) Boissonneault et al., J Biol Chem. 2009;284(4):1971-81; (3) Lukiw et al., J Biol Chem. 2008; 283(46):31315-22; (4) Lukiw Neuroreport, 2007 18(3):297-300 (4) Cogswell Journal of Alzheimer's Disease. 2008;14(1):27–41; (6) Kim et al., Science. 2007;317(5842):1220-4; (7) Lee et al., Nature Neuroscience. 2008;11(10):1137–1139; (8) Saba et al. PLoS One. 2008; 3(11):e3652; (9) Montag et al.,Mol Neurodegener. 2009;4:36; (-) Not altered in the HD sequenced samples. miRNas highlighted in orange show a similar expression deregulation pattern in HD and in the indicated neurodegenerative disease

Table S12. Correlation between the expression patterns of the isomiRs and the corresponding reference miRNAs. Only differently expressed sequences reaching statistical significance were considered (p<0.05, Hochberg and Benjamini correction). 3'-IsomiRs include 3'-trimming and 3'-addition variants; 5'-IsomiRs include 5'-trimming variants and nt-substitution variants affecting the seed region. The isomiRs that present a discordant expression pattern with respect to the reference miRNA are highlighted in red.

	Number of isomiRs differently expressed in the frontal cortex			
	UP in HD-FC	DOWN in HD-FC		
5'-IsomiRs	3	2	UP in HD-FC	**Differential expression pattern of the corresponding reference miRNA**
3'-IsomiRs	77	18		
5'-IsomiRs	-	4	DOWN in HD-FC	
3'-IsomiRs	35	163		

	Number of isomiRs differently expressed in the striatum			
	UP in HD-ST	DOWN in HD-ST		
5'-IsomiRs	5	-	UP in HD-ST	**Differential expression pattern of the corresponding reference miRNA**
3'-IsomiRs	127	23		
5'-IsomiRs	-	3	DOWN in HD-ST	
3'-IsomiRs	17	93		

Table S13. Top transcription factors with a possible role in miRNA expression deregulation

Co-regulated miRNAs	Transcription factor (ID)	Transcription factor (Name)	p value Boostrapping (1000 permutations)
HD-FC DOWN	REST	RE1-silencing transcription factor	0.027972028
	TP53	p53 tumor supressor	0.053946054
	NFYA	nuclear transcription factor Y, alpha	0.097902098
	NR1H2-RXRA	nuclear receptor subfamily 1, group	0.128871129
	Foxa2	forkhead box A2	0.134865135
HD-ST DOWN	id1	inhibitor of DNA binding 1	0.066933067
	REST	RE1-silencing transcription factor	0.092907093
	TLX1-NFIC	T-cell leukemia homeobox 1-nuclear	0.106893107
	NFYA	nuclear transcription factor Y, alpha	0.138861139
	Pax6	paired box 6	0.17982018
HD-FC UP	Myf	myogenic factor	0.052947053
	TAL1-TCF3	T-cell acute lymphocytic leukemia 1·	0.088911089
	Cebpa	CCAAT/enhancer binding protein (C/	0.128871129
	SRY	sex determining region Y	0.14985015
	TEAD1	TEA domain family member 1	0.151848152
HD-ST UP	RORA_2	RAR-related orphan receptor A	0.026973027
	Myf	myogenic factor	0.112887113
	PPARG	peroxisome proliferator-activated re	0.127872128
	TAL1-TCF3	T-cell acute lymphocytic leukemia 1·	0.131868132
	SRF	serum response factor (c-fos serum	0.152847153
HD-DOWN (FC and ST)	REST	RE1-silencing transcription factor	0.005994006
	NR1H2-RXRA	nuclear receptor subfamily 1, group	0.093906094
	Pax6	paired box 6	0.138861139
	RELA	transcription factor p65	0.140859141
	NFYA	nuclear transcription factor Y, alpha	0.163836164
HD-UP (FC and ST)	Cebpa	CCAAT/enhancer binding protein (C/	0.072927073
	SRF	serum response factor (c-fos serum	0.097902098
	EMBP1	DNA-binding protein EMBP-1	0.126873127
	Myf	myogenic factor	0.13986014
	FOXI1	forkhead box I1	0.15984016

The identification of the transcription factors with a possible role in HD expression deregulation is performed by comparing and characterizing the promoter regions of miRNAs with similar expression patterns, using the algorithm described by Blanco et al. (Blano et al. (2006) PLoS Computational Biology. 2 (5): e49). The package requires a list with the co-regulated miRNAs and a list with the total number of miRNAs expressed (we consider miRNAs with more than 10 counts per sample). To identify a possible significant enrichment of cis-regulatory elements in co-regulated miRNAs the algorithm assigns a p-value using the permutation-based simulation. Transcription factors highlighted in red have been shown to participate in HD gene expression deregulation.

Table S14. Predicted targets of the HD-downregulated miRs and seed region IsomiRs involved in HD canonical pathway (p = 5,83E-03) in comparison with the HD gene expression deregulation pattern reported by Hodges et al. (2006)

Predicted Targets	Description	Expression pattern in Hodges et al. (2006)
APAF1	apoptotic peptidase activating factor 1	-
ARFIP2	ADP-ribosylation factor interacting protein 2	-
BCL2L1	BCL2-like 1	UP
BDNF	brain-derived neurotrophic factor	-
CAPN6	calpain 6	-
CDK5R1	cyclin-dependent kinase 5, regulatory subunit 1 (p35)	UP
CLTC	clathrin, heavy chain (Hc)	DOWN
CREB1	cAMP responsive element binding protein 1	UP
CREB5	cAMP responsive element binding protein 5	UP
EP300	E1A binding protein p300	UP
GRB2	growth factor receptor-bound protein 2	DOWN
GRM1	glutamate receptor, metabotropic 1	DOWN
HDAC4	histone deacetylase 4	UP
IGF1R	insulin-like growth factor 1 receptor	-
MAP2K4	mitogen-activated protein kinase kinase 4	DOWN
MAPK1	mitogen-activated protein kinase 1	UP
NEUROD1	neurogenic differentiation 1	-
PDPK1	3-phosphoinositide dependent protein kinase-1	DOWN
PIK3R1	phosphoinositide-3-kinase, regulatory subunit 1 (alpha)	UP
PIK3R3	phosphoinositide-3-kinase, regulatory subunit 3 (gamma)	DOWN
PRKCE	protein kinase C, epsilon	-
PRKD1	protein kinase D1	-
RAC1	ras-related C3 botulinum toxin substrate 1 (rho family, small GTP binding protein Rac1)	-
RCOR1	REST corepressor 1	-
SGK1	serum/glucocorticoid regulated kinase 1	-
SOS1	son of sevenless homolog 1 (Drosophila)	UP
SOS2	son of sevenless homolog 2 (Drosophila)	-
SP1	Sp1 transcription factor	UP
TCERG1	transcription elongation regulator 1	-
ZDHHC17	zinc finger, DHHC-type containing 17	-

The expression pattern of the mRNAs reported by Hodges et al. (Hum Mol Gen, 2006 15 (6): 965-77) considered the values of the Table S1 of the article that show a fold expression change of + 1,2 and -1,2 for the HD-upregulated and downregulated genes, respectively (p<0,001).

Table S15A. Number of overlapping genes in HD deregulated transcriptome (Hodges et al., 2006) and the experimentally validated targets for the HD-deregulated miRNAs

	Number of overlapping genes	
	mRNAs UP (1743)	**mRNAs DOWN** (1798)
V-mRNAs for miRNAs UP (150)	28	20
V-mRNAs for miRNAs DOWN (171)	62	7

mRNA UP: signifficantly upregulated mRNAs (Hodges et al.,(2006), Hum Mol Genet. 15 (6): 965-77; Table S1, p<0,001, 1,2 fold, Table S1); **mRNAs DOWN**: signifficantly downregulated mRNAs (Hodges et al., Table S1, P<0.001; -1,2 fold); **V-mRNAs for miRNAs UP**: expermentally validated mRNAs for the HD-upregulated miRNAs (TarBase, http://diana.cslab.ece.ntua.gr/tarbase/); **V-mRNAs for miRNAs DOWN**: expermentally validated mRNAs for the HD-downregulated miRNAs (TarBase). In each dataset, the total number of genes appears in brackets.

Table S15B. Identification of the overlapping genes in HD deregulated transcriptome (Hodges et al., 2006) and the experimentally validated targets for the HD-deregulated miRNAs ID of overlapping genes

mRNAs UP & V-mRNAs for miRNAs UP	mRNAs UP & V-mRNAs for miRNAs DOWN	mRNAs DOWN & V-mRNAs for miRNAs UP	mRNAs DOWN & V-mRNAs for miRNAs DOWN
ZNF294	SMC1L1	UBE2V1	FBXW1B
TLOC1	NFIA	TPM3	TLN1
SQSTM1	ITGB1	TOMM34	F11R
SLC7A11	CD164	STRN	PLDN
PSAT1	ATP6V0E	SLC25A22	RBMS1
PPIF	LASS2	SEC23A	OSBPL8
PGM1	DNAJC1	RTN4	CDC14B
P4HA2	CAV1	RBMS1	
NOTCH2	TJP2	RAB27B	
NOTCH1	AK2	PPP3R1	
MYO10	TEAD1	PPP3CA	
MBNL1	IQGAP1	PISD	
MAT2A	SWAP70	PEX11B	
JUN	TLN1	PANX1	
IGF2R	SYPL	HARS	
IFRD1	PHF19	GPD2	
HSPA1B	PP1201	CA12	
HSPA1A	THG-1	ATP2A2	
FGF2	GNAI3	ARHGDIA	
EGFR	F11R	AP2A1	
DOCK7	FLJ10420		
CYP1B1	HIC		
CXCL12	KIS		
CHD1	SLC16A1		
CGI-38	SP1		
CBFB	CTNND1		
BCL2	PLP2		
	FLJ21924		
	PTBP1		
	SERPINB6		
	HEBP2		
	GSN		
	PAPSS2		
	SSFA2		
	ALDH9A1		
	C9orf88		
	PARG1		
	KIAA1102		
	LOC339924		
	SLC22A5		
	KATNA1		
	RELA		
	MAN2A1		
	CHSY1		
	RYK		
	NFIC		
	SMAD5		
	CYP1B1		
	PGM1		
	PTTG1IP		
	VAMP3		
	SPC18		
	STOM		
	CDCA7		
	C14orf32		
	CPNE3		
	UHRF1		
	G3BP		
	PTPN12		
	ARPC1B		
	CDC14B		
	CDKN1C		

Table S16. Top biological functions and canonical pathways for predicted targets of HD-deregulated seed region-isomiRs. IPA analysis.

Top biological functions for predicted targets of HD-upreguated miRNAs and isomiRs		
Disease and Disorders	**p-value**	**# molecules**
Genetic Disorder	1,63E-03 - 1,53E-02	40
Neurological Disease	1,63E-03 - 4,47E-02	42
Psychological Disorders	1,63E-03 - 1,63E-03	18
Hematological Disease	3,60E-03 - 3,10E-02	3
Developmental Disorder	1,94E-02 - 2,26E-02	5
Molecular and Cellular Functions	**p-value**	**# molecules**
Cellular Compromise	3,99E-04 - 1,94E-02	5
Cellular Assembly and Organization	7,88E-04 - 4,47E-02	7
Cell Morphology	1,20E-03 - 4,47E-02	10
Cell-to-Cell Signalling and Interaction	3,50E-03 - 4,47E-02	4
Cellular Development	5,65E-03 - 3,10E-02	4
Physiological System Development and function	**p-value**	**# molecules**
Nervous System Development and Function	7,88E-04 - 4,47E-02	38
Organ Development	7,34E-03 - 3,10E-02	13
Tissue Morphology	1,38E-02 - 3,10E-02	12
Embryonic Development	2,38E-02 - 3,10E-02	5
Behavior	4,47E-02 - 4,47E-02	2
Top canonical pathways for predicted targets of HD-upreguated miRNAs and isomiRs		
Wnt/β-catenin Signaling	1.64E-05	23/147
ERK/MAPK Signaling	4.38E-04	21/159
HGF Signaling	5.43E-04	14/87
Factors Promoting Cardiogenesis in Vertebrates	8.51E-04	/12/71
Growth Hormone Signaling	1.10E-03	/10/54

Top biological functions for predicted targets of HD-downreguated miRNAs and isomiRs		
Disease and Disorders	**p-value**	**# molecules**
Genetic Disorder	1,61E-02 - 1,61E-02	19
Neurological Disease	1,61E-02 - 2,93E-02	22
Psychological Disorders	1,61E-02 - 1,61E-02	19
Molecular and Cellular Functions	**p-value**	**# molecules**
Cellular Movement	1,47E-04 - 1,80E-02	14
Cell-To-Cell Signaling and Interaction	6,65E-04 - 3,95E-02	9
Cellular Growth and Proliferation	1,56E-03 - 3,30E-02	10
Cellular Assembly and Organization	1,89E-03 - 6,33E-03	7
Cell Morphology	4,44E-03 - 2,93E-02	8
Physiological System Development and function	**p-value**	**# molecules**
Nervous system Development and Function	4,90E-06 - 3,95E-02	58
Organ Development	1,56E-03 - 1,38E-02	8
Tissue Development	2,00E-03 - 2,84E-02	14
Embryonic Development	6,33E-03 - 3,41E-02	4
Tissue Morphology	1,80E-02 - 3,95E-02	5
Top canonical pathways for predicted targets of HD-downregulated miRNAs and isomiRs		
PTEN Signaling	1.11E-06	21/82
PPARa/RXRa Activation	2.17E-05	67/136
Wnt/β-catenin Signaling	3.20E-05	27/147
B Cell Receptor Signaling	6.12E-05	31/120
RAN Signaling	7.18E-05	7/015

2.4. miRNA variants (IsomiRS) are functionally linked to biological processes in distinct species

The advent in sequencing has permitted to elucidate novel miRNAs, increase the knowledge of miRNA biogenesis, and discover putative post-transcriptional editing processes in miRNAs, ignored until now. Mature miRNA biogenesis is subjected to post-transcriptional modifications that are largely uncharacterized, resulting in variants (IsomiRs) of the reference miRNAs annotated in the miRbase. The majority of these variants are modifications of the 3'- and 5'-terminus of the reference miRNA and to a minor extent changes in selective nucleotides along the sequence. [77, 124, 205, 185, 191]. Moreover, it has been proved that isomiRs are naturally occurring, not been caused by RNA degradation during sample preparation for next generation sequencing [160]. In line with this, isomiRs deregulation has been detected in *D. melanogaster* development, elucidating a putative function in important biological processes [85]. To unravel the putative role of isomiRs in evolution we have performed a complete characterization of isomiRs in different species, using publicly available sRNA datasets obtained by high-throughput sequencing strategies. In addition, we analyzed the dynamic changes in the expression of isomiRs in brain samples, at different points during human life using published sRNA sequencing datasets [250].

2.4.1. IsomiRs in evolution

In order to study the significance of isomiRs through evolution, 24 samples of different species have been analyzed using Seqbuster [205] with standard parameters (see methods). The species included in the study were: *C. elegans* development states [24], *D. melanogaster* development states [58], 2 pulldown of AGO in mouse [262], chicken embryonic stage [95], macaque and human [250]. IsomiRs were detected in all species (see 2.5), presenting the majority of miRNAs (80-

90%) presented isomiRs that affected the 3’-terminus were detected for most miRNAs (80%-90%) in each specie, isomiRs of 5’ terminus were less abundant, affecting only a 40-60% of the total miRNA genes. When determining the presence of the annotated miRNA sequences stored in miR- Base database, only 50% of the miRNA genes were represented by those sequences, from now on called ’reference sequence’. An exception is the worm, wherein the 80% of the miRNAs contained the annotated miRNA sequences.

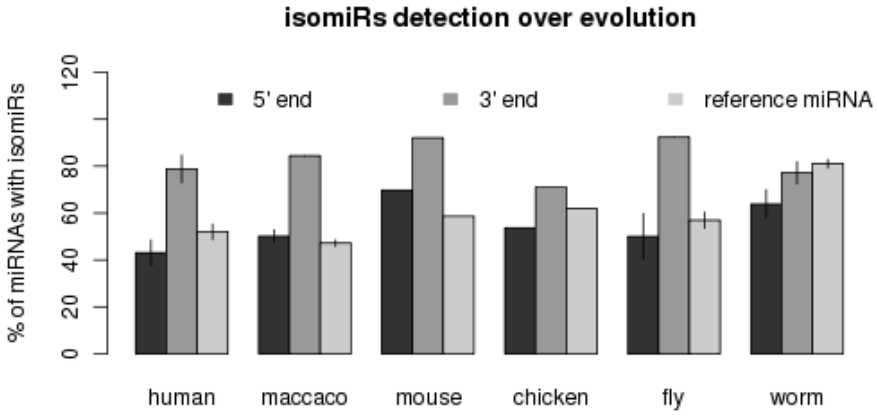


Figure 2.5: Percentage of miRNAs with isomiRs and reference sequence in different species. It is shown isomiRs affecting 5’ and 3 ends separately, dark and medium light grey respectively. The light grey shows the miRNA genes containing the reference sequence. In the case species is contains more than one sample the mean is represented and the standard deviations is shown by dark lines.

2.4.2. IsomiR biogenesis and expression

The formation of 3’- and 5’-trimming isomiRs may be the result of variations in Dicer/Drosha cleavage sites onto the miRNA precursor, during miRNA biogenesis. To test this, we evaluated whether isomiRs were formed in mouse oocytes depleted Dicer and Drosha [237]. Drosha or Dicer knocking-down resulted in a decreased expression of the reference miRNA and the different trimming isomiRs by 80%,

suggesting that these proteins are essential for isomiR biogenesis. Our results have shown that the majority of these 5'- and 3'-trimming variants involve two-three nucleotides upstream or downstream of the mature miRNA (see figure 2.6), being some positions of the precursor preferentially implicated.

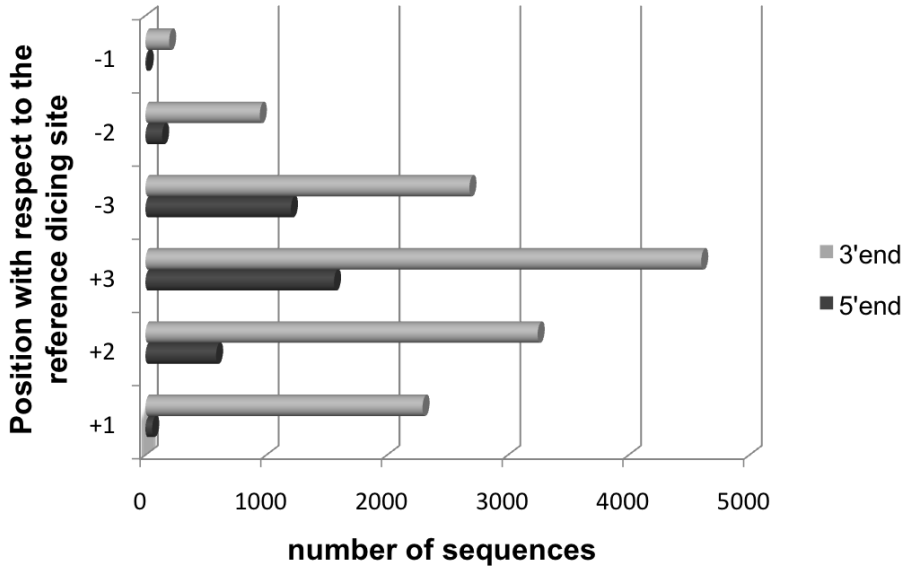


Figure 2.6: Positions affected by isomiRs. Position at the 5'-end (dark grey) and at the 3'-end (light grey) where isomiRs start and end with respect to the reference sequence miRNA annotated in miRBase.

If the abundance of trimming isomiRs is only the direct consequence of Drosha and Dicer activities, highly frequent reference-miRNAs should present a higher proportion of 5'- and 3-trimming isomiRs. However, a poor correlation was found between the expressions (count number) of the reference miRNA and the corresponding trimming isomiRs, especially those affecting the 5'-end (figure 2.7). We observed an heterogeneous landscape were highly expressed reference miRNAs did not present 5-trimming variants and, on the contrary, poorly expressed reference miRNAs showed highly represented 5'-isomiRs. This suggests that other mechanisms, in addition to Drosha/Dicer activities, modulate the extent of 5'-isomiRs expression. It is worth

to mention, that the lack of correlation between the expressions of the reference miRNAs and isomiRs affecting the 3'-ends was lower compared with that of the 5'-trimming variants (q = 0.03 for 5'-isomiRs and q = 0,6 for 3'-isomiRs). These data suggest that, compared with the 5'-trimming variants, a larger proportion of 3'-trimming variants depend on Drosha/Dicer activities. In order to investigate whether specific primary and secondary structures were the reason why some miRNAs present a high proportion of isomiRs, several parameters were analyzed, in the human brain sRNA sequencing dataset. First, the nucleotide population at the 5'- and 3'-ends of the sequences was studied in order to find a possible predominant nucleotide at both sides of the cleavage sites. However, no enrichment was found, either taking into account one or two consecutive terminal nucleotides. We then extended our analysis to motif representations within the mature miRNA and miRNA precursors (data no shown). Nevertheless no differences were found between the population of miRNAs that presented isomiRs (254 in 3'-end, 125 in 5'-end) and that lacking isomiRs (50 in 3'-end,179 in 5'-end). Other structural features were studied that did not present any specific correlation with the group of miRNAs presenting isomiRs or those lacking them, including free energy linked to the secondary structure, secondary structure forms and CG content (see supplementary table S1). Finally, conservation of the miRNA families through evolution was taken into account. Yet the highly conserved miRNA families presented members with a high proportion of isomiRs and others deficient for isomiRs. Together these results suggest that the isomiR expression levels are not related with each miRNA inherent properties, but additional mechanisms, such as cell type or biological status may influence their levels.

2.4.3. Expression pattern of the different sequences in each miRNA gene

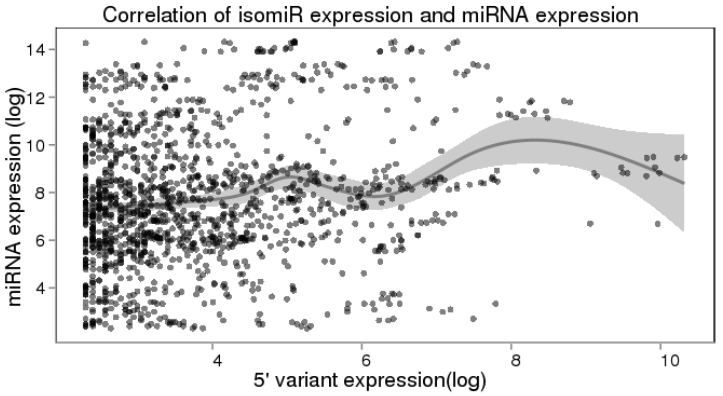
In a deeper study, a regression analysis was performed for miRNA sequencing data of human brains at different ages [250] to deciphering

whether isomiRs have the same expression pattern as that of the corresponding reference sequences.

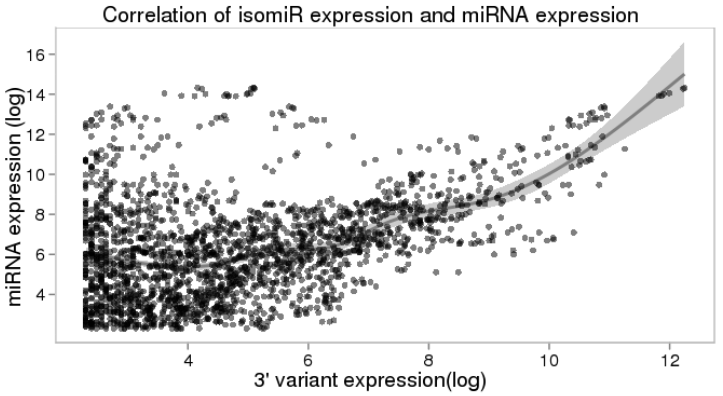
In this analysis we considered the isomiRs contributing in more than a 10% to the total of variants annotating to each specific miRNA locus. In a total of 148 miRNA sequences out of the 463 detected in the brain samples, the expressions of the reference sequence and the corresponding isomiRs along aging, showed a positive correlation (p-value<=0.01, F-test). However, in the remaining miRNAs the aging expression pattern of the isomiRs and the reference miRNA was dissimilar. These results suggest that specific biological states selectively modulate the expression levels of miRNA sequence variants, at least for some miRNAs. (figure 2.8). To investigate whether the biogenesis pathways are the same for isomiRs and reference miRNAs, mouse oocytes with Dicer and Drosha knockdown were analyzed following the same methodology. A background production of miRNAs was found in the cells although Dicer or Drosha were dropped off, in accordance with the original work. The same background of isomiRs were detected, accounting for a reduction of 80% of the reference miRNA and isomiRs when sequestrating Dicer/Drosa. In summary, isomiRs and reference miRNAs share similar complexity in the pathways for their biogenesis.

2.4.4. General characterization

For each miRNA, we studied the contribution of each variant to the total of isomiRs detected. A high percentage of miRNAs presented isomiRs, however the relative abundance of the different variants was not equivalent. In fact, for each miRNA the two more expressed isomiR represented more than 80% of the total isomiRs expression. Therefore, despite of the high diversity of isomiRs, only one or two 5'- or 3'-variants (approximately a 25% of the different isomiRs) were highly expressed when compared to the rest (see figure 2.9). Moreover, considering variation at the 3'-end, the more expressed sequences were represented in a 90% by trimming events, displacing the nucleotide addition variants to those sequences with low expression. Furthermore,



(a) 5 isomiRs



(b) 3 isomiRs

Figure 2.7: Expression correlation between isomiRs and reference miRNAs. Expression is represented by the logarithm transformation of the RPM value assigned to each sequence (isomiR and reference miRNA). The predicted correlation is shown by the grey line and the confidence intervals of the prediction by a shadow grey area.

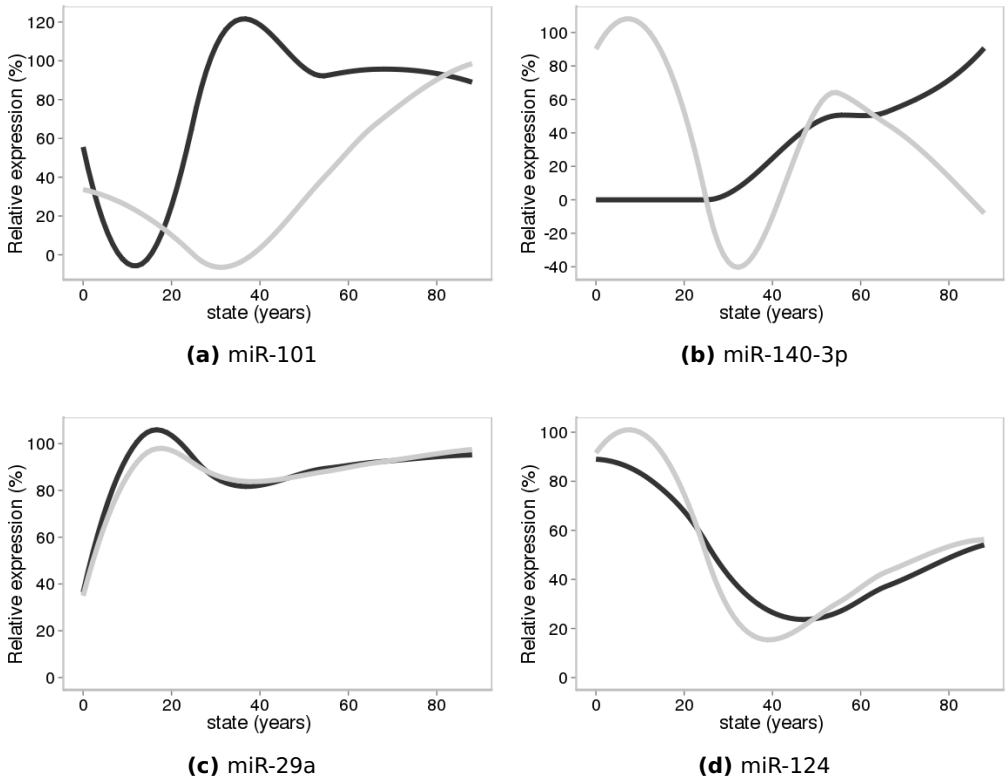
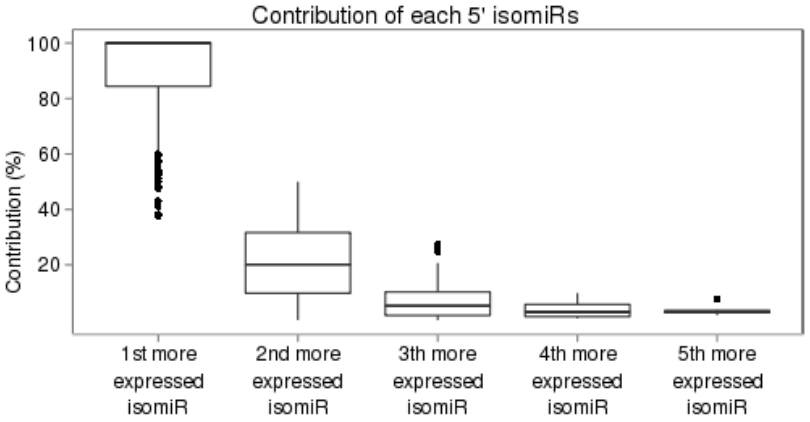
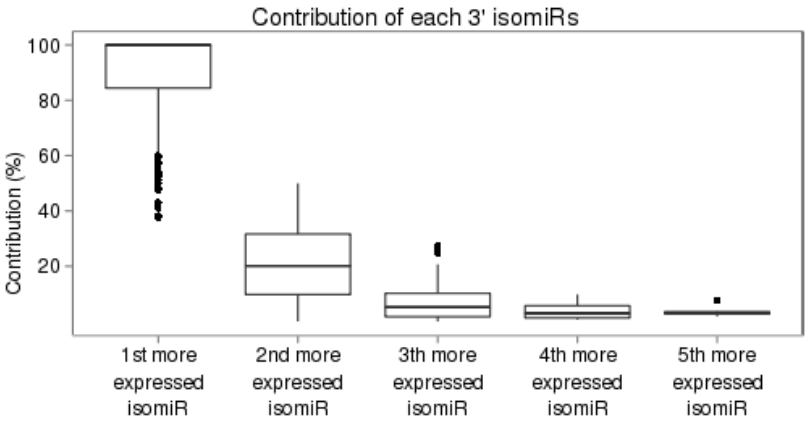


Figure 2.8: Expression pattern of isomiRs/reference miRNAs in human brain aging and development. The x axis represents the different ages and the y axis the expression in logarithm transformation of the RPM value assigned to each sequence (isomiR and reference miRNA). Light grey line represents the reference miRNA sequence and the dark grey line represents the isomiR sequence.



(a) 5 isomiRs



(b) 3 isomiRs

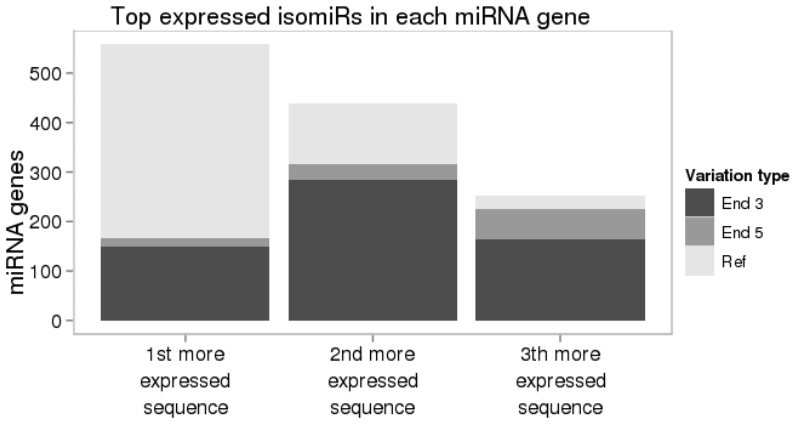
Figure 2.9: Relative expression of each isomiRs in each miRNA gene. IsomiRs contribution to their miRNA locus ranked by their frequencies. The isomiRs of each locus were ranked according to their frequencies where rank 1 is the isomiR more expressed of each miRNA, rank 2 for the second more expressed and so on. (a) The graphic shows the distribution considering 5' isomiRs. (b) Showing contribution of 3' isomiRs.

considering all the sequences mapping onto a specific miRNA locus, the more expressed one was the reference miRNA or a variant in the 3'-end, in a 80% of the cases, (see figure 2.10). From here on, only these more represented sequences were taken into account for subsequent analyses. Taking into account the 3 more expressed sequences in each miRNA, in the 30% of cases, a variation at the 3'-end of the sequence was quantitatively more relevant than the reference sequence. On the contrary, in the majority of cases the abundance of 5'-trimming variants was lower compared with the reference miRNA (see figure 2.10).

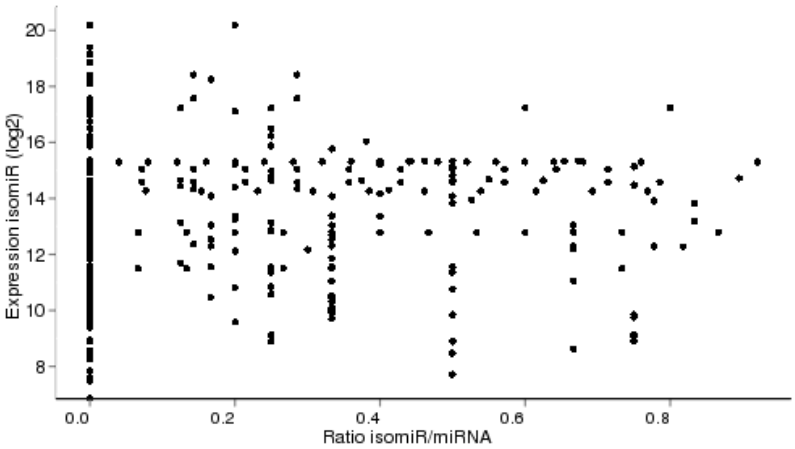
We then investigated whether a relationship exists in the expression extent of the of variants affecting the 5'- and those affecting the 3'-terminus (figure 2.11). Variants showing an upstream trimming at the 5'-end presented in the majority of cases an upstream trimming in 3'-end, or no variation. Similarly, in isomiRs with a downstream trimming at 5'-end, the more represented variation at 3'-end was a downstream trimming or no variation. Therefore, those variations maintaining the mature miRNA size were preferentially expressed.

2.4.5. Conservation across tissues in human

However, 34 miRNA genes had an isomiR as the more expressed sequence. To avoid experimental noise in the comparison, we looked for the relative expression of the sequences in each miRNA gene that showed different sequences as the more expressed form among samples. In the case two samples (A and B) showing different top-expressed sequences for a miRNA gene (sequence SEQ-C and SEQ-D, respectively), we considered a significant difference when the expression of SEQ-C in sample A presented an expression fold change of 1.5 or 0.5 with respect to the expression of the SEQ-D in sample A. After the analysis, 5 miRNAs (showing 100-10,000 counts) presented a differential pattern. Four out of these 5 miRNAs (miR-181a, miR-199a-3p, miR-24, miR-146b-5p and miR-30a-5p) differed in the more expressed sequence at different ages (see supplementary table S2-A), suggesting that selective sequences were preferentially expressed,



(a) Type of isomiR in each position



(b) 5′ isomiR relevance

Figure 2.10: IsomiRs relevance in each miRNA gene. (a) Top 3 sequences more expressed in each miRNA gene. The 3 more expressed sequences of each miRNA genes were classified according to their variation: reference (no variation), variation at 3′-end, and variation at 5′-end. Each column represents the first more expressed sequence, the second more expressed sequence and the third more expressed sequence. (b) The expression of the 5′-isomiRs is represented in the y axis (log2 scale). The ratio between the 5′-isomiRs expression and the total amount of miRNA gene expression is represented in the x axis.

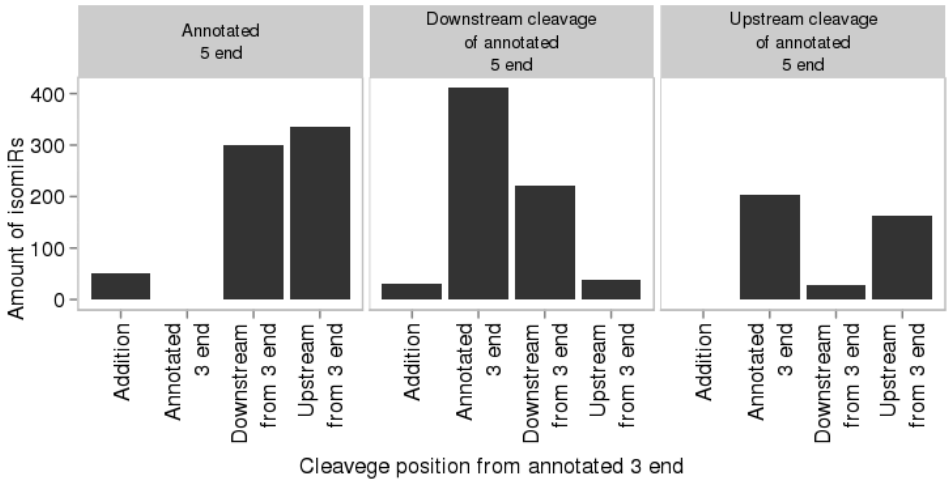


Figure 2.11: Correlation between 5'- and 3'-end variations in isomiRs. The absolute number of sequences that present each type of variation is represented. The 5'-end modifications have been divided in three types: annotated 5'-end, downstream dicing of 5'-end and upstream dicing of the 5'-end. Analogous, the 3'-end have been divided in four different types: addition, annotated 3'-end, downstream dicing of the 3'-end and upstream dicing of the 3'-end.

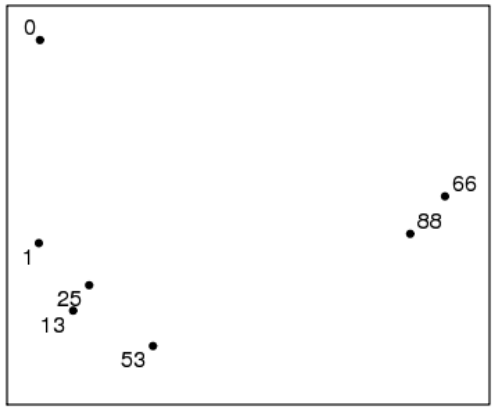
depending on age. In accordance, the top expressed sequence for all miRNAs in brain samples of old individuals from our experiments, was the same as the one identified in the oldest brains (55, 66 and 88 years old) in the Somel study (see supplementary table S2-B). Furthermore, considering the macaque frontal cortex brain samples at different ages, the vast majority of the miRNAs presented the same type of isomiR as the more represented sequence. Three miRNAs (miR-181a, miR-199a-3p and miR-24) presented a similar differential pattern in the type of more abundant isomiR as that observed in the human brain samples (see supplementary table S5). This suggests that sequence plasticity for these miRNAs is important in brain, at different ages. The study of the top- expressed sequence in all miRNA genes was also studied in human undifferentiated stem cells and embryonic bodies [191]. In both samples, for 10 out of 172 miRNAs commonly expressed in both samples, the more expressed sequences differed (see supplementary table S3). We also compared the human brain samples to human stem

cells, resulting in a difference in the diversity pattern of 13 miRNA genes (out of 100 commonly expressed) (see supplementary table S4). Together these results suggest that the relative abundance of the different sequences annotating to selective miRNA genes has a role in specific biological states.

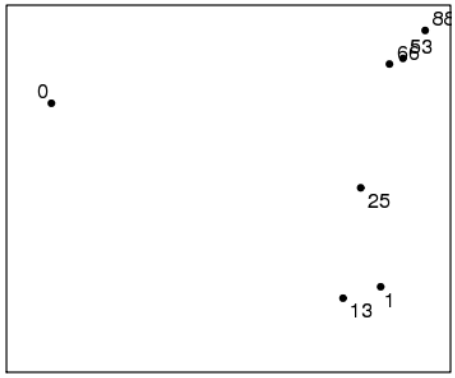
2.4.6. Analyzing the functional role of isomiRs

Our data indicate that isomiRs are not randomly expressed in the different samples, which suggests a functional role for selective miRNA-variants in the human brain, at different ages. To approach this possibility, three different groups of miRNAs sequences were selected in each sample: reference miRNAs, 5'-isomiRs, and 3'-isomiRs. The expression (count number) of these three classes was used as the input data of a principal component analysis, to separate as much as possible the 7 groups corresponding to brain samples at different ages (figure 2.12). The group of sequences corresponding to the reference miRNAs or the 3-isomiRs produced an intuitive aging map, grouping younger and older in two distant clusters. In these PC analyses, the youngest individual presents a distinctive behavior, being specially separated from the rest. The 5'-isomiR map did not show a so clear age-related clustering, except for the two oldest individuals.

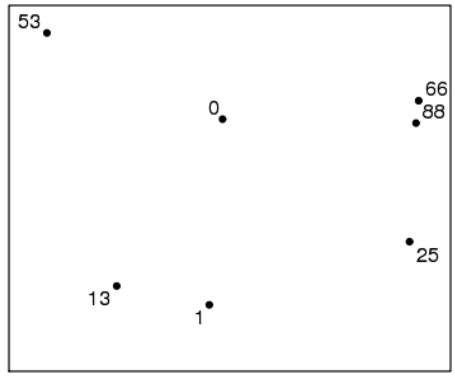
To further explore the possibility of a functional role of the isomiRs in brain aging, we studied the correlation between mRNA and isomiR or reference-miRNAs expressions. To this end we used the publicly available gene expression array data performed on the same sRNA sequenced brain samples [250]. We performed a multiple regression analysis of the gene profile using age as predicted variable, and generated a list of candidate age-related candidate genes (1278 out of 8535 genes expressed). The possibility that these candidate genes are miRNA/isomiR target sequences was then evaluated, using the Targetscan 5.0 prediction algorithm [165]. We obtained 468 (36%) age related genes regulated by miRNA sequences (reference miRNAs and isomiRs), in concordance with the idea that approximately 33% of



(a) reference miRNAs



(b) 3’ isomiRs



(c) 5’ isomiRs

Figure 2.12: MiRNA expression changes during life span. (a-c) The first two principal component analysis of miRNA expression in human brains taking into account reference miRNA, miRNA with 3’ variants, and miRNA with 5’ variants. The analysis was performed by singular value decomposition, using prcmop function in the R stat package, with each sequence scaled to RPM. The number represents each individual’s age in year.

genes are regulated by miRNAs [151, 88]. After that, we analyzed if the expression of the predicted target genes along aging presented a negative correlation with that of the miRNAs or isomiRs. Considering the 468 predicted mRNAs, the expression of a total of 207 showed a negative correlation with that of the reference miRNAs, 391 with 3'-isomiRs and 166 with 5'-isomiRs. Only 66 genes (14%) were common between the three groups, suggesting that miRNA function (indicated by miR-mRNA expression anticorrelation) was specific for some variants. In agreement with this idea if the anti-correlation criteria was ignored, the number of genes commonly targeted by the reference miRNAs, 3'-isomiRs and 5'-isomiRs was significantly increased (30-80% $p < 0{,}01$) (figure 2.13).

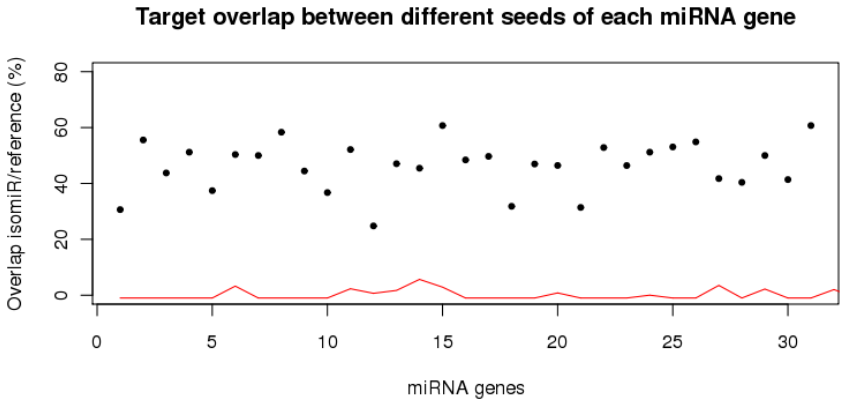


Figure 2.13: Overlapped genes between isomiR and reference miRNA targets. Each point is an 5' isomiR, and the red line indicated the expected overlap assuming a random model (bootstrapping, 1000 permutations).

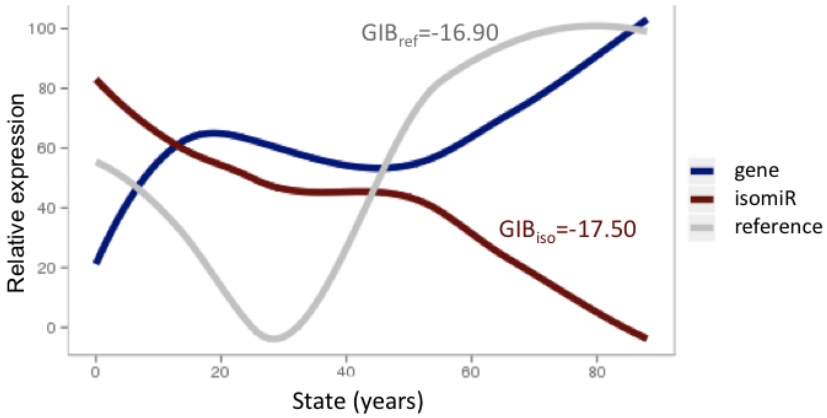
To identify the miRNAs possibly related to age, we considered the proportion of age-related genes and those not related with age that were predicted to be regulated by each miRNA sequence, 468 out of 1278 and (2631 out of 8535, respectively. A total of 87 miRNAs sequences, corresponding to 40 miRNAs genes, were identified as age-related miRNAs (bootstrapping method, $p < 0{,}01$). Interestingly, taking into account all the expressed miRNAs, in the majority of cases

(95%), both the reference miRNA and the corresponding 5'-isomiR were similarly predicted as putative age-related miRNA sequences (see supplementary table S6). In addition, we studied the common miRNAs that were differently expressed between the newborn and oldest age, in human brain and in macaque brain. We found that 29 miRNA genes (57 sequences) showed the same expression pattern, suggesting a putative functional role of isomiRs conserved among species (see supplementary table S7). The three different sets of genes were used as input for the 'DAVID' web-server [118] in order to determine any functional enrichment distinction. Interestingly, the genes targeted by isomiR sequences were statistically enriched in the brain area, suggesting that the regulation of specialized genes were due to isomiRs more than due to the annotated sequences. This agrees with results reported by Somel et al [250], describing a group of genes that were down-regulated at old-ages, and expressed mainly in brain. In other hand, the reference sequences pointed to genes that were enriched in energy metabolism-related pathways, specifically to electron transport chain genes. These genes were also described by Somel et al [250], as one of the main groups that at old-ages their expressions subside. In addition, DAVID web server revealed the transcription factors associated to the 3'-isomiRs and the 5'-isomiRs group of genes were similar. However, the genes targeted by the reference sequences showed to be related to different transcription factors. This agrees with the idea that gene expression modulation along age in these different genes may be related with specific isomiR pos-transcriptional regulation, rather than a transcriptional mediated effect.

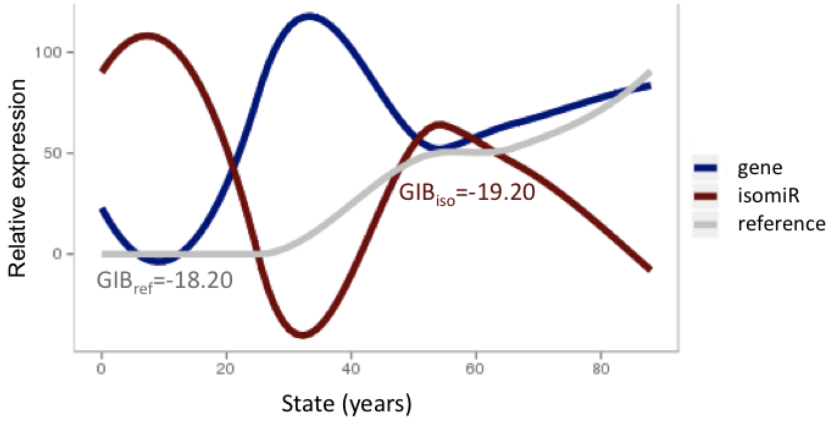
2.4.7. Studying the miRNA-mRNA duplex structure

The previous results suggest that sequences in the same miRNA gene targets a set of different genes. We compared the free energy of the reference miRNA-mRNA duplex and that of the isomiR-mRNA duplex with the same gene. For this analysis we considered isomiRs for which

their expression was not correlated with that of the corresponding reference miRNAs and mRNA targets that were not common for the isomiR and the corresponding reference sequence. For a 92% of the isomiRs (115 out of 125), the duplex energy improved in (at least) one of their target genes that was not predicted for the reference miRNA. If only the most expressed gene (among all targets) was considered for each isomiR, we found that 404 duplex had a better structure stability and 210 duplex had a lower stability. This results were compared to the same study but using the commonly predicted genes between isomiR and reference sequences and isomiRs whose expression correlated to the reference miRNA. In this case, 140 duplex improved the free energy, and 178 decreased their structure stability. Our data showed a stability improvement of the miRNA-mRNA duplex structure (p-value 2x10-6, Fisher test) of those isomiRs that were not correlated to the reference miRNA expression and targeted different genes. For instance, the let-7b gene had a 3' isomiR that differs from the reference miRNA in an extra Uracil at the end of the sequence. DNA Ligase 1 has been predicted for the isomiR form but not for the reference sequence since only the variant is negatively correlated to the gene (see figure 2.14,a). Interestingly, the free energy of the isomiR-mRNA duplex structure was more stable than the reference-mRNA duplex structure (see figure 2.14,a).



(a) hsa-let-7b



(b) hsa-miR-140-3p

Figure 2.14: Gene and miRNA sequences expression profile.(a) Expression profile of the hsa-let-7b family and one of its targets during aging. (b) Expression profile of the hsa-miR-140-3p family and one of its target genes during aging.

2.4.8. Methods

Raw data acquisition: The following samples were downloaded from GEO dataset [78]. Small RNA data were generated with Illumina technology: worm (GSE11738), fly (GSE11624), mouse (GSM41661), macaque (GSE18013) and human (GSE18012). Small RNA from human stem cells and human brain regions were acquired from the ftp server provided by the authors [250, 185]. The human gene expression data came from the same samples than the sRNA data (GSE17757).

MiRNA detection: Last version of SeqBuster was installed to detect miRNAs and their isomiRs for each sample [205] The parameters were set up to detect isomiRs by: a) trimming at the 3' and 5'-end (up to 3 nucleotides shifted from the reference dicing site), b) addition at the 3Õ-end (up to 3 nucleotides added) and c) nucleotide substitution (up to 1 mismatch with respect to the reference miRNA sequence).

Prediction of age-related genes: We applied the same method than the original work used to define age related gene [250]. We applied a lower p-value threshold (0.01 instead of 0.05).

Prediction of age-related miRNA: We considered a gene to be targeted by a miRNA sequences only if: 1) there is a pair complementary between the miRNA seed and the gene according to Targetscan (version 5.1) prediction algorithm [165], and 2) there is a negative expression correlation between the miRNA sequence and the gene. We considered all different seeds of annotated miRNAs and isomiR sequences to run the prediction. Only those genes that have been detected in the gene expression profile of the brain samples, according to the public data from the original work, were used as input for Targetscan algorithm. For each miRNA sequence-gene pair predicted by TargetScan Custom, a linear regression model was performed by the standard R function ('lm'). To consider a gene regulated by a miRNA sequence, the p-value of the model should be less than 0,01 (F-test)

and negative coefficient. The values used for the model were the normalized intensities of microarray data for gene, and logarithms of counts for miRNA sequences. After that, the total number of genes predicted as targets of each miRNA sequence were classified into age-related genes and non age-related genes. A p-value was assigned to each miRNA sequence. The bootstrapping method (1000 permutations) was applied to the data in order to detect a random bias distribution of the age-related genes for each miRNA sequence.

miRNA-gene duplex structure stability: Vienna package were used for the generation of the duplex structure and the free energy calculations [100].

Supplementary material

Table S1: Secondary structure parameters of miRNAs with isomiRs and miRNAs without isomiRs.

	miRNA with isomiRs	**miRNAs without isomiRs**	**p-value (fisher test)**
Hairpin energy structure	-40.27	-40.3	0.9
Nucleotide enrichment: pos -1	A | C | G | U 79|52|18 |108	A | C | G | U 10| 5 | 1| 25	0.2
Nucleotide enrichment: pos 1	A | C | G | U 86| 37| 75| 59	A | C | G | U 16| 5 | 7 |13	0.3
Context information	UTR3 | UTR5 | exon | interg | intron 0 | 1 | 11 | 116 | 130	UTR3 | UTR5 | exon | interg | intron 2 | 0 | 0 | 20 | 19	0.7
CG content	19.7	19.15	0.3

Table S2A: miRNA gene expression in the different ages

Somel et al, 2010

miRNA	Variation (1)	FC2days	FC-1year	FC-13year	FC-25year	FC-55year	FC-66year	FC-88year
hsa-miR-34c-5p	ref	148	235	217	298	2064	477	589
hsa-miR-34c-5p	0 tC0	77	127	215	384	1568	527	571
hsa-miR-34c-5p	0 tGC0	41	90	203	404	1543	595	534
hsa-miR-34c-5p	0 tC qT	26	32	57	96	534	107	123
hsa-miR-181a	ref	3716	3655	1916	1703	2430	1726	1521
hsa-miR-181a	0 qT0	5706	3614	1912	1620	1398	734	492
hsa-miR-181a	0 qTT0	4764	3994	2769	1899	987	314	164
hsa-miR-181a	0 tAGT0	2126	1660	1654	1615	3632	2330	1822
hsa-miR-1974	ref	48	33	650	346	109	81	107
hsa-miR-1974	qG 0	28	18	439	298	98	27	34
hsa-miR-1974	qCG 0	13	6	339	212	56	6	6
hsa-miR-1974	qG tA0	68	15	113	63	49	24	29
hsa-miR-181d	0 qT0	2624	2498	1458	1072	853	484	459
hsa-miR-181d	ref	1480	1653	831	721	530	638	696
hsa-miR-181d	0 tT qGT	392	563	349	122	140	122	141
hsa-miR-181d	0 tT0	439	269	267	205	148	140	133
hsa-miR-146b-5p	0 qGT0	233	1705	2175	2287	870	779	469
hsa-miR-146b-5p	0 qG0	136	982	1137	1107	490	1082	977
hsa-miR-146b-5p	0 qG qG	91	873	1367	842	387	495	327
hsa-miR-146b-5p	ref	64	537	414	496	247	466	562
hsa-miR-99a	ref	4893	3453	2719	3114	2614	3499	3498
hsa-miR-99a	0 tG0	5443	2971	2818	2843	2340	2602	2418
hsa-miR-99a	0 0 qA	3079	825	692	774	672	721	655
hsa-miR-99a	0 0 qT	2240	403	308	306	259	236	246
hsa-miR-99a	qA tG0	333	112	130	112	114	120	93
hsa-miR-143	ref	4461	5665	12391	10415	3698	7718	21706
hsa-miR-143	0 tC0	4324	5040	13131	10166	3402	707	14954
hsa-miR-143	0 0 qT	1330	1183	3206	2801	830	1520	4049
hsa-miR-143	0 tC qA	445	712	1937	1654	568	1215	2704
hsa-miR-378	0 qC qA	1055	2580	1908	1295	1028	1652	1354
hsa-miR-378	0 qC0	935	2741	1864	1107	851	1395	1078
hsa-miR-378	ref	460	2110	1856	1198	890	1086	103
hsa-miR-378	0 qC qAT	375	705	666	418	322	436	343
hsa-miR-378	tA qC qA	76	234	173	121	89	143	107
hsa-miR-378	tA qC0	55	209	128	73	72	108	79
hsa-miR-378	tA qC qT	27	82	62	59	32	49	45
hsa-miR-1185	0 qC0	20	95	89	86	71	86	66
hsa-miR-1185	ref	0	66	88	88	77	90	79
hsa-miR-451	ref	43	778	594	254	357	101	290
hsa-miR-451	0 qT0	75	506	306	264	195	47	169
hsa-miR-451	0 tT0	18	213	248	65	121	24	48
hsa-miR-199a-3p	0 tA0	0	1304	2315	1886	1037	1215	3216
hsa-miR-199a-3p	ref	0	1408	1982	1663	979	1218	3581
hsa-miR-199a-3p	qT tA0	1418	277	447	301	137	229	506
hsa-miR-199a-3p	0 0 qA	0	270	374	276	197	215	470
hsa-miR-199a-3p	0 tA qT	0	194	336	280	160	189	481
hsa-miR-199a-3p	tA 0	379	85	116	110	47	52	142
hsa-miR-199a-3p	qT 0	375	71	115	75	29	51	121
hsa-miR-103	ref	19476	56737	39654	36655	28124	32880	28370
hsa-miR-103	0 tGA0	21680	40869	32565	27184	23938	23053	19040
hsa-miR-103	0 tA0	5265	10725	8559	7760	6242	6633	5777
hsa-miR-103	0 tA qT	3386	4779	3292	3040	2720	2178	2041
hsa-miR-24	0 0 qT	104	1620	2127	1348	1031	1194	1218
hsa-miR-24	ref	504	204	1957	1287	956	1325	1215
hsa-miR-24	0 tG0	165	542	710	520	394	416	334
hsa-miR-24	0 0 qTT	255	275	303	181	125	166	178
hsa-miR-10	ref	2594	783	1052	1129	1832	1324	1761
hsa-miR-10	0 tG0	2039	366	585	619	1323	709	875
hsa-miR-10	0 0 qA	2720	381	50	494	836	606	640
hsa-miR-10	0 0 qT	871	90	132	120	172	101	158
hsa-miR-99b*	0 qT0	97	124	84	81	0	51	0
hsa-miR-99b*	ref	55	108	69	80	0	52	0

hsa-miR-181b	0 qT0	2183	2154	1314	1062	1386	380	308
hsa-miR-181b	0 tGGT0	442	439	904	777	2083	507	406
hsa-miR-181b	0 tT qGT	906	1541	101	392	683	315	213
hsa-miR-181b	ref	390	650	377	325	709	330	333
hsa-miR-181b	qC tGGT0	58	52	59	57	159	44	32
hsa-miR-151-3p	0 qA0	357	331	362	358	872	503	442
hsa-miR-151-3p	0 qA qA	321	294	247	251	711	431	364
hsa-miR-151-3p	ref	293	318	256	366	537	331	315
hsa-miR-151-3p	0 tG0	191	82	123	167	407	138	149
hsa-miR-708	ref	530	723	334	262	251	256	267
hsa-miR-708	0 tG0	576	379	310	293	236	188	184
hsa-miR-708	0 0 qA	80	119	78	59	55	19	17
hsa-miR-374a	ref	139	183	135	133	104	155	150
hsa-miR-374a	0 tG0	190	122	131	106	117	110	106
hsa-miR-30a*	ref	28	176	142	119	103	148	143
hsa-miR-30a*	0 tC qT	36	112	118	85	69	102	89
hsa-miR-30a*	0 tC0	27	79	82	63	55	67	60
hsa-miR-30a	0 qCT0	1170	2695	320	2233	1176	1782	1413
hsa-miR-30a	0 qC0	932	1832	2502	1703	885	1589	1371
hsa-miR-30a	ref	475	1806	2201	1663	90	1647	1571
hsa-miR-30a	0 0 qA	484	1013	1743	1073	489	845	726
hsa-miR-140-3p	tT qA0	6675	6732	10747	0	13960	0	0
hsa-miR-140-3p	tT qAC0	3670	3885	4261	0	3869	0	0
hsa-miR-140-3p	tT qA qA	2695	2940	4187	0	5650	0	0
hsa-miR-140-3p	0 qAC0	0	0	0	0	2836	0	6240
hsa-miR-140-3p	0 qA0	0	0	0	0	2456	0	3786
hsa-miR-140-3p	0 qACA0	0	0	0	0	1673	0	3426
hsa-miR-140-3p	ref	0	0	0	0	653	0	1089
hsa-miR-106b	0 qA0	849	207	199	141	136	162	117
hsa-miR-106b	ref	734	264	234	107	142	145	143
hsa-miR-106b	0 qA qA	241	61	67	53	48	61	44
hsa-miR-15b	ref	127	164	67	45	39	73	48
hsa-miR-15b	0 tCA0	107	85	68	34	26	26	31
hsa-miR-132	ref	30	746	490	226	133	220	136
hsa-miR-132	0 0 qT	34	271	169	124	47	54	51

(1) Label indicating the variation type. The first group refers to the 5' end and the second group refers to the 3' end. 'qTA' indicates that the nucleotides 'TA' are presented on the isomiR, but not on the annotated miRNA in miRBase. 'tTA' indicates that the nucleotides 'TA' are not presented on the isomiR, but they are on the annotated miRNA in miRBase. An addition is indicated on a third group with a 'q'+'nucleotides' contiguously to the 3' end label. '0' means no modifications.

Table S2-B miRNA gene expression in different human brain samples

		Somel et al, 2010							Marti et al,2010	
miRNA	**Variation (1)**	**FC2days**	**FC-1year**	**FC-13year**	**FC-25year**	**FC-55year**	**FC-66year**	**FC-88year**	**FC**	**ST**
hsa-miR-101	0 0 qA	3628	2425	2316	2357	2112	3271	3069	1890	1368
hsa-miR-101	0 0 qT	0	0	0	0	0	0	0	0	0
hsa-miR-101	0 qG0	2065	1243	1304	1323	987	1418	1348	899	776
hsa-miR-101	0 tA0	9306	6241	6867	7470	6154	9969	8703	2970	1745
hsa-miR-101	qG 00	8796	6863	6673	6860	8766	9859	9374	3231	2786
hsa-miR-101	qG 0 qA	4476	3087	2409	2743	2788	3124	2963	1259	1035
hsa-miR-101	qG tA0	14172	10811	10453	13325	14248	14811	14215	4019	2784
hsa-miR-101	ref	6230	6186	5769	5051	5529	9373	8960	3803	2830
hsa-miR-122	0 0 qA	99	120	153	57	37	86	28	0	0
hsa-miR-122	0 qT0	88	74	89	45	20	38	13	0	0
hsa-miR-122	0 tG0	172	169	231	106	79	332	62	68	52
hsa-miR-122	ref	249	313	283	165	104	395	126	74	35
hsa-miR-122	tT 00	0	0	0	0	0	0	0	0	0
hsa-miR-191	0 0 qA	229	354	495	382	286	195	165	441	256
hsa-miR-191	0 qT0	1335	1945	1285	1035	659	428	332	693	482
hsa-miR-191	0 tG0	1580	4078	3453	2856	2556	2742	2752	3759	4283
hsa-miR-191	ref	1949	6405	5137	4338	3924	4910	4548	4205	3559
hsa-miR-191	tC 00	93	102	85	65	68	79	94	104	117
hsa-miR-191	tC qT0	95	63	52	52	30	28	32	0	0
hsa-miR-323-3p	0 tT qA	22	42	62	118	76	81	72	119	56
hsa-miR-323-3p	qG 00	99	268	256	420	153	206	120	722	536
hsa-miR-323-3p	ref	564	3014	1940	2839	1399	1813	1287	989	531
hsa-miR-124	0 qA0	18287	17489	17386	8955	3207	12734	7903	14763	
hsa-miR-124	0 qAA0	15594	9854	10863	6471	2219	8804	5678	23912	
hsa-miR-124	0 qA qT	7583	6115	5332	2904	814	2936	1754	5905	
hsa-miR-124	qT qA0	0	0	0	0	0	0	0	808	
hsa-miR-124	qT qAA0	1056	737	822	500	179	643	399	704	
hsa-miR-124	qT tC0	902	680	766	498	160	514	311	738	
hsa-miR-124	ref	674	873	796	343	134	617	339	1188	
hsa-miR-124	tT qA0	659	573	579	449	340	478	401	0	
hsa-miR-136	0 0 qA	42	255	176	208	168	122	93	381	
hsa-miR-136	ref	49	294	216	224	175	282	244	367	
hsa-miR-221	0 0 qT	0	1149	0	0	0	308	187	0	
hsa-miR-221	0 tC0	0	4346	0	0	0	4117	3252	10169	
hsa-miR-221	0 tC qA	0	0	0	0	0	0	0	742	
hsa-miR-221	0 tC qT	0	812	0	0	0	621	464	0	
hsa-miR-221	0 tTC0	0	0	0	0	0	0	0	816	
hsa-miR-221	ref	0	7935	0	0	0	4765	3805	8685	
hsa-miR-31	0 qG0	0	53	85	53	83	96	93	204	
hsa-miR-31	0 qG qA	0	0	0	0	0	0	0	0	
hsa-miR-31	0 qGT0	0	57	78	52	81	81	58	120	
hsa-miR-31	ref	0	95	167	113	146	159	184	166	
hsa-miR-432	0 tG0	58	362	510	441	184	338	315	600	
hsa-miR-432	0 tGG0	165	1123	1563	1451	644	1020	832	609	
hsa-miR-432	0 tG qT	36	284	228	219	94	118	125	644	
hsa-miR-432	ref	28	281	211	162	94	132	132	689	
hsa-miR-7	0 qT0	2395	29094	37331	23022	6504	5372	2750	1228	
hsa-miR-7	0 tGT qTTT	0	0	0	0	0	0	0	0	
hsa-miR-7	0 tT0	61	949	757	519	186	288	169	98	
hsa-miR-7	0 tTGT0	34	496	535	323	76	180	76	0	
hsa-miR-7	ref	361	15708	11213	6981	2044	3778	2284	1654	
hsa-miR-9*	0 qA0	2390	1606	1133	1200	1036	1076	821	691	
hsa-miR-9*	0 qAA0	528	285	194	191	144	113	71	0	
hsa-miR-9*	0 tT qA	162	128	116	100	102	102	84	0	
hsa-miR-9*	ref	3196	5422	3903	3555	3273	3696	2981	1747	
hsa-miR-9*	tA 00	1104	833	542	444	309	435	295	467	
hsa-miR-9*	tA qA0	2841	2203	1469	1251	948	1276	968	1826	
hsa-miR-9*	tA qAA0	2287	1299	745	691	489	653	504	541	

(1) Label indicating the variation type. The first group refers to the 5' end and the second group refers to the 3' end. 'qTA' indicates that the nucleotides 'TA' are presented on the isomiR, but not on the annotated miRNA in miRBase. 'tTA' indicates that the nucleotides 'TA' are not presented on the isomiR, but they are on the annotated miRNA in miRBase. An addition is indicated on a third group with a 'q'+'nucleotides' contiguously to the 3' end label. '0' means no modifications.

Table S3: miRNA gene expression in human stem cells

miRNA gene	Variation (1)	hESC	EB	miRNA gene	Variation (1)	hESC	EB
hsa-let-7i	0 0 qT	0	542	hsa-miR-30a*	0 tC qT	817	979
hsa-let-7i	0 0 qA	0	222	hsa-miR-30a*	0 tC0	314	246
hsa-let-7i	ref	50	233	hsa-miR-30a*	ref	423	371
hsa-miR-100	ref	81	126	hsa-miR-30a*	tC tC qT	999	647
hsa-miR-100	0 0 qA	116	48	hsa-miR-30a*	tC tC0	187	93
hsa-miR-107	0 tA qT	1713	1212	hsa-miR-30a*	0 tAGC qCGC	97	53
hsa-miR-107	0 tA qC	1089	712	hsa-miR-30a*	tC 00	621	396
hsa-miR-107	0 tA0	7945	7255	hsa-miR-331-3p	0 0 qT	753	869
hsa-miR-107	ref	8887	7149	hsa-miR-331-3p	0 tA0	536	683
hsa-miR-127-3p	0 0 qT	187	0	hsa-miR-331-3p	0 0 qA	417	567
hsa-miR-127-3p	0 0 qA	243	0	hsa-miR-331-3p	ref	592	1639
hsa-miR-127-3p	0 0 qAT	235	82	hsa-miR-372	tA 00	0	4272
hsa-miR-127-3p	ref	486	44	hsa-miR-372	0 tT qA	1433	0
hsa-miR-1287	ref	95	33	hsa-miR-372	ref	34155	2910
hsa-miR-1287	0 0 qA	145	0	hsa-miR-372	0 tT qC	2880	82
hsa-miR-1298	0 qT0	1063	542	hsa-miR-372	0 0 qT	1616	113
hsa-miR-1298	ref	1356	2480	hsa-miR-454	0 qTT0	558	242
hsa-miR-1298	0 tA0	999	970	hsa-miR-454	0 qT0	880	231
hsa-miR-1298	0 0 qA	1970	1281	hsa-miR-454	0 tGGT qTGT	137	0
hsa-miR-135b	ref	153	75	hsa-miR-454	0 0 0	219	48
hsa-miR-135b	0 tGA qTA	195	0	hsa-miR-455-5p	0 qTG0	124	82
hsa-miR-143	0 tC0	1404	242	hsa-miR-455-5p	ref	214	93
hsa-miR-143	0 qA0	153	0	hsa-miR-455-5p	0 qT0	201	122
hsa-miR-143	ref	1367	547	hsa-miR-455-5p	qG tCG0	103	48
hsa-miR-143	0 0 qT	489	206	hsa-miR-484	ref	44	133
hsa-miR-146a	0 qG0	391	133	hsa-miR-484	0 tAT qCT	124	68
hsa-miR-146a	0 0 qA	116	57	hsa-miR-498	ref	111	153
hsa-miR-146a	ref	296	287	hsa-miR-498	0 tC0	137	89
hsa-miR-146a	0 0 qT	124	0	hsa-miR-503	ref	174	157
hsa-miR-191	0 qT0	7506	6043	hsa-miR-503	0 tCAG qA	0	186
hsa-miR-191	0 tG0	3195	3657	hsa-miR-503	0 tCAG qAAA	76	320
hsa-miR-191	tC 0 qA	502	300	hsa-miR-503	0 tCAG0	0	151
hsa-miR-191	tC 00	1097	1003	hsa-miR-548j	ref	140	173
hsa-miR-191	0 0 qA	2425	1900	hsa-miR-548j	qC 00	187	117
hsa-miR-191	tC qT0	648	538	hsa-miR-720	qA 00	0	162
hsa-miR-191	ref	5332	7327	hsa-miR-720	ref	66	129
hsa-miR-199a-3p	tA 00	4433	211	hsa-miR-769-5p	0 tT0	201	82
hsa-miR-199a-3p	0 tA0	31928	1653	hsa-miR-769-5p	0 0 qAA	158	48
hsa-miR-199a-3p	0 tA qT	2272	380	hsa-miR-769-5p	ref	325	198
hsa-miR-199a-3p	ref	23464	1684	hsa-miR-769-5p	0 0 qA	460	142
hsa-miR-199a-3p	0 0 qA	9344	458	hsa-miR-92b	qA 0 qT	166	287
hsa-miR-199a-3p	qT 00	3644	182	hsa-miR-92b	0 tC0	0	13223
hsa-miR-199a-3p	qT tA0	8493	522	hsa-miR-92b	tT 00	439	340
hsa-miR-210	0 tA qG	510	106	hsa-miR-92b	ref	0	20403
hsa-miR-210	0 tA qC	320	48	hsa-miR-92b	0 0 qT	0	5949
hsa-miR-210	0 tA qT	195	53	hsa-miR-92b	0 tCC0	0	8855
hsa-miR-210	ref	621	77	hsa-miR-92b	qA 00	396	429
hsa-miR-221*	ref	618	611	hsa-miR-302b	tT tG qT	0	413
hsa-miR-221*	qA qC0	201	186	hsa-miR-302b	0 tG0	22051	0
hsa-miR-221*	0 qCT0	2383	2647	hsa-miR-302b	0 tAG0	11053	0
hsa-miR-221*	0 qCTG0	137	146	hsa-miR-302b	0 0 qT	9262	0
hsa-miR-221*	qA qCT0	111	113	hsa-miR-302b	ref	16663	0
hsa-miR-221*	0 qC0	2518	1904	hsa-miR-302b	tT 0 qT	0	1706
hsa-miR-221*	qA 00	362	255	hsa-miR-302b	tT 00	0	3924
hsa-miR-28-5p	ref	95	122	hsa-miR-302d	0 tT0	1671	0
hsa-miR-28-5p	0 tG0	137	102	hsa-miR-302d	tT 00	5972	5716
hsa-miR-302a	0 tA qT	2264	0	hsa-miR-302d	ref	12153	0
hsa-miR-302a	0 tGA qTA	2875	0	hsa-miR-302d	tT 0 qT	222	191

hsa-miR-302a	0 tA0	9331	0	hsa-miR-302d	0 tGT qTT	1208	0
hsa-miR-302a	ref	9313	0	hsa-miR-302d	tTA 00	489	458
hsa-miR-302a	tT tGA qTA	158	133	hsa-miR-302d	0 tT qA	1481	0
hsa-miR-302a	tT 00	1306	1875	hsa-miR-18a	0 tAG0	195	113
hsa-miR-302a	qAG tTGA0	256	547	hsa-miR-18a	ref	134	144
				hsa-miR-18a	0 tG0	187	162

(1) Label indicating the variation type. The first group refers to the 5' end and the second group refers to the 3' end. 'qTA' indicates that the nucleotides 'TA' are presented on the isomiR, but not on the annotated miRNA in miRBase. 'tTA' indicates that the nucleotides 'TA' are not presented on the isomiR, but they are on the annotated miRNA in miRBase. An addition is indicated on a third group with a 'q'+'nucleotides' contiguously to the 3' end label. '0' means no modifications.

Table S4: miRNA gene profile in difference cell types and experiments

		Somel et al, 2010							Marti et al,2010		Morin et al, 2008	
miRNA	**Variation(1)**	**FC2days**	**FC-1year**	**FC-13year**	**FC-25year**	**FC-55year**	**FC-66year**	**FC-88year**	**FC**	**ST**	**ESC**	**EB**
hsa-let-7i	0 0 qA	9456	5980	4835	5141	3058	4133	3252	16403	13497	0	222
hsa-let-7i	0 0 qT	31442	15931	9999	9072	6004	6823	5369	10572	7209	0	542
hsa-let-7i	0 tT0	6720	4498	3568	3380	2083	2896	2162	7877	7009	0	0
hsa-let-7i	ref	38532	38319	21111	20289	14976	27328	22516	37025	28078	50	233
hsa-miR-107	0 0 qT	137	138	85	59	28	32	19	0	0	0	0
hsa-miR-107	0 tA0	2612	3673	2586	2522	1667	2246	1721	3413	1333	7945	7255
hsa-miR-107	0 tA qC	0	0	0	0	0	0	0	0	0	1089	712
hsa-miR-107	0 tA qT	1866	1548	995	939	551	786	561	1012	430	1713	1212
hsa-miR-107	ref	3744	9615	6921	6443	4421	7388	5667	11843	4970	8887	7149
hsa-miR-125a-5p	0 tA0	399	305	306	295	126	272	207	1963	1374	1901	829
hsa-miR-125a-5p	0 tGA0	634	501	472	485	258	459	345	2925	2516	1166	315
hsa-miR-125a-5p	0 tGA qA	0	0	0	0	0	0	0	144	173	0	0
hsa-miR-125a-5p	0 tTGA0	0	0	0	0	0	0	0	0	0	116	48
hsa-miR-125a-5p	ref	115	179	117	109	39	82	59	429	297	372	242
hsa-miR-127-3p	0 0 qA	0	0	0	0	0	0	0	1433	809	243	0
hsa-miR-127-3p	0 0 qAT	264	560	1066	706	251	750	457	1603	822	235	82
hsa-miR-127-3p	0 0 qT	287	444	469	408	142	495	370	1156	663	187	0
hsa-miR-127-3p	0 tT qAT	466	802	764	594	206	638	451	0	0	0	0
hsa-miR-127-3p	ref	773	2701	2480	1999	860	3025	2171	6691	3829	486	44
hsa-miR-151-5p	0 qA0	74	64	65	66	83	89	87	481	657	1846	1464
hsa-miR-151-5p	0 qA qA	0	0	0	0	0	0	0	163	241	354	215
hsa-miR-151-5p	0 qAT0	0	0	0	0	0	0	0	0	0	124	97
hsa-miR-151-5p	0 tT0	0	0	0	0	0	0	0	126	140	0	0
hsa-miR-151-5p	qC 00	0	0	0	0	0	0	0	62	40	0	0
hsa-miR-151-5p	ref	113	241	211	184	287	275	257	1304	1205	343	289
hsa-miR-152	0 0 qT	1349	709	1096	501	321	543	654	384	418	7554	1007
hsa-miR-152	0 qG0	772	635	788	441	327	665	594	532	602	7834	1315
hsa-miR-152	0 qG qT	355	215	298	140	86	142	178	277	297	4329	667
hsa-miR-152	ref	3247	3687	4143	2324	1783	3096	3201	1401	1785	3758	683
hsa-miR-192	0 0 qC	0	0	0	0	0	0	0	0	0	774	582
hsa-miR-192	0 qA0	2022	1158	972	758	1209	1181	1145	891	790	6099	4227
hsa-miR-192	0 qA qT	534	232	208	144	224	202	207	185	179	1454	458
hsa-miR-192	0 tC0	444	302	417	283	539	291	282	263	236	0	0
hsa-miR-192	ref	2826	3135	2183	1583	3074	2746	2901	2010	1496	2822	2058
hsa-miR-192	tC 00	168	119	126	65	70	85	75	0	0	0	0
hsa-miR-192	tC qA0	561	360	287	183	202	274	219	187	138	1557	756
hsa-miR-192	tC qAG0	0	0	0	0	0	0	0	119	113	1587	745
hsa-miR-192	tC qAGT0	203	77	74	38	56	51	65	82	96	1140	235
hsa-miR-21	0 qC0	2318	1028	1099	1001	827	1102	1930	4783	5765	130752	178402
hsa-miR-21	0 qC qA	593	344	307	324	194	182	332	642	1050	25619	33560
hsa-miR-21	0 tA0	1021	877	1097	920	1287	1185	2213	3407	5903	0	0
hsa-miR-21	0 tGA qTAC	0	0	0	0	0	0	0	0	0	10305	8410
hsa-miR-21	ref	2542	2388	2567	2336	2603	2750	6558	7384	9686	15375	14560
hsa-miR-221*	0 qC0	858	1397	1329	1315	969	1054	809	0	0	2518	1904

hsa-miR-221*	0 qCT0	742	1413	1285	967	658	459	335	0	0	2383	2647
hsa-miR-221*	0 qCTG0	0	0	0	0	0	0	0	0	0	137	146
hsa-miR-221*	0 tT0	308	502	631	577	430	489	384	0	0	0	0
hsa-miR-221*	qA 00	108	144	115	124	83	105	85	0	0	362	255
hsa-miR-221*	qA qC0	0	0	0	0	0	0	0	0	0	201	186
hsa-miR-221*	qA qCT0	0	0	0	0	0	0	0	0	0	111	113
hsa-miR-221*	ref	3172	7742	7072	6654	5693	6661	5954	0	0	618	611
hsa-miR-222	0 qC0	111	348	398	540	336	323	271	0	0	0	0
hsa-miR-222	0 qCT0	920	3756	3757	4630	3292	3647	3001	3816	3948	8234	14905
hsa-miR-222	0 qCTC0	262	912	729	891	607	732	604	1124	634	8514	23783
hsa-miR-222	0 qCTC qA	0	0	0	0	0	0	0	0	0	1161	2329
hsa-miR-222	0 qCT qT	0	0	0	0	0	0	0	388	273	0	0
hsa-miR-222	ref	200	530	657	928	562	580	477	346	307	248	418
hsa-miR-222	tA qCT0	0	0	0	0	0	0	0	0	0	116	198
hsa-miR-222	tA qCTC0	0	0	0	0	0	0	0	0	0	375	538
hsa-miR-320a	0 0 qT	997	551	423	372	395	291	293	2130	2440	2637	3424
hsa-miR-320a	0 qA0	729	620	526	424	566	432	429	3259	3603	7086	12885
hsa-miR-320a	0 qAA0	0	0	0	0	0	0	0	0	0	2531	3671
hsa-miR-320a	0 tA0	678	679	684	699	836	670	592	0	0	0	0
hsa-miR-320a	0 tA qT	0	0	0	0	0	0	0	1195	1253	0	0
hsa-miR-320a	qG 00	0	0	0	0	0	0	0	228	235	0	0
hsa-miR-320a	ref	3973	4087	2776	2945	3412	4144	4057	14155	14292	5414	9207
hsa-miR-320a	tA 00	256	165	120	106	125	135	99	467	503	481	445
hsa-miR-320a	tA qA0	0	0	0	0	0	0	0	208	237	375	651
hsa-miR-320a	tA qAA0	0	0	0	0	0	0	0	0	0	272	369
hsa-miR-342-3p	0 qC0	474	790	609	448	219	356	265	891	929	2272	1475
hsa-miR-342-3p	0 qCA0	171	254	208	134	55	46	35	191	242	706	458
hsa-miR-342-3p	0 qC qT	182	178	116	75	41	38	18	0	0	690	360
hsa-miR-342-3p	0 tT0	0	0	0	0	0	0	0	149	287	0	0
hsa-miR-342-3p	ref	525	1167	781	508	320	628	501	1387	1949	425	255
hsa-miR-342-3p	tTC qCA0	218	180	132	93	55	101	54	182	232	320	151
hsa-miR-363	0 0 qT	185	152	183	121	94	68	70	73	84	53477	10589
hsa-miR-363	0 qA0	313	404	525	373	284	219	243	349	347	22009	3889
hsa-miR-363	0 tA0	73	90	113	58	68	55	44	168	192	4155	1168
hsa-miR-363	ref	338	539	631	397	376	314	353	444	477	22133	6926
hsa-miR-363	tA 00	0	0	0	0	0	0	0	0	0	3314	938
hsa-miR-363	tA 0 qT	0	0	0	0	0	0	0	0	0	9178	1944
hsa-miR-363	tA 0 qTT	0	0	0	0	0	0	0	0	0	13508	3048
hsa-miR-873	0 qA0	57	163	189	165	62	86	62	234	44	1817	963
hsa-miR-873	0 qA qA	0	0	0	0	0	0	0	0	0	306	113
hsa-miR-873	0 qAT0	0	0	0	0	0	0	0	0	0	830	222
hsa-miR-873	ref	127	789	539	627	215	357	251	356	58	573	280
hsa-miR-92b	0 0 qA	0	0	0	0	0	0	0	1073	2693	0	0
hsa-miR-92b	0 0 qT	71	115	84	49	27	57	52	0	0	0	5949
hsa-miR-92b	0 tC0	181	305	315	185	83	189	158	1155	2016	0	13223
hsa-miR-92b	0 tCC0	101	150	155	111	56	68	66	0	0	0	8855
hsa-miR-92b	0 tC qA	0	0	0	0	0	0	0	610	1450	0	0
hsa-miR-92b	qA 00	0	0	0	0	0	0	0	59	73	396	429

hsa-miR-92b	qA 0 qT	0	0	0	0	0	0	0	0	0	166	287
hsa-miR-92b	ref	383	1399	940	579	302	880	806	7489	12961	0	20403
hsa-miR-92b	tT 00	0	0	0	0	0	0	0	0	0	439	340

(1) Label indicating the variation type. The first group refers to the 5' end and the second group refers to the 3' end. 'qTA' indicates that the nucleotides 'TA' are presented on the isomiR, but not on the annotated miRNA in miRBase. 'tTA' indicates that the nucleotides 'TA' are not presented on the isomiR, but they are on the annotated miRNA in miRBase. An addition is indicated on a third group with a 'q'+'nucleotides' contiguously to the 3' end label. '0' means no modifications.

Table S5: miRNA gene expression in human and macaque brain region

		Human (Somel et al, 2010)							Macaque(Somel et al,2010)				
miRNA gene	**Variation**	**FC2days**	**FC-1year**	**FC-13year**	**FC-25year**	**FC-55year**	**FC-66year**	**FC-88year**	**FC-16days**	**FC-2year**	**FC-5year**	**FC-7year**	**FC-10year**
miR-101	qG tA0	14172	10811	10453	13325	14248	14811	14215	0	0	0	0	0
miR-101	qG tAG0	0	0	0	0	0	0	0	8536	6903	12795	8357	17530
miR-101	ref	6230	6186	5769	5051	5529	9373	8960	1524	1035	1760	0	2500
miR-129-3p	0 qCAT0	0	0	0	0	0	0	0	1519	2059	2530	3386	3827
miR-129-3p	0 qTAT0	26	401	239	179	289	226	162	0	0	0	0	0
miR-129-3p	ref	400	2292	2025	1842	3065	2592	2563	417	535	221	402	227
miR-136	0 0 qA	42	255	176	208	168	122	93	136	184	217	143	248
miR-136	ref ref	49	294	216	224	175	282	244	124	147	114	134	163
miR-29a	0 tA0	444	7285	10487	9628	6730	11695	9540	0	0	0	0	0
miR-29a	ref	3894	55439	69217	77591	55752	101751	89384	522	1044	3673	2022	4243
miR-29a	tC qA0	0	0	0	0	0	0	0	16932	19731	89542	44301	136532

(1) Label indicating the variation type. The first group refers to the 5' end and the second group refers to the 3' end. 'qTA' indicates that the nucleotides 'TA' are presented on the isomiR, but not on the annotated miRNA in miRBase. 'tTA' indicates that the nucleotides 'TA' are not presented on the isomiR, but they are on the annotated miRNA in miRBase. An addition is indicated on a third group with a 'q'+'nucleotides' contiguously to the 3' end label. '0' means no modifications.

Table S6: miRNA sequences predicted as age-related miRNA

miRNA se quence	Age related genes	Non-age related genes	Total	P-value	miRNA se quence	Age related genes	Non-age related genes	Total	P-value
hsa-let-7a,0 0 qA	4	28	32	0.32	hsa-miR-191,0 tG0	0	3	3	1
hsa-let-7a,0 qT0	9	29	38	0.01	hsa-miR-191,ref	0	9	9	1
hsa-let-7a,0 tT0	1	18	19	0.7	hsa-miR-192,0 qA0	0	28	28	1
hsa-let-7a,ref	0	6	6	1	hsa-miR-192,ref	0	2	2	1
hsa-let-7b,0 qT0	7	0	7	0	hsa-miR-192,tC qA0	1	8	9	0.32
hsa-let-7b,ref	0	1	1	1	hsa-miR-1974, qG 00	0	2	2	1
hsa-let-7c,0 0 qA	12	25	37	0	hsa-miR-1974,ref	0	2	2	1
hsa-let-7c,0 qT0	14	17	31	0	hsa-miR-199a-3p,0 tA0	1	76	77	1
hsa-let-7c,0 tT0	7	12	19	0	hsa-miR-199a-3p, qT tA0	2	9	11	0.12
hsa-let-7c, qT 00	14	19	33	0	hsa-miR-199a-3p,ref	0	70	70	1
hsa-let-7c, qT tT0	7	13	20	0	hsa-miR-206,ref	3	74	77	0.98
hsa-let-7c,ref	3	24	27	0.42	hsa-miR-21,0 qC0	0	2	2	1
hsa-let-7d,0 0 qA	1	0	1	0	hsa-miR-21,0 tA0	2	6	8	0.07
hsa-let-7d,0 qT0	5	1	6	0	hsa-miR-21,ref	0	1	1	1
hsa-let-7d,0 tT0	1	42	43	0.98	hsa-miR-219-2-3p,0 tGT0	0	4	4	1
hsa-let-7d,ref	0	78	78	1	hsa-miR-219-2-3p,0 tT0	0	5	5	1
hsa-let-7e,0 0 qA	4	0	4	0	hsa-miR-219-2-3p,ref	0	1	1	1
hsa-let-7e,0 tT0	2	4	6	0.03	hsa-miR-219-5p,0 qT0	4	38	42	0.54
hsa-let-7e,ref	0	6	6	1	hsa-miR-219-5p,0 qTG0	4	39	43	0.58
hsa-let-7f,0 0 qA	5	8	13	0	hsa-miR-219-5p,ref	6	41	47	0.32
hsa-let-7f,0 tT0	0	0	0	1	hsa-miR-22,0 tGT0	0	68	68	1
hsa-let-7f, qA 00	11	28	39	0	hsa-miR-22,ref	0	55	55	1
hsa-let-7f,ref	0	0	0	1	hsa-miR-22*,ref	0	12	12	1
hsa-let-7f,tT 00	0	0	0	1	hsa-miR-221,0 tC0	0	2	2	1
hsa-let-7g,0 0 qA	8	3	11	0	hsa-miR-221,ref	0	2	2	1
hsa-let-7g,0 tT0	1	0	1	0	hsa-miR-221*,0 qC0	0	9	9	1
hsa-let-7g, qC 00	0	0	0	1	hsa-miR-221*,0 qCT0	3	1	4	0
hsa-let-7g,ref	0	2	2	1	hsa-miR-221*,0 tT0	1	20	21	0.7
hsa-let-7i,0 0 qT	22	19	41	0	hsa-miR-221*, qA 00	4	4	8	0
hsa-let-7i,ref	1	10	11	0.4	hsa-miR-221*,ref	0	49	49	1
hsa-miR-1,0 tAT0	0	0	0	1	hsa-miR-222,0 qCT0	0	81	81	1
hsa-miR-1,0 tT0	0	0	0	1	hsa-miR-222,ref	0	67	67	1
hsa-miR-1,0 tT qA	0	0	0	1	hsa-miR-23a,0 qA0	0	8	8	1
hsa-miR-1,ref	0	0	0	1	hsa-miR-23b,0 qA0	1	93	94	1
hsa-miR-100,0 0 qA	0	8	8	1	hsa-miR-23b*,ref	0	33	33	1
hsa-miR-100,0 tG0	0	3	3	1	hsa-miR-24,0 0 qT	0	2	2	1
hsa-miR-100,ref	0	3	3	1	hsa-miR-24,ref	0	19	19	1
hsa-miR-101,0 tA0	0	2	2	1	hsa-miR-25,ref	0	72	72	1
hsa-miR-101, qG 00	1	3	4	0.07	hsa-miR-26a,0 0 qA	0	0	0	1
hsa-miR-101, qG tA0	0	8	8	1	hsa-miR-26a,0 0 qAT	8	1	9	0
hsa-miR-101,ref	0	1	1	1	hsa-miR-26a,0 0 qT	11	8	19	0
hsa-miR-103,0 tGA0	2	3	5	0.01	hsa-miR-26a,0 0 qTT	12	6	18	0
hsa-miR-103,ref	0	13	13	1	hsa-miR-26a,0 tT0	0	0	0	1
hsa-miR-106b,0 qA0	34	146	180	0	hsa-miR-26a,0 tT qA	0	0	0	1
hsa-miR-106b,ref	26	116	142	0	hsa-miR-26a,ref	0	2	2	1
hsa-miR-107,0 tA0	8	7	15	0	hsa-miR-26b,0 qT0	0	0	0	1
hsa-miR-107,ref	0	10	10	1	hsa-miR-26b,ref	4	18	22	0.1
hsa-miR-122,0 tG0	0	0	0	1	hsa-miR-27b,0 qA0	0	107	107	1
hsa-miR-122,ref	0	0	0	1	hsa-miR-27b,0 qA qA	0	136	136	1
hsa-miR-124,0 qA0	5	9	14	0.01	hsa-miR-27b,0 tC0	4	160	164	1
hsa-miR-124,0 qAA0	7	10	17	0	hsa-miR-27b,0 tGC0	8	171	179	1
hsa-miR-124, qT 00	5	3	8	0	hsa-miR-27b,ref	0	88	88	1
hsa-miR-124, qT qA0	4	6	10	0	hsa-miR-28-3p,ref	0	0	0	1
hsa-miR-124, qT qAA0	8	10	18	0	hsa-miR-29a,0 tA0	2	164	166	1
hsa-miR-124, qT qA qT	16	13	29	0	hsa-miR-29a, qC 00	4	82	86	0.98
hsa-miR-124, qT qA qTT	14	9	23	0	hsa-miR-29a, qC tA0	10	86	96	0.58
hsa-miR-124, qT tC0	14	12	26	0	hsa-miR-29a,ref	4	166	170	1
hsa-miR-124, qT tCC0	8	9	17	0	hsa-miR-29b,0 tT0	4	164	168	1
hsa-miR-124,ref	1	3	4	0.05	hsa-miR-29b,ref	4	161	165	1
hsa-miR-124,tT qA0	0	0	0	1	hsa-miR-29c,0 tA0	0	2	2	1
hsa-miR-124,tT qAA0	0	0	0	1	hsa-miR-29c,ref	1	1	2	0.01
hsa-miR-125a-5p,0 tA0	6	6	12	0	hsa-miR-30a,0 0 qA	0	4	4	1
hsa-miR-125a-5p,0 tGA0	8	6	14	0	hsa-miR-30a,0 qC0	0	7	7	1
hsa-miR-125b,0 0 qA	16	47	63	0	hsa-miR-30a,0 qCT0	0	8	8	1
hsa-miR-125b,0 tA0	7	10	17	0	hsa-miR-30a,ref	0	58	58	1
hsa-miR-125b,0 tGA0	12	16	28	0	hsa-miR-30a*,ref	0	117	117	1
hsa-miR-125b,ref	10	12	22	0	hsa-miR-30d,0 qCT0	9	16	25	0
hsa-miR-126*,ref	0	149	149	1	hsa-miR-30d,ref	0	5	5	1
hsa-miR-127-3p,0 0 qA	0	0	0	1	hsa-miR-30e,0 qCT0	4	5	9	0
hsa-miR-127-3p,0 0 qAT	0	0	0	1	hsa-miR-30e*,0 tC qT	1	10	11	0.36
hsa-miR-127-3p,0 0 qT	0	0	0	1	hsa-miR-31,ref	1	96	97	1
hsa-miR-127-3p,0 tT qAT	0	1	1	1	hsa-miR-320a,0 0 qT	45	98	143	0
hsa-miR-127-3p,ref	0	0	0	1	hsa-miR-320a,0 qA0	28	36	64	0
hsa-miR-127-5p,0 qT0	0	22	22	1	hsa-miR-320a,0 tA0	1	2	3	0.04
hsa-miR-127-5p,0 tT0	0	26	26	1	hsa-miR-320a,0 tA qT	11	146	157	0.97
hsa-miR-127-5p,ref	0	47	47	1	hsa-miR-320a,ref	0	2	2	1

hsa-miR-127-5p,tCT qTC0	1	4	5	0.1	hsa-miR-320a,tA 00	10	58	68	0.17
hsa-miR-1271,ref	3	56	59	0.92	hsa-miR-323-3p, qG 00	0	15	15	1
hsa-miR-1277,ref	0	0	0	1	hsa-miR-323-3p,ref	0	53	53	1
hsa-miR-128,0 qT0	14	7	21	0	hsa-miR-330-3p,0 tA0	13	204	217	1
hsa-miR-128,ref	4	12	16	0.02	hsa-miR-330-3p,ref	11	200	211	1
hsa-miR-128,tTC 00	8	4	12	0	hsa-miR-330-3p,tG 00	17	149	166	0.73
hsa-miR-129-3p,0 qTAT0	0	30	30	1	hsa-miR-330-3p,tG qG0	11	172	183	0.99
hsa-miR-129-3p,0 tT0	5	100	105	0.99	hsa-miR-330-3p,tG qGG0	0	48	48	1
hsa-miR-129-3p,ref	2	79	81	1	hsa-miR-330-3p,tG tA0	14	162	176	0.93
hsa-miR-129-5p,0 tC0	0	101	101	1	hsa-miR-330-3p,tG tGA0	12	95	107	0.5
hsa-miR-129-5p,0 tC qT	0	33	33	1	hsa-miR-330-3p,tGC 00	0	0	0	1
hsa-miR-129-5p,ref	0	82	82	1	hsa-miR-330-3p,tGC qG0	0	0	0	1
hsa-miR-130a,0 qT0	30	101	131	0	hsa-miR-330-3p,tGC qGG0	0	0	0	1
hsa-miR-130a,ref	18	74	92	0.01	hsa-miR-335,0 tT0	0	3	3	1
hsa-miR-130b,ref	24	131	155	0.05	hsa-miR-335,ref	0	6	6	1
hsa-miR-132,0 0 qT	2	9	11	0.12	hsa-miR-33a,0 tA0	0	21	21	1
hsa-miR-132,ref	0	16	16	1	hsa-miR-33a,ref	1	0	1	0
hsa-miR-132*,ref	0	5	5	1	hsa-miR-340,0 tT0	2	27	29	0.7
hsa-miR-134,ref	0	4	4	1	hsa-miR-340,ref	8	19	27	0
hsa-miR-136,0 0 qA	0	31	31	1	hsa-miR-342-3p,0 qC0	7	12	19	0
hsa-miR-136,ref	0	55	55	1	hsa-miR-342-3p,ref	2	3	5	0
hsa-miR-136*,ref	0	7	7	1	hsa-miR-342-3p,tTC qCA0	19	11	30	0
hsa-miR-137,0 tG0	13	9	22	0	hsa-miR-34c-5p,0 tC0	35	33	68	0
hsa-miR-137,ref	10	4	14	0	hsa-miR-34c-5p,0 tGC0	38	43	81	0
hsa-miR-138,0 qT0	0	10	10	1	hsa-miR-34c-5p,ref	17	20	37	0
hsa-miR-138,0 tCCG0	0	9	9	1	hsa-miR-361-5p,ref	1	12	13	0.46
hsa-miR-138,0 tCG0	0	25	25	1	hsa-miR-363,0 qA0	0	0	0	1
hsa-miR-138,0 tCG qA	0	19	19	1	hsa-miR-363,ref	0	4	4	1
hsa-miR-138,0 tCG qAT	0	21	21	1	hsa-miR-374a,0 tG0	17	95	112	0.1
hsa-miR-138,0 tG0	0	23	23	1	hsa-miR-374a,ref	0	5	5	1
hsa-miR-138,ref	0	27	27	1	hsa-miR-374a*,0 tT0	1	24	25	0.76
hsa-miR-139-5p,0 qT0	0	18	18	1	hsa-miR-374a*,ref	0	5	5	1
hsa-miR-140-3p,0 qA0	0	0	0	1	hsa-miR-374b,ref	0	2	2	1
hsa-miR-140-3p,0 qAC0	0	0	0	1	hsa-miR-376c,ref	4	12	16	0.03
hsa-miR-140-3p,0 qACA0	0	0	0	1	hsa-miR-378,0 qC0	0	1	1	1
hsa-miR-140-3p,0 qA qA	0	0	0	1	hsa-miR-378,0 qC qA	0	0	0	1
hsa-miR-140-3p,ref	0	0	0	1	hsa-miR-378,ref	0	13	13	1
hsa-miR-140-3p,tT qA0	6	9	15	0	hsa-miR-378,tA qC0	0	3	3	1
hsa-miR-143,0 tC0	0	0	0	1	hsa-miR-378,tA qC qA	0	4	4	1
hsa-miR-143,ref	0	1	1	1	hsa-miR-379,0 0 qA	0	9	9	1
hsa-miR-146b-5p,0 qG0	0	67	67	1	hsa-miR-379,0 0 qAA	0	22	22	1
hsa-miR-146b-5p,0 qG qG	0	27	27	1	hsa-miR-379,0 tG0	0	19	19	1
hsa-miR-146b-5p,0 qGT0	0	15	15	1	hsa-miR-379,ref	0	10	10	1
hsa-miR-146b-5p,ref	0	59	59	1	hsa-miR-382,ref	0	63	63	1
hsa-miR-148a,ref	18	13	31	0	hsa-miR-382,tG 00	1	89	90	1
hsa-miR-148b,ref	0	5	5	1	hsa-miR-383,0 qT0	0	10	10	1
hsa-miR-151-3p,0 qA0	0	0	0	1	hsa-miR-383,ref	0	11	11	1
hsa-miR-151-3p,0 qA qA	0	0	0	1	hsa-miR-409-3p, qC 00	0	46	46	1
hsa-miR-151-3p,ref	0	0	0	1	hsa-miR-409-3p,ref	0	64	64	1
hsa-miR-151-5p,ref	0	14	14	1	hsa-miR-409-5p,ref	0	11	11	1
hsa-miR-152,0 0 qT	2	1	3	0	hsa-miR-410,ref	0	94	94	1
hsa-miR-152,0 qG0	1	1	2	0.01	hsa-miR-411,0 tG0	0	9	9	1
hsa-miR-152,ref	0	0	0	1	hsa-miR-411, qA 00	1	49	50	0.98
hsa-miR-15a,0 tG0	22	34	56	0	hsa-miR-411, qA tG0	0	58	58	1
hsa-miR-15a,ref	11	9	20	0	hsa-miR-411,ref	0	8	8	1
hsa-miR-16,0 qT0	52	95	147	0	hsa-miR-423-3p,ref	0	0	0	1
hsa-miR-16,ref	2	5	7	0.03	hsa-miR-423-5p,0 qT0	3	2	5	0
hsa-miR-17,ref	22	187	209	0.69	hsa-miR-423-5p,ref	0	0	0	1
hsa-miR-17*,ref	10	146	156	0.97	hsa-miR-432,0 tGG0	0	49	49	1
hsa-miR-181a,0 qT0	22	14	36	0	hsa-miR-432,ref	0	28	28	1
hsa-miR-181a,0 qTT0	11	5	16	0	hsa-miR-433,0 tT0	1	134	135	1
hsa-miR-181a,0 tAGT0	0	3	3	1	hsa-miR-433,ref	1	108	109	1
hsa-miR-181a,0 tGT0	0	1	1	1	hsa-miR-451,0 qT0	0	0	0	1
hsa-miR-181a,0 tT0	2	4	6	0.03	hsa-miR-451,ref	0	2	2	1
hsa-miR-181a,ref	18	20	38	0	hsa-miR-485-3p,ref	0	114	114	1
hsa-miR-181b,0 0 qAT	9	2	11	0	hsa-miR-485-5p,0 qG0	0	82	82	1
hsa-miR-181b,0 qT0	12	3	15	0	hsa-miR-485-5p,ref	0	95	95	1
hsa-miR-181b,0 qTT0	7	2	9	0	hsa-miR-487b,ref	0	2	2	1
hsa-miR-181b,0 tGGT0	1	7	8	0.27	hsa-miR-488*,0 tA0	3	6	9	0.02
hsa-miR-181b,0 tGT0	0	4	4	1	hsa-miR-488*,ref	0	3	3	1
hsa-miR-181b,0 tT0	0	2	2	1	hsa-miR-495,0 qT0	0	19	19	1
hsa-miR-181b,0 tT qG	0	1	1	1	hsa-miR-495,ref	1	98	99	1
hsa-miR-181b,0 tT qGT	6	2	8	0	hsa-miR-497,ref	0	10	10	1

hsa-miR-181b,0 tT qGTT	6	2	8	0	hsa-miR-499-5p,0 qA0	1	53	54	0.99
hsa-miR-181b,ref	0	5	5	1	hsa-miR-499-5p,ref	1	30	31	0.91
hsa-miR-181d,0 qT0	21	10	31	0	hsa-miR-532-5p,ref	1	9	10	0.33
hsa-miR-181d,ref	23	22	45	0	hsa-miR-543,0 qT0	0	4	4	1
hsa-miR-185,0 qT0	1	10	11	0.36	hsa-miR-543,ref	0	36	36	1
hsa-miR-185,ref	2	122	124	1	hsa-miR-598,0 qT0	0	0	0	1
hsa-miR-186,0 qT0	18	31	49	0	hsa-miR-598,ref	0	0	0	1
hsa-miR-186,ref	0	2	2	1	hsa-miR-628-5p,ref	0	31	31	1
hsa-miR-92a,ref	5	56	61	0.7	hsa-miR-655,ref	0	33	33	1
hsa-miR-92b,0 tC0	1	7	8	0.23	hsa-miR-7,0 qT0	1	5	6	0.15
hsa-miR-92b,ref	0	2	2	1	hsa-miR-7,ref	0	24	24	1
hsa-miR-93,0 0 qA	35	38	73	0	hsa-miR-708,0 tG0	36	13	49	0
hsa-miR-93,ref	36	182	218	0.02	hsa-miR-708,ref	15	10	25	0
hsa-miR-95,ref	0	0	0	1	hsa-miR-769-5p,ref	0	8	8	1
hsa-miR-98,0 tT0	0	15	15	1	hsa-miR-873,0 qA0	1	12	13	0.46
hsa-miR-98,ref	0	9	9	1	hsa-miR-873,ref	0	18	18	1
hsa-miR-99a,0 tG0	0	7	7	1	hsa-miR-9,0 tA0	22	27	49	0
hsa-miR-99a, qA tG0	1	13	14	0.47	hsa-miR-9,0 tGA0	22	14	36	0
hsa-miR-99a,ref	0	2	2	1	hsa-miR-9,ref	19	21	40	0
hsa-miR-99b,0 tG0	0	1	1	1	hsa-miR-9,tT 00	25	15	40	0
hsa-miR-99b,ref	0	1	1	1	hsa-miR-9*,0 qA0	31	47	78	0
hsa-miR-744,0 0 qA	0	1	1	1	hsa-miR-9*,ref	0	6	6	1
hsa-miR-744,0 0 qT	4	1	5	0	hsa-miR-9*,tA 00	22	23	45	0
hsa-miR-744,0 tA0	0	1	1	1	hsa-miR-9*,tA qA0	23	34	57	0
hsa-miR-744,ref	0	1	1	1	hsa-miR-9*,tA qAA0	23	59	82	0
hsa-miR-885-3p,0 tA0	0	6	6	1	hsa-miR-9*,tA qA qT	23	65	88	0
hsa-miR-889,ref	0	13	13	1					

(1) Label indicating the variation type. The first group refers to the 5' end and the second group refers to the 3' end. 'qTA' indicates that the nucleotides 'TA' are presented on the isomiR, but not on the annotated miRNA in miRBase. 'tTA' indicates that the nucleotides 'TA' are not presented on the isomiR, but they are on the annotated miRNA in miRBase. An addition is indicated on a third group with a 'q'+'nucleotides' contiguously to the 3' end label. '0' means no modifications.

3 Discussion

SeqBuster and SeqCluster bioinformatics tools to characterize the small RNA transcriptome

miRNA analysis: SeqBuster

We have developed SeqBuster, a bioinformatics tool for the analysis of deep sequencing data and, in particular for the analysis of miRNA variants, or isomiRs. Other existing bioinformatics tools address specific questions such as differential expression or miRNA prediction [31, 82, 103, 271]; SeqBuster performs these as well as other types of analyses in a single user-friendly platform. The main analyses that SeqBuster provides are: a) adapter recognition and removal, b) miRNA annotation and c) simultaneous miRNA analysis of multiple samples simultaneously. Furthermore, SeqBuster is the first tool that provides automated pre-analysis for sequence annotation. Several features highlight SeqBuster as an exclusive tool for the characterization of sRNA data generated by large-scale sequencing technologies. First, SeqBuster, in addition to the web-server, includes a stand-alone version that permits the annotation against any custom database installed in the local machine independently of the web server. This means that the analysis is not restricted to the databases stored in the web server, which is a limitation of other web-based bioinformatics tools such as [103, 292, 82]. Second, the R environment, in which the different analysis packages have been developed, permits the incorporation and/or modification of different types of analysis. Such R package additions or modifications could be geared towards the analysis of different types of RNA data generated by large-scale sequencing. This provides a flexible platform that can be easily modified by adding new analysis packages. This is a great bonus, since the field of high-throughput sequencing is changing very fast. Third, SeqBuster is highly versatile offering a wide range of options from raw data processing and normalization to annotation and visualization, therefore offering complete control of the analysis process. Finally, SeqBuester presents a complete picture of the results of mapping a dataset of sRNAs in

the output, including the variation of the different sequences with respect to the annotated miRNAs, as highlighted in recent studies [77, 191, 155, 152], and differential expression data.

The flexibility of SeqBuster is illustrated by the pre-analysis approach (applied to the hESC and EB raw data, see chapter 2.1). In addition to many known miRNAs we discovered 109 novel miRNAs. Of these, 15% were already known at the time of the original analysis [191] and were detected in our study as a consequence of the algorithm used for the adapter recognition/removal and the alignment parameters (see methods in chapter 2.1). For the recognition of adapter sequences, we included a modified version of the Needleman-Wunsch algorithm that allows the detection of more adapters when comparing with algorithms that detect exact matches of the adapter sequence. In the latest version of SeqBuster, to avoid depending on external mapping tools, we have included a custom algorithm written in Java programming language that generates and uses an index of the read sequences. The module maps sequences of 16 to 36 nt long, allowing up to 1 mismatch, up to 3 nt trimmed at the 3'-end matching the precursor sequence, and up to 3 nt at the end of the sequence considered as nucleotide addition. Moreover, a parse filter integrated in the aligner module shows only the best hits (best alignment score) for each sequence. This allows skipping any posterior parsing step for the removal of secondary alignments with worse alignment scores. The parsing step is common when using BLAST or BLAT tools, that were chosen in the first version of SeqBuster due to the higher specificity for the detection of putative nucleotide additions in the alignment than other tools. However, the new algorithm proved to be as specific as BLAST in the detection of miRNAs. The high efficiency of this algorithm is demonstrated by the reduced time required to match 200,000 sequences against the miRBase database (1 min, compared with the 65 min required when using BLAST), achieving 99% of concordance with BLAST. In contrast to other tools, this module does not require the indexing of the database and it can be used directly with a FASTA input file containing the precursor sequences.

The application of SeqBuster tool to several publicly available sRNA datasets and to our own sequenced samples has proven its efficacy in the detection and deep characterization of miRNA variability. In addition, SeqBuster forms part of the tools offered by the European consortium SIROCCO (Silencing RNAs: organisers and coordinators of complexity in eukaryotic organisms) that aims a multidisciplinary approach to study the biological relevance of small silencing RNAs. Since its publication, several groups have used SeqBuster [43, 96, 185, 77, 292] indicating that this tool is useful for the scientific community.

Non-miRNAs sRNAs analysis: SeqCluster

The identification and functional characterization of new sRNAs species is one of the hottest topics in the field of the epigenetic modulation of gene expression. While a number of pipelines to analyze sRNA high-throughput sequencing data are specifically focused in several aspects of miRNA characterization, the unbiased exploration of the whole sRNA transcriptome remains a challenge. We have developed SeqCluster, an extension of the miRNA analysis tool SeqBuster, offering a highly flexible custom analysis of deep sequencing data with emphasis on the non-biased characterization of non-miRNA sRNAs. There are several pipelines for dealing with different aspects of the annotation and classification of sRNAs including DeepBase [284], MiRanalyzer [103], DSAP [82] and mirTools [292]. DeepBase is a database for the storage of the information (sRNA location in the genome) generated by their custom pipeline not integrated for external users. DSAP, miRanalyzer and MiRTools are automated multiple-task web servers for the analysis of deep-sequencing sRNA datasets. However, in these pipelines, the annotation of the non-miRNA sRNAs is limited, based only in the mapping against Rfam and/or RepBase databases with the aiming to remove putative RNA degradation products. A recent tool developed by Buermans *et al* [43] matches sRNAs to genes in the ENSEMBL database [87]. However, none of the current available

sRNA pipelines includes a method to deal with cross-mapping events, i.e. sRNAs that map to multiple locations, therefore ignoring up to 30% of the total mapped sRNAs. Differing from the current sRNA analysis tools, the main advantages of the SeqCluster framework are: 1) the classification and annotation of the data is not restricted to specific databases and can be customized by the user; 2) there is the possibility to inspect sequences highly expressed sequences that have not been successfully classified; 3) the distinction of potential RNA degradation products from putative functional sRNAs, through the visualization of the expression levels and their localization in the genome; and 4) the solving of the problem of ambiguous sRNAs mapping. Furthermore, SeqCluster permits differential expression analysis between two samples or two groups of samples in different biological contexts to highlight functionally important and relevant sRNAs. Time series datasets can also be analyzed using specific statistical methods (linear/multiple regression) to determine which sRNAs correlate with a certain biological conditions. We have shown that SeqCluster is able to detect and classify all types of sRNAs known so far, in different species, including recently discovered classes sRNAs of still unknown function (see below).

Characterization of miRNA variability

The first cloning- and sequencing-based high-throughput experiments identified the majority of the miRNA reference sequences [155]. More than 40 samples were used, and the majority were cell lines from different diseases and seven human tissues. Although several sRNAs were found to annotate in each miRNA gene, the authors of the study defined the more expressed sequence as the reference miRNA, responsible for repression of the target genes. The other highly expressed sequences, not defined as the reference miRNA have largely been ignored until recently. The advent of the new sequencing technologies contributed to the detection of a huge amount of sequences showing slight differences with respect to the reference miRNA. In 2008, the miRNA profiling using

the second generation sequencing technologies, led to the definition of these variants with the term 'IsomiRs.' This new scenario prompted specific research in miRNA/isomiR profiling in different biological states and the possible functional role of these variants in gene expression regulation.

Our studies have been focused on the isomiR characterization and profiling in different paradigms involving human samples and different distantly species related to unravel possible functional signatures associated to miRNAs variants.

To evaluate the relevance of all isomiRs mapping onto a miRNA gene we first studied their relative abundance in several tissues and species. For each miRNA, the average number of variants found was 10; but, only two or three sequences were expressed at considerable levels, representing 80% of the total isomiR expression. Therefore, the majority of variants have negligible expression. Furthermore, the weigh of variations at the 3'-end was generally higher compared with that of 5'-end variants. In fact, we detected low sequence heterogeneity at the 5'-end of the miRNAs, with the abundance of most of these variants been negligible. This suggests that the 5'-terminus of the miRNAs is protected from variations. This observation agrees with the known crucial role of Watson/Crick base pairing of the 5'-seed region of the miRNA with the 3'-UTR of the mRNA, for gene targeting. In line with this, scarce miRNAs presented nucleotide substitutions at positions 1-11 of the miRNA, containing the 5'-seed (nt 2-8) and the cleavage (nt 10-12) sites [76] that are typically base paired in the miR:mRNA duplex.

Interestingly, the more expressed variants have a tendency to conserve the 22 nt miRNA length, meaning that a 3'-trimming event is strongly correlated with a similar 5'-trimming event. This is in agreement with the sRNA loading mechanism that requires specific sequence length to be functional [65].

The deep analysis of sRNA datasets from human stem cells [191], and human brains [250] revealed numerous variation, suggesting that the miRNA transcriptome is more complex than previously thought. Our analysis indicates that miRNA genes express sequences with three different types of modifications: different starting nucleotide (trimming at 5'-end), different end nucleotides (variation at 3'-end, trimming or addition), and different middle nucleotides (nucleotide substitution). Most miRNAs displayed 3'-trimming and 3'-addition events in agreement with previous reports [155, 179, 12, 152]. In particular, the most common single nucleotides added to the 3'-end where A or U. These modifications could influence mRNA expression regulation, since 3'-end pairing has been suggested to contribute to target recognition, particularly when sites have weaker miRNA seed matches [41, 42]. In plants, it has been shown that addition of A residues to the 3'-end plays a negative role in miRNA degradation [178] . It has been also proposed that the combined effects of 5'-deletions and 3'-U extensions in Oryza sativa and Arabidopsis thaliana can alter the specificity by which miRNAs associate with different Argonaute proteins [77]. In addition, recent data show selective stability for different miRNAs in human cells that depends on regions of the 3'-terminus [20]. Whether short nucleotide extensions in animal miRNAs influence their stability or the silencing mechanism needs to be shown in functional studies.

Another type of variability highlighted by SeqBuster is related with significant nucleotide modifications along the mature miRNA. Several lines of evidence argue against RT-PCR and sequencing errors as the main sources of sequence discrepancies with respect to the reference miRNA. First, Lee *et al*, rejected that experimental steps in the common workflow of the next generation sequencing produce artefact sequences [160]. Second, the frequencies of nucleotide modifications were remarkably higher compared to the estimates of Illumina sequencing errors [72]. Third, the positional non-randomness of nucleotide changes along the length of the miRNA was observed

in all the libraries analyzed, being some of them common among the different samples. Finally, nucleotide changes, insertions and deletions have also been reported in other studies using different sequencing strategies. Furthermore, a recent work published in the 'Science' journal on (June 2011), has reported a widespread RNA-DNA differences in the human transcriptome discovering more than 10,000 exonic sites where the RNA sequences do not match that of the DNA, which suggests the existence of a new yet-unexplored layer in genome variation [167]. The nucleotide substitutions have been described in few miRNAs as pri-miR precursor editing changes from A to G attributed in part to A to I deaminations, which lead to a repression in the maturation of the miRNA [35, 155, 77, 285, 133, 132]. Edited adenosines have been also described in mature miRNAs [35, 133, 132]. These variants were confirmed in the human brain libraries; indeed, we found that the A to G change is one of the most common types of nucleotide substitution events in the seed region. However, a considerable proportion of brain miRNAs presented nucleotide changes that were distinct from the classical A to I editing. Common nucleotide substitutions in brain miRNAs included U to G, C to A, G to A and G to U changes. These types of nucleotide substitutions have been also reported in a meta-analysis of sRNA datasets in plants [77, 124] and in the let family of miRNAs in different mouse cells lines [219]. These modifications may result in alternative base pairing between the variant and the target mRNA, possibly affecting the efficiency of gene regulation, as previously described [67, 41] . This has been shown in the human miR-376a that presents an A to G editing event in the seed region that affects gene targeting [133]. Another example has been recently reported in the mouse let-7 family of miRNAs, as several types of nucleotide modifications at the ninth position of mmu-let7a result in an increase of the stability of the predicted miRNA:mRNA duplexes [219].

The nucleotide substitution events are not randomly distributed along the sequence. The less variable positions present key roles in gene target recognition and silencing. For instance, scarce miRNAs presented

nucleotide substitutions at positions 2-4 of the miRNA within the 5'-seed region (nt 2-8). A contiguous and perfect base pairing of the miRNA nucleotides 2-8, representing the seed region, is the most stringent requirement for efficient target recognition [86, 41]. In addition, low variability was detected at positions 14-16, contained in an anchor site (nt 13-16) that has been proposed to hold important determinants for miRNA:mRNA association. However, positions 9-12 of the miRNA were amongst the more variable, which could contribute to the imperfect mRNA:miRNA pairing detected in the central region of the miRNA. For gene expression repression, bugles or mismatches must be present in the central region of the miRNA:mRNA duplex (nt 9-12 of the miRNA), precluding the Argonaute-mediated endonucleolytic cleavage of mRNA [86].

Overall, the present analysis strongly suggests a biological function for the sequence plasticity in miRNAs, which may have broad implications in mRNA targeting, stability and/or gene expression regulation. The exhaustive description of the different types of miRNA variability will be extremely useful to uncover tissue-specific isomiR distributions relevant in development, physiology and disease conditions.

IsomiR expression in different species

Different publications report miRNA profiling in several species using next generation sequencing, including *C. elegans* [24], *D. melanogaster* [58], mouse [262], chicken [95], macaque and human [250, 191]; however these studies were focused on the reference miRNA, therefore ignoring the isomiR landscape. To gain insight into the possible role of isomiRs in evolution, we fully characterized miRNA variants in different species. We found isomiRs in all the species we studied, from worm to human. In all samples, every annotated miRNA gene presented different types of isomiRs, suggesting the existence of conserved mechanisms responsible for isomiR generation and/or maintenance. However, for a given conserved miRNA, the more abundant type of

isomiRs differed among species. For instance, the miR-124 family, where the reference sequence is the more expressed in human, but not in worm. This result suggests that for a given miRNA, specific isomiRs contribute to selective species dependent processes.

To evaluate the isomiR expression pattern conservation across closely related species, we analyzed data from frontal cortex brain samples of macaques [250]. Only 3 miRNA genes out of 100 that presented the same isomiR relative abundance pattern in the human frontal cortex samples differed between these two species. These data suggest that form the same brain area variability of the expression pattern for many miRNA genes is largely conserved between human and macaque, considering the same brain area. If we only consider the human data, in 4 out of 126 miRNA genes expressed in all samples, the most expressed representative showed differences between individuals. In agreement with this concordance between close species, the isomiR expression pattern of the oldest individuals in the Somel study and that of two additional brain samples corresponding to individuals of comparable ages [185], showed that for 100% of the miRNAs, the most abundant variant was the same. In addition, considering human brain samples and stem cells, 13 miRNAs differed in the type of isomiR that was more represented. For instance, the most expressed sequence for the miR-103 gene was different in brain samples and in stem cells. While in brain, the reference miR-103 and the isomiR-103 were equally expressed, in stem cells, the reference miR-103 sequence was expressed 9 fold with respect to. Curiously, this miRNA was among the top ten most expressed miRNAs in stem cells. Similarly, the ten most expressed miRNAs in the brain samples, including the let-7 family, showed that the most expressed sequence was the reference miRNA with a at least 7 times higher expression than the other variants. These data indicate that for a given condition, the most expressed miRNAs are mainly represented by a unique sequence that may dominate gene expression regulation. However, for other miRNAs the expression of the isomiR and the reference miRNA is comparable and therefore, several sequences may participate in gene expression regulation underlying

specific processes.

In summary, these data suggest that isomiRs are not simply a secondary product of miRNA expression, and that specific biological conditions may be associated with the preferential expression of a particular isomiR.

IsomiRs profiling in physiological and pathological processes

We have demonstrated that isomiRs are common in different eukaryotic species, from worn to human, across different states and with similar expression than the reference miRNA sequences. However, the functional role of isomiRs is still unknown. During the course of this thesis, a study was published, correlating the expression of some isomiRs with fly development. This constitutes the first evidence for a possible functional role of miRNA variants in important biological processes [85]. The authors demonstrated that: (1) non-template nucleotide additions of adenosines to miRNA 3'-ends are highly abundant in early development; (2) a subset of miRNAs with nontemplate 3'-U are expressed in adult tissues; and (3) the ammount of at least eight reference-miRNAs varies across tissues and during fly development.

To gain insights into the possible role of isomiRs in different biological processes we studied their profile in different physiological and pathological paradigms, in human samples. These included postnatal brain development and aging, Huntington's disease and stem cell differentiation.

IsomiRs in human brain postnatal development and ageing

To study the possible relevance of isomiRs in brain development and aging we studied sequence diversity for each miRNA using publicly available sRNA datasets of seven samples corresponding to human brains of different ages [250]. Unlike in Drosophila development, [85],

we found that non-template nucleotide additions were not highly represented in the isomiR fraction of the human brain samples. Therefore this type of variants may have a more relevant function in Drosophila, as compared with the human brain.

We studied the differential expression in the newborn and adult brain, and found that 29 miRNAs presented the same type of isomiRs with differential relative abundance at two ages, both in the macaque and in the human brains. These data strongly suggest a relevant role for the differential expression of these isomiRs differential expression in primate brain development.

Somel *et al* demonstrated that miRNA expression changes along life span [250]. They showed a clear aging map when the expression of miRNA sequences was used. This map was replicated in our analysis when reference sequences were used. However, the ismoiRs expression did not show an intuitive map, although the ages still remained well separated. Isomirs with variation at the 3'-end resulted in a map with two groups separated by a huge distance (different expression pattern): birth time and the rest of individuals. A possibility exists that this kind of variation is more relevant during the developmental processes that take place during infancy (normally the first 2 years after birth) than in the process of aging. In the case of 5' isomiRs, the aging map showed that only the oldest individuals were clustered and in general, every sample had a unique expression pattern, suggesting that these sequences may not underlie gene regulation linked to the development or aging of the human brain.

To study a direct influence of isomiRs in gene expression regulation, we first predicted the isomiRs targets using the Targetscan algorithm. Subsequently we considered the anti-correlation between isomiRs and predicted mRNA targets expressions as a filter to identify putative targets. In agreement with previous reports [151, 88], we found that up to 36% of human genes are likely regulated by all miRNAs-

sequences. Interestingly, if we only consider the reference miRNAs, then the percentage of predicted target genes is 15% than it would be lower than expected.

Our analysis showed that the predicted targets, deduced by seed complementary for the different seeds in the same miRNA gene overlapped significantly (30%-80% of overlap). This suggests that there is intrinsic plasticity in the different miRNA-variants for common target recognition, therefore ensuring the regulation of the same set of target genes. Taking into consideration the anti-correlation between miRNA-variants and mRNA expressions, only 66 mRNAs were commonly targeted by the reference miRNAs, 3'-isomiRs and 5'-ismiRs. This suggests that the different types of variants have an intrinsic potential to target a big set of genes. Therefore, other factors may narrow down the number of targets, including, the stability of the miR:mRNA duplex and/or the availability of a specific mRNA in a given condition.

Similarly to the study by Somel *et al.*, [250], functional enrichment analysis of the reference miRNAs targets that are involved in brain development and ageing pointed to an enrichment in energy metabolism-related pathways, specifically to electron transport chain genes. However, we showed that isomiRs targets involved in these processes were more enriched in brain specific genes, suggesting that variants may be more related to brain specific pathways. This agrees with the concept that miRNA sequence plasticity may contribute to different aspects underlying brain development and aging.

Considering the anti-correlation between the expressions of miRNA/isomiR predicted targets and that of mRNAs to define miRNA/isomiR targets, we found a number of possible targets for 3'-isomiRs that were not shared by the corresponding reference miRNA. Since the seed region in 3'-isomiRs and the corresponding reference miRNAs is shared, other explanations may account for this discrepancy. Our results show that considering the most abundant genes whose expression

anti-correlated with that of the isomiR (and not with that of the reference miRNA), in a significant number of cases ($p<0.01$) the stability of the duplex isomiR/mRNA was increased, compared with that of the reference-miRNA/mRNA. This suggests that modifications of the duplex free energy by certain variations in the 3'-end of the miRNA are important in determining the type of miRNA sequence that dominates as a regulator of a specific target. A similar result was found when analyzing mRNAs commonly targeted by 5'-isomiRs and the corresponding reference miRNA. Therefore, stability of the miRNA/mRNA or isomiR/mRNA may be a general important factor to identify the type of sequence participating in gene expression regulation.

Overall, these results support the hypothesis that miRNAs may act as a buffer in gene regulation through a set of sequences for each miRNA gene. The complexity in miRNA diversity make easier the gene regulation network since the way to control gene expression may not be dependent on a unique sequence, but on a set of sequences that participate in gene regulation linked to a specific biological condition. The mechanism involved in this phenotype-related sequence selection will need further studies. We can conclude that isomiR expression is becoming an important aspect of the miRNA biology showing differential expression between different stages in a development process or aging. However many questions about the real relevance and mechanism are still opened, and they will warrant specific functional studies.

miRNAs and IsomiRs profiling in Huntington Disease

A number of recent evidences have shown the importance of miRNAs in the physiology of the central nervous system, including developmental processes, neuron differentiation and specification, cell maintenance and metabolism [56, 26, 208]. Our own analysis strongly suggests that isomiRs may also underlie important aspects of the brain development and aging. Besides, an increasing number of studies

have highlighted the involvement of miRNAs in mechanisms that lead to neurodegenerative processes [112, 158]. However, no specific studies have addressed the possible relevance of miRNA variants in the neuropathology associated to neurodegenerative diseases.

Huntington's disease (HD) is an autosomal dominant neurodegenerative disease caused by a CAG-repeat expansion in the HTT (huntingtin) gene that encodes HTT. The disorder predominantly affects neurons of the forebrain (frontal cortex -FC- and striatum -ST-). Gene expression deregulation has largely been reported in HD. Such deregulation is in part due to the aberrant nuclear localization of the transcriptional repressor RE1-Silencing Transcription Factor (REST) [295, 294]. In addition to protein-coding genes, recent reports have shown that miRNA transcriptome is perturbed in HD [129, 203] affecting miR9/9*, miR-124a, miR-132, miR-29b, miR-29a, miR-330, miR-17, miR-196a, miR-222, miR-485 and miR-486. To deeply characterize the miRNA and isomiR transcriptome linked to HD neurodegeneration, we sequenced and analyzed the sRNA expression pattern in the FC and the ST of patients with HD and control individuals without histopathological lesions. The differential expression analysis of the sequenced samples confirmed the expression pattern of 9 out of 11 miRNAs, whose expression was altered in the FC of HD patients in two previous studies [129, 203] The majority (55-80%) of the HD deregulated miRNAs were common between the FC and the ST samples.

Interestingly, we noted that in some cases, miRNA members of the same family exhibited similar expression trends, suggesting that they are co-regulated in HD. For instance, let-7a, let-7c, let-7d and let-7e were downregulated in HD-FC and HD-ST. In another example, miR-30a, miR-30b, miR30c and miR-30e were all upregulated in HD-FC and HD-ST. These results are in agreement with the idea of a relevant role of certain miRNA families in HD pathology. An interesting opened question is the understanding of the mechanisms accounting for the co-regulation of non-clustered miRNAs of the same family in pathological

processes.

In addition, the majority of the reported miRNAs deregulated in other neurodegenerative disorders (Alzheimer's disease, Parkinson's disease, Prion disease and spinocerebellar ataxia type 1) were altered in the HD sequenced samples showing a similar deregulation pattern to that reported in the other neurodegenerative diseases. This indicates that they may participate in gene expression deregulation underlying common processes in neurodegeneration.

There is strong evidence that the abnormal function of the transcriptional repressor REST plays a major role in gene expression deregulation in HD [129, 203]. Some neuronal miRNAs modulated by REST (miR-124, miR-29a, miR-29b, miR9/9* miR-132 and miR-330-3p) are abnormally expressed in the FC of HD patients [128, 203, 129]. Taking into account all brain-expressed miRNAs showing experimental or predicted evidence for REST modulation (REST-miRNAs) our analysis identified a number of deregulated REST-miRNAs. Furthermore, REST-miRNAs were found to be enriched among those downregulated in HD brain regions. Overall, these results agree with the concept that non-physiological REST activity in HD may affect the expression of protein-coding genes by repressing normally brain-expressed miRNAs.

In addition, we identified P53 tumour suppressor as a transcription factor with a putative role in HD-miRNA deregulation. P53 is found in polyglutamine aggregates both in vivo and in vitro, suggesting a hypofunction of P53 in HD [254]. Therefore, P53 may contribute to miRNA expression down-regulation in HD.

In studying sequence variability in control and HD-brain samples, we observed that the proportion of the different types of variants was similar in all brain samples suggesting that the molecular mechanisms involved in the generation of these variants are not altered in HD. However, a few sequences mapping on the same miRNA were discordant, presenting

the reference miRNA and the isomiR an opposite expression pattern in HD. The majority of these discordant IsomiRs were 3'-variants (3'-trimming and 3'-addition variants). Whether selective forms of a miRNA are relevant in HD and how these variants are selectively deregulated in HD deserve further research.

To gain insights into the biological pathways possibly affected by reference miRNAs and isomiRs expression deregulation, we run functional enrichment analyses using the ingenuity pathway analysis (IPA) tool, for the predicted targets of the reference miRNA HD-deregulated miRNAs and the seed-region-HD-deregulated IsomiRs (5'-trimming variants and nucleotide substitution variants affecting the seed region). The most enriched biological and disease-related processes and canonical pathways amongst the HD-enriched or HD-diminished miRNA/IsomiR targets were analyzed in our study. Significantly enriched functional annotations include Ôgenetic disorder', 'neurological disease', 'developmental disorders', 'cell-to-cell signaling and interaction', 'cell growth and proliferation', 'nervous system development and function' and organ development'. Some of these terms were common between the two groups (miRNAs and IsomiRs), in agreement with the significant overlap of the predicted mRNA targets. HD was identified as one of the significant canonical pathways (P = 6.6E-03) for the predicted HD-downregulated miRNAs/IsomiRs. Interestingly, 16 out of the 30 predicted targets involved HD-canonical pathway are significantly deregulated in an exhaustive study of mRNA profiling in [114].

In addition, the analysis revealed that axonal guidance signaling and actin and cytoskeleton signaling were among the highly significantly enriched pathways for the predicted targets of the upregulated miRNAs, being these pathways also identified in a functional enrichment analysis in a previous study of HD deregulated genes [114].

The results revealed a significant enrichment in genes involved in neurological diseases for both HD-enriched and HD-diminished seed-

region isomiR targets. Furthermore, the HD canonical pathway was significantly enriched for the HD-downregulated seed region isomiR targets. These results underscore the possible contribution of the IsomiRs in the modulation of HD related pathways.

IsomiRs in human stem cell differentiation

The variability in the expression pattern of isomiRs was evaluated using human public data of hESC and EB [191]. Comparing these two conditions we retrieved 32 miRNAs showing a dissimilar pattern of isomiRs relative abundance. This suggests that different isomiRs may participate in selective aspects of stem cell differentiation. As a proof of concept, we evaluated the predicted targets of the reference miRNAs and the 5'-trimming variants that were differentially expressed in hESC and EB, since 5'-trimming phenomena result in changes in the seed region that is essential for target recognition and silencing[165].

Profiling non-miRNA sRNAs from worm to human

The best-known class of small non-coding RNAs is the miRNAs. However, in the last years the high-throughput sequencing strategies have revealed a complex non-coding sRNA transcriptome, with many novel sequences of unknown function. We used SeqCluster to classify the different types of non-miRNA sRNAs sequences in different species, as a primary approach to identify sRNAs with a relevant role in biological processes.

The samples used to test SeqCluster showed diverse sources of sRNAs. In the human brain these clustered sRNAs (usRNAs) were highly represented by repeat and non-coding RNA classes, some of them published by other groups [239, 134, 259, 80]. The proportion of each usRNA type is highly similar between the two human tissues studied (brain and stem cells), with an overlap of 60% of the same usRNAs. In both human and mouse, we found similar results except for repeat-derived sRNAs which

are clearly less frequent in mouse. This is most likely due to the fact that LINE repeats are a major source of sRNAs in the human genome, being twice more abundant in the human genome as compared to the mouse genome. The rest of the repeat derived sRNAs are equally represented in the human and the mouse genomes.

The differences we found between *D. melanogaster* and *C. elegans* are likely due to incomplete annotation of these genomes. In *D. melanogaster*, there is no annotation for tRNA and SNOR families, which comprise 90% of the non-coding derived sRNAs in the human and mouse datasets, explaining the minimum percentage of this class here. In addition, 60% of the genome was annotated as 'predicted genes' in *C. elegans*, hence it is expected that there is a high percentage of sRNAs predicted to lie in apparently inter-genic regions. Some of these sRNA may be classified as junk sRNAs generated as a by-product of cellular activities or sequencing errors, because the sequences forming part of the cluster are randomly distributed in large regions. To determine whether these usRNA show evidence of important new functional roles, it will be necessary to test their function using experimental methods. Nevertheless, our work presented here as well as analyses of other groups [261, 259, 239] show that the investigation of sRNAs can lead to the discovery of novel sRNAs with a functional role in gene expression, such as, tiny RNAs (located in the transcription start site of genes) or spli-RNAs (located closed to splicing site events), published by [261].

4 Conclusions

SeqBuster is extremely useful to evaluate tissue-specific isomiR distributions relevant in different conditions, such as development, physiology and disease conditions.

1. A seed-based algorithm integrated in Java for the miRNA detection provides independence from external tools, and improves the analytical time consuming process, conserving specificity.

2. A pre-analysis step is not restricted to the databases stored in the web service, overcoming the limitations in the storage capacities detected in other web-based bioinformatics tools.

3. The R environment, in which the different analysis packages have been developed, allows the incorporation and/or modification of different types of analyses.

4. The analysis packages are dedicated to generate a dynamic and interactive interface containing a deep characterization of all isomiRs and their relevance in the sample.

SeqCluster provides a novel framework integrating different levels of processes to successfully classify non-miRNA small RNA data from next generation sequencing technologies.

1. SeqCluster is highly flexible, allowing the incorporation of different genomes and databases for mapping and annotation purposes, providing the possibility to use different mapping tools.

2. SeqCluster detects and successfully classifies all types of known sRNAs and keeps all unclassified sequences for manual inspection, therefore providing an unbiased analysis of non-miRNA sRNAs.

MiRNA variability is a widespread phenomena in different species and tissues.

1. The presence of isomiRs detected in different tissues and stages from worm to human, suggest a relevant role of miRNA sequence plasticity in evolution.

2. In most cases only two isomiR sequences where highly represented in each miRNA gene, suggesting that a limited number of variants have a functional role.

3. For some miRNA genes, the most represented sequence differed depending on the specie and biological condition, indicating that different isomiRs may influence selective processes.

4. The miRNA sequence length was strongly conserved when variation occurred, implying that length restriction is an important functional feature.

5. No structural sequence features correlated with the different types of isomiRs, suggesting that other biological processes underlie isomiR expression.

6. The poor variability at the 5'-end of the miRNAs compared with the 3'-end is in accordance with the importance of the 5'-region in mRNA target recognition and silencing.

7. The non-random nucleotide changes along miRNA sequence suggest a new layer of plasticity in mRNA targeting and silencing.

MiRNA variability is a ubiquitous phenomenon in the adult human brain that may influence the mechanism of gene expression modulation underlying physiological and pathological processes.

1. The expression patterns for most isomiRs correlated with that of the corresponding reference miRNA in physiology (human brain development and aging) and pathology (HuntingtonÕs disease), indicating no major alterations of isomiR biogenesis in these processes.

2. The differential expression pattern of some isomiRs, compared with the reference miRNA, may imply a functional role for selective isomiR species in brain pathology and physiology.

3. The isomiR-mRNA and reference miRNA-mRNA duplex stabilities may have a relevant role in the definition of the target transcriptome.

4. The different seeds of each miRNA were predicted to affect many common functions underlying brain aging and neurodegeneration. This agrees with the idea that the different types of isomiRs may differently target the transcriptome (in terms of efficiency and/or type of targets) but produce, in some cases, a similar overall functional output.

5. The isomiRs, differently expressed in brain physiology and pathology identified some targets enriched in specific functions different from those of the reference miRNAs, supporting the idea of a particular functionality of miRNA variants.

Abbrevations

AGO	Argonauta protein
AFM	Atomic force microscopy
CAGE	Cap analysis of gene expression
cDNA	complementary DNA
CNS	Central nervous system
CUT	Cryptic unstable transcript
DCL	Dicer like protein
DCP2	Decapping protein
DGCR8	DiGeorge syndrome critical region gene 8
Dicer	Dicer 1, ribonuclease type III
DNA	Deoxyribonucleic acid
Drosha	double-stranded dsRNA-specific endoribonucleases
dsRNA	Double-stranded RNA
FC	Frontal cortex
FRET	Fluorescence resonance energy transfer
GRO-seq	Global run sequencing
hc-siRNA	Heterochromatic siRNA
HD	Huntington disease
ICGC	International cancer genome consortium
IRES	Internal ribosome entry site
lincRNA	Long interspersed ncRNA
lncRNA	Long non-coding RNA

miRNA	MicroRNA
miRNP	Micro-ribonucleoproteins
mRNA	Messenger RNA
nat-siRNA	Natural antisense transcript-derived siRNA
nasRNA	Non-coding associated sRNA
ncRNA	Non-coding RNA
nt	Nucelotides
OCL	Overlap-layout-consensus
PABPC	Cytoplasmic poly(A)-binding protein
PALR	Promoter-associated lncRNA
PASR	Promoter-associated sRNA
PEM	Pair-end mapping
piRNA	PIWI-interacting RNA
PIWI	CG6122 gene product from transcript CG6122-RA
Pol II-III	RNA polymerase II-III
PROMPT	Promoter upstream transcript
RdRP	RNA-dependent RNA polymerase
REST	RE1-silencing transcription factor
RISC	RNA-induced silencing complex
RNA	Ribonucleic acid
rRNA	Ribosomal RNA
SAGE	Serial analysis of gene expression

siRNA	Small interference RNA
SMRT	Single real time sequencing
SNP	Single nucleotide polymorphism
spli-RNA	Splicing site derived sRNA
sRNA	Small RNA
ssRNA	Single-stranded RNA
ST	Striatum
ta-siRNA	Trans-acting siRNA
tasi-RNA	Gene termini associated human RNA
TE	Transposon element
tiRNA	Transcription initiation RNA
TRBP	TAR RNA binding protein
tRNA	Transfer RNA
tSMS	True single molecule sequencing
TSS	Transcription start site
TUF	Transcript of unknown function
ZMW	Zero-mode waveguides

Annex

List of publications

1. Marti E, Pantano L, Bañez-Coronel M, Llorens F, Miñones-Moyano E, Porta S, Sumoy L, Ferrer I, Estivill X. **A myriad of miRNA variants in control and Huntington's disease brain regions detected by massively parallel sequencing.** Nucleic Acids Res. 2010 Nov 1;38(20):7219-35. Epub 2010 Jun 30.

2. Pantano L, Estivill X, Marti E. **SeqBuster, a bioinformatic tool for the processing and analysis of small RNAs datasets, reveals ubiquitous miRNA modifications in human embryonic cells.Nucleic Acids Res.** 2010 Mar;38(5):e34. Epub 2009 Dec 11.

3. Armengol L,Villatoro S,Gonzalez JR, Pantano L, Garcia-Aragones M, Rabionet K, Estivill X. **Identification of copy number variants defining genomic differences among major human groups.** PLoS One. 2009 Sep 30;4(9):e7230.

4. Varas F, Stadtfeld M, De Andres L, Maherali N, di Tullio A, Pantano L, Notredame C, Hochedlinger K and Graf T. **Fibroblast derived induced pluripotent stem cells show no common retroviral insertions.** Stem Cells. 2009 Feb;27(2):300-6.

5. Pantano L, Armengol L, Villatoro S, Estivill X. **ProSeeK: A web server for MLPA probe.** BMC Genomics. 2008 Nov 28;9:573.

6. Khaja R, Zhang J, MacDonald JR, He Y, Joseph-George AM, Wei J, Rafiq MA, Qian C, Shago M, Pantano L, Aburatani H, Jones K, Redon R, Hurles M, Armengol L, Estivill,X, Mural RJ, Lee C,

Scherer SW, Feuk L. **Genome assembly comparison identifies structural variants in the human genome.** Nat Genet. 2006 Dec;38(12):1413-8.

Communications to scientific meetings

Poster presentations

18th Annual International Conference on Intelligent Systems for Molecular Biology. 7-9 July 2010. Boston. USA.
SeqCluster, small RNA characterization tool. Pantano L, Estivill X, Marti E.

Silencing RNAs: organisers and coordinators of complexity in eukaryotic organisms - SIROCCO meeting. 16-18 November 2009. Hixton. UK.
SeqBuster, a bioinformatic tool for the processing and analysis of small RNAs datasets. Pantano L, Estivill X, Marti E.

The European Society of Human Genetics. May 31 - June 3 2008. Barcelona. Spain.
ProSeeK: A web server for MLPA probe. Pantano L, Armengol L, Villatoro S, Estivill X.

Oral presentation

Silencing RNAs: organisers and coordinators of complexity in eukaryotic organisms - SIROCCO meeting. 16-18 November 2009. Hixton. UK.
SeqBuster, a bioinformatic tool for the processing and analysis of small RNAs datasets. Pantano L, Estivill X, Marti E.

Bibliography

[1] http://seqanswers.com/.

[2] http://www.454.com/.

[3] http://www.appliedbiosystems.com.

[4] http://www.helicosbio.com/.

[5] http://www.illumina.com/.

[6] "prize overview: Archon x prize for genomics".

[7] "visigen biotechnologies inc. - technology overview".

[8] www.pacificbiosciences.com.

[9] Affymetrix ENCODE Transcriptome Project and Cold Spring Harbor Laboratory ENCODE Transcriptome Project. Post-transcriptional processing generates a diversity of 5'-modified long and short rnas. Nature, 457(7232):1028–1032, Feb 2009.

[10] S F Altschul, W Gish, W Miller, E W Myers, and D J Lipman. Basic local alignment search tool. J Mol Biol, 215(3):403–410, Oct 1990.

[11] V Ambros. The functions of animal micrornas. Nature, 431(7006):350–355, Sep 2004.

[12] Alexei Aravin and Thomas Tuschl. Identification and characterization of small rnas involved in rna silencing. FEBS Lett, 579(26):5830–5840, Oct 2005.

[13] et al. Aravin, A. A novel class of small rnas bind to mili protein in mouse testes. Nature, 442:203–207, 2006.

[14] et al. Aravin, A.A. A pirna pathway primed by individual transposons is linked to de novo dna methylation in mice. Molecular Cell, 31:785–799, 2008.

[15] J M Aury, C Cruaud, V Barbe, O Rogier, S Mangenot, G Samson, J Poulain, V Anthouard, C Scarpelli, F Artiguenave, and P Wincker. High quality draft sequences for prokaryotic genomes using a mix of new sequencing technologies. BMC Genomics, 9:603–603, 2008.

[16] J E Babiarz, J G Ruby, Y Wang, D P Bartel, and R Blelloch. Mouse es cells express endogenous shrnas, sirnas, and other microprocessor-independent, dicer-dependent small rnas. Genes Dev, 22(20):2773–2785, Oct 2008.

[17] J P Bachellerie, J Cavaillé, and A Hüttenhofer. The expanding snorna world. Biochimie, 84(8):775–790, Aug 2002.

[18] D Baek, J Villén, C Shin, F D Camargo, S P Gygi, and D P Bartel. The impact of micrornas on protein output. Nature, 455(7209):64–71, Sep 2008.

[19] S Bagga, J Bracht, S Hunter, K Massirer, J Holtz, R Eachus, and A E Pasquinelli. Regulation by let-7 and lin-4 mirnas results in target mrna degradation. Cell, 122(4):553–563, Aug 2005.

[20] Sophie Bail, Mavis Swerdel, Hudan Liu, Xinfu Jiao, Loyal A Goff, Ronald P Hart, and Megerditch Kiledjian. Differential regulation of microrna stability. RNA, 16(5):1032–1039, May 2010.

[21] D P Bartel. Micrornas: genomics, biogenesis, mechanism, and function. Cell, 116(2):281–297, Jan 2004.

[22] D P Bartel. Micrornas: target recognition and regulatory functions. Cell, 136(2):215–233, Jan 2009.

[23] A Bateman and J Quackenbush. Bioinformatics for next generation sequencing. Bioinformatics, 25(4):429–429, Feb 2009.

[24] P J Batista, J G Ruby, J M Claycomb, R Chiang, N Fahlgren, K D Kasschau, D A Chaves, W Gu, J J Vasale, S Duan, D Conte, S Luo, G P Schroth, J C Carrington, D P Bartel, and C C Mello. Prg-1 and 21u-rnas interact to form the pirna complex required for fertility in c. elegans. Mol Cell, 31(1):67–78, Jul 2008.

[25] S Batzoglou, D B Jaffe, K Stanley, J Butler, S Gnerre, E Mauceli, B Berger, J P Mesirov, and E S Lander. Arachne: a whole-genome shotgun assembler. Genome Res, 12(1):177–189, Jan 2002.

[26] Dominik Beck, Steve Ayers, Jianguo Wen, Miriam B Brandl, Tuan D Pham, Paul Webb, Chung-Che Chang, and Xiaobo Zhou. Integrative analysis of next generation sequencing for small non-coding rnas and transcriptional regulation in myelodysplastic syndromes. BMC Med Genomics, 4:19, 2011.

[27] I Behm-Ansmant, J Rehwinkel, T Doerks, A Stark, P Bork, and E Izaurralde. mrna degradation by mirnas and gw182 requires both ccr4:not deadenylase and dcp1:dcp2 decapping complexes. Genes Dev, 20(14):1885–1898, Jul 2006.

[28] S Bennett. Solexa ltd. Pharmacogenomics, 5(4):433–438, Jun 2004.

[29] E Berezikov, W J Chung, J Willis, E Cuppen, and E C Lai. Mammalian mirtron genes. Mol Cell, 28(2):328–336, Oct 2007.

[30] E Berezikov, N Robine, A Samsonova, J O Westholm, A Naqvi, J H Hung, K Okamura, Q Dai, D Bortolamiol-Becet, R Martin, Y Zhao, P D Zamore, G J Hannon, M A Marra, Z Weng, N Perrimon, and E C Lai. Deep annotation of drosophila melanogaster microrna yields insights into their processing, modification, and emergence. Genome Res, Jan 2011.

[31] Philipp Berninger, Dimos Gaidatzis, Erik van Nimwegen, and Mihaela Zavolan. Computational analysis of small rna cloning data. Methods, 44(1):13–21, Jan 2008.

[32] B E Bernstein, J A Stamatoyannopoulos, J F Costello, B Ren, A Milosavljevic, A Meissner, M Kellis, M A Marra, A L Beaudet, J R Ecker, P J Farnham, M Hirst, E S Lander, T S Mikkelsen, and J A Thomson. The nih roadmap epigenomics mapping consortium. Nat Biotechnol, 28(10):1045–1048, Oct 2010.

[33] E Bernstein, A A Caudy, S M Hammond, and G J Hannon. Role for a bidentate ribonuclease in the initiation step of rna interference. Nature, 409(6818):363–366, Jan 2001.

[34] E. et al Birney. Identification and analysis of functional elements in 1% of the human genome by the encode pilot porject. Nature, 447:799–816, 2007.

[35] Matthew J Blow, Russell J Grocock, Stijn van Dongen, Anton J Enright, Ed Dicks, P Andrew Futreal, Richard Wooster, and Michael R Stratton. Rna editing of human microrna. Genome Biol, 7(4):R27, 2006.

[36] M T Bohnsack, K Czaplinski, and D Gorlich. Exportin 5 is a rangtp-dependent dsrna-binding protein that mediates nuclear export of pre-mirnas. RNA, 10(2):185–191, Feb 2004.

[37] E Bonnet, J Wuyts, P Rouzé, and Y Van de Peer. Evidence that microrna precursors, unlike other non-coding rnas, have lower folding free energies than random sequences. Bioinformatics, 20(17):2911–2917, Nov 2004.

[38] G M Borchert, W Lanier, and B L Davidson. Rna polymerase iii transcribes human microrna. Nat Struct Mol Biol, 13(12):1097–1101, Dec 2006.

[39] et al. Brennecke, J. Discrete small rna-generating loci as master regulators of transposon activity in drosophila. Cell, 128:7089–1103, 2007.

[40] et al. Brennecke, J. An epigenetic role for maternally inherited pirnas in transposon silencing. Science, 322:1387–1392, 2008.

[41] Julius Brennecke, Alexander Stark, Robert B Russell, and Stephen M Cohen. Principles of microrna-target recognition. PLoS Biol, 3(3):e85, Mar 2005.

[42] P Brodersen and O Voinnet. Revisiting the principles of microrna target recognition and mode of action. Nat Rev Mol Cell Biol, 10(2):141–148, Feb 2009.

[43] H P Buermans, Y Ariyurek, G van Ommen, J T den Dunnen, and P A 't Hoen. New methods for next generation sequencing based microrna expression profiling. BMC Genomics, 11:716–716, 2010.

[44] A M Burroughs, Y Ando, M J de Hoon, Y Tomaru, T Nishibu, R Ukekawa, T Funakoshi, T Kurokawa, H Suzuki, Y Hayashizaki, and C O Daub. A comprehensive survey of 3' animal mirna modification events and a possible role for 3' adenylation in modulating mirna targeting effectiveness. Genome Res, 20(10):1398–1410, Oct 2010.

[45] Wheeler D Burrows M. A block sorting lossless data compression algorithm. Technical Report, page 124, 1994.

[46] N Bushati and S M Cohen. microrna functions. Annu Rev Cell Dev Biol, 23:175–205, 2007.

[47] J Butler, I MacCallum, M Kleber, IA Shlyakhter, MK Belmonte, ES Lander, C Nusbaum, and DB Jaffe. Allpaths: De novo assembly of whole-genome shotgun microreads. Genome Res, 18:810–820, 2008.

[48] X Cai, C H Hagedorn, and B R Cullen. Human micrornas are processed from capped, polyadenylated transcripts that can also function as mrnas. RNA, 10(12):1957–1966, Dec 2004.

[49] A Calado, N Treichel, E C Müller, A Otto, and U Kutay. Exportin-5-mediated nuclear export of eukaryotic elongation factor 1a and trna. EMBO J, 21(22):6216–6224, Nov 2002.

[50] D Campagna, A Albiero, A Bilardi, E Caniato, C Forcato, S Manavski, N Vitulo, and G Valle. Pass: a program to align short sequences. Bioinformatics, 25(7):967–968, Apr 2009.

[51] P Carninci, A Sandelin, B Lenhard, S Katayama, K Shimokawa, J Ponjavic, C A Semple, M S Taylor, P G Engström, M C Frith, A R Forrest, W B Alkema, S L Tan, C Plessy, R Kodzius, T Ravasi, T Kasukawa, S Fukuda, M Kanamori-Katayama, Y Kitazume, H Kawaji, C Kai, M Nakamura, H Konno, K Nakano, S Mottagui-Tabar, P Arner, A Chesi, S Gustincich, F Persichetti, H Suzuki, S M Grimmond, C A Wells, V Orlando, C Wahlestedt, E T Liu, M Harbers, J Kawai, V B Bajic, D A Hume, and Y Hayashizaki. Genome-wide analysis of mammalian promoter architecture and evolution. Nat Genet, 38(6):626–635, Jun 2006.

[52] M Chalfie, H R Horvitz, and J E Sulston. Mutations that lead to reiterations in the cell lineages of c. elegans. Cell, 24(1):59–69, Apr 1981.

[53] A M Chalk, C Wahlestedt, and E L Sonnhammer. Improved and automated prediction of effective sirna. Biochem Biophys Res Commun, 319(1):264–274, Jun 2004.

[54] S W Chan, D Zilberman, Z Xie, L K Johansen, J C Carrington, and S E Jacobsen. Rna silencing genes control de novo dna methylation. Science, 303(5662):1336–1336, Feb 2004.

[55] S Cheloufi, C O Dos Santos, M M Chong, and G J Hannon. A dicer-independent mirna biogenesis pathway that requires ago catalysis. Nature, 465(7298):584–589, Jun 2010.

[56] D Chen, Y Meng, C Yuan, L Bai, D Huang, S Lv, P Wu, L L Chen, and M Chen. Plant sirnas from introns mediate dna methylation of host genes. RNA, 17(6):1012–1024, Jun 2011.

[57] S W Chi, J B Zang, A Mele, and R B Darnell. Argonaute hits-clip decodes microrna-mrna interaction maps. Nature, 460(7254):479–486, Jul 2009.

[58] W J Chung, K Okamura, R Martin, and E C Lai. Endogenous rna interference provides a somatic defense against drosophila transposons. Curr Biol, 18(11):795–802, Jun 2008.

[59] D Cifuentes, H Xue, D W Taylor, H Patnode, Y Mishima, S Cheloufi, E Ma, S Mane, G J Hannon, N D Lawson, S A Wolfe, and A J Giraldez. A novel mirna processing pathway independent of dicer requires argonaute2 catalytic activity. Science, 328(5986):1694–1698, Jun 2010.

[60] J M Claycomb, P J Batista, K M Pang, W Gu, J J Vasale, J C van Wolfswinkel, D A Chaves, M Shirayama, S Mitani, R F Ketting, D Conte, and C C Mello. The argonaute csr-1 and its 22g-rna cofactors are required for holocentric chromosome segregation. Cell, 139(1):123–134, Oct 2009.

[61] N Cloonan, A R Forrest, G Kolle, B B Gardiner, G J Faulkner, M K Brown, D F Taylor, A L Steptoe, S Wani, G Bethel, A J Robertson, A C Perkins, S J Bruce, C C Lee, S S Ranade, H E Peckham, J M Manning, K J McKernan, and S M Grimmond. Stem cell transcriptome profiling via massive-scale mrna sequencing. Nat Methods, 5(7):613–619, Jul 2008.

[62] L J Core, J J Waterfall, and J T Lis. Nascent rna sequencing reveals widespread pausing and divergent initiation at human promoters. Science, 322(5909):1845–1848, Dec 2008.

[63] D N Cox, A Chao, J Baker, L Chang, D Qiao, and H Lin. A novel class of evolutionarily conserved genes defined by piwi are essential

for stem cell self-renewal. Genes Dev, 12(23):3715–3727, Dec 1998.

[64] D N Cox, A Chao, and H Lin. piwi encodes a nucleoplasmic factor whose activity modulates the number and division rate of germline stem cells. Development, 127(3):503–514, Feb 2000.

[65] B Czech and G J Hannon. Small rna sorting: matchmaking for argonautes. Nat Rev Genet, 12(1):19–31, Jan 2011.

[66] B Czech, C D Malone, R Zhou, A Stark, C Schlingeheyde, M Dus, N Perrimon, M Kellis, J A Wohlschlegel, R Sachidanandam, G J Hannon, and J Brennecke. An endogenous small interfering rna pathway in drosophila. Nature, 453(7196):798–802, Jun 2008.

[67] Asis K Das and Gordon G Carmichael. Adar editing wobbles the microrna world. ACS Chem Biol, 2(4):217–220, Apr 2007.

[68] et al. Das, P.P. Piwi and pirnas act upstream of an endogenous sirna pathway to suppress tc3 transposon mobility in the caenorhabditis elegans germline. Molecular Cell, 31:79–90, 2008.

[69] A M Denli, B B Tops, R H Plasterk, R F Ketting, and G J Hannon. Processing of primary microrna by the microprocessor complex. Nature, 432(7014):231–235, Nov 2004.

[70] M C Derry, A Yanagiya, Y Martineau, and N Sonenberg. Regulation of poly(a)-binding protein through pabp-interacting proteins. Cold Spring Harb Symp Quant Biol, 71:537–543, 2006.

[71] J C Dohm, C Lottaz, T Borodina, and H Himmelbauer. Sharcgs, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing. Genome Res, 17(11):1697–1706, Nov 2007.

[72] Juliane C Dohm, Claudio Lottaz, Tatiana Borodina, and Heinz Himmelbauer. Substantial biases in ultra-short read data sets

from high-throughput dna sequencing. Nucleic Acids Res, 36(16):e105, Sep 2008.

[73] M Droege and B Hill. The genome sequencer flx system–longer reads, more applications, straight forward bioinformatics and more complete data sets. J Biotechnol, 136(1-2):3–10, Aug 2008.

[74] S Duharcourt, G Lepère, and E Meyer. Developmental genome rearrangements in ciliates: a natural genomic subtraction mediated by non-coding transcripts. Trends Genet, 25(8):344–350, Aug 2009.

[75] H L Eaves and Y Gao. Mom: maximum oligonucleotide mapping. Bioinformatics, 25(7):969–970, Apr 2009.

[76] Margaret S Ebert, Joel R Neilson, and Phillip A Sharp. Microrna sponges: competitive inhibitors of small rnas in mammalian cells. Nat Methods, 4(9):721–726, Sep 2007.

[77] H A Ebhardt, H H Tsang, D C Dai, Y Liu, B Bostan, and R P Fahlman. Meta-analysis of small rna-sequencing errors reveals ubiquitous post-transcriptional rna modifications. Nucleic Acids Res, 37(8):2461–2470, May 2009.

[78] R Edgar, M Domrachev, and A E Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. Nucleic Acids Res, 30(1):207–210, Jan 2002.

[79] J Eid, A Fehr, J Gray, K Luong, J Lyle, G Otto, P Peluso, D Rank, P Baybayan, B Bettman, A Bibillo, K Bjornson, B Chaudhuri, F Christians, R Cicero, S Clark, R Dalal, A Dewinter, J Dixon, M Foquet, A Gaertner, P Hardenbol, C Heiner, K Hester, D Holden, G Kearns, X Kong, R Kuse, Y Lacroix, S Lin, P Lundquist, C Ma, P Marks, M Maxham, D Murphy, I Park, T Pham, M Phillips, J Roy, R Sebra, G Shen, J Sorenson, A Tomaney, K Travers, M Trulson, J Vieceli, J Wegener, D Wu, A Yang, D Zaccarin, P Zhao, F Zhong, J Korlach, and S Turner. Real-time dna sequencing from single polymerase molecules. Science, 323(5910):133–138, Jan 2009.

[80] C Ender, A Krek, M R Friedländer, M Beitzinger, L Weinmann, W Chen, S Pfeffer, N Rajewsky, and G Meister. A human snorna with microrna-like functions. Mol Cell, 32(4):519–528, Nov 2008.

[81] C R Faehnle and L Joshua-Tor. Argonautes confront new small rnas. Curr Opin Chem Biol, 11(5):569–577, Oct 2007.

[82] N Fahlgren, C M Sullivan, K D Kasschau, E J Chapman, J S Cumbie, T A Montgomery, S D Gilbert, M Dasenko, T W Backman, S A Givan, and J C Carrington. Computational and analytical framework for small rna profiling by high-throughput sequencing. RNA, 15(5):992–1002, May 2009.

[83] M Faller, D Toso, M Matsunaga, I Atanasov, R Senturia, Y Chen, Z H Zhou, and F Guo. Dgcr8 recognizes primary transcripts of microrna through highly cooperative binding and formation of higher-order structures. RNA, 16(8):1570–1583, Aug 2010.

[84] K K Farh, A Grimson, C Jan, B P Lewis, W K Johnston, L P Lim, C B Burge, and D P Bartel. The widespread impact of mammalian microrna on mrna repression and evolution. Science, 310(5755):1817–1821, Dec 2005.

[85] S L Fernandez-Valverde, R J Taft, and J S Mattick. Dynamic isomir regulation in drosophila development. RNA, 16(10):1881–1888, Oct 2010.

[86] Witold Filipowicz, Suvendra N Bhattacharyya, and Nahum Sonenberg. Mechanisms of post-transcriptional regulation by microrna: are the answers in sight? Nat Rev Genet, 9(2):102–114, Feb 2008.

[87] P Flicek, M R Amode, D Barrell, K Beal, S Brent, Y Chen, P Clapham, G Coates, S Fairley, S Fitzgerald, L Gordon, M Hendrix, T Hourlier, N Johnson, A Kähäri, D Keefe, S Keenan, R Kinsella, F Kokocinski, E Kulesha, P Larsson, I Longden, W McLaren, B Overduin, B Pritchard, H S Riat, D Rios, G R Ritchie, M Ruffier,

M Schuster, D Sobral, G Spudich, Y A Tang, S Trevanion, J Vandrovcova, A J Vilella, S White, S P Wilder, A Zadissa, J Zamora, B L Aken, E Birney, F Cunningham, I Dunham, R Durbin, X M Fernández-Suarez, J Herrero, T J Hubbard, A Parker, G Proctor, J Vogel, and S M Searle. Ensembl 2011. Nucleic Acids Res, 39(Database issue):800–806, Jan 2011.

[88] R C Friedman, K K Farh, C B Burge, and D P Bartel. Most mammalian mrnas are conserved targets of microrna. Genome Res, 19(1):92–105, Jan 2009.

[89] M R Garcia-Silva, M Frugier, J P Tosar, A Correa-Dominguez, L Ronalte-Alves, A Parodi-Talice, C Rovira, C Robello, S Goldenberg, and A Cayota. A population of trna-derived small rnas is actively produced in trypanosoma cruzi and recruited to specific cytoplasmic granules. Mol Biochem Parasitol, 171(2):64–73, Jun 2010.

[90] Michael R. Garey and David S. Johnson. Computers and Intractability: A Guide to the Theory of NP-Completeness. W. H. Freeman and Company W. H. Freeman and Company W. H. Freeman and Company W. H. Freeman and Company W. H. Freeman and Company, 1979.

[91] T W Geisbert, A C Lee, M Robbins, J B Geisbert, A N Honko, V Sood, J C Johnson, S de Jong, I Tavakoli, A Judge, L E Hensley, and I Maclachlan. Postexposure protection of non-human primates against a lethal ebola virus challenge with rna interference: a proof-of-concept study. Lancet, 375(9729):1896–1905, May 2010.

[92] M Georges, A Clop, F Marcq, H Takeda, D Pirottin, S Hiard, X Tordoir, F Caiment, F Meish, B Bibé, J Bouix, J M Elsen, F Eychenne, E Laville, C Larzul, D Milenkovic, J Tobin, and A C Charlier. Polymorphic microrna-target interactions: a novel

source of phenotypic variation. Cold Spring Harb Symp Quant Biol, 71:343–350, 2006.

[93] M Ghildiyal, H Seitz, M D Horwich, C Li, T Du, S Lee, J Xu, E L Kittler, M L Zapp, Z Weng, and P D Zamore. Endogenous sirnas derived from transposons and mrnas in drosophila somatic cells. Science, 320(5879):1077–1081, May 2008.

[94] A J Giraldez, Y Mishima, J Rihel, R J Grocock, S Van Dongen, K Inoue, A J Enright, and A F Schier. Zebrafish mir-430 promotes deadenylation and clearance of maternal mrnas. Science, 312(5770):75–79, Apr 2006.

[95] E A Glazov, P A Cottee, W C Barris, R J Moore, B P Dalrymple, and M L Tizard. A microrna catalog of the developing chicken embryo identified by a deep sequencing approach. Genome Res, 18(6):957–964, Jun 2008.

[96] M Gowda, C C Nunes, J Sailsbery, M Xue, F Chen, C A Nelson, D E Brown, Y Oh, S Meng, T Mitchell, C H Hagedorn, and R A Dean. Genome-wide characterization of methylguanosine-capped and polyadenylated small rnas in the rice blast fungus magnaporthe oryzae. Nucleic Acids Res, 38(21):7558–7569, Nov 2010.

[97] S Griffiths-Jones. The microrna registry. Nucleic Acids Res, 32(Database issue):109–111, Jan 2004.

[98] A Grimson, M Srivastava, B Fahey, B J Woodcroft, H R Chiang, N King, B M Degnan, D S Rokhsar, and D P Bartel. Early origins and evolution of micrornas and piwi-interacting rnas in animals. Nature, 455(7217):1193–1197, Oct 2008.

[99] A Grishok, A E Pasquinelli, D Conte, N Li, S Parrish, I Ha, D L Baillie, A Fire, G Ruvkun, and C C Mello. Genes and mechanisms related to rna interference regulate expression of the small temporal rnas that control c. elegans developmental timing. Cell, 106(1):23–34, Jul 2001.

[100] A R Gruber, R Lorenz, S H Bernhart, R Neuböck, and I L Hofacker. The vienna rna websuite. Nucleic Acids Res, 36(Web Server issue):70–74, Jul 2008.

[101] W Gu, M Shirayama, D Conte, J Vasale, P J Batista, J M Claycomb, J J Moresco, E M Youngman, J Keys, M J Stoltz, C C Chen, D A Chaves, S Duan, K D Kasschau, N Fahlgren, J R Yates, S Mitani, J C Carrington, and C C Mello. Distinct argonaute-mediated 22g-rna pathways direct genome surveillance in the c. elegans germline. Mol Cell, 36(2):231–244, Oct 2009.

[102] L Guo and Z Lu. Global expression analysis of mirna gene cluster and family based on isomirs from deep sequencing data. Comput Biol Chem, 34(3):165–171, Jun 2010.

[103] M Hackenberg, M Sturm, D Langenberger, J M Falcón-Pérez, and A M Aransay. miranalyzer: a microrna detection and analysis tool for next-generation sequencing experiments. Nucleic Acids Res, 37(Web Server issue):68–76, Jul 2009.

[104] M Hafner, M Landthaler, L Burger, M Khorshid, J Hausser, P Berninger, A Rothballer, M Ascano, A C Jungkamp, M Munschauer, A Ulrich, G S Wardle, S Dewell, M Zavolan, and T Tuschl. Transcriptome-wide identification of rna-binding protein and microrna target sites by par-clip. Cell, 141(1):129–141, Apr 2010.

[105] A Hamilton, O Voinnet, L Chappell, and D Baulcombe. Two classes of short interfering rna in rna silencing. EMBO J, 21(17):4671–4679, Sep 2002.

[106] A J Hamilton and D C Baulcombe. A species of small antisense rna in posttranscriptional gene silencing in plants. Science, 286(5441):950–952, Oct 1999.

[107] S M Hammond, E Bernstein, D Beach, and G J Hannon. An rna-directed nuclease mediates post-transcriptional gene silencing in drosophila cells. Nature, 404(6775):293–296, Mar 2000.

[108] J Han, Y Lee, K H Yeom, J W Nam, I Heo, J K Rhee, S Y Sohn, Y Cho, B T Zhang, and V N Kim. Molecular basis for the recognition of primary micrornas by the drosha-dgcr8 complex. Cell, 125(5):887–901, Jun 2006.

[109] J V Hartig, S Esslinger, R Böttcher, K Saito, and K Förstemann. Endo-sirnas depend on a new isoform of loquacious and target artificially introduced, high-copy sequences. EMBO J, 28(19):2932–2944, Oct 2009.

[110] D Haussecker, Y Huang, A Lau, P Parameswaran, A Z Fire, and M A Kay. Human trna-derived small rnas in the global regulation of rna silencing. RNA, 16(4):673–695, Apr 2010.

[111] P Havlak, R Chen, K J Durbin, A Egan, Y Ren, X Z Song, G M Weinstock, and R A Gibbs. The atlas genome assembly system. Genome Res, 14(4):721–732, Apr 2004.

[112] Sebastien S Hebert and Bart De Strooper. Alterations of the microrna network cause neurodegenerative disease. Trends Neurosci, 32(4):199–206, Apr 2009.

[113] D Hernandez, P François, L Farinelli, M Osterås, and J Schrenzel. De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. Genome Res, 18(5):802–809, May 2008.

[114] Angela Hodges, Andrew D Strand, Aaron K Aragaki, Alexandre Kuhn, Thierry Sengstag, Gareth Hughes, Lyn A Elliston, Cathy Hartog, Darlene R Goldstein, Doris Thu, Zane R Hollingsworth, Francois Collin, Beth Synek, Peter A Holmans, Anne B Young, Nancy S Wexler, Mauro Delorenzi, Charles Kooperberg, Sarah J Augood, Richard L M Faull, James M Olson, Lesley Jones, and Ruth Luthi-Carter. Regional and cellular gene expression changes in human huntington's disease brain. Hum Mol Genet, 15(6):965–977, Mar 2006.

[115] D S Horner, G Pavesi, T Castrignanò, P D De Meo, S Liuni, M Sammeth, E Picardi, and G Pesole. Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. Brief Bioinform, 11(2):181–197, Mar 2010.

[116] E Hornstein and N Shomron. Canalization of development by micrornas. Nat Genet, 38 Suppl:20–24, Jun 2006.

[117] et al. Houwing, S. A role for piwi and pirnas in germ cell maintenance and transposon silencing in zebrafish. Cell, 129:69–82, 2007.

[118] d a W Huang, B T Sherman, and R A Lempicki. Systematic and integrative analysis of large gene lists using david bioinformatics resources. Nat Protoc, 4(1):44–57, 2009.

[119] X Huang, J Wang, S Aluru, S P Yang, and L Hillier. Pcap: a whole-genome assembly program. Genome Res, 13(9):2164–2170, Sep 2003.

[120] E Huntzinger and E Izaurralde. Gene silencing by micrornas: contributions of translational repression and mrna decay. Nat Rev Genet, 12(2):99–110, Feb 2011.

[121] A Hüttenhofer, M Kiefmann, S Meier-Ewert, J O'Brien, H Lehrach, J P Bachellerie, and J Brosius. Rnomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger rnas in mouse. EMBO J, 20(11):2943–2953, Jun 2001.

[122] G Hutvágner, J McLachlan, A E Pasquinelli, E Bálint, T Tuschl, and P D Zamore. A cellular function for the rna-interference enzyme dicer in the maturation of the let-7 small temporal rna. Science, 293(5531):834–838, Aug 2001.

[123] G Hutvágner and P D Zamore. A microrna in a multiple-turnover rnai enzyme complex. Science, 297(5589):2056–2060, Sep 2002.

[124] K Iida, H Jin, and J K Zhu. Bioinformatics analysis suggests base modifications of trnas and mirnas in arabidopsis thaliana. BMC Genomics, 10:155–155, 2009.

[125] A Jacquier. The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small rnas. Nat Rev Genet, 10(12):833–844, Dec 2009.

[126] W R Jeck, J A Reinhardt, D A Baltrus, M T Hickenbotham, V Magrini, E R Mardis, J L Dangl, and C D Jones. Extending assembly of short dna sequences to handle error. Bioinformatics, 23(21):2942–2944, Nov 2007.

[127] H Jiang and W H Wong. Seqmap: mapping massive amount of oligonucleotides to the genome. Bioinformatics, 24(20):2395–2396, Oct 2008.

[128] R Johnson and N J Buckley. Gene dysregulation in huntington's disease: Rest, micrornas and beyond. Neuromolecular Med, 11(3):183–199, 2009.

[129] R Johnson, C Zuccato, N D Belyaev, D J Guest, E Cattaneo, and N J Buckley. A microrna-based gene dysregulation pathway in huntington's disease. Neurobiol Dis, 29(3):438–445, Mar 2008.

[130] P Kapranov, J Cheng, S Dike, D A Nix, R Duttagupta, A T Willingham, P F Stadler, J Hertel, J Hackermüller, I L Hofacker, I Bell, E Cheung, J Drenkow, E Dumais, S Patel, G Helt, M Ganesh, S Ghosh, A Piccolboni, V Sementchenko, H Tammana, and T R Gingeras. Rna maps reveal new rna classes and a possible function for pervasive transcription. Science, 316(5830):1484–1488, Jun 2007.

[131] Y Katz, E T Wang, E M Airoldi, and C B Burge. Analysis and design of rna sequencing experiments for identifying isoform regulation. Nat Methods, 7(12):1009–1015, Dec 2010.

[132] Yukio Kawahara, Boris Zinshteyn, Thimmaiah P Chendrimada, Ramin Shiekhattar, and Kazuko Nishikura. Rna editing of the microrna-151 precursor blocks cleavage by the dicer-trbp complex. EMBO Rep, 8(8):763–769, Aug 2007.

[133] Yukio Kawahara, Boris Zinshteyn, Praveen Sethupathy, Hisashi Iizasa, Artemis G Hatzigeorgiou, and Kazuko Nishikura. Redirection of silencing targets by adenosine-to-inosine editing of mirnas. Science, 315(5815):1137–1140, Feb 2007.

[134] H Kawaji, M Nakamura, Y Takahashi, A Sandelin, S Katayama, S Fukuda, C O Daub, C Kai, J Kawai, J Yasuda, P Carninci, and Y Hayashizaki. Hidden layers of human small rnas. BMC Genomics, 9:157–157, 2008.

[135] Y Kawamura, K Saito, T Kin, Y Ono, K Asai, T Sunohara, T N Okada, M C Siomi, and H Siomi. Drosophila endogenous small rnas bind to argonaute 2 in somatic cells. Nature, 453(7196):793–797, Jun 2008.

[136] W J Kent. Blat–the blast-like alignment tool. Genome Res, 12(4):656–664, Apr 2002.

[137] R F Ketting, S E Fischer, E Bernstein, T Sijen, G J Hannon, and R H Plasterk. Dicer functions in rna interference and in synthesis of small rna involved in developmental timing in c. elegans. Genes Dev, 15(20):2654–2659, Oct 2001.

[138] R F Ketting, T H Haverkamp, H G van Luenen, and R H Plasterk. Mut-7 of c. elegans, required for transposon silencing and rna interference, is a homolog of werner syndrome helicase and rnased. Cell, 99(2):133–141, Oct 1999.

[139] A Khvorova, A Reynolds, and S D Jayasena. Functional sirnas and mirnas exhibit strand bias. Cell, 115(2):209–216, Oct 2003.

[140] V N Kim, J Han, and M C Siomi. Biogenesis of small rnas in animals. Nat Rev Mol Cell Biol, 10(2):126–139, Feb 2009.

[141] C King and T Scott-Horton. Pyrosequencing: a simple method for accurate genotyping. J Vis Exp, (11), 2008.

[142] Y. Kirino and Z. Mourelatos. Mouse piwi-interacting rnas are 2[prime]-o-methylated at their 3[prime] termini. Nat Struct Mol Bio, 14:347–348, 347-348.

[143] S Kishore, A Khanna, Z Zhang, J Hui, P J Balwierz, M Stefan, C Beach, R D Nicholls, M Zavolan, and S Stamm. The snorna mbii-52 (snord 115) is processed into smaller rnas and regulates alternative splicing. Hum Mol Genet, 19(7):1153–1164, Apr 2010.

[144] T Kiss. Small nucleolar rnas: an abundant group of noncoding rnas with diverse cellular functions. Cell, 109(2):145–148, Apr 2002.

[145] C. Klattenhoff and W. Theurkauf. Biogenesis and germline functions of pirnas. Development, 135:3–9, 2008.

[146] W P Kloosterman and R H Plasterk. The diverse functions of micrornas in animal development and disease. Dev Cell, 11(4):441–450, Oct 2006.

[147] W P Kloosterman, E Wienholds, R F Ketting, and R H Plasterk. Substrate requirements for let-7 function in the developing zebrafish embryo. Nucleic Acids Res, 32(21):6284–6291, 2004.

[148] S W Knight and B L Bass. A role for the rnase iii enzyme dcr-1 in rna interference and germ line development in caenorhabditis elegans. Science, 293(5538):2269–2271, Sep 2001.

[149] Y W Kong, I G Cannell, C H de Moor, K Hill, P G Garside, T L Hamilton, H A Meijer, H C Dobbyn, M Stoneley, K A Spriggs, A E Willis, and M Bushell. The mechanism of micro-rna-mediated translation repression is determined by the promoter of the target gene. Proc Natl Acad Sci U S A, 105(26):8866–8871, Jul 2008.

[150] J O Korbel, A E Urban, J P Affourtit, B Godwin, F Grubert, J F Simons, P M Kim, D Palejev, N J Carriero, L Du, B E Taillon, Z Chen, A Tanzer, A C Saunders, J Chi, F Yang, N P Carter, M E Hurles, S M Weissman, T T Harkins, M B Gerstein, M Egholm, and M Snyder. Paired-end mapping reveals extensive structural variation in the human genome. Science, 318(5849):420–426, Oct 2007.

[151] A Krek, D Grün, M N Poy, R Wolf, L Rosenberg, E J Epstein, P MacMenamin, I da Piedade, K C Gunsalus, M Stoffel, and N Rajewsky. Combinatorial microrna target predictions. Nat Genet, 37(5):495–500, May 2005.

[152] Florian Kuchenbauer, Ryan D Morin, Bob Argiropoulos, Oleh I Petriv, Malachi Griffith, Michael Heuser, Eric Yung, Jessica Piper, Allen Delaney, Anna-Liisa Prabhu, Yongjun Zhao, Helen McDonald, Thomas Zeng, Martin Hirst, Carl L Hansen, Marco A Marra, and R Keith Humphries. In-depth characterization of the microrna transcriptome in a leukemia progression model. Genome Res, 18(11):1787–1797, Nov 2008.

[153] M Lagos-Quintana, R Rauhut, W Lendeckel, and T Tuschl. Identification of novel genes coding for small expressed rnas. Science, 294(5543):853–858, Oct 2001.

[154] E C Lai. Micro rnas are complementary to 3' utr sequence motifs that mediate negative post-transcriptional regulation. Nat Genet, 30(4):363–364, Apr 2002.

[155] P Landgraf, M Rusu, R Sheridan, A Sewer, N Iovino, A Aravin, S Pfeffer, A Rice, A O Kamphorst, M Landthaler, C Lin, N D Socci, L Hermida, V Fulci, S Chiaretti, R Foà, J Schliwka, U Fuchs, A Novosel, R U Müller, B Schermer, U Bissels, J Inman, Q Phan, M Chien, D B Weir, R Choksi, G De Vita, D Frezzetti, H I Trompeter, V Hornung, G Teng, G Hartmann, M Palkovits, R Di Lauro, P Wernet, G Macino, C E Rogler, J W Nagle, J Ju, F N Papavasiliou, T Benzing, P Lichter, W Tam, M J Brownstein, A Bosio, A Borkhardt,

J J Russo, C Sander, M Zavolan, and T Tuschl. A mammalian microrna expression atlas based on small rna library sequencing. Cell, 129(7):1401–1414, Jun 2007.

[156] B Langmead, C Trapnell, M Pop, and S L Salzberg. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. Genome Biol, 10(3), 2009.

[157] N C Lau, L P Lim, E G Weinstein, and D P Bartel. An abundant class of tiny rnas with probable regulatory roles in caenorhabditis elegans. Science, 294(5543):858–862, Oct 2001.

[158] P Lau and B de Strooper. Dysregulated micrornas in neurodegenerative disorders. Semin Cell Dev Biol, 21(7):768–773, Sep 2010.

[159] E J Lee, S Banerjee, H Zhou, A Jammalamadaka, M Arcila, B S Manjunath, and K S Kosik. Identification of pirnas in the central nervous system. RNA, 17(6):1090–1099, Jun 2011.

[160] L W Lee, S Zhang, A Etheridge, L Ma, D Martin, D Galas, and K Wang. Complexity of the microrna repertoire revealed by next-generation sequencing. RNA, 16(11):2170–2180, Nov 2010.

[161] R C Lee and V Ambros. An extensive class of small rnas in caenorhabditis elegans. Science, 294(5543):862–864, Oct 2001.

[162] R C Lee, R L Feinbaum, and V Ambros. The c. elegans heterochronic gene lin-4 encodes small rnas with antisense complementarity to lin-14. Cell, 75(5):843–854, Dec 1993.

[163] Y Lee, K Jeon, J T Lee, S Kim, and V N Kim. Microrna maturation: stepwise processing and subcellular localization. EMBO J, 21(17):4663–4670, Sep 2002.

[164] Y S Lee, K Nakahara, J W Pham, K Kim, Z He, E J Sontheimer, and R W Carthew. Distinct roles for drosophila dicer-1 and dicer-2 in the sirna/mirna silencing pathways. Cell, 117(1):69–81, Apr 2004.

[165] B P Lewis, C B Burge, and D P Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microrna targets. Cell, 120(1):15–20, Jan 2005.

[166] H Li, J Ruan, and R Durbin. Mapping short dna sequencing reads and calling variants using mapping quality scores. Genome Res, 18(11):1851–1858, Nov 2008.

[167] M Li, I X Wang, Y Li, A Bruzel, A L Richards, J M Toung, and V G Cheung. Widespread rna and dna sequence differences in the human transcriptome. Science, May 2011.

[168] R Li, Y Li, K Kristiansen, and J Wang. Soap: short oligonucleotide alignment program. Bioinformatics, 24(5):713–714, Mar 2008.

[169] R Li, C Yu, Y Li, T W Lam, S M Yiu, K Kristiansen, and J Wang. Soap2: an improved ultrafast tool for short read alignment. Bioinformatics, 25(15):1966–1967, Aug 2009.

[170] Y Li, J Luo, H Zhou, J Y Liao, L M Ma, Y Q Chen, and L H Qu. Stress-induced trna-derived rnas: a novel class of small rnas in the primitive eukaryote giardia lamblia. Nucleic Acids Res, 36(19):6048–6055, Nov 2008.

[171] D D Licatalosi, A Mele, J J Fak, J Ule, M Kayikci, S W Chi, T A Clark, A C Schweitzer, J E Blume, X Wang, J C Darnell, and R B Darnell. Hits-clip yields genome-wide insights into brain alternative rna processing. Nature, 456(7221):464–469, Nov 2008.

[172] L P Lim, N C Lau, P Garrett-Engele, A Grimson, J M Schelter, J Castle, D P Bartel, P S Linsley, and J M Johnson. Microarray analysis shows that some micrornas downregulate large numbers of target mrnas. Nature, 433(7027):769–773, Feb 2005.

[173] et al. Lin, H. The role of the pirna pathway in stem cell self-renewal. Developmental Biology, 319:479–479, 2008.

[174] N K Liu and X M Xu. Microrna in central nervous system trauma and degenerative disorders. Physiol Genomics, Mar 2011.

[175] Q Liu, T A Rand, S Kalidas, F Du, H E Kim, D P Smith, and X Wang. R2d2, a bridge between the initiation and effector steps of the drosophila rnai pathway. Science, 301(5641):1921–1925, Sep 2003.

[176] C Llave, K D Kasschau, M A Rector, and J C Carrington. Endogenous and silencing-associated small rnas in plants. Plant Cell, 14(7):1605–1619, Jul 2002.

[177] S Lu, Y H Sun, and V L Chiang. Adenylation of plant mirnas. Nucleic Acids Res, 37(6):1878–1885, Apr 2009.

[178] Shanfa Lu, Ying-Hsuan Sun, and Vincent L Chiang. Adenylation of plant mirnas. Nucleic Acids Res, 37(6):1878–1885, Apr 2009.

[179] D J Luciano, H Mirsky, N J Vendetti, and S Maas. Rna editing of a mirna precursor. RNA, 10(8):1174–1177, Aug 2004.

[180] E Lund, S Güttinger, A Calado, J E Dahlberg, and U Kutay. Nuclear export of microrna precursors. Science, 303(5654):95–98, Jan 2004.

[181] Y Luo and S Li. Genome-wide analyses of retrogenes derived from the human box h/aca snornas. Nucleic Acids Res, 35(2):559–571, 2007.

[182] C A Maher, C Kumar-Sinha, X Cao, S Kalyana-Sundaram, B Han, X Jing, L Sam, T Barrette, N Palanisamy, and A M Chinnaiyan. Transcriptome sequencing to detect gene fusions in cancer. Nature, 458(7234):97–101, Mar 2009.

[183] C.D. Malone and G.J. Hannon. Small rnas as guardians of the genome. Cell, 136:656–668, 2009.

[184] P A Maroney, Y Yu, J Fisher, and T W Nilsen. Evidence that micrornas are associated with translating messenger rnas in human cells. Nat Struct Mol Biol, 13(12):1102–1107, Dec 2006.

[185] E Martí, L Pantano, M Bañez-Coronel, F Llorens, E Miñones-Moyano, S Porta, L Sumoy, I Ferrer, and X Estivill. A myriad of mirna variants in control and huntington's disease brain regions detected by massively parallel sequencing. Nucleic Acids Res, 38(20):7219–7235, Nov 2010.

[186] H Matsumura, A Ito, H Saitoh, P Winter, G Kahl, M Reuter, D H Krüger, and R Terauchi. Supersage. Cell Microbiol, 7(1):11–18, Jan 2005.

[187] J. S. Mattick. Non-coding rnas: the architects of eukaryotic complexity. EMBO rep, 2:986–91, 2001.

[188] J S Mattick and I V Makunin. Non-coding rna. Hum Mol Genet, 15 Spec No 1:17–29, Apr 2006.

[189] A M Maxam and W Gilbert. A new method for sequencing dna. Proc Natl Acad Sci U S A, 74(2):560–564, Feb 1977.

[190] M J Moore. From birth to death: the complex lives of eukaryotic mrnas. Science, 309(5740):1514–1518, Sep 2005.

[191] R D Morin, M D O'Connor, M Griffith, F Kuchenbauer, A Delaney, A L Prabhu, Y Zhao, H McDonald, T Zeng, M Hirst, C J Eaves, and M A Marra. Application of massively parallel sequencing to microrna profiling and discovery in human embryonic stem cells. Genome Res, 18(4):610–621, Apr 2008.

[192] J C Mullikin and Z Ning. The phusion assembler. Genome Res, 13(1):81–90, Jan 2003.

[193] U Nagalakshmi, Z Wang, K Waern, C Shou, D Raha, M Gerstein, and M Snyder. The transcriptional landscape of the yeast genome defined by rna sequencing. Science, 320(5881):1344–1349, Jun 2008.

[194] N Nagarajan, T D Read, and M Pop. Scaffolding and validation of bacterial genome assemblies using optical restriction maps. Bioinformatics, 24(10):1229–1235, May 2008.

[195] R Niwa and F J Slack. The evolution of animal microrna function. Curr Opin Genet Dev, 17(2):145–150, Apr 2007.

[196] S Nottrott, M J Simard, and J D Richter. Human let-7a mirna blocks protein production on actively translating polyribosomes. Nat Struct Mol Biol, 13(12):1108–1114, Dec 2006.

[197] Sanne Nygaard, Anders Jacobsen, Morten Lindow, Jens Eriksen, Eva Balslev, Henrik Flyger, Niels Tolstrup, Soren Moller, Anders Krogh, and Thomas Litman. Identification and analysis of mirnas in human breast cancer and teratoma samples using deep sequencing. BMC Med Genomics, 2:35, 2009.

[198] K.A. O'Donnell and J.D. Boeke. Mighty piwis defend the germline against genome intruders. Cell, 129:37–44, 2207.

[199] K Okamura, W J Chung, J G Ruby, H Guo, D P Bartel, and E C Lai. The drosophila hairpin rna pathway generates endogenous short interfering rnas. Nature, 453(7196):803–806, Jun 2008.

[200] K Okamura, J W Hagen, H Duan, D M Tyler, and E C Lai. The mirtron pathway generates microrna-class regulatory rnas in drosophila. Cell, 130(1):89–100, Jul 2007.

[201] P H Olsen and V Ambros. The lin-4 regulatory rna controls developmental timing in caenorhabditis elegans by blocking lin-14 protein synthesis after the initiation of translation. Dev Biol, 216(2):671–680, Dec 1999.

[202] M Ono, M S Scott, K Yamada, F Avolio, G J Barton, and A I Lamond. Identification of human mirna precursors that resemble box c/d snornas. Nucleic Acids Res, Jan 2011.

[203] Amy N Packer, Yi Xing, Scott Q Harper, Lesley Jones, and Beverly L Davidson. The bifunctional microrna mir-9/mir-9* regulates rest and corest and is downregulated in huntington's disease. J Neurosci, 28(53):14341–14346, Dec 2008.

[204] J Pak and A Fire. Distinct populations of primary and secondary effectors during rnai in c. elegans. Science, 315(5809):241–244, Jan 2007.

[205] L Pantano, X Estivill, and E Martí. Seqbuster, a bioinformatic tool for the processing and analysis of small rnas datasets, reveals ubiquitous mirna modifications in human embryonic cells. Nucleic Acids Res, 38(5), Mar 2010.

[206] P J Park. Chip-seq: advantages and challenges of a maturing technology. Nat Rev Genet, 10(10):669–680, Oct 2009.

[207] A E Pasquinelli, B J Reinhart, F Slack, M Q Martindale, M I Kuroda, B Maller, D C Hayward, E E Ball, B Degnan, P Müller, J Spring, A Srinivasan, M Fishman, J Finnerty, J Corbo, M Levine, P Leahy, E Davidson, and G Ruvkun. Conservation of the sequence and temporal expression of let-7 heterochronic regulatory rna. Nature, 408(6808):86–89, Nov 2000.

[208] Helena Persson, Anders Kvist, Natalia Rego, Johan Staaf, Johan Vallon-Christersson, Lena Luts, Niklas Loman, Goran Jonsson, Hugo Naya, Mattias Hoglund, Ake Borg, and Carlos Rovira. Identification of new microrna s in paired normal and tumor breast tissue suggests a dual role for the erbb2/her2 gene. Cancer Res, 71(1):78–86, Jan 2011.

[209] C P Petersen, M E Bordeleau, J Pelletier, and P A Sharp. Short rnas repress translation after initiation in mammalian cells. Mol Cell, 21(4):533–542, Feb 2006.

[210] J F Petrosino, S Highlander, R A Luna, R A Gibbs, and J Versalovic. Metagenomic pyrosequencing and microbial identification. Clin Chem, 55(5):856–866, May 2009.

[211] E Pettersson, J Lundeberg, and A Ahmadian. Generations of sequencing technologies. Genomics, 93(2):105–111, Feb 2009.

[212] X Piao, X Zhang, L Wu, and J G Belasco. Ccr4-not deadenylates mrna associated with rna-induced silencing complexes in human cells. Mol Cell Biol, 30(6):1486–1494, Mar 2010.

[213] R S Pillai, S N Bhattacharyya, C G Artus, T Zoller, N Cougot, E Basyuk, E Bertrand, and W Filipowicz. Inhibition of translational initiation by let-7 microrna in human cells. Science, 309(5740):1573–1576, Sep 2005.

[214] K R Pomraning, K M Smith, and M Freitag. Genome-wide high throughput analysis of dna methylation in eukaryotes. Methods, 47(3):142–150, Mar 2009.

[215] M Pop. Genome assembly reborn: recent computational challenges. Brief Bioinform, 10(4):354–366, Jul 2009.

[216] M Pop, D S Kosack, and S L Salzberg. Hierarchical scaffolding with bambus. Genome Res, 14(1):149–159, Jan 2004.

[217] Church GM Porreca GJ, Shendure J. Plony dna sequencing. Curr Protoc Mol Bio, Chapter 7:Unit 7–8, 2006.

[218] Y Qi, X He, X J Wang, O Kohany, J Jurka, and G J Hannon. Distinct catalytic and non-catalytic roles of argonaute4 in rna-directed dna methylation. Nature, 443(7114):1008–1012, Oct 2006.

[219] J G Reid, A K Nagaraja, F C Lynn, R B Drabek, D M Muzny, C A Shaw, M K Weiss, A O Naghavi, M Khan, H Zhu, J Tennakoon, G H Gunaratne, D B Corry, J Miller, M T McManus, M S German, R A Gibbs, M M Matzuk, and P H Gunaratne. Mouse let-7 mirna populations exhibit rna editing that is constrained in the 5'-seed/ cleavage/anchor regions and stabilize predicted mmu-let-7a:mrna duplexes. Genome Res, 18(10):1571–1581, Oct 2008.

[220] J Reinartz, E Bruyns, J Z Lin, T Burcham, S Brenner, B Bowen, M Kramer, and R Woychik. Massively parallel signature sequencing (mpss) as a tool for in-depth quantitative gene

expression profiling in all organisms. Brief Funct Genomic Proteomic, 1(1):95–104, Feb 2002.

[221] J A Reinhardt, D A Baltrus, M T Nishimura, W R Jeck, C D Jones, and J L Dangl. De novo assembly using low-coverage short read sequence data from the rice pathogen pseudomonas syringae pv. oryzae. Genome Res, 19(2):294–305, Feb 2009.

[222] B J Reinhart and D P Bartel. Small rnas correspond to centromere heterochromatic repeats. Science, 297(5588):1831–1831, Sep 2002.

[223] B J Reinhart, E G Weinstein, M W Rhoades, B Bartel, and D P Bartel. Micrornas in plants. Genes Dev, 16(13):1616–1626, Jul 2002.

[224] Kasandra J-L Riley, Gabrielle S Rabinowitz, and Joan A Steitz. Comprehensive analysis of rhesus lymphocryptovirus microrna expression. J Virol, 84(10):5148–5157, May 2010.

[225] A Rodriguez, S Griffiths-Jones, J L Ashurst, and A Bradley. Identification of mammalian microrna host genes and transcription units. Genome Res, 14(10A):1902–1910, Oct 2004.

[226] M Ronaghi, S Karamohamed, B Pettersson, M Uhlén, and P Nyrén. Real-time dna sequencing using detection of pyrophosphate release. Anal Biochem, 242(1):84–89, Nov 1996.

[227] R Ronen, I Gan, S Modai, A Sukacheov, G Dror, E Halperin, and N Shomron. mirnakey: a software for microrna deep sequencing analysis. Bioinformatics, 26(20):2615–2616, Oct 2010.

[228] J G Ruby, C Jan, C Player, M J Axtell, W Lee, C Nusbaum, H Ge, and D P Bartel. Large-scale sequencing reveals 21u-rnas and additional micrornas and endogenous sirnas in c. elegans. Cell, 127(6):1193–1207, Dec 2006.

[229] J G Ruby, C H Jan, and D P Bartel. Intronic microrna precursors that bypass drosha processing. Nature, 448(7149):83–86, Jul 2007.

[230] J G Ruby, A Stark, W K Johnston, M Kellis, D P Bartel, and E C Lai. Evolution, biogenesis, expression, and target predictions of a substantially expanded set of drosophila microrna. Genome Res, 17(12):1850–1864, Dec 2007.

[231] G. Ruvkun. Tiny rna: Where do we come from? what are we? where are we going? Trends in Plant Science, 13:313–316, 2008.

[232] G Ruvkun and J Giusto. The caenorhabditis elegans heterochronic gene lin-14 encodes a nuclear protein that forms a temporal developmental switch. Nature, 338(6213):313–319, Mar 1989.

[233] S Sai Lakshmi and S Agrawal. pirnabank: a web resource on classified and clustered piwi-interacting rnas. Nucleic Acids Res, 36(Database issue):173–177, Jan 2008.

[234] S L Salzberg. Recent advances in rna sequence analysis. F1000 Biol Rep, 2:64–64, 2010.

[235] J R Sanford, X Wang, M Mort, N Vanduyn, D N Cooper, S D Mooney, H J Edenberg, and Y Liu. Splicing factor sfrs1 recognizes a functionally diverse landscape of rna transcripts. Genome Res, 19(3):381–394, Mar 2009.

[236] F Sanger, S Nicklen, and A R Coulson. Dna sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A, 74(12):5463–5467, Dec 1977.

[237] J C Schwamborn, E Berezikov, and J A Knoblich. The trim-nhl protein trim32 activates microrna and prevents self-renewal in mouse neural progenitors. Cell, 136(5):913–925, Mar 2009.

[238] D S Schwarz, G Hutvágner, T Du, Z Xu, N Aronin, and P D Zamore. Asymmetry in the assembly of the rnai enzyme complex. Cell, 115(2):199–208, Oct 2003.

[239] M S Scott, F Avolio, M Ono, A I Lamond, and G J Barton. Human mirna precursors with box h/aca snorna features. PLoS Comput Biol, 5(9), Sep 2009.

[240] K Seggerson, L Tang, and E G Moss. Two genetic circuits repress the caenorhabditis elegans heterochronic gene lin-28 after translation initiation. Dev Biol, 243(2):215–225, Mar 2002.

[241] M Selbach, B Schwanhäusser, N Thierfelder, Z Fang, R Khanin, and N Rajewsky. Widespread changes in protein synthesis induced by micrornas. Nature, 455(7209):58–63, Sep 2008.

[242] A G Seto, R E Kingston, and N C Lau. The coming of age for piwi proteins. Mol Cell, 26(5):603–609, Jun 2007.

[243] J Shendure, G J Porreca, N B Reppas, X Lin, J P McCutcheon, A M Rosenbaum, M D Wang, K Zhang, R D Mitra, and G M Church. Accurate multiplex polony sequencing of an evolved bacterial genome. Science, 309(5741):1728–1732, Sep 2005.

[244] T Shiraki, S Kondo, S Katayama, K Waki, T Kasukawa, H Kawaji, R Kodzius, A Watahiki, M Nakamura, T Arakawa, S Fukuda, D Sasaki, A Podhajska, M Harbers, J Kawai, P Carninci, and Y Hayashizaki. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc Natl Acad Sci U S A, 100(26):15776–15781, Dec 2003.

[245] T Sijen and R H Plasterk. Transposon silencing in the caenorhabditis elegans germ line by natural rnai. Nature, 426(6964):310–314, Nov 2003.

[246] J T Simpson, K Wong, S D Jackman, J E Schein, S J Jones, and I Birol. Abyss: a parallel assembler for short read sequence data. Genome Res, 19(6):1117–1123, Jun 2009.

[247] N R Smalheiser and V I Torvik. Mammalian micrornas derived from genomic repeats. Trends Genet, 21(6):322–326, Jun 2005.

[248] A D Smith, Z Xuan, and M Q Zhang. Using quality scores and longer reads improves accuracy of solexa read mapping. BMC Bioinformatics, 9:128–128, 2008.

[249] T F Smith and M S Waterman. Identification of common molecular subsequences. J Mol Biol, 147(1):195–197, Mar 1981.

[250] M Somel, S Guo, N Fu, Z Yan, H Y Hu, Y Xu, Y Yuan, Z Ning, Y Hu, C Menzel, H Hu, M Lachmann, R Zeng, W Chen, and P Khaitovich. Microrna, mrna, and protein expression link development and aging in human and macaque brain. Genome Res, 20(9):1207–1218, Sep 2010.

[251] J J Song, S K Smith, G J Hannon, and L Joshua-Tor. Crystal structure of argonaute and its implications for risc slicer activity. Science, 305(5689):1434–1437, Sep 2004.

[252] J Starega-Roslan, J Krol, E Koscianska, P Kozlowski, W J Szlachcic, K Sobczak, and W J Krzyzosiak. Structural basis of microrna length variety. Nucleic Acids Res, 39(1):257–268, Jan 2011.

[253] Mitchell S Stark, Sonika Tyagi, Derek J Nancarrow, Glen M Boyle, Anthony L Cook, David C Whiteman, Peter G Parsons, Christopher Schmidt, Richard A Sturm, and Nicholas K Hayward. Characterization of the melanoma mirnaome by deep sequencing. PLoS One, 5(3):e9685, 2010.

[254] J S Steffan, A Kazantsev, O Spasic-Boskovic, M Greenwald, Y Z Zhu, H Gohler, E E Wanker, G P Bates, D E Housman, and L M Thompson. The huntington's disease protein interacts with p53 and creb-binding protein and represses transcription. Proc Natl Acad Sci U S A, 97(12):6763–6768, Jun 2000.

[255] F A Steiner, K L Okihara, S W Hoogstrate, T Sijen, and R F Ketting. Rde-1 slicer activity is required only for passenger-strand cleavage during rnai in caenorhabditis elegans. Nat Struct Mol Biol, 16(2):207–211, Feb 2009.

[256] A Szakmary, D N Cox, Z Wang, and H Lin. Regulatory relationship among piwi, pumilio, and bag-of-marbles in drosophila germline stem cell self-renewal and differentiation. Curr Biol, 15(2):171–178, Jan 2005.

[257] H Tabara, M Sarkissian, W G Kelly, J Fleenor, A Grishok, L Timmons, A Fire, and C C Mello. The rde-1 gene, rna interference, and transposon silencing in c. elegans. Cell, 99(2):123–132, Oct 1999.

[258] R J Taft, E A Glazov, N Cloonan, C Simons, S Stephen, G J Faulkner, T Lassmann, A R Forrest, S M Grimmond, K Schroder, K Irvine, T Arakawa, M Nakamura, A Kubosaki, K Hayashida, C Kawazu, M Murata, H Nishiyori, S Fukuda, J Kawai, C O Daub, D A Hume, H Suzuki, V Orlando, P Carninci, Y Hayashizaki, and J S Mattick. Tiny rnas associated with transcription start sites in animals. Nat Genet, 41(5):572–578, May 2009.

[259] R J Taft, E A Glazov, T Lassmann, Y Hayashizaki, P Carninci, and J S Mattick. Small rnas derived from snornas. RNA, 15(7):1233–1240, Jul 2009.

[260] R J Taft, C D Kaplan, C Simons, and J S Mattick. Evolution, biogenesis and function of promoter-associated rnas. Cell Cycle, 8(15):2332–2338, Aug 2009.

[261] R J Taft, C Simons, S Nahkuri, H Oey, D J Korbie, T R Mercer, J Holst, W Ritchie, J J Wong, J E Rasko, D S Rokhsar, B M Degnan, and J S Mattick. Nuclear-localized tiny rnas are associated with transcription initiation and splice sites in metazoans. Nat Struct Mol Biol, 17(8):1030–1034, Aug 2010.

[262] O H Tam, A A Aravin, P Stein, A Girard, E P Murchison, S Cheloufi, E Hodges, M Anger, R Sachidanandam, R M Schultz, and G J Hannon. Pseudogene-derived small interfering rnas regulate gene expression in mouse oocytes. Nature, 453(7194):534–538, May 2008.

[263] C Trapnell and S L Salzberg. How to map billions of short reads onto genomes. Nat Biotechnol, 27(5):455–457, May 2009.

[264] et al. Vagin, V.V. Large-scale sequencing reveals 21u-rnas and additional micrornas and endogenous sirnas in c. elegans. Science, 313:320–324, 2006.

[265] R P van Rij, M C Saleh, B Berry, C Foo, A Houk, C Antoniewski, and R Andino. The rna silencing endonuclease argonaute 2 mediates specific antiviral immunity in drosophila melanogaster. Genes Dev, 20(21):2985–2995, Nov 2006.

[266] J C van Wolfswinkel, J M Claycomb, P J Batista, C C Mello, E Berezikov, and R F Ketting. Cde-1 affects chromosome segregation through uridylation of csr-1-bound sirnas. Cell, 139(1):135–148, Oct 2009.

[267] V E Velculescu, L Zhang, B Vogelstein, and K W Kinzler. Serial analysis of gene expression. Science, 270(5235):484–487, Oct 1995.

[268] T A Volpe, C Kidner, I M Hall, G Teng, S I Grewal, and R A Martienssen. Regulation of heterochromatic silencing and histone h3 lysine-9 methylation by rnai. Science, 297(5588):1833–1837, Sep 2002.

[269] G. Wang and V. Reinke. A c. elegans piwi, prg-1, regulates 21u-rnas during spermatogenesis. Current Biology, 18:861–867, 2008.

[270] X H Wang, R Aliyari, W X Li, H W Li, K Kim, R Carthew, P Atkinson, and S W Ding. Rna interference directs innate immunity against viruses in adult drosophila. Science, 312(5772):452–454, Apr 2006.

[271] Z Wang, M Gerstein, and M Snyder. Rna-seq: a revolutionary tool for transcriptomics. Nat Rev Genet, 10(1):57–63, Jan 2009.

[272] R L Warren, G G Sutton, S J Jones, and R A Holt. Assembling millions of short dna sequences using ssake. Bioinformatics, 23(4):500–501, Feb 2007.

[273] T Watanabe, Y Totoki, A Toyoda, M Kaneda, S Kuramochi-Miyagawa, Y Obata, H Chiba, Y Kohara, T Kono, T Nakano, M A Surani, Y Sakaki, and H Sasaki. Endogenous sirnas from naturally formed dsrnas regulate transcripts in mouse oocytes. Nature, 453(7194):539–543, May 2008.

[274] M J Weber. Mammalian small nucleolar rnas are mobile genetic elements. PLoS Genet, 2(12), Dec 2006.

[275] D A Wheeler, M Srinivasan, M Egholm, Y Shen, L Chen, A McGuire, W He, Y J Chen, V Makhijani, G T Roth, X Gomes, K Tartaro, F Niazi, C L Turcotte, G P Irzyk, J R Lupski, C Chinault, X Z Song, Y Liu, Y Yuan, L Nazareth, X Qin, D M Muzny, M Margulies, G M Weinstock, R A Gibbs, and J M Rothberg. The complete genome of an individual by massively parallel dna sequencing. Nature, 452(7189):872–876, Apr 2008.

[276] B Wightman, I Ha, and G Ruvkun. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in c. elegans. Cell, 75(5):855–862, Dec 1993.

[277] E Wong and C L Wei. Chip'ing the mammalian genome: technical advances and insights into functional elements. Genome Med, 1(9):89–89, 2009.

[278] L Wu, J Fan, and J G Belasco. Micrornas direct rapid deadenylation of mrna. Proc Natl Acad Sci U S A, 103(11):4034–4039, Mar 2006.

[279] L Wu, Q Zhang, H Zhou, F Ni, X Wu, and Y Qi. Rice microrna effector complexes and targets. Plant Cell, 21(11):3421–3435, Nov 2009.

[280] C Xiao, D P Calado, G Galler, T H Thai, H C Patterson, J Wang, N Rajewsky, T P Bender, and K Rajewsky. Mir-150 controls b cell

differentiation by targeting the transcription factor c-myb. Cell, 131(1):146–159, Oct 2007.

[281] M Xu, D Fujita, and N Hanagata. Perspectives and challenges of emerging single-molecule dna sequencing technologies. Small, 5(23):2638–2649, Dec 2009.

[282] S Yamasaki, P Ivanov, G F Hu, and P Anderson. Angiogenin cleaves trna and promotes stress-induced translational repression. J Cell Biol, 185(1):35–42, Apr 2009.

[283] J H Yang, J H Li, P Shao, H Zhou, Y Q Chen, and L H Qu. starbase: a database for exploring microrna-mrna interaction maps from argonaute clip-seq and degradome-seq data. Nucleic Acids Res, 39(Database issue):202–209, Jan 2011.

[284] J S Yang and E C Lai. Dicer-independent, ago2-mediated microrna biogenesis in vertebrates. Cell Cycle, 9(22):4455–4460, Dec 2010.

[285] Weidong Yang, Thimmaiah P Chendrimada, Qingde Wang, Miyoko Higuchi, Peter H Seeburg, Ramin Shiekhattar, and Kazuko Nishikura. Modulation of microrna processing and expression through rna editing by adar deaminases. Nat Struct Mol Biol, 13(1):13–21, Jan 2006.

[286] S Yekta, I H Shih, and D P Bartel. Microrna-directed cleavage of hoxb8 mrna. Science, 304(5670):594–596, Apr 2004.

[287] R Yi, M N Poy, M Stoffel, and E Fuchs. A skin microrna promotes differentiation by repressing 'stemness'. Nature, 452(7184):225–229, Mar 2008.

[288] Y Zeng and B R Cullen. Efficient processing of primary microrna hairpins by drosha requires flanking nonstructured rna sequences. J Biol Chem, 280(30):27595–27603, Jul 2005.

[289] D R Zerbino and E Birney. Velvet: algorithms for de novo short read assembly using de bruijn graphs. Genome Res, 18(5):821–829, May 2008.

[290] Hua Zhang, Jian-Hua Yang, Yu-Sheng Zheng, Peng Zhang, Xiao Chen, Jun Wu, Ling Xu, Xue-Qun Luo, Zhi-Yong Ke, Hui Zhou, Liang-Hu Qu, and Yue-Qin Chen. Genome-wide analysis of small rna and novel microrna discovery in human acute lymphoblastic leukemia based on extensive sequencing approach. PLoS One, 4(9):e6849, 2009.

[291] R Zhou, B Czech, J Brennecke, R Sachidanandam, J A Wohlschlegel, N Perrimon, and G J Hannon. Processing of drosophila endo-sirnas depends on a specific loquacious isoform. RNA, 15(10):1886–1895, Oct 2009.

[292] E Zhu, F Zhao, G Xu, H Hou, L Zhou, X Li, Z Sun, and J Wu. mirtools: microrna profiling and discovery based on high-throughput sequencing. Nucleic Acids Res, 38(Web Server issue):392–397, Jul 2010.

[293] D G Zisoulis, M T Lovci, M L Wilbert, K R Hutt, T Y Liang, A E Pasquinelli, and G W Yeo. Comprehensive discovery of endogenous argonaute binding sites in caenorhabditis elegans. Nat Struct Mol Biol, 17(2):173–179, Feb 2010.

[294] Chiara Zuccato, Nikolai Belyaev, Paola Conforti, Lezanne Ooi, Marzia Tartari, Evangelia Papadimou, Marcy MacDonald, Elisa Fossale, Scott Zeitlin, Noel Buckley, and Elena Cattaneo. Widespread disruption of repressor element-1 silencing transcription factor/neuron-restrictive silencer factor occupancy at its target genes in huntington's disease. J Neurosci, 27(26):6972–6983, Jun 2007.

[295] Chiara Zuccato, Marzia Tartari, Andrea Crotti, Donato Goffredo, Marta Valenza, Luciano Conti, Tiziana Cataudella, Blair R Leavitt, Michael R Hayden, Tonis Timmusk, Dorotea Rigamonti, and Elena

Cattaneo. Huntingtin interacts with rest/nrsf to modulate the transcription of nrse-controlled neuronal genes. Nat Genet, 35(1):76–83, Sep 2003.