UNIVERSITÀ
DEGLI STUDI
DI PADOVA

UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Biologia

Scuola di Dottorato di Ricerca In Bioscienze E Biotecnologie

Indirizzo Biologia Evoluzionistica

Ciclo XXVII

microRNAs impact in cancer: from non canonical biogenesis and functions to methodological aspects

Direttore della Scuola : Ch.mo Prof. Giuseppe Zanotti

Coordinatore d'indirizzo: Ch.mo Prof. Andrea Pilastro

Supervisore: Ch.mo Prof. Giorgio Casadoro

Co-Supervisore: Dr.ssa Stefania Bortoluzzi

Dottoranda : Claudia Saccoman

Table of contents

Astract

The DNA information appears nowadays more stratified than it was supposed to be a decade before. In this scenario, non coding RNAs have been introduced in the fraction of functional RNA, carrying information and underpinning regulatory circuits of complex genetic phenomena in eukaryotes. microRNAs are endogenous single stranded ~22 nt long transcripts among that unveiled non coding RNA with regulatory functions, detected both in animals and plants. Increasing evidence shows that deregulation of microRNAs (miRNAs) plays an important role in both solid and hematologic malignancies. In this work we considered microRNA non canonical functions and involvement in tumours, integrating computational analyses of genome-wide datasets and targeted experimental results, with a critical approach to the specific adopted computational tools.

First, we studied miRNAs role in myeloproliferative disorders, more specifically in primary myelofibrosis, considering also other small RNAs detected in RNA-seq data. Myeloproliferative neoplasms are chronic myeloid cancers involving CD34+ hematopoietic stem cells alterations, evolving to acute leukemia in the most severe forms. This study deals indeed with an Illumina sequencing of small RNAs samples of CD34+ hematopoietic stem cells of patients affected by primary myelofibrosis and of controls, in order to characterize miRNAs profile and find relevant differentially expressed elements, as putative effectors of a disrupted post-transcriptional regulation involved in PMF initiation and progression.

Then, in order to have a better understanding of each step of a computational analysis of RNA-seq data, we studied the impact on small RNAs differential expression analysis of normalization methods developed for long RNA. We evaluated five commonly used normalization methods to pinpoint a procedure to perform a robust RNA-seq analysis. We estimated statistical distribution parameters from a real microRNA numerous dataset and we simulated a huge number of small RNAs dataset. We controlled datasets characteristics in order to generate 9 different testing scenarios and measure the normalization impact on differentially expressed elements recognition, through ROC and AUC curves. We ascertain that normalization methods still need strong efforts in developing new algorithms in order to fill the wide room for improvement.

Thereafter we evaluated the implication of microRNAs in the gene expression changes observed after H-ferritin silencing. We explored whether different FHC amounts might modulate miRNA expression levels in K562 cells and we studied the impact of miRNAs in gene expression profile modifications. To this aim, we performed a miRNA-mRNA integrative analysis in K562 silenced for FHC ($K562^{shFHC}$) comparing it with K562 transduced with scrambled RNA ($K562^{shRNA}$). The remarkable up-regulation of four miRNAs, hsa-let-7g-5p, hsa-let-7f-5p, hsa-let-7i-5p and hsa-miR-125b-5p, in silenced cells and their down-regulation when FHC expression was rescued supported a specific relation between FHC silencing and miRNA-modulation. The integration of target predictions with miRNA and gene expression profiles led to the identification of a regulatory network. Our data, confirmed by an experimental validation, indicate that, FHC silencing may affect RAF1/pERK1/2 levels through the modulation of a specific set of miRNAs and they add new insights to the relationship among iron homeostasis and miRNAs.

We further explored a putative non canonical role of microRNAs, more specifically, in the context of the always more evident complex cross talk between protein-coding and non-protein coding RNAs. We worked on a preliminary study that deals with the involvement of microRNAs in the regulation of alternative translation (AT) and thus of protein isoform equilibrium. There is an increasing appreciation of the high prevalence of alternative translation in mammals. Complex and regulated translation pattern are achieved thanks to multiple Open Reading Frame (ORFs) and Translation Initiation Sites (TISs) in the same mRNA that can influence each other in different ways. miRNAs were recently demonstrated to be involved in modulation of protein isoform equilibrium binding to TISs. We provided novel data on the overlap of active TISs of mRNAs, experimentally defined using GTI-seq, to miRNA-binding sites, experimentally determined using CLASH technique. The genes whose sites were recognized are supposed to be involved in miRNA-modulated AT and we modelled the interaction mechanism. The miRNA-based regulation of mRNA alternative translation surely deserves further investigation to clarify if and how it impacts on cell processes and on disease.

Sommario

L'informazione contenuta nel DNA appare oggi sempre più stratificata di quanto non si pensasse. In questo scenario, gli RNA non codificanti sono stati riconosciuti come RNA funzionali, portatori di informazione e parti fondamentali dei più complessi circuiti regolativi negli eucarioti. Tra i più studiati RNA non codificanti con funzioni regolative ci sono i microRNA (miRNA), RNA a singolo filamento lunghi circa 22 nucleotidi, presenti sia in piante che animali. Ci sono prove sempre più evidenti che la deregolazione dei miRNA abbia un ruolo fondamentale nei tumori solidi e del sangue. In questo lavoro abbiamo preso in considerazione le funzioni non canoniche dei miRNA, il loro coinvolgimento nei tumori, integrando analisi computazionali di dati genome-wide e dati sperimentali più specifici, con un approccio critico rispetto gli strumenti computazionali.

Abbiamo innanzitutto studiato il ruolo dei miRNA nelle neoplasie mieloproliferative, più specificamente nella mielofibrosi, considerando anche altri small RNA presenti nei dati RNA-seq. Le malattie mieloproliferative sono tumori cronici della linea mieloide che vedono l'alterazione delle cellule emopoietiche CD34+, ed evolvono in leucemia acuta nei casi più gravi. In questo studio abbiamo pertanto analizzato dati di RNA-seq, prodotti con tecnologia Illumina, di cellule raccolte da pazienti affetti da mielofibrosi primaria e da controlli sani, al fine di caratterizzare i profili di microRNA e trovare gli elementi differenzialmente espressi, in quanto possibili elementi di regolazione post trascrizionale alterata e coinvolti nella genesi e nello sviluppo della mielofibrosi.

Successivamente, al fine di aver piena consapevolezza dei vari passi di un'analisi computazionale, abbiamo studiato l'impatto dell'applicazione su dati di RNA corti di algoritmi di normalizzazione, sviluppati per RNA lunghi, valutato a livello dei risultati dell'analisi differenziale. Abbiamo preso in considerazione cinque tra i più comunemente usati algoritmi, per individuare la procedura che permetta di svolgere in modo più robusto l'analisi di dati RNA-seq. Abbiamo stimato i parametri della distribuzione statistica di un dataset reale di microRNA particolarmente numeroso, e abbiamo simulato un numero sostanzioso di dataset. Abbiamo generato nove tipi di data set con diverse caratteristiche controllate e abbiamo misurato l'impatto della normalizzazione nei vari casi, quantificando l'impatto sull'analisi differenziale attraverso curve ROC e AUC. Abbiamo evidenziato la necessità di nuovi algoritmi di

normalizzazione, più specifici per i miRNA, in grado di colmare le grosse lacune dei metodi attuali.

Ci siamo in seguito concentrati sul coinvolgimento dei microRNA nei cambiamenti dei valori di espressione genica, rilevati in cellule K562 in cui fosse silenziata la ferritina FHC. Abbiamo indagato se diversi livelli di FHC potessero modulare i livelli di espressione dei microRNA e abbiamo monitorato l'impatto dei miRNAs rispetto le modificazioni dei livelli d'espressione dei geni. A tal fine abbiamo, condotto un'analisi integrata di miRNA-mRNA in cellule K562 silenziate per la FHC ($K562^{shFHC}$) confrontandole con cellule K562 trasdotte con RNA scrambled ($K562^{shRNA}$). La notevole up-regolazione di quattro miRNA, hsa-let-7g-5p, hsa-let-7f-5p, hsa-let-7i-5p e hsa-miR-125b-5p, nelle cellule silenziate e il fatto che i loro livelli di espressione scendessero quando fosse riattivata l'espressione di FHC, supporta l'esistenza di una relazione tra FHC e la modulazione dei miRNA. Integrando le informazioni sui target dei miRNA e i profili di espressione dei geni, abbiamo identificato dei network regolativi. I nostri dati, confermati con validazioni sperimentali, indicano che il silenziamento di FHC potrebbe impattare sui livelli di RAF1/pERK1/2 attraverso la modulazione di specifici gruppi di microRNA, fornendo nuove informazioni sul rapporto tra omeostasi del ferro e miRNA.

Infine, ci siamo occupati di un ruolo non canonico dei microRNA, più specificamente nel contesto delle sempre più evidenti interazioni tra RNA codificanti e RNA non codificanti. Abbiamo condotto uno studio preliminare sul coinvolgimento dei microRNA nella regolazione della traduzione alternativa e di conseguenza dell'equilibrio delle varie isoforme proteiche. C'è una maggior consapevolezza della diffusione del meccanismo della traduzione alternativa nei mammiferi. Si realizzano pattern complessi di regolazione delle isoforme grazie alla presenza, nello stesso mRNA, di più Open Reading Frame (ORF) e Translation Initiation Sites (TISs) utilizzati. Questi sono in grado di influenzarsi a vicenda in maniera diversa. E' stato recentemente dimostrato che i miRNA sono coinvolti nella modulazione dell'equilibrio delle isoforme proteiche, legandosi ai TIS. Noi abbiamo individuato la corrispondenza di siti TIS attivi nei trascritti di mRNA, trovati sperimentalmente con GTI-seq, e siti di legame di miRNA nelle sequenze di mRNA, determinati sperimentalmente con tecnica CLASH. Questi geni in cui sono stati riconosciuti siti di legame, si suppongono coinvolti in un meccanismo di traduzione alternativa modulata da miRNAs. Alcune interazioni miRNA-tis sono state confermate sperimentalmente,

ma ulteriori studi sono necessari per valutare se il meccanismo di modulazione della traduzione alternativa da parte dei miRNA possa impattare su processi cellulari e nella malattia.

Introduction

The present work aims at exploring multiple aspects of the microRNAs world prevalently with a bioinformatics approach.

We are having an overview on microRNA biogenesis, their biological role and mechanism of action, computational tool of analysis. Finally I'll give you a hint about the four main themes we focused on in this thesis.

The non coding small RNA: a focus on microRNA

RNA-seq technologies offer the possibility to investigate in a wide range of biological application of interrogating the transcriptome[1-3]. Many works and ambitious projects arose. As such, the Encyclopedia of the DNA Elements (ENCODE)[4, 5]which combined efforts aim to characterize RNAs across 15 different human cell lines. They provide new insights into the mechanisms of gene regulation[6]. Many newly identified elements were discovered unveiling the pervasiveness of transcription and many noncoding elements were found to control regulatory networks[7, 8]. Among the non coding elements they studied:

- Micro-RNAs (miRNAs), which are short RNA fragments known to have a role in post-transcription regulation
- Transfer RNAs (tRNAs), the adapter molecules between mRNA and amino acids.
- Small nuclear RNAs (snRNAs) associated with the spliceosome
- Small nucleolar RNAs (snoRNAs), which guide chemical modifications (methylation and pseudouridylation) of ribosomal and transfer RNAs as well as snRNAs.

The DNA information appears nowadays more complex than it was supposed to be a decade before. The commonly accepted modular structure of the DNA has been called into question: the same genetic sequence does not correspond to a single regulatory function or transcript but it is commonly believed that multiple layers of information

are embedded in every sequence[9]. The genome has been compared to a palimpsest by Tuck and Tollervey[9], as it resembles an overwritten text. In figure 1 you can see their interpretation of that concept.

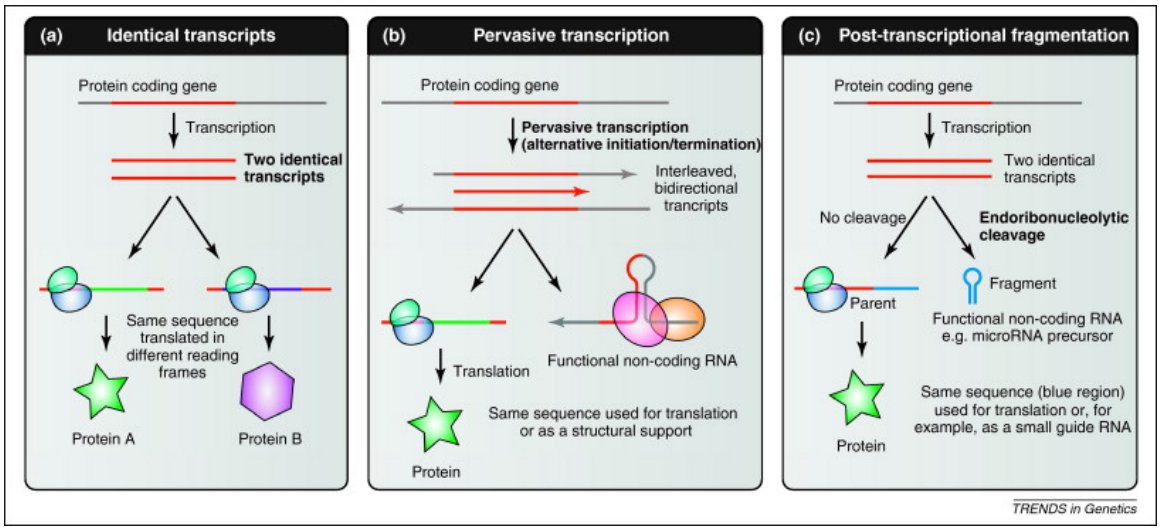

Figure 1: two transcripts from the same locus can use the same sequence to different functional effects. The overlapping arrangement of genetic information enables a single sequence to encode multiple functions. This principle is embodied at many genomic loci, which generate ensembles of transcripts with shared sequences but disparate functions. This raises questions about how a specific function is assigned to a transcript, given the numerous possibilities. There are several explanations, illustrated by the various ways in which overlapping transcripts are generated: **(a)** Two transcripts identical in sequence and length might function differently, perhaps being translated in alternative reading frames (green or purple) to generate distinct proteins. Here, extrinsic factors are responsible for specifying which reading frame should be used. **(b)** Alternative transcription initiation and/or termination generate an ensemble of interleaved transcripts from a single genomic locus. Within this ensemble, a shared sequence (red) can perform distinct functions, perhaps contributing to an open reading frame (green) in one transcript and a structural feature in another. Here, the function of a sequence is governed by its context, with the different lengths and orientations of transcripts perhaps affecting their folding or recruitment of binding factors. **(c)** Many classes of transcript might act as precursors to shorter fragments, excised by post-transcriptional cleavage. These fragments might function in ways distinct from those of their parents. Thus, within the context of the shorter fragment, a shared sequence (blue) can perform an alternative role. This indicates that the length of a transcript might contribute to specifying which of several possible functions is performed by a particular sequence. Other post-transcriptional processes (such as splicing) can also generate alternative transcripts, but are beyond the scope of this review. Figure adapted from Tuck at al[9].

In this complex scenario, non coding RNA must be introduced in the fraction of functional RNA, carrying information and underpinning regulatory circuits of complicated genetic phenomena in eukaryotes[10, 11]. microRNAs are endogenous single stranded ~22 nt long among that unveiled non coding RNA with regulatory functions, detected both in animals and plants[12]. They were first revealed during a characterization of genes that control the timing of larval development in worm *Caernorhabditis elegans*. This was the case of lin-4 and let-7 [13, 14]. Soon in other bilateral animals including mammals, homologs of let-7 were identified. Their

expression followed that observed in *C. elegans* as if let-7 might be playing orthologous roles in different metazoan lineages[15]. Several thousand of other small RNAs were later recognized in worms, flies, plants, green algae, viruses and mammals and they were called microRNA[16-18]. They are evolutionary conserved elements[19], for example miRNA regulatory system in the floral developmental phase has a well conserved patterns for each step of the pathway, suggesting they play important roles in the evolution of flower[20].

Salmena et al. [21] proposed a theory that hypothesizes that key elements of multiple RNAs communication are microRNAs. All the types of RNA transcripts are supposed to exchange information mediated by microRNA-binding sites called "microRNA response elements" (MREs). They based their hypothesis on theoretical and experimental studies in which "RNAs influences each others' level by competiting for a limited pool of microRNAs", that's why they called their theory "competitive endogenous RNA" (ceRNA). The actual regulational machinery mechanism is still debated but al lot of breakthroughs have been made.

miRNAs biogenesis

miRNAs biogenesis starts from RNA polymerase II transcription of several bases long primary transcript called pri-miRNAs. These transcripts fold back themselves to form hairpin structures[12]. The pri-miRNA is then cleaved in the nucleus by the Drosha RNase III endonuclease[22] liberating a ~70 nt long stem loop intermediate, known as pre-miRNA[23, 24]. Secondarily, the pre-miRNA is actively exported to the cytoplasm by Ran-GTP and by the export receptor Exportin-5 [25, 26], where the RNase III endonuclease Dicer lops off the terminal base pairs and the loop of the hairpin precursor. This process release a ~22 nt miRNA duplex[22]. Depending on thermodynamic properties of the duplex, one of the two strands is incorporated into the Argonaute (Ago) protein[27], a component of the ribonucleoprotein complex called RNA-induced silencing complex (RISC). This is the effector complex and it guides the miRNA incorporated to the targets. The other strand that is not incorporated, appears typically to be degraded. It is not clear which are the rules governing strand selection. It has been reported that both miRNA strands are functional[28] and can be both accumulated in specific cell tissues and types.

miRNAs can be transcribed as long polycistronic primary transcripts and the first experimental confirmation was found in total HeLa cells, with two miRNAs clusters transcribed from a transcription unit (TU) [23]. They can also be encoded outside a cluster, as monocistronic transcripts with their own promoters[29, 30]. miRNA genes structure definition is ongoing and was found in different genomic locations: both intronic and exonic. Also a special class of miRNAs has been reported, called mirtrons, whose exact sequence of the pre-miRNA[31] is an intron. Figure 2 from Olena and Patton[30] resumes schematically a classification of miRNAs depending on their genomic location.

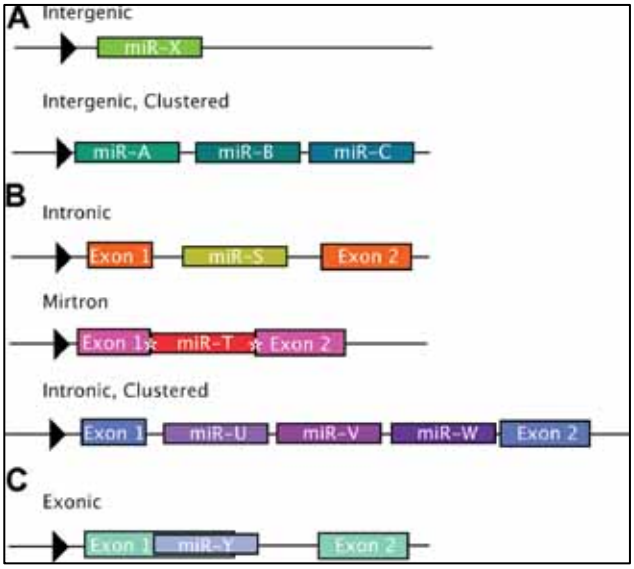

Figure 2. Genomic location of miRNAs. A: Intergenic miRNAs are found in genomic regions distinct from known transcription units. These miRNAs can be monocistronic (top part) with their own promoters (black arrowhead), or polycistronic, where several miRNAs are transcribed as cluster of primary transcripts (bottom part) with a shared promoter (black arrowhead). B: Intronic miRNAs are found in the introns of annotated genes, both protein coding and noncoding. These miRNAs can be present as a single miRNA (top part) or as a cluster of several miRNAs (bottom part). Intronic miRNAs are thought to be transcribed from the same promoter as their host genes (black arrowhead, all parts) and processed from the introns of host gene transcripts. In the special case of mirtrons (middle part), the intron is the exact sequence of the pre-miRNA with splice sites on either side (denoted by white asterisks). In this case, the Microprocessor complex is thought to be unnecessary in mirtron maturation (Okamura et al., 2007). C: Exonic miRNAs are far more rare than either of the types above and often overlap an exon and an intron of a noncoding gene. These miRNAs are also thought to be transcribed by their host gene promoter and their maturation often excludes host gene function. Figure adapted from Olena and Patton[30].

Evolutionary conservation of miRNAs

All miRNAs are characterized by a common biosynthetic pathway and reaction mechanism. Many miRNAs sequences are conserved, in their mature form, among different organisms. Moreover, the evolutionary appearance of multicellular organisms has the same trend of the appearance of the miRNA pathway for regulating gene expression. Some miRNA pathways are conserved virtually intact throughout phylogeny while miRNA diversity also correlates with speciation. The bigger is animal morphological complexity and the growing is the number of miRNA genes, expression of miRNAs and diversities of miRNAs targets, corroborating to the idea that organismal complexity can be estimated by the complexity of the miRNA circuitry. The complexity of the miRNA gene families establishes a link between genotypic complexity and phenotypic complexity in animal evolution[32]. Many works in the area of miRNA phylogenetic conservation and diversity suggests that miRNAs play important roles in animal evolution, by driving phenotypic variation during development[19]. A well-recognized example of evolutionally conserved miRNA is *let-7*. It displays a temporal expression pattern during many organisms development. It belongs to a larger gene family that has been amazingly conserved across almost all groups of bilaterally symmetrical animals, highly conserving its temporal expression pattern, and this is represented in Figure 3. Curiously, *let-7* RNA is not found in more basal metazoans, including non bilaterians suggesting that acquisition of the *let-7* gene was an essential step of evolution from lower metazoan to a higher bilaterians[19].

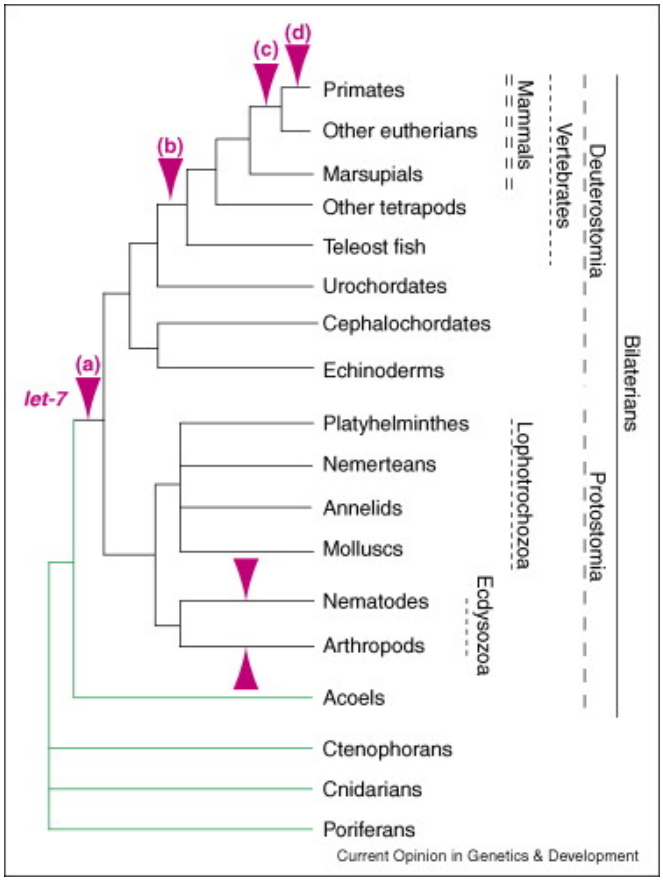

Figure 3. Acquisitions of miRNAs in metazoan phylogeny. An abbreviated phylogeny is modified from earlier studies. Green lines represent animal phyla that do not have let-7 miRNAs. Major innovations of miRNA repertoires are represented by magenta arrowheads, which are based on previous data. A class of approximately 20 miRNAs, including let-7, is common to all bilaterians except acoels (a). The next innovation (b) is the addition of about 56 new families of miRNAs at the branch leading to the vertebrates. A third miRNA innovation (c) occurred at the branch leading to the placental (eutherian) mammals, at which time about 40 new miRNA families were acquired. Furthermore, a large number of primate-specific miRNAs are also identified (d). Each of the nematode and arthropod lineages might also have evolved unique miRNA families. Figure adapted from Niwa et al[19]

Many miRNAs are highly conserved throughout the animal kingdom, as for example mir-1 that is detected from *C. elegans* to human. It operates during muscle development and is essential in maintaining muscle fibers integrity. Moreover, a recent comprehensive study of microRNA gene expression in zebrafish, lists 142 miRNA loci in the genome of *Danio rerio* that are homologous to more than 100 different mammalian microRNAs, belonging to almost 100 different families[33].

It is fascinating how episodes of miRNAs innovation correlate with major introductions of developmental complexity during evolution. That observation suggests that a dramatic expansion of the non coding repertoire, among all miRNAs, could represent the mechanism originating the complexity in higher order organisms, rather than an increase in protein coding inventory. Supporting that hypothesis,

diverse comparative genomic study concluded that both flies and vertebrates witness a growth in their respective number of cell types over geologic time, coherently to the gaining of their respective number of miRNAs[34, 35]. It is reasonable to claim that miRNAs are active driver of animal evolution over the course of animal phylogeny towards gene network complexity, through their regulation of an increasing number of targets. It is noteworthy that both DNA and RNA viruses exploit miRNA potential as key regulatory elements in the control of gene expression. Viruses indeed have evolved mechanisms to degrade, boost, or hijack cellular miRNAs to benefit the viral life cycle[36].

Non canonical miRNAs biogenesis

Canonical biogenesis is not the only mechanism miRNAs are produced with. In the last decade, more evidences emerged of alternative biogenesis mechanisms either Drosha-independent or Dicer-independent, adding complexity to miRNAs regulatory network37, 38. This versatility in miRNA biogenesis may reflect a fine tuned regulation mechanism of specific miRNAs expression in different developmental stages or altered cell states, in order to achieve differential gene regulation.

Drosha cleavage is bypassed both during mirtron production, when small RNAs are generated through mRNA splicing, lariat debranching and folding as pre-miRNAs, and in rare exceptions where small RNAs derive from endogenous short hairpin RNA transcription, as the case of the 7-methylguanosine (m^7G)-capped pre-mir-320.

Another example of bypassing Drosha processing is when miRNAs are produced from other non-coding RNAs. This is the case of tRNA or tRNA like that can be miRNA precursors. Mature miRNAs derived are demonstrated to be functional in modulating proliferation and DNA damage response[39, 40]. Processing of small nucleolar RNAs (snoRNAs), as for ACA45[41], and small nuclear RNA-like viral RNAs[42] can be an additional biogenetic pathway.

Moreover, miRNA biogenesis goes through non canonical miRNA production in presence of unusual pri-miRNA structure. If the miRNA precursor has a shorter 3' overhang needs an additional processing step to generate the mature miRNA: it has to be monouridylated by the uridylyl transferase for efficient Dicer processing. miRNAs characterized by the monouridylation of terminal 3' are called uridylyl transferase (TUTase)-dependent group.

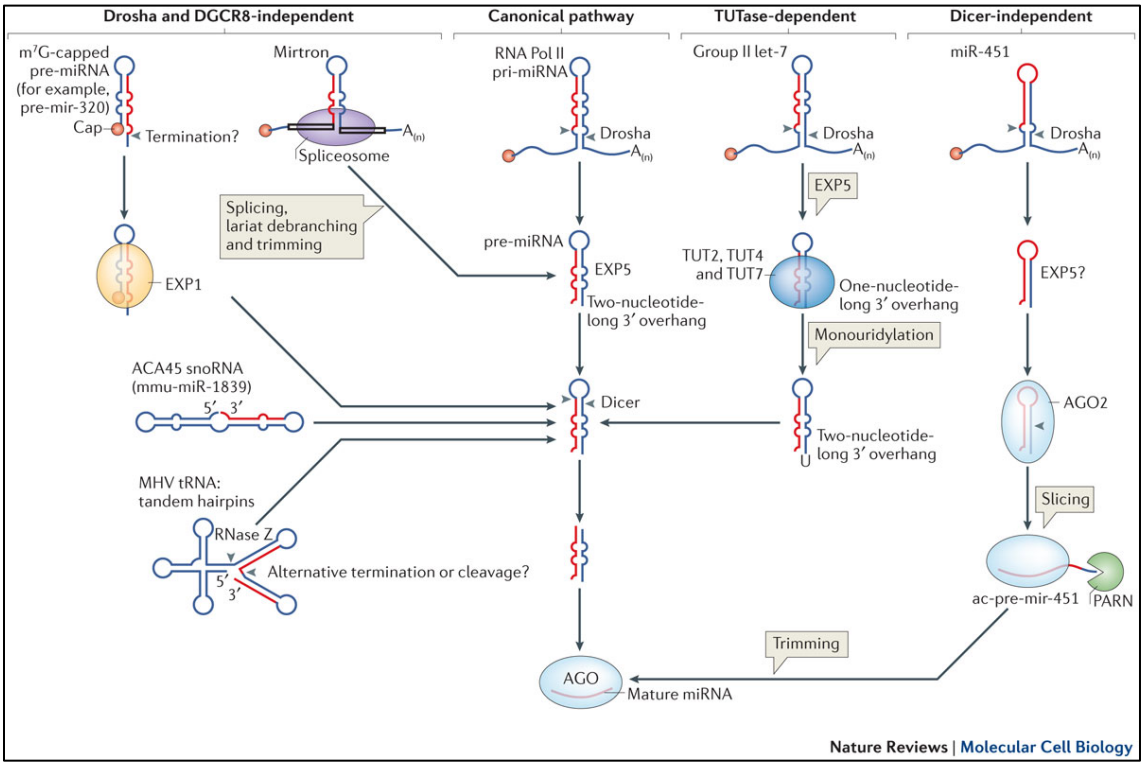

Figure 4: The 7-methylguanosine (m7G)-capped pre-mir-320 is directly generated through transcription, bypassing Drosha processing, and it is exported to the cytoplasm by exportin 1 (EXP1). Mirtron loci produce pre-miRNAs directly through splicing and debranching. Some mirtrons contain 5′ or 3′ single-stranded RNA tails that need to be trimmed before Dicer processing. Some small nucleolar RNAs (snoRNAs), such as ACA45, and tRNAs (or tRNA-like RNAs) may also be cleaved to produce pre-miRNAs. Terminal uridylyl transferase (TUTase)-dependent group II pri-miRNAs produce pre-miRNAs with a shorter 3′ overhang that is suboptimal for Dicer processing. This means that they need to be monouridylated for efficient Dicer processing. In a Dicer-independent pathway, a short pre-mir-451 is produced by Drosha, exported to the cytoplasm (possibly by EXP5) and loaded on Argonaute 2 (AGO2) without Dicer processing. AGO2 cleaves ('slices') the stem of pre-mir-451, generating AGO-cleaved pre-mir-451 (ac-pre-mir-451), which is further trimmed by the 3′–5′ exonuclease poly(A)-specific ribonuclease PARN. The question marks indicate places in which the depicted action has not be fully confirmed. MHV, murine γ-herpesvirus; mmu, Mus musculus; Pol II, polymerase II. Figure adapted from Ha et al37

Alternative miRNA biogeneses could be Dicer-independent pathways, as the case of a short pre-mir-451 produced by Drosha, exported to the cytoplasm and loaded on Argonaute 2. Dicer processing is substituted by a trimming performed by AGO2 and the 3′–5′ exonuclease poly(A)-specific ribonuclease PARN. miR-451 is a erythropoietic miRNA conserved in vertebrates, in confirmation of the importance of its functional role. Figure 4 shows all the different types of alternative miRNA biogeneses.

IsomiRNAs

miRNAs are annotated as a single defined sequences but are actually mixture of sequences, slightly different from the official mature miRNA. As the case of the most part of miRNAs, several length or sequence variants have been detected. These variants are called isomiRNAs[43]. Sequencing-based technology revealed a bona fide repertoire of expressed small RNAs, originally dismissed as sequencing/alignment artifacts or poor quality RNAs. Nowadays isomiRs are confidently detected and they contribute to miRNA expression and qualitative characteristics. They were experimentally validated and characterized as tissues, conditions and cell types specific[44-46]. They were demonstrated to vary in response to stimuli, suggesting their biogenesis to be dynamic and regulated. IsomiRs are categorized into three main classes:

- 5'isomiRs, with variations at the 5' end;
- 3'isomiRs, differing from the canonical miRNA in the 3' end;
- Polymorphic isomiRs, harboring distinct nucleotides composition.

The biological mechanism underlying isomiRs production has not been fully elucidated but several lines of evidence suggest that isomiRs could be processed by variations in Drosha/Dicer cleavage of the pre-miRNA[47-50]. Others pri/pre-miRNA enzymes processing activities could be source of template miRNA variations. Exoribonucleases catalyzed nucleotides trimming, nucleotidyl transferases catalyzed nucleotides addition, RNA editing are considered wellspring of variations[43, 51]. Interestingly, a very low frequency of single nucleotide polymorphisms (SNP) has been identified in genomic miRNA regions[52, 53].

A growing number of reports suggest that isomiRs are biologically functional and to act as canonical miRNAs[44, 46]. Also the fact that 5' isomiRs are also under selection during evolution witnesses their functional importance[46].

For example, has-miR-101 has many different 5'-isomiR-101 ubiquitously detected and highly abundant that interact with RISC complexes, silencing their target[54]. Cloonan et al.[55] biotin-labeled miRNAs and isomiRs to pull down endogenous mRNA targets, detecting highly expressed isomiRs incorporated into the RISC complex and targeting endogenous mRNAs.

Considering that different isomiRs can characterize different tissues or condition, as tumor respect to normal tissues, and that diverse isomiRs could impact differently on target genes and pathways, isomiRs deregulated expression could be implicated in diseases.

moRNAs

High-throughput sequencing revealed overlapping reads aligning to hairpins outside miRNA loci. That small RNAs were called microRNA-offset RNAs[56-58].

MoRNAs were first reported in a simple chordate, the ascidian *Ciona intestinalis*, as ~20-nt-long RNAs derived from the ends of pre-miRNAs. moRNAs were considered as by-products of potentially atypical miRNA processing, possibly generated by RNAse III-like processing[59]. In *C intestinalis* moRNAs displayed a developmental regulated expression. More evidence of moRNAs pervasiveness arrived with Langenberger study[58]. It reported the presence of 78 members of this new class of small RNAs, in short RNA sequencing of human prefrontal cortex. For 71 of the 78 loci, the moRNAs were well-conserved, together with miRNA processed from the same hairpin. The 78 moRNAs loci belonged to only 54 distinct families. Interestingly, studying the families, Langenberger noted that four families showed moRNAs in three or more paralogs, and seven families had two paralogs with evidence for moRNA expression. He inferred association of moRNAs with an early evolutionary origin, as almost all miRNA families with multiple paralogs are evolutionarily old.

moRNAs sequences partially overlap miRNA regions but generally span the Drosha cutting sites, letting us hypothesize a non canonical processing of the hairpin precursor in moRNA biogenesis[60]. However the origin of moRNA is still unclear and many hypothesis were generated. They might arise from exonuclease activity on their precursor [39, 61]or from alternative Drosha processing[59, 60, 62]. For sure moRNAs are detected in different tissues: Taft et al reported moRNAs enriched expression in the nucleus in the human leukemia cell line THP-1[63], Meiri et al detected moRNAs expression in solid tumours[64]while Bortoluzzi and Bisognin in JAK2V617F-mutated SET2 cells[57]. Unfortunately, information about moRNA functional role is still fragmentary. They may act as miRNAs but no experimental evidence has been

reported. Their need to be further characterized and nowadays their mechanism of action remains to be elucidated (Asikainen, personal communication).

miRNA regulatory function and target prediction tools

When incorporated in the RISC complex, miRNAs are known to target mRNAs by imperfect base pairing to the 3'UTR mRNA region. They act to downregulate gene expression by either two posttranscriptional mechanism: by irreversibly triggering mRNA degradation[65-67] or by their translational repression[12]. Recent works show that mRNA destabilization explains most (66%–>90%) miRNA-mediated repression[65].

Plant miRNA target site are located within target genes open reading frames (ORF) and miRNAs bind to mRNA by perfect sequence complementarity[68].

It's more complicated for animals. miRNAs bind to mRNAs by imperfect Watson-Crick pairing of miRNA nucleotides 2-8, called "seed region", to a short region of mRNA termed miRNA recognition elements (MREs), generally situated at 3' untraslated region (UTR) but also sporadically in the 5'UTR or ORF[69]. The seed region of miRNAs is the most conserved in miRNA sequences[70]. Not even the position of mRNA MRE is easily identifiable along the mRNA sequence and it's not unique for each mRNA: many mRNAs has potential multiple sites for the same miRNA[71-73], and it has been reported that multiple sites enhance the degree of downregulation[74]. Target sites can be classified in three main groups: 1) canonical, 2) 3'-supplemetary, 3) 3'-compensatory sites. Among the canonical, three main types can be distinguished:

- the 7mer1A that has an adenine in position 1 at the 5' end of miRNA;
- the 8mer having matched adenine in position 1 and an additional match in position 8;
- the 7mer-m8 that has a match in position 8;

A minor class of canonical sites is represented by 6-nt seed which has a limited impact in downregulating targeted mRNA.

In all the previous cases, there can be an additional binding site at the 3', the so-called 3'-supplementary site. It usually has a weaker effect on target recognition and a lower

efficiency. 3' compensatory sites consist in a binding site that has a mismatch in the seed but an additional extended pairing at the 3' of the miRNA that compensates it.

Figure 5 from Witkos et al[75] shows the just described different miRNA-mRNA representative interactions.

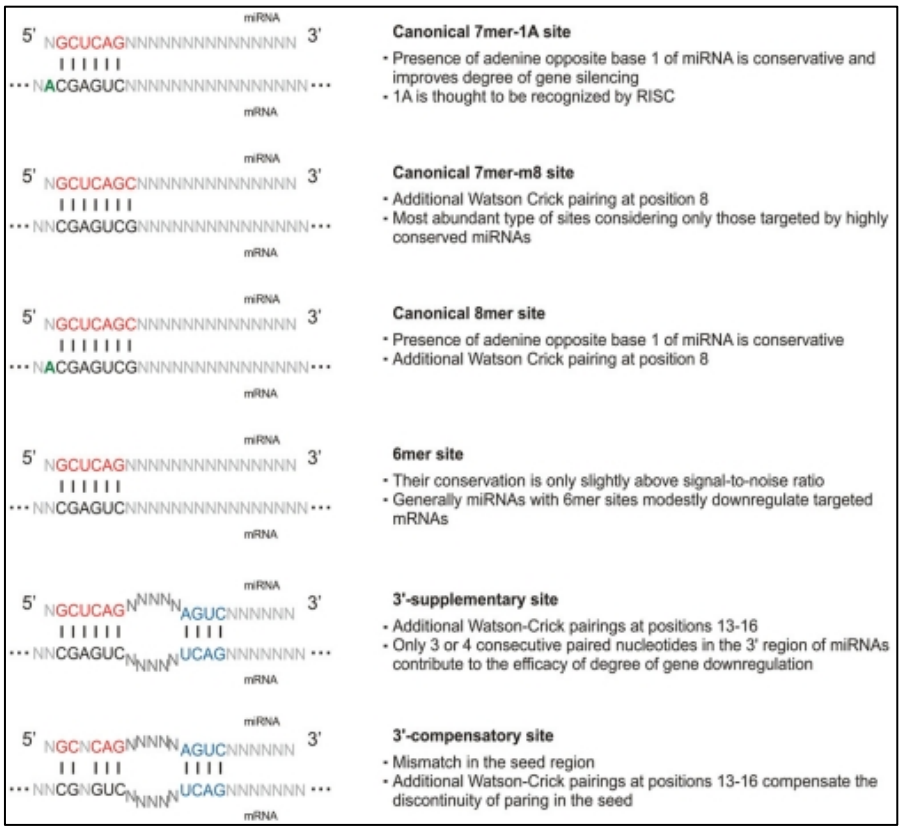

Figure 5. Different classes of miRNA target sites are presented in a schematic way. Vertical dashes represent single Watson-Crick pairing. Nucleotides involved in binding have been arbitrarily defined to depict positions of required complementarity between miRNA and mRNA. Seed regions of miRNAs are marked by red color and the adenine at binding position 1 by green. Interactions between mRNA and the 3' end of miRNA have not been shown because they are sequence-dependent and do not significantly contribute to the miRNA downregulation effect. In the case of 3'-suppelmentary and 3'-compensatory sites two regions of pairing (base pairs colored in blue) force middle mismatches to form a loop structure. Additionally, features of particular site types have been listed. Figure adapted from Witkos et al.[75]

The mechanism governing miRNA targeting is poorly characterized[76, 77].

It is hard to predict which mRNA will be bound by a miRNA also because of the numerous predicted MREs sequences that match the complementarity of the short miRNA seed region. It is also complicated to predict the effect of miRNA binding[78], considering that different miRNAs could act cooperatively to the same target, increasing the sensitivity of repression and enhancing the regulatory effect[71]. miRNAs themselves are recently supposed to be regulated by target interactions that do not

necessarily affect miRNA levels[79], entailing that miRNA expression level could not be representative of its impact. The complexity of miRNA-mRNA interaction causes ambiguity in target prediction results and a high rate of false positive predicted target. Target identification is challenging and more rules than sequence complementarity need to be taken into account[71]. Many algorithms have been developed and available tools for miRNA target prediction encompass a range of different computational approaches, from the modeling of physical interactions to the incorporation of machine learning. Well-known methods are miRanda[80], DIANA-microT [81], RNAhybrid[82], MicroInspector[83], Target Scan and TargetScans[84, 85], PicTar[86], MicroTar[87], PITA[88], RNA22[89] .Target prediction algorithms have been extensively discussed[76, 90-92] but the debate always ended wishing for powerful predictive algorithm. Each method takes into account diverse weighted features to compute a score. That score is used to rank predictions and call putative mRNA targets. They differently use parameters as free energy of binding, folding energy, the presence of multiple binding sites within 3'UTR, secondary structure of 3'UTR that influence accessibility, local AU content, pattern recognition, target site accessibility energy. Especially, target prediction programs can be divided in two classes, distinguished on the basis of the use or not of the information about evolutionary conservation of interactions[75]. miRanda, DIANA-microT, RNAhybrid, MicroInspector, PicTar, TargetScan, are all based on conservation criteria, while PITA, MicroTar and RNA22 are not.

miRNAs can be grouped depending on their seed sequence, that are supposed to target the same mRNAs. miRNAs with same seed sequence are gathered together in the same miRNA family. It has been reported that miRNA families are well conserved among related species and have the same target. Friedman et al[93] gave a comprehensive overview of miRNA target conservation: "In total, >45,000 miRNA target sites within human 3′UTRs are conserved above background levels, and >60% of human protein-coding genes have been under selective pressure to maintain pairing to miRNAs. Mammalian-specific miRNAs have far fewer conserved targets than do the more broadly conserved miRNAs, even when considering only more recently emerged targets. Although pairing to the 3′ end of miRNAs can compensate for seed mismatches, this class of sites constitutes less than 2% of all preferentially conserved sites detected".

Alternative miRNA localizations and functions

miRNAs are known endogenous posttranscriptional downregulators of gene expression by either irreversibly triggering mRNA degradation or inducing translational repression. Proper miRNAs' functioning requires their assembly into an RNA-induced silencing complex (RISC), a multiprotein complex.

There are increasing evidences that miRNAs can act in a non-canonical way to modulate gene expression or to perform different biological effect. This is the case of microRNAs found in bloodstream and body fluids, associated with extracellular vesicles (EVs), which are small membrane vesicles secreted from various types of cells, in cell-free microvesicles (MVs) or in complexes with other factors, such as RNA-binding proteins and high-density lipoprotein (HDL) particles. For example, EVs released by cells of the immune system can play a regulatory role in the induction and suppression of immune responses[94]. miRNA binding, in that context, was hypothesised to act as cell-to-cell communication mediators[95].

Besides posttranscriptional regulation mechanism, by targeting 3'UTR mRNAs, microRNAs are transcriptional modulator of epigenetic remodelling events at targeted gene promoters, playing a role in a more stable and heritable form of gene regulation[96]. The well conserved miR-10a, for example, targets a homologous DNA region in the promoter region of the hoxd4 gene, involved in animal development as well as in in tumour invasion and metastasis. microRNA-10a inhibits hoxd4 gene expression by targeting the promoter region and mediating chromatin remodelling DNA histone methylation, involving Dicer and Ago1-3[97].

Similarly, miRNAs can bind to evolutionary conserved loci that are complementary genomic seed-matches, corresponding to promoters, lineage-specific transcription factors and/or members of their epigenetic machinery. They bind polycomb proteins (PcGs) binding sites affecting gene expression at transcriptional level, not only at translational mRNA level. They were reported to guide chromatin remodelling complexes to specific genome sites in the nucleus[98] or promoting de-novo methylation of DNA[99].

miRNAs are found to associate with two kind of PcGs, PRC1 and PRC2, that are transcriptional repressors, strongly evolutionary conserved, that modify chromatin structure by covalent modification of histone proteins. PcGs own RNA binding properties and a leading role in PRC promoter targeting is played by microRNAs, as

the part of policomb-RNA complex that modulates PcG-promoter pairing[100]. The polycomb-microRNA complex catalyzes the histone H3 trimethylation or histone H2A monoubiquination of the “bivalent domain" of the NFI-A promoter, modulating NFI-A expression. Upregulation of NFI-A levels in primary hematopoietic stem/progenitor cells (HSC/HPCs) induces differentiation along the erythroid lineage, while downregulation leads toward the granulocytic linage. More specifically miR-223 was identified to target NFI-A promoter, thus mediating cell-lineage faith through PcGs recruitment[101].

A non-canonical miRNA function more related to miRNA-based post-transcriptional silencing can be related to the control of alternative translation.

There is a growing appreciation of widespread of alternative translation (AT) in mammals[102-107]. It’s a fine regulated mechanism and many mRNAs display a complex translation pattern, having multiple open reading frames (ORFs). ORFs can be mutually alternative, influence each other and/or the formation of secondary structures that can modulate ribosome activity, giving birth to alternative protein isoforms[108, 109]. miRNAs were reported to bind 5’UTR[110] and coding sequences (CDS) regions of mRNAs[111-113]. A recent work of Sonda et al[114] demonstrated that C/EBPβ mRNA alternative translation is regulated by miR-142-3p. The miRNA binds to non canonical site of the mRNA coding sequence, in a region including one of three in frame translation initiation site (TIS). The miRNA binding changes the ratio between protein isoforms with different properties, thus impacting on the cell phenotype.

This is just a short list of non canonical miRNA functions and this is an area of active ongoing study.

miRNAs deregulation and cancer

It is well known that gene expression is intricately regulated and there are multiple layers of expression level controls. microRNAs are active player of gene post transcriptional regulation and cellular homeostasis. They play important roles in many biological processes including cell differentiation, organogenesis, development, regulation of cell cycle and apoptosis. Dysregulation of miRNAs has been observed in many diseases but the cause-effect mechanism has not always been established, nor their role in tumour initiation and progression[115]. Moreover, dysregulated miRNAs of a tumour expression profile are not necessary pathogenetic and needlessly activate a

specific mechanism of action in oncogenesis[116]. So that dysregulation detection constitutes only the starting point for disease-relevant studies[117]. The deregulation of miRNAs in cancer can be due to epigenetic changes, as altered DNA methylation or histone acetylation, to defect in miRNAs biogenesis, as altered Dicer or Drosha activity, or to chromosomal abnormalities[118]. They can act either as tumour suppressor genes, promoting cancer cell death and/or to inhibit cancer cell growth, or as oncogenes, actively contributing to cancer cell proliferation. The first evidence of miRNA involvement in human cancer derived from studies on chronic lymphocytic leukemia (CLL). They were studying a deleted region in CLL at chromosome 13q14. They were expected to find a tumour suppressor gene, being the region deleted in tumour phenotype. They rather detected the miR-15a and miR-16-1 loci, transcribed from the same polycistronic RNA. These miRNAs target several factors that promote cell-cycle progression, as CDK6, CARD10 and CDC27. Lacking miRNAs post transcriptional regulation of these factors, due to deletion in haematopoietic and solid malignancies, there is an enhanced proliferative response of cancer cells to a variety of mitogenic stimuli[116]. In this case miRNAs influence tumour biology through their action as signal modulators along cancer-relevant pathways. Another way miRNAs can be involved in disease is through the increased or decreased activity of their transcription factors at the promoter. The well-known p53 tumour suppressor gene directly transactivates the miR-34a, miR-34b and miR-34c family transcription units, which are able to mediate some aspects of the cellular response to p53 activation, including cell-cycle arrest and apoptosis[119-121]. Another miRNA able to impair cell proliferation or induce apoptosis through oncogenes targeting is let-7, targeting RAS and MYC[122, 123], while opposite effect has miR-21 targeting tumour suppressor proteins in breast cancer, glioblastomas and pancreas[118, 124-126]. A number of microRNAs, that Hurst et al called metastamir, are demonstrated to play a role in the metastatic program, both displaying a pro- and anti-metastatic effects[127]. They showed involvement in epithelial-mesenchymal transition, migration and angiogenesis. miR-10b for example, is highly expressed in ~50% of metastatic tumours. It suppresses the homeobox D10 expression, leading to an increase in RHOC, a pro-metastatic gene, and initiation of breast cancer invasion and metastasis. miR-373 and miR-520c were studied by Agami et al[128], found to promote migration and to increase *in vivo* metastasis at least in part by targeting the adhesion molecule, CD44. On the other hand, miR-373 showed higher expression in lymph-node metastasis compared with

the primary tumours. Conversely, miR-146 family of miRNA could profoundly inhibit invasion and metastasis of MDA-MB-231 human breast carcinoma cells, similarly to miR-126 and miR-335 that are active regulators of tumour invasion and metastasis in human breast cancer.

The list of validated microRNAs involved in tumour development and progression is still long, acting through a huge variety of mechanisms. Although significant advances have been made so far, only fewer success in the development of miRNAs for use in therapy have been reported. A great step forward will be the use of circulating miRNAs in body fluids are accessible and allow a non invasive inquiry. They have already been detected in serum of women characterized of being expected and disappeared after the baby birth, as proof of pregnancy status. Circulating miRNAs expression level was further used to monitor the existence of cancer cell in patients[95, 129], as for early diagnosis with serum miRNAs in colorectal adenocarcinoma[130]or lung cancer[131]. Indeed, human body fluids, as blood, urine, saliva amniotic fluid, colostrum, breast milk, bronchial lavage, cerebrospinal fluid, peritoneal fluid, pleural fluid, tears and seminal fluid, have been shown to harbour extracellular miRNAs that are emerging as effective biomarkers for detection of diseases[132, 133]. However, the entire spectrum of miRNAs in the fluids has not been fully characterized.

More efforts must be put in that issue for an effective translation of miRNAs knowledge into clinical practices.

Aim of the work

In this work we considered different aspect of the microRNA word, integrating computational analyses of genome-wide datasets and targeted experimental results.

First, we studied miRNA role in myeloproliferative disorders, more specifically in primary myelofibrosis, considering also other small RNAs detected in RNA-seq data. This study deals indeed on a Illumina sequencing of small RNAs samples of CD34+ hematopoitic stem cells of patients affected by primary myelofibrosis and of controls, in order to characterize miRNA profiles and find relevant differentially expressed elements.

Then, in order to have a better understanding of each step of a computational analysis of RNA-seq data, we studied the impact on small RNAs differential expression analysis of long RNA-contrived normalization methods.

Thereafter we evaluated the implication of microRNAs in the gene expression changes observed after H-ferritin silencing in K562 cells.

The last part of this thesis report a preliminary study of the involvement of microRNAs in the regulation of alternative translation and thus of protein isoform equilibrium.

References

1. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5**, 621-628 (2008).

2. Nagalakshmi, U. *et al*. The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. *Science* **320**, 1344-1349 (2008).

3. Lister, R. *et al*. Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis. *Cell* **133**, 523-536 (2008).

4. Consortium, T. E. N. C. O. D. E. P. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**, 636-640 (2004).

5. Consortium, T. E. N. C. O. D. E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).

6. Djebali, S. *et al*. Landscape of transcription in human cells. *Nature* **489**, 101-108 (2012).

7. Gerstein, M. B. *et al*. Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91-100 (2012).

8. Celniker, S. E. *et al*. Unlocking the secrets of the genome. *Nature* **459**, 927-930 (2009).

9. Tuck, A. C. & Tollervey, D. RNA in pieces. *Trends in Genetics* **27**, 422-432 (2011).

10. Mattick, J. S. The genetic signatures of noncoding RNAs. *PLoS Genetics* **5** (2009).

11. Costa, F. F. Non-coding RNAs, epigenetics and complexity. *Gene* **410**, 9-17 (2008).

12. Bartel, D. P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**, 281-297 (2004).

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

14. Reinhart, B. J. *et al*. The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature* **403**, 901-906 (2000).

15. Pasquinelli, A. E. *et al*. Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature* **408**, 86-89 (2000).

16. Griffiths-Jones, S., Saini, H. K., Dongen, S. v. & Enright, A. J. miRBase: tools for microRNA genomics. *Nucleic Acids Res*. **36**, D154-D158 (2008).

17. Lagos-Quintana, M., Rauhut, R., Lendeckel, W. & Tuschl, T. Identification of Novel Genes Coding for Small Expressed RNAs. *Science* **294**, 853-858 (2001).

18. Lau, N. C., Lim, L. P., Weinstein, E. G. & Bartel, D. P. An Abundant Class of Tiny RNAs with Probable Regulatory Roles in Caenorhabditis elegans. *Science* **294**, 858-862 (2001).

19. Niwa, R. & Slack, F. J. The evolution of animal microRNA function. *Curr. Opin. Genet. Dev.* **17**, 145-150 (2007).

20. Luo, Y., Guo, Z. & Li, L. Evolutionary conservation of microRNA regulatory programs in plant flower development. *Dev. Biol.* **380**, 133-144 (2013).

21. Salmena, L., Poliseno, L., Tay, Y., Kats, L. & Pandolfi, P. A ceRNA Hypothesis: The Rosetta Stone of a Hidden RNA Language? *Cell* **146**, 353-358 (2011).

22. Lee, Y. *et al*. The nuclear RNase III Drosha initiates microRNA processing. *Nature* **425**, 415-419 (2003).

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

24. Zeng, Y., Yi, R. & Cullen, B. R. MicroRNAs and small interfering RNAs can inhibit mRNA expression by similar mechanisms. *Proceedings of the National Academy of Sciences* **100**, 9779-9784 (2003).

25. Yi, R., Qin, Y., Macara, I. G. & Cullen, B. R. Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. *Genes Dev.* **17**, 3011-3016 (2003).

26. Lund, E., GÃ¼ttinger, S., Calado, A., Dahlberg, J. E. & Kutay, U. Nuclear Export of MicroRNA Precursors. *Science* **303**, 95-98 (2004).

27. Dueck, A., Ziegler, C., Eichner, A., Berezikov, E. & Meister, G. microRNAs associated with the different human Argonaute proteins. *Nucleic Acids Res.* **40**, 9850-9862 (2012).

28. Ro, S., Park, C., Young, D., Sanders, K. M. & Yan, W. Tissue-dependent paired expression of miRNAs. *Nucleic Acids Res.* **35**, 5944-5953 (2007).

29. Axtell, M. J., Westholm, J. O. & Lai, E. C. Vive la diffÃ©rence: biogenesis and evolution of microRNAs in plants and animals. *Genome Biol.* **12** (2011).

30. Olena, A. F. & Patton, J. G. Genomic organization of microRNAs. *J. Cell. Physiol.* **222**, 540-545 (2010).

31. Westholm, J. O. & Lai, E. C. Mirtrons: microRNA biogenesis via splicing. *Biochimie* **93**, 1897-1904 (2011).

32. Lee, C., Risom, T. & Strauss, W. M. Evolutionary Conservation of MicroRNA Regulatory Circuits: An Examination of MicroRNA Gene Complexity and Conserved MicroRNA-Target Interactions through Metazoan Phylogeny. *DNA Cell Biol.* **26**, 209-218 (2007).

33. Wienholds, E. *et al*. MicroRNA Expression in Zebrafish Embryonic Development. *Science* **309**, 310-311 (2005).

34. Sempere, L. F., Cole, C. N., Mcpeek, M. A. & Peterson, K. J. The phylogenetic distribution of metazoan microRNAs: insights into evolutionary complexity and constraint. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* **306B**, 575-588 (2006).

35. Hertel, J. *et al*. The expansion of the metazoan microRNA repertoire. *BMC Genomics* **7** (2006).

36. Guo, Y. E. & Steitz, J. A. Virus Meets Host MicroRNA: the Destroyer, the Booster, the Hijacker. *Mol. Cell. Biol.* **34**, 3780-3787 (2014).

37. Ha, M. & Kim, V. N. Regulation of microRNA biogenesis. *Nature Reviews Molecular Cell Biology* **15**, 509-524 (2014).

38. Mingyi Xie, Joan A. Steitz. Versatile microRNA biogenesis in animals and their viruses. *RNA biology* **11** (2014).

39. Babiarz, J. E., Ruby, J. G., Wang, Y., Bartel, D. P. & Blelloch, R. Mouse ES cells express endogenous shRNAs, siRNAs, and other Microprocessor-independent, Dicer-dependent small RNAs. *Genes Dev.* **22**, 2773-2785 (2008).

40. Maute, R. L. *et al*. tRNA-derived microRNA modulates proliferation and the DNA damage response and is down-regulated in B cell lymphoma. *Proceedings of the National Academy of Sciences* **110**, 1404-1409 (2013).

41. Ender, C. *et al*. A Human snoRNA with MicroRNA-Like Functions. *Mol. Cell* **32**, 519-528 (2008).

42. Cazalla, D., Xie, M. & Steitz, J. A Primate Herpesvirus Uses the Integrator Complex to Generate Viral MicroRNAs. *Mol. Cell* **43**, 982-992 (2011).

43. Neilsen, C. T., Goodall, G. J. & Bracken, C. P. IsomiRs, the overlooked repertoire in the dynamic microRNAome. *Trends in Genetics* **28**, 544-549 (2012).

44. Azuma-Mukai, A. *et al*. Characterization of endogenous human Argonautes and their miRNA partners in RNA silencing. *Proceedings of the National Academy of Sciences* **105**, 7964-7969 (2008).

45. Fernandez-Valverde, S., Taft, R. J. & Mattick, J. S. Dynamic isomiR regulation in Drosophila development. *RNA* **16**, 1881-1888 (2010).

46. Tan, G. C. *et al*. 5' isomiR variation is of functional and evolutionary importance. *Nucleic Acids Res.* **42**, 9424-9435 (2014).

47. Jaskiewicz, L. & Zavolan, M. Dicer partners expand the repertoire of miRNA targets. *Genome Biol.* **13** (2012).

48. Neilsen, C. T., Goodall, G. J. & Bracken, C. P. IsomiRs, the overlooked repertoire in the dynamic microRNAome. *Trends in Genetics* **28**, 544-549 (2012).

49. Fukunaga, R. *et al*. Dicer Partner Proteins Tune the Length of Mature miRNAs in Flies and Mammals. *Cell* **151**, 533-546 (2012).

50. Wu, H., Ye, C., Ramirez, D. & Manjunath, N. Alternative Processing of Primary microRNA Transcripts by Drosha Generates 5â€² End Variation of Mature microRNA. *PLoS ONE* **4** (2009).

51. Morin, R. D. *et al*. Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Res*. **18**, 610-621 (2008).

52. Chen, K. & Rajewsky, N. Natural selection on human microRNA binding sites inferred from SNP data. *Nat*. *Genet*. **38**, 1452-1456 (2006).

53. Saunders, M. A., Liang, H. & Li, W. Human polymorphism at microRNAs and microRNA target sites. *Proceedings of the National Academy of Sciences* **104**, 3300-3305 (2007).

54. Llorens, F. *et al*. A highly expressed miR-101 isomiR is a functional silencing small RNA. *BMC Genomics* **14** (2013).

55. Cloonan, N. *et al*. MicroRNAs and their isomiRs function cooperatively to target common biological pathways. *Genome Biol*. **12** (2011).

56. Bortoluzzi, S., Biasiolo, M. & Bisognin, A. MicroRNA-offset RNAs (moRNAs): by-product spectators or functional players? *Trends Mol*. *Med*. **17**, 473-474 (2011).

57. Bortoluzzi, S. *et al*. Characterization and discovery of novel miRNAs and moRNAs in JAK2V617F-mutated SET2 cells. *Blood* **119**, e120-e130 (2012).

58. Langenberger, D. *et al*. Evidence for human microRNA-offset RNAs in small RNA sequencing data. *Bioinformatics* **25**, 2298-2301 (2009).

59. Shi, W., Hendrix, D., Levine, M. & Haley, B. A distinct class of small RNAs arises from pre-miRNAâ€“proximal regions in a simple chordate. *Nature Structural & Molecular Biology* **16**, 183-189 (2009).

60. Berezikov, E. *et al*. Deep annotation of Drosophila melanogaster microRNAs yields insights into their processing, modification, and emergence. *Genome Res*. **21**, 203-215 (2011).

61. Ruby, J. G. *et al*. Evolution, biogenesis, expression, and target predictions of a substantially expanded set of Drosophila microRNAs. *Genome Res*. **17**, 1850-1864 (2007).

62. Zhou, H. *et al*. Deep annotation of mouse iso-miR and iso-moR variation. *Nucleic Acids Res*. **40**, 5864-5875 (2012).

63. Taft, R. J. *et al*. Nuclear-localized tiny RNAs are associated with transcription initiation and splice sites in metazoans. *Nature Structural & Molecular Biology* **17**, 1030-1034 (2010).

64. Meiri, E. *et al*. Discovery of microRNAs and other small RNAs in solid tumors. *Nucleic Acids Res*. **38**, 6234-6246 (2010).

65. Eichhorn, S. *et al*. mRNA Destabilization Is the Dominant Effect of Mammalian MicroRNAs by the Time Substantial Repression Ensues. *Mol*. *Cell* **56**, 104-115 (2014).

66. Djuranovic, S., Nahvi, A. & Green, R. miRNA-Mediated Gene Silencing by Translational Repression Followed by mRNA Deadenylation and Decay. *Science* **336**, 237-240 (2012).

67. Guo, H., Ingolia, N. T., Weissman, J. S. & Bartel, D. P. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* **466**, 835-840 (2010).

68. Voinnet, O. Origin, Biogenesis, and Activity of Plant MicroRNAs. *Cell* **136**, 669-687 (2009).

69. Lytle, J. R., Yario, T. A. & Steitz, J. A. Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5'-UTR as in the 3'- UTR. *Proceedings of the National Academy of Sciences* **104**, 9667-9672 (2007).

70. Lim, L. P. *et al*. The microRNAs of Caenorhabditis elegans. *Genes Dev*. **17**, 991-1008 (2003).

71. Grimson, A. *et al*. MicroRNA Targeting Specificity in Mammals: Determinants beyond Seed Pairing. *Mol*. *Cell* **27**, 91-105 (2007).

72. Lai, E. C. Predicting and validating microRNA targets. *Genome Biol*. **5** (2004).

73. Brennecke, J., Stark, A., Russell, R. B. & Cohen, S. M. Principles of MicroRNAâ€“Target Recognition. *PLoS Biol* **3** (2005).

74. Doench, J. G., Petersen, C. P. & Sharp, P. A. siRNAs can function as miRNAs. *Genes Dev*. **17**, 438-442 (2003).

75. Witkos, T. M., Koscianska, E. & Krzyzosiak, W. J. Practical aspects of microRNA target prediction. *Curr*. *Mol*. *Med*. **11** (2011).

76. Hofacker, I. L. How microRNAs choose their targets. *Nat*. *Genet*. **39**, 1191-1192 (2007).

77. Bartel, D. P. MicroRNAs: Target Recognition and Regulatory Functions. *Cell* **136**, 215-233 (2009).

78. Baek, D. *et al*. The impact of microRNAs on protein output. *Nature* **455**, 64-71 (2008).

79. Pasquinelli, A. E. MicroRNAs and their targets: recognition, regulation and an emerging reciprocal relationship. *Nature Reviews Genetics* **13**, 271-282 (2012).

80. Enright, A. J. *et al*. MicroRNA targets in Drosophila. *Genome Biol*. **5**, R1-R1 (2004).

81. Kiriakidou, M. A combined computational-experimental approach predicts human microRNA targets. *Genes Dev*. **18**, 1165-1178 (2004).

82. Rehmsmeier, M. Fast and effective prediction of microRNA/target duplexes. *RNA* **10**, 1507-1517 (2004).

83. Rusinov, V., Baev, V., Minkov, I. N. & Tabler, M. MicroInspector: a web tool for detection of miRNA binding sites in an RNA sequence. *Nucleic Acids Res*. **33**, W696-W700 (2005).

84. Lewis, B. P. *et al*. Prediction of mammalian microRNA targets. *Cell* **115**, 787-798 (2003).

85. Lewis, B. P., Burge, C. B. & Bartel, D. P. Conserved Seed Pairing, Often Flanked by Adenosines, Indicates that Thousands of Human Genes are MicroRNA Targets. *Cell* **120**, 15-20 (2005).

86. Krek, A. *et al*. Combinatorial microRNA target predictions. *Nat. Genet.* **37**, 495-500 (2005).

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

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

89. Miranda, K. C. *et al*. A Pattern-Based Method for the Identification of MicroRNA Binding Sites and Their Corresponding Heteroduplexes. *Cell* **126**, 1203-1217 (2006).

90. Yue, D., Liu, H. & Huang, Y. Survey of computational algorithms for MicroRNA target prediction. *Curr. Genomics* **10** (2009).

91. Reyesâ Herrera, P. H. & Ficarra, E. One Decade of Development and Evolution of MicroRNA Target Prediction Algorithms. *Genomics, Proteomics & Bioinformatics* **10**, 254-263 (2012).

92. Peterson, S. M. *et al*. Common features of microRNA target prediction tools. *Frontiers in Genetics* **5** (2014).

93. Friedman, R. C., Farh, K. K., Burge, C. B. & Bartel, D. P. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.* **19**, 92-105 (2009).

94. van, d. G. & Nolte-â€™t Hoen, Esther N. M. â€œSmall Talkâ€ in the Innate Immune System via RNA-Containing Extracellular Vesicles. *Frontiers in Immunology* **5** (2014).

95. Kosaka, N. *et al*. Trash or Treasure: extracellular microRNAs and cell-to-cell communication. *Non-Coding RNA* **4** (2013).

96. Hawkins, P. & Morris, K. V. RNA and transcriptional modulation of gene expression. *Cell cycle (Georgetown, Tex.)* **7**, 602-607 (2008).

97. Tan, Y. *et al*. Transcriptional inhibiton of Hoxd4 expression by miRNA-10a in human breast cancer cells. *BMC Molecular Biology* **10** (2009).

98. Zardo, G. *et al*. Polycombs and microRNA-223 regulate human granulopoiesis by transcriptional control of target gene expression. *Blood* **119**, 4034-4046 (2012).

99. Sinkkonen, L. *et al*. MicroRNAs control de novo DNA methylation through regulation of transcriptional repressors in mouse embryonic stem cells. *Nature Structural & Molecular Biology* **15**, 259-267 (2008).

100. Zardo, G. *et al*. Transcriptional targeting by microRNA-Polycomb complexes: A novel route in cell fate determination. *Cell Cycle* **11**, 3543-3549 (2012).

101. Zardo, G. *et al*. Polycombs and microRNA-223 regulate human granulopoiesis by transcriptional control of target gene expression. *Blood* **119**, 4034-4046 (2012).

102. Kochetov, A. V. Alternative translation start sites and hidden coding potential of eukaryotic mRNAs. *Bioessays* **30**, 683-691 (2008).

103. Kochetov, A. V. *et al*. uORFs, reinitiation and alternative translation start sites in human mRNAs. *FEBS Lett*. **582**, 1293-1297 (2008).

104. Menschaert, G. *et al*. Deep proteome coverage based on ribosome profiling aids mass spectrometry-based protein and peptide discovery and provides evidence of alternative translation products and near-cognate translation initiation events. *Molecular & cellular proteomics: MCP* **12**, 1780-1790 (2013).

105. Vanderperre, B. *et al*. Direct Detection of Alternative Open Reading Frames Translation Products in Human Significantly Expands the Proteome. *PLoS ONE* **8** (2013).

106. Smith, E. *et al*. Leaky ribosomal scanning in mammalian genomes: significance of histone H4 alternative translation in vivo. *Nucleic Acids Res*. **33**, 1298-1308 (2005).

107. Wang, Y. *et al*. Gene Expression Profiles and Molecular Markers To Predict Recurrence of Dukes' B Colon Cancer. *Journal of Clinical Oncology* **22**, 1564-1571 (2004).

108. Morris, D. R. & Geballe, A. P. Upstream Open Reading Frames as Regulators of mRNA Translation. *Mol. Cell. Biol*. **20**, 8635-8642 (2000).

109. Skabkin, M., Skabkina, O., Hellen, C. T. & Pestova, T. Reinitiation and Other Unconventional Posttermination Events during Eukaryotic Translation. *Mol. Cell* **51**, 249-264 (2013).

110. Lytle, J. R., Yario, T. A. & Steitz, J. A. Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5′ UTR as in the 3′ UTR. *Proceedings of the National Academy of Sciences* **104**, 9667-9672 (2007).

111. Helwak, A., Kudla, G., Dudnakova, T. & Tollervey, D. Mapping the Human miRNA Interactome by CLASH Reveals Frequent Noncanonical Binding. *Cell* **153**, 654-665 (2013).

112. Tay, Y., Zhang, J., Thomson, A. M., Lim, B. & Rigoutsos, I. MicroRNAs to Nanog, Oct4 and Sox2 coding regions modulate embryonic stem cell differentiation. *Nature* **455**, 1124-1128 (2008).

113. Liu, C. *et al*. CLIP-based prediction of mammalian microRNA binding sites. *Nucleic Acids Res*. **41**, e138-e138 (2013).

114. Sonda, N. *et al*. miR-142-3p Prevents Macrophage Differentiation during Cancer-Induced Myelopoiesis. *Immunity* **38**, 1236-1249 (2013).

115. Garzon, R., Calin, G. A. & Croce, C. M. MicroRNAs in Cancer. *Annu. Rev. Med*. **60**, 167-179 (2009).

116. Mendell, J. & Olson, E. MicroRNAs in Stress Signaling and Human Disease. *Cell* **148**, 1172-1187 (2012).

117. Palanichamy, J. K. & Rao, D. S. miRNA dysregulation in cancer: towards a mechanistic understanding. *Frontiers in Genetics* **5** (2014).

118. Iorio, M. V. & Croce, C. M. MicroRNA dysregulation in cancer: diagnostics, monitoring and therapeutics. A comprehensive review. *EMBO Molecular Medicine* **4**, 143-159 (2012).

119. Bommer, G. T. *et al*. p53-Mediated Activation of miRNA34 Candidate Tumor-Suppressor Genes. *Current Biology* **17**, 1298-1307 (2007).

120. He, L. *et al*. A microRNA component of the p53 tumour suppressor network. *Nature* **447**, 1130-1134 (2007).

121. Chang, T. *et al*. Transactivation of miR-34a by p53 broadly influences gene expression and promotes apoptosis. *Mol. Cell* **26**, 745-752 (2007).

122. Johnson, S. M. *et al*. RAS Is Regulated by the let-7 MicroRNA Family. *Cell* **120**, 635-647 (2005).

123. Sampson, V. B. *et al*. MicroRNA Let-7a Down-regulates MYC and Reverts MYC-Induced Growth in Burkitt Lymphoma Cells. *Cancer Res*. **67**, 9762-9770 (2007).

124. Le Quesne, J. & Caldas, C. Micro-RNAs and breast cancer. *Molecular Oncology* **4**, 230-241 (2010).

125. Iorio, M. V. *et al*. MicroRNA Gene Expression Deregulation in Human Breast Cancer. *Cancer Res*. **65**, 7065-7070 (2005).

126. Iorio, M. V. & Croce, C. M. microRNA involvement in human cancer. *Carcinogenesis* **33**, 1126-1133 (2012).

127. Hurst, D. R., Edmonds, M. D. & Welch, D. R. Metastamir - the field of metastasis-regulatory microRNA is spreading. *Cancer Res*. **69**, 7495-7498 (2009).

128. Huang, Q. *et al*. The microRNAs miR-373 and miR-520c promote tumour invasion and metastasis. *Nat. Cell Biol*. **10**, 202-210 (2008).

129. Mittal, N. & Zavolan, M. Seq and CLIP through the miRNA world. *Genome Biol*. **15** (2014).

130. Zheng, G. *et al*. Serum microRNA panel as biomarkers for early diagnosis of colorectal adenocarcinoma. *Br. J. Cancer* **111**, 1985-1992 (2014).

131. Qin, X., Xu, H., Gong, W. & Deng, W. The tumor cytosol miRNAs, fluid miRNAs, and exosome miRNAs in lung cancer. *Cancer Endocrinology* **4** (2015).

132. Bahn, J. H. *et al*. The Landscape of MicroRNA, Piwi-Interacting RNA, and Circular RNA in Human Saliva. *Clin. Chem*. **61**, 221-230 (2015).

133. Weber, J. A. *et al*. The MicroRNA Spectrum in 12 Body Fluids. *Clin. Chem*. **56**, 1733-1741 (2010).

Chapter 1

We expanded the knowledge of miRNAs and moRNAs expressed by CD34+ cells, we identified and validated a few elements that can contribute to PMF pathogenesis. We considered small RNA sequencing data of 6 CD34+ cells, including 3 samples collected from 3 pools of bone marrow CD34+ cells of healthy subjects, and 3 samples of circulating CD34+ cells of patients affected by primary myelofibrosis (PMF), two of which were from a single patient and one was from a pool of 4 patients. In addition to 784 miRNAs annotated in miRBase, our in-house pipeline miR&moRe let us discover 34 new miRNAs expressed in our samples. miRNAs are *de facto* mixtures of isomiRs, specific variations of isomiRs expression impact also on miRNAs expression. Thus, we considered isomiR counts for miRNA expression calculations, recognizing that most of miRNAs detected are expressed in their isoform variants, not as the annotated sequence. We also detected in our samples sequences aligning to hairpins outside known and novel miRNAs that correspond to expressed microRNA-offset RNAs, called moRNAs. Myeloproliferative disorders are clonal hematopoietic stem cell neoplasias, miRNA and moRNA deregulation can be implied in tumor physiopathology. We then looked for differentially expressed small RNAs in PMF CD34+ samples respect to control samples. We recognized 37 sRNAs with significant differentially expressed (DE) in patient respect to control CD34+. Noteworthy, among the differentially expressed sRNAs, 2 moRNAs are included. hsa-3'-moR-128-2 was highly expressed in normal CD34+ cells and dramatically downregulated in PMF patients: the moRNA was not detected in considered PMF samples. We excluded multiple matching loci and ruled out mapping or annotations artifacts and made sure that the detected small RNA was a moRNA derived from the non-canonical processing of the human mir-128-2 hairpin. moRNA biological roles and mechanisms of function still deserve investigation. Very likely, moRNAs can function as miRNAs in post-transcriptional gene silencing, guiding RISC to complementary target mRNAs. Six of the selected small RNAs differentially expressed in PMF considering small RNA sequencing data in CD34+ resulted significantly differentially expressed, also in PMF granulocytes samples, we thus validated the differential expression of miR-10b-5p, miR-19b-3p, miR-29a-3p, miR-379-5p, miR-543 and moR-128-2. Target predictions of these validated small RNA were performed by using two different programs, miRanda and PITA. A functional

enrichment analysis, based on Reactome annotation maps, of targets predicted by both methods was obtained using a hypergeometric test. miRNA targets are enriched in many interesting pathways involved in tumor development and progression, as signaling by FGFR, DAP12 and Oncogene Induced Senescence. Hopefully identified and validated elements will help in the understanding the mechanisms that contribute to PMF pathogenesis and in formulate new targeted therapies.

small RNAs sequencing in CD34+ cells identifies miRNAs and moRNAs variable in PMF patients

Saccoman C[1], Bisognin A[1], Mannarelli C[2], Guglielmelli P[2], Vannucchi AM[2], Bortoluzzi S[1]

- [1]*Department of Biology, University of Padova, Padova, Italy;*
- [2]*Department of Hematology, University of Florence, Florence, Italy*

Abstract

Myeloproliferative neoplasms are chronic myeloid cancers involving CD34+ hematopoietic stem cells alterations. They include essential thrombocythemia (ET), polycythemia vera (PV) and myelofibrosis (MF). MF is the most severe form and can evolve in acute leukemia. Increasing evidence shows that deregulation of microRNAs (miRNAs) plays an important role in both solid and hematologic malignancies. Moreover, previous studies showed that microRNA-offset RNAs (moRNAs) could be expressed by processing of miRNA precursors. To attain deeper knowledge of small RNAs expressed in CD34+ cells and of the possible miRNA/moRNA-mediated post-transcriptional regulation in PMF, we sequenced with Illumina HiSeq2000 technology CD34+ cells from healthy subjects and from of patients affected by primary myelofibrosis (PMF).

We detected the expression of 784 known miRNAs, discovered 34 new miRNAs and 99 new miRNA-offset RNAs (moRNAs), expressed in in CD34+ cells. We then identified 37 small RNAs (DEMs) differentially expressed in patients respect to healthy subjects, with a prevalence of miRNA up-regulation in the disease. Six of the 37 identified small RNAs resulted significantly differentially expressed also in PMF granulocytes samples, we thus validated the differential expression of miR-10b-5p, miR-19b-3p, miR-29a-3p, miR-379-5p, miR-543 and moR-128-2. Target predictions of these validated small RNA and a functional enrichment analysis were performed. miRNA targets are enriched in many interesting pathways involved in tumor

development and progression, as signaling by FGFR, DAP12 and Oncogene Induced Senescence.

In this study, we expanded the knowledge of miRNAs and moRNAs expressed in CD34+ cells, identified and validated a few elements that can contribute to PMF pathogenesis. Hopefully this information will help in the understanding the mechanisms that contribute to PMF pathogenesis and in formulate new targeted therapies.

Background

Philadelphia-negative chronic myeloproliferative neoplasms (MPNs) are a heterogeneous group of clonal hematopoietic stem cell (HSC) disorders associated with overproduction of mature myeloid cells[1, 2].

MPNs are chronic myeloid cancers that include essential thrombocythemia (ET), polycythemia vera (PV) and primary myelofibrosis (PMF). MPNs can be complicated by thrombosis and/or hemorrhage and they may evolve into acute myeloid leukemia.

MF, the most serious MPN form, can arise primarily (PMF) or follow PV and ET onset (post-PV/ET MF). In primary myelofibrosis (PMF) the abnormal proliferation of megakaryocytes is accompanied by deposition of fibrous connective tissues in the bone marrow, abnormal stem cell trafficking, and extramedullary hematopoiesis (myeloid metaplasia).[1, 2]

PMF is associated with marked hepatosplenomegaly, anemia and profound constitutional symptoms including fatigue, weight loss, cachexia, pruritus, night sweats, low-grade fever, and bone and joint pain. Treatment, apart from "conventional" allogeneic stem cell transplantation (SCT), is guided by risk stratification and the patient's clinical needs.[2-4]

In 2005 the first mutation related to MPNs was identified in the Janus Kinase 2 (JAK2)[5-9].

The JAK2 V617F mutation is present in approximately 95% of patients with PV, and in 50% to 60% of those with ET or primary MF (PMF). Additional mutations have been identified in patients who have myeloproliferative neoplasms with or without JAK2 mutations: in particular a signaling mutation that activates the thrombopoietin receptor (MPL) and in epigenetic regulators, but also chromosomal aberrations.[2, 10-12]

In 2013, somatic mutations of *CALR*, the gene encoding calreticulin, have been found in 20% to 25% of patients with essential thrombocythemia (ET) or PMF. Like *JAK2* and *MPL* mutations, somatic mutations of *CALR* behave as driver mutations responsible for the myeloproliferative phenotype.

Despite the fact that the mutational landscape of MPNs has been extensively investigated, the molecular etiology of the disease has not been fully elucidated. Indeed several lines of evidence indicate that the identified mutations are not sufficient for disease initiation and progression. Although murine models have provided unequivocal evidence that *JAK2*^{V617F} is able to cause MPNs[13], disease

phenotype is significantly heterogeneous between different murine lines and even within the same line, suggesting that disease phenotype is affected by other unknown genetic or epigenetic factors[14].

MicroRNAs are endogenous small non-coding RNAs, approximately 22 nt in length, crucial for post-transcriptional gene regulation. They are loaded into the RNA-induced silencing complex (RISC), directing the complex (including Argonaute proteins) to downregulate target mRNA expression by either triggering mRNA degradation or translational repression.[15] Recent studies show that mRNA destabilization explains most (66%–>90%) miRNA-mediated repression[16].

It is known that that deregulation of miRNAs plays an important role in both solid and hematologic malignancies[17, 18]. Indeed, hematopoietic differentiation is tightly governed by gene expression that is strictly regulated at multiple cell-fate decision levels. miRNAs regulate hematopoiesis acting both in HSC and in committed progenitor cells[18-20]. At the stem cell level, some miRNAs evolutionally conserved are responsible for expanding HSCs by inhibiting apoptosis[21-23]. At the progenitor cell level, miRNAs regulate the developmental fate of the megakaryocyte-erythroid progenitor (MEP) cell, the common progenitor of the erythroid and megakaryocytic lineages[24, 25]. At the more committed hematopoietic cell level, specific miRNAs are expressed in different blood cell lineages and in different stages of hematopoietic differentiation. For example Chen et al.[26] reported that miR-142s expression was lower in the erythroid and T-lymphoid lineages and higher in B-lymphoid and myeloid lineages, while miR-223 expression was confined to myeloid lineages, with a very low detectable expression in T- and B-lymphoid and erythroid lineages.

miRNAs have an important role in regulation of hematopoiesis[27-30]. miR-16, miR-451 upregulation and miR-150, miR-155, miR-221 and miR-222 downregulation are associated with different stages of erythropoiesis[31, 32]. miR-223 expression level, determined by two regulatory regions on its gene, fine-tunes lineage commitment of myeloid precursor[33]. miR-181 family was detected during granulocytic and macrophage-like differentiation and its level decrease along the hematopoietic lineage. They modulate differentiation by targeting and negatively regulating PRKCD mRNA, an upstream regulator of a pathway of the myeloid differentiation, and CAMKK1 mRNA, involved in the granulocytic and PMA-induced macrophage-like differentiation[34, 35].

Recent studies highlighted aberrant miRNA expression in MPNs, and specific miRNA signatures that distinguish MPN granulocytes from those of healthy donors[18, 36].

A recent study characterized both gene and microRNA (miRNA) expression profiles in CD34+ cells from PMF patients[37]. It identified several biomarkers and putative molecular targets such as FGR, LCN2, and OLFM4. By means of miRNA-gene expression integrative analysis, the study suggested that JARID2 downregulation, mediated by miR-155-5p overexpression, might contribute to MK hyperplasia in PMF.

High-throughput analysis of miRNA expression levels in MPN $CD34^+$ cells were previously reported only by Lin et al.[35,38] and by Zhan et al.[26].

In a preliminary study, we performed short RNA massive sequencing and extensive bioinformatic analysis in the JAK2V617F-mutated SET2 cell line[39], detected and quantified 652 known mature miRNAs, of which 21 were highly expressed, thus being responsible of most of miRNA-mediated gene repression. In the same study, we showed that the majority of miRNAs were mixtures of sequence variants (isomiRs) and we identified 78 novel miRNAs. Indeed, we discovered that SET2 cells express a number of miRNA-offset RNAs (moRNAs), short RNAs derived from genomic regions flanking mature miRNAs, whose biological role needs to be elucidated.

In this study, we characterized miRNA and moRNA expression in CD34+ stem cells using massive small RNA-seq. The observed specificities in small RNAs expression of PMF CD34+ cells were subsequently confirmed considering granulocytes from PMF, PV and ET patients and from healthy controls. We thus provided new information regarding the possible role of miRNAs and new moRNAs in the disease.

Materials and Methods

Small RNA-seq library construction and sequencing

We deep-sequenced small RNAs libraries using Illumina HiSeq2000 technology, single reads from 49 to 57 bp. We sequenced 3 samples of pooled CD34+ bone marrow cells from healthy subjects (unknown genders) and 3 samples of circulating CD34+ cells of patients affected by primary myelofibrosis (PMF), a myeloproliferative neoplasm, two of which were from a single patient and one was from a pool of 4 patients, for a total number of 6 sequenced samples. All samples and raw reads information are summarized in Supplementary Table 1.

Small RNA data analysis: preprocessing

First step of data analysis is data preprocessing. The starting point is adapter removal. Reads “adapter-only”, too short or unclipped have been discarded. Unclipped reads are discarded because they can't represent miRNA or miRNA like short RNAs.
We admitted a read length range between 15 and 30 nt, slightly wider than the human annotated miRNAs length in miRBase to conserve also possible new longer isomiRs. We therefore discarded raw reads out of the range 15-30 bps in length. We then filtered out low quality reads, keeping all that reads displaying a base mean quality higher than 30, and allowing no more than 2 nucleotides per read with quality under 20. To complete data preprocessing we eliminated ground noise, considered as reads belonging to unique sequences with less than 10 reads counts each.

Small RNA data analysis: reads mapping and comparative filtering

Reads have been mapped using Bowtie v. 1.1.0 both to the GRCh38 genome assembly and the known hairpins sequences extended in both directions by additional 30 bp to accommodate moRNAs mapping at the extremities of known hairpins. Reads mapping to more than 5 different loci on the genome, out of miRNA hairpins, are unlikely to be real miRNAs, and they have been thus discarded.

Barplot on Supplementary Figure 1 shows filtering effects on absolute reads counts for each sample.

We processed each sample data with our in-house pipeline miR&more. The output consists of lists of known miRNA read counts, lists of new miRNAs and moRNAs and lists of variants (isomiRs) for all the small RNAs found in each sample.

Expression data normalization and sample cluster analysis

Sample merging and carefully conducted steps of data normalization and transformation are needed to guarantee the comparability of samples, to allow descriptive unsupervised analyses and differential expression tests. We performed normalization using R/Bioconductor package DESeq. Inference of differential expression in DESeq relies on the estimation of the typical relationship between the data variance and their mean, or, equivalently, between the data dispersion and their mean. Variance dependence to the mean can be modeled following different ways using DESeq, with several algorithms and very different results. The selection of the method used is crucial, since variance estimation influences unsupervised classification, differential expression and all following analyses. We tried two different methods for fitting data variance: 1) a parametric model, 2) a local regression model.

The first is the recommended default but in some data sets could fail to give optimal results. Sum of square of residuals for local regression is 15311.52 whereas for GLM is 156849.3, so we can conclude that local regression fits better our data.

We performed cluster analysis using R to check whether samples were correctly classified in their own biological class. We clustered samples using both full small RNAs expression matrix (904 small RNAs) and filtered, computing euclidean distance and complete linkage as clustering method.

To filter the expression matrix at different levels of small RNAs expression we calculated, for each small RNA in the matrix, the sum of expression vector values. Then we performed clustering analysis under the following conditions 1) considering all the small RNAs found 2) selecting only small RNAs over the median (430 miRNAs and 22 moRNAs), 3) filtering only small RNAs over the third quartile (219 miRNAs and 7 moRNAs).

Differentially expressed sRNAs

We performed a differential expression analysis using DESeq R/Bioconductor package. We considered those short RNAs that had a total expression throughout all samples higher than the median. We performed a multiple test correction according to the Benjamini Hochberg (FDR). We considered a corrected p-value of 0.05 as threshold to identify differentially expressed elements.

Validation of differentially expressed sRNAs

We performed Real-time PCR (RT-PCR) assays on granulocytes using single TaqMan MicroRNA Assay (Applied Biosystem). These were carried out in an independent cohort of normal controls (n=10) and patients with PMF (n=50), polycythemia vera (PV)(N=30) and essential thrombocythemia (ET)(n=30).

Target prediction of validated small RNAs and functional enrichment

The complexity of miRNA-mRNA interactions causes ambiguity in target prediction results. Target genes identification is indeed challenging and many algorithms have been developed. Target prediction programs can be divided in two classes, distinguished on the basis of the use or not of the information about evolutionary conservation of interaction[68]. We chose to perform a target prediction using two different programs, miRanda[55] and PITA[63], which implement orthogonal target prediction strategies. Our choice was determined also by code availability that allowed us to make custom predictions using as query sequences also isomiRs and moRNA sequences.

We performed a target prediction using both miRanda 3.3a and PITA executable version 6 (31-Aug-08). We applied default parameters of miRanda to predict target of selected small RNA sequences since these settings are reported to optimize the dynamic programming miRanda algorithm. We used default parameters for PITA target prediction too. According to PITA documentation, we considered a binding site

with score <= -10 likely to be functional in endogenous microRNA expression levels. We performed a hypergeometric test using an in-house modified version of the R Category package of Bioconductor that supports Reactome annotation maps via the reactome.db R package.

Results and Discussion

sRNA sequencing libraries

Small RNA libraries were prepared from 1 μg of total RNA according to TruSeq Small RNA kit.

Quality control was performed on a high sensitivity DNA chip (Agilent).

The purified cDNA libraries was used for cluster generation on Illumina's Cluster Station and sequenced on an Illumina HiSeq2000 instrument.

small RNA expressed in and PMF CD34+ cells and in PMF patients

We considered small RNA sequencing data of 6 CD34+ cells, including 3 samples collected from 3 pools of bone marrow CD34+ cells of healthy subjects (CTR; unknown gender), and 3 samples of circulating CD34+ cells of patients affected by primary myelofibrosis (PMF), two of which were from a single patient and one was from a pool of 4 patients.

The Illumina 2000 sequencing produced a total of 787,913,722 raw reads (131,318,954 per sample mean). After a stringent filtering during the preprocessing and quality control steps (Supplementary Figure 1), 349,372,534 were aligned to GRCh38 genome “extended” hairpins. Aligned reads corresponded to 44.3% of initial raw reads, but due to high sequencing depth, the number of considered reads was still high.

Table 1 reports a summary of different types of small RNAs detected in the considered samples, according to current miRNA annotations.

	CTR	ONLY CTR	PMF	ONLY PMF	ALL
known miRNAs	568	24	760	216	784
new miRNAs	20	3	31	14	34
moRNAs	52	3	96	47	99
Total new sRNAs	72	6	127	61	133
Total sRNAs	640	30	887	277	917

Table1 Summary of small RNAs expressed in considered CD34+ cells.

We detected a total of 917 small RNAs expressed in at least one of the 6 considered CD34+ samples, including 784 know miRNAs. Notably, 8 known miRNAs are highly expressed and contribute to the total miRNA expression throughout all samples from 2.5% to 25%, representing all together the 80% of the total expression. Moreover, we detected 133 new small RNAs, including 34 new miRNAs produced from known hairpins and 99 microRNA-offset RNAs (moRNAs).

Descriptive sample clustering analysis was conducted to check if differences in mapped read numbers across samples may affect expression estimation and small RNAs profile sample characterization. Cluster analysis and heatmaps are represented in Supplementary Figure 2. Two heatmap plots were generated, by considering normalized expression profiles of all the small RNAs expressed and considering only those expressed over the median level. Both unsupervised analyses show the same result: samples are correctly clustered, with CTR samples clustered together and separately from PMF samples. Looking at the distances, CTRs are closer to each other than the MFs in their group pointing out an increased variability of small RNA expression profiles in patient samples respectively to controls.

Cluster analyses and heatmaps demonstrate that patients and controls small RNAs profiles are significantly different from each other and highlight a characteristic miRNA and moRNA expression profile in PMF.

New miRNAs

In addition to 784 miRNAs annotated in miRBase, our in-house pipeline miR&moRe let us discover 34 new miRNAs expressed in our samples (Supplementary Table 2).

To find new miRNAs, we considered all the hairpins precursors annotated in miRBase, to be used as reference for read mapping and small RNA detection and quantification. Some of the known hairpin precursors were known to only generate one mature miRNA with only a handful of reads reported in mirBase across all NGS experiments surveyd. After identifying the hairpin region that would most likely pair with annotated mature we classify as new miRNAs all the clusters of reads that map there. A consistent number of reads were attributed to new miRNAs. Since these reads passed stringent quality filtering and mapping criteria, it is improbable that they could be sequencing errors. Furthermore, two of these new miRNAs: hsa-miR-2110* and hsa-miR-548ag-2* are highly expressed and show a mean expression over the median of the mean expression values distribution calculated on all the detected small RNAs.

miRNAs are mixtures of isoforms contributing to miRNAs expression

Our and other previous studies showed that miRNAs are mixtures of sequences, slightly different from the official mature miRNA, called isomiRs [39, 40]. Microarray technology, relying on sequence hybridization to appropriately designed annealing probes, can only detect annotated miRNA sequences. Sequencing-based technology reveals instead all the repertoire of expressed small RNAs, both the unknown and the annotated miRNAs, and is able to detect isomiRs, that contribute to miRNA expression and qualitative characteristics.

In our dataset, miRNA expression counts indeed consider for each miRNA a group of reads that not only perfectly match the annotated miRNA (“exact”), but also match the precursor with a 1-2 nt shorter or longer sequence than the mature miRNA in the 3’ region (“shorter or longer at 3’ ”), in the 5’ end (“shorter or longer at 5’”) or at both the ends (“both”).

Expressed miRNAs with unique sequence are very few, only 165 out of the 784 annotated miRNAs detected (21%). Unique sequence miRNAs are in general weakly expressed: they were detected at level under the median of the mean expression values

distribution. Notably, the remaining 619 expressed miRNAs have more than one isomiR. We also detected reads aligning to hairpins precursor with one or two mismatches but we excluded these reads from the total miRNAs expression estimation.

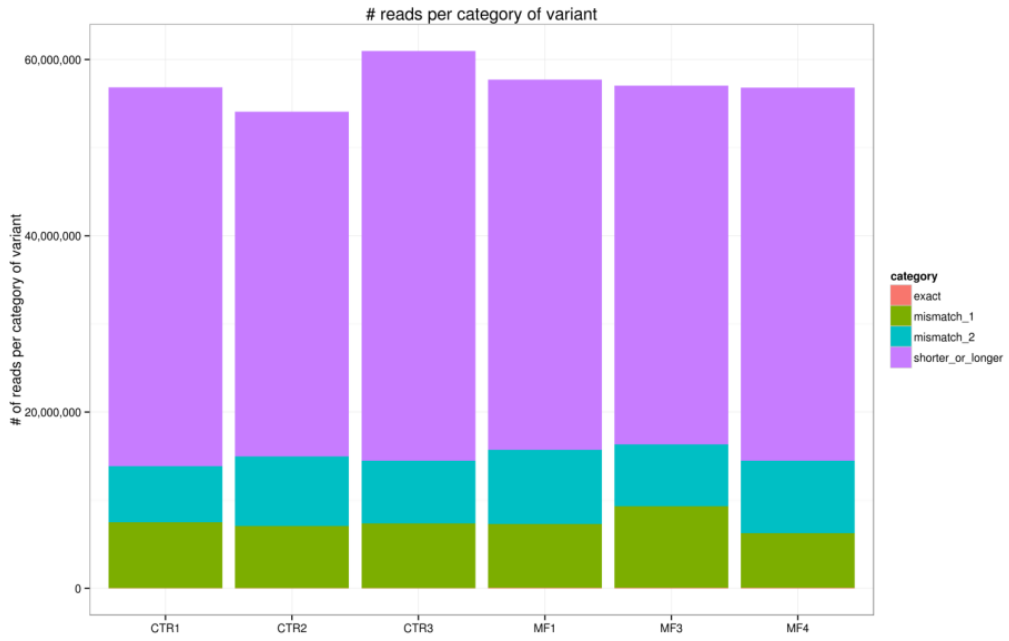

Figure 1 Distribution of reads per category across sample. "Exact" reads are identical to the mature miRNA sequence annotated in miRBase, whereas "mismatch" reads present respectively one or two nucleotides different from the annotated sequence but identical length; the last category includes reads perfecly matching the miRNA precursor (and genomic) sequence but shorter or longer than the annotated mature miRNA. "Exact" reads are rare, while shorter or longer are very abundant.

Figure 1 and Supplementary Table 3 show that the annotated sequence, called "exact" isomiR, rarely is the dominantly expressed form, while most of the total expression contribute is given by shorter or longer isomiR sequences. Reads aligning with mismatches are also represented in display items for comparison. These considerations are based on statistics considering all expressed miRNAs. Looking at singular miRNAs, the ratio between isomiR types changes. For example miR-10b-5p is mainly detected in its "shorter or longer" variant (Figure 2).

We considered the possibility that peculiar genetic characteristics of the considered PMF cells would result in specific isomiR sequences, or can be related to variations of isomiRs expression level in disease. Our results did not identified isomiRs expressed exclusively in PMF samples. Anyway, since miRNAs are *de facto* mixtures of isomiRs, specific variations of isomiRs expression impact also on miRNAs

expression. Thus, we considered isomiR counts for miRNA expression calculations and for the following analyses.

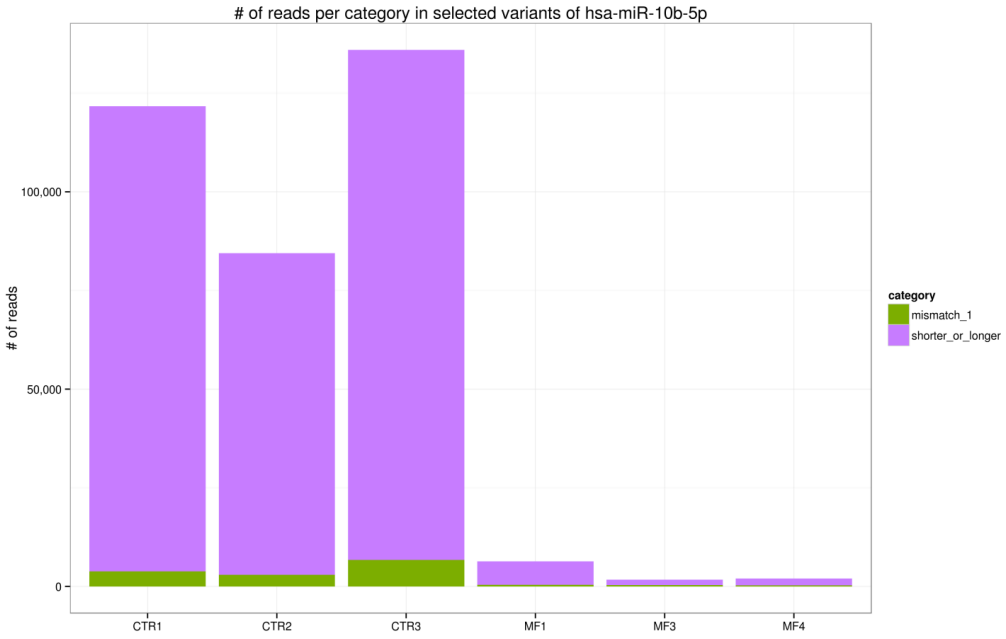

Figure 2. Abundance of reads falling in different isomiR types fro hsa-miR-10b-5p. Reads different from the annotated mature miRNA sequence are the most abundant.

moRNAs discovery

As anticipated, we also detected in our samples sequences aligning to hairpins outside known and novel miRNAs, that correspond to expressed microRNA-offset RNAs, called moRNAs (Table 1).

moRNAs sequences partially overlap miRNA regions but generally span the Drosha cutting sites, letting us hypothesize a non canonical processing of the hairpin precursor in moRNA biogenesis.[39]

A complete list of all the detected moRNAs is in Supplementary Table 4. Noteworthy, 28 moRNAs were highly expressed, 26 of them over the median of the short RNAs expression values distribution and 12 of them even over the third quartile (Table 2).

name	CTR	MF	strand	position	seq
hsa-moR-128-2-3p	2493	0	+	chr3:35744548-35744568	CCCTACTGTGTCACACTCCTA
hsa-moR-21-5p	2946	2437	+	chr17:59841243-59841271	ACATCTCCATGGCTGTACCACCTTGTCGG
hsa-moR-24-2-5p	5971	1719	-	chr19:13836350-13836376	TGCCTGGCCTCCCTGGGCTCTGCCTCC
hsa-moR-27a-5p	1752	312	-	chr19:13836510-13836534	CGAAGCCTGTGCCTGGCCTGAGGAG
hsa-moR-3651-5p	0	1266	-	chr9:92292537-92292565	ATGGACAGCTCTCCAGTGGATTCGATGGG
hsa-moR-421-5p	848	2784	-	chrX:74218449-74218472	CCTAATCCGGTGCACATTGTAGGC
hsa-moR-6724-1-5p	683	280	+	chr21:8205298-8205332	TGTGGGGGAGAGGCTGTCGCTGCGCTTCTGGGCCC
hsa-moR-6724-2-5p	683	280	+	chr21:8249488-8249522	TGTGGGGGAGAGGCTGTCGCTGCGCTTCTGGGCCC
hsa-moR-6724-3-5p	683	280	+	chr21:8388345-8388379	TGTGGGGGAGAGGCTGTCGCTGCGCTTCTGGGCCC
hsa-moR-6724-4-5p	683	280	+	chr21:8432513-8432547	TGTGGGGGAGAGGCTGTCGCTGCGCTTCTGGGCCC
hsa-moR-941-4-5p	6564	2531	+	chr20:63919746-63919768	CACCCGGCTGTGTGCACATGTGC
hsa-moR-941-5-5p	9780	3800	+	chr20:63919858-63919880	CACCCGGCTGTGTGCACATGTGC

Table 2. List of most abundant moRNAs in considered CD34+ samples, which are expressed over the third quartile of all sRNAs expression.

We classified moRNAs on the basis of the hairpin precursor arm they where processed from: 5'-moRNAs mapping to the 5' hairpin arm, and 3'-moRNAs spanning over the 3' hairpin arm. 5'-moRs were significantly more abundant respect to 3'-moRs. Out of 99 moRNAs expressed in considered samples, only 16 (16.2%) were processed from the 3' hairpin arm, while 83 (91.1%) were 5'-moRs. According to our data, seven hairpins were processed producing two moRNAs each (Supplementary Table 4).

5'-moRs estimated expression values were 10 times higher than 3'-moRs, ranging from summed up normalized values over all samples of 5 to 40,739, compared to a 3'-moRs range of 6 to 7,478. Both 3'-moRs and 5'-moRs are more expressed in controls than in PMF patient samples.

Considering all the small RNAs processed from the same hairpins precursors from which the 9 most expressed moRNAs are derived, we see that whereas miRNAs are

more expressed in PMF samples compared to CTRs, the relation is flipped in moRNAs, that are more expressed in normal CD34+ than in PMF samples (Figure 3). For the same set of moRNAs, Figure 4a shows in detail all the expressed small RNAs that are produced from the same hairpins precursor.

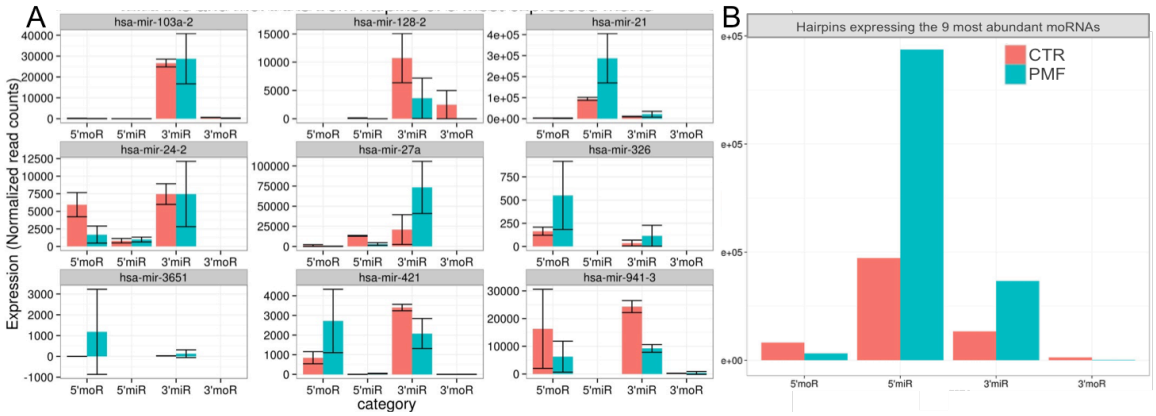

Figure 3. Expression of miRNAs and moRNAs produced from the same hairpin, considering the hairpins expressing most abundant moRNAs considered separately (A) and together (B).

We observed that moRNAs sequences partially overlap mature miRNA regions and generally span the Drosha cutting sites, with moRNAs protruding from the canonical hairpin precursor.

Identification of sRNA differentially expressed in PMF vs CTR

Since myeloproliferative disorders are clonal hematopoietic stem cell neoplasias, miRNA and moRNA deregulation can be implied in tumor physiopathology. We then looked for differentially expressed small RNAs in PMF CD34+ samples respect to control samples. We recognized 37 sRNAs with significant differentially expressed (DE) in patient respect to control CD34+ (Table 3).

name	CTR	PMF	log2(FC)	P value	P-val-adj
hsa-miR-1185-5p	0	136	15.00	1.24E-05	6.98E-04
hsa-miR-127-5p	0	317	15.00	4.66E-07	7.01E-05
hsa-miR-1277-5p	0	98	15.00	1.53E-04	6.28E-03
hsa-miR-299-3p	0	210	15.00	1.78E-04	6.69E-03
hsa-miR-323a-3p	0	121	15.00	5.34E-03	7.29E-02

hsa-miR-377-3p	0	206	15.00	9.34E-04	1.92E-02
hsa-miR-377-5p	0	155	15.00	8.32E-06	6.25E-04
hsa-miR-379-3p	0	89	15.00	2.76E-04	9.56E-03
hsa-miR-382-5p	0	205	15.00	2.97E-04	9.56E-03
hsa-miR-431-3p	0	80	15.00	1.72E-03	3.11E-02
hsa-miR-490-3p	0	719	15.00	2.48E-03	3.99E-02
hsa-miR-539-3p	0	204	15.00	1.37E-03	2.58E-02
hsa-miR-543	0	262	15.00	4.64E-04	1.40E-02
hsa-miR-654-5p	0	99	15.00	6.37E-04	1.60E-02
hsa-miR-656	0	78	15.00	8.90E-04	1.91E-02
hsa-miR-665	0	317	15.00	2.67E-03	4.15E-02
hsa-miR-758-3p	0	171	15.00	3.04E-03	4.56E-02
hsa-miR-873-5p	0	211	15.00	6.57E-04	1.60E-02
hsa-miR-25-3p	37	226192	12.59	2.85E-27	1.29E-24
hsa-miR-29a-3p	206	44817	7.77	6.73E-04	1.60E-02
hsa-miR-136-5p	34	1516	5.47	1.24E-04	5.57E-03
hsa-miR-495-3p	8	281	5.09	6.43E-04	1.60E-02
hsa-miR-873-3p	7	239	5.04	5.15E-03	7.27E-02
hsa-miR-485-5p	12	353	4.84	1.17E-03	2.30E-02
hsa-miR-19b-3p	3795	99173	4.71	7.13E-07	8.03E-05
hsa-miR-432-5p	14	343	4.62	2.09E-03	3.62E-02
hsa-5'-moR-542	9	232	4.62	6.72E-03	8.66E-02
hsa-miR-379-5p	31	561	4.16	7.86E-03	9.58E-02
hsa-miR-19a-5p	15	265	4.16	5.16E-03	7.27E-02
hsa-miR-33b-5p	26	394	3.90	6.09E-03	8.08E-02
hsa-miR-1307-5p	779	10914	3.81	8.23E-04	1.86E-02
hsa-miR-142-3p	5954	33867	2.51	7.73E-03	9.58E-02
hsa-miR-3150b-3p	196	7	-4.90	2.23E-03	3.73E-02
hsa-miR-10b-5p	119504	2855	-5.39	2.15E-08	4.84E-06
hsa-3'-moR-128-2	2489	0	-15.00	1.15E-05	6.98E-04
hsa-miR-128-2*	102	0	-15.00	1.87E-05	9.36E-04
hsa-miR-5008-3p	149	0	-15.00	4.17E-06	3.76E-04

Table 3. List of 37 most expressed miRNAs

Figure 4A shows the logarithm of the mean expression ratio in PMF and control cells for DE miRNAs and moRNAs. As shown in Figure 4A, the most part of DE sRNAs are upregulated in patients, while only five small RNAs are downregulated.

Noteworthy, among the differentially expressed sRNAs, 2 moRNAs (hsa-5'-moR-542 and hsa-3'-moR-128-2) are included.

hsa-5'-moR-542 results up-regulated in PMF with a log_2FC of 3.5.

hsa-3'-moR-128-2 highly expressed in normal CD34+ cells (at levels over the third quartile of the overall small RNA expression distribution) and dramatically downregulated in PMF patients: the moRNA was not detected in considered PMF samples. We mapped hsa-3'-moR-128-2 sequence to the whole human genome to exclude multiple matching multiple loci and to rule out mapping or annotations artifacts. We can thus exclude that moRNA-associated reads could come from different or contaminating RNAs. An additional UCSC Blat[41] analysis confirmed that the moRNA sequence only aligned to chr3:35786042-35786062. We are therefore confident that the detected small RNA is a moRNA derived from the non-canonical processing of the human mir-128-2 hairpin.

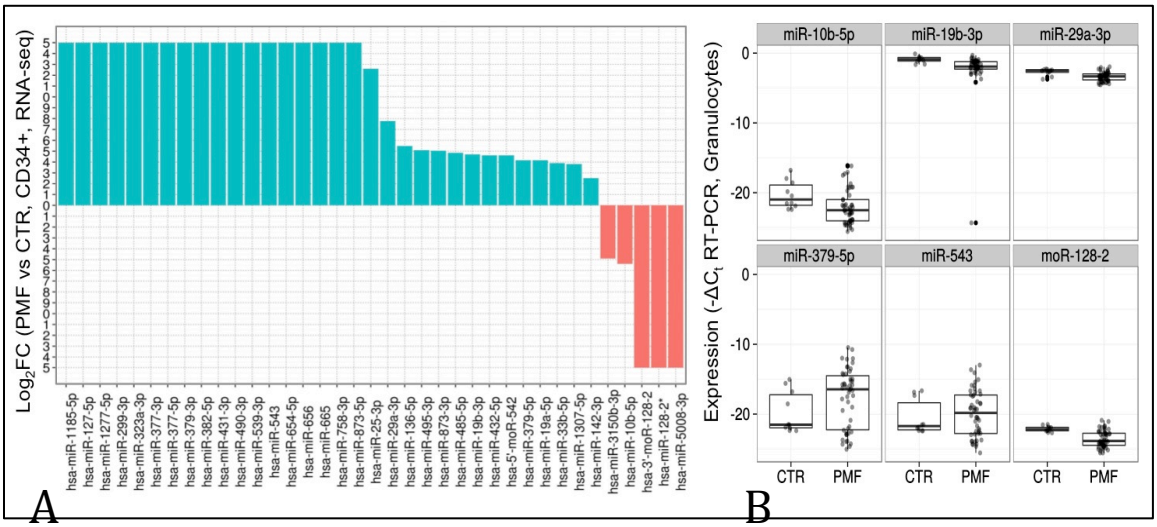

Figure 4. Differential expression of small RNAs in PMF vs CTR CD34+. A) Log2 FC of small RNA differentially expressed considering PMF vs CTR CD34+, according to RNA-seq data. When a small RNA was not expressed in one sample category, the ratio was infinite and we represent it as the arbitrary maximum value of 15. B) RT-PCR expression calculation in granulocytes collected from an independent and sizeable cohort of normal controls (n=10) and of PMF (50), PV (30) or ET (30) patients.

Validations confirmed 6 differentially expressed sRNAs

We selected the most significantly deregulated and highly expressed detected differentially expressed miRNAs for further analysis. Specifically, we considered 21

small RNAs among the 37 differentially expressed for a quantification with Real-time PCR (RT-PCR) in granulocytes collected from an independent and sizeable cohort (total n=120) of normal controls and PMF, PV or ET patients.

Six of the selected small RNAs differentially expressed in PMF considering small RNA sequencing data in CD34+ resulted significantly differentially expressed, also in PMF granulocytes samples: miR-10b-5p and moR-128-2 from RT-PCR, and miR-19b-3p, miR-29a-3p, miR-379-5p, miR-543 from a previous study of Norfo et al[37] (Figure 4B).

For these small RNAs the evidence of differential expression in PMF was robust, since it was detected by NGS and also validated technically (by qRT—PCR) and biologically (in independent samples). For the remaining miRNAs, the observed differences were not confirmed by PMF granulocytes analysis.

Of the sRNAs for which DE in PMF was detected both in CD34+ cells and in granulocytes, all resulted to follow the same trend also in PV or ET granulocytes (Data not shown). Considering the significance of the observed variation, we would like to mention that miR-10b-5p is down regulated both in PV and ET samples (with a significant p-value, not reaching significance when the p-value is adjusted), whereas miR-19b-3p and miR-543 are respectively down and up regulated only in ET. Regarding the moR-128-2, very downregulated in PMF CD34+ and granulocytes, it decreases, but at a lower extent, without reaching statistical significance, also in PV and ET granulocytes.

miR-10b-5p, downregulated both in PMF CD34+ and granulocytes, has been previously reported to be deregulated in breast cancer[42-44]and involved in chemoresistance related pathway[45] . It has been validated as downregulated in endometrial carcinoma[46], bladder cancer [47], in advanced stage of small cell carcinoma of the cervix (SCCC)[48] and in clear cell renal cell carcinoma (ccRCC) and its expression level has been also included in a linear model that capture the metastatic tumor signature and patient prognosis[49].

We found miR-29a-3p upregulated in patients CD34+ respect to controls. Han et al.[50]previously demonstrated that miR-29a is downregulated in hematopoietic progenitors respect to lineage-committed progenitors, including granulocytes. That means that its expression level grows along the committed lineage. In our validation set of granulocyte cells, we recognized miR-29a as significantly downregulated in patients compared to controls. It results therefore deregulated in the validations too,

but it reversed the direction of differential expression in the committed granulocytes of the validation set, respect to the HSCs of sequencing set. Assuming that miR-29a expression in controls follows the highlighted trend reported by Han et al., in our patient data it results highly expressed in HSC when it should have a low expression level, while it has a low expression level in committed granulocytes when it should be highly expressed. We can thus classify miR-29a as a deregulated small RNA, which expression is modulated along differentiation. Figure 5 shows miR-29a trend in different cell and sample types.

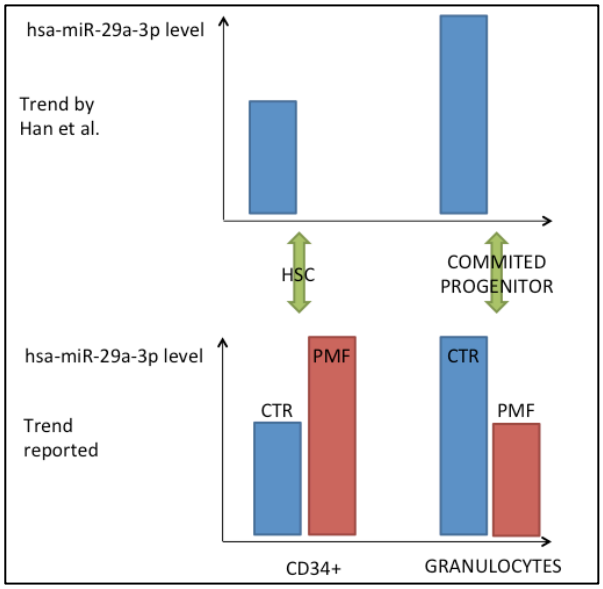

Figure 5. Schematic trend of miR-29a in CD34+ and granuocytes

Coherently with our finding, Han et al also showed that sustained expression of miR-29a-3p in mouse HSC/progenitors leads myeloid progenitors to self-renewal capacity, to biased myelopoiesis and myeloproliferative disorder that progress to acute myeloid leukemia. Additional data supporting miR-29a-3p deregulation comes from a previous study by Norfo et al.[37], in which miRNA expression profiling was obtained by Affymetrix miRNA 2.0 array analysis on a cohort of 42 PMF patients CD34+ cells and 31 healthy donors. In the same study, miR-29a-3p was found upregulated in CD34+ patients, in agreement with our findings in PMF CD34+ patients. miR-29a-3p upregulation in PMF CD34+ cells was validated by RT-PCR (with TaqMan probes) in an independent set of CD34+ cells from 10 PMF patients and 8 healthy subjects.

In the same study, also the upregulation of miR-379-5p, miR-543 and miR-19b-3p were validated in PMF CD34+. Our validations on granulocytes do not confirm a statistically significant upregulation of these miRNas in PMF, showing only a trend toward, but we are confident that the validation of Norfo et al. on CD34+ is solid, considering that they were carried out on our same cell type. Interestingly Norfo et al. validated microarray data using two sets of experiments, conducted on CD34+ cells and on granulocytes, that showed that PMF-specific variations of a few miRNAs are observed in CD34+ and not in granulocytes.

Indeed miR-486-3p was significantly downregulated in PMF granulocytes and upregulated in PMF in CD34+.

The 3'-moR-128-2, a newly annotated small RNA, results expressed in CD34+ and is not detected in PMF. It is also downregulated in PMF granulocytes respectively to controls.

Supplementary Figure 3 shows the expression profiles in considered RNA-seq samples, of all the detected sRNAs expressed from mir-128-1 and mir-128-2 loci: of them, the unique miRNA expressed from 128-1 hairpin is equally expressed in PMF and normal samples, 3'-moR-128-2 is highly expressed in normal stem cells and down regulated in PMF, whereas miR-128-3p (miR-128-2-3p) is weakly expressed.

In Figure 6 we show additional information regarding miR-128-3p and 3'-moR-128-2. The moRNA sequence is not contained in the canonical hairpin (Figure 6A). Thus, the moRNA probably derives from the processing of an alternative hairpin precursor. In Figure 6B we show the RNAfold predicted minimum free energy (MFE) folding structure of the canonical hairpin and of the longer one from which the moRNA is derived.

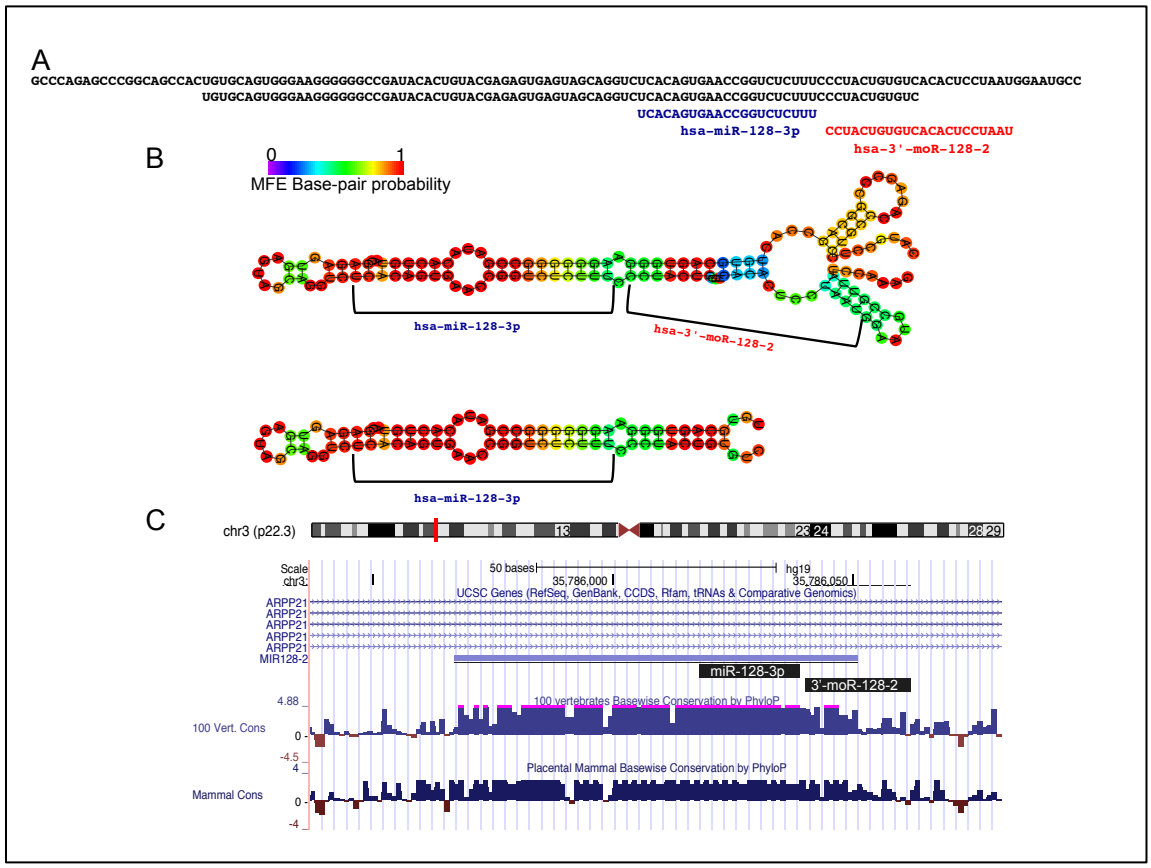

Figure 6. The 3'-moR-128-2 is produced by the precursor sequence of miR-128-3p. Panel A) shows that the moRNA is derived from a region of the primary miRNA sequence exceeding the canonical hairpin precursor sequence, and that the moRNA is not exaclty adjactent to the annotated miRNA. Panel B) shows the minimum free energy (MFE) folding structure predicted by RNAfold for the canonical hairpin seqeunce and for the longer one, from which the moRNA is probably derived. Panel C) shows that the considered small RNAs are conserved in evolution through vertebrates.

Figure 6C shows that mir-128 locus is inside an intron of ARPP21 gene, and displays that the genomic region encoding the 5' region of the moRNA is as conserved as that corresponding to the miRNA, according to Vertebrate and Mammals UCSC base-wise conservation score.

moRNA biological roles and mechanisms of function still deserve investigation. Very likely, moRNAs can function as miRNAs in post-transcriptional gene silencing, guiding RISC to complementary target mRNAs. This was first demonstrated by Umbach and colleagues, that used a luciferase-based indicator assay to demonstrate that a viral moRNA (moR-rR1-3-5p) has inhibitory activity against an artificial mRNA bearing a perfect target site [51, 52]. Beyond this proof of principle experiment, a recent study reported moRNA specific expression in human embryonic stem cells (hESCs; Asikainen et al., personal communication, 2015). In the same study, moRNA and miRNA transfection experiments and microarray quantification of gene

expression were conducted and identified gene silenced by moR-103a-2-3p, one of the most abundantly expressed moRNAs in hESCs, and by miR-103a.

In line with these previous studies, we assumed that 3'-moR-128-2 can act as a miRNA, and investigated its possible impact on target gene silencing and on specific pathways or biological processes.

We conducted a preliminary functional characterization of the possible biological role of sRNAs DE in PMF, by a double strategy.

First we investigated possible target genes and pathways of the group of validated DE sRNAs, considered as a whole.

Then, we focused more on one of the most novel elements emerged by our results, 3'-moR-128-2, to get specific insights on its possible functions in CD34+ and, in turn, in PMF disease.

Genes and pathways targeted by the sRNAs deregulated in PMF

Target predictions of miR-10b-5p, miR-19b-3p, miR-29a-3p, miR-379-5p, miR-543 and of moR-128-2 were performed by using two different programs, miRanda[53] and PITA[54], which implement orthogonal target prediction strategies, and for which the code availability allowed us to make custom predictions of possible miRNAs and moRNA target genes, by using as query sequences the identified isomiRs and isomoRs sequences.

Among different isomiRs detected for each considered miRNA, we considered the most expressed, even if it was different from the annotated sequence (Supplementary Table 6). We also considered those isomiRs that were significantly contributing to miRNA total expression, and which were differently expressed in patients respect to controls (t-test<0.05 and $|\log_2 FC| > 1$). Accordingly, both isomoRs were considered for moR-128-2.

A functional enrichment analysis, based on Reactome annotation maps, of targets predicted by both methods was obtained using a hypergeometric test.

Supplementary Table 6 includes details on considered sequences and on the number of identified target genes per sequence. It lists the significantly (p-value <= 0.05) enriched pathways, and the number of distinct genes represented in small RNAs

targets. We choose to include in the table only those pathways that appear enriched in targets of at least half of the small RNAs.

miRNA targets are enriched in many interesting pathways involved in tumor development and progression, as signaling by FGFR, DAP12 signaling and Oncogene Induced Senescence.

Human fibroblast grow factor receptors (FGFRs) are a family of four tyrosine kinase receptors (FGFR1–4) involved in a variety of cellular processes. They are indeed key regulators of fibrogenesis, embryogenesis, angiogenesis, metabolism, and many other processes of proliferation and differentiation[55, 56]. Deregulation of FGFR signaling has been observed in numerous tumors.[57, 58]

DAP10 is an immunoreceptor tyrosine-based activation motif (ITAM)-bearing transmenbrane adapter molecule and it is reported to be signaling partner of activating natural killer receptors. DAP12 complex to TREM-1 and MDL-1 receptors to form receptor complexes involved in macrophage differentiation[59] and apoptosis in M1 leukemia cells[60], significant monocytic activation of myeloid cell, calcium mobilization and inflammatory response[61, 62]. Its elevated expression levels are associates with enhanced cytotoxic characteristics in large granular lymphocyte leukemia[63].

Senescence is the stable cell growth arrest. Oncogene senescence (OIS) occurs when the activation of an oncogene is triggered, in this case it is termed oncogene-induced senescence. OIS acts as a barrier against tumour progression by driving stable growth arrest of cancer progenitor cells [64-66].

3'-moR-182-2

Intrigued by the striking expression pattern of the newly discovered 3'-moR-182-2 we looked in details into sequence, structure, expression and functional differences of 3'-moR-128-2 and miR-128-3p.

First we considered how sequence variants (isomiRs) of these two small RNAs relate to each other (Figure 7). For miR-128-3p, we identified 7 variants expressed in considered CD34+ samples: one exact isomiR, corresponding to the miRBase annotated mature form, and 6 “shorter or longer” variants (miR-128-3p-SL-1 to miR-128-3p-SL-6), whereas only 2 3'-moR-128-2 isomoR were found out (3'-moR-182-2-1 and 3'-moR-182-2-2)(Figure 7A).

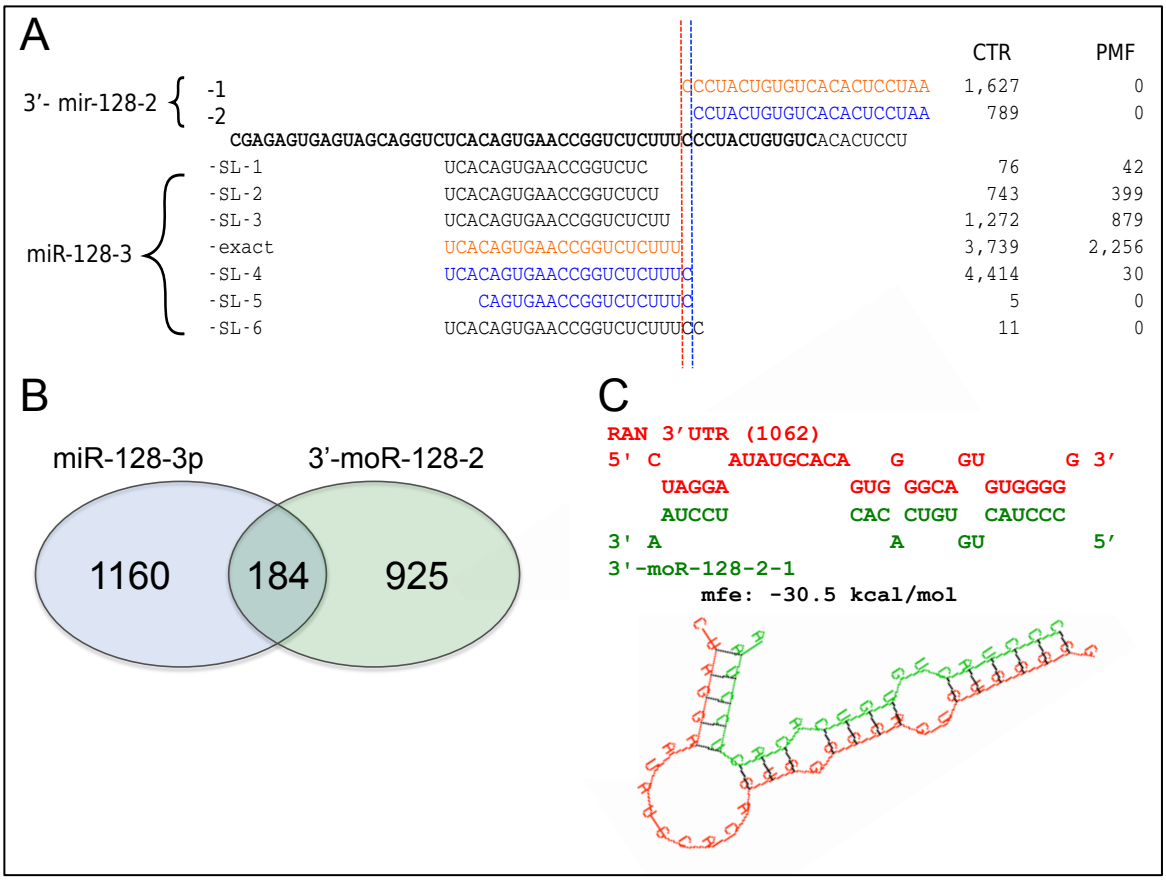

Figure 7. Origin, sequence variability, and relations beween 3'-moR-128-2 and the adjacent miR-128-3p. Panel A) shows that 3'-moR-128-2 and miR-128-3p map to the same locus and that both shows sequence variability (isomiRs and isomoRs). Both the major and the minoR isomoRs are found in normal CD34+ cells and not in PMF samples. Red and blue colors indicate isomiR and isomoR groups that can be produced with an unique sequence cutting sites. The most expressd isomoR is not associated to the corresponding most expressed isomiR. Moreover, expression levels, in CTR and PMF samples, of isomiRs and isomoRs are poorly correlated intragroup. These observations, point against the moRNA being simply a by-product of the miRNA biogenesis. A similar indication is given by the fact that some abundant isomiRs are not associated to detected isomoR sequences. Panel B) shows that 3'-moR-128-2 and miR-128-3p have different, poorly overlapping, sets of predicted targets. Panel C) 3-moR-128-2 can stably bind RAN 3'UTR, schematic representation of the binding.

According to the conservative hypothesis that interprets moRNA as byproducts of Drosha cleavage [67], one should expect comparable mean levels of miRNA and moRNA cognate variants (i.e. obtained from a single endonucleolytic cleavage cut) in control and PMF samples. 3'-moR-182-2-1, the most abundant isomoR, is expressed only in control samples and its only viable cognate partner is the miR-128-3p-exact variant that is highly and nearly equally expressed in control and PMF samples (with 3,739 and 2,256 normalized reads, respectively)(Figure 7A). This observation does not point in this direction. Neither do the poor correlation of expression levels, in CTR and PMF samples, of cognate isomiRs and isomoRs, the fact that some abundant isomiRs are not associated to detected isomoR sequences, the good conservation of

moRNA sequence (Figure 6C), and the previously reported observation that both isomoRs are not contained in the canonical hairpin (Figure 6A).

At the functional level, we predicted the targets of most expressed miR-128-3p isomiRs (miR-128-3p-exact, miR-128-3p-SL-2, miR-128-3p-SL-3, and miR-128-3p-SL-4) and targets of the two 3'-moR-182-2 isomoRs (3'-moR-182-2-1 and 3'-moR-182-2-2) assuming that they would act as miRNA, as indicated by the available experimental data[51].

We compared the union of predicted targets of miR-128-3p variants and the union of predicted targets of of 3-moR-128-2 variants, to understand how the moRNA function can be related to that of the cognate miRNA, as previously supposed (Asikainen personal communication). As shown in the Venn diagram in Figure 7B, only a small fraction of 3-moR-128-2 target genes, less than 17%, is putatively targeted also by at least one of the miR-128-3p isomiRs.

According to Reactome-based functional enrichments, performed as explained in the previous paragraph, different pathways are enriched in predicted targets of 3-moR-128-2 and of miR-128-3p. miR-128-3p targets are enriched in genes that are part of cellular pathways for the most part related to NGF, FGFR, ERBB4, ERBB2 signaling and transduction and to calcium ion homeostasis and signal transduction.

Targets of 3-moR-128-2 are enriched too in genes part of several, distinct, pathways related to cellular signaling in growth and proliferation as “Signaling by Notch“, “Signaling by ERBB4”, “Signaling by FGFR in disease” but also, quite interestingly, in genes part of the “Post-transcriptional silencing by small RNAs” and of the more general “Regulatory RNA pathways”. Remembering that 3-moR-128-2 is highly expressed in normal and not detected in PMF CD34+, it is worth notice that, 4 out of 7 genes of the “Post-transcriptional silencing by small RNAs” path, namely AGO1, AGO3, TNRC6A, and TNRC6B can be targeted by at least one isomoR of 3-moR-128-2. Moreover, both considered 3-moR-128-2 isomoRs can also target RAN, the RAS-related nuclear protein, member of the RAS Oncogene Family, that is required for RNA export from the nucleus. Table 4 shows regulatory pathways identified for the two expressed moR-128 variants.

Short RNA	**Variant**	**Regulatory RNA pathways targets**
3'-moR-128-2	3'-moR-128-2-1	AGO3, RAN, POLR2H
	3'-moR-128-2-2	AGO1, RAN, TNRC6A, TNRC6B

Table 4. regulatory RNA pathways of moR-128 predicted target

In principle 3-moR-128-2, where it is expressed, as in CD34+ hematopoietic stem cells, could affect the expression of genes important for the entire process of miRNA-based silencing. It can indeed target genes essential for post-transcriptional silencing both by translation repression, as AGO1/3, and by mRNA degradation, as TNRC6A/B. AGO1 and AGO3 are required for post-transcriptional translation repression activity; AGO1 is also involved in transcriptional silencing of promoters [68], and AGO3 is additionally putatively involved into the modulation of mature miRNA incorporation to the RISC complex, thus controlling the ratio between microRNA guide and passenger strand [69].

TNRC6A, and TNRC6B play a role in miRNA-dependent translation repression and endonucleolytic cleavage, by recruiting specific deadenylase complexes.

Moreover, 3-moR-128-2 can stably bind RAN 3'UTR. RAN is a multifunctional protein, involved in many processes and diseases. RAN controls cell cycle progression and it is a potential therapeutic target for treatment of cancers with activation of the PI3K/Akt/mTORC1 and Ras/MEK/ERK pathways [70].

Specifically in relation to the above mentioned findings, as known, RAN play a key role in RNA export from the nucleus and for the biogenesis of all miRNAs. Thus, RAN silencing by 3-moR-128-2 can impair pre-miRNA transportation to the cytoplasm and output a reduction of miRNA biogenesis, a situation someway similar to that documented by a recent study that identified, in B. mori, a virus-encoded miRNA that suppresses the host miRNA biogenesis exactly by targeting the host exportin-5 RAN cofactor [71].

Conclusion

In this study, we characterized miRNA and moRNA expression in CD34+ stem cells using massive small RNA-seq. The observed specificities in small RNAs expression of PMF CD34+ cells were subsequently confirmed considering granulocytes from PMF, PV and ET patients and from healthy controls. We thus provided new information regarding the possible role of miRNAs and new moRNAs in the disease. An interesting findings is the validated differentially expressed 3-moR-128-2 that we suppose could affect the expression of genes important for the entire process of miRNA-based silencing. It can indeed target genes essential for post-transcriptional silencing both by translation repression, as AGO1/3, and by mRNA degradation, as TNRC6A/B. It can also stably bind RAN 3'UTR that controls cell cycle progression and it is a potential therapeutic target for treatment of cancers with activation of the PI3K/Akt/mTORC1 and Ras/MEK/ERK pathways. Hopefully this information will help in the understanding the mechanisms that contribute to PMF pathogenesis and in formulate new targeted therapies.

References

1. Tefferi A and Vardiman JW. Classification and diagnosis of myeloproliferative neoplasms: The 2008 World Health Organization criteria and point-of-care diagnostic algorithms. Leukemia. 2007;22(1):14-22.

2. Vannucchi AM, Guglielmelli P, Tefferi A. Advances in Understanding and Management of Myeloproliferative Neoplasms. CA: A Cancer Journal for Clinicians. 2009;59(3):171-191.

3. Vannucchi AM. Management of myelofibrosis. ASH Education Program Book. 2011;2011(1):222-230.

4. Vannucchi AM. From Palliation to Targeted Therapy in Myelofibrosis. N Engl J Med. 2010;363(12):1180-1182.

5. Baxter EJ, Scott LM, Campbell PJ, et al. Acquired mutation of the tyrosine kinase JAK2 in human myeloproliferative disorders. The Lancet. 2005;365(9464):1054-1061.

6. Jones AV, Kreil S, Zoi K, et al. Widespread occurrence of the JAK2 V617F mutation in chronic myeloproliferative disorders. Blood. 2005;106(6):2162-2168.

7. James C, Ugo V, Le CouÃ©dic J, et al. A unique clonal JAK2 mutation leading to constitutive signalling causes polycythaemia vera. Nature. 2005;434(7037):1144-1148.

8. Kralovics R, Passamonti F, Buser AS, et al. A Gain-of-Function Mutation of JAK2 in Myeloproliferative Disorders. N Engl J Med. 2005;352(17):1779-1790.

9. Zhao R, Xing S, Li Z, et al. Identification of an Acquired JAK2 Mutation in Polycythemia Vera. J Biol Chem. 2005;280(24):22788-22792.

10. Klampfl T, Harutyunyan A, Berg T, et al. Genome integrity of myeloproliferative neoplasms in chronic phase and during disease progression. Blood. 2011;118(1):167-176.

11. Vannucchi AM, Lasho TL, Guglielmelli P, et al. Mutations and prognosis in primary myelofibrosis. Leukemia. 2013;27(9):1861-1869.

12. Vannucchi AM and Biamonte F. Epigenetics and mutations in chronic myeloproliferative neoplasms. Haematologica. 2011.

13. Mullally A, Lane SW, Ball B, et al. Physiological Jak2V617F expression causes a lethal myeloproliferative neoplasm with differential effects on hematopoietic stem and progenitor cells. Cancer Cell. 2010;17(6):584-596.

14. Chen E, Beer PA, Godfrey AL, et al. Distinct Clinical Phenotypes Associated with JAK2V617F Reflect Differential STAT1 Signaling. Cancer Cell. 2010;18(5):524-535.

15. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell. 2004;116(2):281-297.

16. Eichhorn S, Guo H, McGeary S, et al. mRNA Destabilization Is the Dominant Effect of Mammalian MicroRNAs by the Time Substantial Repression Ensues. Mol Cell. 2014;56(1):104-115.

17. Iorio MV and Croce CM. MicroRNA dysregulation in cancer: diagnostics, monitoring and therapeutics. A comprehensive review. EMBO Molecular Medicine. 2012;4(3):143-159.

18. Guglielmelli P, Tozzi L, Pancrazzi A, et al. MicroRNA expression profile in granulocytes from primary myelofibrosis patients. Exp Hematol. 2007;35(11):1708.e1-1708.e12.

19. Hussein K, Theophile K, Dralle W, Wiese B, Kreipe H, Bock O. MicroRNA expression profiling of megakaryocytes in primary myelofibrosis and essential thrombocythemia. Platelets. 2009;20(6):391-400.

20. Guglielmelli P, Tozzi L, Bogani C, et al. Overexpression of microRNA-16-2 contributes to the abnormal erythropoiesis in polycythemia vera. Blood. 2011;117(25):6923-6927.

21. Guo S, Lu J, Schlanger R, et al. MicroRNA miR-125a controls hematopoietic stem cell number. Proceedings of the National Academy of Sciences. 2010;107(32):14229-14234.

22. O'Connell RM, Chaudhuri AA, Rao DS, Gibson WSJ, Balazs AB, Baltimore D. MicroRNAs enriched in hematopoietic stem cells differentially regulate long-term hematopoietic output. Proceedings of the National Academy of Sciences. 2010;107(32):14235-14240.

23. Ooi AGL, Sahoo D, Adorno M, Wang Y, Weissman IL, Park CY. MicroRNA-125b expands hematopoietic stem cells and enriches for the lymphoid-balanced and lymphoid-biased subsets. Proceedings of the National Academy of Sciences. 2010;107(50):21505-21510.

24. Lu J, Guo S, Ebert BL, et al. MicroRNA-Mediated Control of Cell Fate in Megakaryocyte-Erythrocyte Progenitors. Developmental Cell. 2008;14(6):843-853.

25. Kumar MS, Narla A, Nonami A, et al. Coordinate loss of a microRNA and protein-coding gene cooperate in the pathogenesis of 5qâˆ’ syndrome. Blood. 2011;118(17):4666-4673.

26. Chen C, Li L, Lodish HF, Bartel DP. MicroRNAs Modulate Hematopoietic Lineage Differentiation. Science. 2004;303(5654):83-86.

27. Zhan H, Cardozo C, Raza A. MicroRNAs in myeloproliferative neoplasms. Br J Haematol. 2013;161(4):471-483.

28. Zhang L, Sankaran VG, Lodish HF. MicroRNAs in erythroid and megakaryocytic differentiation and megakaryocyteâ€“erythroid progenitor lineage commitment. Leukemia. 2012;26(11):2310-2316.

29. BÃ¡ez A, MartÃn-Antonio B, Piruat JI, et al. Gene and miRNA Expression Profiles of Hematopoietic Progenitor Cells Vary Depending on Their Origin. Biology of Blood and Marrow Transplantation. 2014;20(5):630-639.

30. Raghavachari N, Liu P, Barb JJ, et al. Integrated analysis of miRNA and mRNA during differentiation of human CD34+ cells delineates the regulatory roles of microRNA in hematopoiesis. Exp Hematol. 2014;42(1):14-27.e2.

31. Bruchova H, Yoon D, Agarwal AM, Mendell J, Prchal JT. Regulated expression of microRNAs in normal and polycythemia vera erythropoiesis. Exp Hematol. 2007;35(11):1657-1667.

32. Bruchova H, Merkerova M, Prchal JT. Aberrant expression of microRNA in polycythemia vera. Haematologica. 2008;93(7):1009-1016.

33. Vian L, Di Carlo M, Pelosi E, et al. Transcriptional fine-tuning of microRNA-223 levels directs lineage choice of human hematopoietic progenitors. Cell Death & Differentiation. 2013;21(2):290-301.

34. Su R, Lin H, Zhang X, et al. MiR-181 family: regulators of myeloid differentiation and acute myeloid leukemia as well as potential therapeutic targets. Oncogene. 2014.

35. Lin X, Rice KL, Buzzai M, et al. miR-433 is aberrantly expressed in myeloproliferative neoplasms and suppresses hematopoietic cell growth and differentiation. Leukemia. 2013;27(2):344-352.

36. Slezak S, Jin P, Caruccio L, et al. Gene and microRNA analysis of neutrophils from patients with polycythemia vera and essential thrombocytosis: down-regulation of micro RNA-1 and -133a. Journal of Translational Medicine. 2009;7(1).

37. Norfo R, Zini R, Pennucci V, et al. miRNA-mRNA integrative analysis in primary myelofibrosis CD34+ cells unveils the role of miR-155/JARID2 axis in abnormal megakaryopoiesis. Blood. 2014.

38. Zhan H, Cardozo C, Yu W, et al. MicroRNA deregulation in polycythemia vera and essential thrombocythemia patients. Blood Cells, Molecules, and Diseases. 2013;50(3):190-195.

39. Bortoluzzi S, Bisognin A, Biasiolo M, et al. Characterization and discovery of novel miRNAs and moRNAs in JAK2V617F-mutated SET2 cells. Blood. 2012;119(13):e120-e130.

40. Gaffo E, Zambonelli P, Bisognin A, Bortoluzzi S, Davoli R. miRNome of Italian Large White pig subcutaneous fat tissue: new miRNAs, isomiRs and moRNAs. Anim Genet. 2014;45(5):685-698.

41. Kent WJ. BLAT--the BLAST-like alignment tool. Genome Res. 2002;12(4):656-664.

42. Gee HE, Camps C, Buffa FM, et al. MicroRNA-10b and breast cancer metastasis. Nature. 2008;455(7216):E8-E9.

43. Ma L, Teruya-Feldstein J, Weinberg RA. Tumour invasion and metastasis initiated by microRNA-10b in breast cancer. Nature. 2007;449(7163):682-688.

44. Chan M, Liaw CS, Ji SM, et al. Identification of Circulating MicroRNA Signatures for Breast Cancer Detection. Clinical Cancer Research. 2013;19(16):4477-4487.

45. Ouyang M, Li Y, Ye S, et al. MicroRNA Profiling Implies New Markers of Chemoresistance of Triple-Negative Breast Cancer. PLoS ONE. 2014;9(5).

46. Tsukamoto O, Miura K, Mishima H, et al. Identification of endometrioid endometrial carcinoma-associated microRNAs in tissue and plasma. Gynecol Oncol. 2014;132(3):715-721.

47. Zaravinos A, Radojicic J, Lambrou GI, et al. Expression of miRNAs Involved in Angiogenesis, Tumor Cell Proliferation, Tumor Suppressor Inhibition, Epithelial-Mesenchymal Transition and Activation of Metastasis in Bladder Cancer. J Urol. 2012;188(2):615-623.

48. Huang L, Lin J, Yu Y, Zhang M, Wang H, Zheng M. Downregulation of Six MicroRNAs Is Associated with Advanced Stage, Lymph Node Metastasis and Poor Prognosis in Small Cell Carcinoma of the Cervix. PLoS ONE. 2012;7(3).

49. Wu X, Weng L, Li X, et al. Identification of a 4-microRNA Signature for Clear Cell Renal Cell Carcinoma Metastasis and Prognosis. PLoS ONE. 2012;7(5).

50. Han Y-, Park CY, Bhagat G, et al. microRNA-29a induces aberrant self-renewal capacity in hematopoietic progenitors, biased myeloid development, and acute myeloid leukemia. J Exp Med. 2010;207(3):475-489.

51. Umbach JL, Strelow LI, Wong SW, Cullen BR. Analysis of rhesus rhadinovirus microRNAs expressed in virus-induced tumors from infected rhesus macaques. Virology. 2010;405(2):592-599.

52. Bortoluzzi S, Biasiolo M, Bisognin A. MicroRNAâ€“offset RNAs (moRNAs): by-product spectators or functional players? Trends Mol Med. 2011;17(9):473-474.

53. Enright AJ, John B, Gaul U, et al. MicroRNA targets in Drosophila. Genome Biol. 2004;5(1):R1-R1.

54. Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E. The role of site accessibility in microRNA target recognition. Nat Genet. 2007;39(10):1278-1284.

55. Brooks AN, Kilgour E, Smith PD. Molecular Pathways: Fibroblast Growth Factor Signaling: A New Therapeutic Opportunity in Cancer. Clinical Cancer Research. 2012;18(7):1855-1862.

56. Tiong KH, Mah LY, Leong C. Functional roles of fibroblast growth factor receptors (FGFRs) signaling in human cancers. Apoptosis. 2013;18(12):1447-1468.

57. Katoh M and Nakagama H. FGF Receptors: Cancer Biology and Therapeutics. Med Res Rev. 2014;34(2):280-300.

58. Tan L, Wang J, Tanizaki J, et al. Development of covalent inhibitors that can overcome resistance to first-generation FGFR kinase inhibitors. Proceedings of the National Academy of Sciences. 2014.

59. Aoki N, Kimura S, Takiyama Y, et al. The Role of the DAP12 Signal in Mouse Myeloid Differentiation. The Journal of Immunology. 2000;165(7):3790-3796.

60. Aoki N, Kimura S, Oikawa K, et al. DAP12 ITAM Motif Regulates Differentiation and Apoptosis in M1 Leukemia Cells. Biochem Biophys Res Commun. 2002;291(2):296-304.

61. Gingras M, Lapillonne H, Margolin JF. TREM-1, MDL-1, and DAP12 expression is associated with a mature stage of myeloid development. Mol Immunol. 2002;38(11):817-824.

62. Bakker ABH, Baker E, Sutherland GR, Phillips JH, Lanier LL. Myeloid DAP12-associating lectin (MDL)-1 is a cell surface receptor involved in the activation of myeloid cells. Proceedings of the National Academy of Sciences. 1999;96(17):9792-9796.

63. Chen X, Bai F, Sokol L, et al. A critical role for DAP10 and DAP12 in CD8+ T cellâ€“mediated tissue damage in large granular lymphocyte leukemia. Blood. 2009;113(14):3226-3234.

64. Aird KM and Zhang R. Nucleotide metabolism, oncogene-induced senescence and cancer. Cancer Lett.

65. Grasso D and Vaccaro MI. Macroautophagy and the oncogene-induced senescence. Endocrinology of Aging. 2014;5.

66. Hills S and Diffley JX. DNA Replication and Oncogene-Induced Replicative Stress. Current Biology. 2014;24(10):R435-R444.

67. Ma H, Wu Y, Choi JG, Wu H. Lower and upper stem-single-stranded RNA junctions together determine the Drosha cleavage site. Proc Natl Acad Sci U S A. 2013;110(51):20687-20692.

68. Romero-Cordoba SL, Salido-Guadarrama I, Rodriguez-Dorantes M, Hidalgo-Miranda A. miRNA biogenesis: Biological impact in the development of cancer. Cancer Biol Ther. 2014;15(11):1444-1455.

69. Winter J and Diederichs S. Argonaute-3 activates the let-7a passenger strand microRNA. RNA Biol. 2013;10(10):1631-1643.

70. Yuen HF, Chan KK, Grills C, et al. Ran is a potential therapeutic target for cancer cells with molecular changes associated with activation of the PI3K/Akt/mTORC1 and Ras/MEK/ERK pathways. Clin Cancer Res. 2012;18(2):380-391.

71. Singh CP, Singh J, Nagaraju J. A baculovirus-encoded MicroRNA (miRNA) suppresses its host miRNA biogenesis by regulating the exportin-5 cofactor Ran. J Virol. 2012;86(15):7867-7879.

72. Azuma-Mukai A, Oguri H, Mituyama T, et al. Characterization of endogenous human Argonautes and their miRNA partners in RNA silencing. Proceedings of the National Academy of Sciences. 2008;105(23):7964-7969.

73. Fernandez-Valverde S, Taft RJ, Mattick JS. Dynamic isomiR regulation in Drosophila development. RNA. 2010;16(10):1881-1888.

74. Tan GC, Chan E, Molnar A, et al. 5' isomiR variation is of functional and evolutionary importance. Nucleic Acids Res. 2014;42(14):9424-9435.

75. Jaskiewicz L and Zavolan M. Dicer partners expand the repertoire of miRNA targets. Genome Biol. 2012;13(11).

76. Neilsen CT, Goodall GJ, Bracken CP. IsomiRs â€“ the overlooked repertoire in the dynamic microRNAome. Trends in Genetics. 2012;28(11):544-549.

77. Fukunaga R, Han B, Hung J, Xu J, Weng Z, Zamore P. Dicer Partner Proteins Tune the Length of Mature miRNAs in Flies and Mammals. Cell. 2012;151(3):533-546.

78. Morin RD, O'Connor MD, Griffith M, et al. Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. Genome Res. 2008;18(4):610-621.

79. Cloonan N, Wani S, Xu Q, et al. MicroRNAs and their isomiRs function cooperatively to target common biological pathways. Genome Biol. 2011;12(12).

80. Chan Y, Lin Y, Lin R, et al. Concordant and Discordant Regulation of Target Genes by miR-31 and Its Isoforms. PLoS ONE. 2013;8(3).

81. Langenberger D, Bermudez-Santana C, Hertel J, Hoffmann S, Khaitovich P, Stadler PF. Evidence for human microRNA-offset RNAs in small RNA sequencing data. Bioinformatics. 2009;25(18):2298-2301.

Chapter 2

In order to have more awareness in applying normalization methods when managing RNA-seq data of small RNA dataset, we evaluated the performance of normalization algorithms formulated for long RNAs, applied to human small RNA datasets. We simulated multiple matrixes with a controlled number of differentially expressed elements. We chose five normalization methods among the most cited and widespread, implemented in R packages: DESeq, edgeR, Quantile, NBPSeq, TBT. Each algorithm is based on different hypothesis on statistical shape and characteristics of data and we tested their impact on the downstream analysis in a differential expression test. To quantify normalization algorithms performances we calculated ROC curves and AUC curves. In this way we could compare the power of discovery differentially expressed small RNAs under different dataset characteristic, even when the assumption each method makes were violated. ROC curves showed that algorithms do not perform significantly differently applied to the same simulated scenarios and we are not able to definitively prefer a normalization algorithm as the best. The most successful situation for all the algorithm was when small RNAs strongly differentiate between the two groups. That coincides with a simulation scenario with the widest generated mean fold change. All the algorithms produced a high false positive rate and AUC curves showed us only small differences in the performances. We are far from reaching the consensus on the best normalization algorithm and there is still room to improve normalization methods for RNA-seq analysis.

Normalization impact on small RNA-seq data

Saccoman C.[1], Bisognin A.[1], Romualdi C.[1], Bortoluzzi S.[1]

1-Department of Biology, University of Padova, via G. Colombo 3, 35131, Padova, Italy

Abstract

RNA-seq technologies are useful for a wide range of biological investigations. While providing the highest throughput and the biggest discovery potential than ever, count data analysis is getting always more challenging. Many normalization methods have been developed to manage massive RNA-seq datasets and draw the information encrypted. All methods rely on statistical assumptions on data characteristics but no one of them has been specifically full-fledged for small RNAs. We comprehensively evaluated five commonly used normalization methods to pinpoint a procedure to perform a robust RNA-seq analysis. We thus considered DESeq, edgeR, Quantile, NBPSeq and TBT algorithms.

We simulated a huge number of small RNAs dataset with controlled characteristics in order to generate 9 different testing scenarios. With the aim of constructing a scenario the more realistic as possible, we estimated the distribution parameter of small RNA profiles from a real small RNAs dataset. We paid attention to choose a numerous real dataset to have an estimated value as representational of the reality as possible. We introduced in the simulated distribution a fixed number of differentially expressed elements and we applied five normalization methods. We evaluated the impact of the algorithms as the ability to recognize differentially expressed elements, conducting a differential expression analysis to all the normalized datasets.

Our results show that there is not yet a performing algorithm to use and there is an impelling need for more powerful normalization methods. Appropriate statistical and computational methods could improve the accuracy of results that still include a high number of false positive elements.

Introduction

High-throughput technologies are today in common use in biology. In the last decade we witnessed a very quick transformation in the methods for studying the transcriptome, offering a growing spectrum of applications. RNA-seq technology is the innovative element and since the very first papers in which it was applied[1-3], the potentials to understand the transcriptome have been appreciated. The evolution started from the relative low throughput sequence-based approach of Sanger sequencing of cDNA or EST libraries[4-7] and led to tag-based methods that improved the throughput, as serial analysis of gene expression (SAGE) [8, 9], cap analysis of gene expression (CAGE) [10-12] and massively parallel signature sequencing (MPSS) [13-15]. These technologies were still expensive and the true revolutionary tool for transcriptomics was the advent of RNA-seq[16], performed by the Illumina technologies[17-21], the Applied Biosystems SOLiD[22] and by Roche 454 Life Science[23-25]. The use of these new technologies slushed the sequencing costs and improved the throughput, opening the doors to a wide application field. Examples of the use of NGS are chromatin immunoprecipitation coupled to sequencing (ChIP-seq), whole genome genotyping, genome wide structural variation, de novo assembling and re-assembling of genome, mutation detection and carrier screening, detection of inherited disorders and complex human diseases, paired ends and genomic captures, sequencing of mitochondrial genome, personal genomics[26] and post-transcriptional gene regulation[27].

Regarding transcriptome characterization, RNA-seq has a huge discovery potential that previous microarray technology or PCR missed. Indeed, it does not depend on prior knowledge about genome or transcriptome sequence. In addition, differently from microarray technology, RNA-seq is not affected by background noise nor by signal saturation signal[1], it has a wide dynamic range of expression estimations, enabling the detection of weakly expressed genes[27]. Datasets produced are so heavily large and complex that the information contained is not clear-cut. Millions of short reads are produced from an RNA-seq experiment and a computational pipeline is always necessary to manage raw-data and to draw out information. Processing methods to extract and to release the information are pivotal. As well elucidated by Oshlack at al. [28] [28], a typical pipeline is organized in three main different steps: 1) Reads mapping to reference genome or transcriptome; 2) Summary of mapped reads at gene-level or transcriptome-level

depending on the experiment; 3) Data normalization, to compare different biological samples. The previous steps are always preliminary respect to further and more specific applications of methods that are determined by the experiment aim and scientific question of interest. A common issue in RNA-seq experiments is to determine differentially expressed (DE) elements among compared samples, generally two or more conditions or treatments. The goal is hence to identify genes or small RNAs that have changed significantly across samples and that can characterize distinct experimental groups of samples. In these specific applications of RNA-seq experiment, the normalization step is critical. It affects the good performance of the issue and DE genes could not actually be recognized due to an inadequate normalization[29-31]. Its aim is to correct for technical bias that affect data and make multiple samples comparable, while maintaining true biological signal[32]. Bias have impact upon between-sample distributional differences in read counts[29, 33]. These bias are due to differences in sample preparation, in library size (or in sequencing depth) [1], in variable GC contents[34], in gene length or in relative abundances of the genes[35, 36]. Despite many efforts to develop and evaluate normalization methods, there is not a recognized gold standard procedure. Many statistical methods have been formulated and discussed for microarray data analysis[37]. Due to the different principle of functioning upon which RNA-seq and microarray rely, and to deep differences in the distributions of expression data obtained with the two methods, microarray normalization algorithms cannot be applied to RNA-seq data. Microarrays record intensities as continuous measurements, assumed to follow a logarithmic or Gamma distribution[38-40], while RNA-seq data are count, and thus discrete, values. Many algorithms have been devised to achieve normalisation of RNA-seq-derived long RNA expression data but no one has been developed specifically for small RNA. There are only few statistical studies evaluating normalization methods on mRNA-seq[29, 35, 41-44]. Only one focus on normalization methods applied specifically to small RNAs dataset[45] but Zhou et al. [32] took a dig at it. All the procedures that deal with undesired variations among samples make assumption about true shape of data, in order to correct for differences in data shapes respect to the assumption. The most commonly used statistical models for RNA-seq are Negative Binomial and Poisson distributions[46-49], while more frequently they used a beta-binomial distribution[50]. It is also accepted that the fraction of deregulated mRNA respect to the total mRNAs expressed is negligible and that the deregulation is equally arranged between up-regulation and down-regulation[41, 49]. mRNA data and small RNA

data are somehow different in their expression and it not guarantee that normalization methods conceived for mRNA fit also for small RNA data. More specifically microRNAs (miRNA) are small RNA ~22 nt long[51], post-transcriptional regulator of gene expression[52]. Their deregulated expression is implicated in tumour onset and progression[53-55] and recognizing differentially expressed miRNAs can thus be crucial. Although many relevant normalization methods for RNA-seq have been generated, there is no guidance about how each algorithm impacts on the downstream analysis of microRNA data, as differentially expressed elements detection. The aim of this study is to evaluate normalization effect on DE downstream analysis in microRNA datasets.

Materials and Methods

In this section we describe the normalization methods considered, the real dataset used and the statistical model adopted in the simulation model. We performed a systematic analysis in order to test the robustness of normalization algorithms when their assumptions on data distribution are violated.

Normalization methods

We chose five normalization methods among the most cited and widespread, implemented in R packages: DESeq[56], edgeR[57, 58], Quantile[59], NBPSeq[60], TBT[61]. Each algorithm is based on different hypothesis on statistical shape and characteristics of data and we briefly describe their assumptions.

DESeq

It considers RNA-seq count data to follow a negative binomial distribution of witch parameters are variance and mean. It differentiates from other normalization methods because of its data-driven relationships of variance and mean. DESeq is based on the hypothesis that the majority of genes are not differentially expressed and that the proportion of up-regulated and down-regulated genes is equal. Differentially expressed elements impact on statistical data distribution is then negligible and they do not hence change the assumed regular negative distribution. It is implemented in the DESeq Bioconductor package and is easily applicable for R programming language users.

edgeR

As DESeq, edgeR methods considers data to have a negative binomial distribution and the most part of it to be not DE. This information is taken into account to model the overdispersion parameter relative to the Poisson distribution, and uses a conditional weighted likelihood to moderate the level of overdispersion across genes. To achieve the moderation, they share information over all reads and the most evident effect is to stabilize dispersion estimation in small samples. The normalization methods implements the Trimmed Mean of M values (TMM) [31] method to calculate a scaling factor as a weighted trimmed mean of the log ratios between two classes of samples.

Quantile

The Quantile normalization method was developed for microarray data and it has been later applied to RNA-seq count data too. Its aim is to make two data distribution identical in their statistical properties and it works modifying data to the mean of the corresponding ranked values. It makes hypothesis neither on the proportion of up regulated and down regulated genes nor on a specific statistical distribution shape. It just assumes data to have the same distribution across samples. It is implemented in the R Bioconductor Limma package.

NBPSeq

NBPSeq has been formulated for RNA-Seq data normalization of DE analysis experiments. Di et al.[60, 62] claim this method to be a parameterized negative binomial distribution based. It was derived from Robinson and Smyth algorithms[57] and it is asserted to solve edgeR inappropriate estimation of data overdispersion. Di et al model has indeed an additional parameter to allow the dispersion parameter to depend on the mean. Their parametric method complements nonparametric regression approaches for modelling the dispersion parameter.

TBT

TBT is developed by Kadota et al. and consist of a double normalization step of Trimmed Mean of M values (TMM) [31] algorithm alternated by the baySeq procedure[62]. They start normalizing data with a first run of TMM normalization. Then they estimate the percentage of DE using the empirical Bayesian method implemented in baySeq and they

exclude the corresponding DE for the last step that is a second TMM normalization. In that way they eliminate DE elements for TMM normalization that have been demonstrated to shift the median log-ratio of data from the expected zero mean[61]. In this manner they do not need to make assumption on the proportion of up/down regulated elements. On the other hand, they consider a negative binomial statistical model for RNA-seq count data.

Real Dataset

We downloaded and considered a real data set from the Sequence Read Archive (SRA) database. It is a microRNA high throughput sequencing dataset of 63 head and neck squamous cell carcinoma, submitted by the French National League Against Cancer - Research Dept. It is a part of an integrative study of the Cartes d'Identité des Tumeurs (CIT) project (Accesion Study ERP001908).

We chose this public dataset because of its large number of samples of Human tissue, belonging all to the same biological type, all sequenced with high depth. We indeed needed the more samples as possible in order to have the more representative statistical population of small RNA expression profiles. In the 63 high throughput samples we found 970 small RNAs expressed, more precisely annotated microRNAs, new microRNAs and new moRNAs. We assume data to have a negative binomial distribution with an overdispersed variance. The head and neck tumour count data allowed us to make a realistic estimation of distribution parameters of a miRNA expression profile.

Comparison procedure

The aim was to evaluate the performance of normalization algorithms formulated for long RNAs, applied to human small RNA datasets. We simulated multiple matrixes with a controlled number of differentially expressed elements, keeping track of them with a flag on data. We then measured the impact of normalization methods as the ability to recognize differentially expressed elements in a differentially expressed analysis, after the normalization step. To this aim we applied the normalization methods and performed the same differential expression analysis to all the normalized data. The analysis had identified some elements as differentially expressed and, knowing the truth about the nature of elements, we easily determined how many true positive, true negative, false

positive and false negative elements each test produced, when applied to a datasets normalized with a specific method.

From these values we derived measures of the goodness of normalization algorithms in maintaining the intrinsic information the data contain. To quantify normalization algorithms performances we calculated ROC curves and AUC curves. In this way we compared the power of discovery differentially expressed small RNAs under different dataset characteristic, even when the assumption each method makes were violated.

Simulation model

In order to significantly assess the impact of the normalization methods on small RNAs downstream analysis, we simulated many groups of small RNAs dataset, each made up of 1000 simulated datasets of 1000 small RNAs expression valules (Ngenes=1000) in 10 normal samples compared against 10 treated samples (Nsamples=20). Each group of datasets has specific and controlled statistical distribution of small RNA expression values.

In the "head and neck squamous cell carcinoma" small RNAs dataset, we detected 970 small RNAs (Ngenes=970) in the 63 samples (Nsamples=63). The dataset can thus be represented as a matrix 970x63, with sample values on the columns and gene values across samples on the rows. We started from the evaluation of the mean and the dispersions of small RNA expression values across samples, drawing two vectors of 970 values (970 means and 970 dispersions). To this aim, we used both a customized version of the Bioconductor DESeq package and the standard Bioconductor edgeR package.

The function "estimateTagwiseDisp.R" implements the empirical Bayes strategy proposed by Robinson and Smyth[57] for estimating the tagwise negative binomial dispersions. The empirical Bayes method is based on a weighted conditional maximum likelihood.

In the DESeq package, the dispersion is considered as the square of the coefficient of biological variation, and additive weighted component of the variance, together with the uncertainty in measuring a concentration by counting reads. The function "estimateDispersions" performs three steps: it first estimate a temporary dispersion value for each gene, then it fit a curve for each group and then it assign a final estimated dispersion value. It implicitly normalizes data and we modified it in order to pass over this step.

In Figure 1 we compared the estimated dispersion

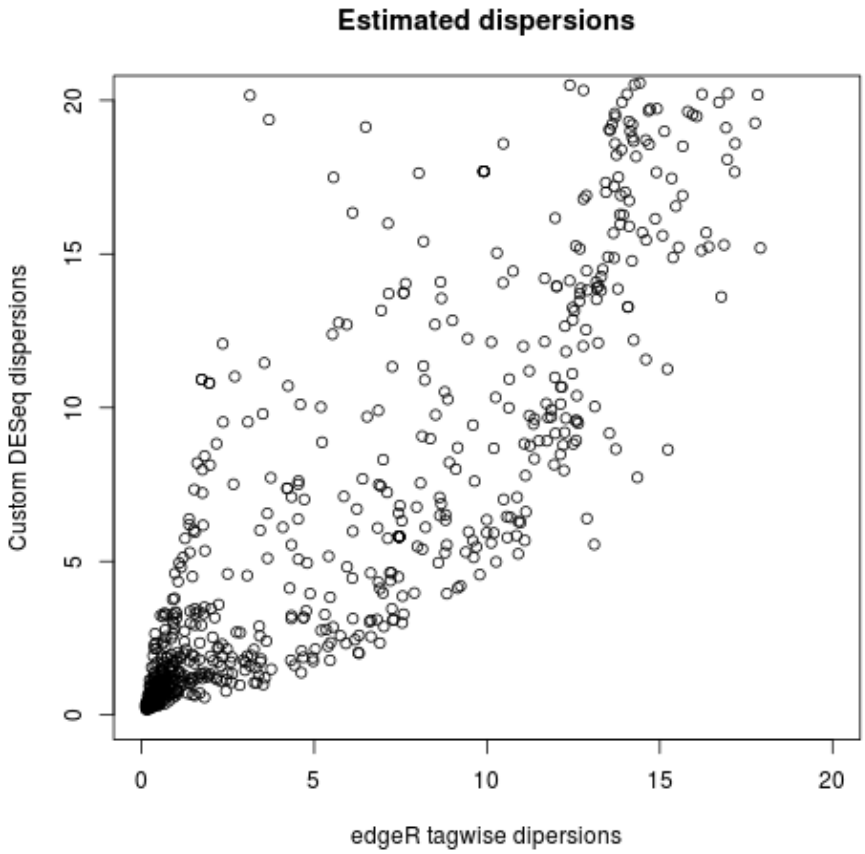

Figure 1. Comparison of estimated dispersion with different methods, a customized version of DESeq and the edgeR tagwise dispersion.

We then fitted the relationship between all the couples mean-variance. It is a characteristic parameter of small RNAs dataset[56, 58]and we considered the fitted distribution function rather than directly the mean-variance couples values, in order to be able to sample as many couples mean-variance as we wanted. We therefore needed a multiple datasets owing the same characteristic to have a representative population and give a statistical significance to the results. We chose the custom DESeq values as estimated variance because we could more easily fit the relation between data dispersion and mean. The fitted variance-mean relationship is represented in Figure 2.

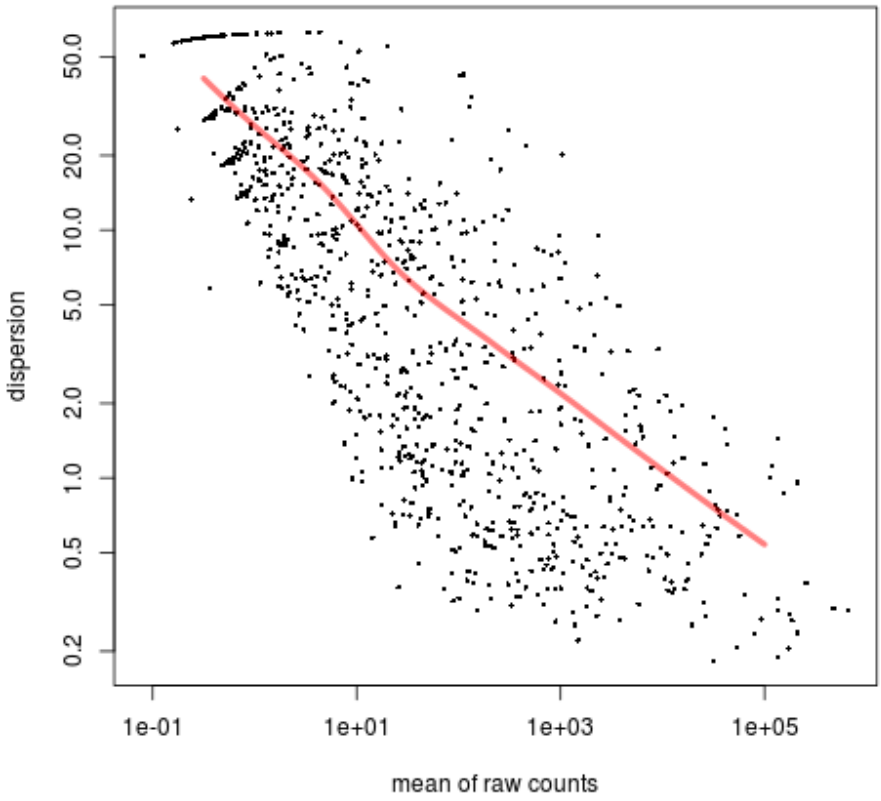

Figure 2. Fitted relation of estimated mean and dispersion values

To create a matrix 1000x20 (Ngenes=1000, Nsamples=20) with realistic expression values, we sampled 1000 times from the quartile of the distribution of the estimated means. Having estimated the relation with dispersion, for every sampled mean value, we calculated the related dispersion, finding 1000 couples of mean and dispersion values, corresponding to the parameters of 1000 small RNAs expression values of the assumed binomial negative distributions. Having fixed 1000 corresponding distributions, we sampled from each distribution 20 values, drawing a matrix 1000x20. We repeated the sampling of 20 expression values 1000 times, obtaining 1000 matrixes of 1000x20.

After that, we have always considered the same 1000 matrixes 1000x20 and we always introduced a 20% of elements differentially expressed in one of the two classes of 10 samples, modifying the 20% of small RNA expression values in 10 of the 20 samples. We implicitly assume that differentially expressed elements are in a relative small number, more specifically the 20% of the total. To choose which small RNA should be differentially expressed we randomly sampled 200 indexes of the matrix, for each matrix.

Every normalization method assumes data to respond to some specific statistical characteristics but actually, in most cases, we do not know which is the truth about data. To this aim we tuned the proportion of up/down regulated small RNA (respectively upregulated-downregulated: 50%-50%, 70%-30%, 90%-10%) and the mean extent of the shift from normal values, measured as fold change (FC: 2, 4, 8).

The combination of different values of the parameters FC and up/down differentiated our dataset in 9 groups of 1000 matrixes each.

Each count data matrix was then normalized using all the normalization algorithms chosen: DESeq, edgeR, quantile, NBPSeq, TBT.

Differential expression analysis

To assess the impact of normalization methods we performed a differential expression analysis for each normalized matrix.

To test which small RNAs were recognized as differentially expressed, we proceeded in a similar way the "nbinomTest" DESeq package does. Unlike it, we did not normalize data in the testing for differentially expressed genes. We indeed similarly calculated the mean expression value for each small RNA, both considering all the classes and individually for every class of samples. Then we computed the fold change for each small RNA, as the logarithm to basis 2 of the ratio for the first to the second condition values, and a t-test to get the statistical significance for the change. Since we worked with matrixes of 1000 miRNAs, it is necessary to control the false discovery rate (FDR). Raw P-values were thus adjusted for multiple testing with the Benjamini-Hochberg procedure[63].

We considered as differentially expressed all the small RNAs with an adjusted P-value < 0.05.

Results and discussion

We aimed to investigate the impact of the five chosen normalization methods on small RNA downstream analysis. All the normalization methods were developed for count data of long RNA transcripts. Small RNA dataset exhibit a negative binomial distribution with an overdispersed variance. In order to measure the performances of different normalization algorithms we simulated different population of small RNA count data. All the populations have the same dimensions: they are all composed by 1000 elements and each of it is a matrix of count expression values. The matrixes are all made up of two groups of samples of 10 members each, the groups corresponding to the classes we compare in the differential analysis. For each member we measure the expression values of 1000 small RNAs.

All the populations have the 20% of differentially expressed small RNAs in one of the two groups, with controlled characteristics. They indeed vary in the proportion of up/down regulated and in the mean fold change. The ratio of up/down regulated small RNA let us investigate the robustness of the normalization methods respect to the assumption that there are the same proportion of up/down regulated small RNAs. Introducing a high mean fold change in the differential expressed elements, we can analyse differences in normalization methods in presence of high count small RNAs.

The goal of a differential expression analysis is to assess whether two groups of sample belong (null hypothesis) or not (reject the null hypothesis) to the same population, in that sense the test is a binary classifier. Formally an instance is mapped to one element of the set {p,n} of positive and negative class labels. A classifier is a mapping from instances to predicted classes. We label as positive all the instances that belong to different population (differentially expressed elements) and negative all the elements that confirm the null hypothesis of belonging to the same population.

The performance of a binary classifier can be measured with the power of the test: it's the test's probability to correctly rejecting the null hypothesis, called also "sensitivity", "recall" or "true positive rate" (TPr). In our scenario it corresponds to the correct classification of a small RNA as differentially expressed. We are also interested in the false positive rate (FPr), or "false alarm rate", that means the probability of error in rejecting the null hypothesis. In other word, how often we classify as differentially expressed small RNAs that are not. Testing for differentially expressed small RNAs in a simulated scenario where we know the truth about belonging to differentially expressed

elements or not, let us evaluate the effective false positive rate and false positive rate in a representative situation of the reality.
We proceeded with a first qualitative evaluation of the impact of normalization methods, drawing Receiver Operating Characteristics (ROC) curves. ROC graph shows TPr on Y axis and FPr. Every point of the graph correspond to a pair of TPr and FPr varying the threshold of the scoring that we use to assign a label {p,n} to an instance. We can see all the curves in Figure 3.

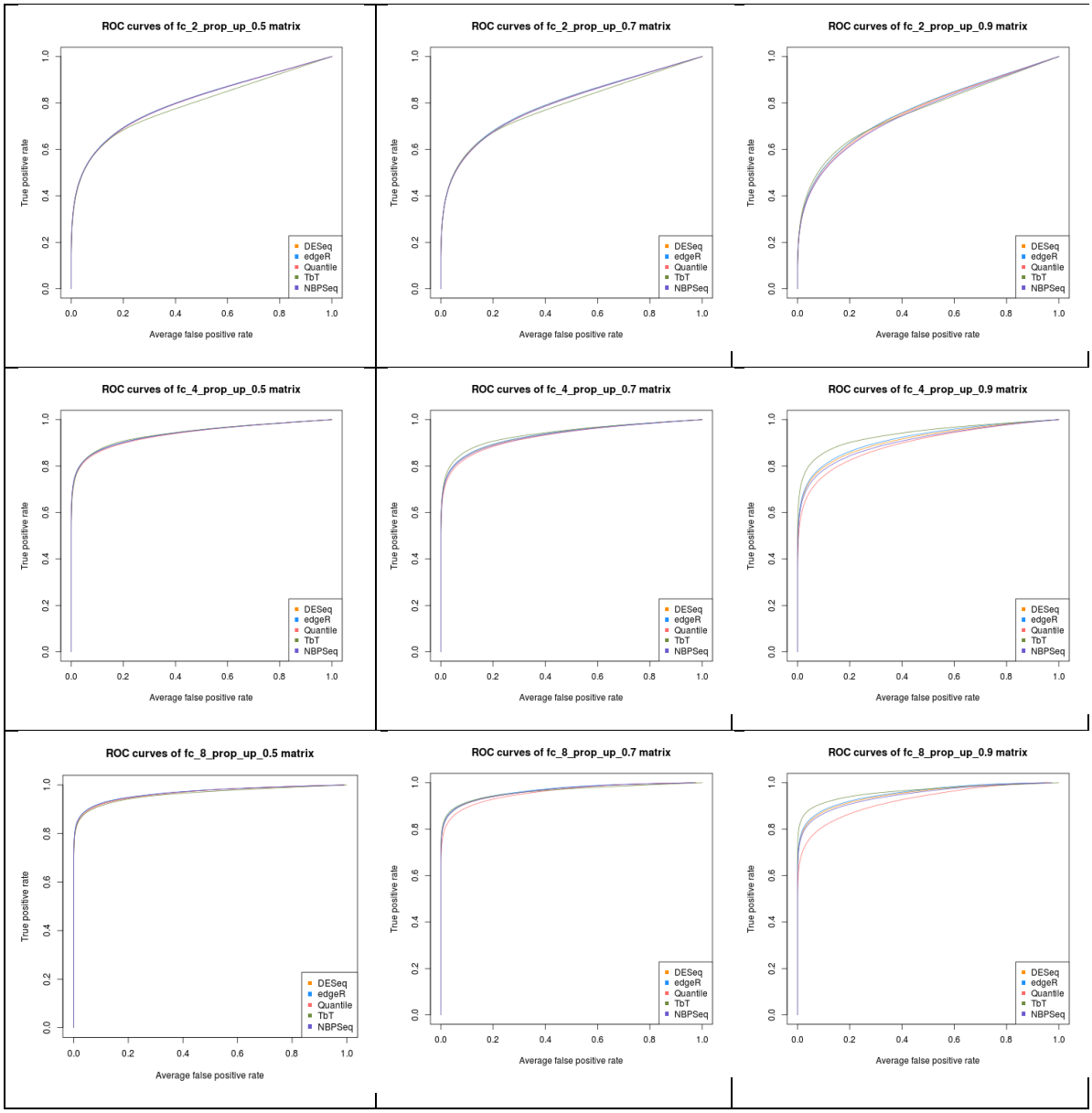

Figure 3. ROC graph in 9 different scenarios, ROC graph shows TPr on Y axis and FPr. Every point of the graph correspond to a pair of TPr and FPr varying the threshold of the scoring.

We can see that the algorithms do not perform significantly differently applied to the same scenarios and we are not able to definitively prefer a normalization algorithm as the best. They all have the same trends, varying the fold change value and the ratio between up regulated and down regulated small RNAs. More specifically we can see that an unbalanced proportion of up-regulated small RNAs respect to the down regulated small RNAs worsen the performance. As previously mentioned, an imbalance of up- and down-regulated miRNAs is very likely to occur in many real word datasets.

Considering a up-down regulated ratio of 50%, independently from the fold change value, the methods led to the same condition. The choice of the normalization procedure in that scenario does not weigh on DE recognition.

Fold change differentiate a little how the normalization methods have a bearing on DE test and we decided to quantify the performance by comparing the values of areas under the ROC curves. To compare classifier we may want to reduce ROC performances to a single scalar value. It is commonly accepted to calculate the Area Under the ROC Curve (AUC) that is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher then a randomly chosen negative instance. The higher the AUC value is and better is the classifier performance.

In Tables 1A, 1B, 1C are represented all the mean values of the areas under the curves that we draw for each TPR-FPR pair, varying the p-values threshold.

The most successful situation is when small RNAs strongly differentiate between the two groups. That coincides with a simulation scenario where a mean fold change is equal to 8. Considering that the attainment falls short when the proportion of up regulated small RNAs grows, it is not surprising that the best performance is achieved in case of fold change equal to 8 and 50% of proportion. The normalization algorithm that allows the best compromise between TPr and FPr is edgeR but it does not move far away from other methods.

To have a dimension of the differences, we arbitrary considered as threshold a p-value adjusted of 0.05 and in Table 2A, 2B, 2C we gathered a complete calculation of TPr and FPr for each of the 9 scenarios. We can see that all the normalization algorism produce a high false positive rate and this strengthen the needs of experimental validation in case of biological data.

	FC2UP0.5	**FC2UP0.7**	**FC2UP0.9**
auc_deseq	0,8059	0,7974	0,7677
auc_edger	0,8060	0,7981	0,7693
auc_quantile	0,8037	0,7948	0,7631
auc_tbt	0,7918	0,7861	0,7638
auc_nbpseq	0,8044	0,7944	0,7590

Table 1A AUC (area under ROC curves) values of FC=2 and proportion of up regulated=50%, 70%, 90%

	FC4UP0.5	FC4UP0.7	FC4UP0.9
auc_deseq	0,9386	0,9331	0,9109
auc_edger	0,9391	0,9341	0,9165
auc_quantile	0,9362	0,9284	0,8945
auc_tbt	0,9397	0,9391	0,9363
auc_nbpseq	0,9380	0,9305	0,9043

Table 1B AUC (area under ROC curves) values of FC=4 and proportion of up regulated=50%, 70%, 90%

	FC8UP0.5	FC8UP0.7	FC8UP0.9
auc_deseq	0,9674	0,9647	0,9505
auc_edger	0,9683	0,9664	0,9538
auc_quantile	0,9664	0,9585	0,9231
auc_tbt	0,9622	0,9625	0,9613
auc_nbpseq	0,9678	0,9643	0,9458

Table 1C AUC (area under ROC curves) values of FC=8 and proportion of up regulated=50%, 70%, 90%

	FC2UP05		FC2UP07		FC2UP09	
	TPR	**FPR**	**TPR**	**FPR**	**TPR**	**FPR**
deseq	0,3505	0,8200	0,2766	0,7789	0,4088	0,7800
edger	0,3656	0,7850	0,2396	0,7100	0,4492	0,8050
quantile	0,2638	0,7650	0,3717	0,7588	0,3292	0,6750
tbt	0,1691	0,6133	0,1100	0,6111	0,1397	0,6270
nbpseq	0,3304	0,7850	0,3108	0,7286	0,3922	0,7550

Table 2A. TPR, FPR of FC=2 and proportion of up regulated=50%, 70%, 90% p-value adjusted = 0.05.

	FC4UP05		FC4UP07		FC4UP09	
	TPR	**FPR**	**TPR**	**FPR**	**TPR**	**FPR**
deseq	0,4313	0,9296	0,4593	0,9500	0,5859	0,9400
edger	0,5670	0,9600	0,5788	0,9500	0,6349	0,9700
quantile	0,4020	0,9548	0,5144	0,9246	0,6851	0,9300
tbt	0,2382	0,9293	0,2434	0,9184	0,3945	0,9694
nbpseq	0,3497	0,9450	0,4862	0,9300	0,5551	0,9347

Table 2B. TPR, FPR of FC=4 and proportion of up regulated=50%, 70%, 90% p-value adjusted = 0.05.

	FC8UP05		FC8UP07		FC8UP09	
	TPR	**FPR**	**TPR**	**FPR**	**TPR**	**FPR**
deseq	0,3108	0,9700	0,3852	0,9450	0,7009	0,9949
edger	0,3484	0,9700	0,5038	0,9849	0,6270	0,9850
quantile	0,3568	0,9450	0,5707	0,9747	0,7997	0,9849
tbt	0,2808	0,9444	0,2831	0,9646	0,2886	0,9596
nbpseq	0,1541	0,9750	0,3509	0,9497	0,7186	0,9450

Table 2C. TPR, FPR of FC=8 and proportion of up regulated=50%, 70%, 90% p-value adjusted = 0.05.

Conclusions

In this study, we systematically evaluated the impact of five normalization algorithms formulated to manage RNA count data, when applied to small RNA count data.

To this aim we simulated nine different realistic scenarios of 1000 data set each, each scenario differing in the proportion of up and down regulated small RNAs and in the extent of deregulation.

We have not been able to identify a method that was robust respect to the diverse characteristics of the datasets. All the algorithms produce a high false positive rate and AUC curves show us only small differences in the performances.

We are far from reaching the consensus on the best normalization algorithm and there is still room to improve normalization methods for RNA-seq analysis.

References

1. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5**, 621-628 (2008).

2. Nagalakshmi, U. *et al*. The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. *Science* **320**, 1344-1349 (2008).

3. Lister, R. *et al*. Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis. *Cell* **133**, 523-536 (2008).

4. The Status, Quality, and Expansion of the NIH Full-Length cDNA Project: The Mammalian Gene Collection (MGC). *Genome Res*. **14**, 2121-2127 (2004).

5. Boguski, M. S., Tolstoshev, C. M. & Bassett, D. E. Gene discovery in dbEST. *Science* **265**, 1993-1994 (1994).

6. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences* **74**, 5463-5467 (1977).

7. Sanger, F. & Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol*. **94**, 441-448 (1975).

8. Velculescu, V. E., Zhang, L., Vogelstein, B. & Kinzler, K. W. Serial Analysis of Gene Expression. *Science* **270**, 484-487 (1995).

9. Harbers, M. & Carninci, P. Tag-based approaches for transcriptome research and genome annotation. *Nature Methods* **2**, 495-502 (2005).

10. Kodzius, R. *et al*. CAGE: cap analysis of gene expression. *Nature Methods* **3**, 211-222 (2006).

11. Nakamura, M. & Carninci, P. Cap analysis gene expression: CAGE]. *Tanpakushitsu Kakusan Koso.Protein, Nucleic Acid, Enzyme* **49**, 2688-2693 (2004).

12. Shiraki, T. *et al*. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proceedings of the National Academy of Sciences* **100**, 15776-15781 (2003).

13. Brenner, S. *et al*. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol*. **18**, 630-634 (2000).

14. Peiffer, J. A. *et al*. A spatial dissection of the Arabidopsis floral transcriptome by MPSS. *BMC Plant Biology* **8** (2008).

15. Reinartz, J. *et al*. Massively parallel signature sequencing (MPSS) as a tool for in-depth quantitative gene expression profiling in all organisms. *Briefings in Functional Genomics & Proteomics* **1**, 95-104 (2002).

16. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10**, 57-63 (2009).

17. Nagalakshmi, U. *et al*. The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. *Science* **320**, 1344-1349 (2008).

18. Wilhelm, B. T. *et al*. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* **453**, 1239-1243 (2008).

19. Lister, R. *et al*. Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis. *Cell* **133**, 523-536 (2008).

20. Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. & Gilad, Y. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*. **18**, 1509-1517 (2008).

21. Morin, R. *et al*. Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *BioTechniques* **45**, 81-94 (2008).

22. Cloonan, N. *et al*. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nature Methods* **5**, 613-619 (2008).

23. Barbazuk, W. B., Emrich, S. J., Chen, H. D., Li, L. & Schnable, P. S. SNP discovery via 454 transcriptome sequencing. *The Plant Journal* **51**, 910-918 (2007).

24. Vera, J. C. *et al*. Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Mol. Ecol*. **17**, 1636-1647 (2008).

25. Emrich, S. J., Barbazuk, W. B., Li, L. & Schnable, P. S. Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Res*. **17**, 69-73 (2007).

26. Pareek, C. S., Smoczynski, R. & Tretyn, A. Sequencing technologies and genome sequencing. *J. Appl. Genet*. **52**, 413-435 (2011).

27. Marguerat, S. & BÃ¤hler, J. RNA-seq: from technology to biology. *Cellular and Molecular Life Sciences* **67**, 569-579 (2010).

28. Oshlack, A., Robinson, M. D. & Young, M. D. From RNA-seq reads to differential expression results. *Genome Biol*. **11** (2010).

29. Srivastava, S. & Chen, L. A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. *Nucleic Acids Res*. **38**, e170-e170 (2010).

30. Taslim, C. *et al*. Comparative study on ChIP-seq data: normalization and binding pattern characterization. *Bioinformatics* **25**, 2334-2340 (2009).

31. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*. **11** (2010).

32. Zhou, X., Oshlack, A. & Robinson, M. D. miRNA-Seq normalization comparisons need improvement. *RNA* **19**, 733-734 (2013).

33. Risso, D., Ngai, J., Speed, T. P. & Dudoit, S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol*. **32**, 896-902 (2014).

34. Pickrell, J. K. *et al*. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768-772 (2010).

35. Bullard, J., Purdom, E., Hansen, K. & Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11** (2010).

36. Oshlack, A. & Wakefield, M. J. Transcript length bias in RNA-seq data confounds systems biology. *Biology Direct* **4** (2009).

37. Bolstad, B. M., Irizarry, R. A., Ã…strand, M. & Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185-193 (2003).

38. Ogunnaike, B. A., Gelmi, C. A. & Edwards, J. S. A probabilistic framework for microarray data analysis: Fundamental probability models and statistical inference. *J. Theor. Biol.* **264**, 211-222 (2010).

39. Sebastiani, P., Gussoni, E., Kohane, I. S. & Ramoni, M. F. Statistical Challenges in Functional Genomics. *Statistical Science* **18**, 33-70 (2003).

40. Farztdinov, V. & McDyer, F. Distributional fold change test, a statistical approach for detecting differential expression in microarray experiments. *Algorithms for Molecular Biology* **7** (2012).

41. Dillies, M. *et al*. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics* (2012).

42. Sun, Z. & Zhu, Y. Systematic comparison of RNA-Seq normalization methods using measurement error models. *Bioinformatics* **28**, 2584-2591 (2012).

43. Soneson, C. & Delorenzi, M. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics* **14** (2013).

44. Guo, Y., Li, C., Ye, F. & Shyr, Y. Evaluation of read count based RNAseq analysis methods. *BMC Genomics* **14** (2013).

45. Garmire, L. X. & Subramaniam, S. Evaluation of normalization methods in mammalian microRNA-Seq data. *RNA* **18**, 1279-1288 (2012).

46. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11** (2010).

47. Auer, P. L. & Doerge, R. W. Statistical Design and Analysis of RNA Sequencing Data. *Genetics* **185**, 405-416 (2010).

48. Robinson, M. D. & Smyth, G. K. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics* **9**, 321-332 (2008).

49. Hardcastle, T. J. & Kelly, K. A. baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics* **11** (2010).

50. Zhou, Y., Xia, K. & Wright, F. A. A powerful and flexible approach to the analysis of RNA sequence count data. *Bioinformatics* **27**, 2672-2678 (2011).

51. Bartel, D. P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**, 281-297 (2004).

52. Bartel, D. P. MicroRNAs: Target Recognition and Regulatory Functions. *Cell* **136**, 215-233 (2009).

53. Iorio, M. V. & Croce, C. M. MicroRNA dysregulation in cancer: diagnostics, monitoring and therapeutics. A comprehensive review. *EMBO Molecular Medicine* **4**, 143-159 (2012).

54. Iorio, M. V. & Croce, C. M. microRNA involvement in human cancer. *Carcinogenesis* **33**, 1126-1133 (2012).

55. Pasquinelli, A. E. MicroRNAs and their targets: recognition, regulation and an emerging reciprocal relationship. *Nature reviews.Genetics* **13**, 271-282 (2012).

56. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11** (2010).

57. Robinson, M. D. & Smyth, G. K. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* **23**, 2881-2887 (2007).

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

59. Smyth, G. K. in (eds Gentleman, R., Carey, V. J., Huber, W., Irizarry, R. A. & Dudoit, S.) 397-420 (Springer New York, 2005).

60. Di, Y., Schafer, D. W., Cumbie, J. S. & Chang, J. H. The NBP Negative Binomial Model for Assessing Differential Gene Expression from RNA-Seq. *Statistical Applications in Genetics and Molecular Biology* **10**, 1-28 (2011).

61. Kadota, K., Nishiyama, T. & Shimizu, K. A normalization strategy for comparing tag count data. *Algorithms for Molecular Biology* **7** (2012).

62. Di, Y., Schafer, D. W. & Di, M. Y. Package "NBPSeq". *Mol. Biol. (N. Y.)* **10** (2012).

63. Hardcastle, T. J. & Kelly, K. A. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics* **11** (2010).

64. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society.Series B (Methodological)* **57**, 289-300 (1995).

Chapter 3

We were later interested in studying whether different intracellular amounts of FHC might affect gene expression profile. As first step toward the dissection of the molecular basis of FHC-modulated gene expression, we performed an integrated analysis of miRNA and mRNA expression patterns in K562 FHC-silenced cells. We used K562 cells in which the expression of FHC has been stably knocked-down by shRNA interference and whose transcriptome profile was already established. By using a microRNA PCR Panel, we found that 4 out of 84 analysed miRNAs, namely hsa-let-7g-5p, hsa-let-7f-5p, hsa-let-7i-5p and hsa-miR-125b-5p, were consistently and significantly up-regulated in FHCsilenced K562 cells compared to control cells. The correlation among FHC amounts and the expression of the four miRNAs was further supported by the transient silencing and the reconstitution experiments. The profile of the four up-regulated miRNAs has been integrated with the transcriptome analysis by combining data obtained from the microRNA targets prediction software with a correlation-based approach. This test is based on the assumption that, since miRNAs tend to down-regulate the expression of their targets, the expression profiles of miRNAs are expected to be inversely related with those of their true target genes. This analysis led to the identification of 91 down-regulated genes, the majority of whom appear to be candidate targets of a single miRNA, while 15 are subjected to multiple miRNA regulation. IPA revealed that the highest scored pathways in which these genes are involved are: “Cell Death and Survival, Hematological System Development and Function, Hematopoiesis” and “DNA Replication, Recombination and Repair, Cell Cycle, Cancer”. We believe that the identification of FHC-dependent miRNA/mRNA networks implies that different amounts of the ferritin subunit contribute, in K562 cells, to the remodelling of gene expression taking place during these cellular processes through the action of let-7g, let-7f, let-7i and miR-125b. This observation is further strengthened by the comparison of the miRNAs-regulated pathways, reported in this work. It is interesting to note that, among the common pathways, “Cell Death and Survival” and “Hematological System Development and Function” rely on the ERK1/2 activation that our results demonstrate to be severely affected by FHC modulation. In conclusion, the data presented in this study add a further level of complexity to the relationship among iron and miRNAs, since it appears that the

intracellular amounts of FHC subunit are able to regulate let-7g, let-7f, let-7i and miR-125b expression, as well as the repertoire of their down-stream genes. Recent reports suggest that the redox state of the cell might influence Let7 and 125-b levels, but certainly the FHC interference on miRNA expression deserves further analysis.

H-ferritin-regulated microRNAs modulate gene expression in K562 cells

Flavia Biamonte[1¶], Fabiana Zolea[1¶], Andrea Bisognin[2], Maddalena Di Sanzo[1], Claudia Saccoman[2], Domenica Scumaci[1], Ilenia Aversa[1], Mariafranca Panebianco[1], Maria Concetta Faniello[1], Stefania Bortoluzzi[2], Giovanni Cuda[1&], Francesco Saverio Costanzo[1&*]

[1] Department of Experimental and Clinical Medicine, Magna Græcia University of Catanzaro, Salvatore Venuta Campus, Viale Europa, 88100 Catanzaro, Italy.

[2] Department of Biology, University of Padua, Via G. Colombo 3, 35131 Padua, Italy.

* Corresponding author

E-mail: fsc@unicz.it (FSC)

[¶] These authors contributed equally to this work.

[&]These authors also contributed equally to this work.

Abstract

In a previous study, we showed that the silencing of the heavy subunit (FHC) of ferritin, the central iron storage molecule in the cell, is accompanied by a modification in global gene expression. In this work, we explored whether different FHC amounts might modulate miRNA expression levels in K562 cells and studied the impact of miRNAs in gene expression profile modifications. To this aim, we performed a miRNA-mRNA integrative analysis in K562 silenced for FHC (K562shFHC) comparing it with K562 transduced with scrambled RNA (K562shRNA). Four miRNAs, namely hsa-let-7g, hsa-let-7f, hsa-let-7i and hsa-miR-125b, were significantly up-regulated in silenced cells. The remarkable down-regulation of these miRNAs, following FHC expression rescue, supports a specific relation between FHC silencing and miRNA-modulation. The integration of target predictions with miRNA and gene expression profiles led to the identification of a regulatory network which includes the miRNAs up-regulated by FHC silencing, as well as 91 down-regulated putative target genes. These genes were further classified in 9 networks; the highest scoring network, "Cell Death and Survival, Hematological System Development and Function, Hematopoiesis", is composed by 18 focus molecules including RAF1 and ERK1/2. We confirmed that, following FHC silencing, ERK1/2 phosphorylation is severely impaired and that RAF1 mRNA is significantly down-regulated. Taken all together, our data indicate that, in our experimental model, FHC silencing may affect RAF1/pERK1/2 levels through the modulation of a specific set of miRNAs and add new insights in to the relationship among iron homeostasis and miRNAs.

Introduction

A tight regulation of iron homeostasis is essential for life in eukaryotic cells. The availability of iron is required for critical pathways such as ATP generation and DNA synthesis. Deregulated iron levels contribute indeed to the generation of free radicals that, in turn, damage cellular proteins and nucleic acids [1]. Ferritin, a 24-mer protein, is devoted to keep intracellular iron in a bio-available and non-toxic form [2], thus playing a central role in intracellular iron equilibrium.

The nano-cage of the ferritin molecule is composed by a well-defined array of heavy-type (FHC) and light-type (FLC) subunits, coded by two different genes [3] that share both extensive aminoacid sequence (55%) and structural similarity. The two subunits perform different functions in iron metabolism: FHC is involved in rapid iron uptake and release and it has ferroxidase activity, while FLC, devoid of enzymatic activity, essentially contributes to long-term iron storage [4]. Recently, several lines of evidence have demonstrated that FHC is a multi-functional protein, that might play a central role in proliferation [5], angiogenesis [6], chemokine signalling [7] and neoplastic transformation [8]. *FHC* expression is modulated, at transcriptional level, by proteins involved in tumorigenesis; among them, E1A [9], p53 [10], and c-Myc [11] act as repressors, while c-Jun is an inducer [12]. FHC itself binds to p53 and is able to activate *p53* transcription under oxidative stress conditions [13]. Moreover, *FHC* transcription is activated by TNFα and interleukin 1α (IL-1α) [14], suggesting that pathways related to inflammation and stress can impact on ferritin regulation. The ferritin H subunit also physically interacts with, and regulates the activity of the chemokine receptor CXCR4 [7], highly expressed in a variety of human malignancies. *FHC* down-regulation by shRNA interference strongly modifies, *in vivo* and *in vitro*, the proliferation of human melanoma cells [15].

This scenario is even more complicated when considering the relationship between ferritin and cellular proliferation. *FHC* up-regulation has been associated with induction of differentiation and growth arrest in hematopoietic systems [16], differentiation of the Caco-2 enterocytic cell line [17] and switch from pre-adipocytes to adipocytes [18].

The last decade has witnessed a tremendous increase of knowledge on the role of microRNAs (miRNAs) in regulating gene expression in normal and pathological conditions. These non-coding RNAs, with an average length of 19-25 nucleotides, are able to modulate the expression of thousands of genes by inhibiting translation or

inducing degradation of transcripts. Moreover, one target transcript can be controlled by more than one miRNA [19]. It has been suggested that miRNAs might regulate more than 60% of the protein coding genes [20]. Key functional roles for miRNAs have been demonstrated in development, organogenesis and cell differentiation [21]. In hematopoietic stem and progenitor cells miR-221, miR-222, miR-223 and miR-150 act as master regulators, contributing to the hematopoietic development and the lineage specification [22, 23].

The role of miRNAs in cancer has been deeply investigated. Specific patterns of miRNA expression (*miRNome*) and variations have been established in different tumour stages and subtypes. miRNAs can play oncogenic (oncomiRNAs) and/or tumor suppressive role in almost all the aspects of cancer biology [24]. Moreover, a specific miRNA can play opposite roles in different contexts: for example, miR-29 acts as a tumor-suppressor in lung cancer, while it plays oncogenic functions in breast cancer [25]. Like virtually all other cellular processes, also iron homeostasis is regulated by specific miRNAs. miR-210 acts on the transferrin receptor and is involved in iron acquisition. Iron storage and utilization are controlled by miR-200b, targeting FHC, while iron release is regulated by miR-485-3p, through its action on ferroportin (Fpn) [26].

We have recently found that, in a metastatic melanoma cell line [15] and in the K562 erythroleukemia cell line [27], the silencing of *FHC* subunit is accompanied by profound modifications of gene expression. The molecular basis of the link among *FHC* levels and gene expression profile in these cells have not been established yet.

In this study, we profiled both mRNA and miRNA expression in K562 cells silenced for the ferritin H subunit and compared these expression profiles with that of control cells. We identified specific miRNAs and genes differentially expressed upon *FHC*-knock down and studied the relations thereof.

Materials and Methods

miRNA isolation and quantitative real-time PCR

miRNA-enriched total RNA was extracted from cultured FHC-silenced K562 ($K562^{shFHC}$) cells and K562 transduced with scrambled RNA (shRNA) using miRCURY^{TM} RNA Isolation Kit Cell and Plant (EXIQON, Woburn, USA) following the manufacturer's protocol. The concentration of RNA and the RNA quality (260/280 and 260/230 absorbance ratios) of the samples were measured using Nanodrop (Thermo SCIENTIFIC, Waltham, MA, USA). We designed a double-step analysis for identification and quantification of abnormally expressed miRNAs in K562 shFHC compared to K562 shRNA cells. The first was a "panel" procedure that simultaneously evaluated expression level of different mature miRNAs by quantitative real-time PCR (qRT-PCR); the second was performed on individual miRNAs, which eventually resulted differentially expressed in panel experiments. For panel analysis, we used Cancer Focus microRNA PCR Panel that assesses the expression levels of 84 onco-miRNAs.

Each sample was assayed in triplicate, and the experimental data were normalized to the expression levels of the housekeeping small nuclear RNA,U6.

Identification of differentially expressed miRNAs

The fold change of miRNAs expression among the tested samples was calculated using $2^{-\Delta\Delta Ct}$ formula. Differences among the two sets of samples were analyzed by the Student *t*-test. Those differences with a $p<0.05$ were considered statistically significant.

From this first analysis we decided to focus on those miRNAs that were found to be up-regulated in K562 shFHC compared to K562 shRNA cells. cDNA synthesis, was performed using TaqMan® MicroRNA Reverse Transcription Kit (Life Technologies, Carlsbad, CA, USA) containing microRNA-specific RT primers and Taqman miRNA assay. To measure miRNAs expression levels, 1.33 µL of each cDNA was added to the specific TaqMan microRNA Assay (20X) and TaqMan 2X Universal PCR Master MiX (Life Technologies, Carlsbad, CA, USA). The amplification conditions for miRNA qRT-PCR were the following: 10 min at 95 °C, 40 cycles at 95 °C for 15 s, and 60 °C for 60 s. The experiments were performed in duplicate and the analysis was performed using the $2^{-\Delta\Delta Ct}$ formula.

Transfection of K562 cells

K562 cells were transfected using electroporation. In particular, over-expression of FHC was performed using the expression vector containing the full length of human FHC cDNA (pc3/FHC); transient silencing of K562 cells was obtained using a homemade FHC siRNA, kindly provided by Prof. Sonia Levi from the Vita-Salute San Raffaele University Milano, Italy. For rescue of FHC expression, approximately 6×106 K562-silenced cells ($K562^{shFHC}$) were resuspended in 600µL of Opti-MEM (Gibco BRL). Subsequently, 3×106 of cell suspension was mixed with 30 µg of pc3/FHC (K562shFHC/pc3FHC). The remaining 3×106 of cell suspension was mixed with the control plasmid, pcDNA™3.1 (K562shFHC/pcDNA™3.1). For transient silencing of FHC, 6x106 K562 cells were resuspended in 600µL of Opti-MEM and then, half was mixed with 15µg of a GFP-positive control siRNA (K562 Ctrl siRNA) and half with 15µg of FHC siRNA (K562 FHC siRNA). After 15 minutes of incubation at room temperature, each sample was electroporated in a sterile electroporation cuvette (Bio-Rad Gene Pulser cuvette, 0.4 cm) using Gene Pulser Xcell Electroporation System (Bio-Rad). Electroporation was performed at 285V and 975µFa. After electroporation, cell suspensions were centrifugated at maximum speed and the pellets were left at room temperature for 20 minutes. Then, fresh complete medium was added to the pellets and cells were further incubated at 37°C in a humidified atmosphere supplemented with 5% CO2. Transfection efficiency was measured after 72h using real-time PCR.

Identification of differentially expressed genes

Genes modulated after FHC silencing have been identified using Limma package [28]. Differential expression analysis was obtained by a t-statistic, which is computed for each gene and for each contrast, with standard errors moderated across genes, exploiting the Empirical Bayes shrinkage method to stabilize the variance estimate.

Only genes with absolute log(FC) of at least 1 and a FDR q-value lower than 0.1 have been considered differentially expressed.

Identification of anticorrelated predicted targets of miRNAs

We identified the predicted regulatory relations significantly supported by expression data, integrating target predictions with miRNA and gene expression profiles in silenced and un-silenced cells. Only differentially expressed miRNAs were considered. Target predictions were computed with TargetScan.

Pairwise Spearman correlations between miRNA and predicted target gene expression profiles were calculated. The supported relationships associated to statistically significant correlations ($r <= -0.81$ and p-value $<=0.05$) were selected.

Pathways visualization

Network visualization and annotation have been performed using Cytoscape [29].

Functional analysis of target genes

In order to infer the potential functions of the differentially expressed miRNAs, we performed the functional analysis of their target genes using Ingenuity Pathway Analysis (IPA) database. IPA maps each gene within a molecular network and defines it as "focus molecule". Ingenuity Pathway Analysis (IPA) software program was used as described elsewhere [30]. Following IPA analysis, Panther (Protein ANalysis THrough Evolutionary Relationships) was used to also classify genes in specific signalling and metabolic pathways (http://www.pantherdb.org/).

RNA extraction and quantitative real-time PCR for *FHC* and *c-Myc* and *RAF1* detection

Total RNA was extracted from two distinct batches of K562 shRNA and K562 shFHC cells using the Trizol method (Life Technologies, Carlsbad, CA, USA). Real-time PCR was performed using 10X SYBR Green PCR Master mix (Life Technologies, Carlsbad, CA, USA), 400 nM of each primer pair, 20 ng of cDNA (total RNA equivalent) and nuclease-free water. The thermal profile consisted of 1 step at 95 °C for 10 min followed by 45 cycles at 95 °C for 30 s, 60 °C for 60 s. Human glyceraldehyde 3-phosphate

dehydrogenase (GAPDH) was used as housekeeping. Each reaction was performed in duplicate. The primer sequences for FHC and GAPDH have been already published (27). The primer sequences for RAF1 and c-MYC were as follow:

RAF1 FW: TGCTGCGTCTTTGATTGGAG

RAF1 REV: TGGTGCTACAGTGCTCATGA

c-MYC FW: CCTCGGATTCTCTGCTCTCC

c-MYC REV: TGTGAGGAGGTTTGCTGTGG

Protein Extraction and Western Blotting Analysis

K562 shRNA and K562 shFHC cells were lysed in the following buffer [20 mM Hepes pH 7.9, 420 mM NaCl, 1% Triton X-100, 1 mM EDTA, 25% glycerol, 1 mM PMSF, 1 mM Na_3VO_4, 1 mM DTT, 1 μg/ml aprotinin, 1 μg/ml leupeptin] for 30 min on ice. After removal of the cell debris by centrifugation (12,000 ×*g*, 30 min), the concentration of proteins in the supernatant was measured by the Bio-Rad protein assay according to the manufacturer's instructions (Bio-Rad Laboratories, Hercules, CA, USA) [31]. A total of 50 μg protein extract was boiled for 10 min in SDS sample buffer, separated by 12% SDS-PAGE and the proteins were transferred to a nitrocellulose membrane by electroblotting. Non-specific reactivity was blocked by incubating the membrane in nonfat dry milk in TPBS [5% (w/v) milk in PBS (pH 7.4) and 0.005% Tween 20] for 2 h at room temperature. The membrane was incubated with primary mouse anti-Phospho-p44/42 MAPK (Erk1/2) (Thr202/Tyr204) antibody (1:1000; Cell Signaling Technology, Danvers, MA, USA) overnight at 4°C. Being washed in TPBS, the membranes were subsequently incubated with anti-mouse secondary antibody (1:3000 Cell Signaling Technology, Danvers, MA, USA) for 2 hours. The membrane was developed by ECL-Western blot detection reagents according to the manufacturer's instructions (Santa Cruz Biotechnology, Texas, USA). γ-Tubulin was used as a loading control.

Assessment of cell proliferation

3-[4,5-Dimethylthiaoly]-2,5-diphenyltetrazolium bromide (MTT) assay was performed to detect proliferation of K562 shRNA and K562 shFHC cells. The experiments were performed on starved cells that were obtained culturing proliferating cells with RPMI

1640 without FBS for 24h. A total of $4.5X10^4$ cells/well were seeded into 96-well plate and let to grow for 72h in RPMI medium. There were octuplicates for each cell type. Fresh MTT (Sigma Aldrich, Saint Louis, MO, USA), re-suspended in PBS was added to each well. After 2h incubation, culture medium was discarded and replaced with 200µL of DMSO. Optical density was measured at 570 nm in a spectrophotometer. Each experiment was performed in triplicate.

Results

miRNA and transcriptome analysis in K562 cells

Our main goal in the past years has been the identification and classification of genes whose expression is directly or indirectly modulated by FHC in different human cell lines. In the present study, we evaluated if the silencing of FHC may also alter oncomiRNAs expression in the K562 erythroleukemic cell line with the aim of identifying the potentially regulated genes. To this, we utilized a cell clone, already described, in which FHC expression has been knocked-down with specific shRNA [27]. Using Cancer Focus microRNA PCR Panel, we have identified 59 miRNAs, 12 of which were up-regulated and 3 down-regulated, with an absolute Log fold-change (LogFC) greater than 1, in K562 cells silenced for H ferritin (shFHC) versus K562 cells transduced with scrambled RNA (shRNA) (Table S1). The analysis was performed in triplicate from both cell types using quantitative real-time PCR (qRT-PCR). Four miRNAs namely hsa-let-7g-5p, hsa-let-7f-5p, hsa-let-7i-5p and hsa-miR-125b-5p resulted significantly up-regulated, with LogFC variation of at least five and a t-test p-value <0.05, after FHC knock-down (Table 1).

microRNA	LogFC (shFHC vs shRNA)	p-value
hsa-let-7g-5p	7.38	0.0133
hsa-let-7f-5p	5.47	0.0218
hsa-let-7i-5p	4.95	0.0340
hsa-miR-125b-5p	5.82	0.0470

Table 1 .Four miRNAs are significantly up-regulated after H ferritin silencing

The expression of these four miRNAs was assessed by TaqMan assay in an independent set of RNA obtained from silenced cells and from K562 cells in which the expression of FHC has been restored. The results of a duplicate set of experiments are shown in Panels A and B of Figure 1. Panel A shows the extent of FHC-silencing and reconstitution. In Panel B are reported the expression levels of the four miRNAs in the silenced and reconstituted K562 cells. It appears that the four miRNAs are indeed up-regulated in the cells in which the FHC subunit was present at a significantly lower level (FC= 0.3)

compared to shRNA cells (p-value<0.05) and down-regulated in the cells where FHC expression levels were rescued.

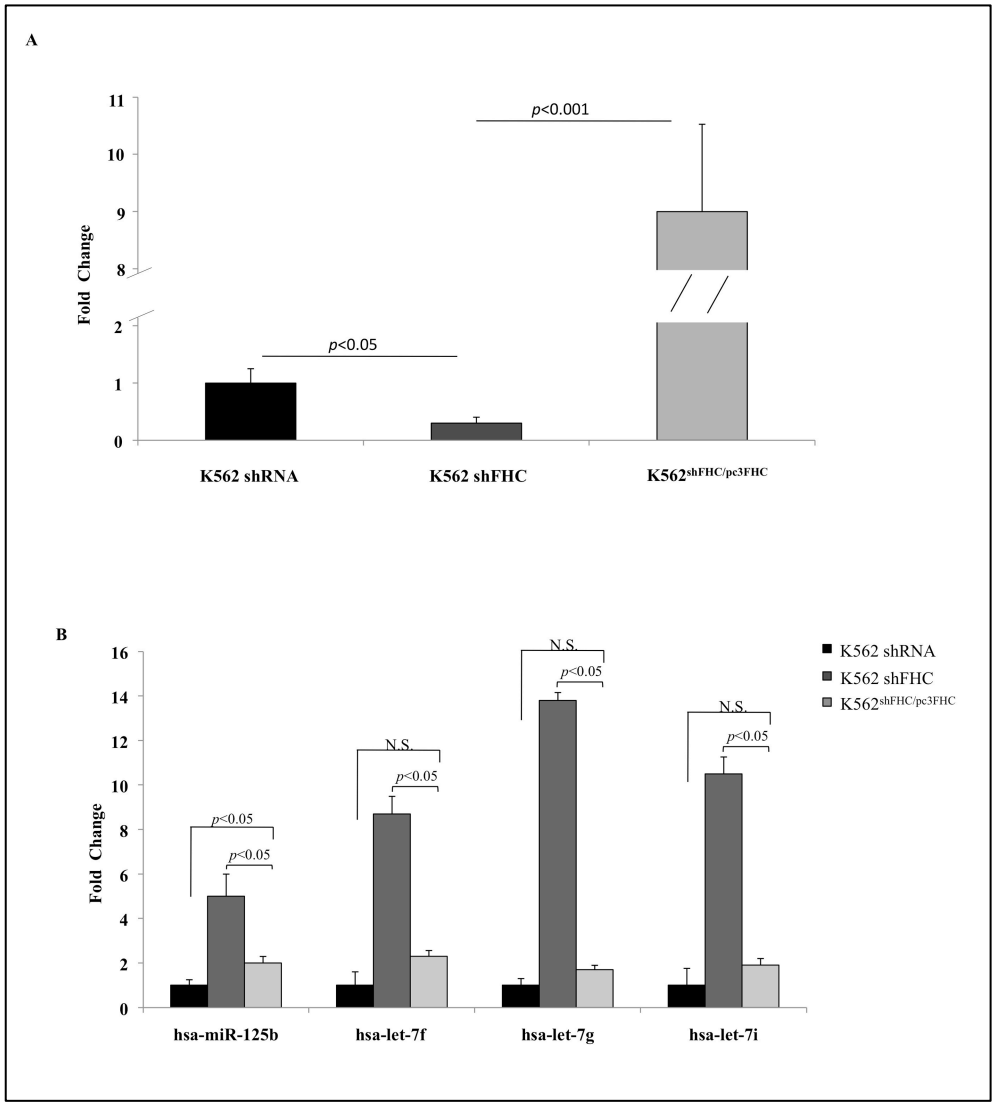

Figure 1. Four miRNAs are significantly modulated by FHC amounts. A) Real-time PCR analysis of FHC mRNA performed on total RNA from K562 shRNA, K562 shFHC and K562$^{shFHC/pc3FHC}$. Results are representative of two different experiments. B) TaqMan analysis of hsa-miR-125b, hsa-let-7f, hsa-let-7g, hsa-let-7i in K562 shRNA, K562 shFHC and K562$^{shFHC/pc3FHC}$. Results are representative of two different experiments. N.S.: Not Significant

According with the microRNA PCR panel, the greatest increase was observed for hsa-let-7g, whose expression is about 14-fold higher in the silenced cells compared to the control. We also transiently trasfected K562 cells with a homemade FHC siRNA kindly provided by Professor S. Levi, and compared the expression levels of the four miRNAs with those of control cells. In two independent experiments, a different silencing efficiency, in the order of about 20 and 40%, was obtained. Panels A and B of Figure 2

show that, in both samples, *FHC* silencing is accompanied by an up-regulation of hsa-let-7g-5p, hsa-let-7f-5p, hsa-let-7i-5p and hsa-miR-125b-5p.

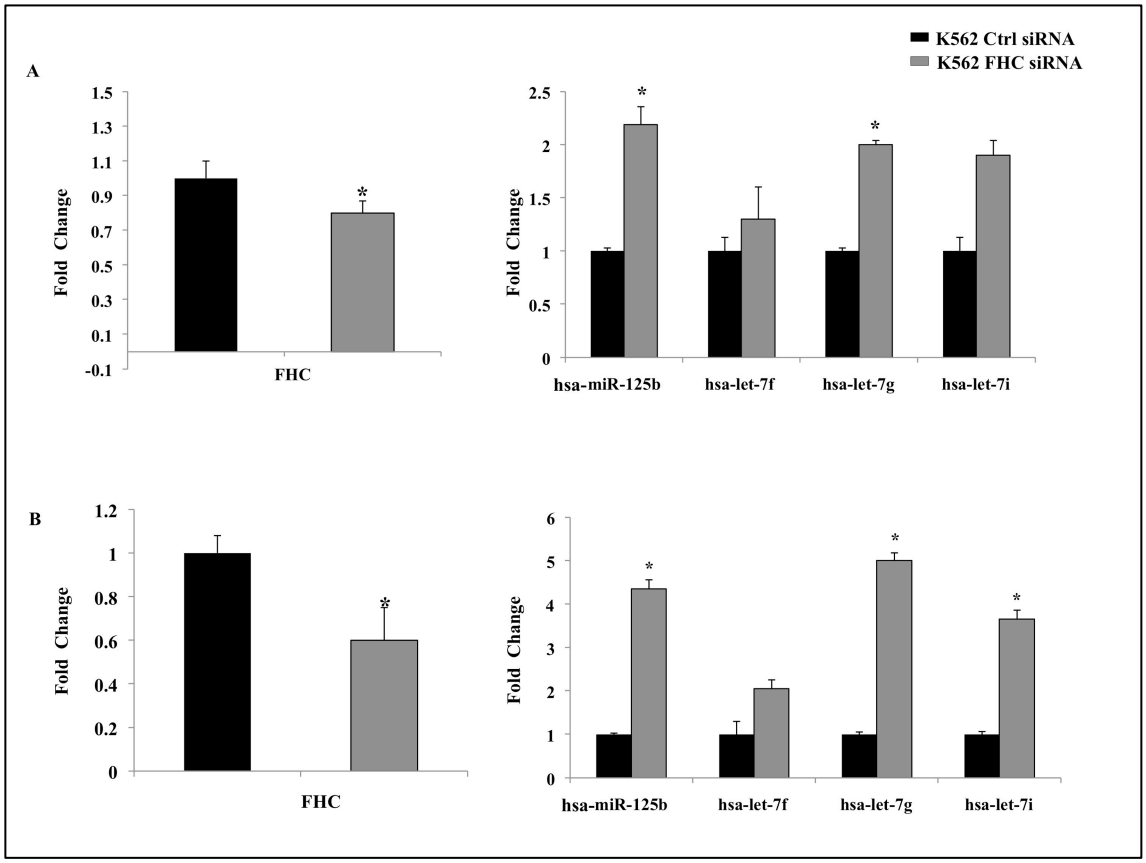

Figure 2. Transient silencing of FHC induce up-regulation of hsa-miR-125b, hsa-let-7f, hsa-let-7g, hsa-let-7i. A transient silencing of FHC of about A) 20% and B) 40% is accompanied by the up-regulation of hsa-miR-125b, hsa-let-7f, hsa-let-7g, hsa-let-7i. Results are representative of two different experiments performed by TaqMan analysis. **p value*<0.05

The gene expression profile of FHC-silenced versus un-silenced K562 cells has been already determined in a previous work [27]. Here, in order to integrate miRNA and mRNA transcriptome findings, we have re-analyzed the raw microarray data. The Limma differential expression analysis identified 219 transcripts with a significantly altered expression in the FHC-silenced cells, including 64 up- and 53 down-regulated genes with an absolute LogFC greater than 2. The full cast of the FHC-dependent mRNAs is reported in Supplementary Table 2 (Table S2).

miRNA-mRNA regulatory network

Next, for differentially expressed miRNAs, we integrated target predictions with miRNA and gene expression profile, to identify the regulatory relationships significantly supported by expression data. Combining TargetScan predictions of miRNA-target interactions with a correlation-based analysis of miRNA and transcript expression profiles (see Materials and Methods), we obtained 108 interactions supported by expression data, involving hsa-let-7g-5p, hsa-let-7f-5p, hsa-let-7i-5p and hsa-miR-125b-5p and 91 down-regulated genes. In particular, the expression of hsa-let-7i resulted to be negatively correlated with that of 13 transcripts; hsa-let-7f and hsa-let-7g with that of 20 transcripts and 25 transcripts, respectively; finally, hsa-miR-125b negatively correlated with 50 transcripts. As shown in the reconstructed regulatory network (Figure 3), the majority of these genes were supported targets of only one miRNA, whereas 15 genes were putatively regulated by two or more up-regulated miRNAs.

The list of the 91 down-regulated transcripts with their cognate miRNAs is reported in Supplementary Table 3 (Table S3).

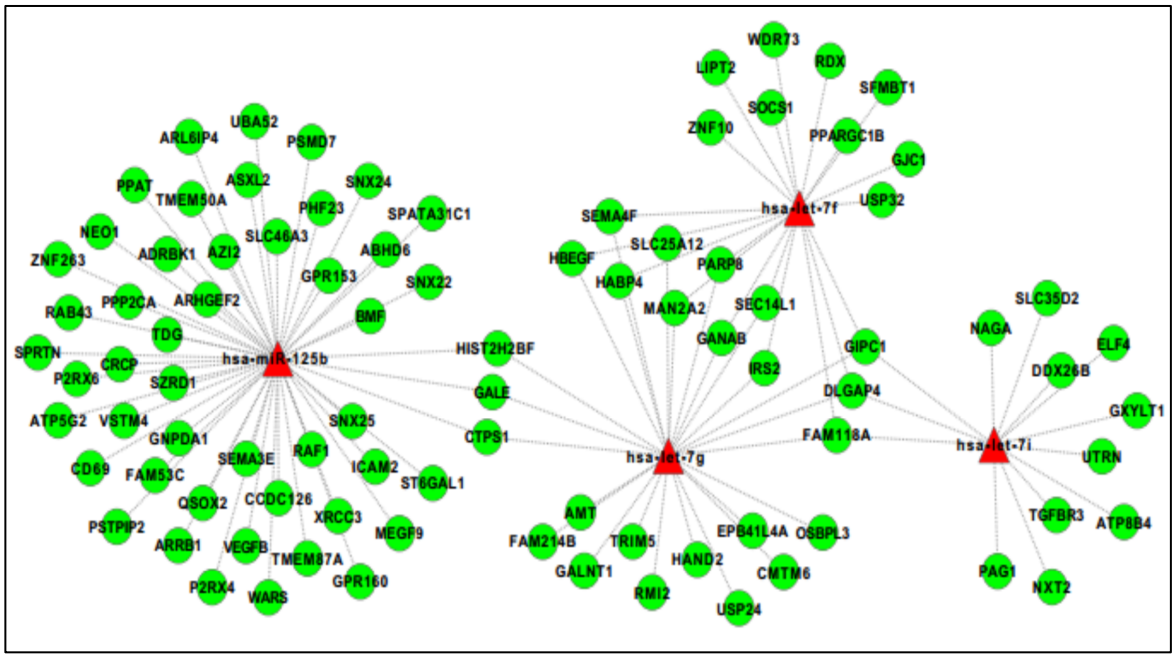

Figure 3. miRNA-mRNA interaction networks. miRNA-mRNA interaction networks built by Cytoscape. We identified a total of 108 miRNA-mRNA significantly negatively correlated interaction. The four up-regulated miRNAs are colored in red and the 91 down-regulated target mRNAs are in green. let-7i is correlated with 13 transcripts; let-7f and let-7g with 20 and 25 transcripts, respectively; miR-125b negatively correlates with 50 transcripts. The majority of genes are supported targets of only one specific miRNA, whereas 15 genes are putatively regulated by two or more distinct up-regulated miRNAs.

miRNAs modulated by FHC silencing impact on specific pathways

The 91 down-regulated supported target genes of hsa-let-7g-5p, hsa-let-7f-5p, hsa-let-7i-5p and hsa-miR-125b-5p were studied with two knowledge-based approaches to better characterize the networks potentially modulated by *FHC* silencing. Ingenuity Pathway Analysis tool (IPA) highlighted the 9 networks reported in Table 2; of them, the highest scoring is "Cell Death and Survival, Hematological System Development and Function, Hematopoiesis" with a significance score of 37 and 18 focus molecules (Panel A of Figure 4), followed by "DNA Replication, Recombination and Repair, Cell Cycle, Cancer" with a significance score of 32 and 16 focus molecules (Panel B of Figure 4). The significance scores of these networks (estimating the probability that a collection of genes equal to or greater than the number in a network can be achieved by chance alone) are very high, since a score of 3 indicates a 1/1000 chance that the focus genes are in a specific network due to random chance.

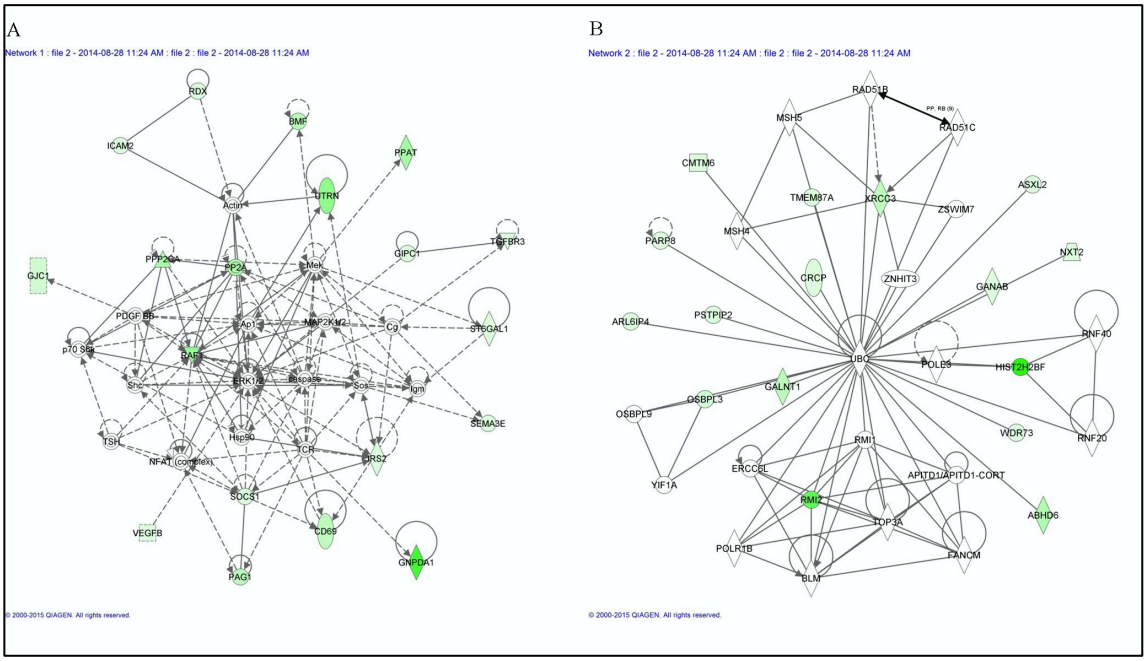

Figure 4. The two highest scoring networks identified by IPA, that correlate genes target of the miRNAs differentialy expressed after FHC silencing. Ingenuity Pathway Analysis was used to investigate the networks potentially affected by the down-regulated genes. (A) Cell Death and Survival, Hematological System Development and Function, Hematopoiesis" is the highest scoring network with a significance score of 37 and 18 focus molecules (B) DNA Replication, Recombination and Repair, Cell Cycle, Cancer" has a significance score of 32 and 16 focus molecules. The target down-regulated genes are shaded in green. Intensity of shading correlates with the degree of down-regulation. A solid line represents a direct interaction between two genes, while a dotted line indicates an indirect interaction.

Molecules in Network	Score	Focus Molecules	Top Diseases and Functions
BMF, CD69, GIPC1, GJC1, GNPDA1, ICAM2, IRS2, PAG1, PPAT, PPP2CA, RAF1, RDX, SEMA3E, SOCS1, ST6GAL1, TGFBR3, UTRN, VEGFB	37	18	Cell Death and Survival, Hematological System Development and Function
ABHD6, ARL6IP4, ASXL2, CMTM6, CRCP, GALNT1, GANAB, HIST2H2BF, NXT2, OSBPL3, PARP8, PSTPIP2, RMI2, TMEM87A, WDR73,XRCC3	32	16	DNA Replication, Recombination and Repair, Cell Cycle, Cancer
DDX26B, GXYLT1, PHF23, QSOX2, RAB43, SLC46A3, SNX22, SNX24, SNX25, SZRD1, TMEM50A, VSTM4, ZNF10	24	13	Cancer, Gastrointestinal disease, Cell death and Survival
ADRBK1, ARHGEF2, ARRB1, ATP5G2, AZI2, HAND2, HBEGF, P2RX4, P2RX6, SFMBT1, SLC25A12, UBA52, USP24	23	13	Cardiovascular system development and function, Developmental disorders, Organ morphology
AMT, DLGAP4, ELF4, EPB41L4A, FAM214B, MEGF9, SEMA4F, SPRTN, USP32, ZNF263	16	10	Cancer, Gastrointestinal disease, Cell to cell signaling and interaction
CCDC126, FAM118A, FAM53C, GALE, GJC1, GPR153, GPR160, MAN2A2, NEO1, PPARGC1B,	16	10	Cancer, Cellular Movement, Tissue Morphology
CTPS1, HABP4, NAGA, PSMD7, SEC14L1, TDG, TRIM5, WARS, LIPT2	13	8	Carbohydrate Metabolism, Developmental Disorder, Hereditary Disorder
LIPT2	2	1	Organ, Morphology, Riproductive System Development and Function, Endocrine System Development and Function
ATP8B4	2	1	Cancer, Organismal Injury and Abnormalities, Reproductive System Disease

Table 2 Top 9 molecular networks predicted by IPA, by analysis of genes with expression profiles significantly negatively correlated with that of miRNAs differentially expressed after FHC silencing

In parallel, ANalysis THrough Evolutionary Relationship (PANTHER) showed that hsa-let-7g, hsa-let-7f, hsa-let7i and hsa-miR-125b up-regulation might determine changes in 29 metabolic pathways 18 of which are signalling pathways (Figure 5) mostly involving *RAF1*. Notably, *RAF1* is one of the 18 focus molecules identified by IPA in the network reported in Figure 4A.

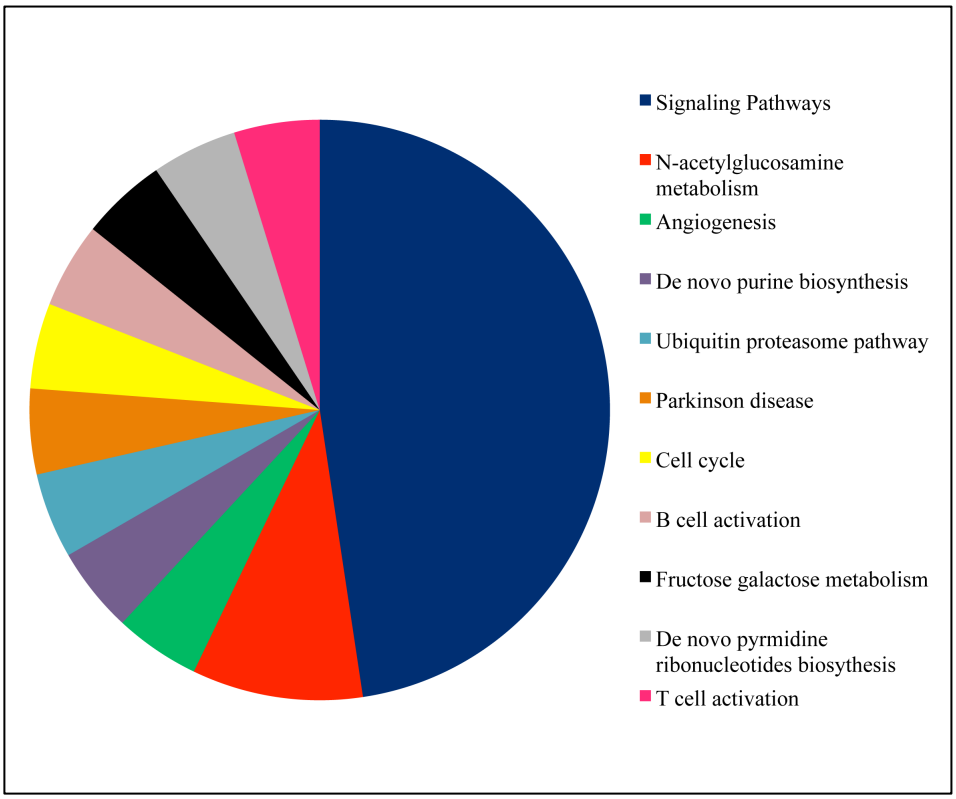

Figure 5. Pathway analysis performed using PANTHER. Panther gene ontology (GO) analysis for the 91 down-regulated target genes. Several metabolic pathways are affected, the majority of them is represented by signalling pathways , shaded in blue.

RAF1, pERK1/2 and c-*Myc* expression in K562 FHC-silenced cells

We noticed that, in the "Cell Death and Survival, Hematological System Development and Function, Hematopoiesis" network (Figure 4A), the 18 focus molecules potentially modulated by hsa-let-7g-5p, hsa-let-7f-5p, hsa-let-7i-5p and hsa-miR-125b-5p, converge on a central hub represented by the ERK1/2 kinase. In particular, 6 of them directly impact on this kinase; RAF-1, VEGFB, CD69 and IRS2 are known activators, PPP2CA acts as inhibitor, while SOCS1 might act either as inhibitor or activator depending on the cellular context. The potential involvement of ERK1/2 is further supported by the

observation that the pathways identified by PANTHER analysis all depend on the activation of this molecule. Thus, we decided to investigate RAF1 expression and ERK1/2 activation in the FHC-silenced cells by real-time PCR and western blot analysis, respectively. The experiments were performed in duplicate on RNAs and protein extracts from two independent sets of cells..

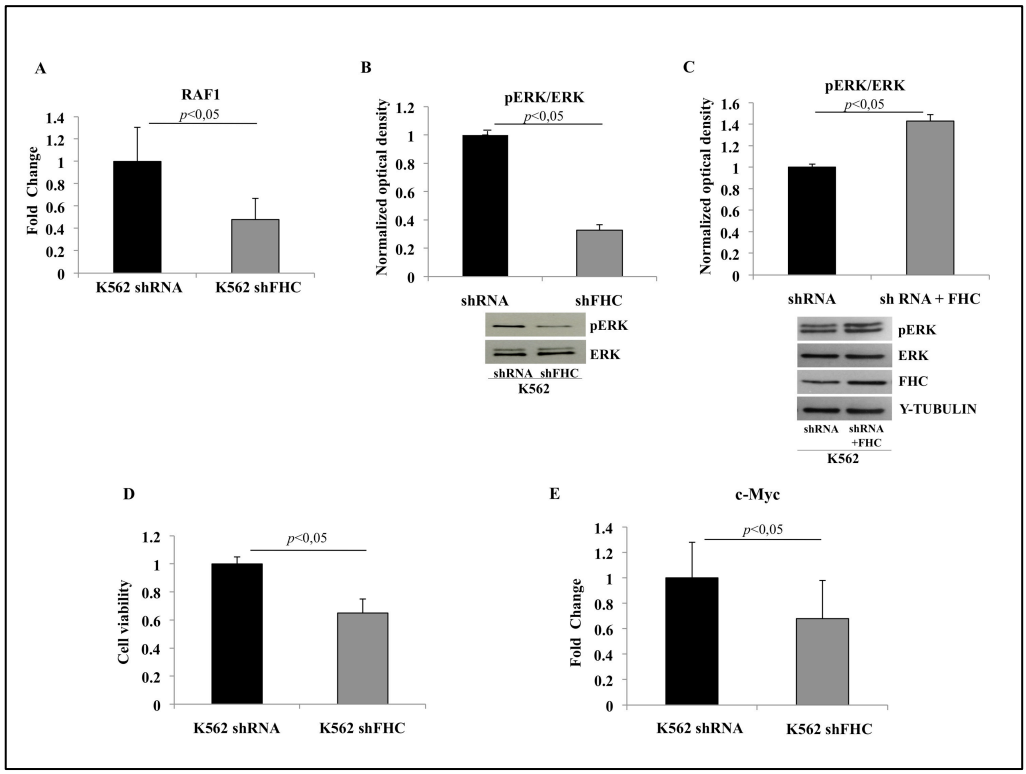

Figure6. FHC silencing in K562 cells reduces proliferation rate via RAF1/MAPK pathway inhibition and is associated with *c-Myc* down-regulation. A) Real-time PCR of *RAF1* mRNA performed on K562 shRNA, K562 shFHC and K562$^{shFHC/pc3FHC}$. Results are representative of two different experiments B) Western Blot analysis for pERK1/2 was performed on 50µg of total protein extract from K562shRNA and K562shFHC cells. Total ERK1/2 was used as loading control. Results are representative of three different experiments. C) Western Blot analysis for pERK1/2 and FHC was performed on 50µg of total protein extract from K562shRNA and K562 shRNA+FHC. Total ERK1/2 and γ-Tubulin were used as loading controls. Results are representative of two different experiments. D) Equal number of starved silenced and un-silenced cells were plated into a 96-well plate, incubated for 72 h and analysed by MTT assay. Proliferation of FHC-silenced cells is reduced of about 35% compared to controls. Data are presented as mean ± standard deviation. E) Real-time PCR of c-*Myc* mRNA performed on K562 shRNA, K562 shFHC and K562$^{shFHC/pc3FHC}$. Results are representative of two different experiments.

Panel A of Figure 6 shows that RAF1 levels are significantly altered by FHC silencing, thus confirming the microarray data. Panel B shows that ERK1/2 phosphorylation is also severely impaired in the silenced cells compared to control. To further correlate ERK1/2 phosphorylation and FHC expression levels, we analysed pERK1/2 after FHC over-expression. Panel C of Figure 6 shows that, in the control cells, an FHC over-expression

of the order of about 37% is accompanied by an increased ERK1/2 phosphorylation. The role of ERK1/2 in the control of cell proliferation has been largely demonstrated [32]. Therefore, we analysed the proliferation rate of the silenced and un-silenced K562 cells by MTT assayThe experiments were performed in triplicate and the results, reported in Panel D of Figure 6, indicate that the proliferation of FHC silenced cells is reduce of about 35% compared to the controls It has been reported that in the 3' untranslated region of *c-Myc* mRNA there are multiple potential binding sites for Let-7 miRNAs family members. Moreover, the overexpression of Let-7 in cell cultures is accompanied by a decrease in *c-Myc* mRNA levels. Consequently, we determined by qRT-PCR the amounts of *c-Myc* mRNA in K562 cells silenced or not for FHC, finding that *c-Myc* mRNA was down-regulated to an extent of about 35% following FHC-silencing (Panel E of Figure 6).

Discussion

While the biochemical bases of ferritin function in iron uptake and deposition have been clearly established, and the respective roles of the two subunits determined, other aspects of its biological functions still remain to be clarified. Since the middle of last century, a robust body of data indicates that intracellular FHC is not only essential for iron metabolism but is also involved in critical metabolic pathways from the signalling cascades of CXCR4 [7] and G-CSFR [33] to Apo-B biogenesis [34].

We are interested in studying whether different intracellular amounts of FHC might affect gene expression profile of a given cell; proteome and transcriptome analysis has already revealed that the silencing of FHC is accompanied, in different cell types, by profound modifications in the steady-state amount of key proteins and transcripts [15, 27]. This phenomenon can be at least partially attributed to perturbations of the oxidative state of the cell induced by FHC-silencing, but the type and the amount of transcripts potentially regulated by FHC suggest the existence of additional mechanisms that still require to be investigated. In this work, as first step toward the dissection of the molecular basis of FHC-modulated gene expression, we performed an integrated analysis of miRNA and mRNA expression patterns in K562 FHC-silenced cells.

We have utilised K562 cells in which the expression of FHC has been stably knocked-down by shRNA interference and whose transcriptome profile is already established [27]. By using a microRNA PCR Panel, we found that 4 out of 84 analysed miRNAs, namely hsa-let-7g-5p, hsa-let-7f-5p, hsa-let-7i-5p and hsa-miR-125b-5p, are consistently and significantly up-regulated in FHC-silenced K562 cells compared to control cells. The correlation among FHC amounts and the expression of the four miRNAs is further supported by transient silencing and reconstitution experiments.

The Let-7 human miRNA family is composed by 14 members widely considered as tumor suppressors. Let-7 miRNAs regulate, among others, the expression of the oncogenes *Ras* [35], *Myc* [36, 37] and *HMGA2* [38]; accordingly, we found that, in FHC-silenced K562 cells, the up-regulation of Let7-g, -f and -i, is accompanied by an important reduction of *Myc* expression.

Different members of the Let-7 family regulate highly overlapping set of genes, thus suggesting a redundant function. On the other hand, the regulation of their expression is elicited at multiple levels, and, in certain cancers, only specific members appear to be deregulated [39]. Therefore, an emerging question is whether different members of a

miRNA family undergo a differential regulation within the same cell. Our data point in this direction, since FHC-silencing is accompanied, in K562 cells, by a selective up-regulation of three out of 9 Let-7 miRNAs analysed.

miR-125b, a member of the miR-125 family, is an intriguing molecule, acting either as tumor suppressor or as an oncogene in different cancer types [40, 41]. Recently, miR-125b has been utilized as biomarker to distinguish cell lines derived from acute (HL60) and chronic (K562) myeloid leukemias and it has been proposed that in K562 it may act as a tumor promoting agent [42].

Our results demonstrate that *RAF1*, one of the target genes of miR-125b, is down-regulated in the FHC-silenced K562 cells. Moreover, we have shown, in these cells, a reduced activation of pERK1/2, that plays a central role in all the pathways in which the miRNA-regulated genes are involved. ERK1/2 MAP kinases regulate growth, survival and cell cycle progression in mammalian cells upon phosphorylation-induced activation [43]. Our data show that *FHC* knock-down may negatively regulate, through the modulation of miR-125b expression, the ERK activation thus suggesting that, in our experimental model, hsa-miR-125b may prevalently act as tumor suppressor molecule. Consistent with this hypothesis is also the significant reduction in proliferation rate of FHC-silenced cells. A correlation among FHC levels, has-miR-125b and ERK1/2 activation is further supported by the decreased miRNAs amount (data not shown) and the augmented phosphorylation of the MAPK in FHC over-expressing cells (Panel C of Figure 6).

In this study, the profile of the four up-regulated miRNAs has been integrated with the transcriptome analysis by combining data obtained from the microRNA targets prediction software with a correlation-based approach. This test is based on the assumption that, since miRNAs tend to down-regulate the expression of their targets, the expression profiles of miRNAs are expected to be inversely related with those of their true target genes. This analysis led to the identification of 91 down-regulated genes, the majority of whom appear to be candidate targets of a single miRNA, while 15 are subjected to multiple miRNA regulation. IPA revealed that the highest scored pathways in which these genes are involved are: “Cell Death and Survival, Hematological System Development and Function, Hematopoiesis” and “DNA Replication, Recombination and Repair, Cell Cycle, Cancer”. The role of FHC in the processes of cell differentiation and neoplastic transformation has been investigated for a long time, starting from the observation that its intracellular amounts can significantly vary when comparing differentiated with

undifferentiated cells, or transformed versus non transformed cells [44, 5]. Both the central role of FHC in iron homeostasis and its ability in modulating different transduction pathways, are consistent with its increased expression during differentiation and neoplastic transformation. We believe that the identification of FHC-dependent miRNA/mRNA networks implies that different amounts of the ferritin subunit contribute, in K562 cells, to the remodelling of gene expression taking place during these cellular processes through the action of let-7g, let-7f, let-7i and miR-125b. This observation is further strengthened by the comparison of the miRNAs-regulated pathways, reported in this manuscript, with those highlighted in our previous work on FHC-silenced K562 undergoind differentiation [27]. It is interesting to note that, among the common pathways, "Cell Death and Survival" and "Hematological System Development and Function" rely on the ERK1/2 activation which is severely affected by FHC silencing.

In conclusion, the data presented in this study add a further level of complexity to the relationship among iron and miRNAs, demonstrating that the intracellular amounts of FHC subunit are able to regulate let-7g, let-7f, let-7i and miR-125b expression, as well as the repertoire of their down-stream genes. Even though an increasing body of evidence suggests that the redox state of the cell might significantly influence Let7 and 125-b levels [45, 46], we believe that the FHC interference on miRNA expression deserves further analysis.

Acknowledgments

This study was supported by the Italian Ministry of Education, University and Research (PON01_02834 – Prometeo grant from to GC; PRIN 2010-11, project number 2010NYKNS7 to SB) and by Fondazione Cariparo (Excellence Projects 2011-2012) to SB.

References

1. Beard JL, Connor JR, Jones BC. (1993) Iron in the brain. Nutr Rev 51: 157–170.
2. Arosio P, Ingrassia R, Cavadini P. (2009) Ferritins: a family of molecules for iron storage, antioxidation and more. Biochim Biophys Acta 1790(7):589-599.
3. Costanzo F, Colombo M, Staempfli S, Santoro C, Marone M, et al. (1986) Structure of gene and pseudogenes of human apoferritin HNucleic Acids Res 14:721–736.
4. Levi S, Luzzago A, Cesareni G, Cozzi A, Franceschinelli F. (1988) Mechanism of ferritin iron uptake: activity of the H-chain and deletion mapping of the ferro-oxidase site.A study of iron uptake and ferro-oxidase activity of human liver, recombinant H-chain ferritins, and of two H-chain deletion mutants. J Biol Chem 263(34):18086-18092.
5. Alkhateeb AA, Connor JR. (2013) The significance of ferritin in cancer: anti-oxidation, inflammation and tumorigenesis. Biochim Biophys Acta 1836(2):245-254.
6. Coffman LG, Parsonage D, D'Agostino R Jr, Torti FM, Torti SV. (2009) Regulatory effects of ferritin on angiogenesis. Proc Natl Acad Sci U S A 106(2):570-575.
7. Li R, Luo C, Mines M, Zhang J, Fan GH.(2006). Chemokine CXCL12 induces binding of ferritin heavy chain to the chemokine receptor CXCR4, alters CXCR4 signaling, and induces phosphorylation and nuclear translocation of ferritin heavy chain. J Biol Chem. 281(49):37616-27.
8. Bevilacqua MA, Costanzo F, Buonaguro L, Cimino F. (1988) Ferritin H and L mRNAs in human neoplastic tissues. Ital J Biochem 37(1):1-7.
9. Bevilacqua MA, Faniello MC, Quaresima B, Tiano MT, Giuliano P, et al. (1997) A common mechanism underlying the E1A repression and the cAMP stimulation of the H ferritin transcription. J Biol Chem 272(33):20736-20741.
10. Faniello MC, Di Sanzo M, Quaresima B, Baudi F, Di Caro V, et al. (2008) p53-mediated downregulation of H ferritin promoter transcriptional efficiency via NF-Y. Int J Biochem Cell Biol 40(10):2110-2109.
11. Wu KJ, Polack A, Dalla-Favera R. (1999) Coordinated regulation of iron-controlling genes, H-ferritin and IRP2, by c-MYC. Science 283(5402):676-679.

12. Faniello MC, Chirico G, Quaresima B, Cuda G, Allevato G, et al. (2002). An alternative model of H ferritin promoter transactivation by c-Jun. Biochem J 363(Pt 1):53-58.
13. Lee JH, Jang H, Cho EJ, Youn HD. (2009) Ferritin binds and activates p53 under oxidative stress Biochem Biophys Res Commun 389(3):399-404.
14. Tsuji Y, Miller LL, Miller SC, Torti SV, Torti FM. (1991) Tumor necrosis factor-alpha and interleukin 1-alpha regulate transferrin receptor in human diploid fibroblasts. Relationship to the induction of ferritin heavy chain.J Biol Chem. 266(11):7257-7261.
15. Di Sanzo M, Gaspari M, Misaggi R, Romeo F, Falbo L, et al.(2011) H ferritin gene silencing in a human metastatic melanoma cell line: a proteomic analysis. J Proteome Res 10(12):5444-5453.
16. Iwasaki K, Mackenzie EL, Hailemariam K, Sakamoto K, Tsuji Y. (2006) Hemin-mediated regulation of an antioxidant-responsive element of the human ferritin H gene and role of Ref-1 during erythroid differentiation of K562 cells. Mol Cell Biol 26(7):2845-56.
17. Bevilacqua MA, Faniello MC, Iovine B, Russo T, Cimino F et al. (2002) Transcription factor NF-Y regulates differentiation of CaCo-2 cells. Arch Biochem Biophys 407: 39–44.
18. Festa M, Ricciardelli G, Mele G, Pietropaolo C, Ruffo A, et al. (2000) Overexpression of H ferritin and up-regulation of iron regulatory protein genes during differentiation of 3T3-L1 pre-adipocytes. J Biol Chem 275(47):36708-36712.
19. Kim VN, Han J, Siomi MC. (2009) Biogenesis of small RNAs in animals. Nat Rev Mol Cell Biol 10(2):126-139.
20. Friedman RC, Farh KK, Burge CB, Bartel DP. (2009) Most mammalian mRNAs are conserved targets of microRNAs. Genome Research 19: 92–105.
21. Kim VN. (2005) MicroRNA biogenesis: coordinated cropping and dicing. Nature Reviews Molecular Cell Biology 6: 376–385.
22. Montagner S, Dehó L, Monticelli S. (2014) MicroRNAs in hematopoietic development BMC Immunol 15:14.
23. O'Connell RM, Zhao JL, Rao DS. (2011) MicroRNA function in myeloid biology. Blood 118(11):2960-2969.

24. Calin GA, Croce CM. (2006) MicroRNA-cancer connection: the beginning of a new tale. Cancer Res 66(15):7390-7394.
25. Jansson MD, Lund AH. 2012 MicroRNA and cancer. Mol Oncol 6(6):590-610.
26. Davis M, Clarke S. (2013) Influence of microRNA on the maintenance of human iron metabolism. Nutrients 5(7):2611-2628.
27. Misaggi R, Di Sanzo M, Cosentino C, Bond HM, Scumaci D, et al. (2014) Identification of H ferritin-dependent and independent genes in K562 differentiating cells by targeted gene silencing and expression profiling. Gene 535(2):327-335.
28. Smyth GK. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 3:Article3.Wu et al. 2010
29. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13(11):2498-2504.
30. J. Wang, S. Sen. (2011) MicroRNA functional network in pancreatic cancer: from biology to biomarkers of disease. J. Biosci 36(3):481-91.
31. Bradford MM. (1976) A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. Anal Biochem 72:248-254.
32. Zhang W, Liu HT. (2002) MAPK signal pathways in the regulation of cell proliferation in mammalian cells. Cell Research 12(1):9-18.
33. Yuan X, Cong Y, Hao J, Shan Y, Zhao Z, et al. (2004) Regulation of LIP level and ROS formation through interaction of H-ferritin with G-CSF receptor. J Mol Biol 339(1):131-144.
34. Rashid KA, Hevi S, Chen Y, Le Cahérec F, Chuck SL. (2002) A proteomic approach identifies proteins in hepatocytes that bind nascent apolipoprotein B. J Biol Chem 277(24):22010-22017.
35. Johnson SM, Grosshans H, Shingara J, Byrom M, Jarvis R, et al. (2005) RAS is regulated by the let-7 microRNA family. Cell 120(5):635-647.
36. Akao Y, Nakagawa Y, Naoe T. (2006) let-7 microRNA functions as a potential growth suppressor in human colon cancer cells. Biol Pharm Bull 29(5):903-906.

37. Sampson VB, Rong NH, Han J, Yang Q, Aris V, et al (2007) MicroRNA let-7a down-regulates MYC and reverts MYC-induced growth in Burkitt lymphoma cells. Cancer Res 67(20):9762-9770.
38. Lee YS, Dutta A. (2007) The tumor suppressor microRNA *let-7* represses the HMGA2 oncogene. Genes Dev 21(9):1025-1030.
39. Boyerinas B, Park SM, Hau A, Murmann AE, Peter ME. (2010) The role of let 7 in cell differentiation and cancer. Endocr Relat Cancer 17:F19-F36.
40. Banzhaf-Strathmann J, Edbauer D. (2014) Good guy or bad guy: the opposing roles of microRNA 125b in cancer. Cell Commun Signal 12:30.
41. Bousquet M et al. (2008) Myeloid cell differentiation arrest by miR-125b-1 in myelodysplastic syndrome and acute myeloid leukemia with the t(2;11)(p21;q23) translocation. J Exp Med. 205(11):2499-2506.
42. Xiong Q, Yang Y, Wang H, Li J, Wang S et al. (2014) Characterization of miRNomes in acute and chronic myeloid leukemia cell lines. Genomics Proteomics Bioinformatics 12(2):79-91.
43. Roskoski R Jr. (2012) ERK1/2 MAP kinases: structure, function, and regulation. Pharmacol Res 66(2):105-143.
44. Coccia EM, Stellacci E, Orsatti R, Testa U, Battistini A. (1995) Regulation of ferritin H-chain expression in differentiating Friend leukemia cells. Blood. 86(4):1570-1579.
45. Hou W, Tian Q, Steuerwald NM, Schrum LW, Bonkovsky HL. (2012) The let-7 microRNA enhances heme oxygenase-1 by suppressing Bach1 and attenuates oxidant injury in human hepatocytes. Biochim Biophys Acta. 2012 Nov-Dec;1819(11-12):1113-1122.
46. Manca S, Magrelli A, Cialfi S, Lefort K, Ambra R, et al. (2011) Oxidative stress activation of miR-125b is part of the molecular switch for Hailey-Hailey disease manifestation. Exp Dermatol 20(11):932-937.

Chapter 4

The hypothesis that miRNAs could regulate alternative translation of many mRNAs is indirectly supported by the integration of the two types of evidence: the multiple and non-canonical active ORFs in human mRNAs, provided by GTI-seq data, and miRNA-mRNA binding outside 3'UTRs thanks to CLASH technique. Looking for overlapping region identified in both the experimental evidences, we identified many genes in which one or more miRNAs could interfere with translation of main annotated ORF or with ORFs located in the 5' UTR respectively to the annotated ORF, or even downstream it. These regions are evolutionary conserved and the miRNA footprints tend to overlap mRNA regions involved in RNA fold stabilization. We selected most significant cases of genes and we provided experimental evidence of TIS activity. We obtained direct evidence of miRNA-TIS interaction causing suppression of protein expression in 60% of tested cases. This non canonical miRNAs function surely deserve further investigation, to better characterize the mechanisms of AT regulation, more specifically we need to understand if and how the miRNA-based regulation of mRNA alternative translation impact on cell processes and on disease.

Nowadays multiple proofs of functional and regulatory role of the non coding have been adduced and it is always more evident that there are a huge redundancy of control throughout all the different steps of extended variety of biological networks. miRNAs appear key modulators of information and understanding the interplay of miRNAs and DNA, coding RNA or other targeted elements is crucial. Further studies need to be performed to elucidate post-transcriptional and transcriptional role of miRNAs. Profiling miRNAs that are bound to their target have already been useful, as we discovered potential new miRNAs function thanks to the integration of data from CLASH technique and from GTI-seq. Further experimental data on active miRNA binding, as HITS-CLIP or PAR-CLIP, would adduce proof of their functionality and would not be mere prediction, rather experimental evidences. Experimental evidences of new functions would drive to targeted sequencing experiments and, improving the technique potential.

Are miRNAs also important regulators of alternative translation?

Bortoluzzi S[1,2]*, Saccoman C[1], Cagnin S[1,2], Chemello F[1], Bronte V[3]*, and Bisognin A[1]

1 Department of Biology, University of Padova, via G. Colombo 3, 35131, Padova, Italy
2 Centro di Ricerca Interdipartimentale per le Biotecnologie Innovative CRIBI, University of Padova, via G. Colombo 3, 35131, Padova, Italy
2 Department of Pathology and Diagnostics, Verona University Hospital, Piazzale A. Scuro 10 - 37134 Verona, Italy

*Corresponding Authors
Dr. Stefania Bortoluzzi, Department of Biology, University of Padova, via G. Colombo 3, 35131, Padova, Italy
Phone +39 049 827 6502
Fax +39 049 827 6209
Email: stefania.bortoluzzi@unipd.it

Prof. Vincenzo Bronte
Department of Pathology and Diagnostics, Verona University Hospital, Piazzale A. Scuro 10 - 37134 Verona, Italy
Phone: +39-045-8124007;
Fax: +39-045-8126455;
E-mail: vincenzo.bronte@univr.it

Abstract

We previously demonstrated that non-canonical binding of a miR-142-3p to a translation initiation site of C/EBPβ multi-ORF mRNA is able to change the ratio between protein isoforms with different properties, thus impacting on the cell phenotype. There is an increasing appreciation of the high prevalence of alternative translation in mammals. Complex translation patterns are known, with multiple ORFs in the same mRNA that can influence each other in different ways. New evidence indicates that miRNAs can frequently bind 5'UTR and coding regions of mRNAs. We provide novel data on the overlap of active translation initiation sites (TISs) of mRNAs, experimentally defined using GTI-seq, with miRNA-binding sites, experimentally determined using CLASH technique. We identified many genes in which one or more miRNAs could interfere with translation of the main annotated ORF, or with ORFs starting in the 5' UTR and downstream the main TIS. We propose a new mechanism relevant to increase our understanding of the complex cross talk between protein-coding and non-protein coding RNAs. We model how the binding of a miRNA to a TIS can produce different regulatory effects, according to the involved ORF types coexisting in mRNAs, and to their regulatory relations. Increased evolutionary conservation of miRNA footprints overlapping TISs, their propensity to fall in regions stabilizing mRNA fold, as well as direct experimental evidence of miRNA-mRNA interactions corroborate the proposed hypothesis. The miRNA-based regulation of mRNA alternative translation surely deserves further investigation to clarify if and how it impacts on cell processes and on disease.

ABBREVIATIONS LIST

5’TIS: 5’ Translation Initiation Site
AT: Alternative Translation
aTIS: annotated Translation Initiation Site
BP: Biological Process (Gene Ontology)
CC: Cellular Components (Gene Ontology)
CDS: Coding Sequence
CLASH: Cross Linking Ligation And Sequencing of Hybrids
CLIP: Cross Linking Immuno Precipitation
dTIS: downstream Translation Initiation Site
GTI-seq: Global Translation Initiation sequencing
IRES: Internal Ribosome Entry Site
LRS: Leak Ribosomal Scanning
MF: Molecular Function (Gene Ontology)
miRNA: microRNA
NMD: Nonsense Mediated Decay
ORF: Open Reading Frame
PIC: Pre Initiation Complex
TIS: Translation Initiation Site
uORF: upstream Open Reading Frame
UTR: Untraslated Region

Introduction

Non-canonical binding of a miRNA to the translation initiation site is able to change isoforms ratio of an mRNA whose alternative translation can generate multiple protein isoforms with different properties. This equilibrium consequently determines cell fate and ultimately induces phenotype changes relevant to cancer-induced immune tolerance (Sonda et al., 2013).

There is an increasing appreciation of the high prevalence of alternative translation (AT) in mammals (Kochetov, 2008;Menschaert et al., 2013;Smith et al., 2005;Vanderperre et al., 2013;Wang et al., 2004). The importance of protein synthesis studies is increasing, with the growing list of genetic diseases caused by mutations that affect mRNA translation (Valasek, 2012). The emerging scenario shows complex translation patterns of many mRNAs, in which multiple ORFs influence each other and/or the formation of secondary structures that can modulate ribosome activity (Morris and Geballe, 2000;Skabkin et al., 2013). On the other hand, new data indicate that miRNAs act in a non-canonical ways to regulate gene expression at different levels (Kosaka et al., 2013;Mittal and Zavolan, 2014) and they can bind more frequently 5'UTR and coding sequences (CDS) regions of mRNAs than 3'UTR, where canonical target sites are expected.

In this manuscript we discuss whether and how much miRNAs can contribute to the regulation of AT and we provide new data supporting the pervasiveness and perhaps the biological significance of this intriguing miRNA function.

AT: a widespread post-transcriptional regulation mechanism with underappreciated complexity

Proteome complexity in terms of multiplicity of different functional protein isoforms emerges not only from alternative splicing of RNA transcripts and the use of alternative promoters, but also from use of alternative translation initiation sites (TIS). The importance of AT in eukaryotes, both in terms of prevalence and functional relevance, is becoming apparent from recent studies, which added considerable evidence to earlier reports on AT in either specific genes or gene categories.

Translation can be divided mechanistically into three steps: initiation, elongation and termination. As the rate-limiting step of translation, initiation involves ribosome loading,

scanning, and TIS selection before elongation commitment. Normally, translation initiation requires the pre-initiation complex (PIC) assembly and recruitment to the mRNA, which is a CAP-dependent process. The PIC then scans the 5' UTR of the mRNA until it encounters a start codon. Normally it is assumed that the first AUG encountered by the ribosome serves as TIS. However, one or more potential initiation sites could exist upstream of the main ORF, forming uORFs. Likewise, TIS downstream the main start codon could also potentially serve as initiators.

After the elongation step, the translation stops when the ribosome encounters a stop codon, and termination occurs with the concerted action of release factors, that triggers peptide release, tRNA dissociation and ribosome separation. In some cases, the 40S subunit remains associated with the mRNA and could re-initiate translation from a downstream TIS (re-initiation).

Inefficient recognition of a TIS can lead to initiation downstream (leaky ribosomal scanning; LRS). On the basis of LRS model, the presence of uORFs in mRNAs (at least one uORF was reported in 50% of mammalian transcripts), can suppress the translation efficiency of the main ORF. In other cases, the uORF translation could also stimulate the translation of the main ORF (ATF4) (Kochetov et al. 2008).

According to the LRS model, the TIS sequence context is deemed to be important in determining the strength of a given AUG codon, with optimal (GCCRCCAUGG) and suboptimal consensus sequences known. Bioinformatic analyses showed that downstream alternative TIS (dTIS) following a suboptimal TIS are under negative selection in vertebrates, with stronger selection in genes with weaker primary TIS (Bazykin and Kochetov, 2011). Moreover, the authors of this study noticed that genes with multiple conserved TISs are enriched for olfactory receptors and transcription factors. It was also proposed that other mRNA features, such as stable secondary structures over or near the TIS, could also influence AUG recognition. In addition to these *cis* sequence elements, the stringency of TIS selection is also subject to regulation by *trans-acting* factors such as eIF1 and eIF1A.

Apparently, different ORFs of the same mRNA can regulate or influence each other in complex ways.

By definition, uORFs encode peptide different from that encoded by the main ORF, whereas alternative overlapping ORFs can be in frame or out of frame. According to the leaky ribosomal scanning model, suboptimal 5' TIS are not used with 100% efficiency and the ribosome goes ahead to start translation from downstream TIS. Different in frame

ORFs *(with shorter ORFs being suffixes of the longer ORF)* can generate proteins sharing the C-terminal part and differing in the N-terminal domains. The alternative inclusion of functional domains in different isoforms may differentiate, to some extent, their functional activity. For instance, two isoforms (Ora1α and Ora1β) arising from in frame alternative translation of the plasma membrane store-operating channel Ora1 mRNA display distinct plasma membrane mobility, since the shorter isoforms lack the phospholipid-binding domain (Fukushima et al., 2012). Recently, PTENα, a N-terminal extended isoform of PTEN, translated through alternative initiation at an upstream in frame TIS was discovered (Liang et al., 2014). Isoforms with different N-terminal signals can be targeted to distinct cell compartments: RNase Z isoforms display dual nuclear/mitochondrial targeting since the shortest isoform does not include the mitochondrial targeting sequence.

Moreover, ORFs in different frames may produce, by AT of the same mRNA, proteins with completely different sequences and functions. A relevant example is AltPrP, a protein produced by alternative use of a downstream out of frame TIS of the PrP (prion protein) mRNA, with a different amino acid sequence from the PrP (Vanderperre et al., 2011). AltPrP is a protein integrated in the outer mitochondrial membrane. AltPrP expression from PrP cDNA is constitutively negatively regulated and it is increased by proteasome inhibition and endoplasmic reticulum stress.

Recently, it was shown that cells proteome includes many previously disregarded small peptides (Slavoff et al., 2013) and the products of AT were directly detected by proteomic studies (Slavoff et al., 2013; Vanderperre et al., 2013), thus supporting the high prevalence of AT in Eukaryotes.

Many human genes undergoing alternative translation play important roles in tumorigenesis and development. Notable examples include OCT4, a transcription factor with a pivotal role in embryonic stem cells self renewal (Cao et al., 2009;Gao et al., 2012;Wang et al., 2009), and the massively regulated, multi isoform, and multi functional P53 gene (Candeias et al., 2006).

It is worth notice also that intronless genes cannot use alternative splicing to generate multiple proteins and AT is used instead as main mechanism for diversity generation. Many human genes as histone genes and G-protein-coupled receptor genes are predominantly intronless (Grzybowska, 2012). Besides, about 70% of single-copy, primate-specific human transcriptional units are intronless (Tay et al., 2009).

As anticipated, another important mechanism of AT is re-initiation (Jackson et al., 2012). Translation can be described as a cyclical process, consisting of initiation, elongation, termination, and ribosome recycling. In some cases, the recycling of post-translation complexes is incomplete: the 40S ribosomal subunit remains bound to mRNA, and termination is followed by re-initiation, usually downstream of the stop codon (Skabkin et al., 2013). In yeast, eIFs binding to the large ribosomal subunit lasts for several rounds of elongation and critically enhances the re-initiation capacity of post-termination 40S ribosomes (Szamecz et al., 2008;Valasek, 2012). In this view, the translation of small upstream ORFs (uORFs) can modulate the efficiency of use of downstream start sites. The uORFs are translated using TISs located 5' (5' TIS) to the main/annotated TIS (aTIS). As many as 44% of human mRNAs 5'UTRs include one or more uORFs (Kochetov et al., 2008). An uORFs shorter than 30 codons and the intercistronic distance (between uORF and the next TIS) longer than 37 nucleotides have been shown to favor efficient re-initiation (Luukkonen et al. 1995).

The uORFs are quite common in certain classes of genes, including two-thirds of oncogenes and many other genes involved in the control of cellular growth and differentiation. They can influence the translation of downstream ORFs with different mechanisms (Morris and Geballe, 2000). At least three possible fates are available to a ribosome after translating an uORF: 1) The ribosome may remain associated with the mRNA, continue scanning, and re-initiate further downstream, at either a proximal or distal AUG codon; 2) Another option for the ribosome is to stall during either the elongation or termination phase of uORF translation, creating a blockade to additional ribosome scanning (in the known cases, ribosome stalling is mediated by the peptide structure encoded by the uORF); 3) In addition to influencing the action of ribosomes during and after termination, uORFs may affect gene expression by altering mRNA stability.

Translation can be also initiated in a 5'-CAP recognition-independent way by the usage of secondary structures known as internal ribosome entry sites (IRES). IRES sequence elements form complex secondary structures that directly recruit ribosomes and drive translation particularly in special conditions as cellular stress, apoptosis or hypoxia (Liu and Qian, 2014).

After the use of an IRES, both LRS and re-initiation are possible, in principle (Kochetov et al., 2008). In addition, non AUG codons can be used for TIS, and they are frequently

used in 5'TIS (Touriol et al., 2003), an example being the extended-C/EBPα isoform, which is translated from a non-AUG codon (Muller et al., 2010).

A prediction of all possible short and long ORF for each mRNA produced by the human genome is feasible, but is expected to result in a massive number of putative AUG and non-AUG TISs, including an exceedingly large proportion of false positives. A recent study exploited global translation initiation sequencing (GTI-seq), to achieve detection of alternative translation initiation in mammalian cells with single-nucleotide resolution (Lee et al., 2012). This study used cycloheximide (CHX) and lactimidomycin (LTM) to selectively inhibit ribosome translocation and thus capture initiation events by deep sequencing of ribosome-protected mRNA fragments (RPF). RPF reads alignment to mRNA sequences defined read peaks pinpointing 16,863 active TIS in about 10,000 human transcripts and showed that 50% of the transcripts contain multiple TISs, supporting the idea of AT prevalence. In addition, 42% of transcripts showed an inactive aTISs, with translation initiating from different TISs, at least in cells considered in the study. Translation initiation downstream the aTIS was observed in 27% of transcripts with TIS peaks, with upstream TIS associated to suboptimal consensus sequences when both 5'TIS and aTIS were active. Besides, increasingly stable folded structures, shortly after the 5'TIS of transcripts with repressed aTIS initiation, were observed. In particular, more stable structure were present in correspondence of suboptimal 5'TIS codons, indicating the correlation of TIS selection with the sequence context of an optimal 5'TIS or with the intervention of secondary structures in TIS selection.

miR-142-3p non-canonical binding on a TIS controls AT of the C/EBPβ mRNA, thus promoting macrophage differentiation and acquisition of immunosuppressive function in cancer.

In a recent work, we identified a mini-circuit involving miR-142-3p, which is activated by tumor-released IL-6 family cytokines during tumor-induced myelopoiesis (Sonda et al., 2013). We demonstrated that miR-142-3p downregulation eventually promotes macrophage differentiation and acquisition of an immunosuppressive function in the tumor environment. The circuit includes the C/EBPβ intronless transcription factor gene. C/EBPs are a family of transcription factors that regulate the expression of tissue specific genes during differentiation of many cell types (Morris and Geballe, 2000).

The C/EBPβ mRNA is a known example of in frame AT. It encodes three protein isoforms that are produced by alternative use of aTIS and in-frame dTISs. These isoforms have different molecular weight and transcription activity: LAP* and LAP are considered to act as activating proteins, whereas the truncated LIP isoform, lacking the transcription activation domain is deemed an inhibitory isoform (Abreu and Sealy, 2010;Albergaria et al., 2013;Park et al., 2013). Indeed, the C/EBPβ mRNA presents a small out-of-frame regulatory uORFs immediately upstream the major TIS, modulating the ratio of isoforms produced by AT of in-frame ORFs (Morris and Geballe, 2000).

The miR-142-3p downregulates gp130 (the common subunit of the interleukin-6 cytokine receptor family involved in signaling after ligand interaction) by canonical binding to its mRNA 3' UTR and impacts on the ratio of the three C/EBPβ transcription factor isoforms, mainly by moving the equilibrium towards LIP. In turn, miR-142-3p transcription is controlled by the isoforms ratio, with LIP acting as transcriptional inducer.

A peculiar characteristic of this regulatory circuit is that miR-142-3p exerts its regulatory activity on C/EBPβ by a non canonical binding to the mRNA coding sequence, in a region including one of three in frame TIS (aTIS an dTIS1-2) of the C/EBP mRNA. The RNA duplex is stabilized by an imperfect pairing, involving 17/23 nt of the miRNA, with two bulges in the so-called seed region in the 5’ of the miRNA. In this way, miR-142-3p interferes with a complex equilibrium of regulated re-initiation, possibily involving, secondary structures and other modulator factors such as eIF2, and it plays a key role in the activation the macrophage differentiation and the acquisition of their immunosuppressive functions in cancer.

We asked whether this non-canonical function of a microRNA was a special case or rather the regulation of pervasive AT by miRNAs was a more widespread phenomenon. In this paper, we present new data supporting the latter hypothesis, suggesting that novel, non-canonical regulatory roles for miRNA related to AT regulation are possible.

miRNAs regulate gene expression at different levels with diverse mechanisms and they frequently bind mRNAs before 3'UTR

Several, non-canonical miRNA functions have been discovered in last years in terms of target recognition mode and regulatory actions. Regarding unusual regulatory actions, it is known that miRNAs can be imported in a signal-dependent way into the nucleus (Liao et al., 2010) and subsequently guide chromatin remodeling complexes to specific genome sites, eventually silencing gene transcription (Zardo et al., 2012) or promoting *de-novo* methylation of DNA (Sinkkonen et al., 2008). Furthermore, miRNAs may act as cell-to-cell communication mediators through exosomal vesicles, which are delivered to other cells and are detectable in body fluids (Chiba, 2012).

Regarding target recognition, it is generally accepted that miRNAs usually bind to the 3'UTR sequence of their mRNA targets, but targeting 5' untranslated regions (5'UTR) (Lytle et al., 2007) and CDS regions (Helwak et al., 2013;Huang et al., 2008;Tay et al., 2008) has been reported. Moreover, microRNA targets containing simultaneous 5'-UTR and 3'-UTR interaction sites are also known (Lee et al., 2009).

Anyway, the majority of characterized miRNA target sites are in the 3'UTRs of mRNAs, and large-scale studies examining the effects of either introducing or deleting a miRNA have shown that sites in 3'UTRs generally are more effectively suppressor of target expression than those in either 5'UTRs or CDS (Bartel, 2009). Besides, highly repetitive ORFs containing many miRNA sites can generally be subject to significant and, in some cases, substantial repression by the cognate miRNA and that repeats occur frequently within families of paralogous C_2H_2 zinc-finger genes (Schnall-Levin et al., 2011), suggesting the potential for their coordinated regulation.

Recent reports support that miRNA binding outside the 3'UTR is more common than previously expected. Liu et al. (2013) (Liu et al., 2013) used crosslinking immunoprecipitation (CLIP) for identification of a conspicuous number of argonaute-bound target sequences that contain miRNA binding sites. They described sequences, thermodynamic and target structure features essential for target binding by miRNAs in the 3' UTR, CDS and 5' UTR regions of target messenger RNA (mRNA) and showed that, out of 6,666 AGO tags, 61% mapped to 3′ UTRs, 37% to CDS, and 2% to 5′ UTRs.

In a recent seminal study, Helwak et al., 2013 (Helwak et al., 2013) exploited crosslinking, ligation, and sequencing of hybrids (CLASH) technique to unveil miRNA-

target pairs as chimeric reads in deep-sequencing data. More than 18,000 high-confidence miRNA-mRNA interactions were reported; around 60% of seed interactions are non-canonical, containing bulged or mismatched nucleotides. Overall, 60% of all target sites were mapped to the CDS, whereas 35% and 5% were mapped to the 3'UTR and to the 5'UTR, respectively. The proportions of miRNA targets associated by CLASH to the CDS are slightly higher compared to previous CLIP-seq experiments, likely because in this study the mapping of sequencing reads to a transcriptome database can consider target sites overlapping splice junctions. Notably, different miRNAs vary in the relative proportions of targets in 5′ UTRs, coding sequences, and 3′ UTRs. Targets in CDSs were shown to be significantly up-regulated upon miRNA depletion, and up-regulation of sites in the CDS is about half of that in 3′ UTRs (Helwak et al., 2013).

These reports suggest that miRNA binding outside the 3'UTR allow a miRNA-target interaction, which is non-canonical in terms of the mRNA region involved but canonical in terms of functional consequences, i.e. suppression or negative regulation of target expression. Interesting exceptions are provided by the previously cited study by (Sonda et al., 2013), that indicated a miR-142-3p-based isoform ratio regulation of the target gene and by Orom et al. (2008) (Ørom et al., 2008), who showed that miR-10a binds a TOP motif in the 5′UTR of ribosomal protein mRNAs and alleviates its translational repression during amino acid starvation, exerting eventually a positive control of global protein synthesis. In this case, the miRNA contributes to a complex translational regulatory system, and acting together with other, less characterized, regulators (Miloslavski et a., 2014).

Emerging data about the prevalence of AT and about the complexity of translation patterns, coupled with the new findings of frequent binding of miRNAs to 5'UTR and CDS regions of mRNAs suggest that, as we demonstrated for C/EBPβ mRNA and miR-142-3p, the miRNA binding to alternatively translated target mRNAs can interfere with AT regulation.

Hypothesis: miRNAs binding to mRNAs can interfere with alternative translation regulation in different ways

miRNA binding to either mRNA 5'UTR or to CDS may:

- Interfere with ribosome scanning of mRNAs, thus impacting on TIS recognition efficiency;
- Perturb the equilibrium among the efficiency of translation of in frame or out of frame AT ORFs;
- Reduce uORF translation, thus secondarily affecting the translation efficiency of downstream ORFs whose translation depends on re-initiation;
- Reduce the translation of uORFs inhibitory of downstream ORFs translation, ultimately increasing its translation efficiency.

As previously said, N-terminal proteomics and ribosome profiling data showed that alternative translation is highly prevalent in mammals. TISDdb reports that in average 1.4 and 2 translation initiation sites per mRNA are used in human and mice, respectively, considering only one cell type per species (Wan and Qian, 2014). According to Vanderperre et al. 2013, the majority (88%) of mRNAs present alternative ORFs (3.8 per mRNA in average) and alternative proteins represent the 55% of the proteome. Many eukaryotic proteins show N-terminal heterogeneity presumably due to AT (Liu and Qian, 2014) and, as previously discussed, studies on several genes, as C/EBP factors, Ora1 genes, PTEN, PrP and OCT4 isoforms, state that AT is has particularly relevant biological roles. In Figure 1 we exemplify possible multi ORF mRNAs structures and propose how miRNA interactions with different TISs can produce diverse regulatory effects, under different scenarios. Conceivably, the miRNA binding over a TIS can directly induce ribosome stalling and/or interact with secondary structures in the mRNA, and ultimately modulate the TIS recognition efficiency. miRNAs binding to the main TIS (aTIS) can interfere with ribosomal scanning (Figure 1A). In mRNA with multiple in frame, interdependent ORFs the miRNA can recognize and bind the aTIS, reduce the aTIS recognition efficiency and favor the usage of dTIS, perturbing the ratio of long and N-truncated isoforms. If the different ORFs are out of frame, the miRNA binding can act

as a switch and trigger the production of protein products with completely different sequence and properties.

It has been estimated that about 50% of mammalian transcripts contain at least one uORF (Calvo et al. 2009). In the same study, uORF-mediated translational regulation has been experimentally validated for one hundred eukaryotic transcripts, including around 30 human transcripts. uORFs can positively or negatively influence the translation of the following ORF(s) with different mechanisms, briefly reviewed above. The binding of a miRNA to a 5'TIS in an inhibitory uORF can affect downstream TIS usage (Figure 1 B): the miRNA-induced inhibition of uORF translation can stimulate the downstream aTIS usage. This example can fit either uORF whose translation is inhibitory of aTIS usage and uORF encoding peptides inhibitory of aTIS usage. If instead the translation starting from the 5'TIS increases the efficiency of the downstream TIS usage by the so-called re-initiation mechanism, the miRNA-induced inhibition of the uORF translation can decrease also the aTIS usage (Figure 1C).

Many mRNAs display both uORFs, multiple in frame (and/or out of frame) ORFs, and also ORFs located in the 3'UTR regions (defined according to the position on the main ORF). Figure 1 C sketches a more complex structure of mRNAs. The translation of these ORFs can be interdependently regulated and the binding of a miRNA to one of the TISs can produce different regulatory effects, according to the involved ORF types, and to their regulatory relations.

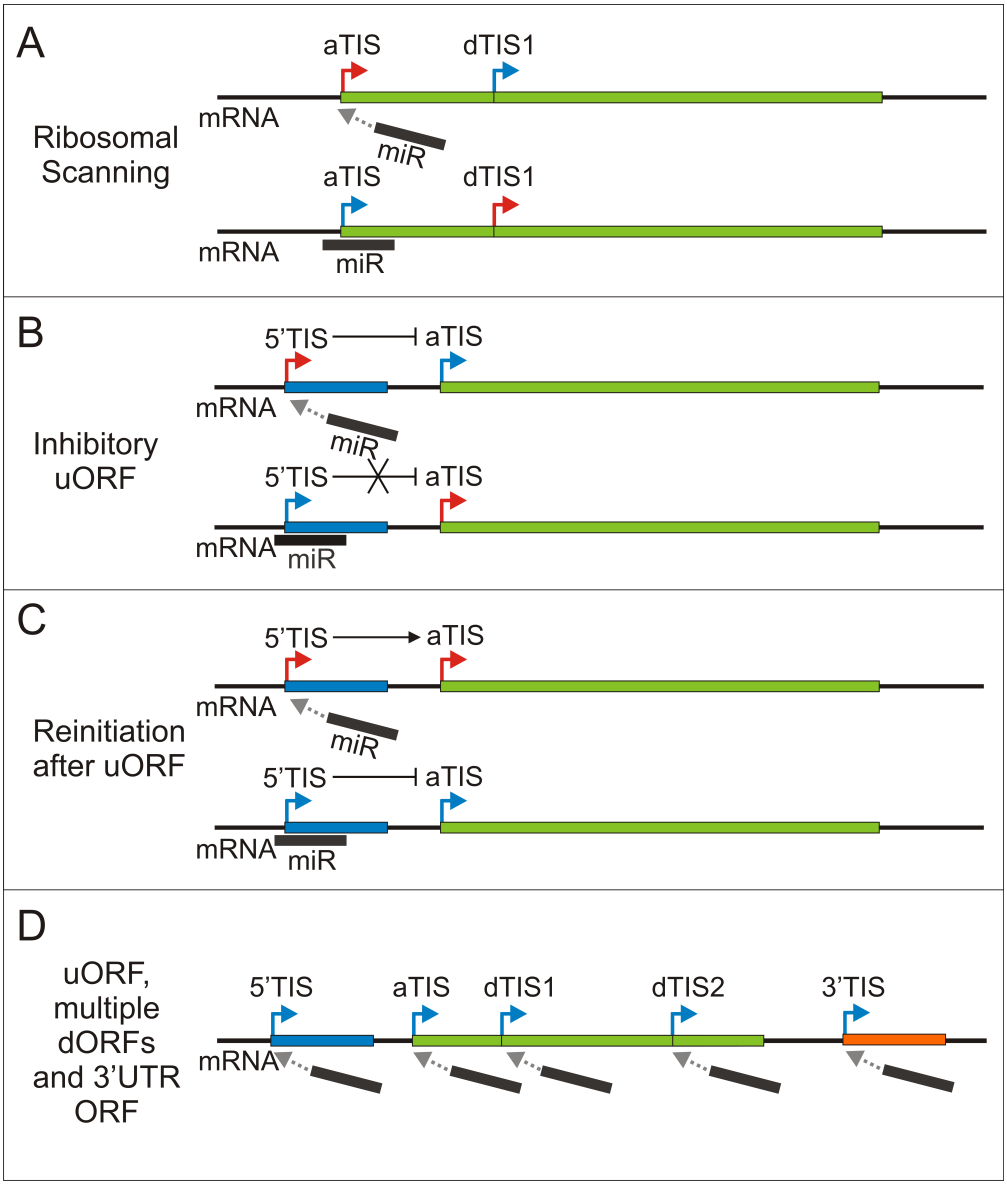

Figure 1. Examples of possible regulatory effects due to interactions between miRNAs and different multi ORF mRNAs structures, under the assumption that the miRNA binding over a TIS can induce either ribosome stalling or interact with secondary structures in the mRNA, and ultimately modulate the TIS recognition efficiency. A) miRNAs interference with ribosomal scanning. The pre-initiation complex scans the 5' UTR of the mRNA until it encounters a start codon: normally the first TIS is used. In mRNA with two alternative TISs, the miRNA binding to the most 5' TIS (annotated TIS, aTIS) can reduce its usage in favor of the downstream one (dTIS), thus perturbing the ratio between long and truncated protein isoforms. If the different ORFs are out of frame, the miRNA binding can trigger the production of a protein product with completely different sequence and properties. Many mRNAs present uORFs in the 5'UTR, which can affect the main TIS usage. B) Generally, 5' ORFs are inhibitory: in this case the miRNA binding to the 5'TIS can reduce uORF translation and thus stimulate the downstream aTIS usage. C) In other cases, the translation of the uORF starting from the 5'TIS is able to increase the efficiency of the downstream TIS usage, by the socalled re-initiation mechanism. The miRNA binding to the 5'TIS in mRNAs presenting re-initiation can decrease the aTIS usage. D) As known for several genes, mRNAs may present complex structures, including 5' and 3'TIS, along with one or more dTIS in addition to the aTIS: the translation of multiple ORFs can be interdependently regulated. In these cases, the binding of a miRNA to one of the TISs can produce different regulatory effects, according to the involved ORF types, and to their regulatory relations.

Experimentally determined miRNA binding sites overlap with active TISs in human mRNAs

We considered the information about 6,693 genes with at least one proven active TIS, according to GTI-seq experiments, and merged these data with results on miRNA binding to human RNAs (belonging to 6,957 genes) experimentally determined using clash technique, obtaining a group of 3,624 unique genes and 4,064 mRNAs in the merged dataset. As done before, and following the study by Lee et al (2012) providing the original translation initiation data, the position of the main ORF as reference defined the aTIS. We thus tagged as 5'TIS the TIS upstream the aTIS, dTIS the TIS downstream the aTIS (i.e. included in the main ORF), and 3'TIS the TIS associated to ORFs starting after the end of the main ORF and located in the 3'UTR (Figure 2A). It is worth notice that, in the considered group of genes, the majority (2,164, 59.7%) is associated with 2 or more different active TISs (Figure 2B). The ring plot in the same figure shows the proportions of different types of TISs in the considered merged datasets (Figure 2C).

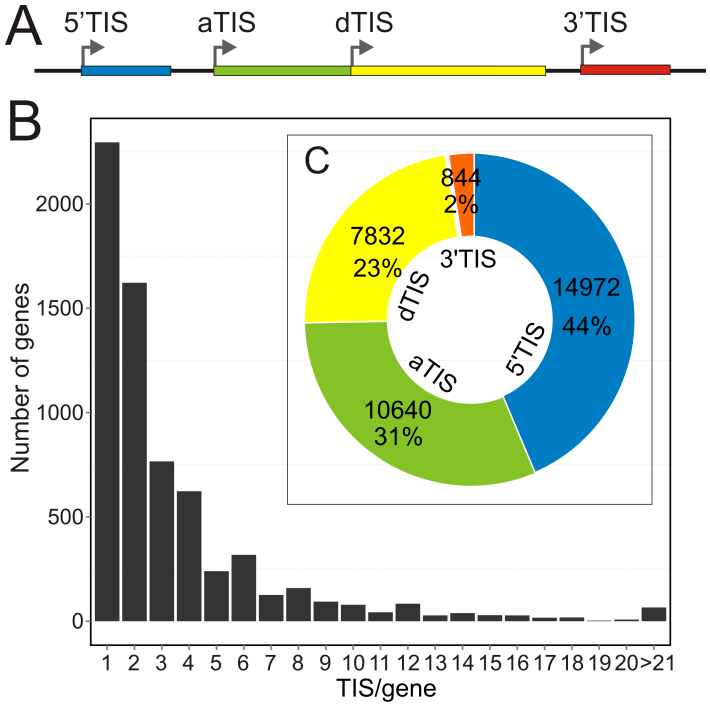

Figure 2. A) Schematization of the different types of TIS considered: the annotated TIS (aTIS) was defined by the position of the main annotated ORF, as in Lee et al (2012); 5'TIS are upstream the aTIS; dTISs are located downstream the aTIS but they are included in the main ORF; 3'TISs are located after the end of the main ORF, (i.e. in the mRNA region annotated as 3'UTR). B) Distribution of number of genes according to the number of experimentally determined TISs per gene. C) Proportions of different types of TISs in the 4,064 considered mRNAs.

We believe that the observation of numerous genes with multiple TISs can be only in minimal part due to the existence of several mRNAs belonging to the same gene, since our data include 1.2 mRNAs per gene, in average. Anyway, the following results are all based on analyses considering the correct data granularity, i.e. mRNA positions of TIS and miRNA binding sites.

We considered 10,775 mRNAs with at least one active TIS and 7,388 mRNAs that can bind a miRNA, for a total of 396 different miRNAs represented. The first data integration focused on the identification of the subset of mRNAs in which the miRNA binding site is overlapping an active TIS (i.e. at least one nt in the start codon is included in the miRNA binding region), found 264 possible interactions between 96 unique miRNAs (28.1% of the 342 unique miRNAs in the merged dataset) and active TISs. Involved TISs are located into 197 mRNAs, belonging to 197 unique genes, accounting for 5.4% of the 3,624 unique genes in the merged dataset.

Many genes with roles in translation and chromatin structure regulation are involved in possible miRNA-TIS interactions. The group of mRNAs in which at least one TIS is bound by a miRNA include many histone proteins, proteins involved in nucleosome and chromatin assembly and/or in DNA conformation change, ribosomal proteins, and other proteins involved in translation and translation regulation. We wondered whether the observed gene types could be biased due to the set of 3,624 genes included in the "merged dataset", restricted to genes represented both in the Lee et al (2012) GTI-seq dataset and in the Helwak et al., 2013 CLASH dataset. Thus we analysed the GO terms and Reactome pathways enrichment of the 197 genes involved in possible miRNA-TIS interactions using as background the whole set of genes in the "merged dataset". In this way, we exclude that the enriched categories were due to a bias in the composition of the dataset. Table 1 shows a selection of non-redundant GO categories and pathways significantly enriched (p-value <0.0001) in the set of 197 genes perhaps involved in miRNA-TIS interactions.

Functional Category Type	ID	Term	P value	Odds Ratio	Gene Count
GO BP	GO:0006414	translational elongation	9.45E-23	15.87	33
	GO:0000184	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	3.47E-22	14.07	34
	GO:0006415	translational termination	5.71E-22	16.72	31
	GO:0006614	SRP-dependent cotranslational protein targeting to membrane	4.00E-20	12.80	32
	GO:0006413	translational initiation	5.97E-19	10.23	34
	GO:0006612	protein targeting to membrane	2.43E-18	9.64	34
	GO:0000956	nuclear-transcribed mRNA catabolic process	6.92E-18	8.84	35
	GO:0016071	mRNA metabolic process	1.36E-12	3.84	53
	GO:0006886	intracellular protein transport	1.28E-09	3.36	45
	GO:0046907	intracellular transport	1.85E-09	2.85	61
	GO:0006334	nucleosome assembly	2.40E-07	7.17	14
	GO:0034728	nucleosome organization	5.99E-07	5.99	15
	GO:0090304	nucleic acid metabolic process	3.06E-05	1.88	100
	GO:0006323	DNA packaging	4.62E-05	4.14	14
GO CC	GO:0005840	Ribosome	7.93E-18	8.39	36
	GO:0030529	ribonucleoprotein complex	6.64E-18	4.99	59
	GO:0065010	extracellular membrane-bounded organelle	9.57E-09	2.68	61
	GO:0005829	Cytosol	9.92E-09	2.41	87
	GO:0000786	nucleosome	2.37E-08	11.95	12
	GO:0044815	DNA packaging complex	1.23E-07	9.76	12
GO MF	GO:0003735	structural constituent of ribosome	8.38E-19	10.57	33
	GO:0003723	RNA binding	3.67E-11	2.88	82
	GO:0003676	nucleic acid binding	2.01E-08	2.36	103
	GO:0097159	organic cyclic compound binding	2.96E-05	1.90	117
	GO:1901363	heterocyclic compound binding	4.28E-05	1.87	116
Reactome	156842	Eukaryotic Translation Elongation	3.12E-20	14.16	33
	975956	Nonsense Mediated Decay (NMD) independent of the Exon Junction	3.12E-20	14.16	33

		Complex (EJC)			
	72764	Eukaryotic Translation Termination	3.14E-19	14.29	31
	975957	Nonsense Mediated Decay (NMD) enhanced by the Exon Junction Complex (EJC)	3.49E-19	11.93	34
	157279	3' -UTR-mediated translational regulation	1.57E-18	11.13	34
	156827	L13a-mediated translational silencing of Ceruloplasmin expression	1.57E-18	11.13	34
	72689	Formation of a pool of free 40S subunits	2.22E-18	12.17	32
	72737	Cap-dependent Translation Initiation	1.60E-17	10.00	34
	72613	Eukaryotic Translation Initiation	1.60E-17	10.00	34
	72706	GTP hydrolysis and joining of the 60S ribosomal subunit	2.70E-17	10.24	33
	72766	Translation	5.80E-16	7.62	37
	74160	Gene Expression	1.04E-10	3.30	70
	1643685	Disease	1.67E-10	3.36	62
	72702	Ribosomal scanning and start codon recognition	5.68E-07	7.12	14
	212300	PRC2 methylates histones and DNA	3.91E-07	7.42	14
	4839726	Chromatin organization	4.44E-07	5.06	19
	171306	Packaging Of Telomere Ends	8.57E-06	8.25	10
	2559580	Oxidative Stress Induced Senescence	1.83E-05	4.52	15
	774815	Nucleosome assembly	2.30E-05	6.21	11
	606279	Deposition of new CENPA-containing nucleosomes at the centromere	2.30E-05	6.21	11
	157579	Telomere Maintenance	4.32E-05	5.68	11

Table 1. GO terms and Reactome pathways significantly enriched in the set of genes with active TISs overlapping miRNA-binding sites.

Interestingly, these genes are highly enriched in genes encoding for ribosomal proteins or for playing roles in translational initiation, elongation and termination, in the post-

transcriptional regulation of gene expression and in non-sense-mediated decay. Genes involved in possible miRNA-TIS interactions are also enriched in histone genes (nucleosome CC, nucleic acid binding MF, and DNA packaging BP), which are intronless, as mentioned above and other genes involved in gene expression regulation, as SET, encoding a protein that inhibits nucleosomes acetylation. At least 62 of the considered genes are known to be involved in one “disease” pathway, 70 in gene expression pathway, 15 in oxidative stress induced senescence and 11 in telomere maintenance and in nucleosome assembly.

Characterization of miRNA-TIS interactions

Position relative to TISs

In considered miRNA-TIS interactions, the overlap between miRNA-binding sites and TIS positions shows a slight but significant propensity for microRNA binding sites to lie more frequently downstream the TIS (11.1 and 13.4 Nt respectively in the region 5’ and 3’ to the TIS; binomial two sided test p-value=0.002).

Evolutionary conservation of miRNA footprints

We evaluated the conservation of miRNA binding sites in coding regions of mRNAs, using PhyloP 100 Vertebrate conservation as local basewise conservation score. Very interestingly, we found that the third (wobble) base conservation scores were significantly higher (11% more conserved) in miRNA binding sites than in the rest of coding sequence (t-test p-value = 0.002), suggesting that miRNA binding sites are under evolutionary selection.

Possible meddling in the mRNA folding

We investigated if miRNA binding sites tend to fall in regions that appear particularly important for the stabilization on the mRNA structure. To answer this question, for each mRNA, we predicted the minimum free energy (mfe) structure of the whole sequence, using RNAfold. We considered the part of the RNAfold "dot plot" base pairing matrix,

that gives pairing in the minimum free energy structure and compared it with the miRNA binding sites.

We observed that in 215/264 (81.4%) interactions at least one continuous stretch of 5 or more paired nucleotides is included in the miRNA binding region. In 72/264 (27.3%) interactions at least one stretch of 10 or more paired nucleotides is included in the miRNA binding region. In these cases, the miRNA attachment is predicted to perturb the folding. Panel C of the figure in Supplementary File 2 shows, for each miRNA-mRNA interaction (in abscissa), the boxplot of the distribution of the length of continuous stretches (1 or more nt) of paired nucleotides in the minimum free energy structure of the considered mRNA. For comparison, the longest stretch of paired nucleotides overlapping with the miRNA binding site is shown as red dot. Red dots fall over the median and over the third quartile value in respectively 82.6 and 57.6% of interactions. Since when the miRNA footprint contains two or more continuous stretches near each other we counted only the longest one, we probably underestimated the number of paired nucleotides involved in miRNA-binding that could interfere with mRNA 3D structure. We can conclude that miRNA binding regions tend to fall in regions stabilizing the RNA folding.

miRNA-TIS interactions classified by TIS type

The identified 264 putative miRNA-TIS interactions involve 64 5'TIS, 68 aTIS, 128 dTIS and 4 3'TIS (Figure 3A). Figure 3B reports the numbers of unique miRNAs according to the category of the TIS overlapping the miRNA-binding site. A specific miRNA can bind more than one position in the mRNA and thus the miRNA-binding sites can overlap several TISs of different categories, in the same or in different mRNAs. The Venn diagram in Figure 3C shows that some miRNAs have binding sites overlapping only one TIS category (we observed 27 miRNAs whose binding sites overlap only dTISs), whereas other groups of miRNA-binding sites overlap to two or more TIS categories.

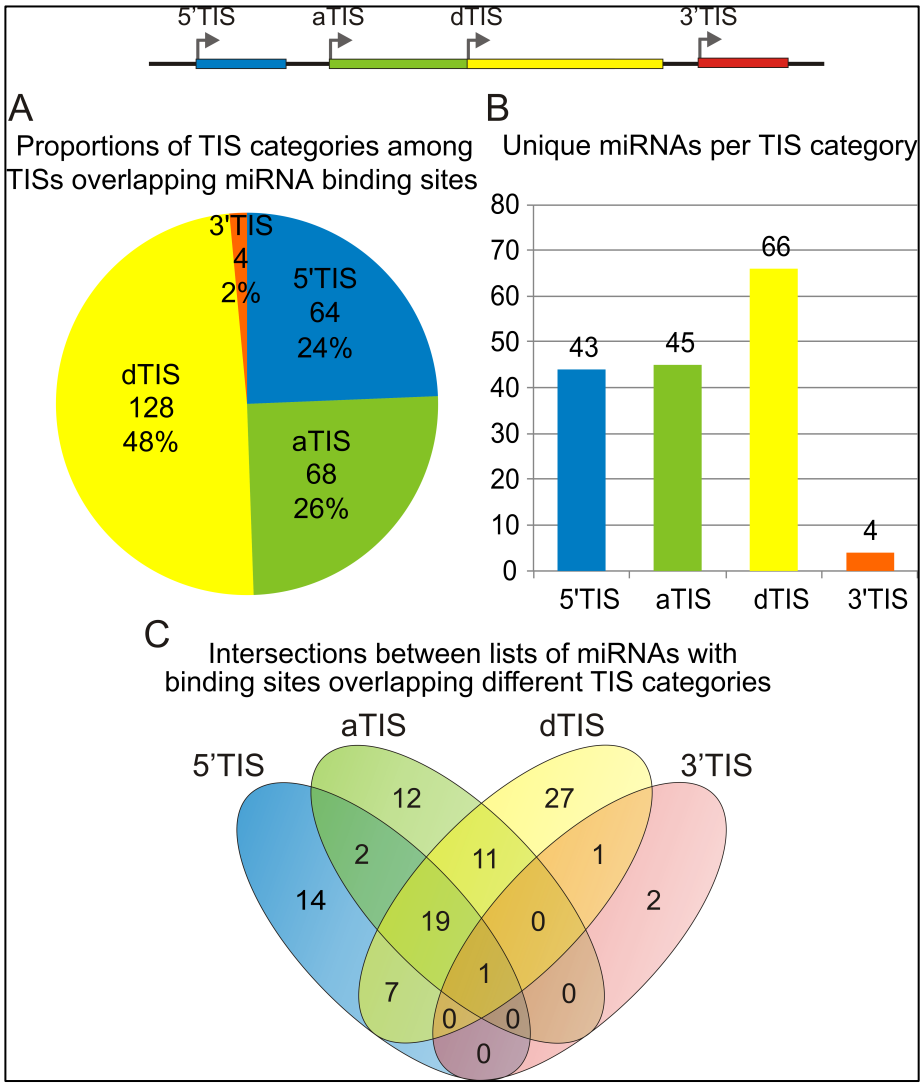

Figure 3. A total of 264 putative miRNA-TIS interactions were identified. A) Proportions of TISs types included in the set of active TISs overlapping miRNA-binding sites. B) Number of different miRNAs putatively interacting with each TIS type. C) Intersections of miRNA sets interacting with considered TIS types.

Only miR-92a-3p has experimentally determined binding sites that overlap all the four TIS types. Interestingly, many miRNAs can potentially interact with a few TISs in specific mRNAs, whereas others exhibit more pervasive potential interactions with many TISs, as miR-615-3p that overlaps 17 different TISs (Figure 4).

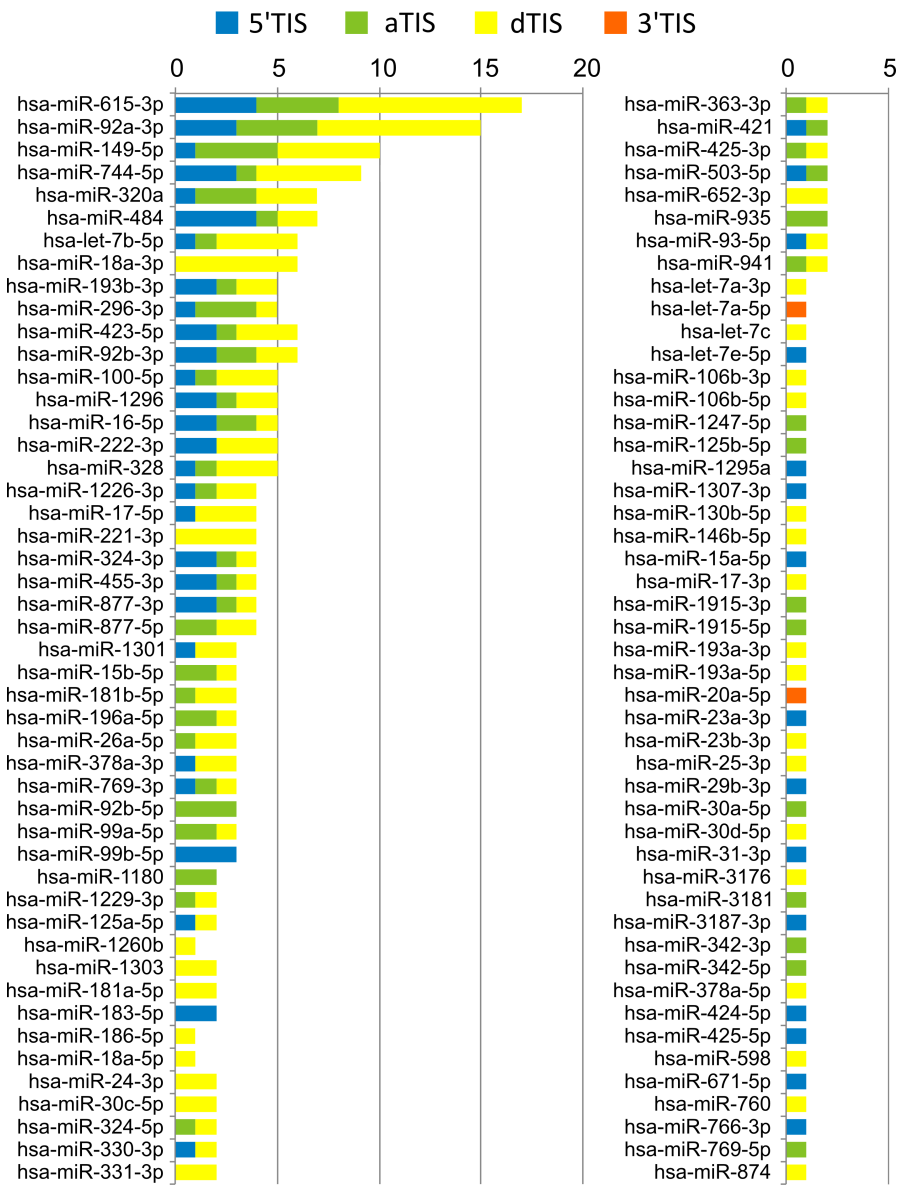

Figure 4. For each miRNA, the number of putative miRNA-TIS interactions are shown, with colour code indicating the TIS type.

According to our data, miR-92a-3p binding sites overlap with 15 different TIS (in 14 mRNAs) including 2 5'TISs, 4 aTIS and 8 dTIS and 1 3'TIS. On the other hand, we observed that 44 mRNAs have two or more miRNAs with a binding site overlapping a TIS. Of them, specific mRNAs, as those encoding HIST1H2BK and HIST1H3B can potentially interact with up to 6 different miRNAs (Table 2).

Gene symbol	mRNA refseq	miRNAs	
		Names	N
HIST1H2BK	NM_080593	hsa-miR-15a-5p, hsa-miR-16-5p, hsa-miR-320a, hsa-miR-424-5p, hsa-miR-423-5p, hsa-miR-1180	6
HIST1H3B	NM_003537	hsa-let-7b-5p, hsa-miR-320a, hsa-miR-330-3p, hsa-miR-193b-3p, hsa-miR-652-3p, hsa-miR-877-5p	6
RPS16	NM_001020	hsa-miR-17-3p, hsa-miR-24-3p, hsa-miR-149-5p, hsa-miR-1296, hsa-miR-1303	5
HIST1H4C	NM_003542	hsa-miR-17-5p, hsa-miR-100-5p, hsa-miR-425-3p, hsa-miR-874	4
RPS9	NM_001013	hsa-miR-342-3p, hsa-miR-423-5p, hsa-miR-935, hsa-miR-1296	4
EEF1G	NM_001404	hsa-miR-17-5p, hsa-miR-193a-3p, hsa-miR-18a-3p	3
HIST1H2BO	NM_003527	hsa-miR-320a, hsa-miR-296-3p, hsa-miR-92b-5p	3
RPL31	NM_000993	hsa-miR-100-5p, hsa-miR-18a-3p, hsa-miR-615-3p	3
RPL36A	NM_021029	hsa-miR-331-3p, hsa-miR-1226-3p, hsa-miR-1260b	3
RPS17	NM_001021	hsa-miR-222-3p, hsa-miR-320a, hsa-miR-615-3p	3
ASCC3	NM_006828	hsa-miR-425-5p, hsa-miR-1226-3p	2
ASNS	NM_001673	hsa-miR-17-5p, hsa-miR-23a-3p	2
CHCHD2	NM_016139	hsa-miR-221-3p, hsa-miR-1229-3p	2
DLST	NM_001933	hsa-miR-503-5p, hsa-miR-92b-5p	2
EIF4G1	NM_182917	hsa-miR-615-3p, hsa-miR-455-3p	2
GHITM	NM_014394	hsa-miR-92a-3p, hsa-miR-484	2
GNB2L1	NM_006098	hsa-miR-29b-3p, hsa-miR-183-5p	2
HIST1H3H	NM_003536	hsa-miR-320a, hsa-miR-378a-3p	2
HIST2H2BF	NM_001024599	hsa-miR-18a-3p, hsa-miR-760	2
LARS	NM_020117	hsa-miR-92a-3p, hsa-miR-92b-3p	2
NDUFA2	NM_002488	hsa-miR-99a-5p, hsa-miR-100-5p	2
NPM1	NM_002520	hsa-miR-320a, hsa-miR-296-3p	2
RPL12	NM_000976	hsa-let-7b-5p, hsa-miR-222-3p	2
RPL13A	NM_012423	hsa-miR-26a-5p, hsa-miR-92a-3p	2
RPL27	NM_000988	hsa-miR-92a-3p, hsa-miR-324-5p	2
RPL32	NM_001007074	hsa-miR-186-5p, hsa-miR-769-3p	2

RPL9	NM_000661	hsa-let-7a-5p, hsa-miR-18a-5p	2
RPLP1	NM_001003	hsa-miR-877-5p, hsa-miR-1180	2
RPS10	NM_001204091	hsa-miR-1296, hsa-miR-3176	2
RPS12	NM_001016	hsa-miR-30c-5p, hsa-miR-181b-5p	2
RPS23	NM_001025	hsa-miR-92a-3p, hsa-miR-1915-5p	2
RRM2	NM_001165931	hsa-miR-26a-5p, hsa-miR-30c-5p	2
TMX2	NM_015959	hsa-miR-196a-5p, hsa-miR-296-3p	2

Table 2. List of genes and mRNAs with binding sites for two or more different miRNAs overlapping with active TISs.

As shown in Table 3 A the mRNA of the intronless gene HIST1H4C, encoding a member of the histone H4 family, includes an active aTIS and 12 dTISs, three of which overlap the binding sites of four different miRNAs.

Gene Symbol	Gene ID	Gene description	RefSeq mRNA	TIS type	TIS position in mRNA	Position relatively to aTIS	Frame	ORF length	Start codon	overlapping miRNAs
A) mRNA possibly interacting with multiple miRNAs										
HIST1H4C	8364	Histone cluster 1, H4c	NM_003542	aTIS	1	1	1	104	ATG	-
				dTIS	85	85	1	76	GGC	hsa-miR-874
				dTIS	146	146	1	15	GTC	hsa-miR-100-5p, hsa-miR-425-3p
				dTIS	241	241	1	7	CTG	hsa-miR-17-5p
B) Active 5' TIS overlapping miRNA binding sites										
ASNS	440	Asparagine Synthetase	NM_001101	5'TIS	150	-102	0	30	CTG	miR-23A-3p
				5'TIS	210	-42	0	10	CTG	miR17-5p
				aTIS	252	1	0	562	ATG	-

C) Active aTIS, overlappping miRNA-binding sites and followed by active dTIS(s)										
TRAP1	10131	TNF Receptor-Associated Protein 1	NM_016292	aTIS	90	1	1	705	ATG	miR-149-5p
				dTIS	1338	1249	1	289	GTT	-
D) Not active aTIS, overlapping miRNA-binding sites and followed by active dTIS(s)										
SRSF9	8683	Serine/arginine-rich splicing factor 9	NM_003769	aTIS	147	1	1	221	ATG	miR-935
				dTIS	549	403	1	88	GGG	miR-30d-5p
E) Active dTISs overlapping with miRNA-binding sites										
VMP1	81671	Vacuole Membrane Protein 1	NM_030938	aTIS	274	1	1	407	ATG	-
				dTIS	316	43	1	393	ATG	miR-25-3p

Table 3. Selected examples of genes with miRNA-mRNA interactions involving TISs. In the table we report active TISs overlapping miRNA-binding sites, plus the reference aTIS (main ORF) position also if non active and/or non overlapping miRNA binding sites. A) The mRNA of the HIST1H4C presents 12 different active dTIS (data not shown) and 3 of them overlap with the binding site of a miRNA. B) In the ASNS mRNA, two miRNAs can interact with two active 5'TISs. C) In the TRAP1 mRNA the active aTIS overlaps the miR-149-5p binding site and it is followed by at least one active dTIS. D) SRSF9 is an example of mRNA presenting a miRNA binding site overlapping an inactive aTIS, which is followed by active dTIS(s). E) VMP1 mRNA presents two active dTIS (data not shown), one of which, in frame with the aTIS. possibly interacts with a hsa-miR-25-3p.*The aTIS (main ORF) is reported for reference also if it is non-active and/or non-overlapping miRNA binding sites.

miRNA binding sites overlapping active 5' TISs

We identified 43 miRNAs whose binding sites overlap one of 64 TISs located in the 5'UTR of one of 55 different mRNAs. Supplementary Files 3 and 4 report details regarding 55 involved genes and the results of functional enrichments. This group comprises genes encoding proteins regulating translation and involved in RNA splicing. Enriched Reactome pathways comprise Gap junction trafficking and regulation, Gap junction degradation, as well as Cell-extracellular matrix interactions, Adherens junctions interactions, Deadenylation-dependent mRNA decay, and Nonsense-Mediated Decay (NMD).

An interesting example of mRNA in which a 5' TISs can interact with miRNAs is ASNS, encoding Asparagine Synthetase (Table 3 B). hsa-miR-23a-3p and hsa-miR-17-5p binding sites overlap two different 5'TISs, directing the translation of uORFs.

miRNA binding sites overlapping active aTISs

As said before, aTIS in 58 mRNAs (58 genes) are overlapping binding sites of 45 different miRNAs, for a total of 68 possible interactions. Of these aTIS, 33 (48.5%) are associated to one or more additional active TIS included in the main ORF (dTIS). Figure 5 shows that the majority of dTIS (22, 66.7% of mRNAs with active aTIS and dTIS(s) are in frame with the aTIS, and thus encode proteins differing for the N-terminal region from the main isoform. In 14 mRNAs we observed multiple dTIS in frame with the aTIS. The remaining cases are out of frame and encode totally different proteins, 7 with single dTIS out of frame and 4 mRNAs presenting more than one dTIS out of frame. In total, 18 mRNAs present multiple dTISs active after an active aTIS (Figure 5)..

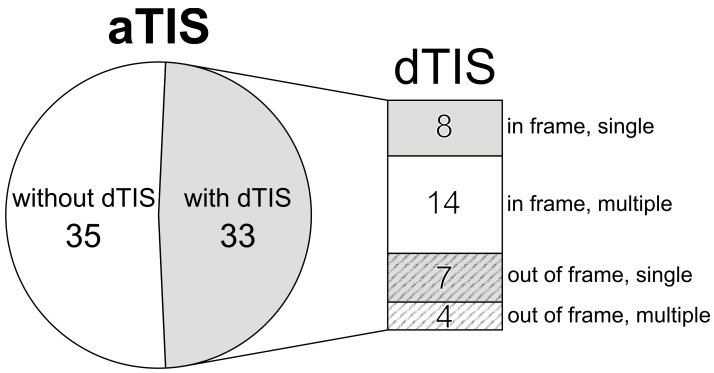

Figure 5. Among mRNAs with aTIS putatively interacting with miRNAs in our dataset, one half are associated to active dTIS. The majority of dTIS are in frame with the aTIS, and thus encode proteins differing for the N-terminal region from the main isoform. In 23 mRNAs we observed multiple active dTIS in frame with the aTIS. The remaining cases are out of frame and encode totally different proteins, partly with a single dTIS out of frame and partly presenting more than one dTIS out of frame. Moreover, 28 mRNAs present multiple dTISs active after an active aTIS.

The enrichment results regarding genes with aTISs putatively interacting with miRNAs are similar to those obtained considering the whole set of genes. The most represented elements are constituent of the ribosome and proteins involved in either DNA packaging or in non-sense mediated RNA decay (Supplementary Files 3 and 4). Among possible miRNA-aTIS interactions we considered the example of TRAP1 (Table 3 C). This gene encodes for TNF Receptor Associated Protein 1, a mitochondrial chaperone (member of the heat shock protein 90 family), with ATPase activity that interacts with tumor necrosis factor type I and may regulate cellular stress responses. hsa-miR-149-5p overlaps the TRAP1 main ORF start codon, that is followed by an active dTIS in frame with the main

ORF (705 amino acids), from which a N-term truncated peptide of 289 amino acids may be produced

miRNA binding sites overlapping not active aTISs followed by active dTIS(s)

We then considered those mRNAs with inactive aTISs followed by active dTIS (CDS). These case are of special interest since we can hypothesize that they represent mRNAs in which the aTIS was found inactive by GTI-seq experiments due to the binding of a miRNA covering the canonical aTIS. We found two genes in which the inactive aTIS is overlapping a miRNA-binding site and it is followed by one or more active dTIS: SRSF9 (serine/arginine-rich splicing factor 9) and COG4 (component of oligomeric golgi complex 4), respectively interacting with hsa-miR-935 and hsa-miR-615-3p. SRSF9 mRNA in an interesting example of a complex combination of ORFs: the same mRNA hosts, in addition to the aTIS one 5'TIS and three dTIS. According to data, the aTIS encoding a 221 residues peptide and overlapping hsa-miR-935 binding site is inactive, whereas the other TISs are active. In particular, among dTISs, two encode short (91 and 64 residues) peptides and are out of frame respectively to the main ORF, whereas the third dTIS is in frame with the aTIS, corresponds to a 88 amino acids ORF and overlaps the binding site for hsa-miR-30d-5p.

miRNA binding sites overlapping active dTISs

Many considered mRNAs present ORFs included in the main annotated ORF, starting with dTISs. Among them, 93 mRNAs present one or more dTISs overlapped by the binding site one of 66 different miRNAs. Also in this group of genes we found many ribosomal proteins and histone genes, as well a few genes involved in cellular responses to stress (Supplementary Files 3 and 4).

Table 3 E shows, as an example, that in the VMP1 mRNA, presenting two active dTISs, the second one, in frame with the main ORF (aTIS) overlaps hsa-miR-25-3p binding site. The main VMP1 ORF encodes the (multi-pass) vacuolar Membrane protein 1, a stress-induced protein that plays a role in the initial stages of the autophagic process. When overexpressed, VMP1 promotes formation of intracellular vacuoles followed by cell

death. The alternative dTIS overlapping with hsa-miR-25-3p binding site hypothetically encodes a VMP1 isoform lacking the 14 N-terminal residues.

Direct experimental evidence of functional miRNA-TIS interactions

To further test the robustness of our hypothesis, we selected 10 putative miRNA-TIS interactions for experimental investigation using luciferase reporter assay. Selected interactions included 9 miRNA footprints putatively interacting with one of three different miRNAs (hsa-miR-23a-3p, hsa-miR-615-3p and has-miR-1226-3p) and overlapping different types of TISs (3 5' TISs, 1 aTISs, and 6 dTISs) according to the above-described set of interactions.

Moreover, we also considered JUNB (Jun B proto-oncogene, alias AP1) gene, in which the hsa-miR-1226-3p binding site (positions 32-49 in the mRNA, covering 18 nt) is only one nucleotide far from a 5' TIS (positions 51-53 in the mRNA). As previously explained, the gene was not included in previous results since we used stringent criteria for the definition of putative miRNA-TIS interactions strictly focusing on cases in which the TIS is comprised in the stretch of nucleotides pairing with the miRNA. Since miR-1226-3p is 26 nt long, we reasoned that the occupancy due to miRNA binding could anyhow influence the TIS usage. JunB plays key biological roles, since it regulates gene activity following the primary growth factor response. It is also associated with many diseases, including anaplastic large cell lymphoma, and it is related to key pathways, as mucin expression in cystic fibrosis via IL-6, IL-17 signaling pathways and G-protein signaling Ras family GTPases in kinase cascades. Moreover, a recent study about human endothelial cells activation (Schmid et al. 2013) reported evidence of translational control of JunB expression, and demonstrated that the variations of protein expression following activation are not attributable to transcriptional control through TFs. Due to the importance of JunB, and since this putative miRNA-TIS interaction represented an "extreme" case, we thought that it deserved experimental investigation.

Supplementary File 5 contains detailed materials and methods regarding the validation experiments. In brief, for seven mRNAs (ASNS, EIF3B, EIF4G1, PPIA, RPS17, TAGLN2 and UQCC1) wild type miRNA footprint regions of around 100 nt were

obtained by PCR amplification of the HEK293T cells cDNA. For other three mRNAs (BAG66, JUNB, and RPL5), by annealing of synthetic oligonucleotides, both wild type and mutated miRNA footprint regions were obtained. Footprint regions were cloned in Dual-Luciferase miRNA Target Expression Vector. Each construct was co-transfected both with miRNA mimics and with scramble miRNA. All experiments were done in triplicate.

The comparison of luciferase activity observed in mimics and scramble RNA co-transfections showed a reduction of luciferase expression in 9 out of 10 tested wild type footprint sequences. In 6 cases, the difference is statistically significant with p-value < 0.05 (Figure 6 A-B). These results indicate that all tested TIS (originally belonging to annotated, 5' or downstream ORFs) are active and that in 60% of them the tested miRNA really binds to the footprint region suppressing the protein expression.

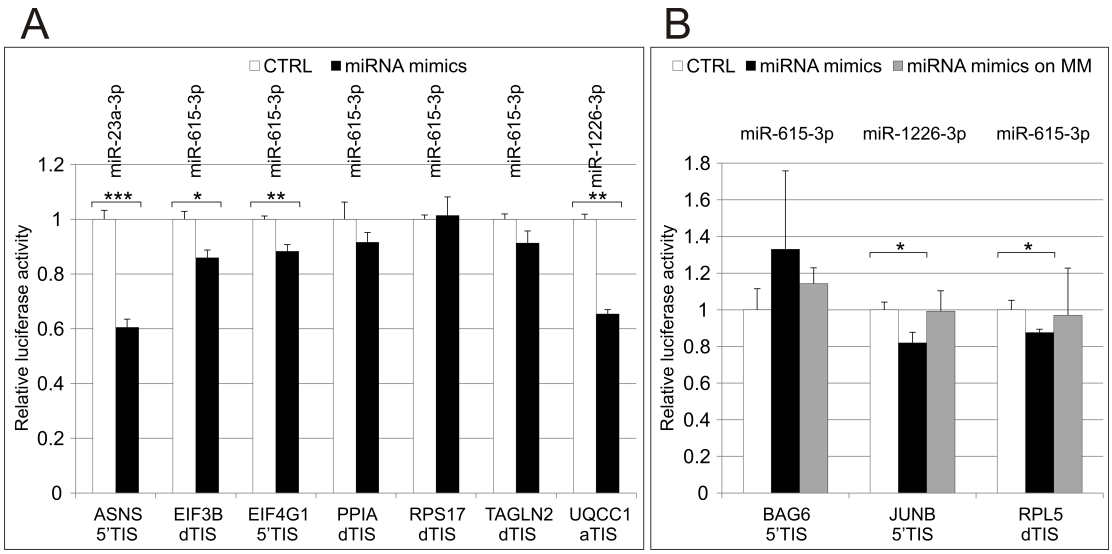

Figure 6. Validations by Luciferase reporter assay of miRNA-mRNA interactions involving miRNA binding sites overlapping TISs. Panel A reports relative luciferase activity measurements (average and standard error) for co-transfections with miRNA mimics and with scramble miRNA (CTRL) of constructs containing, for each gene, the wild type miRNA footprint region overlapping the TIS. Below the gene name the type of TIS tested is indicated, whereas tested miRNAs are indicated on the top. Asterisks indicate statistically significant reductions of reporter expression (t-test p-value <0.05 *; <0.01 **, <0.001 ***). Panel B reports relative luciferase activity measurements for co-transfections with miRNA mimics and with scramble miRNA (CTRL) of constructs containing, for each gene, the wild type and the mutated (MM) miRNA footprint region overlapping the TIS, providing a rescue experiment for tested interactions.

Moreover, rescue experiments, carried out for BAG66, JUNB, and RPL5, showed a complete restoration of expression when the miRNA mimics was co-transfected with constructs containing the mutated footprints. For the two interactions (RPL5/miR-615-3p and JUNB/miR-1226-3p) in which the mimics-induced silencing was statistically significant this result is particularly relevant. Experimental evidence demonstrates that

tested footprint regions are directly responsible for the miRNA-mRNA interaction probably interfering with protein synthesis.
In summary, the validation experiments provided direct functional evidence of miRNA-TIS interactions in 60% of tested cases, giving further support to the hypothesis that miRNA binding in regions overlapping with TISs can interfere with protein expression.

Outlook

The hypothesis that miRNAs could regulate alternative translation of many mRNAs is indirectly supported both by GTI-seq data providing information about the frequency of multiple and non-canonical active ORFs in human mRNAs and by the commonness of miRNA-mRNA binding outside 3'UTRs. The integration of the two types of evidence allowed us to identify many genes in which one or more miRNAs could interfere with translation of main annotated ORF or with ORFs located in the 5' UTR respectively to the annotated ORF, or even downstream it. We demonstrated that miRNA-binding regions overlapping TISs are evolutionary conserved and that the miRNA footprints tend to overlap mRNA regions involved in RNA fold stabilization. We displayed how the binding of a miRNA to one of the TISs can produce different regulatory effects, according to the involved ORF types co(existing) in mRNAs, and to their regulatory relations. For each ORF type, we selected a few genes for which we provided experimental evidence of TIS activity. We obtained direct evidence of miRNA-TIS interaction causing suppression of protein expression in 60% of tested cases. Both in the experimentally investigated set of genes and in the largest group of putative interactions collected in this study, many interesting genes and miRNAs are represented that surely deserve further investigation, to better characterize the mechanisms of AT regulation by miRNAs. These studies will tell us if and how the miRNA-based regulation of mRNA alternative translation impact on cell processes and on disease.

References

Abreu, M.M. and Sealy, L. 2010. The C/EBPbeta isoform, liver-inhibitory protein (LIP), induces autophagy in breast cancer cell lines. *Exp. Cell Res.* **316**: 3227-3238.

Albergaria, A., Resende, C., Nobre, A.R., Ribeiro, A.S., Sousa, B., Machado, J.C., Seruca, R., Paredes, J. and Schmitt, F. 2013. CCAAT/enhancer binding protein β (C/EBPβ) isoforms as transcriptional regulators of the pro-invasive CDH3/P-cadherin gene in human breast cancer cells. *PloS One* **8**.

Bartel, D.P. 2009. MicroRNAs: Target Recognition and Regulatory Functions. *Cell* **136**: 215-233.

Bazykin, G.A. and Kochetov, A.V. 2011. Alternative translation start sites are conserved in eukaryotic genomes. *Nucleic Acids Res.* **39**: 567-577.

Candeias, M.M., Powell, D.J., Roubalova, E., Apcher, S., Bourougaa, K., Vojtesek, B., Bruzzoni-Giovanelli, H. and Fåhraeus, R. 2006. Expression of p53 and p53/47 are controlled by alternative mechanisms of messenger RNA translation initiation. *Oncogene* **25**: 6936-6947.

Calvo SE, Pagliarini DJ, Mootha VK. Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. ProcNatl Acad Sci U S A. 2009;106(18):7507-12.

Cao, D., Li, J., Guo, C.C., Allan, R.W. and Humphrey, P.A. 2009. SALL4 is a novel diagnostic marker for testicular germ cell tumors. *Am. J. Surg. Pathol.* **33**: 1065-1077.

Chiba, M. 2012. Exosomes secreted from human colorectal cancer cell lines contain mRNAs, microRNAs and natural antisense RNAs, that can transfer into the human hepatoma HepG2 and lung cancer A549 cell lines. *Oncol. Rep.*

Fukushima, M., Tomita, T., Janoshazi, A. and Putney, J.W. 2012. Alternative translation initiation gives rise to two isoforms of Orai1 with distinct plasma membrane mobilities. *J. Cell. Sci.* **125**: 4354-4361.

Gao, Y., Wei, J., Han, J., Wang, X., Su, G., Zhao, Y., Chen, B., Xiao, Z., Cao, J. and Dai, J. 2012. The Novel Function of OCT4B Isoform-265 in Genotoxic Stress. *Stem Cells* **30**: 665-672.

Grzybowska, E.A. 2012. Human intronless genes: Functional groups, associated diseases, evolution, and mRNA processing in absence of splicing. *Biochem. Biophys. Res. Commun.* **424**: 1-6.

Helwak, A., Kudla, G., Dudnakova, T. and Tollervey, D. 2013. Mapping the Human miRNA Interactome by CLASH Reveals Frequent Noncanonical Binding. *Cell* **153**: 654-665.

Huang, D.W., Sherman, B.T. and Lempicki, R.A. 2008. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* **4**: 44-57.

Jackson, R.J., Hellen, C.U.T. and Pestova, T.V. 2012. Termination and post-termination events in eukaryotic translation. *Advances in Protein Chemistry and Structural Biology* **86**: 45-93.

Kochetov, A.V. 2008. Alternative translation start sites and hidden coding potential of eukaryotic mRNAs. *Bioessays* **30**: 683-691.

Kochetov, A.V., Ahmad, S., Ivanisenko, V., Volkova, O.A., Kolchanov, N.A. and Sarai, A. 2008. uORFs, reinitiation and alternative translation start sites in human mRNAs. *FEBS Lett.* **582**: 1293-1297.

Kosaka, N., Yusuke, Y., Hagiwara, K., Tominaga, N., Katsuda, T. and Ochiya, T. 2013. Trash or Treasure: extracellular microRNAs and cell-to-cell communication. *Non-Coding RNA* **4**.

Lee, I., Ajay, S.S., Yook, J.I., Kim, H.S., Hong, S.H., Kim, N.H., Dhanasekaran, S.M., Chinnaiyan, A.M. and Athey, B.D. 2009. New class of microRNA targets containing simultaneous 5′-UTR and 3′-UTR interaction sites. *Genome Res.* **19**: 1175-1183.

Lee S, Liu B, Lee S, Huang SX, Shen B, Qian SB. Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. Proc Natl Acad Sci U S A. 2012;109(37):E2424-32.

Liang H, He S, Yang J, Jia X, Wang P, Chen X, Zhang Z, Zou X, McNutt MA, Shen WH, Yin Y. PTENα, a PTEN isoform translated through alternative initiation, regulates mitochondrial function and energy metabolism. Cell Metab. 2014 19(5):836-48.

Liao, J., Ma, L., Guo, Y., Zhang, Y., Zhou, H., Shao, P., Chen, Y. and Qu, L. 2010. Deep Sequencing of Human Nuclear and Cytoplasmic Small RNAs Reveals an Unexpectedly Complex Subcellular Distribution of miRNAs and tRNA 3′ Trailers. *PLoS ONE* **5**.

Liu, C., Mallick, B., Long, D., Rennie, W.A., Wolenc, A., Carmack, C.S. and Ding, Y. 2013. CLIP-based prediction of mammalian microRNA binding sites. *Nucleic Acids Res.* **41**: e138-e138.

Lytle, J.R., Yario, T.A. and Steitz, J.A. 2007. Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5′ UTR as in the 3′ UTR. *Proceedings of the National Academy of Sciences* **104**: 9667-9672.

Menschaert, G., Van Criekinge, W., Notelaers, T., Koch, A., Crappé, J., Gevaert, K. and Van Damme, P. 2013. Deep proteome coverage based on ribosome profiling aids mass spectrometry-based protein and peptide discovery and provides evidence of alternative translation products and near-cognate translation initiation events. *Molecular & Cellular Proteomics: MCP* **12**: 1780-1790.

Miloslavski R, Cohen E, Avraham A, Iluz Y, Hayouka Z, Kasir J, Mudhasani R,Jones SN, Cybulski N, Rüegg MA, Larsson O, Gandin V, Rajakumar A, Topisirovic I, Meyuhas O. Oxygen sufficiency controls TOP mRNA translation via the TSC-Rheb-mTOR pathway in a 4E-BP-independent manner. J Mol Cell Biol. 2014; 6(3):255-66.

Mittal, N. and Zavolan, M. 2014. Seq and CLIP through the miRNA world. *Genome Biol.* **15**.

Morris, D.R. and Geballe, A.P. 2000. Upstream Open Reading Frames as Regulators of mRNA Translation. *Mol. Cell. Biol.* **20**: 8635-8642.

Ørom, U.A., Nielsen, F.C. and Lund, A.H. 2008. MicroRNA-10a Binds the 5′UTR of Ribosomal Protein mRNAs and Enhances Their Translation. *Mol. Cell* **30**: 460-471.

Park, B., Kook, S., Lee, S., Jeong, J., Brufsky, A. and Lee, B. 2013. An isoform of C/EBPβ, LIP, regulates expression of the chemokine receptor CXCR4 and modulates breast cancer cell migration. *The Journal of Biological Chemistry* **288**: 28656-28667.

Schnall-Levin, M., Rissland, O.S., Johnston, W.K., Perrimon, N., Bartel, D.P. and Berger, B. 2011. Unusually effective microRNA targeting within repeat-rich coding regions of mammalian mRNAs. *Genome Res.* **21**: 1395-1403.

Sinkkonen, L., Hugenschmidt, T., Berninger, P., Gaidatzis, D., Mohn, F., Artus-Revel, C., Zavolan, M., Svoboda, P. and Filipowicz, W. 2008. MicroRNAs control de novo DNA methylation through regulation of transcriptional repressors in mouse embryonic stem cells. *Nature Structural & Molecular Biology* **15**: 259-267.

Skabkin, M., Skabkina, O., Hellen, C.T. and Pestova, T. 2013. Reinitiation and Other Unconventional Posttermination Events during Eukaryotic Translation. *Mol. Cell* **51**: 249-264.

Slavoff, S.A., Mitchell, A.J., Schwaid, A.G., Cabili, M.N., Ma, J., Levin, J.Z., Karger, A.D., Budnik, B.A., Rinn, J.L. and Saghatelian, A. 2013. Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nature Chemical Biology* **9**: 59-64.

Smith, E., Meyerrose, T.E., Kohler, T., Namdar-Attar, M., Bab, N., Lahat, O., Noh, T., Li, J., Karaman, M.W., Hacia, J.G. et al. 2005. Leaky ribosomal scanning in mammalian genomes: significance of histone H4 alternative translation in vivo. *Nucleic Acids Res.* **33**: 1298-1308.

Sonda, N., Simonato, F., Peranzoni, E., Calì, B., Bortoluzzi, S., Bisognin, A., Wang, E., Marincola, F., Naldini, L., Gentner, B. et al. 2013. miR-142-3p Prevents Macrophage Differentiation during Cancer-Induced Myelopoiesis. *Immunity* **38**: 1236-1249.

Szamecz, B., Rutkai, E., Cuchalová, L., Munzarová, V., Herrmannová, A., Nielsen, K.H., Burela, L., Hinnebusch, A.G. and Valášek, L. 2008. eIF3a cooperates with sequences 5′ of uORF1 to promote resumption of scanning by post-termination ribosomes for reinitiation on GCN4 mRNA. *Genes Dev.* **22**: 2414-2425.

Tay, S., Blythe, J. and Lipovich, L. 2009. Global discovery of primate-specific genes in the human genome. *Proceedings of the National Academy of Sciences* **106**: 12019-12024.

Tay, Y., Zhang, J., Thomson, A.M., Lim, B. and Rigoutsos, I. 2008. MicroRNAs to Nanog, Oct4 and Sox2 coding regions modulate embryonic stem cell differentiation. *Nature* **455**: 1124-1128.

Touriol, C., Bornes, S., Bonnal, S., Audigier, S., Prats, H., Prats, A. and Vagner, S. 2003. Generation of protein isoform diversity by alternative initiation of translation at non-AUG codons. *Biol. Cell* **95**: 169-178.

Valasek, L.S. 2012. 'Ribozoomin' - Translation Initiation from the Perspective of the Ribosome-bound Eukaryotic Initiation Factors (eIFs). *Curr. Protein Peptide Sci.* **13**: 305-330.

Vanderperre, B., Lucier, J., Bissonnette, C., Motard, J., Tremblay, G., Vanderperre, S., Wisztorski, M., Salzet, M., Boisvert, F. and Roucou, X. 2013. Direct Detection of Alternative Open Reading Frames Translation Products in Human Significantly Expands the Proteome. *PLoS ONE* **8**.

Vanderperre, B., Staskevicius, A.B., Tremblay, G., McCoy, M., O'Neill, M.,A., Cashman, N.R. and Roucou, X. 2011. An overlapping reading frame in the PRNP gene encodes a novel

polypeptide distinct from the prion protein. *FASEB Journal: Official Publication of the Federation of American Societies for Experimental Biology* **25**: 2373-2386.

Wan J, Qian SB. TISdb: a database for alternative translation initiation in mammalian cells. Nucleic Acids Res. 2014;42(Database issue):D845-50.

Wang, X., Zhao, Y., Xiao, Z., Chen, B., Wei, Z., Wang, B., Zhang, J., Han, J., Gao, Y., Li, L. et al. 2009. Alternative Translation of OCT4 by an Internal Ribosome Entry Site and its Novel Function in Stress Response. *Stem Cells* **27**: 1265-1275.

Wang, Y., Jatkoe, T., Zhang, Y., Mutch, M.,G., Talantov, D., Jiang, J., McLeod, H.,L. and Atkins, D. 2004. Gene expression profiles and molecular markers to predict recurrence of Dukes' B colon cancer. *J. Clin. Oncol.* **22**: 1564-1571.

Zardo, G., Ciolfi, A., Vian, L., Starnes, L.M., Billi, M., Racanicchi, S., Maresca, C., Fazi, F., Travaglini, L., Noguera, N. et al. 2012. Polycombs and microRNA-223 regulate human granulopoiesis by transcriptional control of target gene expression. *Blood* **119**: 4034

Conclusions

In this work we have extensively characterized miRNAs and other small RNAs under diverse points of view.

We expanded the knowledge of miRNAs and moRNAs expressed by CD34+ cells and identified and validated a few elements that can contribute to PMF pathogenesis. We considered small RNA sequencing data of CD34+ cells of healthy subjects and PMF patients. In addition to 784 miRNAs annotated in miRBase, our in-house pipeline miR&moRe let us discover 34 new miRNAs expressed in our samples. Most miRNAs were expressed in their isoform variants, not as the annotated sequence. We also detected in CD34+ sequences aligning to hairpins outside known and novel miRNAs that correspond to expressed microRNA-offset RNAs, called moRNAs. Myeloproliferative disorders are clonal hematopoietic stem cell neoplasias, miRNAs and moRNAs deregulation can be implied in tumor physiopathology. We then looked for differentially expressed small RNAs in PMF CD34+ samples respect to control samples recognizing 37 sRNAs with significant differentially expression (DE). Noteworthy, among them 2 moRNAs are included and one was highly expressed in normal CD34+ cells but not detected at all in considered PMF samples. Very likely, moRNAs can function as miRNAs but biological roles and mechanisms of function still deserve investigation. We validated the differential expression of six selected DE on PMF granulocytes samples. Target predictions of these validated small RNA and functional enrichment analysis showed that miRNA targets are enriched in many interesting pathways involved in tumor development and progression, as signaling by FGFR, DAP12 signaling and Oncogene Induced Senescence. Hopefully identified and validated elements can help in the understanding the mechanisms that contribute to PMF pathogenesis and in formulating new targeted therapies.

In order to have more awareness in applying normalization methods when managing RNA-seq data of small RNA dataset, we evaluated the performance of normalization algorithms formulated for long RNAs, applied to human small RNA datasets. We simulated multiple matrixes with a controlled number of differentially expressed elements. We chose five normalization methods among the most cited and widespread, implemented in R packages. Each algorithm is based on different hypothesis on statistical shape and characteristics of data and we tested their impact on the downstream analysis in

a differential expression test. To quantify normalization algorithms performances we calculated ROC curves and AUC curves. ROC curves showed that algorithms do not perform significantly differently applied to the same simulated scenarios and we are not able to definitively prefer a normalization algorithm as the best. All the algorithms produced a high false positive rate and AUC curves showed us only small differences in the performances. We are far from reaching the consensus on the best normalization algorithm and there is still room to improve normalization methods for RNA-seq analysis.
We were later interested in studying whether different intracellular amounts of FHC might affect miRNA and gene expression profiles. As first step toward the dissection of the molecular basis of FHC-modulated gene expression, we performed an integrated analysis of miRNA and mRNA expression patterns in K562 FHC-silenced cells. By using a microRNA PCR Panel, we found that 4 out of 84 analysed miRNAs were consistently and significantly up-regulated in FHC-silenced cells.
The profile of the four up-regulated miRNAs has been integrated with the transcriptome analysis by combining data obtained from the microRNA targets prediction software with a correlation-based approach. This analysis led to the identification of 91 down-regulated targets. IPA revealed that the highest scored pathways in which these genes are involved are: “Cell Death and Survival, Hematological System Development and Function, Hematopoiesis” and “DNA Replication, Recombination and Repair, Cell Cycle, Cancer”. It is interesting to note that, among the common pathways, “Cell Death and Survival” and “Hematological System Development and Function” rely on the ERK1/2 activation that our results demonstrate to be severely affected by FHC modulation. In conclusion, we believe that the identification of FHC-dependent miRNA/mRNA networks implies that different amounts of the ferritin subunit contribute, in K562 cells, to the remodelling of gene expression taking place during these cellular processes through the action of let-7g, let-7f, let-7i and miR-125b
The hypothesis that miRNAs could regulate alternative translation of many mRNAs is indirectly supported by the integration of the two types of evidence: the multiple and non-canonical active ORFs in human mRNAs, provided by GTI-seq data, and miRNA-mRNA binding outside 3’UTRs using CLASH technique data. Looking for overlapping regions in both the experimental evidences, we identified many genes in which one or more miRNAs could interfere with translation of ORF. These regions are evolutionary conserved and the miRNA footprints tend to overlap mRNA regions involved in RNA fold stabilization. We selected most significant genes and we provided experimental

evidence of miRNA-TIS interaction causing suppression of RNA expression in 60% of tested cases. This non canonical miRNA function surely deserve further investigation, to better characterize the mechanisms of AT regulation.

miRNAs appear key modulators of information and understanding the interplay of miRNAs and DNA, coding RNA or other targeted elements is crucial. Although a general picture of miRNAs-mediated pathway is emerging, many questions remain. Further studies need to be performed to elucidate post-transcriptional and transcriptional role of miRNAs in different processes and diseases.