Aus der Klinik für Anaesthesiologie

Klinikum der Ludwig-Maximilians-Universität München

Direktor: Professor Dr. Bernhard Zwißler

Die Rolle intragenischer miRNA in der Regulation ihrer Host-Gene

Als kumulative Habilitationsschrift

Zur Erlangung des akademischen Grades eines habilitierten Doktors der Medizin an der

Ludwig-Maximilians-Universität München

Vorgelegt von

Ludwig Christian Giuseppe Hinske

(2017)

Hintergrund

Micro-RNAs (miRNAs) sind kleine, nicht-kodierende RNA-Sequenzen. Lee und Kollegen beschrieben in den 90er Jahren erstmals, dass das für die larvale Entwicklung notwendige Gen *lin-4* nicht in ein Protein translatiert wird, sondern dessen Transkript über basenkomplementäre Interaktion mit dem 3´-Ende des Transkripts des Gens *lin-14* dessen Translation epigenetisch negativ regulieren kann (R. C. Lee, Feinbaum, & Ambros, 1993). Die Entdeckung war revolutionär, da bis zu diesem Zeitpunkt davon ausgegangen wurde, dass nicht-translatierte RNA lediglich ein Abfallprodukt ohne relevante biologische Funktion sei. Erst einige Zeit später, im Jahr 2001, zeigten Lagos-Quintana und seine Kollegen, dass miRNAs in einer Vielzahl von Organismen nachweisbar sind, unter anderem in menschlichen Zellen. Außerdem beschrieben sie, dass miRNAs nicht nur organismus-, sondern auch gewebespezifisch exprimiert werden (Lagos-Quintana, Rauhut, Lendeckel, & Tuschl, 2001). Daraufhin stieg die Anzahl der Arbeiten, die sich mit der Rolle von miRNAs in der Pathogenese verschiedenster Erkrankungen beschäftigten, exponentiell an (Chan, Krichevsky, & Kosik, 2005; Hammond, 2006; Xie et al., 2005). Heutzutage ist man sich der zentralen Rolle von miRNAs als Regulatoren physiologischer Signalkaskaden bewusst (Ledderose et al., 2012; Martin et al., 2011; Tranter et al., 2011; Yan, Hao, Elton, Liu, & Ou, 2011).

Trotz intensiver Forschungsbemühungen sind allerdings viele Fragen im Bereich der miRNA-Forschung nicht ausreichend beantwortet. Bereits anhand der genomischen Lokalisation müssen drei Arten von miRNAs unterschieden werden. Erstens gibt es miRNAs, wie beispielsweise miR-21, die wie protein-kodierende Gene als solitäre transkriptionelle Einheit vorliegen und somit eigenständig reguliert werden (Long et al., 2011). Weiterhin gibt es miRNAs, die als Polycistron vorliegen. Polycistrone sind Gencluster, die als ein Primärtranskript abgelesen werden. Dies ist insbesondere interessant, da Polycistrone im menschlichen Organismus untypisch sind (Baskerville & Bartel, 2005). Und als dritte Gruppe existieren intragenische miRNAs, d.h. miRNA-Gene, die innerhalb von proteinkodierenden, sogenannten Host-Genen, gelegen sind. Diese miRNA-Gene können zwar eigene regulatorische Elemente besitzen, wie beispielsweise miR-107 oder miR-126 (Monteys et al., 2010), sind aber häufig funktionell an die Expression ihrer Host-Gene gekoppelt (Rodriguez, Griffiths-Jones, Ashurst, & Bradley, 2004). Dieses Phänomen eröffnet einige wichtige Fragen: Ist diese Kopplung biologisch relevant? Wenn ja, wie kann diese untersucht und charakterisiert werden? Welche Rolle spielt eine solche Beziehung für mögliche pathogenetische Prozesse? Einige Arbeiten zeigen bereits, dass eine aufgehobene Kopplung intragenischer miRNAs und ihrer Host-Gene ein entscheidender Schritt in der Tumorpathogenese sein könnte (Mayr & Bartel, 2009; Singh et al., 2009).

Das hier vorgestellte Habilitationsverfahren befasst sich deshalb mit der Untersuchung der funktionellen Beziehung intragenischer miRNAs zu ihren Host-Genen. Zur Bearbeitung dieser Thematik wurde ein breites Spektrum an bioinformatischen und molekularbiologischen Methoden etabliert und eingesetzt. Die Thematik wurde in drei Schritten bearbeitet:

1. **Bioinformatische Grundlagen**: Erstellung einer Datenbank und Evaluation bioinformatischer Methoden zur Untersuchung von Fragestellungen in Bezug auf intragenische miRNAs.

MiRNAs sind circa 20 Nukleotide kurze, einzelsträngige Nukleinsäuremoleküle. Sie binden an die nicht übersetzte 3´-RNA Region (3´-UTR) proteinkodierender Gene und reduzieren in Folge dessen deren Translation, entweder via translationaler Inhibition oder via mRNA Degradation (Baek et al., 2008; Lagos-Quintana et al., 2001; Lau, Lim, Weinstein, & Bartel, 2001; R. C. Lee & Ambros, 2001). Letzteres scheint dabei das dominierende Prinzip zu sein (Guo, Ingolia, Weissman, & Bartel, 2010). Der Mechanismus der Zielerkennung ist bis heute nicht vollständig verstanden und es existieren keine molekularbiologischen Methoden zur Large-Scale Detektion und Validierung von miRNA-Zielgen-Interaktionen. Dies ist nicht zuletzt dem Umstand geschuldet, dass eine miRNA mehrere hundert Zielgene haben und ein Gen von vielen miRNAs reguliert werden kann. Daher wurden diverse miRNA Zielvorhersage-Algorithmen entwickelt, die auf verschiedenen Prinzipien der Zielvorhersage beruhen. Einer der ältesten und am meisten benutzten Algorithmen ist TargetScan (Lewis, Burge, & Bartel, 2005; Lewis, Shih, Jones-Rhoades, Bartel, & Burge, 2003). TargetScan basiert auf dem Prinzip der sogenannten "Seed-Komplementarität". Als "Seed" einer miRNA bezeichnet man die Basen 2 - 8 (vom 5´-Ende der miRNA gezählt). Er ist der stärkste Prädiktor für die Zielgenerkennung einer miRNA (Brennecke, Stark, Russell, & Cohen, 2005). Andere Zielvorhersagealgorithmen nutzen die freie Bindungsenergie (Kertesz, Iovino, Unnerstall, Gaul, & Segal, 2007), Methoden der Mustererkennung

(Miranda et al., 2006) oder prädiktive Modelle basierend auf Transfektionsexperimenten (Krek et al., 2005; Wang & El Naqa, 2008), die jeweils in Sensitivität und Spezifität deutliche Unterschiede zeigen. Intragenische miRNAs können des Weiteren je nach Lokalisation in intronisch und exonisch untergliedert werden. Dabei muss allerdings die Strangspezifität, die Existenz möglicher alternativer Transkripte und das Vorhandensein möglicher miRNA-Promotoren berücksichtigt werden (Monteys et al., 2010). In einer ersten Arbeit entwickelten wir daher ein Datenbankmodell, in dem die verschiedenen Informationen aus unterschiedlichsten Quellen integriert wurden (L. C. Hinske, Heyn, Galante, Ohno-Machado, & Kreth, 2013). In diesem Modell werden die Informationen für proteinkodierende Gene des National Center for Biotechnology Information (NCBI) mit Transkript-Informationen aus der Reference Sequence Collection (RefSeq) verbunden (Pruitt, Tatusova, & Maglott, 2007). RefSeq ist eine kurierte, nicht-redundante Sammlung von Gentranskripten, die unter anderem durch den Genome Browser der University of California Santa Cruz zum Download bereitsteht. Das Mapping dieser Sequenzen auf das jeweils aktuelle menschliche Referenzgenom erlaubt insbesondere die Extraktion der 3´-UTR-Sequenz der jeweiligen Transkripte. Die jeweils aktuell registrierten miRNA-Gen-Koordinaten können von miRBase extrahiert werden (Griffiths-Jones, 2006; Griffiths-Jones, Grocock, van Dongen, Bateman, & Enright, 2006; Griffiths-Jones, Saini, van Dongen, & Enright, 2008). Dann müssen die miRNA-Gen-Koordinaten mit den RefSeq-Koordinaten verglichen werden, um miRNAs in intragenisch, und spezifischer in intronisch und exonisch klassifizieren zu können. Dabei definieren wir intragenische miRNAs als miRNAs, deren Genkoordinaten vollständig zwischen der Transkriptionsstartseite und dem Transkriptionsende liegen.

Exonisch sind miRNAs dann, wenn ein Teil des miRNA-Gens mit einem kodierenden Sequenzbereich überlappt. Zudem wurden Zielvorhersagealgorithmen integriert. Nach der entsprechenden Klassifikation der miRNAs wurden miRNA-Zielvorhersagen implementiert. Weiterhin wurden Vorhersagealgorithmen, Gen- und miRNA-Expressionsdaten und Protein-Interaktionsdaten integriert (L. C. Hinske et al., 2014; L. C. G. Hinske, Galante, Kuo, & Ohno-Machado, 2010). Abbildung 1 gibt eine grafische Übersicht über die verarbeiteten Datenquellen.

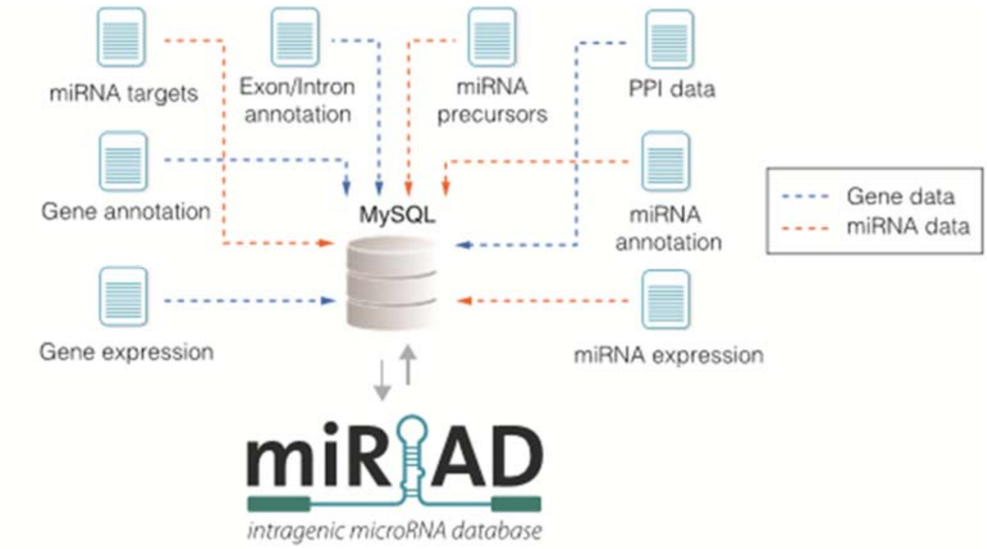


Abbildung 1. Übersicht über die verschiedenen Datenquellen, die in der von uns entwickelten Datenbank miRIAD integriert wurden (aus L.C. Hinske et al., 2014).

Zur besseren Benutzbarkeit dieser Datenbank haben wir in einer weiteren Studie eine Web-Applikation entwickelt (L. C. Hinske et al., 2014).

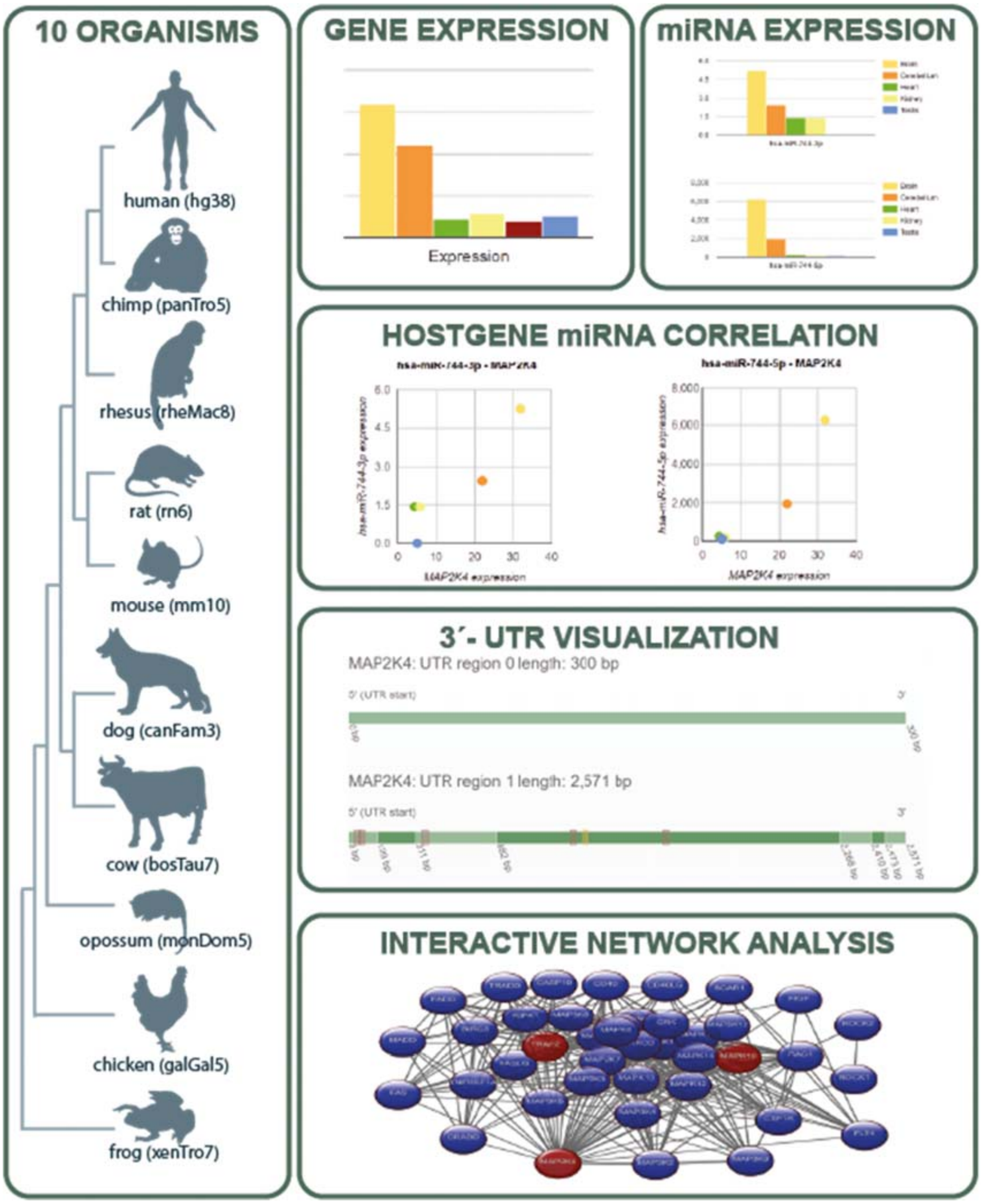


Abbildung 2. Übersicht über die Funktionalität der Web-Applikation zur effektiven Nutzung der miRIAD-Datenbank (aus L.C. Hinske et al., 2017).

Diese Software dient gezielt der Untersuchung der Rolle intronischer miRNAs und ihrer Host-Gene und erlaubte verschiedenste Analysen (Abbildung 2). Die Oberfläche ist einfach aufgebaut und besteht vornehmlich aus einem Suchfeld. Dort kann sowohl nach Genen als auch nach miRNAs gesucht werden, einzeln oder als Liste. In dem resultierenden Suchergebnis werden übersichtlich alle proteinkodierenden Gene und/oder miRNAs dargestellt. Diejenigen Gene, die eine intragenische miRNA beinhalten beziehungsweise die miRNAs, die intragenisch gelegen sind, sind gekennzeichnet. Zudem kann auch die Liste aller intragenischen miRNAs oder aller Host-Gene aufgerufen werden.

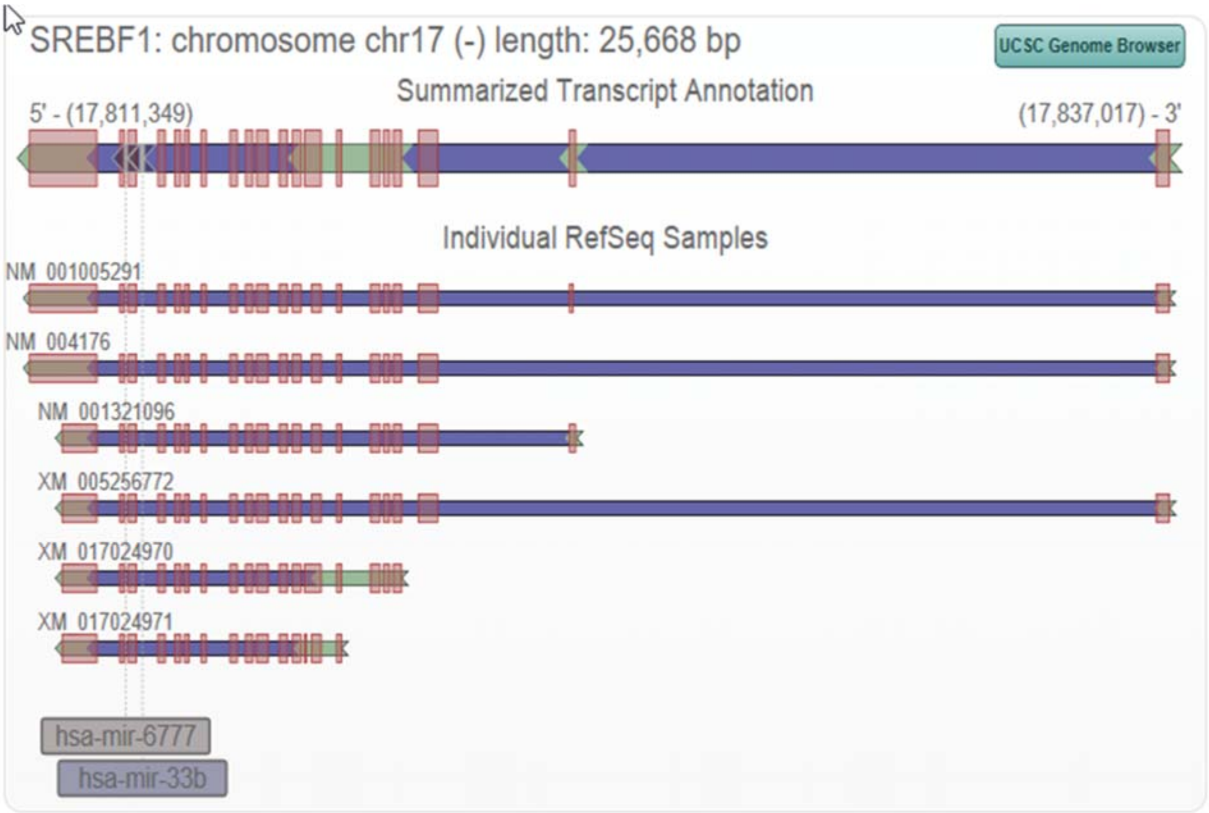


Abbildung 3. Visualisierung des Genmodells. Die Leserichtung ist durch die Pfeilrichtung gekennzeichnet, exonische Bereiche entsprechen den roten Kästchen, intronische Sequenzen den blauen und UTR-Sequenzen den grünen Bereichen (aus L.C. Hinske et al., 2017).

In der darauf folgenden Einzelansicht werden detaillierte Informationen über das Gen und die entsprechende miRNA angezeigt (Abbildung 3). Die Gen-Ansicht beginnt mit einer kurzen Beschreibung. Darunter befindet sich die Visualisierung des Genmodells, in der die Leserichtung, exonische, intronische und nicht-translatierte Bereiche gekennzeichnet sind. Auch mögliche intronische miRNAs werden mit Leserichtung dargestellt. Oben in der Grafik befindet sich die summative Darstellung, die die Informationen aus den darunter dargestellten Einzeltranskripten zusammenfasst. Diese Übersicht erlaubt die schnelle Erfassung von Informationen, wie der Distanz der intronischen miRNA zum nächsten Exon, des Vorhandenseins der miRNA in verschiedenen Isoformen des Gens sowie der Leserichtung der miRNA verglichen mit ihrem Host-Gen. MiRNAs, die laut Prädiktionsmodell ihr eigenes Host-Gen regulieren, werden blau dargestellt.

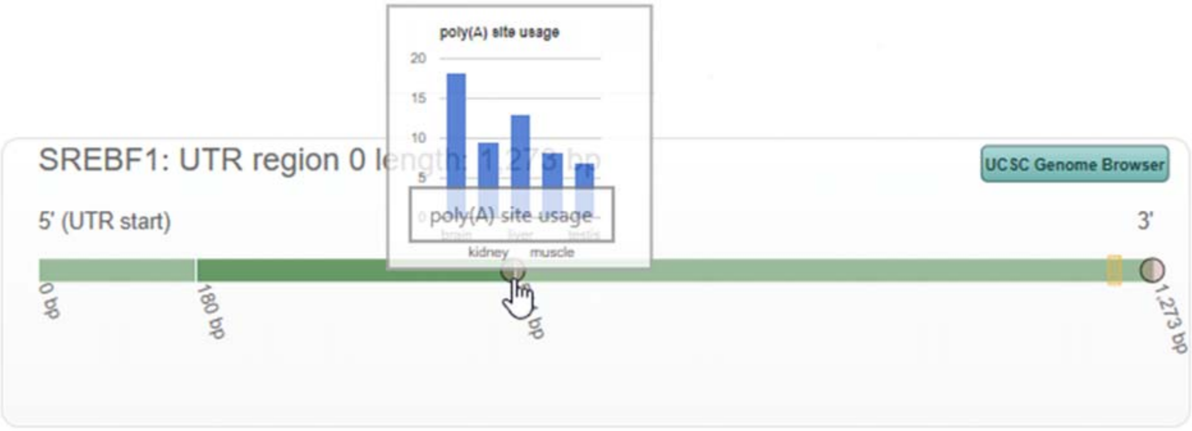
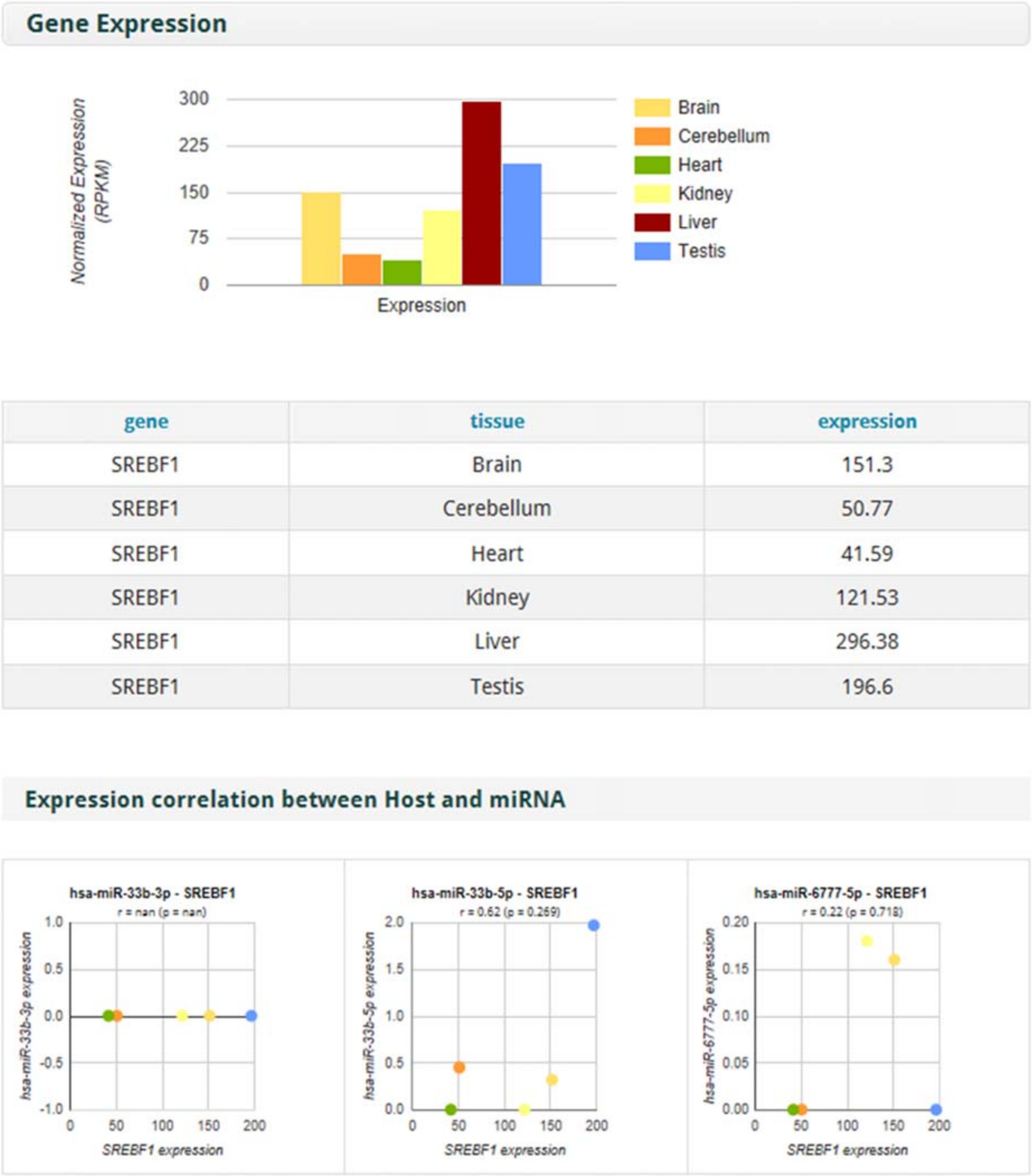


Abbildung 4. Visualisierung der 3´-UTR Region. Alternative Polyadenylierungsisoformen sind durch alternierende Grünbereiche gekennzeichnet, die mögliche miRNA-Bindungsstelle durch ein gelbes Kästchen, Informationen über die gewebeabhängige Verteilung der APA-Isoformen durch die Kreise (aus L.C. Hinske et al., 2017).

Der Genmodellgrafik folgt die gesonderte Darstellung des für die Genregulation durch miRNAs wichtigen 3´-UTRs (Abbildung 4). Diese für das Transkript besondere Region kann ebenfalls in verschiedenen Isoformen vorkommen, was man als alternative Polyadenylierung (APA) bezeichnet (Di Giammartino, Nishida, & Manley, 2011). Dieser Mechanismus ist mitunter für gewebespezifische miRNA-Regulation verantwortlich und scheint insbesondere bei der Aktivierung von Tumorgenen eine zentrale Rolle zu spielen (Mayr & Bartel, 2009; Zhang, Lee, & Tian, 2005).

Visuell werden diese alternativen Polyadenylierungsformen durch alternierende Farben gekennzeichnet. Um alternative Polyadenylierungsmuster darstellen zu können, haben wir einerseits die Daten von Derti et al. prozessiert und integriert, die für mehrere Spezies und Gewebe speziell poly(A)-Sequenzen isoliert und mithilfe von Next-Generation Sequencing sequenziert haben (Derti et al., 2012). Zum anderen haben wir einen Algorithmus entwickelt, um potentielle alternative poly(A) Varianten aus regulären RNA-Sequencing Daten zu extrahieren. Basierend auf dem Next-Generation Sequencing Datensatz von Brawand und Kollegen (Brawand et al., 2011) haben wir nach entsprechendem Mapping der Read-Sequenzen auf das jeweilige Referenzgenom nach Sequenzen gesucht, die mindestens vier aufeinander folgende Adenosin-Nukleotide ohne Korrelat im Referenzgenom aufweisen. Die entsprechende Stelle wurde nur dann als alternative Polyadenylierungsstelle in unsere Datenbank eingefügt, wenn sie von mindestens zwei nicht identischen Reads gestützt wurde. APA-Stellen, für die die Expression in verschiedenen Geweben quantifizierbar war, erscheinen in der Visualisierung als kleine Kreise. Durch Anklicken erscheint das Expressions-Balkendiagramm für diese Stelle. Zudem können vorhergesagte miRNA-

Bindungsstellen angezeigt werden, zusammen mit der dazugehörigen UTR-Sequenz und der relativen Position innerhalb des UTRs.



Gene Expression

gene	tissue	expression
SREBF1	Brain	151.3
SREBF1	Cerebellum	50.77
SREBF1	Heart	41.59
SREBF1	Kidney	121.53
SREBF1	Liver	296.38
SREBF1	Testis	196.6

Expression correlation between Host and miRNA

Abbildung 5. Darstellung der Genexpression in verschiedenen Geweben, sowie die Korrelation zwischen Host-Gen und den intronisch gelegenen miRNAs (aus L.C. Hinske et al., 2017).

Für einen Großteil der aktuell zehn in unserer Datenbank implementierten Spezies sind Expressionsdaten sowohl von miRNAs als auch von mRNA für die verschiedenen Gewebe enthalten. Insbesondere für die Beurteilung einer möglichen transkriptionellen Koregulation ist die Korrelation der einzelnen intragenischen miRNAs und ihrer Host-Gene relevant. Expressions- und Korrelationsdaten werden in unserer Applikation sowohl grafisch als auch tabellarisch dargestellt (Abbildung 5).

Da eine miRNA nicht nur ein einzelnes Gen reguliert, haben wir zudem einen Algorithmus entwickelt, um den Effekt einer miRNA auf ein Netzwerk von Genen zu quantifizieren. Dieses Netzwerk kann entweder vom Benutzer selbst definiert werden, beispielsweise als beobachtete Gen-Signatur für einen bestimmten Krankheitszustand. Da in unsere Datenbank aber auch Protein-Protein-Interaktionsdatenbanken eingebunden wurden, können miRNAs gesucht werden, die nicht nur das Zielgen selbst, sondern das mit dem Zielgen interagierende Netzwerk regulieren können. Für jede miRNA wird ein Score zu dem entsprechenden Netzwerk erstellt.

Zuerst wird die Wahrscheinlichkeit eines zufälligen Auftretens der Seed-Sequenz einer miRNA berechnet:

p(S) = $\prod_{i=1}^{n} p(N_i|D)$, wobei

S = Seed-Sequence

n = Länge(S)

N_i = i-tes Nukleotid of S

D = Nukleotid-Verteilung.

Die Wahrscheinlichkeit, dass diese Sequenz mindestens r mal innerhalb einer zufälligen Sequenz der Länge N (UTR-Sequenz eines jeden Gens innerhalb eines Interaktionsnetzwerks) auftritt, ist gegeben durch:

$p(x_t) = (1 - \sum_{i=0}^{r-1} \left(\binom{L_x}{i} * p(S)^i * (1 - p(S))^{L_x - i} \right))$, wobei

L_x = (Länge des 3´-UTRs von Element x_t) - (Länge der Seed-Sequenz n) + 1

r = erwartetes Auftreten der Sequenz (in unserer Anwendung wurde r=1 gesetzt)

Somit kann die zufällig erwartete Anzahl an Genen mit Seed-Site komplementärer Sequenz $E(X_t)$ innerhalb des Netzwerks X durch die Summe der Einzelwahrscheinlichkeiten x errechnet werden:

$$E(X_t) = \sum_{x_t \epsilon X} p(x_t)$$

Die erwartete Anzahl kann dann mit der beobachteten Anzahl verglichen und statistisch mit der Log-Odds Ratio quantifiziert werden:

$$Score(X|S) = log(\frac{\frac{E(X_t)+O(X_t)}{E(X_t)} * \frac{E(X_n)+O(X_n)}{E(X_n)}}{\frac{|X|}{E(n)}}) .$$

Zusammengefasst haben wir ein Webtool entwickelt, das eine Plattform zur Untersuchung miRNA-bezogener Fragestellungen bietet und die Visualisierung komplexer Zusammenhänge zwischen den verschiedenen Datensätzen erlaubt.

2. Die Beziehung zwischen Host-Gen und miRNA: Bioinformatische Evidenz für eine funktionelle Beziehung zwischen intronischen miRNAs und ihren Host-Genen

Intronische miRNAs werden zusammen mit ihrem Host-Gen als Primärtranskript exprimiert und vor der Splicing-Reaktion aus dem Transkript extrahiert (Kim & Kim, 2007). Danach erfolgt in mehreren Prozessierungsschritten die Reifung zur aktiven miRNA (Denli, Tops, Plasterk, Ketting, & Hannon, 2004; Han et al., 2004; Y. Lee et al., 2003). So reizvoll die Annahme einer funktionellen Beziehung zwischen Host-Gen und miRNA auch sein mag, könnte die Kolokalisation ebenso ein stochastisches Phänomen oder lediglich eine Informationskompression auf der DNA sein. Daher haben wir untersucht, ob es Evidenz für eine funktionelle Beziehung zwischen miRNAs und ihren

Host-Genen gibt (L. C. G. Hinske et al., 2010). Dazu haben wir die oben beschriebene Datenbank benutzt, um zuerst im Rahmen einer strukturellen Analyse genomische Charakteristika intronischer miRNAs zu extrahieren. Es zeigte sich, dass Host-Gene mit intronischen miRNAs insgesamt ca. dreimal länger als protein-kodierende Gene ohne miRNAs sind. Aber nicht nur die Gene an sich, sondern auch die entsprechenden 3´-UTR Sequenzen sind länger und enthalten mehr AU-reiche Regionen. AU-reiche Regionen wiederum sind maßgeblich für die Transkriptstabilität verantwortlich und eine Häufung in der UTR-Sequenz mit schnellerem Abbau und engmaschiger epigenetischer Kontrolle assoziiert (Jing et al., 2005). Danach untersuchten wir die Position intronischer miRNAs innerhalb der Gene. Unsere Hypothese war, dass im Falle eines funktionellen Zusammenhangs ein Positionsbias intronischer miRNAs zugunsten des 5´-Endes der Gene existieren müsste, damit möglichst viele alternative Transkripte mit der miRNA koexprimiert würden. Tatsächlich befinden sich ca. 60% aller intronischen miRNAs in den ersten fünf Introns und zeigen einen starken Leserichtungsbias: Die Leserichtung der intronischen miRNA und des Host-Gens ist deutlich häufiger dieselbe, als es durch Zufall erklärbar wäre. Im nächsten Schritt führten wir eine funktionelle Analyse durch, um die Hypothese eines funktionellen Zusammenhangs näher zu beleuchten. Zuerst zeigten wir, dass die Anzahl von Host-Genen, die laut Vorhersage von ihrer intragenischen miRNA reguliert werden, signifikant höher ist, als durch Zufall erklärbar (negative Regulation erster Ordnung). Dieses Ergebnis war konstant für alle in der Analyse benutzten Zielvorhersage-Algorithmen. Zur Beurteilung der Wahrscheinlichkeit einer Regulation auf höherer Ebene nutzten wir die Genkarten der Kyoto Encyclopedia of Genes and Genomes (KEGG), in der die Interaktionen zwischen Genen bezogen auf bekannte Signalkaskaden kodiert sind (Kanehisa & Goto, 2000). Dann ermittelten wir für jedes Host-Gen, die intragenische miRNA und den dazugehörigen Signalweg die Anzahl der in der entsprechenden Signalkaskade vohergesagten Ziel-Gene. Um die so ermittelte Anzahl statistisch bewerten zu können, führten wir pro Host-Gen 1000 Simulationen durch, in dem die Gene innerhalb der Kaskade durch zufällig ausgewählte Gene ersetzt wurden und erneut die Anzahl möglicher Ziel-Gene ermittelt wurde. Wir fanden eine hoch-signifikante Anreicherung von Zielgenen innerhalb der Host-Gen Signalkaskade. Basierend auf diesen Ergebnissen haben wir die Hypothese aufgestellt, dass ein möglicher Aspekt der Beziehung zwischen intronischen miRNAs und ihren Host-Genen die Verhinderung überschießender Expression im Sinne einer negativen Rückkopplung ist. Dabei haben wir ein mögliches Feedback erster Ordnung, das heißt die miRNA reguliert direkt ihr eigenes Host-Gen, von einem Feedback höherer Ordnung, bei dem die miRNA die Transkription (beispielsweise via Regulation eines Transkriptionsfaktors) oder den funktionellen Zustand (beispielsweise via Regulation einer Kinase) ihres Host-Gens beeinflusst, unterschieden (Abbildung 6).

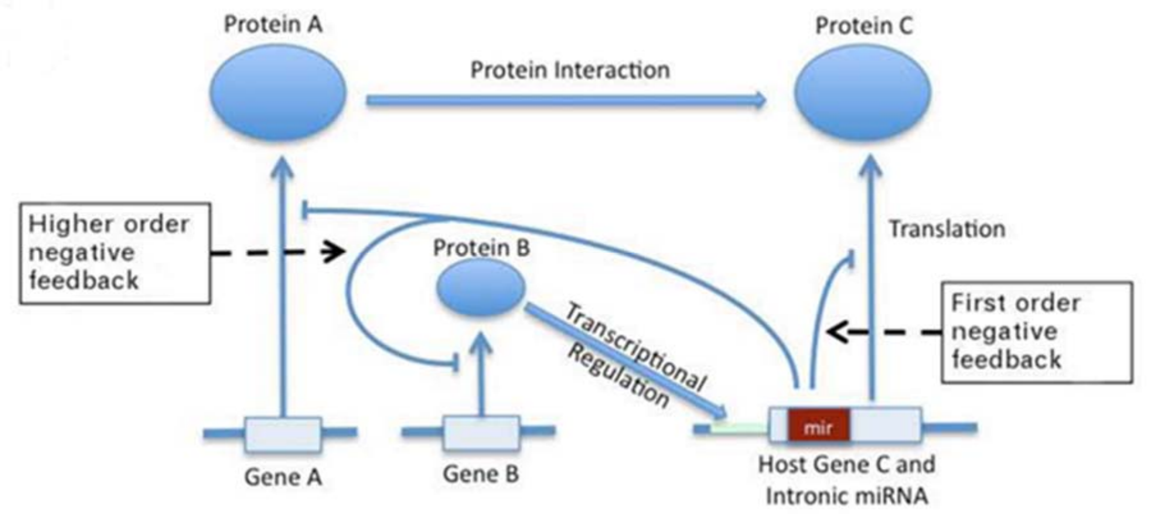

Protein A
Protein Interaction
Protein C
Higher order negative feedback
Protein B
Translation
First order negative feedback
Transcriptional Regulation
mir
Gene A
Gene B
Host Gene C and Intronic miRNA

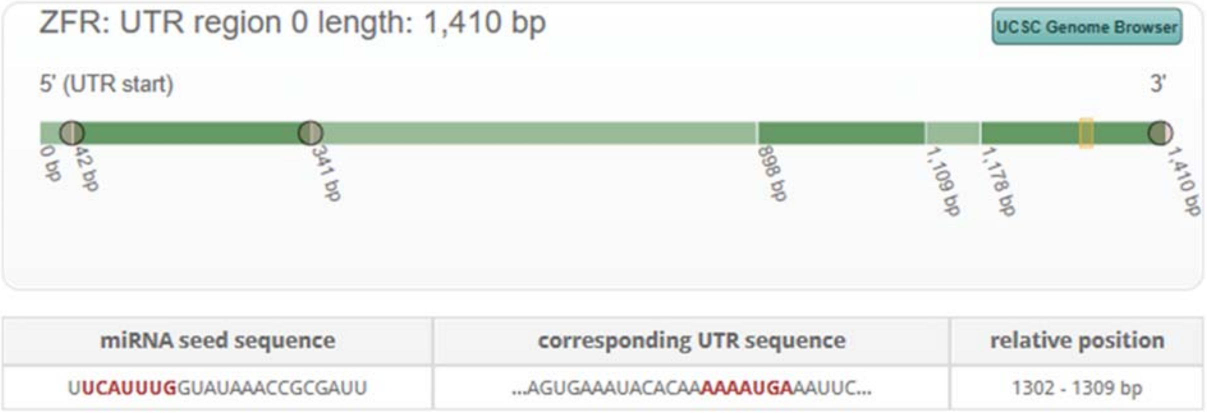
Abbildung 6. Modell der negativen Rückkopplung intronischer miRNAs auf die Signalkaskade der Host-Gene (aus L.C.G. Hinske et al., 2010).

3. Direkte und funktionelle negative Rückkopplung: Alternative Polyadenylierung als Mechanismus zur Rückkopplungsadjustierung

Eine Variante einer negativen Rückkopplung einer miRNA auf ihr Host-Gen ist die direkte Regulation (L. C. G. Hinske et al., 2010). Diese Form der Rückkopplung ist allerdings nur dann sinnvoll, wenn es Mechanismen gibt, mit Hilfe derer diese Beziehung an- und abgeschaltet werden kann. Dill und Kollegen haben erstmals anhand von miR-26b ein Beispiel einer intronischen miRNA zeigen können, die erst im Verlauf des zellulären Differenzierungsprozesses überhaupt in die biologisch aktive Form übersetzt wird und somit ihr Host-Gen reguliert (Dill, Linder, Fehr, & Fischer, 2012). Während die differenzielle Prozessierung intronischer miRNAs einen möglichen Weg der Regulation einer direkten Rückkopplung darstellt, wäre ein dynamischerer Prozess die Modifikation der Länge des 3´-UTR via alternativer Polyadenylierung.

Erreicht wird dies durch das Vorhandensein alternativer Polyadenylierungssignale. Diese wiederum werden unterteilt in “starke” und “schwache” Signale, die von einem Komplex verschiedener Proteine erkannt werden (Beaudoing, Freier, Wyatt, Claverie, & Gautheret, 2000; Di Giammartino et al., 2011). Wir haben daher die 3´-UTR-Sequenzen von Host-Genen mit Seed-komplementärer Sequenz für die eigene intronische miRNA mit denen ohne Seed-komplementäre Sequenz verglichen (L. C. Hinske et al., 2015). Dabei zeigt sich, dass erstere längere 3´-UTR Sequenzen mit mehr alternativen Polyadenylierungssignalen besitzen. Während sich bei Genen ohne intronische miRNA und Host-Genen ohne Seed-komplementäre Sequenz die “starken” Polyadenylierungssignale AAUAAA und AUUAAA zumeist am 3´-Ende des 3´-UTRs befanden, lagen diese bei Host-Genen mit Seed-komplementärer Sequenz präferenziell vor dieser Sequenz. Zudem zeigte sich, dass miRNAs, die potentiell ihr eigenes Host-Gen regulieren, auch vermehrt Transkripte von Genen regulieren, die für den Polyadenylierungsapparat wichtige Proteine kodieren, insbesondere *CPSF2* (Cleavage and Polyadenylation Specific Factor 2). *CPSF2* wiederum ist bereits mit der Veränderung der Signalerkennung in Verbindung gebracht worden (Herr, Molnàr, Jones, & Baulcombe, 2006; Kolev, Yario, Benson, & Steitz, 2008). Wir haben daher U87-Zellen nach Transfektion mit siCPSF2 sequenziert. Bei 97%-iger Reduktion von CPSF2-RNA zeigten sich, wie vorbeschrieben, grundsätzlich verlängerte UTR-Sequenzen, außer bei Host-Genen mit Seed-komplementärer Sequenz. Diese wurden tendenziell kürzer. Mittels Motiv-Detektionsanalyse konnten wir zudem zeigen, dass in den 3´-UTR Sequenzblöcken, die eine besonders starke Expressionszunahme verzeichneten, die “starken” Polyadenylierungssignale angereichert waren.

Zusammengefasst bedeutet dies, dass intronische miRNAs, die ihren eigenen Host regulieren, mitunter die eigene Wirkung auf das Host-Gen begrenzen.



miRNA seed sequence	corresponding UTR sequence	relative position
U**UCAUUUG**GUAUAAACCGCGAUU	...AGUGAAAUACACAA**AAAAUGA**AAUUC...	1302 - 1309 bp

Abbildung 7. Die 3´-UTR von ZFR. Die Bindungsstelle für miR-579 ist als gelbes Kästchen dargestellt, die Bindungssequenz im unteren Bildbereich (aus L.C. Hinske et al., 2017).

Wir haben diese Hypothese an einem Beispiel evaluiert. Das Gen *ZFR* (Zinc Finger Recombinase) besitzt eine intronische miRNA, hsa-miR-579, sowie in der 3´-UTR Region eine Seed-komplementäre Sequenz für diese miRNA (Abbildung 7). Mittels Luciferase-Assay und im Western-Blot konnten wir zeigen, dass hsa-miR-579 sowohl ZFR, als auch CPSF2 direkt reguliert. Des Weiteren entdeckten wir mehrere potentielle Polyadenylierungsstellen in der 3´-UTR, die wir mittels 3´-RACE (Rapid Amplification of cDNA Ends) validierten. Wir konnten zeigen, dass von der Regulation lediglich die längste Transkriptvariante betroffen ist. Zusammenfassend konnten wir zeigen, dass eine direkte negative Rückkopplung durch intronische miRNAs nicht nur existiert, sondern durch alternative Polyadenylierung reguliert werden kann (Abbildung 8).

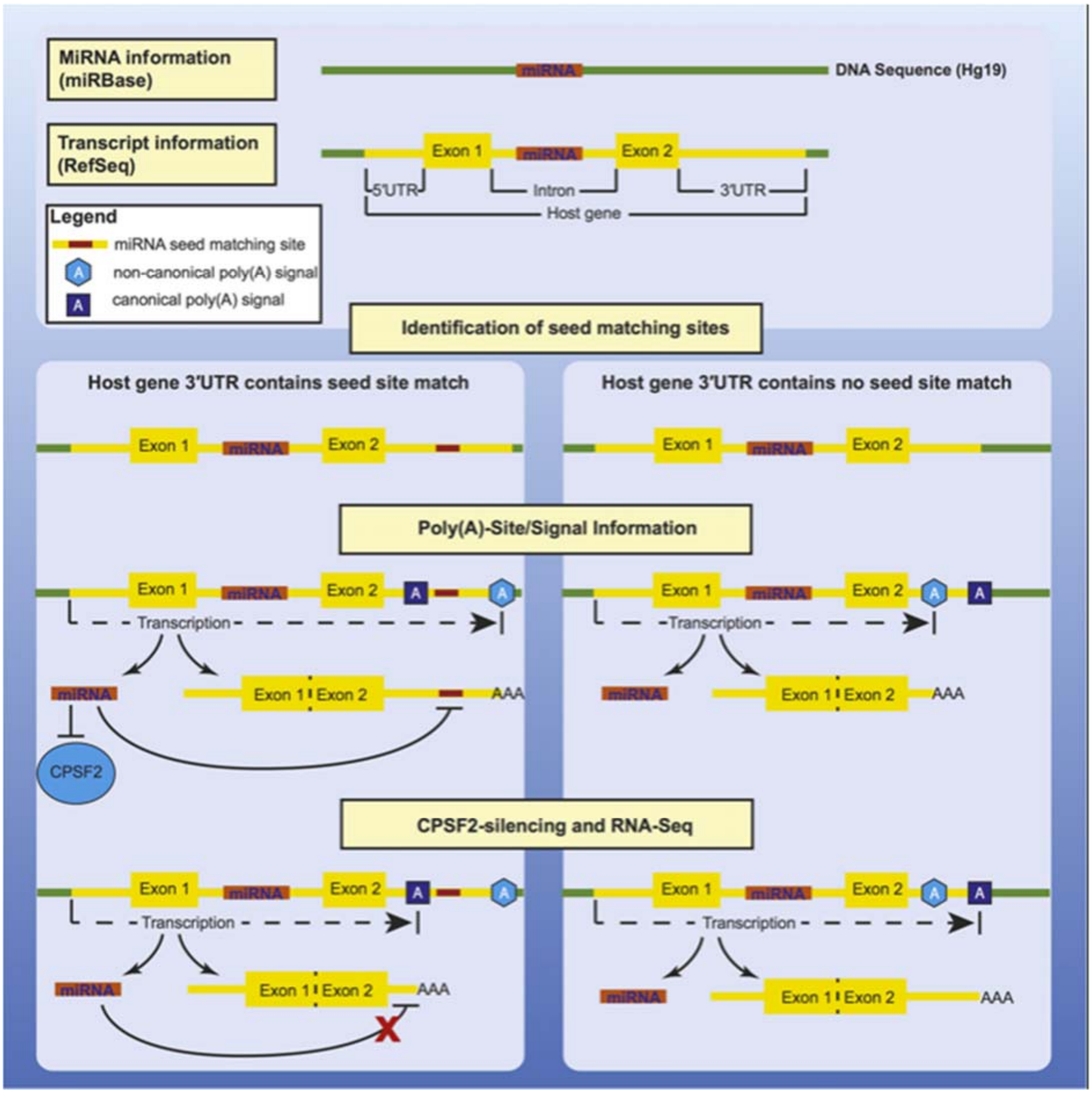

MiRNA information (miRBase)
miRNA
DNA Sequence (Hg19)
Transcript information (RefSeq)
Exon 1
miRNA
Exon 2
5'UTR
Intron
3'UTR
Host gene
Legend
miRNA seed matching site
non-canonical poly(A) signal
canonical poly(A) signal
Identification of seed matching sites
Host gene 3'UTR contains seed site match
Host gene 3'UTR contains no seed site match
Poly(A)-Site/Signal Information
Transcription
Exon 1 Exon 2
AAA
CPSF2
CPSF2-silencing and RNA-Seq

Abbildung 8. Modell der differenziellen Regulierung der direkten Regulation von Host-Genen durch ihre intronischen miRNAs (aus L.C. Hinske et al., 2015).

Ein Großteil der Host-Gene besitzt keine seed-komplementäre Sequenz in der 3´-UTR, was eine mögliche indirekte Regulation nahe legen könnte. Diese Beziehung intragenischer miRNAs zu ihren Host-Genen scheint mitunter sogar von größerer klinischer Relevanz zu sein. Bereits im Jahr 2010 publizierten Tie und Kollegen, dass die intronische miRNA hsa-miR-218, die im menschlichen Genom sowohl in *SLIT2* als auch in *SLIT3* intronisch vorkommt, mit ihrem Host-Gen *SLIT3* koexprimiert wird (Tie et al., 2010). Die Bindung von SLIT an den ROBO1-Rezeptor ist wiederum wichtig für die Invasion und Metastasierung kolorektaler Karzinome. Die Autoren zeigten, dass die ROBO1-expression von miR-218 reprimiert wird und dass in einer hochinvasiven Tumorzelllinie diese Regulation durch Silencing von SLIT3 und miR-218 entfällt. Die Autoren konnten zudem zeigen, dass eine erniedrigte miR-218-Expression mit fortgeschrittenem Tumorstadium, lymphatischer Metastasierung, sowie schlechter Prognose korreliert (Tie et al., 2010). Kürzlich erst konnten Schmitt et al. zeigen, dass die miR-4728, die intronisch in dem insbesondere für das Mamma-Karzinom wichtigen Rezeptor ERBB2/HER2 des MAPK-Signalwegs gelegen ist, ein negativer Regulator dieser Signalkaskade ist (Schmitt et al., 2015). Die Autoren fanden auch hier einen ausgeprägten Zusammenhang zwischen der Expression der miRNA und der Überlebensraten der Patienten.

Ein weiteres, für viele Tumoren sehr zentrales Gen, ist AKT2 (AKT Serine/Threonine Kinase 2) (Agarwal, Brattain, & Chowdhury, 2013; Chautard, Ouédraogo, Biau, & Verrelle, 2014; Emdad, Hu, Das, Sarkar, & Fisher, 2015). AKT wird durch Phosphorylierung aktiviert, vermittelt Zellwachstum und -überleben und inhibiert Apoptose (Chautard et al., 2014; Cui et al., 2015; Emdad et al., 2015; Hu et al., 2014). Interessanterweise ist die intronisch gelegene miRNA hsa-miR-641 kaum untersucht. Wir haben die Hypothese aufgestellt, dass miR-641 ein negativer Regulator des PI3K/AKT-Signalweges ist und dass diese Beziehung im Rahmen der Glioblastompathogenese gestört sein könnte.

In einem ersten Schritt untersuchten wir daher die Expression von miR-641 in Geweproben von Glioblastom-Patienten und verglichen diese mit Normalhirn-Gewebe. Tatsächlich zeigte sich eine deutlich erniedrigte miR-641-Expression. Danach überprüften wir, ob miR-641 möglicherweise AKT2 direkt reguliert. Allerdings enthält die AKT2 3´-UTR keine Seed-komplementäre Sequenz für miR-641 und die AKT2-mRNA änderte sich nicht signifikant nach Transfektion dieser miRNA. Trotz unveränderter AKT2-Expression beobachteten wir aber eine deutliche Zunahme der Apoptoserate der transfizierten Zelllinie, vereinbar mit einer indirekten Regulation des PI3K/AKT-Signalwegs. Daher untersuchten wir den AKT2-Aktivierungszustand in diesen Zellen und fanden eine signifikant reduzierte AKT2-Phosphoryllierung. Basierend auf unserer Interaktionsdatenbank identifizierten wir drei Kinasen, die AKT2 aktivieren (Frias et al., 2006; Jacinto et al., 2006; Laplante & Sabatini, 2012; Scheid, Marignani, & Woodgett, 2002) und deren Expressionslevel durch miR-641-Transfektion deutlich reduziert wurden: PIK3R3, PDK2 und MAPKAP1. Für zwei der drei Kinasen (PIK3R3 und MAPKAP1) wiesen wir eine direkte Regulation durch miR-641 nach, die allerdings die ausgeprägten Expressionsänderungen insbesondere von PIK3R3 nicht erklären konnte. Daraufhin suchten wir nach Transkriptionsfaktoren, die möglicherweise durch miR-641 reguliert werden und sowohl mit PIK3R3 als auch PDK2 interagieren. NFAT5 wurde als wahrscheinlicher Kandidat identifiziert und in der Folge von uns validiert. Zusammenfassend konnten wir am Beispiel von AKT2 und miR-641 zeigen, dass intronische miRNAs zentral für die Regulation ihrer Host-Gen Signalwege sein können und dass ein Wegfall der Regulation mit direkten Implikationen für die Tumorpathogenese verbunden sein kann.

Das Verständnis der Beziehung intronischer miRNAs zu deren Host-Genen ist aber nicht nur von pathogenetischer Bedeutung, sondern könnte auch klinische Anwendung finden. Eine kürzlich veröffentlichte Arbeit befasst sich beispielsweise mit der intronisch gelegenen miR-4722, die spezifisch für den menschlichen Organismus ist. Das Gen dieser miRNA liegt in Intron 5 von IL-18RAP. Sowohl intronische miRNA als auch Host-Gen sind koreguliert und in der Sepsis verstärkt exprimiert. Die Autoren konnten zeigen, dass die Bestimmung der Expression von miR-4722 eine Unterscheidung von Patienten mit Systemic Inflammatory Response Syndrome von Patienten mit Sepsis erlaubt (Ma et al., 2013).

In dieser Habilitationsarbeit wurden die nötigen Methoden entwickelt, um die Beziehung intragenischer miRNAs zu ihren Host-Genen und die Bedeutung für die Klinik zu bearbeiten. Diese Methoden wurden erfolgreich eingesetzt, um die zentrale Hypothese einer negativen Rückkopplung der miRNAs innerhalb der Signalwege ihrer Host-Gene zu etablieren sowie Beispiele sowohl für direkte als auch indirekte Rückkopplung zu validieren und regulative Mechanismen zu identifizieren. Abschließend konnte die klinische Relevanz dieser Mechanismen am Beispiel der Glioblastompathogenese gezeigt werden. Es ist davon auszugehen, dass die Beziehung intronischer miRNAs und ihrer Host-Gene eine zunehmend zentrale Rolle in vielen klinischen Bereichen einnehmen wird und möglicherweise interessante und neue diagnostische wie therapeutische Optionen bietet.

Literaturverzeichnis

Agarwal, E., Brattain, M. G., & Chowdhury, S. (2013). Cell survival and metastasis regulation by Akt signaling in colorectal cancer. *Cellular Signalling*, *25*(8), 1711–1719.

Baek, D., Villén, J., Shin, C., Camargo, F. D., Gygi, S. P., & Bartel, D. P. (2008). The impact of microRNAs on protein output. *Nature*, *455*(7209), 64–71.

Baskerville, S., & Bartel, D. P. (2005). Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA* , *11*(3), 241–247.

Beaudoing, E., Freier, S., Wyatt, J. R., Claverie, J. M., & Gautheret, D. (2000). Patterns of variant polyadenylation signal usage in human genes. *Genome Research*, *10*(7), 1001–1010.

Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csárdi, G., Harrigan, P., … Kaessmann, H. (2011). The evolution of gene expression levels in mammalian organs. *Nature*, *478*(7369), 343–348.

Brennecke, J., Stark, A., Russell, R. B., & Cohen, S. M. (2005). Principles of MicroRNA–Target Recognition. *PLoS Biology*, *3*(3), e85.

Chan, J. A., Krichevsky, A. M., & Kosik, K. S. (2005). MicroRNA-21 is an antiapoptotic factor in human glioblastoma cells. *Cancer Research*, *65*(14), 6029–6033.

Chautard, E., Ouédraogo, Z. G., Biau, J., & Verrelle, P. (2014). Role of Akt in human malignant glioma: from oncogenesis to tumor aggressiveness. *Journal of Neuro-Oncology*, *117*(2), 205–215.

Cui, Y., Lin, J., Zuo, J., Zhang, L., Dong, Y., Hu, G., … Lu, Y. (2015). AKT2-knockdown suppressed viability with enhanced apoptosis, and attenuated chemoresistance to temozolomide of human glioblastoma cells in vitro and in vivo. *OncoTargets and Therapy*, *8*, 1681–1690.

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

Derti, A., Garrett-Engele, P., Macisaac, K. D., Stevens, R. C., Sriram, S., Chen, R., … Babak, T. (2012). A quantitative atlas of polyadenylation in five mammals. *Genome Research*, *22*(6), 1173–1183.

Di Giammartino, D. C., Nishida, K., & Manley, J. L. (2011). Mechanisms and consequences of alternative polyadenylation. *Molecular Cell*, *43*(6), 853–866.

Dill, H., Linder, B., Fehr, A., & Fischer, U. (2012). Intronic miR-26b controls neuronal differentiation by repressing its host transcript, ctdsp2. *Genes & Development*, *26*(1), 25–30.

Emdad, L., Hu, B., Das, S. K., Sarkar, D., & Fisher, P. B. (2015). AEG-1-AKT2: A novel complex controlling the aggressiveness of glioblastoma. *Molecular & Cellular Oncology*, *2*(3), e995008.

Frias, M. A., Thoreen, C. C., Jaffe, J. D., Schroder, W., Sculley, T., Carr, S. A., & Sabatini, D. M. (2006). mSin1 is necessary for Akt/PKB phosphorylation, and its isoforms define three distinct mTORC2s. *Current Biology: CB*, *16*(18), 1865–1870.

Griffiths-Jones, S. (2006). miRBase: the microRNA sequence database. *Methods in Molecular Biology* , *342*, 129–138.

Griffiths-Jones, S., Grocock, R. J., van Dongen, S., Bateman, A., & Enright, A. J.

(2006). miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Research*, *34*(Database issue), D140–4.

Griffiths-Jones, S., Saini, H. K., van Dongen, S., & Enright, A. J. (2008). miRBase: tools for microRNA genomics. *Nucleic Acids Research*, *36*(Database issue), D154–8.

Guo, H., Ingolia, N. T., Weissman, J. S., & Bartel, D. P. (2010). Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature*, *466*(7308), 835–840.

Hammond, S. M. (2006). MicroRNAs as oncogenes. *Current Opinion in Genetics & Development*, *16*(1), 4–9.

Han, J., Lee, Y., Yeom, K.-H., Kim, Y.-K., Jin, H., & Kim, V. N. (2004). The Drosha-DGCR8 complex in primary microRNA processing. *Genes & Development*, *18*(24), 3016–3027.

Herr, A. J., Molnàr, A., Jones, A., & Baulcombe, D. C. (2006). Defective RNA processing enhances RNA silencing and influences flowering of Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(41), 14994–15001.

Hinske, L. C., dos Santos, F. R. C., Ohara, D. T., Ohno-Machado, L., Kreth, S., & Galante, P. A. F. (2017). miRIAD update: using alternative polyadenylation, protein interaction network analysis and additional species to enhance exploration of the role of intragenic miRNAs and their host genes. *Database: The Journal of Biological Databases and Curation*, *2017*,1-8. https://doi.org/10.1093/database/bax053

Hinske, L. C., França, G. S., Torres, H. A. M., Ohara, D. T., Lopes-Ramos, C. M., Heyn, J., … Galante, P. A. F. (2014). miRIAD-integrating microRNA inter- and intragenic data. *Database: The Journal of Biological Databases and Curation*, *2014*.

https://doi.org/10.1093/database/bau099

Hinske, L. C., Galante, P. A. F., Limbeck, E., Möhnle, P., Parmigiani, R. B., Ohno-Machado, L., … Kreth, S. (2015). Alternative Polyadenylation Allows Differential Negative Feedback of Human miRNA miR-579 on Its Host Gene ZFR. *PloS One*, *10*(3), e0121507.

Hinske, L. C. G., Galante, P. A. F., Kuo, W. P., & Ohno-Machado, L. (2010). A potential role for intragenic miRNAs on their hosts' interactome. *BMC Genomics*, *11*, 533.

Hinske, L. C., Heyn, J., Galante, P. A. F., Ohno-Machado, L., & Kreth, S. (2013). Setting up an intronic miRNA database. *Methods in Molecular Biology* , *936*, 69–76.

Hu, B., Emdad, L., Bacolod, M. D., Kegelman, T. P., Shen, X.-N., Alzubi, M. A., … Fisher, P. B. (2014). Astrocyte elevated gene-1 interacts with Akt isoform 2 to control glioma growth, survival, and pathogenesis. *Cancer Research*, *74*(24), 7321–7332.

Jacinto, E., Facchinetti, V., Liu, D., Soto, N., Wei, S., Jung, S. Y., … Su, B. (2006). SIN1/MIP1 maintains rictor-mTOR complex integrity and regulates Akt phosphorylation and substrate specificity. *Cell*, *127*(1), 125–137.

Jing, Q., Huang, S., Guth, S., Zarubin, T., Motoyama, A., Chen, J., … Han, J. (2005). Involvement of microRNA in AU-rich element-mediated mRNA instability. *Cell*, *120*(5), 623–634.

Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, *28*(1), 27–30.

Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U., & Segal, E. (2007). The role of site accessibility in microRNA target recognition. *Nature Genetics*, *39*(10), 1278–1284.

Kim, Y.-K., & Kim, V. N. (2007). Processing of intronic microRNAs. *The EMBO Journal*, *26*(3), 775–783.

Kolev, N. G., Yario, T. A., Benson, E., & Steitz, J. A. (2008). Conserved motifs in both CPSF73 and CPSF100 are required to assemble the active endonuclease for histone mRNA 3′-end maturation. *EMBO Reports*, *9*(10), 1013–1018.

Krek, A., Grün, D., Poy, M. N., Wolf, R., Rosenberg, L., Epstein, E. J., … Rajewsky, N. (2005). Combinatorial microRNA target predictions. *Nature Genetics*, *37*(5), 495–500.

Lagos-Quintana, M., Rauhut, R., Lendeckel, W., & Tuschl, T. (2001). Identification of novel genes coding for small expressed RNAs. *Science*, *294*(5543), 853–858.

Laplante, M., & Sabatini, D. M. (2012). mTOR signaling in growth control and disease. *Cell*, *149*(2), 274–293.

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

Ledderose, C., Möhnle, P., Limbeck, E., Schütz, S., Weis, F., Rink, J., … Kreth, S. (2012). Corticosteroid resistance in sepsis is influenced by microRNA-124–induced downregulation of glucocorticoid receptor-α*. *Critical Care Medicine*, *40*(10), 2745.

Lee, R. C., & Ambros, V. (2001). An extensive class of small RNAs in Caenorhabditis elegans. *Science*, *294*(5543), 862–864.

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

Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., … Kim, V. N. (2003). The nuclear RNase III Drosha initiates microRNA processing. *Nature*, *425*(6956), 415–419.

Lewis, B. P., Burge, C. B., & Bartel, D. P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, *120*(1), 15–20.

Lewis, B. P., Shih, I.-H., Jones-Rhoades, M. W., Bartel, D. P., & Burge, C. B. (2003). Prediction of mammalian microRNA targets. *Cell*, *115*(7), 787–798.

Long, Y.-S., Deng, G.-F., Sun, X.-S., Yi, Y.-H., Su, T., Zhao, Q.-H., & Liao, W.-P. (2011). Identification of the transcriptional promoters in the proximal regions of human microRNA genes. *Molecular Biology Reports*, *38*(6), 4153–4157.

Martin, J., Jenkins, R. H., Bennagi, R., Krupa, A., Phillips, A. O., Bowen, T., & Fraser, D. J. (2011). Post-transcriptional regulation of Transforming Growth Factor Beta-1 by microRNA-744. *PloS One*, *6*(10), e25044.

Mayr, C., & Bartel, D. P. (2009). Widespread shortening of 3′ UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell*, *138*(4), 673–684.

Ma, Y., Vilanova, D., Atalar, K., Delfour, O., Edgeworth, J., Ostermann, M., … Lord, G. M. (2013). Genome-wide sequencing of cellular microRNAs identifies a combinatorial expression signature diagnostic of sepsis. *PloS One*, *8*(10), e75918.

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

Monteys, A. M., Spengler, R. M., Wan, J., Tecedor, L., Lennox, K. A., Xing, Y., &

Davidson, B. L. (2010). Structure and activity of putative intronic miRNA promoters. *RNA* , *16*(3), 495–505.

Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, *35*(Database issue), D61–5.

Rodriguez, A., Griffiths-Jones, S., Ashurst, J. L., & Bradley, A. (2004). Identification of mammalian microRNA host genes and transcription units. *Genome Research*, *14*(10A), 1902–1910.

Scheid, M. P., Marignani, P. A., & Woodgett, J. R. (2002). Multiple phosphoinositide 3-kinase-dependent steps in activation of protein kinase B. *Molecular and Cellular Biology*, *22*(17), 6247–6260.

Schmitt, D. C., Madeira da Silva, L., Zhang, W., Liu, Z., Arora, R., Lim, S., … Tan, M. (2015). ErbB2-intronic microRNA-4728: a novel tumor suppressor and antagonist of oncogenic MAPK signaling. *Cell Death & Disease*, *6*, e1742.

Singh, P., Alley, T. L., Wright, S. M., Kamdar, S., Schott, W., Wilpan, R. Y., … Graber, J. H. (2009). Global changes in processing of mRNA 3’ untranslated regions characterize clinically distinct cancer subtypes. *Cancer Research*, *69*(24), 9422–9430.

Tie, J., Pan, Y., Zhao, L., Wu, K., Liu, J., Sun, S., … Fan, D. (2010). MiR-218 inhibits invasion and metastasis of gastric cancer by targeting the Robo1 receptor. *PLoS Genetics*, *6*(3), e1000879.

Tranter, M., Helsley, R. N., Paulding, W. R., McGuinness, M., Brokamp, C., Haar, L., … Jones, W. K. (2011). Coordinated post-transcriptional regulation of Hsp70.3 gene expression by microRNA and alternative polyadenylation. *The Journal of Biological Chemistry*, *286*(34), 29828–29837.

Wang, X., & El Naqa, I. M. (2008). Prediction of both conserved and nonconserved microRNA targets in animals. *Bioinformatics* , *24*(3), 325–332.

Xie, X., Lu, J., Kulbokas, E. J., Golub, T. R., Mootha, V., Lindblad-Toh, K., … Kellis, M. (2005). Systematic discovery of regulatory motifs in human promoters and 3’ UTRs by comparison of several mammals. *Nature*, *434*(7031), 338–345.

Yan, L., Hao, H., Elton, T. S., Liu, Z., & Ou, H. (2011). Intronic microRNA suppresses endothelial nitric oxide synthase expression and endothelial cell proliferation via inhibition of STAT3 signaling. *Molecular and Cellular Biochemistry*, *357*(1-2), 9–19.

Zhang, H., Lee, J. Y., & Tian, B. (2005). Biased alternative polyadenylation in human tissues. *Genome Biology*, *6*(12), R100.

Manuskripte der kumulativen Habilitationsschrift

Folgende Manuskripte sind aufgrund von Copyright-Beschränkungen lediglich online einsehbar.

Hinske, L. C., Heyn, J., Galante, P. A. F., Ohno-Machado, L., & Kreth, S. (2013). Setting up an intronic miRNA database. Methods in Molecular Biology , 936, 69–76. DOI: 10.1007/978-1-62703-083-0_5

Hinske, L. C., Heyn, J., Hübner, M., Rink, J., Hirschberger, S., & Kreth, S. (2017). Intronic miRNA-641 controls ist host Gene’s pathway PI3K/AKT and this relationship is dysfunctional in glioblastoma multiforme. *Biochemical and Biophysical Research Communications, 2017 Aug 5;489(4):477-483*. DOI: 10.1016/j.bbrc.2017.05.175

Hinske *et al. BMC Genomics* 2010, **11**:533
http://www.biomedcentral.com/1471-2164/11/533

BMC Genomics

RESEARCH ARTICLE Open Access

A potential role for intragenic miRNAs on their hosts' interactome

Ludwig Christian G Hinske[1,2*], Pedro AF Galante[3], Winston P Kuo[4,5], Lucila Ohno-Machado[2]

Abstract

Background: miRNAs are small, non-coding RNA molecules that mainly act as negative regulators of target gene messages. Due to their regulatory functions, they have lately been implicated in several diseases, including malignancies. Roughly half of known miRNA genes are located within previously annotated protein-coding regions ("intragenic miRNAs"). Although a role of intragenic miRNAs as negative feedback regulators has been speculated, to the best of our knowledge there have been no conclusive large-scale studies investigating the relationship between intragenic miRNAs and host genes and their pathways.

Results: miRNA-containing host genes were three times longer, contained more introns and had longer 5′ introns compared to a randomly sampled gene cohort. These results are consistent with the observation that more than 60% of intronic miRNAs are found within the first five 5′ introns. Host gene 3′-untranslated regions (3′-UTRs) were 40% longer and contained significantly more adenylate/uridylate-rich elements (AREs) compared to a randomly sampled gene cohort. Coincidentally, recent literature suggests that several components of the miRNA biogenesis pathway are required for the rapid decay of mRNAs containing AREs. A high-confidence set of predicted mRNA targets of intragenic miRNAs also shared many of these features with the host genes. Approximately 20% of intragenic miRNAs were predicted to target their host mRNA transcript. Further, KEGG pathway analysis demonstrated that 22 of the 74 pathways in which host genes were associated showed significant overrepresentation of proteins encoded by the mRNA targets of associated intragenic miRNAs.

Conclusions: Our findings suggest that both host genes and intragenic miRNA targets may potentially be subject to multiple layers of regulation. Tight regulatory control of these genes is likely critical for cellular homeostasis and absence of disease. To this end, we examined the potential for negative feedback loops between intragenic miRNAs, host genes, and miRNA target genes. We describe, how higher-order miRNA feedback on hosts' interactomes may at least in part explain correlation patterns observed between expression of host genes and intragenic miRNA targets in healthy and tumor tissue.

Background

microRNAs (miRNAs) are small (~22-nt) functional RNA species that provide a newly appreciated layer of gene regulation with an important role in development, cellular homeostasis and pathophysiology. miRNAs are encoded in the genome and transcribed primarily in a Pol II-dependent manner [1], although Pol III-dependent transcription has also been reported [2,3]. Roughly half of the known human microRNAs are found in intergenic regions of the genome, suggesting production of unique primary transcripts (pri-miRNAs) containing one or more miRNA hairpins under the control of independent promoter elements. The overwhelming majority of the other ~50% map to previously annotated intronic regions of protein coding genes, while a small number are even found within exons. The relationship between intragenic miRNAs and their host genes presents many unique questions regarding genomic organization, transcriptional regulation, processing and function.

The genomic organization of intragenic miRNAs exhibits a strong directional bias, such that these species are predominantly oriented on the same strand of the DNA as that of the host gene. The directional bias may prevent steric interference between RNA polymerases

* Correspondence: ludwig.hinske@med.uni-muenchen.de
[1]Department of Anaesthesiology, Clinic of the University of Munich, Marchioninistrasse 15, 81377 Munich, Germany
Full list of author information is available at the end of the article

transcribing the host gene and the miRNA gene(s) [4]; however, the existence of individual antisense miRNA genes and miRNA gene clusters argues that the primary evolutionary pressure for the positional bias is co-regulation of the intronic miRNA and the host gene. Indeed, microarray analyses supports the hypothesis that intronic miRNAs are usually expressed in coordination with the host gene mRNA in human tissues [4,5], strongly suggesting that co-transcription from the host gene promoter is the most common transcriptional mechanism under normal conditions. This assumption has lately successfully been employed to identify new miRNA targets [6]. However, recent findings demonstrate that transcription of a subset of intronic miRNAs in *H. sapiens* can be initiated from internal promoters within operons independently from the host gene [3], suggesting that utilization of internal promoters must also be considered a viable alternative strategy for intronic miRNA gene transcription.

Large portions of miRNA processing are understood (for review see [7]). In brief, a ~70 nucleotide stem-loop precursor pre-miRNA is excised from a relatively long primary miRNA transcript, followed by export from the nucleus via Exportin-5 in a Ran-GTP-dependent manner. In the cytoplasm, pre-miRNAs are further processed into a ~22-nt miRNA/miRNA* duplex. In the case of intronic miRNAs, early steps in the miRNA biogenesis pathway are complicated by the requirement for proper pre-mRNA splicing and mature mRNA assembly of the host message. Recent bioinformatics and experimental work demonstrates that intronic miRNAs can be processed from intronic regions co-transcriptionally [8] prior to the splicing reaction [9]. Interestingly, recent work suggests that several intragenic miRNAs undergo post-transcriptional regulation [10], and defects in this process have been associated with tumor development [10-14]. The nature of the differences in miRNA processing and associated defects between intergenic and intragenic miRNA species is not currently elucidated.

miRNA target recognition in mammals is mainly mediated via imperfect Watson-Crick base-pairing to cognate sites primarily located in the 3'-UTR of mRNA targets. Predicted and validated miRNA targets include a functionally diverse suite of genes that include many transcription factors and cell signaling proteins, suggesting a role for miRNAs in regulatory feedback loops [15-17]. Intragenic miRNAs present unique regulatory possibilities based on functional relationships with their host genes. It has been speculated that intronic miRNAs may directly target their host message or regulate transcription factors, in what is commonly designated “first-order” or “second-order” negative feedback, respectively [18]. Recently published work [19] demonstrates that miR-338, encoded in an intron of the apoptosis-associated tyrosine kinase (AATK) gene, targets several genes that are functionally antagonistic to the AATK protein. Therefore, miR-338 serves the functional interest of the host in this case via a higher-order positive feedback system that downregulates expression of AATK repressors and enforces neuronal differentiation downstream of the kinase.

In the current manuscript, large-scale bioinformatics analyses of human intronic miRNAs related to genomic organization and characterization of miRNA host and target genes are presented. We identify characteristics of host genes and predicted targets, and present evidence that intragenic miRNAs may act as negative feedback regulatory elements of their hosts' interactome (i.e., they can regulate host gene neighbours in addition to host genes).

Results

We integrated genomic and transcriptomic information to analyze properties of intragenic miRNAs themselves, their host genes, as well as their targets. We used all known miRNAs (based on miRBase), all known human transcripts (based on RefSeq), six different and highly established miRNA target prediction algorithms, as well as the gene and pathway annotation ontologies GO and KEGG.

Classification of miRNAs

Based on mapping miRNA genomic coordinates to genomic position of all known genes and their exons and introns (based on RefSeq sequences [20]), we could classify miRNAs into three classes: intergenic, exonic, and intronic (Table 1). For *H. sapiens*, 296 miRNAs were located within intronic regions, and 37 within exonic regions of known genes. We also classified miRNAs from other species (Table 1). Interestingly, organisms that have a well-annotated set of protein-coding genes present distributions that resemble that of the miRNA distribution in humans, showing 33-48% of intronic miRNAs and 0.6-6% of exonic miRNAs (Table 1, organisms *M. musculus, D. melanogaster* and *C. elegans*). On the other hand, organisms containing a smaller number of annotated genes presented a higher number of intergenic miRNAs (Table 1, organisms *C. familiaris, G. gallus* and *D. rerio*), some of which however may become intragenic as more genes will be identified in these organisms. Additional file 1 contains details of miRNAs classification, their genomic position and host genes.

Positional Bias of Intragenic miRNAs

The orientation of the gene for an intronic miRNA depends significantly on the transcription direction of its host gene (p-value = 1.3 × 10-36 in χ^2 test) as shown in Table 1. We found that 65.5% of host genes had

Table 1 Classification of miRNAs in the Genome of Different Species

Organism	Intragenic miRNAs		Intergenic miRNAs	Intragenic miRNAs	
	Intronic	Exonic		miRNAs on Host Gene Strand	miRNAs on Opposite Host Strand
Homo sapiens	296 (42.6%)	37 (5.3%)	362 (52.1%)	282 (84.7%)	51 (15.3%)
Mus musculus	171 (35.4%)	30 (6.2%)	282 (58.4%)	163 (78.2%)	38 (21.8%)
Canis familiaris	3 (1.5%)	0 (0%)	201 (98.5%)	2 (66.7%)	1 (33.3%)
Gallus gallus	50 (10.7%)	1 (0.2%)	418 (89.1%)	46 (90.2%)	5 (9.8%)
Danio rerio	48 (15.0%)	1 (0.3%)	271 (84.7%)	39 (79.6%)	10 (20.4%)
Drosophila melanogaster	65 (42.8%)	2 (1.3%)	85 (55.9%)	53 (79.1%)	14 (20.9%)
Caenorhabditis elegans	51 (33.1%)	1 (0.6%)	102 (66.2%)	33 (63.6%)	19 (36.5%)

Intragenic miRNAs are found in many different species. However, the distribution of intra- and intergenic miRNAs differs. These numbers are obtained by crossing miRNA genomic coordinates with known transcript coordinates (based on RefSeq sequences).

miRNAs in the first five introns. Also, we confirmed that the observed distribution differs significantly from the expected distribution within the first five introns (p = 0.030 in χ^2 test, additional file 2).

Characterization of Host Genes

Assuming, as is widely accepted, that intragenic miRNAs share a common regulatory control with their host genes, we can infer functional aspects of this class of miRNAs by characterizing features of those host genes. To confirm that the position of miRNAs has a particular bias, and is not the result of chance, we randomly sampled genes that matched the set of miRNA host genes (in terms of chromosome and strand distribution) and compared the positions of host, target, and randomly sampled genes. The findings are summarized in Table 2. Host genes are almost three times longer than the randomly sampled cohort and have more introns. When comparing the intron size in different positions (Figure 1a), we found that the first five 5' introns are significantly longer, consistent with our previous finding that most host genes' intronic miRNAs are found in the 5' introns (Figure 1b).

Gene expression can be pre- and post-transcriptionally controlled through regulatory motifs in their 3'-UTRs. Even though regulatory mechanisms are not well understood, two important concepts include regulation through miRNAs, and the role of adenylate/uridylate-rich elements (AREs) in mediating mRNA decay, which plays a significant role in cancer development [21-23]. We first compared the length of the 3'-UTRs of host genes to the length of 3'-UTRs of the random sample. Host genes have 40% longer 3'-UTRs (*p-value* < 0.01). In a second step, we counted occurrences of the pentamer AUUUA in these regions, normalized by the length of the 3'-UTRs. We found significantly more ARE units

Table 2 Properties of Host and Target Genes

Property	Gene Set	Median[Range] Host/Target	Median[Range] Control	Ratio	p-Value
Total length (basepairs)	Host Genes	84871.0 [2792-2220381]	29324.5 [599-2304633]	2.89	< 2.2e-16
	Target genes	83747.5 [2366-2220381]	30232.5 [218-2220381]	2.77	< 2.2e-16
Introns	Host Genes	13[1-88]	8[1-105]	1.62	4.3e-13
	Target genes	10.5[0-78]	8[0-311]	1.31	9.77e-07
Length 5'UTR (basepairs)	Host Genes	279.5[0-385608]	298.5[0-1098107]	0.94	0.25
	Target genes	439.5[0-460277]	282.5[0-1098107]	1.56	2.32e-08
Length 3'UTR (basepairs)	Host Genes	1218.5[0-535884]	872[0-321862]	1.4	4.71e-05
	Target genes	1764[171-11799]	872[0-72058]	2.2	< 2.2e-16
ARE (absolute)	Host Genes	2.0[0-1794]	1.0[0-592]	2.0	3.89e-04
	Target genes	5.0[0-47]	2.0[0-187]	2.5	< 2.2e-16
ARE (per kb)	Host Genes	1.9[0-2.74]	1.49[0-0.045]	1.26	0.012
	Target genes	2.69[0-14.22]	1.63[0-76.92]	1.65	< 2.2e-16
5' UTR GC content	Host Genes	0.6[0.31-0.95]	0.59[0-1]	1.06	0.015
	Target genes	0.59[0.26-1]	0.58[0-1]	1.01	0.71

Host and target genes display similar properties, compared to a set of control genes, including increased length, higher number of total introns, longer 3'UTRs and higher frequency of "AU-rich elements" (AREs).

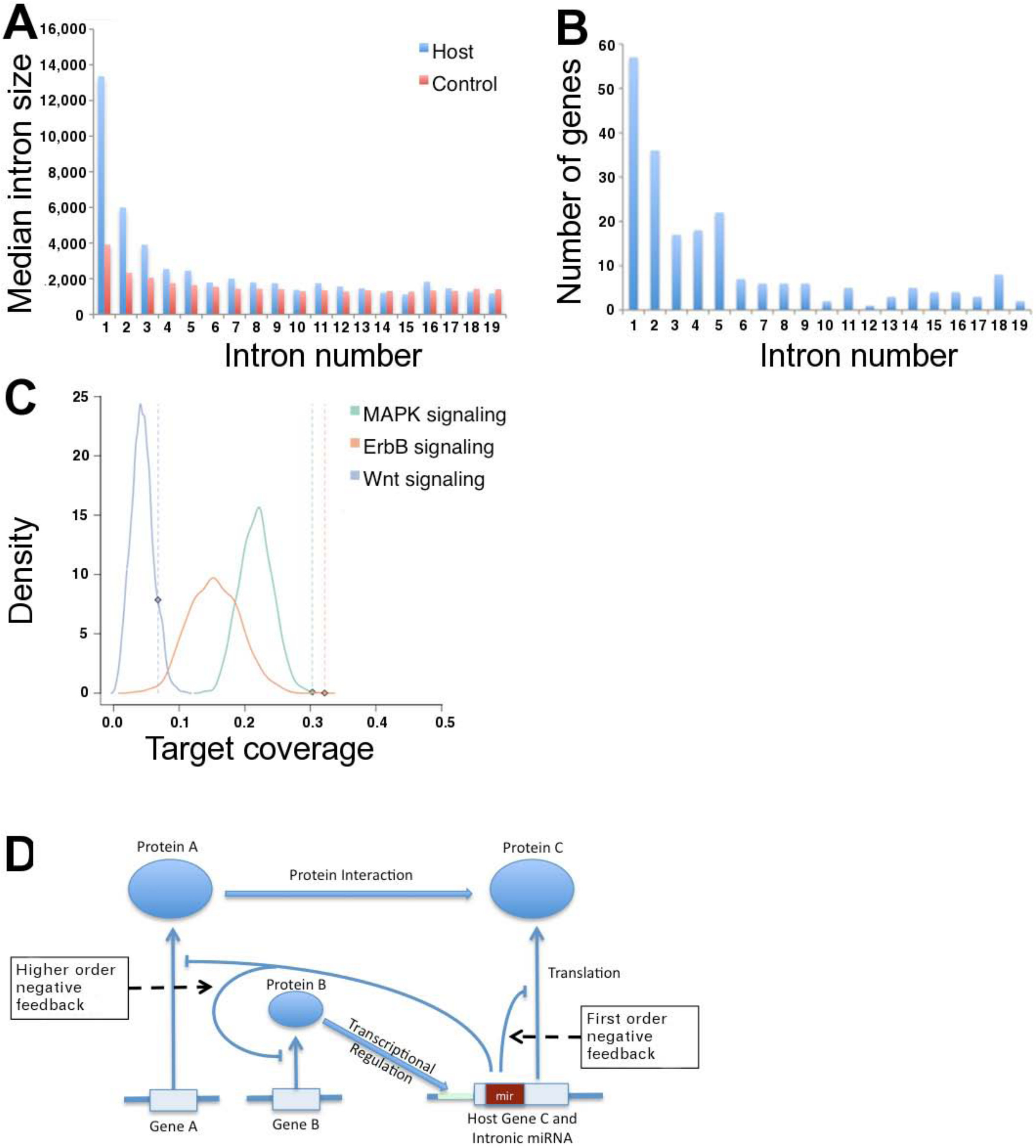


Figure 1 Intragenic miRNA Properties. (a) The sizes of host gene introns closest to the TSS are significantly larger than those of respective controls. (b) Intronic miRNAs appear to be unevenly distributed across the intronic regions of their host genes. More then half of intronic miRNAs are located within the first five 5' introns of their hosts. (c) The observed target coverage (diamond) is at the upper end of the random distribution of target coverages for different signaling pathways, indicating that intragenic miRNAs have more targets within the pathway than would be expected by chance. (d) Intragenic miRNAs may control their host in the setting of a negative feedback circuitry not only through direct regulation of the host's transcription, but also on the order of the interactome, by controlling other genes in the host's pathway.

in host genes than in the random sample (*p-value* < 0.01). Since recently miRNA target genes have been shown to be larger than non-target genes [24], we analyzed total lengths and lengths of 3'-UTRs [25] for host genes predicted to be targeted by their intragenic miRNA and the remaining host genes separately. No significant difference in lengths between the two groups of host genes was observed (*p-value* = 0.3939), but genes in both groups were longer than genes in the control group (*p-value* = 1.552e-07 and *p-value* < 2.2e-16).3'-UTRs were longer in host genes predicted to be targets of their intronic miRNA than in host genes not predicted to be targets (*p-value* = 0.001) and control genes (*p-value* = 5.012e-07). In contrast, the 5'-UTRs of host genes were not significantly longer than the ones in the control group (*p-value* > 0.05).

The GO Biological Process (GOBP) and KEGG are ontologies that associate genes with, cellular processes and biochemical pathways, respectively, including disease pathways. When surveying GOBP for overrepresentation of miRNA host genes in certain categories, we found significant enrichment in gene regulatory, metabolic, neurogenic, and cytoskeletal processes, which reflects the broad range of diseases with which miRNAs have been associated [12,26-34]. Additionally, we found that host genes were overrepresented in several signaling pathways, such as the *MAPK*, *ErbB*, *VEGF*, and the calcium signaling pathway.

Genomic Properties of Target Genes

We looked at genomic properties of a high-confidence set of targets for hosts of intronic miRNAs (prediction agreement ≥ 6) that would give us a set of similar size as the host genes. We then randomly sampled RefSeq transcripts to match chromosome and strand distribution as a control set and performed the analysis analogously to the analysis of genomic properties of the host genes themselves. Table 2 summarizes the results, revealing that the predicted targets have properties that are highly similar to those of host genes.

Relationship Between Intragenic miRNAs and Host Genes

We found that approximatelly 20% of intragenic miRNAs (56 of them, hosted in 49 distinct genes) are predicted to target their own host by at least two methods. This number is significantly higher than would be expected by chance alone (*p-value* < 0.001, obtained by random sampling). Furthermore, we assessed the robustness of our approach by following the above procedure while applying a voting method as the gold standard. We assigned each of the target prediction methods to one of two groups of equal size (n = 3) and required at least one vote from each group to consider that a prediction of a miRNA-host interaction. TarBase did not contain a single instance of miRNA-host interaction, so it was excluded from the analysis. Although the numbers of miRNAs predicted to target their own host varied (12 - 55), depending on which group they had been assigned to, in each case the observed number was significantly higher than would be expected by chance (*p-value* < 0.05, see also additional file 3). Given that host genes that were predicted to be targets of their intragenic miRNA have longer 3'-UTR regions, statistical significance of the number of hosts being targeted by their intronic miRNAs was assessed by repeated creation of sets of non-host control genes with similar 3'-UTR distribution (see Materials and Methods). In line with our previous observations, the number of hosts predicted to be targeted by their intragenic miRNAs (49) was significantly higher than expected by chance (*p-value* = 0.032).

In order to test the hypothesis that intronic miRNAs might act as regulators even in the global functional context of a negative feedback loop circuitry, the KEGG pathway analysis was extended to identify targets within the respective biomolecular pathway. We defined the target coverage as the number of genes within a pathway that were predicted targets (prediction agreement ≥ 2) of miRNAs residing in host genes within that pathway, over the total number of genes in the pathway. To check whether the observed target coverage could be expected by chance, the original genes contained in the pathway were replaced by a set of randomly sampled genes and the expected target coverage of intronic miRNAs with host genes in a particular pathway was calculated. The distributions of expected target coverage for three signaling pathways are visualized in Figure 1c. At a false discovery rate (FDR) of 10%, 22 out of 74 pathways with which host genes were associated showed a significant overrepresentation of targets in the hosts' pathways (Table 3, Additional File 5). Interestingly, many signalling and malignancy-related pathways ranked high.

Implications for Cancer Pathogenesis

Integration of major KEGG pathway information with expression data from two publicly available datasets [35,36] helped us investigate the idea of loss of negative feedback circuitry.

KEGG ID "05215 - Prostate Cancer" contains a single known miRNA host (*AKT2*), and it is not predicted to be targeted by its intronic miRNA (*hsa-miR-641*). The correlation between the expressions of host and predicted targets involved in the pathway were calculated. Figure 2 shows a simplified representation based on the KEGG pathway information. Host and corresponding targets are color-coded, where the green oval indicates the host, *AKT2*, and yellow, orange, and red indicate whether two, three or four methods agreed on the target prediction.

Table 3 Pathways with Overrepresentation of Genes Targeted by an Intronic miRNA

Pathway	Host Genes in Pathway	Target Coverage	p-Value	q-Value
MAPK Signaling	ATF2; DDIT3; AKT2; FGF13; ARRB1; PPP3CA; PRKCA; CACNG8; RPS6KA2; MAP2K4; RPS6KA4	61.4%	< 0.001	< 0.001
Axon Guidance	PPP3CA; PTK2; SEMA4G; SEMA3F; SLIT3; ABLIM2; SLIT2	70.3%	< 0.001	< 0.001
Ubiquitin Mediated Proteolysis	HUWE1; WWP2; BIRC6; ITCH	53.8%	< 0.001	< 0.001
Focal Adhesion	COL3A1; AKT2; PRKCA; PTK2; TLN2	49.5%	< 0.001	< 0.001
Glioma	AKT2; PRKCA	52.3%	< 0.001	< 0.001
Melanoma	AKT2; FGF13	50.7%	< 0.001	< 0.001
Regulation of Actin Cytoskeleton	CHRM2; FGF13; SSH1; PTK2	41.0%	< 0.001	< 0.001
Chronic Myloid Leukemia	AKT2	38.2%	< 0.001	< 0.001
Colorectal Cancer	AKT2	35.7%	< 0.001	< 0.001
Prostate Cancer	AKT2	34.8%	0.001	0.007
Melanogenesis	PRKCA	21.6%	0.001	0.007
Pancreatic Cancer	AKT2	35.6%	0.002	0.01
ErbB Signaling	ERBB4; AKT2; PRKCA; PTK2; MAP2K4	51.7%	0.003	0.02
Glycan Structures Biosynthesis	MGAT4B; FUT8; CSGLCA-T; GALNT10; HS3ST3A1	50.8%	0.003	0.02
Gap Junction	HTR2C; PRKCA; PRKG1	47.9%	0.005	0.02
Non-Small Cell Lung Cancer	AKT2; PRKCA	42.6%	0.007	0.03
Small Cell Lung Cancer	AKT2; PTK2	35.6%	0.013	0.05
Long-Term Depression	PRKCA; PRKG1	33.3%	0.014	0.05
Insulin Signaling	AKT2; SREBF1	36.0%	0.014	0.05
Long-Term Potentiation	PPP3CA; PRKCA; RPS6KA2	27.1%	0.005	0.06
T-Cell Receptor Signaling	AKT2; PPP3CA	32.3%	0.016	0.09
Wnt Signaling	PPP3CA; PRKCA	21.6%	0.020	0.09

22 out of 74 pathways containing host genes show a significant overrepresentation of targets within the pathway at a FDR of 10%. Host genes that were predicted targets of their own miRNA were removed from the count. Interestingly, the list of pathways contains many pathways crucial for development and signal transduction, or associated with neoplastic transformation.

In line with the hypothesis of an interactome feedback circuitry, predicted targets of *hsa-miR-641* appear to be in close proximity and in functional synergy with its host. A similar target pattern is displayed by both miRNAs, *hsa-miR-641* and *hsa-mir-634*, in the non-small-cell lung cancer pathway (additional file 4).

The correlation between host and target expression levels is shown in a two-bar plot. The first bar, labeled "N", represents the correlation between host and target in normal tissue. The second bar, labeled "T", represents the correlation between host and target in cancerous tissue. In the prostate cancer dataset, seven of the fifteen targets are more negatively correlated in healthy tissue than in cancer. In four cases (*AKT3*, *AR*, MAPK1, and *CTNNB1*), we could observe a significant negative correlation in normal tissue, which was either non-significant or was significantly positive in cancer. A similar pattern could be observed in the non small cell lung cancer pathway.

Discussion and Conclusions

Since the first discovery of miRNAs, our understanding of biogenesis and regulation has exponentially grown. In the recent past, it has been estimated that miRNAs that reside in intronic or exonic regions of other genes may be the dominating class [9]. However, functional aspects of intragenic miRNAs are still largely unknown.

It is generally believed that both host and miRNA share regulatory control [4-6], although a recent study found that transcription of roughly 30% of intragenic miRNAs may be initiated independently [3]. After mapping miRNAs to known genes, we found that most intronic miRNAs are oriented in the same direction as their host gene, significantly more than would be expected by chance. Several hypotheses related to this preferential orientation have been suggested. First, most of intragenic miRNAs may not present their own promoter and be dependent to the transcription of their host gene. Second, miRNAs may present their own promoter, and directional bias may prevent physical interference between RNA polymerases transcribing the host gene and RNA polymerases transcribing the miRNA gene [4].

Baskerville and Bartel identified significant correlation between the expression levels of intronic miRNAs and their host genes, suggesting co-regulation [4]. We

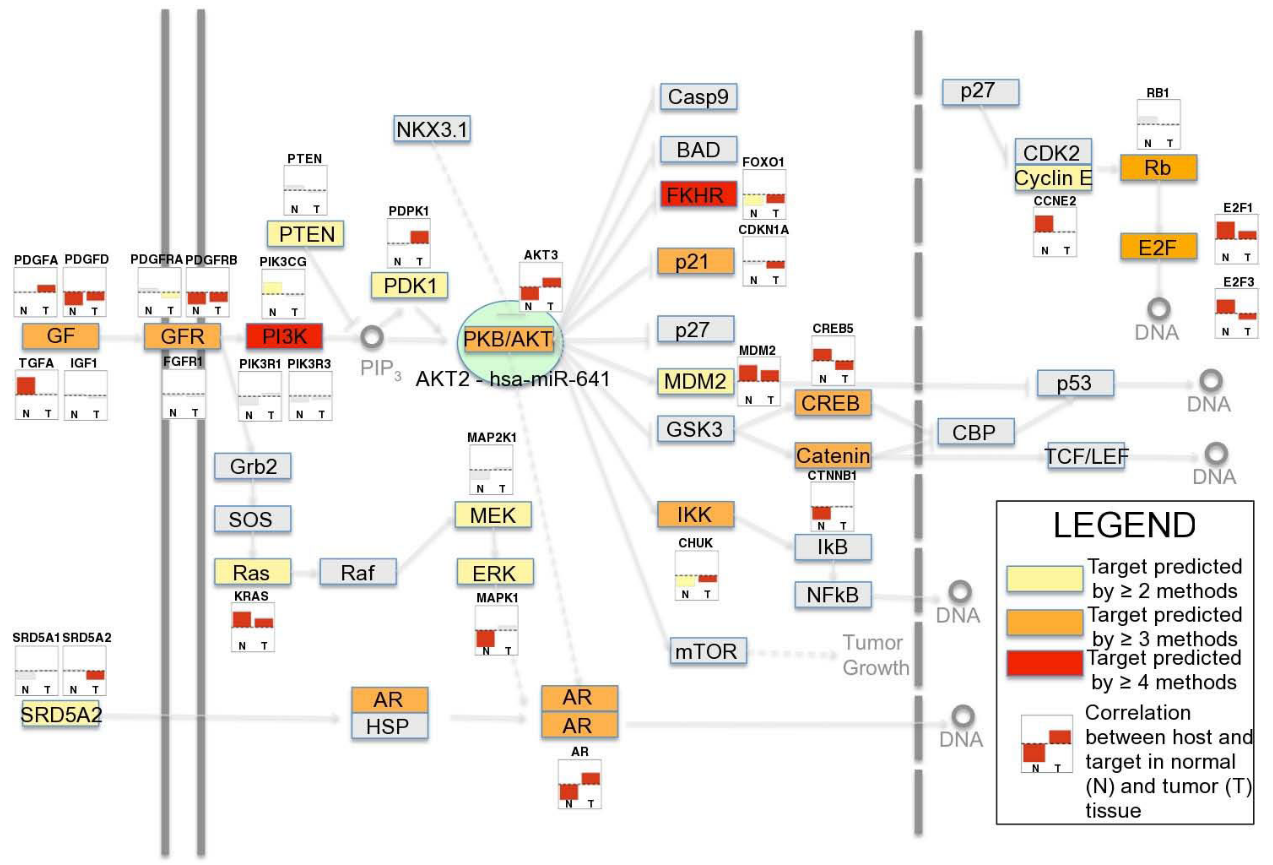


Figure 2 Correlation of Predicted miRNA Targets with Hosts. The *PI3K - PKB/AKT* pathway is believed to be a key component in cancer development. We compared correlation of miRNA predicted targets (prediction agreement ≥ 2) to the respective host in normal (N) and tumor (T) tissue. Several of the hypothesized targets display features predicted by our model, such as *AR*, *PDGFB*, *PDGFRB*, *AKT3* and *MAPK1*.

furthermore found that more than half of intronic miRNAs are found in the 5' regions of their host genes, where introns are firstly excised. It is well known that transcriptional activity is higher towards the 5' region of a gene [37] and also that regulatory motifs tend to reside in these regions [38]. From a functional perspective, these findings may suggest dependency between host and miRNA transciption. In order to characterize the relationship between intronic miRNAs and their hosts, we identified properties of the set of host genes, as well as a set of high confidence targets. Whereas Golan et al. [39] showed in a recent work that intronic miRNA density is lower in large host genes, we provided evidence that the class of host genes in general is significantly longer and contains more and larger introns. This increases transcriptional efforts for the cell and is considered a characteristic of tightly regulated genes [40]. Interestingly, these features can also be found in a high-confidence set of targets (i.e. prediction agreement ≥ 6 methods), which may support the idea of miRNAs as regulators of their own host genes. Additionally, the 3'-UTRs of host genes predicted to be targeted by their own miRNA are significantly longer, exposing the message to more regulatory control mechanisms, such as targeting by miRNAs or ARE mediated mRNA decay. Interestingly, host genes contain significantly more AREs. Many of these properties have been shown to be features of proto-oncogenes and the sum of these findings may suggest tight regulatory control of these genes [21,23,41]. Surveying GOBP and KEGG pathways, we found host genes to be associated with metabolic, biosynthetic, gene regulative processes, and signaling pathways. These categories capture major functional aspects of miRNAs in general, as is reflected by miRNA involvement in diseases such as cancer [32], muscle disorders [27], or neurodegenerative diseases [42]. We then assessed predicted targets, using agreement between six distinct prediction algorithms and a database of validated miRNA targets as a measure of confidence. First, we identified 56 miRNAs predicted to target their own host. Interestingly, more of these miRNA-host gene pairs are conserved than of the remaining miRNA-host gene pairs (Table 4). Recently, Sun et al. validated the predicted interaction between *hsa-miR-126* and its host

Table 4 Conservation of miRNA-Host Pairs

Organism	miRNA-Host Pairs Predicted to Target own Host		miRNA-Host Pairs Not Predicted to Target own Host		p-Value
	Conserved	Total	Conserved	Total	
Homo sapiens - Mus musculus	18 (35.2%)	51	41 (24.5%)	167	0.18
Homo sapiens - Canis familiaris	1 (2.12%)	47	0	142	0.56
Homo sapiens - Gallus gallus	5 (12.5%)	40	1 (0.71%)	139	0.001

The subset of intragenic miRNA host pairs where the miRNA is predicted to target its own host shows a tendency to be more conserved. However, statistical significance can only be shown for conservation between human and chicken (2-sample test for equality of proportions with continuity correction).

EGFL7 [43,44]. By integrating KEGG pathways with these predictions, we observed for 22 of the 74 pathways that host genes were associated with a higher number of targets within the pathway than would be expected by chance alone. A visual representation of the targets of *AKT2*'s intronic miRNA *hsa-miR-641*, for example, showed how components of many protein complexes involved in the signal transduction of growth factor signalling may be potential targets of *hsa-miR-641* (Figure 2). The combination of these findings indicates that intragenic miRNAs may play a role in interactome feedback circuitries, as visualized in Figure 1d, as an additional security switch for genes requiring narrow control. A subset of up to 20% of intragenic miRNAs may directly regulate the host expression (we referred to this phenomenon as "first-order feedback"). Moreover, intragenic miRNAs display targeting patterns that appear not only to influence their hosts' expression levels, but also their functional environment. The observation that structural properties of a set of high-confidence-prediction target genes, such as long 3'-UTRs, length, and number of AREs, resemble those of host genes emphasize the concept of regulation of interacting gene products in highly restricted settings.

Loss of negative feedback control systems is a well-known mechanism by which cancer develops. Blenkiron and coworkers [11] recently suggested that miRNA processing might be disturbed in cancer. If expression levels of intragenic miRNAs are reduced, as observed by some authors [13,45], subsequently important signalling pathways may lose inhibition and this may facilitate uncontrolled cell growth. In a recent study, Tavazoie et al. analyzed six miRNAs that were significantly underexpressed in breast cancer LM2 cells, as compared to normal breast tissue. Four of these miRNAs were intragenic [46]. The authors reported that loss of the intronic miRNA *hsa-miR-335*, which resides in intron 2 of its host gene *MEST*, led to increased migration and invasion rates and hence increased metastatic capacity. Additionally, they could show that *hsa-miR-126* (intron 7, host *EGFL7*) significantly reduced proliferation of breast cancer cells. Likewise, *hsa-miR-151* has been shown to be downregulated in chronic myloid leukemia through *BCR/ABL* [47], and silencing its host gene *PTK2* inhibits leukemogenesis [48]. A similar pattern can for example be found for *hsa-miR-504* and *FGF13* [49,50].

Changes in miRNA biosynthesis such as those found in cancer can interfere with the coordination of expression of miRNA and host. Thus, a negative correlation between expression levels of host and genes targeted by its intragenic miRNA in normal tissue (given that the host is not targeted by the miRNA it contains) and a less negative or even positive correlation in cancerous tissue might be expected. This phenomenon was observed in two distinct datasets in different malignancies (Figure 2, additional file 4 and additional file 6). A key to pathogenesis of both entities is the phosphatidylinositol 3-kinase(PIK3)/AKT signaling pathway, deregulation of which has been reported in several cancers, including prostate cancer [51], lung cancer [52], ovarian cancer [53,54], breast cancer [53,55], and colon tumors [54]. Whereas Noske et al. discovered that silencing *AKT2* through RNA interference leads to reduction in ovarian cancer cell proliferation [56], Maroulakou and coworkers reported accelerated development of polyoma middle T and ErbB2/Neu-driven mammary adenocarcinomas in mice after *AKT2* ablation [57]. Although these findings would appear to be contradictory at first, they can be explained by an intragenic miRNA-driven negative regulatory loop that is disturbed in cancer. Whereas in the first experiment *AKT2* was targeted on mRNA level (and therefore mimicking the role of the corresponding intronic miRNA), in the second experiment both host mRNA and miRNA (if it exists in mouse) were downregulated, and therefore may have disabled a potential negative feedback regulation by *hsa-miR-641*.

One must remember, however, that regulatory networks are far more complex in reality than what we are currently able to model. Transcription factors, enhancers, silencers, and epigenetic modifications play major roles in cancer development and may influence correlation among expression levels of hosts and targets. Also, target prediction methods are error prone, and at this point we can only speculate about the true nature of events and therefore plan to conduct further experiments in which the hypotheses presented can be tested. For example, by integrating different target prediction methods, roughly 20% of intragenic miRNAs were predicted to target their own host. Though this number is significantly higher than expected by chance, it still does not cover the majority of miRNAs. For one, this number may underestimate the true number of miRNAs targeting their own host due to limitations of target prediction methods. Additionally, it has lately been shown that transcription of one third of intronic miRNAs can be initiated independently of the host's transcription [3], in which case direct feedback cannot be claimed. Also, we only investigated feedback on the level of direct miRNA-host interaction and on the order of the interactome based on the KEGG database. However, knowledge about interaction of proteins is still limited and cotranscription of host and miRNA may enable more complex mechanisms. Limitations to current knowledge may also justify, why a significant fraction of predicted targets in do not show the expected behaviour. Indeed, *Cyclin E* and *E2F* in Figure 2 show opposite behavior than what we would expect. Neither of these genes might actually be a target of *hsa-miR-641*; there may also exist stronger regulating elements that control their expression, or the primary mode of silencing in that specific situation may be through translational repression. Nevertheless, it is interesting how key molecules in two different datasets displayed predicted correlation patterns.

Further experiments and biological validation of computational evidence presented here may have great implications, especially in cancer therapy. Modern therapies usually target central molecules, such as *AKT* and *PI3K* with some success. However, these techniques control only single elements in a cascade of complex signalling events. In summary, our findings encourage more focused research on intragenic miRNAs and their targets.

Methods

Classification of miRNAs

miRNA genomic coordinates from miRBase release 11 (April 2008) [58-60] were crossed to genomic coordinates of RNA Reference Sequences (http://www.ncbi.nlm.nih.gov/RefSeq; Release 31) [20] downloaded from UCSC Genome Browser http://genome.ucsc.edu. To each genomic mapped RefSeq sequence, a single gene was assigned. The subset of miRNAs whose coordinates mapped to an annotated gene was defined as intragenic. Intragenic miRNAs were classified as exonic when their coordinates overlapped with any observed exonic region, and intronic otherwise.

Host Genes' Intronic miRNA Distribution

Introns were sequentially enumerated based on gene orientation. For each intron number, host genes containing miRNAs in this intron were counted. We calculated the expected number of genes containing an intronic miRNAs in a given intron number by adding all intron lengths of introns with the respective intron number and dividing it by the summed length of all host genes' introns, thus accounting for intron frequency and length.

Gene Ontology

The Gene Ontology [61] classifications of all 246 host genes of intragenic miRNA genes that were located on the same strand as their host gene were surveyed using Cytoscape 2.6.0 [62] and BiNGO 2.3 [63]. We focused our attention on those categories that were disproportionately overrepresented. The setting "Hypergeometric test" was chosen to calculate the probability of observing an equal or greater number of genes in a given functional category than in the test set. The False Discovery Rate (FDR), which is the standard setting in BiNGO 2.3 [63], was controlled.

Pathways identification

The statistical programming software R 2.7.1 was used in combination with Bioconductor [64,65] packages AnnBuilder 1.18.0, KEGG.db version 2.2.0, and GOStats version 1.7.4 to acquire a list of pathways that were associated with one or more of the 246 host gene proteins.

Target Predictions

Strategies to perform high-throughput miRNA target validation are still very limited. Therefore, target prediction algorithms are employed to allow large-scale assessment of miRNA-target interaction. However, usage of target prediction methods raises two difficulties. First, target prediction methods are known to suffer from a significant number of false positive predictions. We reasoned that a possible way to address this problem would be to estimate statistical significance by generating background distributions by the very same methods. Hence, if target predictions were too close to random, the mean of the generated background distribution should be close to the observed number, whereas a significant finding should not be affected by the absolute number of false positives. Second, different target prediction algorithms incorporate different types of information about miRNA target interactions. To overcome individual biases that may be introduced by one specific method and use the wide range of experimental knowledge gained, we integrated predictions from six current algorithms. Precalculated target predictions for TargetScan release 4.2 [66] (April 2008), PITA [67] catalog version 6 (August 2008), MirTarget2 (mirDB) version 2.0 [68,69] (December 2007), miRanda [70] (September 2008), RNA22 [71] (November 2006) and PicTar 5-way [72] were downloaded. We also included TarBase version 5.0c [73] (June

2008) as a reference database for miRNA target interactions with published evidence; only targets with a "Support Type" value of either "True" or "Microarray" were selected. Some miRNA symbols did not exactly match entries in the database for various reasons, including use of non-official names or older miRBase releases. Whenever a miRNA symbol could not be found, matching was attempted to an extension such as "-1" or "a" (for example, *hsa-mir-511* in mirTarget2 was matched to *hsa-mir-511-1* and *hsa-mir-511-2*). If the miRNA symbol ended with a letter, it was removed to check for other matches (from the PicTar prediction list *hsa-mir-128a* matched to *hsa-mir-128-1*, *hsa-mir-128-2*, and *hsa-mir-128-3* for example). Predictions for a miRNA symbol were ignored if no matches could be found. Due to the diversity of underlying principles, assumptions, and scoring systems, we defined the prediction agreement, i.e. the number of methods that agree on a certain miRNA target prediction, as a measure of confidence in the target prediction. In recent work, Selbach et al. measured changes in protein and mRNA expression after transfection and overexpression of five different miRNAs (hsa-miR-1, hsa-miR-16, hsa-miR-30a, hsa-miR-255, hsa-let-7b) in HeLa cells [74]. We evaluated the different target prediction methods used in this study by measuring the abundance of predicted products (mRNA or the proteins encoded by these mRNAs, a continuous value) and assessing discrimination by areas under the ROC curve using the predicted targets as the binary outcome. All five miRNA datasets were pooled (see additional file 3 for details). The AUC (Area under Receiver Operator Characteristic (ROC) Curve) measures how well predictions and non-predictions can be discriminated at all possible thresholds, with a value of 0.5 indicating no discrimination and a value of 1 indicating perfect discrimination. Target prediction methods varied greatly in AUCs, ranging from 0.55 to 0.92 in protein measurements. With increasing prediction agreement, an almost linear increase in AUC can be observed, indicating that prediction agreement may be used as a proxy for the confidence of a predicted miRNA target interaction (Figure 3).

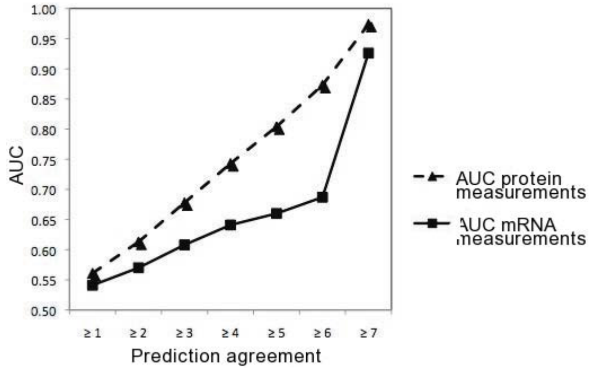


Figure 3 Prediction Agreement as a Measure of Confidence. When constructing an ROCs on protein measurements, there is an almost linear relationship of the resulting AUCs and prediction agreement. This is also true for mRNA measurements, though the slope is less steep.

Gene Expression Datasets

Two publicly available mRNA expression datasets (GSE6956, GSE7670) were downloaded from the Gene Expression Omnibus http://www.ncbi.nlm.nih.gov/geo. We included 87 prostate samples (69 tumor and 18 healthy tissue samples) [35] and 60 lung samples (31 non-small-cell lung cancer and 29 healthy lung tissue samples) [36]. Preprocessing was carried out using BioConductor packages [64,65]. Data from protein and mRNA expression change after miRNA transfection experiments were downloaded from http://psilac.mdc-berlin.de[74].

Genomic Host and Target Gene Properties

In order to assess genomic properties of host genes (n = 246), we constructed a set of control genes (n = 2460) that would match chromosome and strand distribution of host genes in order to exclude structural differences due to chromosomal specificities. We defined miRNA target interactions predicted by at least 6 methods as "high confidence targets" (n = 326). These predictions cover 33 host genes and 43 miRNAs when at least six methods are required and 239 hosts and 272 miRNAs when at least 2 methods are required. Statistical testing was done using Mann-Whitney-U test. For the analysis of total length and 3'-UTR length of host genes, hosts were additionally split into two groups, dependent on whether they were predicted to be targets of their intragenic miRNA. We combined the Kruskal-Wallis rank sum test with post-hoc pairwise Mann-Whitney-U test with Bonferroni correction ($p < 0.016$ defined as significance cut-off for three pairwise comparisons). Assessment of host genes predicted to be targeted by their intragenic miRNAs was carried out as follows: Out of the 2460 control genes, we sampled 1000 sets of genes of size 246 that would match the host gene 3'-UTR length distribution (no significant difference in Mann-Whitney-U test). Intragenic miRNAs were assigned to genes in the sets and the number of genes predicted to be targeted by that miRNA was calculated. Similarly, the observed number of miRNAs predicted to target their own host was assessed by exchanging host genes for randomly chosen genes from predicted targets and recalculation of the number of miRNAs predicted to target their host. Robustness of this approach was tested by additionally requiring a vote from each of two groups of three prediction methods each.

Host-miRNA Conservation

HomoloGene database NCBI release 61 http://www.ncbi.nlm.nih.gov/homologene and mirBase release 11 (April 2008) [58-60] were used to identify homologous host genes in *Homo sapiens*, *Canis familiaris*, and *Gallus gallus*. Proportions of conserved miRNA-host gene pairs for miRNAs predicted and miRNAs not predicted to target their own host were calculated. Similarly, we used information on target site conservation from TargetScan to calculate the proportion of conserved targetsites of predicted target host interactions in the hosts' pathway and of those not in the hosts' pathway. Statistical significance was assessed using the 2-sample test for equality of proportions with continuity correction.

Target Coverage

The union of predicted targets included more than 90% of all known human genes. Since target prediction methods are very different, they are difficult to compare. In this work, only targets that were predicted by at least two different methods were considered in the calculation of target coverage. This reduced the total number of predictions by almost 70%.

We defined the set S_p as the set of genes linked to a pathway and S_t as the set of predicted targets of the miRNAs associated with the pathway through their host genes. The target coverage (*C*) for a pathway was defined as

$$C = \frac{|S_p \cap S_t|}{|S_p|}.$$

Statistical significance of target enrichment within a pathway was tested by randomly sampling $|S_p|$ genes from a universe of all known genes, replacing the genes within the pathway with the set of genes in the random sample (S_i), and subsequently calculating a new "random" target coverage C_i'. This procedure was repeated 1000 times, allowing estimation of the probability as the number of times a target coverage C_i' greater or equal to *C* was observed. We defined the indicator function I (C_i',*C*) as

$$I(C_i', C)\begin{cases} 1 & if\ C_i{'} \geq C \\ 0 & otherwise \end{cases}.$$

Hence, the probability of observing greater or equal target coverage for a given pathway could be estimated as

$$p(C' \geq C) = \frac{\sum_{i=1}^{1000} I\left(\frac{|S_i \cap S_t|}{|S_i|}, C\right)}{1000}, \text{ where } |S_i| = |S_p|.$$

Analogously, the enrichment statistics for miRNAs targeting their own hosts were calculated, where S_p was defined as the set of host genes, S_t as the set of targets of the intragenic miRNAs of these host genes, and S_i as the set of $|S_p|$ randomly sampled genes (out of the non-redundant set of predicted targets for these miRNAs). The R-package *'q-value'* was used to account for multiple hypothesis testing by controlling the False Discovery Rate (FDR) to be < 10%.

Additional material

Additional file 1: Additional Information on Intragenic miRNAs. The table in additional file 1 contains information on intragenic miRNAs, such as genomic position, name and RefSeq ID of the host gene, and orientation.

Additional file 2: Distribution of intragenic miRNAs. Additional file 2 contains an additional barplot showing the distribution of intronic miRNAs across their hosts' introns, as well as a theoretically expected distribution taking intron frequency and size into consideration. The first figure on page 1 shows a barplot of expected and observed distribution of intragenic miRNAs across their hosts' interactome. The second figure is a repetition of Figure 1b, for better comparison. The second page contains the underlying data in table format.

Additional file 3: Evaluation of Target Prediction Methods. Based on protein and mRNA expression measurements in miRNA transfection experiments, we evaluated the target prediction methods used in this study, as well as prediction agreement as a method of its own. We estimated sensitivity, specificity, and AUC for target prediction methods used and prediction agreement based on the Selbach data [74] for changes in mRNA and protein expression after miRNA overexpression.

Additional file 4: Non Small Cell Lung Cancer. The figure is analogous to Figure 2, for a non small cell lung cancer mRNA expression microarray dataset.

Additional file 5: Full Pathway Information. The table provides all 74 KEGG pathways associated with one or more host genes. For each KEGG pathway KEGG ID, pathway name, p-value and odds ratio for the observed number of host genes, total number of expected genes, total number of observed genes, total number of genes in that pathway, pathway url, host gene names and Entrez-IDs, target coverage and p-value, proportion of targets with conserved target sites within the hosts' pathway, proportion of targets with conserved target sites not within the hosts' pathway, q-value of the difference of these two proportions, and Entrez gene IDs for all predicted targets are provided. The asterisk behind a gene ID indicates a conserved target site for that target. It is important to note, however, that a proportion calculated from these gene IDs may differ from the proportion given, as the gene IDs are based on agreement of two prediction methods, whereas the proportion of conserved targets was calculated on predictions made by targetscan only.

Additional file 6: Additional file 6 **contains correlation data from which Figure** 2 **and additional file** 4 **have been generated**. For both pathways, hosts and their predicted targets as well as correlation and p-value are provided.

Acknowledgements

We would like to thank David Singerman, Staal Vinterbo, Jihoon Kim, Erik Pitzer, and Leo Anthony Celi for all the helpful discussions and their valuable feedback.

Part of this work was included in L.C. Hinske's doctoral thesis at Ludwig-Maximilians-University Munich. This work was funded in part by grant

FAS0703850 from the Komen Foundation. P. Galante was funded in part by grant D43TW007015 from the Fogarty International Center, NIH. L. Ohno-Machado was funded in part by grants R01LM009520 and U54HL108460 from the NIH. This work was furthermore conducted with support from Harvard Catalyst | The Harvard Clinical and Translational Science Center (NIH Award #UL1 RR 025758 and financial contributions from Harvard University and its affiliated academic health care centers). The content is solely the responsibility of the authors and does not necessarily represent the official views of Harvard Catalyst, Harvard University and its affiliated academic health care centers, the National Center for Research Resources, or the National Institutes of Health.

Author details

[1]Department of Anaesthesiology, Clinic of the University of Munich, Marchioninistrasse 15, 81377 Munich, Germany. [2]Division of Biomedical Informatics, University of California San Diego, 9500 Gilman Dr, La Jolla, California 92093, USA. [3]Ludwig Institute for Cancer Research, Hospital Alemão Oswaldo Cruz, Rua João Julião, São Paulo 01323-903, Brazil. [4]Harvard School of Dental Medicine, 188 Longwood Avenue, Boston, Massachusetts 02115, USA. [5]Harvard Catalyst - Laboratory for Innovative Translational Technologies, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, Massachusetts 02115, USA.

Authors' contributions

LCGH contributed by developing the design of the study, performing bioinformatics analyses, and writing the paper.
PAFG contributed by developing the design of the study and performing bioinformatics analyses.
WPK contributed by providing background knowledge and writing the paper.
LOM contributed by developing the design of the study, mentoring the project and writing the paper.
All authors have read and approved the final manuscript.

Received: 19 April 2010 Accepted: 1 October 2010
Published: 1 October 2010

References

1. Lee Y, Kim M, Han J, Yeom K-H, Lee S, Baek SH, Kim VN: **MicroRNA genes are transcribed by RNA polymerase II.** *EMBO J* 2004, **23**:4051-4060.
2. Borchert GM, Lanier W, Davidson BL: **RNA polymerase III transcribes human microRNAs.** *Nat Struct Mol Biol* 2006, **13**:1097-1101.
3. Ozsolak F, Poling LL, Wang Z, Liu H, Liu XS, Roeder RG, Zhang X, Song JS, Fisher DE: **Chromatin structure analyses identify miRNA promoters.** *Genes Dev* 2008, **22**:3172-3183.
4. Baskerville S, Bartel DP: **Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes.** *RNA* 2005, **11**:241-247.
5. Rodriguez A, Griffiths-Jones S, Ashurst JL, Bradley A: **Identification of mammalian microRNA host genes and transcription units.** *Genome Research* 2004, **14**:1902-1910.
6. Gennarino VA, Sardiello M, Avellino R, Meola N, Maselli V, Anand S, Cutillo L, Ballabio A, Banfi S: **MicroRNA target prediction by expression analysis of host genes.** *Genome Research* 2009, **19**:481-490.
7. Kim VN, Han J, Siomi MC: **Biogenesis of small RNAs in animals.** *Nature Reviews Molecular Cell Biology* 2009, **10**:126-139.
8. Morlando M, Ballarino M, Gromak N, Pagano F, Bozzoni I, Proudfoot N: **Primary microRNA transcripts are processed co-transcriptionally.** *Nat Struct Mol Biol* 2008.
9. Kim Y-K, Kim VN: **Processing of intronic microRNAs.** *EMBO J* 2007, **26**:775-783.
10. Obernosterer G, Leuschner PJF, Alenius M, Martinez J: **Post-transcriptional regulation of microRNA expression.** *RNA* 2006, **12**:1161-1167.
11. Blenkiron C, Goldstein LD, Thorne NP, Spiteri I, Chin S-F, Dunning MJ, Barbosa-Morais NL, Teschendorff AE, Green AR, Ellis IO, *et al*: **MicroRNA expression profiling of human breast cancer identifies new markers of tumor subtype.** *Genome Biol* 2007, **8**:R214.
12. Lee EJ, Baek M, Gusev Y, Brackett DJ, Nuovo GJ, Schmittgen TD: **Systematic evaluation of microRNA processing patterns in tissues, cell lines, and tumors.** *RNA* 2008, **14**:35-42.
13. Thomson JM, Newman M, Parker JS, Morin-Kensicki EM, Wright T, Hammond SM: **Extensive post-transcriptional regulation of microRNAs and its implications for cancer.** *Genes Dev* 2006, **20**:2202-2207.
14. Merritt WM, Lin YG, Han LY, Kamat AA, Spannuth WA, Schmandt R, Urbauer D, Pennacchio LA, Cheng J-F, Nick AM, *et al*: **Dicer, Drosha, and outcomes in patients with ovarian cancer.** *N Engl J Med* 2008, **359**:2641-2650.
15. Fujita S, Ito T, Mizutani T, Minoguchi S, Yamamichi N, Sakurai K, Iba H: **miR-21 Gene expression triggered by AP-1 is sustained through a double-negative feedback mechanism.** *Journal of Molecular Biology* 2008, **378**:492-504.
16. Li X, Carthew RW: **A microRNA mediates EGF receptor signaling and promotes photoreceptor differentiation in the Drosophila eye.** *Cell* 2005, **123**:1267-1277.
17. Martinez NJ, Ow MC, Reece-Hoyes JS, Barrasa MI, Ambros VR, Walhout AJM: **Genome-scale spatiotemporal analysis of Caenorhabditis elegans microRNA promoter activity.** *Genome Research* 2008, **18**:2005-2015.
18. Li S-C, Tang P, Lin W-C: **Intronic microRNA: discovery and biological implications.** *DNA and Cell Biology* 2007, **26**:195-207.
19. Barik S: **An intronic microRNA silences genes that are functionally antagonistic to its host gene.** *Nucleic Acids Research* 2008, **36**:5232-5241.
20. Pruitt KD, Tatusova T, Maglott DR: **NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.** *Nucleic Acids Research* 2005, **33**:D501-504.
21. Grzybowska E, Wilczynska A, Siedlecki J: **Regulatory Functions of 3' UTRs.** *Biochemical and Biophysical Research Communications* 2001.
22. Mazumder B, Seshadri V, Fox PL: **Translational control by the 3'-UTR: the ends specify the means.** *Trends in Biochemical Sciences* 2003, **28**:91-98.
23. Mignone F, Gissi C, Liuni S, Pesole G: **Untranslated regions of mRNAs.** *Genome Biol* 2002.
24. Hu Z: **Insight into microRNA regulation by analyzing the characteristics of their targets in humans.** *BMC Genomics* 2009, **10**:594.
25. Ulitsky I, Laurent LC, Shamir R: **Towards computational prediction of microRNA function and activity.** *Nucleic Acids Research* 2010.
26. Divakaran V, Mann DL: **The emerging role of microRNAs in cardiac remodeling and heart failure.** *Circ Res* 2008, **103**:1072-1083.
27. Eisenberg I, Eran A, Nishino I, Moggio M, Lamperti C, Amato AA, Lidov HG, Kang PB, North KN, Mitrani-Rosenbaum S, *et al*: **Distinctive patterns of microRNA expression in primary muscular disorders.** *Proc Natl Acad Sci USA* 2007, **104**:17016-17021.
28. Fiore R, Siegel G, Schratt G: **MicroRNA function in neuronal development, plasticity and disease.** *Biochim Biophys Acta* 2008, **1779**:471-478.
29. Grassmann R, Jeang K-T: **The roles of microRNAs in mammalian virus infection.** *Biochim Biophys Acta* 2008, **1779**:706-711.
30. Hansen T, Olsen L, Lindow M, Jakobsen K, Ullum H: **Brain Expressed microRNAs Implicated in Schizophrenia Etiology.** *PLoS ONE* 2007.
31. Kuhn DE, Nuovo GJ, Martin MM, Malana GE, Pleister AP, Jiang J, Schmittgen TD, Terry AV, Gardiner K, Head E, *et al*: **Human chromosome 21-derived miRNAs are overexpressed in down syndrome brains and hearts.** *Biochemical and Biophysical Research Communications* 2008, **370**:473-477.
32. Ma L, Weinberg RA: **MicroRNAs in malignant progression.** *Cell Cycle* 2008, **7**:570-572.
33. Shi X-B, Tepper CG, deVere White RW: **Cancerous miRNAs and their regulation.** *Cell Cycle* 2008, **7**:1529-1538.
34. Tili E, Michaille J-J, Costinean S, Croce CM: **MicroRNAs, the immune system and rheumatic disease.** *Nature clinical practice Rheumatology* 2008, **4**:534-541.
35. Ambs S, Prueitt RL, Yi M, Hudson RS, Howe TM, Petrocca F, Wallace TA, Liu C-G, Volinia S, Calin GA, *et al*: **Genomic profiling of microRNA and messenger RNA reveals deregulated microRNA expression in prostate cancer.** *Cancer Research* 2008, **68**:6162-6170.
36. Carè A, Catalucci D, Felicetti F, Bonci D, Addario A, Gallo P, Bang M-L, Segnalini P, Gu Y, Dalton ND, *et al*: **MicroRNA-133 controls cardiac hypertrophy.** *Nat Med* 2007, **13**:613-618.
37. ENCODE Project Consortium, Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, *et al*: **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.** *Nature* 2007, **447**:799-816.
38. Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M: **Systematic discovery of regulatory motifs in human promoters**

and 3′ UTRs by comparison of several mammals. *Nature* 2005, **434**:338-345.
39. Golan D, Levy C, Friedman B, Shomron N: **Biased hosting of intronic microRNA genes.** *Bioinformatics* 2010.
40. Castillo-Davis CI, Mekhedov SL, Hartl DL, Koonin EV, Kondrashov FA: **Selection for short introns in highly expressed genes.** *Nat Genet* 2002, **31**:415-418.
41. Pickering BM, Willis AE: **The implications of structured 5′ untranslated regions on translation and disease.** *Seminars in Cell & Developmental Biology* 2005, **16**:39-47.
42. Niwa R, Zhou F, Li C, Slack FJ: **The expression of the Alzheimer's amyloid precursor protein-like gene is regulated by developmental timing microRNAs and their targets in Caenorhabditis elegans.** *Dev Biol* 2008, **315**:418-425.
43. Sun Y, Bai Y, Zhang F, Wang Y, Guo Y, Guo L: **miR-126 inhibits non-small cell lung cancer cells proliferation by targeting EGFL7.** *Biochemical and Biophysical Research Communications* 2010, **391**:1483-1489.
44. Nikolic I, Plate K-H, Schmidt MH: **EGFL7 meets miRNA-126: an angiogenesis alliance.** *J Angiogenes Res* 2010, **2**:9.
45. Muralidhar B, Goldstein LD, Ng G, Winder DM, Palmer RD, Gooding EL, Barbosa-Morais NL, Mukherjee G, Thorne NP, Roberts I, *et al*: **Global microRNA profiles in cervical squamous cell carcinoma depend on Drosha expression levels.** *J Pathol* 2007, **212**:368-377.
46. Tavazoie SF, Alarcón C, Oskarsson T, Padua D, Wang Q, Bos PD, Gerald WL, Massagué J: **Endogenous human microRNAs that suppress breast cancer metastasis.** *Nature* 2008, **451**:147-152.
47. Agirre X, Jiménez-Velasco A, San José-Enériz E, Garate L, Bandrés E, Cordeu L, Aparicio O, Saez B, Navarro G, Vilas-Zornoza A, *et al*: **Down-regulation of hsa-miR-10a in chronic myeloid leukemia CD34+ cells increases USF2-mediated cell growth.** *Mol Cancer Res* 2008, **6**:1830-1840.
48. Le Y, Xu L, Lu J, Fang J, Nardi V, Chai L, Silberstein LE: **FAK silencing inhibits leukemogenesis in BCR/ABL-transformed hematopoietic cells.** *Am J Hematol* 2009, **84**:273-278.
49. Missiaglia E, Dalai I, Barbi S, Beghelli S, Falconi M, della Peruta M, Piemonti L, Capurso G, Di Florio A, delle Fave G, *et al*: **Pancreatic endocrine tumors: expression profiling evidences a role for AKT-mTOR pathway.** *J Clin Oncol* 2010, **28**:245-255.
50. Kano M, Seki N, Kikkawa N, Fujimura L, Hoshino I, Akutsu Y, Chiyomaru T, Enokida H, Nakagawa M, Matsubara H: **miR-145, miR-133a and miR-133b: Tumor suppressive miRNAs target FSCN1 in esophageal squamous cell carcinoma.** *Int J Cancer* 2010.
51. Boormans JL, Hermans KG, van Leenders GJLH, Trapman J, Verhagen PCMS: **An activating mutation in AKT1 in human prostate cancer.** *Int J Cancer* 2008, **123**:2725-2726.
52. Forgacs E, Biesterveld EJ, Sekido Y, Fong K, Muneer S, Wistuba II, Milchgrub S, Brezinschek R, Virmani A, Gazdar AF, Minna JD: **Mutation analysis of the PTEN/MMAC1 gene in lung cancer.** *Oncogene* 1998, **17**:1557-1565.
53. Bellacosa A, de Feo D, Godwin AK, Bell DW, Cheng JQ, Altomare DA, Wan M, Dubeau L, Scambia G, Masciullo V, *et al*: **Molecular alterations of the AKT2 oncogene in ovarian and breast carcinomas.** *Int J Cancer* 1995, **64**:280-285.
54. Philp AJ, Campbell IG, Leet C, Vincan E, Rockman SP, Whitehead RH, Thomas RJ, Phillips WA: **The phosphatidylinositol 3′-kinase p85alpha gene is an oncogene in human ovarian and colon tumors.** *Cancer Research* 2001, **61**:7426-7429.
55. Sun M, Paciga JE, Feldman RI, Yuan Z, Coppola D, Lu YY, Shelley SA, Nicosia SV, Cheng JQ: **Phosphatidylinositol-3-OH Kinase (PI3K)/AKT2, activated in breast cancer, regulates and is induced by estrogen receptor alpha (ERalpha) via interaction between ERalpha and PI3K.** *Cancer Research* 2001, **61**:5985-5991.
56. Noske A, Kaszubiak A, Weichert W, Sers C, Niesporek S, Koch I, Schaefer B, Sehouli J, Dietel M, Lage H, Denkert C: **Specific inhibition of AKT2 by RNA interference results in reduction of ovarian cancer cell proliferation: increased expression of AKT in advanced ovarian cancer.** *Cancer Lett* 2007, **246**:190-200.
57. Maroulakou IG, Oemler W, Naber SP, Tsichlis PN: **Akt1 ablation inhibits, whereas Akt2 ablation accelerates, the development of mammary adenocarcinomas in mouse mammary tumor virus (MMTV)-ErbB2/neu and MMTV-polyoma middle T transgenic mice.** *Cancer Research* 2007, **67**:167-177.
58. Griffiths-Jones S, Saini HK, Van Dongen S, Enright AJ: **miRBase: tools for microRNA genomics.** *Nucleic Acids Research* 2008, **36**:D154-158.
59. Griffiths-Jones S, Grocock RJ, Van Dongen S, Bateman A, Enright AJ: **miRBase: microRNA sequences, targets and gene nomenclature.** *Nucleic Acids Research* 2006, **34**:D140-144.
60. Griffiths-Jones S: **miRBase: the microRNA sequence database.** *Methods Mol Biol* 2006, **342**:129-138.
61. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25**:25-29.
62. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Research* 2003, **13**:2498-2504.
63. Maere S, Heymans K, Kuiper M: **BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks.** *Bioinformatics* 2005, **21**:3448-3449.
64. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, *et al*: **Bioconductor: open software development for computational biology and bioinformatics.** *Genome Biol* 2004, **5**:R80.
65. Gentleman R, Carey V, Dudoit S, Ellis B, Gautier L: **The Bioconductor Project.** *bepresscom* 2003.
66. Lewis BP, Shih I-h, Jones-Rhoades MW, Bartel DP, Burge CB: **Prediction of mammalian microRNA targets.** *Cell* 2003, **115**:787-798.
67. Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E: **The role of site accessibility in microRNA target recognition.** *Nat Genet* 2007, **39**:1278-1284.
68. Wang X, El Naqa IM: **Prediction of both conserved and nonconserved microRNA targets in animals.** *Bioinformatics* 2008, **24**:325-332.
69. Wang X: **miRDB: a microRNA target prediction and functional annotation database with a wiki interface.** *RNA* 2008, **14**:1012-1017.
70. John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS: **Human MicroRNA targets.** *PLoS Biol* 2004, **2**:e363.
71. Miranda K, Huynh T, Tay Y, Ang Y, Tam W, Thomson A, Lim B, Rigoutsos I: **A Pattern-Based Method for the Identification of MicroRNA Binding Sites and Their Corresponding Heteroduplexes.** *Cell* 2006, **126**:1203-1217.
72. Krek A, Grün D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, MacMenamin P, da Piedade I, Gunsalus KC, Stoffel M, Rajewsky N: **Combinatorial microRNA target predictions.** *Nat Genet* 2005, **37**:495-500.
73. Sethupathy P, Corda B, Hatzigeorgiou AG: **TarBase: A comprehensive database of experimentally supported animal microRNA targets.** *RNA* 2006, **12**:192-197.
74. Selbach M, Schwanhäusser B, Thierfelder N, Fang Z, Khanin R, Rajewsky N: **Widespread changes in protein synthesis induced by microRNAs.** *Nature* 2008, **455**:58-63.

doi:10.1186/1471-2164-11-533
Cite this article as: Hinske *et al.*: A potential role for intragenic miRNAs on their hosts' interactome. *BMC Genomics* 2010 **11**:533.

DATABASE
The Journal of Biological Databases and Curation

Database, 2014, 1–9
doi: 10.1093/database/bau099
Database tool

Database tool

miRIAD—integrating microRNA inter- and intragenic data

Ludwig Christian Hinske[1,*,†], Gustavo S. França[2,3,†], Hugo A. M. Torres[2], Daniel T. Ohara[2], Camila M. Lopes-Ramos[2], Jens Heyn[1], Luiz F. L. Reis[2], Lucila Ohno-Machado[4], Simone Kreth[1] and Pedro A. F. Galante[2,*]

[1]Clinic of Anaesthesiology, Clinic of the University of Munich, Munich, Germany

[2]Centro de Oncologia Molecular, Hospital Sírio-Libanês, São Paulo, SP 01308-060, Brazil

[3]Departamento de Bioquímica, Instituto de Química, Universidade de São Paulo, São Paulo, Brazil

[4]Division of Medial Informatics, University of California San Diego, La Jolla, CA 93093-0505, USA

*Corresponding author: Tel: +55 11 3155 3704; Fax: +55 11 3155 4220; Email: pgalante@mochsl.org.br

Correspondence may also be addressed to Ludwig Christian Hinske. Tel: +49 89 4400 73410; Fax: +49 89 4400 78886; Email: christian.hinske@med.uni-muenchen.de

†These authors contributed equally to this work.

Citation details: Hinske,L.C., França,G.S., Torres,H.A.M., *et al*. miRIAD—integrating microRNA inter- and intragenic data. *Database* (2014) Vol. 2014: article ID bau099; doi:10.1093/database/bau099

Received 9 January 2014; Revised 4 September 2014; Accepted 9 September 2014

Abstract

MicroRNAs (miRNAs) are a class of small (∼22 nucleotides) non-coding RNAs that post-transcriptionally regulate gene expression by interacting with target mRNAs. A majority of miRNAs is located within intronic or exonic regions of protein-coding genes (host genes), and increasing evidence suggests a functional relationship between these miRNAs and their host genes. Here, we introduce miRIAD, a web-service to facilitate the analysis of genomic and structural features of intragenic miRNAs and their host genes for five species (human, rhesus monkey, mouse, chicken and opossum). miRIAD contains the genomic classification of all miRNAs (inter- and intragenic), as well as classification of all protein-coding genes into host or non-host genes (depending on whether they contain an intragenic miRNA or not). We collected and processed public data from several sources to provide a clear visualization of relevant knowledge related to intragenic miRNAs, such as host gene function, genomic context, names of and references to intragenic miRNAs, miRNA binding sites, clusters of intragenic miRNAs, miRNA and host gene expression across different tissues and expression correlation for intragenic miRNAs and their host genes. Protein–protein interaction data are also presented for functional network analysis of host genes. In summary, miRIAD was designed to help the research community to explore, in a user-friendly environment, intragenic miRNAs, their host genes and functional annotations with minimal effort, facilitating hypothesis generation and *in-silico* validations.

Database URL: http://www.miriad-database.org

Introduction

Amongst regulatory mechanisms of gene expression in eukaryotes, microRNAs (miRNAs) have established a central role in the past two decades (1). These 22-nt short single-stranded RNA molecules guide the RNA-induced silencing complex to modulate the expression of target mRNAs (2). MicroRNA binding sites are most likely recognized by nucleotide sequences in the 3'-untranslated regions (3′-UTR) of target mRNAs. Binding of the miRNA–protein complexes to their targets results in either degradation or translational inhibition of the mRNA transcripts (2).

For humans, ~1900 miRNA genes have been identified (3), and more than half are located within genomic regions containing protein-coding genes (4–6). Hence, miRNA genes can be classified as either inter- or intragenic, and the latter sub-classified as intronic or exonic (4, 5). A substantial number of these intragenic miRNAs are co-transcribed, and consequently co-regulated with their host genes (4, 5, 7). Recent evidence suggests a functional linkage between intragenic miRNAs and their hosts on multiple levels, including direct and indirect interaction (8–10).

Despite the importance of these intragenic miRNAs, their exploration can be daunting, as much of the necessary information is not readily available and requires manual integration from multiple data sources (6, 11, 12). Although other databases exist that provide information related to intra- and intergenic miRNAs (12–15), some tools don't appear to be frequently updated (14), contain only an elementary set of information related to intragenic miRNAs and their host genes (13, 15) and/or their usage is complex and requires in-depth bioinformatics skills (12).

In the current manuscript, we present miRIAD, a web-service designed to examine intragenic miRNAs, their host genes and their functional annotations with a streamlined graphical data representation and an efficient information query system. miRIAD provides information regarding genomic context, gene function, gene interaction, miRNA targets and gene expression for five species, including human and mouse. miRIAD is publicly available at http://www.miriad-database.org.

Materials and Methods

Database architecture and raw data

Because miRIAD integrates a large set of data, processed information is stored in a MySQL relational database. Supplementary Figure S1 provides an overview of the miRIAD database schema, its tables and their relations. To date, miRIAD consists of 60 tables in total, comprising 12 tables for each of the five species (human, rhesus monkey, mouse, opossum and chicken), containing ~10 million records of integrated information.

To construct miRIAD, we used several sets of publicly available data. The reference genomes (human genome sequence—GRCh37/hg19; rhesus genome—rheMac3; mouse genome—mm10/GRCm38; opossum genome—MonDom5; chicken genome—galGal4) were downloaded from UCSC Genome Browser (http://genome.ucsc.edu). The transcriptome sets were downloaded from the RefSeq project (http://www.ncbi.nih.gov/refseq) for all species. MicroRNA genomic coordinates, seed sequences and family information were retrieved from miRBase (http://www.mirbase.org/, release #20). Protein–protein interaction data were acquired from the Human Protein Reference Database (HPRD, downloaded from NCBI http://www.ncbi.nih.gov) and from EMBL's STRING database (http://string-db.org/). Gene expression data were obtained from Brawand *et al.* (16) (coding genes) and Meunier *et al.* (17) (miRNAs).

Host gene and miRNA information

All known genes were classified either as host or non-host based on the presence of overlapping miRNAs for each species. This classification and additional information regarding known genes were stored in three tables (GeneInformation, GeneRegions and GeneSynonyms), as shown in Supplementary Figure S1.

All miRNA genes were classified either as intra- or intergenic, based on their genomic localization. The 'MirnaInformation' table contains the official name, genomic coordinates of the stem loop sequence and, if applicable, the host gene to which the miRNA is related. In case of multiple genes, the host gene assigned was the one on the same strand as the miRNA. If intronic, the intron number and the region length between the miRNA coordinates and the next exon upstream were calculated and stored.

miRNA target prediction

miRIAD contains all conserved target sites within 3'UTRs from TargetScan (http://www.targetscan.org/, release #6.2) for human and mouse. In brief, TargetScan defines miRNA targets by searching, within 3' UTR regions, for 8mer (exact match) and 7mer sites that match the seed region (position 2–7) of mature miRNAs. Information regarding interspecies conservation and match/mismatch profile are also used to define the final set of conserved targets (for further information, see http://www.targetscan.org/). miRIAD contains a total of 1141 miRNAs binding to 466569 mRNA targets from 14867 known protein coding genes for human. Target prediction information for human and mouse were directly downloaded from the TargetScan homepage (file Conserved_Site_Context_Scores.txt, release #62) and calculated for rhesus monkey, opossum and chicken miRNAs using the TargetScan tool kit, applied to all miRNAs and the 3'UTRs from these organisms.

Gene and miRNA expression

To obtain expression for protein-coding genes, data from Brawand *et al.* (16) were downloaded from GEO (GSE30352) and aligned to the genome of each species using TopHat (version 2.0.8b) with default parameters (18). Normalized gene expression values for six tissues (brain, cerebellum, heart, liver, kidney and testis) from all species were computed by means of FPKM (19) with Cufflinks [version 2.2.1; (20)] using transcript annotations from Ensembl (version 71). To determine miRNA expression available for five tissues (brain, cerebellum, heart, kidney and testis) from all species, data from Meunier *et al.* (17) were downloaded from GEO (GSE40499) and reads were aligned to each genome with Bowtie version 1.0.0 using the following parameters: -m 5 -v 0 -a –best –strata. Only exact matches were considered, and reads aligned to >5 different loci were discarded. The 3' adaptors were removed using a sequential trimming strategy (21). Reads totally overlapping to mature miRNA coordinates annotated from miRBase (release 20) were counted and normalized for each species with EdgeR package version 2.6.12 (22). Host gene and intragenic miRNA expression correlations were calculated by Spearman's rank correlation using the normalized values (FPKM and CPM (counts per million) for coding genes and miRNAs, respectively).

Results

Database overview

Figure 1 summarizes the main features, data sets and how information is presented in the miRIAD web tool. Most of miRIAD data related to intragenic miRNAs and their host genes is summarized in Table 1. To provide a useful platform, miRIAD integrates all known protein-coding genes (~22k genes on average, for all five species), all known miRNAs (~900 on average, for all five species), miRNAs targets, validated and predicted protein–protein interactions and expression data for miRNAs and coding genes across five and six tissues, respectively. miRIAD classifies all miRNAs as intragenic or intergenic. It contains a total of 1072 (57%) for human; 167 (29%) for rhesus; 745 (63%) for mouse; 179 (40%) for opossum; 299 (52%) for chicken, additionally specifying whether or not they are transcribed in the same orientation as that of their host genes (84, 54, 87, 92 and 76% of intragenic miRNAs for human, rhesus, mouse, opossum and chicken, respectively). It is worth mentioning that some of the discrepancies between these percentages are likely due to the incompleteness of miRNA and gene annotation for individual species. As we can observe for human and mouse, which have the most complete annotated sets of coding genes and miRNAs, the values are quite similar. Additional complex information is also provided, such as the visualization of intragenic miRNAs within their host genes and positioning along the isoforms, expression correlation between intragenic miRNAs and their host genes, intragenic miRNAs binding to their own host genes and intragenic miRNAs binding to genes that are directly interacting with their host genes. These data are necessary in the identification and evaluation of putative negative or positive feedback mechanisms between miRNAs and host genes, (5, 23–25) and can offer a starting point for future analyses to reveal novel regulatory pathways.

miRIAD query system

The miRIAD query system was developed and optimized to be fast, intuitive and functional. It lets the user search for several terms, such as miRNA symbol, gene name (Official Symbol, Ensembl ID, Entrez ID, HGNC ID or Gene Synonyms) and gene annotation keywords (e.g. 'oncogene', 'kinase', etc.). Searching for miRNAs follows the same principles as those used for coding genes, allowing for non-exact inputs (according to miRNA official nomenclature). It is also possible to query for multiple genes or miRNAs at once. The query system works in the same way for all five species.

The output for each searched term is a list of query matches organized by relevance, containing basic gene information for rapid inspection and selection. Names of host genes and intragenic miRNAs are readily identified by a particular tag (see web page for details). Moreover, non-host genes and intergenic miRNAs are also shown, because they may have indirect associations to intragenic miRNAs or host genes and are therefore also important. By clicking on a gene name, the user can access more

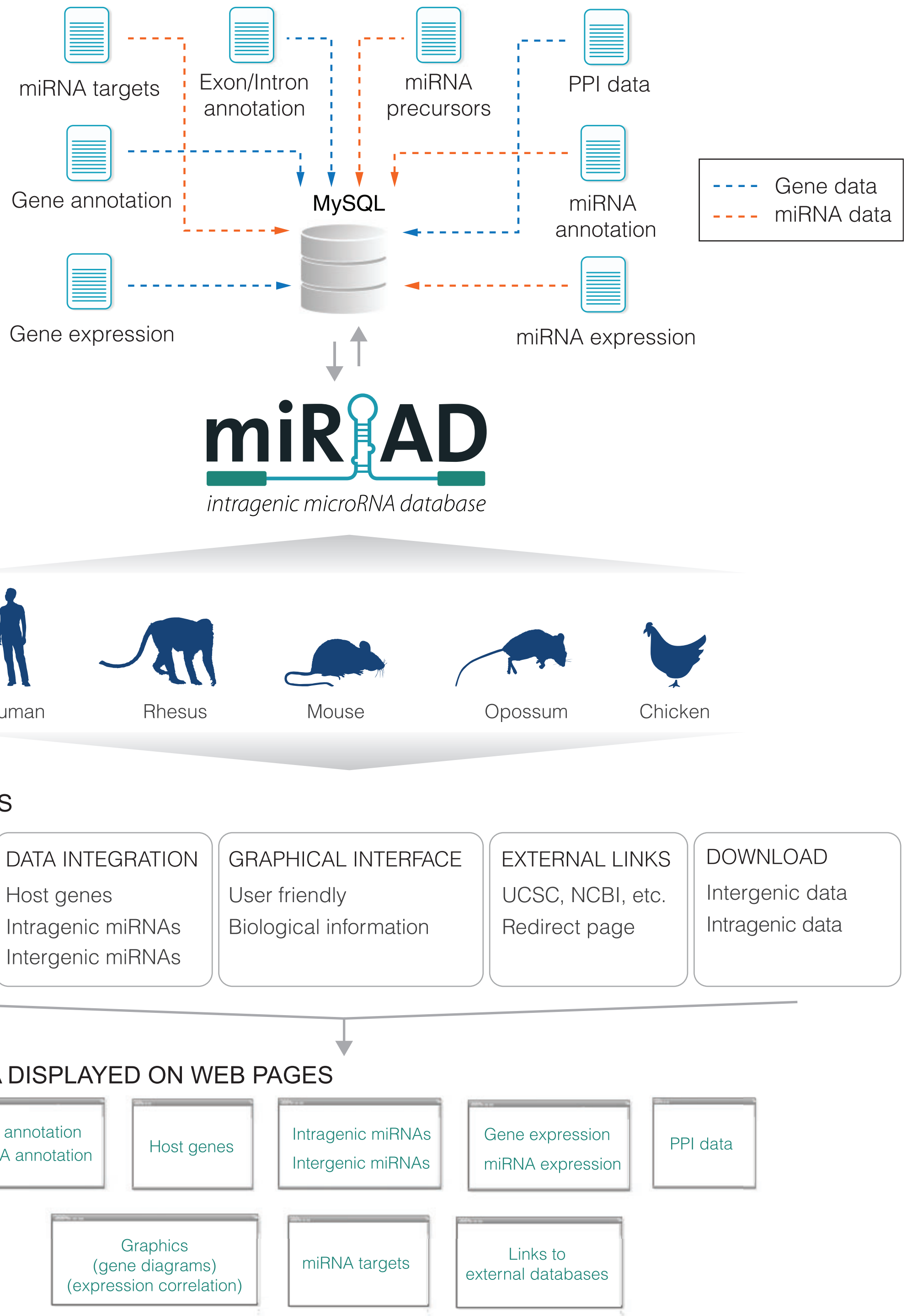


Figure 1. Overview of the miRIAD platform. Schematic representation of the main data presented in the web tool and how they are integrated and displayed. Blue arrows denote data related to protein-coding genes and orange arrows indicate data related to miRNAs. PPI: protein–protein interaction.

detailed information about any known coding or miRNA gene.

Exploring host genes

In the recent past, it has become clear that functional aspects of intragenic miRNAs have to be viewed in the context of their host genes (5, 7, 23, 24, 26). Therefore, information about all known protein-coding genes has been integrated into miRIAD to allow contextual search. For each protein-coding gene, miRIAD provides a 'Summary' section showing annotation data, such as official gene symbol, full gene name and gene name aliases, gene type and a gene function summary when publicly available.

Table 1. Summary of main miRIAD data

Data class	Human	Rhesus	Mouse	Opossum	Chicken
Known protein-coding genes	20 530	22 553	29 664	20 550	16 953
Known miRNA precursors	1871	582	1181	443	573
Intragenic miRNAs	1072	167	745	179	299
Intergenic miRNAs	799	415	435	264	272
Host genes	930	141	613	143	273
Sense miRNAs in respect to host orientation	902	90	645	145	90
Antisense miRNAs in respect to host orientation	170	77	95	12	28
Expressed coding genes	18 442	8112	19 029	12 079	11 278
Expressed miRNAs	1111	475	784	405	465

Moreover, information regarding the genomic context, including the genomic position, transcription 'start' and 'end' and transcription orientation, is provided, as well as a graphical representation of the exon–intron structure of transcripts (Figure 2). If applicable, miRIAD presents miRNA name, genomic region (intronic/exonic), the intron/exon number where they are inserted, the distance to the closest upstream exon and transcriptional orientation, sense (miRNA and host in the same transcriptional orientation), or antisense (in opposite orientation). To facilitate the generation and evaluation of research hypotheses, expression data (based on RNA-Seq) of mRNAs across six tissues (brain, cerebellum, heart, kidney, liver and testis) as well as expression correlation between host genes and their intragenic miRNAs were included. All miRNAs potentially binding to a target gene are displayed under 'miRNA binding sites'. Finally, the last section shows all known protein–protein interaction data for each gene. Cases in which interaction partners of a given host gene are targeted by its intragenic miRNA are explicitly shown. This kind of information is noteworthy because it can reveal unusual regulatory loops and may support findings or suggest future investigations. All these information are exemplified for the oncogene ERBB2 containing mir-4728 (Figure 2).

The gene section also provides links to external databases, such as NCBI Gene (http://www.ncbi.nlm.nih.gov/gene), UCSC Genome Browser (http://genome.ucsc.edu/), Ensembl (http://www.ensembl.org/), KEGG (http://www.genome.jp/kegg/) and Targetscan (http://www.targetscan.org/). Most of these links are context-sensitive, easily redirecting the user to the gene of interest on the web page containing complementary data.

Intragenic miRNAs

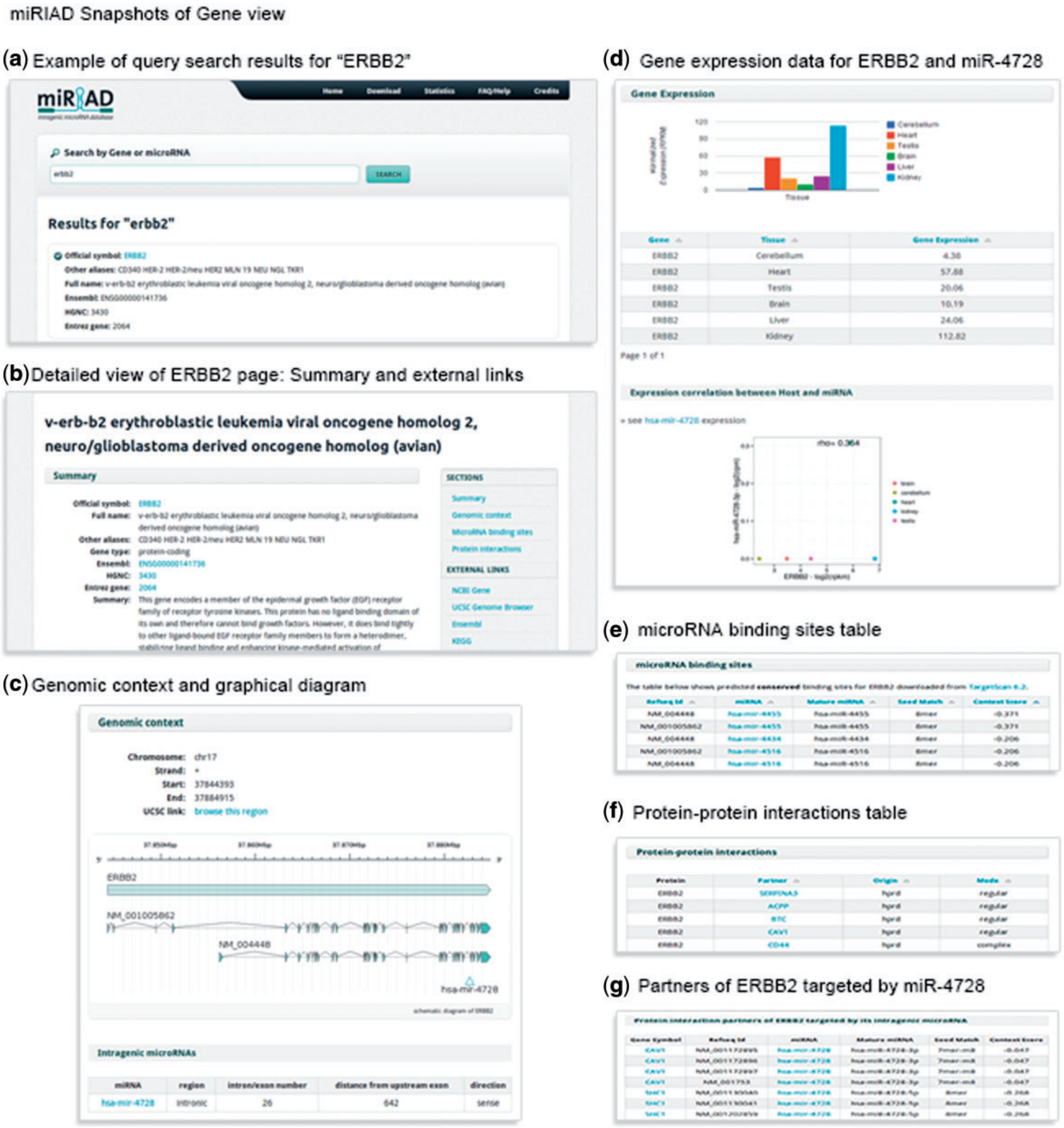
Intragenic miRNAs are the main focus of our web tool, even though we present information for all known miRNAs and protein-coding genes. For each pre-miRNA, miRIAD provides a 'Summary' section with the official miRNA symbol, its full name, miRBase ID, target genes and the genomic context where each miRNA is mapped (Figure 3). For intragenic miRNAs, information about their intragenic position and location along the host genes are depicted by a graphical representation (Figure 3). Cases where an intragenic miRNA potentially targets its own host are highlighted for fast identification. Similar to the presentation of information about protein-coding genes, there are also expression data (based on RNAseq) for six tissues (brain, cerebellum, heart, kidney, liver and testis) and an expression correlation between intragenic miRNAs and their host genes. A set of context-sensitive links to external databases in the top right corner to access complementary information (miRBase, miRDB, Targetscan, mirgen, Magia, miRWalk and miRò) are also presented.

Figure 3 exemplifies the use of this information for mir-483 and its host IGF2. IGF2 produces the insulin-like growth factor 2, an essencial protein for growth and development of the fetus and it is upregulated in several malignancies (27). According to our data, the expression of IGF2 and miR-483-5p are highly correlated (rho = 0.7). Accordingly, a recent report has uncovered a positive feedback between IGF2 and its intragenic mir-483, where the mature miR-483-5p molecule binds to the 5'UTR of IGF2 mRNA, promoting IGF2 transcription by facilitating the association of the helicase DHX9 (24).

Using miRIAD to explore a set of genes

In the following paragraph, we briefly illustrate how miRIAD can be used to explore a gene or a set of genes. Recently, da Cunha *et al.* (28) defined the set of all human genes coding for cell surface proteins (called surfaceome genes). These genes can be considered as potential targets for diagnostic and therapeutic interventions (28, 29).

The set of 3702 human surfaceome genes was retrieved from (28, 29) and submitted to miRIAD to initially be classified as host or non-host genes. In total, 119 surfaceome



miRIAD Snapshots of Gene view

(a) Example of query search results for "ERBB2"

Results for "erbb2"

Official symbol: ERBB2
Other aliases: CD340 HER-2 HER-2/neu HER2 MLN 19 NEU NGL TKR1
Full name: v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian)
Ensembl: ENSG00000141736
HGNC: 3430
Entrez gene: 2064

(b) Detailed view of ERBB2 page: Summary and external links

v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian)

Summary

Official symbol: ERBB2
Full name: v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian)
Other aliases: CD340 HER-2 HER-2/neu HER2 MLN 19 NEU NGL TKR1
Gene type: protein-coding
Ensembl: ENSG00000141736
HGNC: 3430
Entrez gene: 2064
Summary: This gene encodes a member of the epidermal growth factor (EGF) receptor family of receptor tyrosine kinases. This protein has no ligand binding domain of its own and therefore cannot bind growth factors. However, it does bind tightly to other ligand-bound EGF receptor family members to form a heterodimer, stabilizing ligand binding and enhancing kinase-mediated activation of

(c) Genomic context and graphical diagram

Genomic context

Chromosome: chr17
Strand: +
Start: 37844393
End: 37884915
UCSC link: browse this region

Intragenic microRNAs

miRNA	region	intron/exon number	distance from upstream exon	direction
hsa-mir-4728	intronic	26	642	sense

(d) Gene expression data for ERBB2 and miR-4728

Gene Expression

Gene	Tissue	Gene Expression
ERBB2	Cerebellum	4.38
ERBB2	Heart	57.88
ERBB2	Testis	20.06
ERBB2	Brain	10.19
ERBB2	Liver	24.06
ERBB2	Kidney	112.82

Page 1 of 1

Expression correlation between Host and miRNA

» see hsa-mir-4728 expression

(e) microRNA binding sites table

microRNA binding sites

The table below shows predicted **conserved** binding sites for ERBB2 downloaded from TargetScan 6.2.

RefSeq Id	miRNA	Mature miRNA	Seed Match	Context Score
NM_004448	hsa-mir-4455	hsa-miR-4455	8mer	-0.371
NM_001005862	hsa-mir-4455	hsa-miR-4455	8mer	-0.371
NM_004448	hsa-mir-4434	hsa-miR-4434	8mer	-0.206
NM_001005862	hsa-mir-4516	hsa-miR-4516	8mer	-0.206
NM_004448	hsa-mir-4516	hsa-miR-4516	8mer	-0.206

(f) Protein-protein interactions table

Protein-protein interactions

Protein	Partner	Origin	Mode
ERBB2	SERPINA3	hprd	regular
ERBB2	ACPP	hprd	regular
ERBB2	BTC	hprd	regular
ERBB2	CAV1	hprd	regular
ERBB2	CD44	hprd	complex

(g) Partners of ERBB2 targeted by miR-4728

Protein interaction partners of ERBB2 targeted by its intragenic microRNA

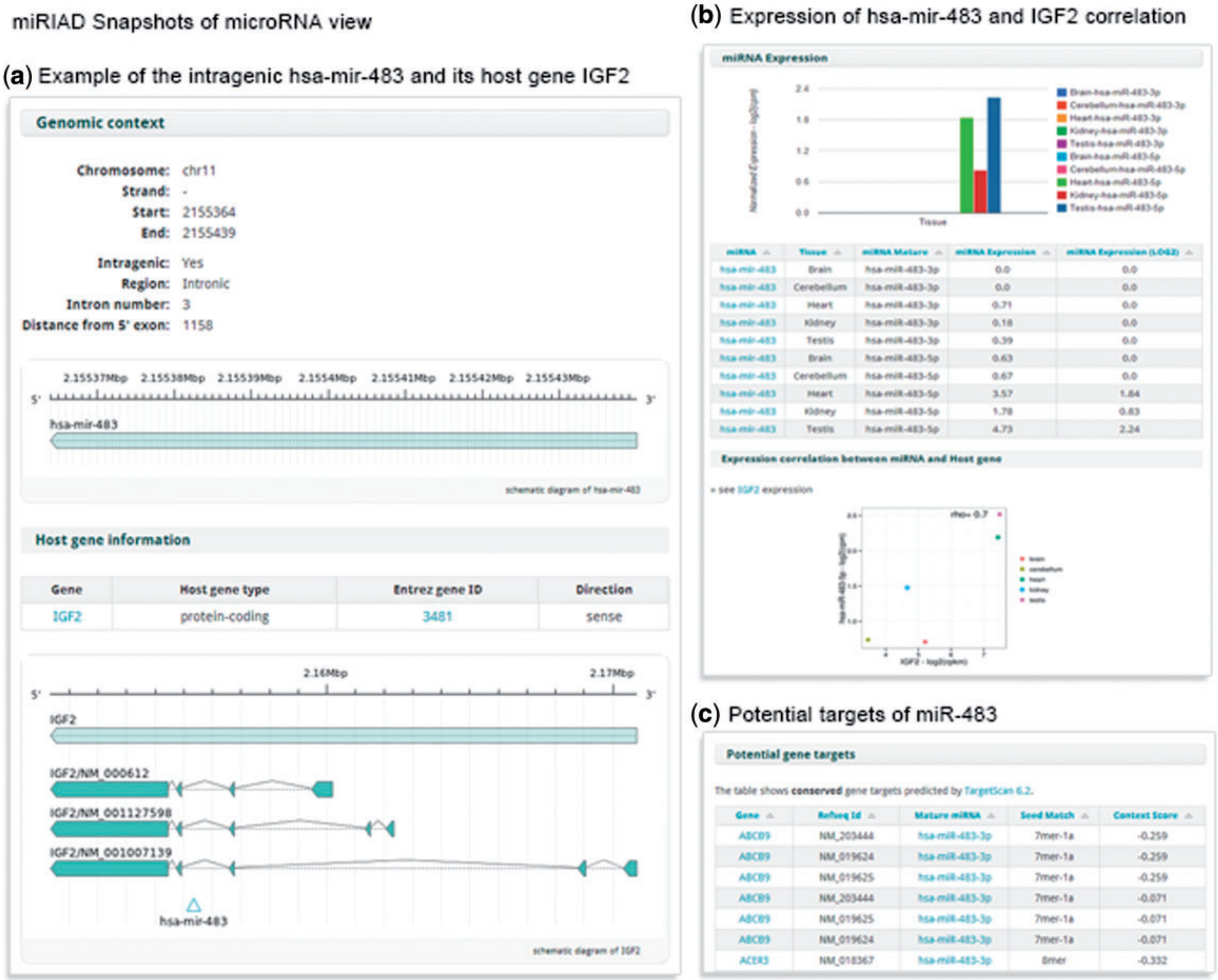
Gene Symbol	RefSeq Id	miRNA	Mature miRNA	Seed Match	Context Score
CAV1	NM_001172895	hsa-mir-4728	hsa-miR-4728-3p	7mer-m8	-0.047
CAV1	NM_001172896	hsa-mir-4728	hsa-miR-4728-3p	7mer-m8	-0.047
CAV1	NM_001172897	hsa-mir-4728	hsa-miR-4728-3p	7mer-m8	-0.047
CAV1	NM_001753	hsa-mir-4728	hsa-miR-4728-3p	7mer-m8	-0.047
SHC1	NM_001130040	hsa-mir-4728	hsa-miR-4728-5p	8mer	-0.268
SHC1	NM_001130041	hsa-mir-4728	hsa-miR-4728-5p	8mer	-0.268
SHC1	NM_001202859	hsa-mir-4728	hsa-miR-4728-5p	8mer	-0.268

Figure 2. A summary of the main information presented in miRIAD for the coding gene ERBB2 and its intragenic mir-4728.

genes are host genes for 150 intragenic miRNAs. Interestingly, most of these miRNAs (87.3%) are transcribed on the same orientation of their host genes, suggesting possible co-transcription (5). 140 of these intragenic miRNAs are actually inserted within intronic regions of surfaceome genes.

Next, we examined two genes in more detail. We selected the genes containing the largest number of intronic miRNAs, CLCN5 and HTR2C. In respect to CLCN5, mutations in its sequence have been proven to be associated with diseases of renal tubules, resulting in chronic renal failure (30). This gene has eight intronic miRNAs, and surprisingly, some of their transcripts may be targeted by their intronic miR-502 (see miRIAD).

It is also striking that this host gene has isoforms starting transcription upstream of the miRNAs, which possibly could prevent co-expression between a CLCN5 transcript and those intronic miRNAs in some tissues or pathologies. Analysis of the expression data suggests co-expression or at least co-regulation between CLCN5 and its intronic



miRIAD Snapshots of microRNA view

(a) Example of the intragenic hsa-mir-483 and its host gene IGF2

Genomic context

Chromosome:	chr11
Strand:	-
Start:	2155364
End:	2155439
Intragenic:	Yes
Region:	Intronic
Intron number:	3
Distance from 5' exon:	1158

Host gene information

Gene	Host gene type	Entrez gene ID	Direction
IGF2	protein-coding	3481	sense

(b) Expression of hsa-mir-483 and IGF2 correlation

miRNA Expression

miRNA	Tissue	miRNA Mature	miRNA Expression	miRNA Expression (LOG2)
hsa-mir-483	Brain	hsa-miR-483-3p	0.0	0.0
hsa-mir-483	Cerebellum	hsa-miR-483-3p	0.0	0.0
hsa-mir-483	Heart	hsa-miR-483-3p	0.71	0.0
hsa-mir-483	Kidney	hsa-miR-483-3p	0.18	0.0
hsa-mir-483	Testis	hsa-miR-483-3p	0.39	0.0
hsa-mir-483	Brain	hsa-miR-483-5p	0.63	0.0
hsa-mir-483	Cerebellum	hsa-miR-483-5p	0.67	0.0
hsa-mir-483	Heart	hsa-miR-483-5p	3.57	1.84
hsa-mir-483	Kidney	hsa-miR-483-5p	1.78	0.83
hsa-mir-483	Testis	hsa-miR-483-5p	4.73	2.24

Expression correlation between miRNA and Host gene

» see IGF2 expression

(c) Potential targets of miR-483

Potential gene targets

The table shows **conserved** gene targets predicted by TargetScan 6.2.

Gene	Refseq Id	Mature miRNA	Seed Match	Context Score
ABCB9	NM_203444	hsa-miR-483-3p	7mer-1a	-0.259
ABCB9	NM_019624	hsa-miR-483-3p	7mer-1a	-0.259
ABCB9	NM_019625	hsa-miR-483-3p	7mer-1a	-0.259
ABCB9	NM_203444	hsa-miR-483-3p	7mer-1a	-0.071
ABCB9	NM_019625	hsa-miR-483-3p	7mer-1a	-0.071
ABCB9	NM_019624	hsa-miR-483-3p	7mer-1a	-0.071
ACER3	NM_018367	hsa-miR-483-3p	8mer	-0.332

Figure 3. A summary of the main information presented in miRIAD for the intragenic mir-483 and its host gene IGF2.

miRNAs. CLCN5, as well as its intronic miRNAs are highly expressed in kidney. The expression correlations are high (rho > 0.7, Spearman's rank correlation) for most of the intragenic miRNAs. The functional relationships between CLCN5 and its intronic miRNAs have not been explored yet, though, and deserve further exploration. Suggesting a conserved regulation, a similar pattern is found for Clcn5 gene in mouse, which has five annotated intragenic miRNAs and also a high expression correlation between miRNAs and the host gene.

The second gene, HTR2C, encodes the 2C subtype of serotonin receptor and contains six intronic miRNAs (Figure 4). Similar to CLCN5, host and miRNAs have the same transcriptional orientation (see miRIAD web page for details). As reported by (10), up-regulation of HTR2C is involved in adipocyte differentiation by repressing the KLF5 gene through the expression of miR-448, a miRNA located in the fourth intron of HTR2C. Interestingly, our expression data show a highly positive (rho > 8.5, Spearman's rank correlation) correlation between miRNAs and host gene, being expressed specifically in cerebellum and brain (Figure 4). The patterns of co-expression are also conserved in opossum and mouse. Moreover, HTR2C is tightly involved in important neuropsychiatric disorders (31); thus, the functional consequences of the concomitant expression of HTR2C and its intragenic miRNAs is tempting to investigate.

miRIAD helped us to identify two interesting gene loci involved in complex human diseases with this quick and unpretentious gene survey. We speculate that many other crucial host/miRNA regulatory mechanisms could be revealed by taking advantage of using miRIAD for initial and/or advanced exploration.

Discussion and conclusion

As the number of newly discovered miRNAs is constantly increasing, our understanding of the importance and the

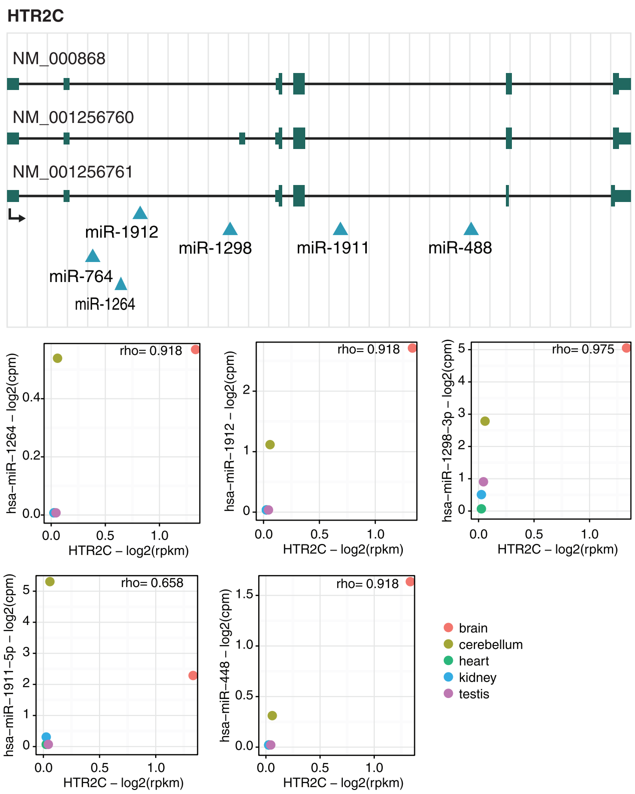


Figure 4. HTR2C gene locus. Genomic mapping of HTR2C transcripts (NM_000868, NM_001256760 and NM_001256761) and their six intragenic miRNAs (miR-1912, miR-764, miR-1264, miR-1298, miR-1911 and miR-488) as well as the expression correlation between HTR2C and these miRNAs. The diagram represents the gene structure according to UCSC genome browser. Because the expression of miR-764 could not be detected, expression correlation for this miRNA and its host gene is not shown.

frequency of intragenic miRNAs has also been expanding (5, 10, 13, 15). For example, the past miRBase release 11 (April 2008) had around 47% of intragenic miRNAs (3), and this proportion increased to 53% in the miRBase 19 (August 2012) and to 57% in the miRBase (20). miRIAD was created to help dealing with the challenges of unraveling the functional relationships between miRNAs and their host genes.

miRIAD data are organized in five layers of information. The first layer contains annotation for protein-coding and miRNA genes, including the official gene name, gene aliases and annotation. The second layer provides genomic information for host and miRNAs. The third layer contains gene expression for miRNAs and coding genes and expression correlation between intragenic miRNAs and their host genes. The fourth layer includes miRNA target prediction information (providing binding sites as well). The fifth layer contains additional information, which extends to protein–-protein interaction data for host genes as well as interaction partners that are targeted by host's intragenic miRNA. Additionally, a set of useful external links to other databases is given. All these information are organized in a streamlined graphical web tool and full integrated into a MySQL relational database. For users who want to manipulate miRIAD information in a local environment, we provide links to download raw data and python code. Specific information not found in those files can be obtained upon request. Therefore, miRIAD can be used to investigate miRNAs in a very integrative context, with special attention to functional features, such as protein–protein interaction, miRNAs targeting host mRNAs or their partners in a functional network. We believe that our web tool can be used as a starting point for developing and testing new hypotheses related to miRNA gene regulation, for one gene or for large-scale data. Importantly, scripts have been developed and pipelined to deal with forthcoming updates.

miRIAD improvements, updates and further development will be ongoing. For example, we envision including additional species and other useful data, such as expression from unhealthy samples. Information on the last and upcoming updates can be found on the miRIAD website.

In conclusion, miRIAD provides a systematic, integrative, user-friendly, and easy-to-use platform to investigate inter- and intragenic miRNAs, host genes and their relationships for five species, including human and mouse. Users can query for and clearly retrieve miRNA and host gene information. Therefore, we believe that miRIAD can substantially improve the way in which we investigate intragenic miRNA and host genes.

Supplementary Data

Supplementary data are available at *Database* Online.

Acknowledgements

We thank all members of the Bioinformatics lab for suggestions.

Funding

This work and some of the authors (P.A.F.G. and L.O.-M.) were partially supported by grants from National Institutes of Health (grant numbers D43TW007015, UH2HL108785 and U54HL108460). G.S.F. was supported by Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP, process number 2011/19645-6). Funding for open access charge: National Institutes of Health [D43TW007015].

Conflict of interest. None declared.

References

1. Krol,J., Loedige,I. and Filipowicz,W. (2010) The widespread regulation of microRNA biogenesis, function and decay. *Nat. Rev. Genet.*, **11**, 597–610.
2. Bartel,D.P. (2009) MicroRNAs: target recognition and regulatory functions. *Cell*, **136**, 215–233.

3. Kozomara,A. and Griffiths-Jones,S. (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.*, **39**, D152–D157.
4. Rodriguez,A., Griffiths-Jones,S., Ashurst,J.L. *et al.* (2004) Identification of mammalian microRNA host genes and transcription units. *Genome Res.*, **14**, 1902–1910.
5. Hinske,L.C.G., Galante,P.A.F., Kuo,W.P. *et al.* (2010) A potential role for intragenic miRNAs on their hosts' interactome. *BMC Genomics*, **11**, 533.
6. Hinske,L.C., Heyn,J., Galante,P.A.F. *et al.* (2013) Setting up an intronic miRNA database. *Methods Mol. Biol.*, **936**, 69–76.
7. Baskerville,S. and Bartel,D.P. (2005) Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA*, **11**, 241–247.
8. Monteys,A.M., Spengler,R.M., Wan,J. *et al.* (2010) Structure and activity of putative intronic miRNA promoters. *RNA*, **16**, 495–505.
9. Yan,L., Hao,H., Elton,T.S. *et al.* (2011) Intronic microRNA suppresses endothelial nitric oxide synthase expression and endothelial cell proliferation via inhibition of STAT3 signaling. *Mol. Cell. Biochem.*, **357**, 9–19.
10. Kinoshita,M., Ono,K., Horie,T. *et al.* (2010) Regulation of adipocyte differentiation by activation of serotonin (5-HT) receptors 5-HT2AR and 5-HT2CR and involvement of microRNA-448-mediated repression of KLF5. *Mol. Endocrinol.*, **24**, 1978–1987.
11. Cho,S., Jang,I., Jun,Y. *et al.* (2013) MiRGator v3.0: a microRNA portal for deep sequencing, expression profiling and mRNA targeting. *Nucleic Acids Res.*, **41**, D252–D257.
12. Meyer,L.R., Zweig,A.S., Hinrichs,A.S. *et al.* (2013) The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Res.*, **41**, D64–D69.
13. Godnic,I., Zorc,M., Jevsinek Skok,D. *et al.* (2013) Genome-wide and species-wide in silico screening for intragenic MicroRNAs in human, mouse and chicken. *PLoS One*, **8**, e65165.
14. Maselli,V., Di Bernardo,D. and Banfi,S. (2008) CoGemiR: a comparative genomics microRNA database. *BMC Genomics*, **9**, 457.
15. He,C., Li,Z., Chen,P. *et al.* (2012) Young intragenic miRNAs are less coexpressed with host genes than old ones: implications of miRNA-host gene coevolution. *Nucleic Acids Res.*, **40**, 4002–4012.
16. Brawand,D., Soumillon,M., Necsulea,A. *et al.* (2011) The evolution of gene expression levels in mammalian organs. *Nature*, **478**, 343–348.
17. Meunier,J., Lemoine,F., Soumillon,M. *et al.* (2013) Birth and expression evolution of mammalian microRNA genes. *Genome Res.*, **23**, 34–45.
18. Kim,D., Pertea,G., Trapnell,C. *et al.* (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.*, **14**, R36.
19. Trapnell,C. and Salzberg,S.L. (2009) How to map billions of short reads onto genomes. *Nat. Biotechnol.*, **27**, 455–457.
20. Trapnell,C., Williams,B.A., Pertea,G. *et al.* (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–515.
21. Marco,A. and Griffiths-Jones,S. (2012) Detection of microRNAs in color space. *Bioinformatics*, **28**, 318–323.
22. Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.
23. Dill,H., Linder,B., Fehr,A. *et al.* (2012) Intronic miR-26b controls neuronal differentiation by repressing its host transcript, ctdsp2. *Genes Dev.*, **26**, 25–30.
24. Liu,M., Roth,A., Yu,M. *et al.* (2013) The IGF2 intronic miR-483 selectively enhances transcription from IGF2 fetal promoters and enhances tumorigenesis. *Genes Dev.*, **27**, 2543–2548.
25. Zhu,Y., Lu,Y., Zhang,Q. *et al.* (2012) MicroRNA-26a/b and their host genes cooperate to inhibit the G1/S transition by activating the pRb protein. *Nucleic Acids Res.*, **40**, 4615–4625.
26. Radfar,M.H., Wong,W. and Morris,Q. (2011) Computational prediction of intronic microRNA targets using host gene expression reveals novel regulatory mechanisms. *PLoS One*, **6**, e19312.
27. Pollak,M. (2008) Insulin and insulin-like growth factor signalling in neoplasia. *Nat. Rev. Cancer*, **8**, 915–928.
28. da Cunha,J.P.C., Galante,P.A.F., de Souza,J.E. *et al.* (2009) Bioinformatics construction of the human cell surfaceome. *Proc. Natl Acad. Sci. USA*, **106**, 16752–16757.
29. de Souza,J.E.S., Galante,P.A.F., de Almeida,R.V.B. *et al.* (2012) SurfaceomeDB: a cancer-orientated database for genes encoding cell surface proteins. *Cancer Immun.*, **12**, 15.
30. Gorvin,C.M., Wilmer,M.J., Piret,S.E. *et al.* (2013) Receptor-mediated endocytosis and endosomal acidification is impaired in proximal tubule epithelial cells of Dent disease patients. *Proc. Natl Acad. Sci. USA*, **110**, 7014–7019.
31. Mickey,B.J., Sanford,B.J., Love,T.M. *et al.* (2012) Striatal dopamine release and genetic variation of the serotonin 2C receptor in humans. *J. Neurosci.*, **32**, 9344–9350.

Database, 2017, 1–8
doi: 10.1093/database/bax053
Database update

Database update

MiRIAD update: using alternative polyadenylation, protein interaction network analysis and additional species to enhance exploration of the role of intragenic miRNAs and their host genes

Ludwig C. Hinske[1,*], Felipe R. C. dos Santos[2,3], Daniel T. Ohara[2], Lucila Ohno-Machado[4], Simone Kreth[1] and Pedro A. F. Galante[2,*]

[1]Department of Anaesthesiology, University Hospital of the Ludwig-Maximilians-University Munich, Munich, Germany, [2]Centro de Oncologia Molecular, Hospital Sírio-Libanês, São Paulo SP 01308-060, Brazil, [3]Inter Unidades em Bioinformática, Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, SP, Brazil and [4]Health System Department of Biomedical Informatics, University of California San Diego, La Jolla, CA 93093, USA

*Corresponding author: Tel: +49 89 4400 76423; Fax: +49 89 4400 78886; Email: ludwig.hinske@med.uni-muenchen.de

Correspondence may also be addressed to Pedro A.F. Galante. Tel: +55 11 3394 4167; Fax: +55 11 3394 5304; Email: pgalante@mochsl.org.br

Citation details: Hinske,L.C., dos Santos,F.R.C., Ohara,D.T. *et al.* MiRIAD update: using alternative polyadenylation, protein interaction network analysis and additional species to enhance exploration of the role of intragenic miRNAs and their host genes. *Database* (2017) Vol. 2017: article ID bax053; doi:10.1093/database/bax053

Received 24 April 2017; Revised 29 May 2017; Accepted 20 June 2017

Abstract

MicroRNAs have established their role as potent regulators of the epigenome. Interestingly, most miRNAs are located within protein-coding genes with functional consequences that have yet to be fully investigated. MiRIAD is a database with an interactive and user-friendly online interface that has been facilitating research on intragenic miRNAs. In this article, we present a major update. First, data for five additional species (chimpanzee, rat, dog, cow and frog) were added to support the exploration of evolutionary aspects of the relationship between host genes and intragenic miRNAs. Moreover, we integrated data from two different sources to generate a comprehensive alternative polyadenylation dataset. The miRIAD interface was therefore redesigned and provides a completely new gene model representation, including an interactive visualization of the 3′ untranslated region (UTR) with alternative polyadenylation sites, corresponding signals and potential miRNA binding sites. Furthermore, we expanded on functional host gene network analysis. Although the previous version solely reported protein interactions, the update features a separate network analysis view that can either be accessed through the submission of a list of genes of interest or directly from a gene's list of protein interactions. In addition to statistical properties of the submitted gene set, the interaction network graph is presented and miRNAs with seed site over- and underrepresentation are identified. In summary, the update of miRIAD provides novel datasets and bioinformatics resources with a significant increase in functionality to facilitate intragenic miRNA research in a user-friendly and interactive way.

Database URL: http://www.miriad-database.org

Introduction

MiRNAs are well-known as small molecules that are involved in controlling regulatory networks of the gene expression (1). Interestingly, most (e.g. 61.5% for human and 66.2% for mouse) miRNA genes are positioned within protein-coding genes in vertebrates (2, 3). These miRNAs are called intragenic miRNAs and their enclosing genes 'host genes'. Accumulating evidence suggests that this special relationship of genomic colocalization between an intragenic miRNA and its host gene is of biological relevance. Negative feedback loops of intragenic miRNAs regulating their host genes have recently been described, ranging from first-order (i.e. direct) negative feedback (4–6) to indirect feedback loops (2, 7, 8).

Using a myriad of data from different sources and databases focused on the analysis of intragenic miRNAs (3, 9–13), we and others have found further functional implications of intragenic miRNAs and their host genes. Recent research suggests evolutionary implications of intragenic miRNA development (14, 15), yielding that novel miRNAs seem to benefit from intragenic colocalization by utilizing existing regulatory circuitries of their host genes (14). Furthermore, increasing evidence highlights the importance of the role of alternative polyadenylation (APA) to characterize the relationship between intragenic miRNAs and their host genes (5, 6). These novel discoveries prompted us to develop a major update of the miRIAD database and interface to account for these new aspects of intragenic miRNA–host gene relationship.

In this article, we provide a detailed description of the updated version of miRIAD. In its first version, miRIAD integrated genomic data for five species to classify miRNAs into intergenic, intronic and exonic, allowing easy identification of intragenic miRNAs and host genes (3). In the updated version, miRIAD contains five additional species (chimpanzee, rat, dog, cow and frog). Among other changes, it was redesigned to include APA information from two different sources (16, 17) for 8 of 10 included species (human, rhesus, chimpanzee, mouse, rat, dog, opossum and chicken). To maximize utility of these new data, the gene model visualization was completely redesigned to implement interactive vector graphics. Interaction network analysis functionality was added to allow evaluation of a set of genes (e.g. gene signatures) with respect to host gene over- or underrepresentation, visualization of protein interactions with respect to intragenic miRNA targeting and identification of over- or underrepresented miRNA target sites in a network. We also show, how to use the new functionality to derive hypotheses about the relationship between a host gene (AKT2) and its intragenic miRNA (hsa-miR-641). To the best of our knowledge, miRIAD is the first public resource to allow these analyses to investigate the role of intragenic miRNAs.

Materials and methods

MiRIAD construction and integration of additional species

Selection of species to be integrated in miRIAD was based on several factors. First, we required the availability of high quality genome assemblies and a good RefSeq coverage. Second, we searched for available polyadenylation, gene and miRNA expression data. Construction of the miRIAD database was performed with the newest genome assemblies (human: hg38/GRCh38, rhesus: rheMac8, chimp: panTro5, frog: xenTro7, cow: bosTau8, opossum: monDom5, rat: rn6, chicken: galGal5, dog: canFam3 and mouse: mm10; Figure 1A) and mirBase version 21 (12), as described in (3). Coding gene and miRNA expression was calculated from RNA-Seq data from Brawand *et al.* (17), Gene Expression Omnibus (GSE30352). RNA-Seq data processing was carried out as previously described (3).

APA information for eight species

We combined APA data from a previously published dataset from (16) (human, rhesus, dog, mouse and rat) with APA information that we derived from processing the dataset obtained by Brawand *et al.* (17). Poly(A) coordinates from Derti *et al.* were mapped to the respective current

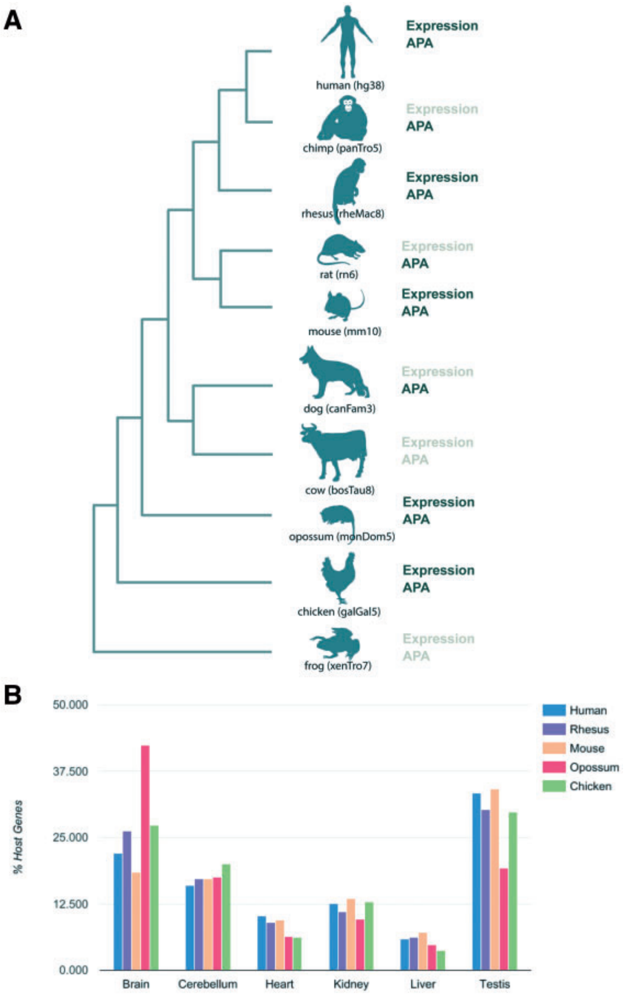


Figure 1. Summary of species in the miRIAD and host genes expression. (**A**) Species present in the miRIAD. (**B**) Host genes expression in six tissues.

genome assemblies using the liftOver tool provided by the Genome Browser from the University of California Santa Cruz (UCSC) (18). Identification of APA sites from RNA-Seq data from Brawand *et al.* was carried out as follows. After data preprocessing [for details see (3)], reads were filtered for those starting or ending with at least four untemplated 'A's or 'T's. Reads with an extremely high A/T/N-content were ignored (cut off ratio was set to 0.8). Potential APA sites were considered, if they (i) mapped to a untranslated region (UTR)-annotated region based on RefSeq and (ii) were supported by at least two independent reads. APA sites within 40 nucleotides were considered to be a single APA site. For benchmarking, expressed sequence tags based alternative poly(A) site information from APADB (19) were downloaded for human and mouse. For human, APA site information was converted to hg38 using the liftOver tool (18). Only APA sites mapping to RefSeq UTR models were considered.

Target predictions and protein interaction network

The protein interaction network feature visualizes relationships between gene products as an interactive scalable vector graphics (SVG) image. Using an enrichment–calculation-based target prediction network score, it may also help to identify miRNAs relevant for regulation of this network, which yet lacks experimental support. Target predictions are based on canonical seed matching used by Targetscan on the 3′-UTR sequences of protein-coding RefSeq transcripts (10). In brief, 3′-UTRs are scanned for base complementarity to Bases 2–7 of the mature miRNA sequences (seed region). Hybridization energy between miRNA and UTR sequence was calculated using the Vienna RNA library (20). The impact of a miRNA on a set of genes is quantified as follows: First, the probability of random occurrence of a given seed sequence is calculated by $P(S) = \prod_{i=1}^{n} P(N_i|D)$, where $S =$ seed sequence, $n =$ length(S), $N_i = i$th nucleotide of S, $D =$ nucleotide distribution.

The probability that this sequence occurs at least r times in a random sequence of length N (UTR sequences for each gene in the network) is given by:

$$P(x_t) = \left(1 - \sum_{i=0}^{r-1}\left(\binom{L_x}{i} * P(S)^i * (1 - P(S))^{L_x - i}\right)\right),$$

where $L_x =$ (length of 3′-UTR of element x_t) − (length of seed sequence n) + 1, $r =$ desired minimum number of occurrences, miRIAD is using $r = 1$.

The expected number of genes containing seed-matching sites $E(X_t)$ in the network X can then be estimated by the sum of probabilities for each gene x.

$$E(X_t) = \sum_{x_t \in X} P(x_t)$$

This number of expected random target genes in the network can be compared with the observed number of genes with seed matches. Statistical evaluation is possible using Fisher's exact test. The score reported in miRIAD equals the log-odds ratio, given by:

$$\mathrm{Score}(X\,|\,S) = \log\left(\frac{\frac{E(X_t)+O(X_t)}{E(X_t)} * \frac{E(X_n)+O(X_n)}{E(X_n)}}{\frac{|X|}{E(n)}}\right).$$

Results

Database statistics

The current version of miRIAD contains 10 species, with a total of 284 374 protein-coding genes and 7369 miRNAs.

Table 1. Summarized statistics for miRNAs and host genes

Organism (genome assembly)	miRNAs (total)	Total number of genes	Total number APA sites	Intragenic miRNAs			Intergenic miRNAs	Host genes	Host genes with APA sites	APA sites in host genes	Number of predicted target interactions	Number of protein interactions
				intronic (sense in %)	exonic (sense in %)	all (sense in %)						
Homo sapiens (hg38)	1881	20 204	206 656	988 (82.4%)	169 (81.1%)	1157 (82.2%)	724	984	949	16 921	22 930 745	7 553 352
Pan troglodytes (panTro5)	628	33 003	39 860	280 (86.8%)	22 (45.5%)	302 (83.8%)	326	253	115	1066	4 230 060	8 182 491
Mulatta macaca (rheMac8)	486	28 336	101 385	214 (83.2%)	21 (57.1%)	235 (80.9%)	251	201	161	1515	3 804 170	11 339 842
Mus musculus (mm10)	1187	36 058	124 297	644 (87.4%)	142 (81.0%)	786 (86.3%)	401	649	599	6377	12 405 027	11 090 779
Rattus norvegicus (rn6)	485	32 387	48 198	153 (73.2%)	45 (62.2%)	198 (70.7%)	287	143	111	725	4 281 941	12 781 450
Bos taurus (bosTau8)	811	27 121	NA	396 (79.5%)	45 (68.9%)	441 (78.5%)	370	356	NA	NA	4 371 894	9 953 188
Canis familiaris (canFam3)	513	24 782	79 089	192 (81.3%)	25 (12.0%)	217 (73.3%)	269	171	121	982	2 025 226	9 551 812
Gallus gallus (galgal5)	709	25 610	6561	350 (87.7%)	43 (58.1%)	393 (84.5%)	316	353	63	217	4 366 456	2 845 210
Monodelphis domesticus (monDom5)	460	33 101	5299	165 (89.1%)	2 (0.0%)	167 (88.0%)	293	131	14	56	6 056 838	11 585 890
Xenopus tropicalis (xenTro7)	209	23 772	NA	61 (60.7%)	0	61 (60.7%)	148	46	NA	NA	1 066 564	0
	7369	284 374	611 345	3344	481	3825	3385	3287	2133	27 859	65 538 921	84 884 014

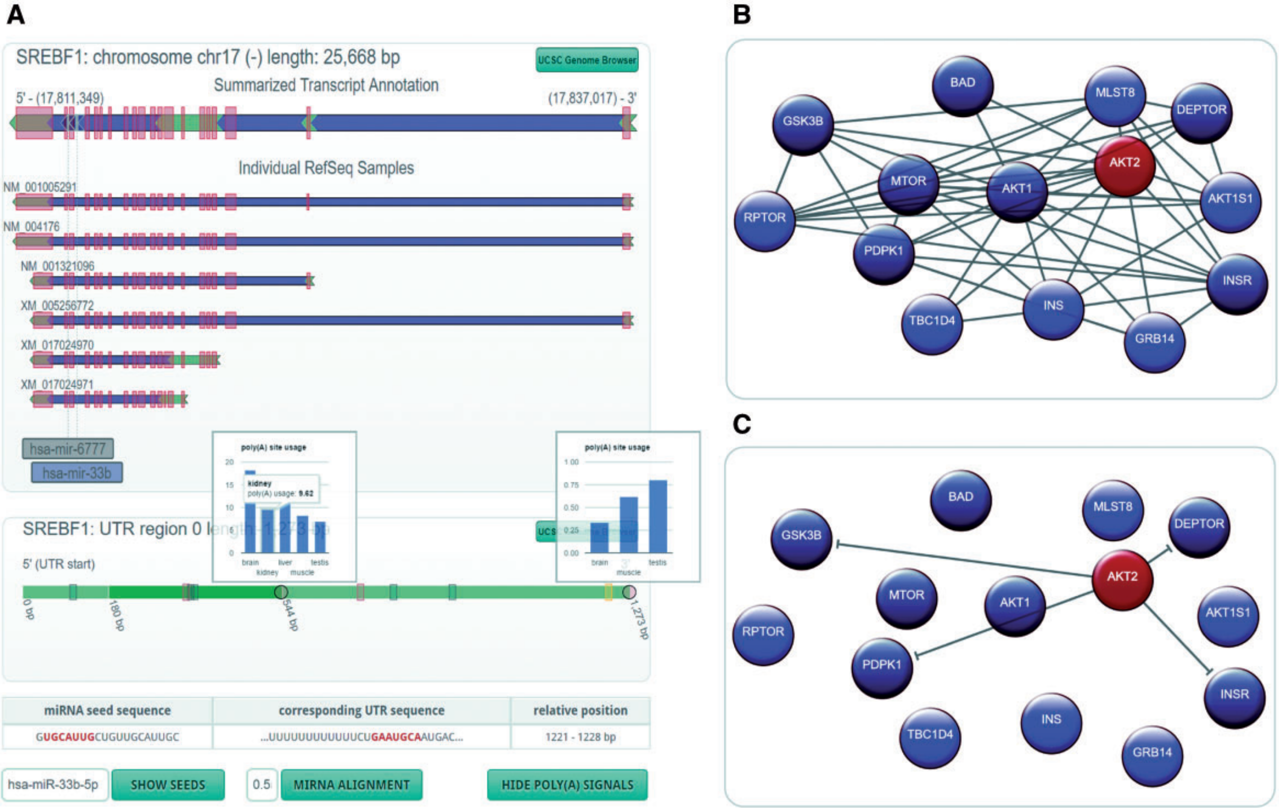
NA, not available

In total, 61.5% of human and 66.2% of mouse miRNAs are intragenic. Expression data for miRNAs as well as for mRNAs are available for six organs (brain, cerebellum, heart, kidney, liver and testis) from human, mouse, rhesus, opossum and chicken (Figure 1A). Investigating the distribution across tissues in human, we found that host genes of intragenic miRNAs are predominantly expressed in neuronal tissue and testis across all organisms (Figure 1B). We were able to extract APA information for 8 of 10 species in miRIAD (Figure 1A). According to our database, 94.6% of human host genes have annotated APA sites, which is more than expected compared with 83% of all human genes (*P* value = 4.6e-38, Fisher exact test). Similarly, 92.3% of murine host genes and have annotated APA sites (72% of all murine genes). This relationship is true with varying degrees for chicken (18% of host genes, expected 11%, *P* value = 3.5e-4), rat (78 vs 59%, *P* value = 5e-06), rhesus (80 vs 65%, *P* value = 2e-06) and chimpanzee (45 vs 28%, *P* value = 3e-09). We did not find significant differences in dog (71 vs 69%, *P* value = 0.68) and opossum (11 vs 8%, *P* value = 0.26). Summarized statistics are available in Table 1. We used the previously published database APADB to benchmark APA sites for mouse and human included in miRIAD (19). APA site information for a total of 14 143 human and 13 472 murine genes was compared. miRIAD includes 29 349 of the 34 753 events registered in APADB mappable to our UTR models (84.5%). Similarly, 82% of murine APA sites were covered by miRIAD (20 826 of 25 323).

Interactive structural representation of UTR, miRNA and host gene relationship

The representation of structural properties of a host gene and its intragenic miRNA is of great importance when investigating their relationship (2, 14). In the new miRIAD-version, we developed a representation based on interactive SVG to visualize the gene structure, highlighting exonic, intronic and UTRs (Figure 2A). It contains a summarized representation, in which region information is merged, followed by individual RefSeq transcripts of the gene of interest. The positions of intragenic miRNAs are shown in the summarized transcript and relative to individual transcripts. This allows the researcher to check for transcripts devoid of the intronic miRNA, proximity to upstream exons as an indicator of cotranscription or organization of miRNA genes in mirtrons. Figure 2A shows the gene model representation of SREBF1 with its intronic miRNAs *miR-6777* and *miR-33b*. The latter is highlighted in blue to indicate that the host gene has at least one seed-matching site within its 3′-UTR.

In addition to structural properties of the gene, the organization of the 3′-UTR further characterizes the



miRNA seed sequence	corresponding UTR sequence	relative position
GUGCAUUGCUGUUGCAUUGC	...UUUUUUUUUUUUUCUGAAUGCAAUGAC...	1221 - 1228 bp

Figure 2. miRIAD representation for a host gene, its intragenic miRNAs, protein–protein interaction (PPI) data and an intragenic miRNA target. (**A**) Genomic representation (including polyadenylation information) for a host gene (SREBF1) and its intragenic miRNAs (hsa-mir-6777 and hsa-mir-33b). (**B**) PPI network for AKT2; (**C**) gene targets for hsa-miR-641, which is an intragenic miRNA for AKT2.

relationship between an intragenic miRNA and its host gene. We therefore included a novel representation of 3′-UTR variants based on published and self-constructed APA information. Segmentation of the UTR by APA sites is symbolized by alternating shades of green. If the user moves the mouse cursor over an intragenic miRNA highlighted in blue, the position of the seed within the UTR will show up. In the case of SREBF1 and hsa-miR-33b, the seed site is located on the 3′ extremity of the transcript with at least one isoform without this seed-matching site. A click on the button 'show poly(A) signals' reveals canonical polyadenylation signals, in this case indicating that there might actually be another APA site that has not yet been described. To support investigation of differential miRNA targeting, APA site utilization across tissues can be visualized where available. A click on a gray circle in the 3′-UTR will open utilization information for this site.

Additionally, the 'show seeds' option allows the identification of seed-matching sites for any miRNA. If no seed matches are found, the user can choose 'miRNA alignment' to search for regions of high similarity to the mature miRNA sequence, helping to identify non-canonical miRNA binding sites. Clicking on a potential miRNA binding site in the UTR (either yellow for seed sites or gray for non-canonical sites) will show the sequence of the miRNA and the sequence of the region of interest in the 3′-UTR (Figure 2A).

Each of the 3′-UTR model representation displays a button on the right top corner that will open the UCSC Genome Browser (18, 21) for the specific UTR region or the full gene model. In this way, a plethora of additional information can be gained, such as evolutionary conservation, without sacrificing simplicity of miRIAD interface usage.

The interactive, visual representation of the gene model is followed by expression information of the gene across the tissues cerebellum, brain, heart, liver, kidney and testis, as well as a figure correlating the expression of the intragenic miRNA with the host gene across these tissues, providing Spearman's rank correlation coefficient and a *P* value. These figures help to rapidly identify tissue specificity, as well as coregulation (indicated by high correlation of expression). An interesting example is that of MAP2K4 and intragenic miRNA 'hsa-miR-744', in which both miRNA transcripts (hsa-miR-744-3p and hsa-miR-744-5p) correlate extremely well with their host gene's expression. Similar to the gene view, the miRNA view yields the structural representation of the miRNA gene, expression across tissues and correlation graphics with their host genes.

Filtering of targeting miRNAs and protein interactions

Although some decades ago, research was focusing on the exploration of single genes only, evaluation of protein interactions and regulation through miRNAs has become increasingly important. Although both protein interaction and target prediction information were already available in the first version of miRIAD, it now includes more data and supports filtering of these. Targeting miRNAs for example can be filtered by score or by name, in case the user wants to check a specific location for a miRNA target interaction or just wants to find miRNAs with a high binding probability. Differential miRNA targeting can be assessed by identifying miRNAs that bind only to a specific APA isoform through filtering for a specific poly(A) index. A click on the miRNA symbol will highlight the seed match(es) within the UTR of the gene (Figure 2A). Also, the list of miRNAs can be significantly reduced by filtering for the tissue of maximum expression. This is especially useful, when looking for potential regulators of a gene that shows strong tissue specificity.

Similarly, genes whose products interact with the gene of interest can be filtered by gene name, score and type of interaction (for STRING), evidence of binding (e.g. two hybrid system or direct interaction for BIND) and by data source [Bind (22), STRING (23), HPRD (24) and BioGRID (25)].

If the gene of interest contains intragenic miRNAs, information on interacting proteins that are potentially targeted by this miRNA will appear. This allows the user to estimate the impact of the intragenic miRNA on the host gene better. The filtered selection of interacting genes can then be submitted to the newly introduced network analysis view for extended evaluation.

Network view: analysis of complex interactions

It is known that intragenic miRNAs have a special impact on their host genes' surrounding network (2, 14). We therefore implemented an algorithm that helps identify miRNAs relevant for networks of genes. If a researcher identifies a set of interesting genes, e.g. a cancer gene signature, it might be of great interest, whether host genes are over- or underrepresented in this gene signature, how these genes interact with one another and if there are miRNAs relevant to this gene signature as a whole. A miRIAD query with a preceding colon followed by the gene symbols of the signature (separated by spaces) will load the network view to help answer these biologically relevant questions. First, statistics on the number of host genes in the submitted gene list (including an estimative of the significance of over-/underrepresentation), their intragenic miRNAs (if any) and the most relevant properties of each relationship (same strand, seed site within host UTR) are shown. The most central part is the network representation (Figure 2B and C), which visualizes regular genes (blue), host genes (red) and protein interactions between them. Network nodes can be rearranged by the user for better visualization, and mouseover will highlight all nodes with direct interactions, which makes it easy to identify hubs in large networks. Interactions can be filtered by score or data origin. Also, if the network contains host genes, interaction arrows can be replaced by predicted target interactions of the intragenic miRNA(s) (Figure 2C).

Exploring the relationship between AKT2 and its intronic miRNA miR-641

AKT2 hosts intragenic miRNA *hsa-miR-641* but the relationship between these two being largely unknown. The gene structure representation shows that miR-641 is located on the same strand as its host gene, and that it is positioned in the first intron. Although this fact *per se* might suggest coregulation, there are four (predicted) RefSeq transcripts that don't include miR-641. Correlation between miRNA and host gene cannot be well-characterized, since miR-641 seems to be only expressed in neuronal tissue. Filtering AKT2's interaction partners for STRING-reported interactions with a minimum score of 900 reveals the network in Figure 2B. MiRNA hsa-miR-637 ranks high in the list of miRNAs that potentially impact the network (score 1.34; targeted genes are dark-blue/dark-red). It is known to control the AKT-pathway (26). Interestingly, targets are very similar to hsa-miR-641 (Figure 2C), indicating a similar function for these two miRNAs. Moreover, miR-641 is only also found in chimp in our dataset, suggesting a relatively new evolutionary role. This example shows how miRIAD can be used to derive hypotheses about the relationship between a miRNA and its host gene.

Discussion

Nowadays, in the era of large scale data generation in genomics and transcriptomics, it is essential to have powerful and user-friendly tools to mine the right information, to propose and to test hypotheses regarding the studied model. The special genomic colocalization of most vertebrate miRNAs intragenically is of great relevance and current studies have been revealing that the functional implications of this coupling extend beyond simple feedback regulatory mechanisms but seems to support miRNA evolution (2, 5, 14). This revelation expands the focus of research requiring tools to study intragenic miRNAs and genes in an evolutionary context.

The new version of miRIAD was therefore extended to a total of 10 species, covering major phylogenetic branches. Statistics on APA show that significantly more host genes contain APA sites than would be expected. This is even true for chicken, the most distant specie investigated. Interestingly, dog and opossum, both being closer to human, don't display this phenomenon. This discovery might be biased by the fact that genome annotation of dog and opossum is not as complete as other genomes but it may also be a starting point for the investigation of a potentially underlying biological principle.

These analyses are complemented by newly implemented data and functionality to accommodate complex data investigation, such as miRNA-host gene centered network analysis and visualization of APA with respect to miRNA binding sites. MiRIAD can now be used to derive interesting hypotheses about the relationship between a miRNA and its host gene. As it was illustrated for AKT2 and its intragenic miRNA miR-641, e.g. miRIAD allowed us to generate the hypothesis that miR-641 might control the AKT pathway in neuronal tissue in human and chimp. It also allows rapid identification of miRNAs that may bind to specific UTR regions or target only specific alternatively polyadenylated isoforms.

At this point, complete gene and miRNA expression and APA information is available only for 8 of 10 species. This is owed to fact that currently only Brawand *et al.* (17) provide a dataset that contains RNA sequencing information on miRNAs and mRNAs from the same individuals, across multiple species and tissues. However, we expect to be able to include additional datasets in future versions. We hope to provide additional poly(A) site information for frog and cow, as well as miRNA and mRNA expression data. Furthermore, miRIAD currently implements target predictions only through seed site matching, ignoring non-canonical sites. This strategy is necessary for the implementation of our model that quantifies the probability of a miRNA-network effect. However, miRIAD is an ongoing project and we are planning to present an extended model that includes non-canonical sites, tissue specificity and APA information in upcoming releases.

In summary, the new version of miRIAD adds important new data and functionality to enhance the exploration of the role of intragenic miRNAs through providing APA information and network analysis in the light of phylogeny.

Acknowledgements

The authors thank Gustavo S. França for working on the previous miRIAD version as well as Anamaria A. Camargo and Luiz F. L. Reis for their valuable suggestions.

Funding

F.R.C.S. and P.A.F.G. were partially supported by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). Funding for open access charge: Hospital SÚrio-Libanês.

Conflict of interest. None declared.

References

1. Bartel,D.P. (2009) MicroRNAs: target recognition and regulatory functions. *Cell*, 136, 215–233.
2. Hinske,L.C.G., Galante,P.A.F., Kuo,W.P. *et al*. (2010) A potential role for intragenic miRNAs on their hosts' interactome. *BMC Genomics*, 11, 533.
3. Hinske,L.C., França,G.S., Torres,H.A.M. *et al*. (2014) miRIAD-integrating microRNA inter- and intragenic data. *Database*, 2014. doi: 10.1093/database/bau099
4. Dill,H., Linder,B., Fehr,A. *et al*. (2012) Intronic miR-26b controls neuronal differentiation by repressing its host transcript, ctdsp2. *Genes Dev*., 26, 25–30.
5. Hinske,L.C., Galante,P.A.F., Limbeck,E. *et al*. (2015) Alternative polyadenylation allows differential negative feedback of human miRNA miR-579 on its host gene ZFR. *PLoS One*, 10, e0121507.
6. Paraboschi,E.M., Cardamone,G., Rimoldi,V. *et al*. (2017) miR-634 is a Pol III-dependent intronic microRNA regulating alternative-polyadenylated isoforms of its host gene PRKCA. *Biochim. Biophys. Acta*, 1861, 1046–1056.
7. Horie,T., Nishino,T., Baba,O. *et al*. (2014) MicroRNA-33b knock-in mice for an intron of sterol regulatory element-binding factor 1 (Srebf1) exhibit reduced HDL-C in vivo. *Sci. Rep*., 4, 5312.
8. Kos,A., Olde Loohuis,N.F.M., Wieczorek,M.L. *et al*. (2012) A potential regulatory role for intronic microRNA-338-3p for its host gene encoding apoptosis-associated tyrosine kinase. *PLoS One*, 7, e31022.
9. Kent,W.J., Sugnet,C.W., Furey,T.S. *et al*. (2002) The human genome browser at UCSC. *Genome Res*., 12, 996–1006.
10. Lewis,B.P., Burge,C.B. and Bartel,D.P. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120, 15–20.
11. Dweep,H. and Gretz,N. (2015) miRWalk2.0: a comprehensive atlas of microRNA-target interactions. *Nat. Methods*, 12, 697.
12. Kozomara,A. and Griffiths-Jones,S. (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res*., 42, D68–D73.
13. Griffiths-Jones,S., Grocock,R.J., van Dongen,S. *et al*. (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res*., 34, D140–D144.
14. França,G.S., Vibranovski,M.D. and Galante,P.A.F. (2016) Host gene constraints and genomic context impact the expression and evolution of human microRNAs. *Nat. Commun*., 7, 11438.
15. França,G.S., Hinske,L.C., Galante,P.A.F. *et al*. (2017) Unveiling the impact of the genomic architecture on the evolution of vertebrate microRNAs. *Front. Genet*., 8, 34.
16. Derti,A., Garrett-Engele,P., Macisaac,K.D. *et al*. (2012) A quantitative atlas of polyadenylation in five mammals. *Genome Res*., 22, 1173–1183.
17. Brawand,D., Soumillon,M., Necsulea,A. *et al*. (2011) The evolution of gene expression levels in mammalian organs. *Nature*, 478, 343–348.

18. Tyner,C., Barber,G.P., Casper,J. *et al.* (2017) The UCSC Genome Browser database: 2017 update. *Nucleic Acids Res.*, 45, D626–D634.
19. Müller,S., Rycak,L., Afonso-Grunz,F. *et al.* (2014) APADB: a database for alternative polyadenylation and microRNA regulation events. *Database*, 2014. doi: 10.1093/database/bau076.
20. Hofacker,I.L. (2009) RNA Secondary structure analysis using the vienna RNA package. *Curr. Protoc. Bioinformatics*. doi: 10.1002/0471250953.bi1202s04.
21. Karolchik,D., Baertsch,R., Diekhans,M. *et al.* (2003) The UCSC Genome Browser Database. *Nucleic Acids Res.*, 31, 51–54.
22. Bader,G.D., Betel,D. and Hogue,C.W.V. (2003) BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.*, 31, 248–250.
23. Szklarczyk,D., Franceschini,A., Wyder,S. *et al.* (2015) STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.*, 43, D447–D452.
24. Prasad,T.S.K., Goel,R., Kandasamy,K. *et al.* (2009) Human protein reference database—2009 update. *Nucleic Acids Res.*, 37, D767–D772.
25. Chatr-Aryamontri,A., Oughtred,R., Boucher,L. *et al.* (2017) The BioGRID interaction database: 2017 update. *Nucleic Acids Res.*, 45, D369–D379.
26. Que,T., Song,Y., Liu,Z. *et al.* (2015) Decreased miRNA-637 is an unfavorable prognosis marker and promotes glioma cell growth, migration and invasion via direct targeting Akt1. *Oncogene*, 34, 4952–4963.

RESEARCH ARTICLE

Alternative Polyadenylation Allows Differential Negative Feedback of Human miRNA miR-579 on Its Host Gene ZFR

Ludwig Christian Hinske[1]*, Pedro A. F. Galante[2], Elisabeth Limbeck[2], Patrick Möhnle[1], Raphael B. Parmigiani[2], Lucila Ohno-Machado[3], Anamaria A. Camargo[2], Simone Kreth[1]*

1 Research Group Molecular Medicine, Department of Anaesthesiology, Clinic of the University of Munich, Munich, Germany, **2** Molecular Oncology Center, Sírio Libanês Hospital, São Paulo, Brazil, **3** Division of Biomedical Informatics, University of California San Diego, La Jolla, California, United States of America

* simone.kreth@med.uni-muenchen.de (SK); ludwig.hinske@med.uni-muenchen.de (LH)

Citation: Hinske LC, Galante PAF, Limbeck E, Möhnle P, Parmigiani RB, Ohno-Machado L, et al. (2015) Alternative Polyadenylation Allows Differential Negative Feedback of Human miRNA miR-579 on Its Host Gene ZFR. PLoS ONE 10(3): e0121507. doi:10.1371/journal.pone.0121507

Academic Editor: Dong-Yan Jin, University of Hong Kong, HONG KONG

Received: September 26, 2014

Accepted: January 31, 2015

Published: March 23, 2015

Data Availability Statement: All relevant data are within the paper and its Supporting Information files, or were deposited at Next Gen Sequencing data, and are accessible at the following URL: (http://www.ncbi.nlm.nih.gov/sra/?term=SRP053217).

Funding: This work and two of the authors (PAFG and LOM) were partially supported by grants from the National Institutes of Health (grant numbers D43TW007015 and U54HL108460). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

About half of the known miRNA genes are located within protein-coding host genes, and are thus subject to co-transcription. Accumulating data indicate that this coupling may be an intrinsic mechanism to directly regulate the host gene's expression, constituting a negative feedback loop. Inevitably, the cell requires a yet largely unknown repertoire of methods to regulate this control mechanism. We propose APA as one possible mechanism by which negative feedback of intronic miRNA on their host genes might be regulated. Using in-silico analyses, we found that host genes that contain seed matching sites for their intronic miRNAs yield longer 32UTRs with more polyadenylation sites. Additionally, the distribution of polyadenylation signals differed significantly between these host genes and host genes of miRNAs that do not contain potential miRNA binding sites. We then transferred these in-silico results to a biological example and investigated the relationship between ZFR and its intronic miRNA miR-579 in a U87 cell line model. We found that ZFR is targeted by its intronic miRNA miR-579 and that alternative polyadenylation allows differential targeting. We additionally used bioinformatics analyses and RNA-Seq to evaluate a potential cross-talk between intronic miRNAs and alternative polyadenylation. CPSF2, a gene previously associated with alternative polyadenylation signal recognition, might be linked to intronic miRNA negative feedback by altering polyadenylation signal utilization.

Introduction

In the recent past, miRNAs have gained significant attention as regulators of the transcriptome. MiRNA genes are found throughout the genome, and about half of them are located in genomic regions that contain protein-coding information. They can be classified as either intergenic or intragenic, and the latter can be subclassified as *exonic* or *intronic* [1]. While some intronic miRNAs may be regulated by their own promoter sequences [2], the expression of the majority of intronic miRNAs depends on transcriptional activation of the host gene: When a protein-coding

Competing Interests: The authors have declared that no competing interests exist.

gene is transcribed into mRNA, this primary transcript also contains the miRNA sequence that may subsequently be processed into a mature miRNA [3]. Consequently, the expression of a miRNA can be coupled to the expression of its host gene. Increasing evidence suggests that this miRNA—host gene relationship is of functional importance: Intronic miRNAs may affect their hosts' expression or the expression of host-interacting proteins [1]. In both cases, intronic miRNAs were shown to influence the molecular activities of their hosts. Recently, Dill et al. experimentally validated an example of an intronic miRNA targeting its host gene, hence uncovering a direct negative feedback mechanism [4]. Interestingly, the miRNA was processed only after differentiation of the cell, showing that this mechanism was time-dependent. This clearly proved the existence of functional relationships between intronic miRNAs and their host genes. Furthermore, this work identified a first example for regulation of this coupling. However, the described model was limited to cell differentiation processes. So far it remains unclear whether there exist more general mechanisms that may enable control of host gene expression by intronic miRNAs.

Whereas differential processing of the intronic miRNA constitutes one way to control activity of a negative feedback mechanism, modulation of miRNA target-site accessibility may be another option. Many protein-coding genes bear multiple polyadenylation sites in their 32UTRs, enabling the transcription of variable size mRNAs that may or may not contain specific miRNA target sites [5]. Poly(A)-site selection is determined by context and type of polyadenylation signals. In general, canonical polyadenylation signals ("AAUAAA", "AUUAAA") are distinguished from non-canonical polyadenylation signals. Several enzymes have been identified that are linked to 3´UTR processing and are commonly referred to as 3´-processing factors, the stoichiometry of which seems to be very influential (for a detailed summary of alternative polyadenylation see [6]). We hypothesized that miRNA target-site accessibility could be modulated by alternative polyadenylation (APA) processes as an additional mechanism of intronic miRNA-driven negative feedback loops. First, we used a bioinformatics approach to investigate, whether APA-motif distribution differs in the 32UTRs of host genes with and without an intronic miRNA seed matching site. We then chose ZFR and its intronic miRNA miR-579 as an example and could show that ZFR is in fact targeted by miR-579. Moreover, we show that there are at least two 32UTR isoforms, one of which contains the miRNA target site while the other doesn't, proving that alternative polyadenylation is a way for the cell to scale the degree of immediate negative feedback. We also investigated, whether intronic miRNAs targeting their own host gene may interfere with polyadenylation machinery. Using bioinformatics screening for overrepresented potential miRNA targets within the APA machinery, we identified CPSF2 as a potential intronic miRNA target. We show that ZFR targets CPSF2, and that silencing of CPSF2 lead to an increased utilization of canonical polyadenylation signals. These data indicate an interesting link between intronic miRNA feedback and alternative polyadenylation.

Results and Discussion

APA regulates the impact of intronic miRNAs on the expression of their host genes

To investigate the hypothesis that APA regulates a negative feedback mechanism imposed by miRNAs targeting their own hosts, we first classified intronic miRNAs into host-targeting (HT) miRNAs or non-host-targeting (NT) miRNAs by searching for seed site matches within the respective 32UTR sequences of the host genes. A total of 203 HT miRNAs were located in 168 host genes, with 583 seed site matches. 601 NT miRNAs were located within 351 host genes (see also S1 Fig.). We found that HT miRNA host genes possess longer 32UTR sequences (median = 2553 nt vs median = 1198 nt, $P < 2.2E-16$) and contain significantly more poly(A) sites than NT miRNA host genes (median = 5 vs median = 3, P = 6.7E-9) (Fig. 1A). Of 583 total seed site matches, 435 HT miRNA-matching seed sites are potentially influenced by APA,

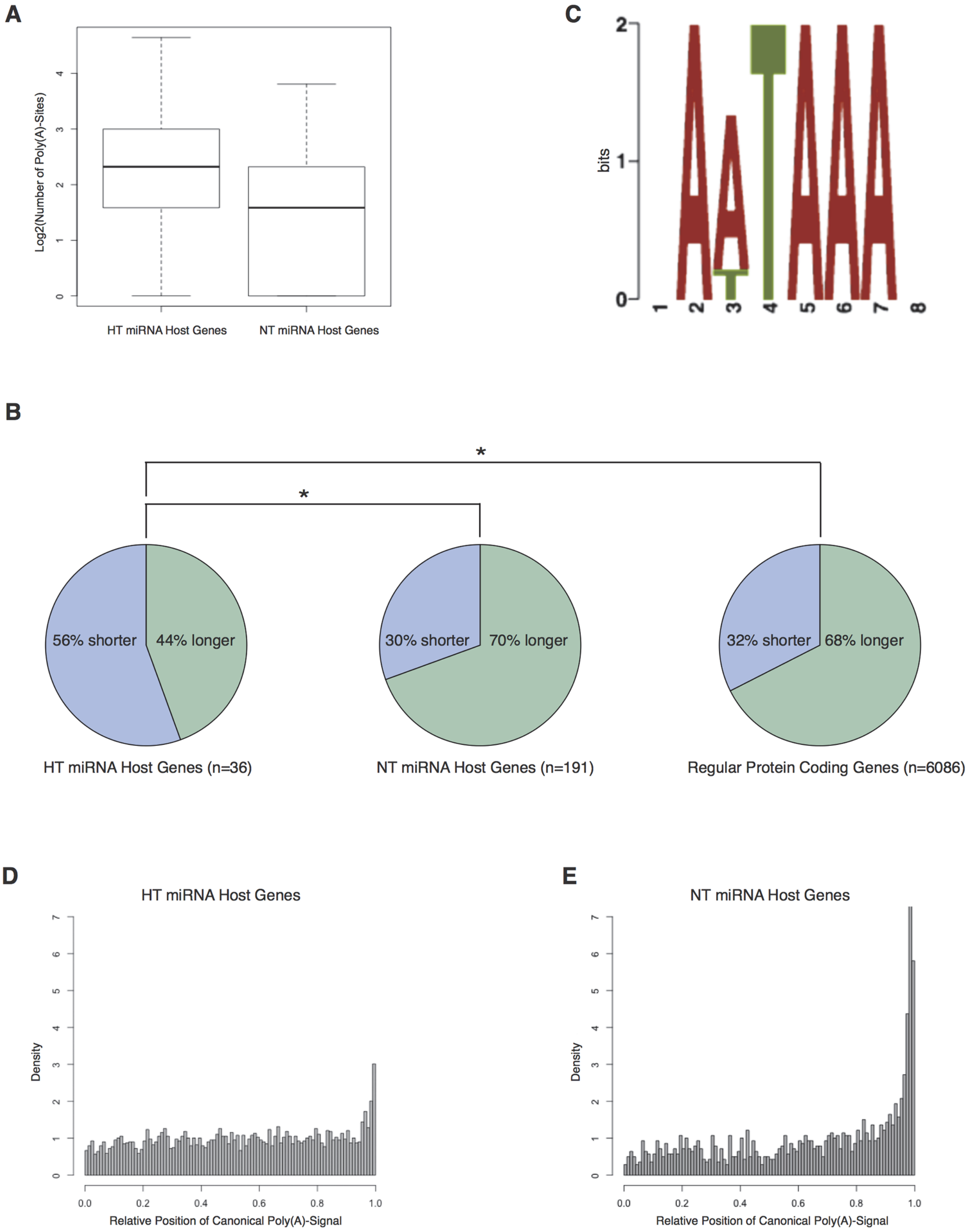


Fig 1. Bioinformatics and biomolecular analyses indicate a role for APA in regulation of negative feedback. A) Comparison of APA-sites for HT miRNA host genes and NT miRNA host genes. B) After CPSF2 silencing HT miRNA host gene UTRs display a different poly(A)-site usage pattern compared to NT miRNA host gene UTRs and regular protein-coding genes' UTRs. C) The motif discovered in upregulated APA regions after CPSF2 silencing resembles the two canonical polyadenylation sites. D) Distribution of canonical poly(A) signals across the 32UTR of HT miRNA host genes and E) NT miRNA host genes.

doi:10.1371/journal.pone.0121507.g001

affecting 124 of the 168 HT host genes. In summary, our results illustrate that 32UTRs of HT miRNA host genes are longer and contain more APA sites. Long 32UTRs have been shown to preferably occur in genes in which slight expression changes can be detrimental to the cell, thus requiring tight regulation [6]. We then mapped the here analyzed host genes to KEGG (Kyoto Encyclopedia of Genes and Genomes), a database of known biological pathways. We found that many of the here analyzed host genes are linked to signal transduction pathways (S1 Table), thus representing a group of genes in which tight expression control is vital. Furthermore it has been shown that shortening of 32UTRs by APA is a highly effective method to escape regulatory control [7, 8]. Thus, our findings point to a potential regulation of HT miRNA host genes by APA. Based on previous publications [4,7], it is tempting to speculate that differential miRNA maturation, as described by Dill and colleagues, could be primarily used for developmental regulation, while APA might be a primary mechanism in short-term processes, such as immunoactivation [7].

ZFR is targeted and differentially regulated by its intronic miRNA hsa-miR-579

After evaluation of binding probabilities and UTR-lengths of potential candidate host genes harboring intronic miRNAs with a seed-matching motif in their 32UTR, *ZFR* (Zink-finger recombinase) was chosen as the example molecule for further evaluation.

ZFR encodes a three zinc-finger protein [9] with a total length of 90,389 base pairs, 19 intronic regions and a 32UTR length of 1,409 nucleotides (Fig. 2A). It hosts the human-specific miRNA gene hsa-mir-579 in intron 11 (intron length: 4,722 bp, distance to the upstream exon: 684 bp), which appears to be co-expressed with its host gene, as there is no bioinformatic evidence of an individual promoter region for this miRNA. Even though not well characterized, recent literature suggests an important role for ZFR in neuron development [10]. It contains a seed site for hsa-miR-579 at position-chr5:32,354,558–32,354,564 and, according to our database, APA sites at positions chr5:32,354,730, chr5:32,355,524, and chr5:32,355,823 (Fig. 2B). Importantly, only the longest UTR isoform harbors the binding site for hsa-miR-579 at nucleotide position 1301 after the CDS. Canonical polyadenylation signal motifs appear at 135, 314 (AUUAAA), and 738 (AAUAAA) nucleotides. These isoforms were validated using 32RACE with subsequent sequencing (S2 Fig.).

To experimentally validate the direct binding and targeting of hsa-miR-579 to its host ZFR, we subcloned its 32UTR into the MCS of the psiCheck-2 vector. This vector contains both *Renilla reniformis* luciferase (Rluc) and *Photinus pyralis* (Firefly) luciferase (Fluc) on a single plasmid with the MCS located downstream of the *Renilla* encoding region. The reporter vectors were co-transfected with pre-miR-579 (or with scrambled control) and Rluc/Fluc ratios were calculated. Luciferase activity was significantly repressed (inhibition by 21.3 ± 11.9%); this effect could be counteracted by introducing a single-nucleotide mutation in the seed matching sequence (Fig. 2C). After pre-miR-579 transfection of U87 cells, a decrease of mRNA levels of ZFR (29%) was observed (Fig. 2D). Western blotting confirmed a significant protein reduction (Fig. 2E). These data show that miR-579 not only targets its host ZFR, but due to the position of the polyadenylation sites, this interaction might be differentially controlled. To investigate this assumption, we transfected pre-miR-579 into U87 cells and measured the expression of both the short and the long, miR-579-seed site match-containing UTR of the ZFR transcript during a time period extending from 24 h to 72 h after transfection. As shown in Fig. 2F, the abundance of the long UTR decreases over time (median expression after 72h was decreased by 38% [range 32%–52% decrease] compared to normal control), while the short variant is not affected (median decrease 16% [range 26% decrease—13% increase]).

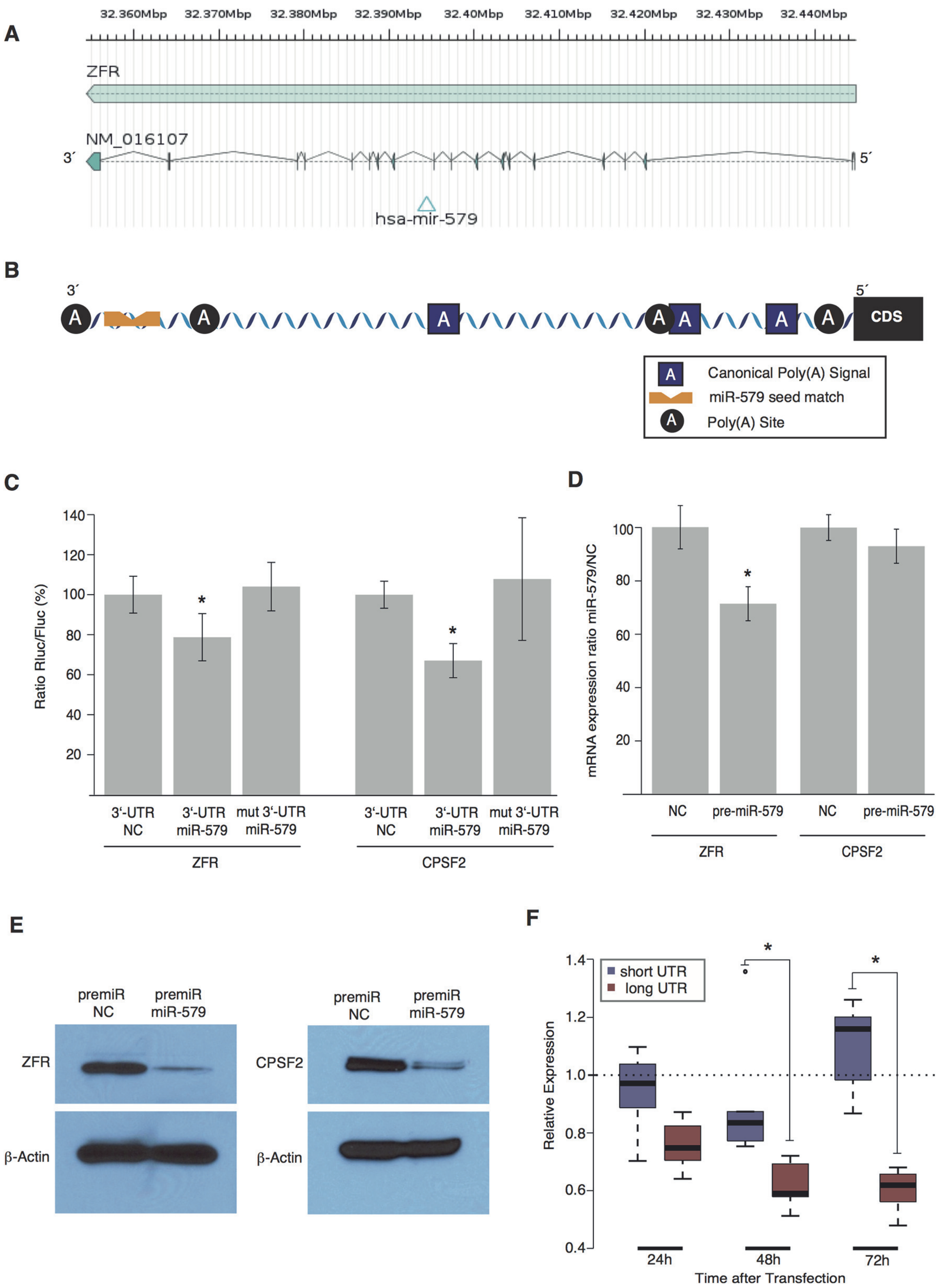


Fig 2. miR-579 targets its host, ZFR, and the APA associated gene CPSF2. A) Schematic diagram of the *ZFR* gene. B) Schematic diagram of the ZFR 32UTR including polyadenylation sites and the seed matching site for miR-579. C) U87 cells were co-transfected with reporter constructs containing wildtype ZFR-32UTR or ZFR-32UTR lacking the miR-579 binding site (mut 32UTR) along with pre-miR-579 or negative control (NC). Results are expressed as Rluc/Fluc ratio relative to NC (mean ± 95% CI; n = 6; *, $p < 0.05$). D) In U87 cells transiently transfected with scrambled control or pre-miR-579, ZFR and CPSF2

mRNA expression was analyzed by quantitative RT-PCR. Values are mean ± 95% CI; n = 5; *, $p < 0.05$. E) Western blot analysis of the same samples using specific antibodies as indicated (β-Actin served as loading control; one representative experiment of three is shown). F) In U87 cells, expression changes of the long (miRNA binding site containing; red) and short (without miRNA binding site; blue) alternatively polyadenylated UTRs after transfection with pre-miR-579 or with scrambled control was determined by quantitative RT-PCR. Values are shown as miR-579 transfection relative to scrambled control (n = 5; *, $p < 0.05$).

doi:10.1371/journal.pone.0121507.g002

APA may thus be a mechanism for the cell to selectively enable and disable direct negative feedback of host genes by their intronic miRNAs.

HT miRNAs influence the host gene's accessibility by targeting the APA machinery

Given the potential influence of APA on miRNA targeting we hypothesized that some miRNAs themselves might actually influence the decision of which polyadenylation site is chosen. One such mechanism would be the targeting of components of the APA machinery, which, via a change of stoichiometry of APA components, might influence the target accessibility of their host genes. We thus analyzed a set of 11 genes that have recently been associated with polyadenylation signal recognition (Table 1) [11]. 32UTR regions were searched in-silico for miRNA seed site matches. Generally, all investigated genes exhibited seed site matches for a larger fraction of HT miRNAs when compared to NT miRNAs or to intergenic miRNAs. Among these genes, CPSF2, a gene linked to the recognition of polyadenylation signals [12, 13], yielded the most significant difference in potential binding sites. Since CPSF2's 32UTR contains a seed-matching motif for miR-579 at 168 bp after the CDS, we first investigated, if CPSF2 is a target of miR-579. Using the aforementioned reporter vector assay, luciferase activity was significantly repressed (inhibition of 33.0 ± 8.5%) and recovered by introduction of a single-point mutation (Fig. 2C). While CPSF2 mRNA levels were unaffected after miR-579 transfection (Fig. 2D), western blotting revealed a significant reduction in CPSF2 protein abundance (Fig. 2E). These results could be interpreted that either miR-579 regulates CPSF2 expression via translational repression or that mRNA changes may occur outside of the analyzed time window. To further elucidate the role of CPSF2 in the context of alternative polyadenylation, U87 cells were transfected with specific siRNAs against CPSF2 resulting in a reduction of CPSF2 mRNA of more than 90%. Subsequently, cells' transcriptome was sequenced using an AB-SOLiD platform. First, potential polyadenylation sites were identified and the reads were mapped to the respective polyadenylation areas. Genes were then filtered for sequencing depth

Table 1. Identification of APA genes preferentially targeted by HT miRNAs.

Gene Symbol	HT versus NT miRNAs	q-value	HT versus intergenic miRNAs	q-value
CSTF1	28 (14%) vs 61 (10%)	0.371	28 (14%) vs 100 (10%)	0.333
CSTF2	76 (37%) vs 172 (29%)	0.171	76 (37%) vs 288 (29%)	0.171
CSTF3	23 (11%) vs 63 (11%)	0.827	23 (11%) vs 135 (13%)	0.495
CPSF1	7 (3%) vs 6 (1%)	0.171	7 (3%) vs 31 (3%)	0.827
CPSF2	79 (38%) vs 158 (27%)	0.021	79 (38%) vs 258 (26%)	0.01
CPSF3	6 (3%) vs 9 (2%)	0.371	6 (3%) vs 14 (1%)	0.333
CPSF4	28 (14%) vs 62 (10%)	0.371	28 (14%) vs 105 (10%)	0.371
NUDT21	81 (39%) vs 201 (34%)	0.371	81 (39%) vs 353 (35%)	0.371
CPSF6	136 (66%) vs 365 (61%)	0.371	136 (66%) vs 618 (61%)	0.371
CPSF7	138 (67%) vs 380 (64%)	0.55	138 (67%) vs 631 (63%)	0.371
FIP1L1	19 (9%) vs 30 (5%)	0.171	19 (9%) vs 55 (5%)	0.171

doi:10.1371/journal.pone.0121507.t001

and significant changes in 32UTR poly(A) region usage (at least one significant increased and at least one significant decreased poly(A) region per 32UTR), a total of 6313 genes were subject to further analysis (36 HT miRNA host genes, 191 NT miRNA host genes, 6086 regular protein coding genes). On average, the mapped reads-count for poly(A)-regions that were more distant from the CDS increased, whereas the mapped reads-count for closer regions decreased after CPSF2-silencing, suggesting an elongation of the 32UTR. Surprisingly, the majority of HT miRNA host genes displayed a significant opposite effect: 32UTRs were shortened (Fig. 1B, Table 2). To find an explanation for these observations, we analyzed the sequence-blocks that most significantly gained read counts using the MEME web tool for overrepresented motifs [14]. The most significant motif found resembles the consensus sequence of the two known canonical polyadenylation signals (Fig. 1C), strongly suggesting a role of CPSF2 in utilization of non-canonical polyadenylation signals. As it is known, that canonical polyadenylation signals tend to be located near the outmost 32 region of a UTR [15], the supposed general tendency towards longer 32UTRs could be well explained by a model where CPSF2 is responsible for the recognition of non-canonical poly(A)-signals. As HT miRNA host genes did not follow that general rule, we compared distributions of the relative position of canonical polyadenylation signals within HT host gene UTRs and NT host gene UTRs. Indeed, distribution patterns for canonical poly(A)-signals in HT miRNA host genes significantly differed from NT miRNA host genes (median = 0.55 vs median = 0.73, $p < 2.2E-16$): While poly(A)-signals in NT miRNA host genes accumulate at the 32 end of the UTR, thus resembling the distribution of the majority of protein-coding genes, they tend to be more evenly distributed in HT miRNA host genes (Fig. 1D and1E). In fact, 473 of the 583 HT seed matching motifs were preceded by a canonical poly(A) signal, offering an explanation why more than half of the significantly affected HT host gene UTRs showed a pattern of utilization of more proximal poly(A)-sites.

We thus identified CPSF2 as a molecule that is potentially targeted by several intronic miRNAs. When silenced, polyadenylation seemed to be biased towards recognition of canonical poly(A)-signals, suggesting 32UTR elongation for the majority of genes, and 32UTR shortening in a significant fraction of HT host genes.

These findings may point to a new model for regulation of miRNA host gene expression via alternative polyadenylation (Figs. 3 and 4): After co-expression of host gene and its intronic miRNA, the miRNA is able to regulate its host gene by binding to the 32UTR. Simultaneously, the miRNA targets CPSF2, thereby changing the stoichiometry of polyadenylation factors. Subsequently, canonical poly(A)-signals are preferred over non-canonical signals leading to a shortening of the host gene UTR with consecutive loss of the seed site match. This leads to a decoupling of the negative feedback circuitry.

Conclusions

The persistent transcriptional coupling of a miRNA with its host that is also its target would per se not be very useful. Thus, mechanisms allowing a differential regulation need to exist. While previous authors described differential intronic miRNA processing as one mechanism [4], we investigated the relationship between *ZFR* and its intronic miRNA *hsa-mir-579* and found another possibility of regulation. We could show that miR-579 targets its host ZFR, and that via APA two ZFR transcripts exist, one that is targeted by its intronic miRNA, and another one that is not. As an addition, we provide evidence that APA in turn might be influenced by intronic miRNAs through interfering with the expression of CPSF2, suggesting that at least some intronic miRNAs might even be able to turn negative feedback off themselves.

It is tempting to speculate that differential miRNA processing is a technique primarily employed during organism development and cell differentiation, while alternative

Table 2. HT miRNA host genes with significant 3´UTR changes after CPSF2-silencing.

host gene symbol	miRNA symbol	HT miRNA	3´UTR change
CHM	hsa-miR-361-5p	yes	shorter UTR
CHM	hsa-miR-361-3p	no	shorter UTR
DKC1	hsa-miR-644b-5p	no	shorter UTR
DKC1	hsa-miR-644b-3p	yes	shorter UTR
GPC1	hsa-miR-149-5p	yes	shorter UTR
GPC1	hsa-miR-149-3p	yes	shorter UTR
HNRNPK	hsa-miR-7-5p	no	shorter UTR
HNRNPK	hsa-miR-7-1-3p	yes	shorter UTR
TNPO1	hsa-miR-4804-5p	no	shorter UTR
TNPO1	hsa-miR-4804-3p	yes	shorter UTR
LPP	hsa-miR-28-5p	no	shorter UTR
LPP	hsa-miR-28-3p	yes	shorter UTR
MLLT6	hsa-miR-4726-5p	yes	shorter UTR
MLLT6	hsa-miR-4726-3p	no	shorter UTR
NHS	hsa-miR-4768-3p	no	shorter UTR
NHS	hsa-miR-4768-5p	yes	shorter UTR
SREBF1	hsa-miR-33b-5p	yes	shorter UTR
SREBF1	hsa-miR-33b-3p	no	shorter UTR
PPFIA1	hsa-miR-548k	yes	shorter UTR
ALDH4A1	hsa-miR-4695-5p	yes	shorter UTR
ALDH4A1	hsa-miR-1290	no	shorter UTR
ALDH4A1	hsa-miR-4695-3p	yes	shorter UTR
CTDSP2	hsa-miR-26a-5p	yes	shorter UTR
CTDSP2	hsa-miR-26a-2-3p	no	shorter UTR
COPZ1	hsa-miR-148b-3p	yes	shorter UTR
COPZ1	hsa-miR-148b-5p	yes	shorter UTR
DPY19L1	hsa-miR-548n	yes	shorter UTR
ZFR	hsa-miR-579	yes	shorter UTR
GALNT7	hsa-miR-548t-5p	yes	shorter UTR
GALNT7	hsa-miR-548t-3p	no	shorter UTR
RBM47	hsa-miR-4802-3p	yes	shorter UTR
RBM47	hsa-miR-4802-5p	no	shorter UTR
GALNT10	hsa-miR-1294	yes	shorter UTR
C9orf3	hsa-miR-23b-3p	no	shorter UTR
C9orf3	hsa-miR-24-3p	no	shorter UTR
C9orf3	hsa-miR-24-1-5p	yes	shorter UTR
C9orf3	hsa-miR-27b-5p	no	shorter UTR
C9orf3	hsa-miR-2278	yes	shorter UTR
C9orf3	hsa-miR-23b-5p	no	shorter UTR
C9orf3	hsa-miR-27b-3p	no	shorter UTR
LASS6	hsa-miR-4774-3p	yes	shorter UTR
LASS6	hsa-miR-4774-5p	no	shorter UTR
ADCY6	hsa-miR-4701-3p	yes	longer UTR
ADCY6	hsa-miR-4701-5p	no	longer UTR
CD58	hsa-miR-548ac	yes	longer UTR
NFYC	hsa-miR-30c-5p	no	longer UTR
NFYC	hsa-miR-30c-1-3p	yes	longer UTR

(*Continued*)

Table 2. (*Continued*)

host gene symbol	miRNA symbol	HT miRNA	3´UTR change
NFYC	hsa-miR-30e-3p	no	longer UTR
NFYC	hsa-miR-30e-5p	no	longer UTR
SCP2	hsa-miR-1273g-3p	yes	longer UTR
SCP2	hsa-miR-1273g-5p	no	longer UTR
SCP2	hsa-miR-5095	no	longer UTR
SCP2	hsa-miR-1273f	yes	longer UTR
ZRANB2	hsa-miR-186-5p	yes	longer UTR
ZRANB2	hsa-miR-186-3p	yes	longer UTR
BRE	hsa-miR-4263	yes	longer UTR
ARHGEF11	hsa-miR-765	yes	longer UTR
AP3S2	hsa-miR-5094	yes	longer UTR
AP3S2	hsa-miR-5009-3p	yes	longer UTR
AP3S2	hsa-miR-5009-5p	yes	longer UTR
IGF2BP2	hsa-miR-548aq-3p	yes	longer UTR
IGF2BP2	hsa-miR-548aq-5p	no	longer UTR
HBS1L	hsa-miR-3662	yes	longer UTR
C9orf5	hsa-miR-32-3p	yes	longer UTR
C9orf5	hsa-miR-32-5p	no	longer UTR
PITPNC1	hsa-miR-548aa	yes	longer UTR
ATAD2	hsa-miR-548d-5p	yes	longer UTR
ATAD2	hsa-miR-548d-3p	yes	longer UTR
FBXW7	hsa-miR-3140-5p	no	longer UTR
FBXW7	hsa-miR-3140-3p	yes	longer UTR
NMNAT1	hsa-miR-5697	yes	longer UTR
RASSF3	hsa-miR-548c-5p	no	longer UTR
RASSF3	hsa-miR-548c-3p	yes	longer UTR

doi:10.1371/journal.pone.0121507.t002

polyadenylation appears to be a mechanism for responding to environmental factors, such as described by Sandberg and colleagues.

As an abstraction of our results, we depict a hypothetical model of intronic miRNA feedback regulation in Fig. 4: After expression of the host gene and its intronic miRNA, the miRNA is able to regulate its host gene by binding to the 3´UTR. Simultaneously, the miRNA targets the 3´UTR-processing factor CPSF2, thereby changing the stoichiometry of polyadenylation factors. Subsequently canonical poly(A)-signals are preferred over non-canonical signals, leading to a shortening of host gene UTRs of these miRNAs with subsequent loss of the seed site match. This leads to decoupling of the negative feedback circuitry.

Due to the nature of miRNAs as fine-tuners of gene expression, it is unlikely that expressional changes of a single miRNA in vivo are enough to sufficiently change CPSF2 expression. Additional miRNAs and further regulatory mechanisms are needed to exert the proposed effect.

Even though reality is doubtless more complex than appreciated in the current work, our results may unveil an important piece in the understanding of miRNA based negative feedback circuitries.

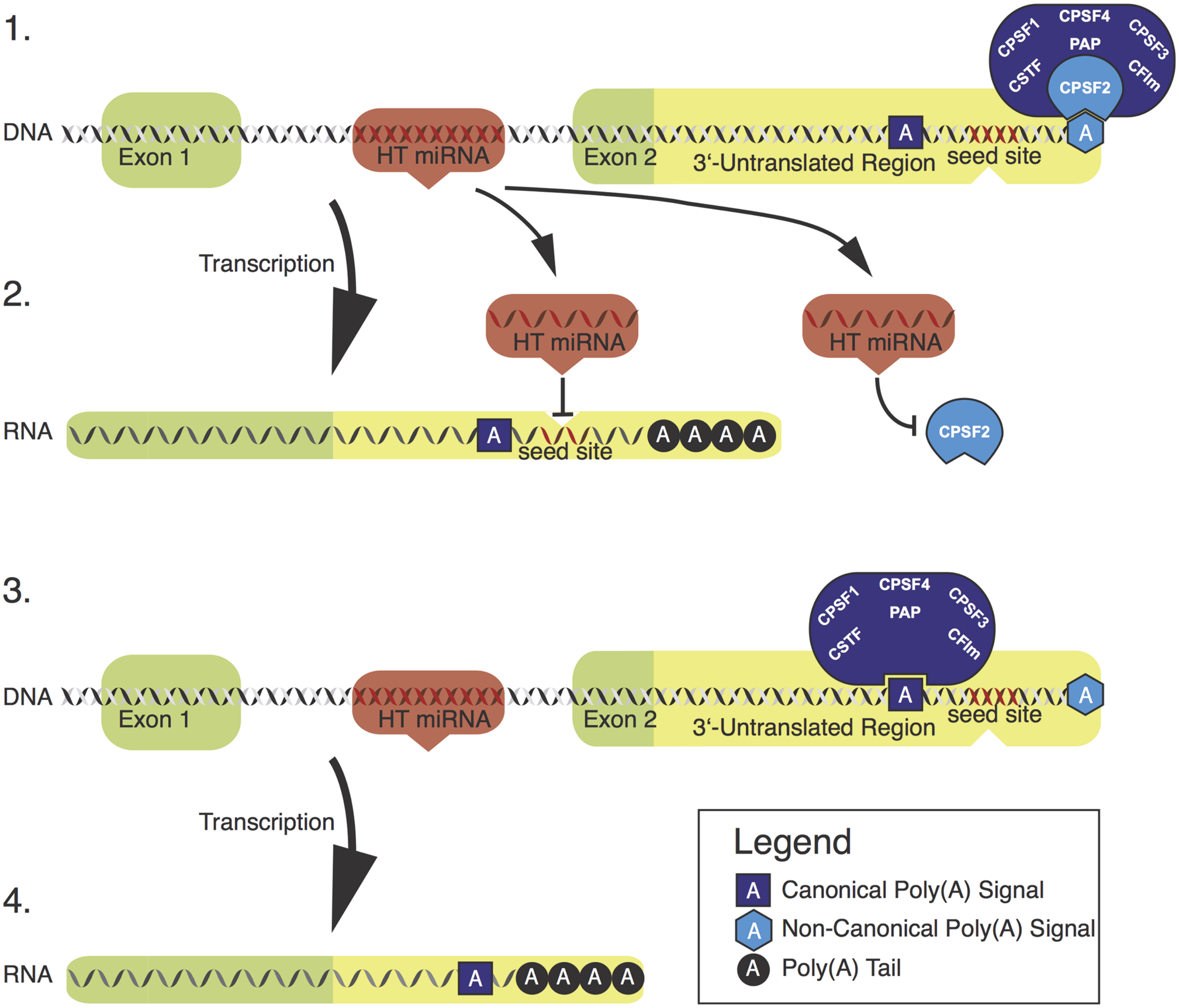


Fig 3. Model of intronic negative feedback regulation. After coexpression of miRNA and host gene, the miRNA directly regulates its host gene as well as CPSF2. After removal of CPSF2 the polyadenylation-complex is biased towards recognition of canonical sites. In the next transcription cycle, the canonical site that precedes the miRNA binding site is utilized. Hence, regulation of the host gene by its intronic miRNA is disabled.

doi:10.1371/journal.pone.0121507.g003

Methods

Datasources

MySQL version 5.0 was used on a dual core server running Ubuntu Linux. The database was accessed using Python 2.7 with the Pygr and MySQLdb libraries. MiRNA seed complementary sites were identified by searching 32UTRs for a complete complementary match of nucleotides 2–8 of the mature miRNA sequence or a match of nucleotides 2–7 followed by an adenine ('A'). The human reference genome sequence (hg19/GRCh37), gene transcription annotation information and human transcriptome data from the Reference Sequence Project (RefSeq; Release #49) [16], were downloaded from the UCSC Genome Browser [17, 18] and retrieved

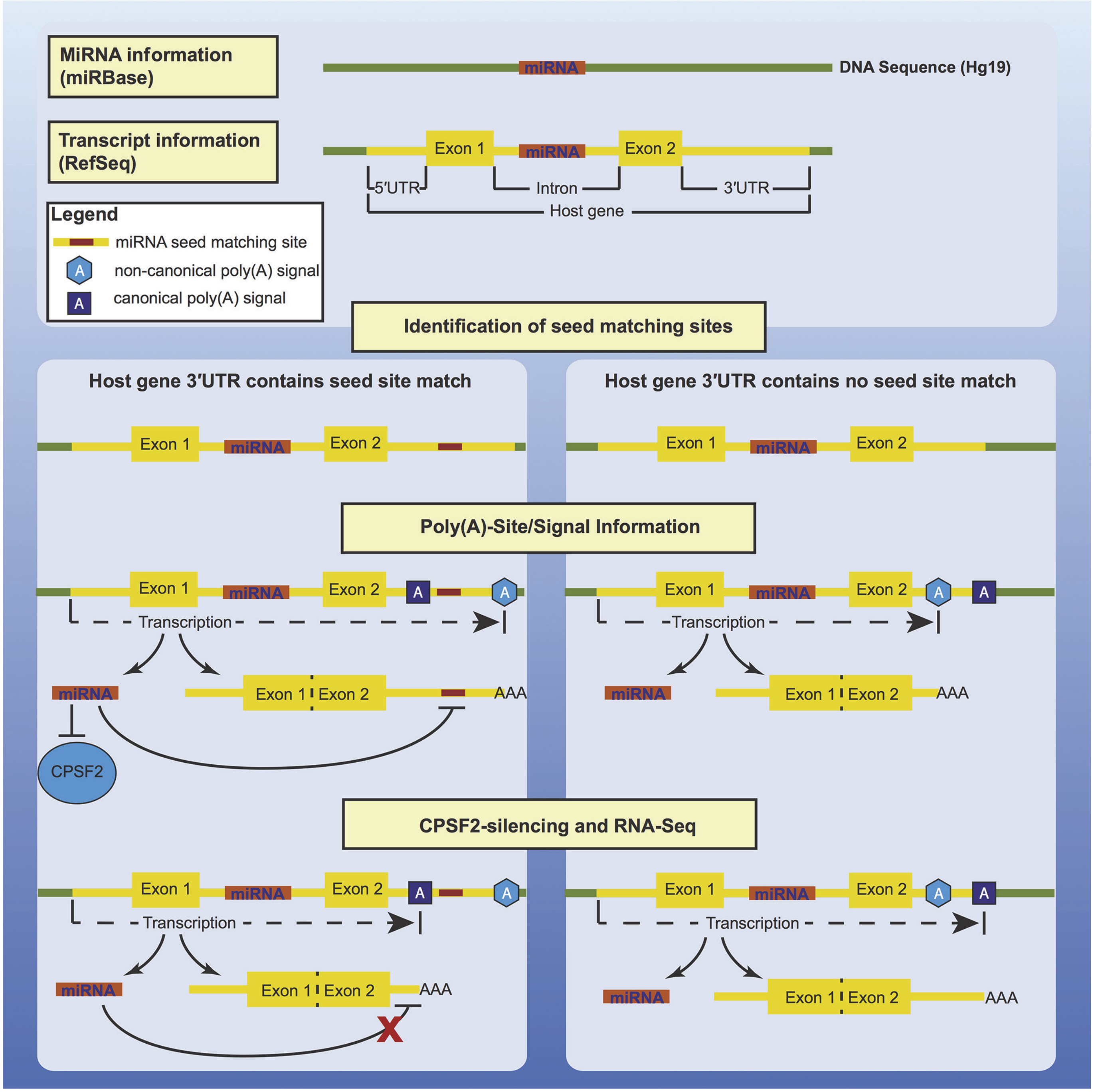

MiRNA information (miRBase)
miRNA
DNA Sequence (Hg19)
Transcript information (RefSeq)
Exon 1
miRNA
Exon 2
5′UTR
Intron
3′UTR
Host gene
Legend
miRNA seed matching site
A non-canonical poly(A) signal
A canonical poly(A) signal
Identification of seed matching sites
Host gene 3′UTR contains seed site match
Host gene 3′UTR contains no seed site match
Poly(A)-Site/Signal Information
Transcription
Exon 1 Exon 2
AAA
CPSF2
CPSF2-silencing and RNA-Seq

Fig 4. Summary.

doi:10.1371/journal.pone.0121507.g004

from the NCBI's ftp-server. miRNA genomic coordinates, seed sequences, and family information were derived from miRBase version 18 [19, 20]. The database was constructed as previously described [21].

Identification of APA Sites

Three different datasources were integrated for the analysis. First, we mapped all expressed sequence tag (EST) sequences to the human reference genome using a previously described protocol [22, 23]. Only sequences with an adenine stretch of more than 10 untemplated nucleotides in the 3´ extremity were selected. Internally primed ESTs were removed and chimeras and paralogs were controlled for. Second, APA site data across five human tissues derived from PolyA-Seq were integrated into this data source [5]. Third, RNA-Seq data (see below) were used to identify potential APA sites. Color code reads were required to contain at least two untemplated "0"s as well as at least two reads for the same site of different mapping length. APA sites within a distance of 40 nucleotides were subsumed into one site. Only sites within the longest annotated RefSeq transcript were considered.

Poly(A)+ libraries construction and sequencing

To prepare Poly(A)+ libraries, we started with 500 ng Poly(A)+ RNA from each sample. The RNA was fragmented using RNAse III, followed by ligation of SOLiD adaptors, reverse transcription, and size selection for subsequent amplification, according to the manufacturers' instructions (Life Technologies). After assessing the amplified DNA for yield and size distribution on the Bioanalyzer instrument (Agilent), libraries were submitted to emulsion PCR followed by sequencing on a SOLiD4 System.

Bioinformatics analysis of RNA-Seq data

A total of $\sim$ 50 million color code reads for CPSF2-silenced cells (study data) and $\sim$ 100 million color code reads for cells transfected with a non-functional pre-miRNA (control data) were analyzed. Data were deposited at [SRA-ACC:SRP053217]. All generated reads were mapped against the human reference genome using the genome mapping pipeline from Bioscope (standard parameters). All alignments were converted to BAM format and only alignments with a quality score ≥ 20 (guaranteeing an alignment error-rate of at most 1% and a unique genome match per read) were selected. These mapped reads were crossed with gene annotation and APA information, and read counts for each poly(A) region were calculated. Statistical significance of read count changes was assessed using the binomial test. A gene's 32UTR was considered prolonged in the study group when the median index of significantly upregulated poly(A)-blocks was greater than the median index of significantly downregulated poly(A)-blocks and shortened otherwise. Only genes that contained both significantly up- and downregulated poly(A)-blocks were considered. The MEME tool was used with standard parameters (motif occurrences per sequence: 0 or 1, motif-width: 6–20, number of motifs: 0–5) on the 292 most significantly upregulated poly(A)-region sequences as positive and 89 most significantly downregulated poly(A)-region sequences as negative controls [12]. Of each of these regions, 40 nucleotides upstream of the poly(A)-site were used.

Statistical analysis

We performed all statistical calculations using the statistical programming software R or the Stats-library from the python scientific computing project SciPy [24]. The Mann-Whitney-U test was used for the assessment of statistical significance of differences in 32UTR lengths and number of APA sites between intronic host-targeting (HT) miRNAs and intronic non-host-targeting (NT) miRNAs. We applied the Fisher's exact test for identification of genes preferentially targeted by HT miRNAs. Correction for multiple hypothesis testing done using the Benjamini-Hochberg algorithm where appropriate. We followed the seed matching motif algorithm of popular target prediction tools and required either a base-complementary match of nucleotides 2–8, or Mapping of HT miRNAs to the Kyoto Encyclopedia of Genes and Gene Products (KEGG) and to the Gene Ontology biological function was carried out using R's bioconductor packages GOstats, KEGG.db, GO.db, org.Hs.eg.db, and Cytoscape in combination with the Bingo plugin [25–29]. qPCR and Luciferase measurements were normalized across the three replicates of the normal control. Statistical significance was assessed using the Mann-Whitney-U test. Throughout the whole manuscript a significance level of < 0.05 was used.

Cell culture

U87 cells (American Type Culture Collection) were grown at 37°C and 5% CO2 in Dulbecco's modified Eagle medium (Lonza) supplemented with 10% heat-inactivated FCS, 1% penicillin/streptomycin/glutamine (v/v) and 1% NEAA.

Transfection and reporter gene assay

Cell transfection experiments were performed using the Neon Transfection System (Invitrogen). U87 cells were transiently transfected with ON-TARGETplus SMARTpool siRNA against CPSF2 or negative control (Dharmacon) at final concentrations of 50 nM. Cells were harvested 96 hours later. The psiCheck-2 Dual-Luciferase Vector (Promega) was used for the generation of reporter constructs (for details see S1 File). U87 cells were co-transfected with 1 μg psi-Check-2 reporter vector containing ZFR or CPSF2 32UTR variants with pre-miR miRNA precursor molecules (Ambion) at final concentrations of 50 nM. After 40 hours, luciferase activity was analyzed using the Dual-Glo Luciferase Assay System (Promega) and Renilla luciferase activities were normalized to Firefly luciferase activities. All data resulted from five or more independent experiments.

RNA isolation and synthesis of cDNA

Total RNA was isolated using the RNAqueos Kit (Ambion) with subsequent DNase treatment (Turbo DNA-free Kit, Ambion). RNA quantity was determined using the NanoDrop ND-1000 spectrophotometer (Peqlab). cDNA was synthesized from 1 μg of total RNA using the SuperScriptIII First Strand Synthesis System (Invitrogen) and random hexamers. For quantification of ZFR long and short UTRs, a primer-specific reverse transcription was performed using the poly(A)-Linker listed in S1 File.

PCR experiments

Quantitative real-time PCR was performed on a Light Cycler 480 (Roche Diagnostics) using Roche´s UPL probes. For quantification of ZFR long and short UTRs, a reverse primer specifically annealing on the poly(A)-linker in combination with specific forward primers was used for qPCR together with Roche's SYBR Green. Cycling conditions were 45 cycles of 95°C for 10 s, 60°C for 10 s, and 72°C for 15 s. Specificity was verified by melting point analysis. In all cases, reference gene normalization to SDHA and TBP as previously described [30]. All qPCR primers are listed in S1 File. 32RLM-RACE was performed using the FirstChoice RLM-RACE Kit (Ambion) and the primers listed in S1 File. PCR products were subcloned into the StrataClone Blunt Vectoramp/kan (Stratagene) and sequenced.

Western blot analysis

Western blotting was performed with 30 μg of total protein extract and antibodies against ZFR or CPSF2 (both: Abcam). Mouse monoclonal anti-β-actin antibody served as a loading control.

Immunoreactive bands were detected using goat anti-rabbit or goat anti-mouse HRP conjugates (Cell Signaling Technologies).

Supporting Information

S1 Fig. Classification of miRNAs into intronic, exonic, and intergenic miRNAs.
(TIFF)

S2 Fig. 3´RACE.
(TIFF)

S1 File. Extended information on Luciferase vector construction and primer sequences.
(PDF)

S1 Table. Mapping of host-targeting intronic miRNA host genes to the KEGG ontology pathways.
(XLS)

Acknowledgments

We would like to thank Jessica Rink for her great help with the miRNA target validation. We would also like to thank Friedrich Kreth and Patricia Hinske for their helpful feedback during manuscript preparation.

Author Contributions

Conceived and designed the experiments: LCH PAFG SK LOM AAC. Performed the experiments: EL PM RBP. Analyzed the data: LCH PAFG. Wrote the paper: LCH PAFG SK LOM AAC PM.

References

1. Hinske LCG, Galante PAF, Kuo WP, Ohno-Machado L. A potential role for intragenic miRNAs on their hosts' interactome. BMC genomics 2010; 11:533. doi: 10.1186/1471-2164-11-533 PMID: 20920310
2. Monteys AM, Spengler RM, Wan J, Tecedor L, Lennox KA, Xing Y, et al. Structure and activity of putative intronic miRNA promoters. RNA (New York, NY) 2010; 16:495–505. doi: 10.1261/rna.1731910 PMID: 20075166
3. Kim Y-K, Kim VN. Processing of intronic microRNAs. The EMBO Journal 2007; 26:775–783. PMID: 17255951
4. Dill H, Linder B, Fehr A, Fischer U. Intronic miR-26b controls neuronal differentiation by repressing its host transcript, ctdsp2. Genes & development 2012; 26:25–30.
5. Derti A, Garrett-Engele P, Macisaac KD, Stevens RC, Sriram S, Chen R, et al. A quantitative atlas of polyadenylation in five mammals. Genome Research 2012.
6. Di Giammartino DC, Nishida K, Manley JL Mechanisms and consequences of alternative polyadenylation. Molecular cell 2011; 43:853–866. doi: 10.1016/j.molcel.2011.08.017 PMID: 21925375
7. Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB. Proliferating cells express mRNAs with shortened 3' untranslated regions and fewer microRNA target sites. Science (New York, NY) 2008; 320:1643–1647. doi: 10.1126/science.1155390 PMID: 18566288
8. Mayr C, Bartel DP. Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. Cell 2009; 138:673–684. doi: 10.1016/j.cell.2009.06.016 PMID: 19703394
9. Elvira G, Massie B, DesGroseillers L. The zinc-finger protein ZFR is critical for Staufen 2 isoform specific nucleocytoplasmic shuttling in neurons. Journal of neurochemistry 2006; 96:105–117. PMID: 16277607
10. Barber JCK, Huang S, Bateman MS, Collins AL. Transmitted deletions of medial 5p and learning difficulties; Does the cadherin cluster only become penetrant when flanking genes are deleted? American journal of medical genetics Part A 2011.

11. Martin G, Gruber AR, Keller W, Zavolan M. Genome-wide Analysis of Pre-mRNA 3'End Processing Reveals a Decisive Role of Human Cleavage Factor I in the Regulation of 3&amp;prime; UTR Length. CellReports 2012; 1:753–763.
12. Kolev NG, Yario TA, Benson E, Steitz JA. Conserved motifs in both CPSF73 and CPSF100 are required to assemble the active endonuclease for histone mRNA 3&apos;-end maturation. EMBO reports 2008; 9:1013–1018. doi: 10.1038/embor.2008.146 PMID: 18688255
13. Herr AJ, Molnàr A, Jones A, Baulcombe DC. Defective RNA processing enhances RNA silencing and influences flowering of Arabidopsis. Proceedings of the National Academy of Sciences of the United States of America 2006; 103:14994–15001. PMID: 17008405
14. Bailey TL, Williams N, Misleh C, Li WW. MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Research 2006; 34:W369–373. PMID: 16845028
15. Beaudoing E, Freier S, Wyatt JR, Claverie JM, Gautheret D. Patterns of variant polyadenylation signal usage in human genes. Genome Research 2000; 10:1001–1010. PMID: 10899149
16. Pruitt KD, Tatusova T, Maglott DR. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Research 2005; 33:D501–504. PMID: 15608248
17. Karolchik D, Hinrichs AS, Kent WJ. The UCSC Genome Browser. In: Current protocols in bioinformatics 2009, Edited by Andreas D Baxevanis [et al] Chapter 1:Unit1.4.
18. Mangan ME, Williams JM, Kuhn RM, Lathe WC. The UCSC genome browser: what every molecular biologist should know. In: Current protocols in molecular biology 2009, Edited by Frederick M Ausubel [et al] Chapter 19:Unit19.19.
19. Griffiths-Jones S, Saini HK, Van Dongen S, Enright AJ. miRBase: tools for microRNA genomics. Nucleic Acids Research 2008; 36:D154–158. PMID: 17991681
20. Griffiths-Jones S. The microRNA Registry. Nucleic Acids Research 2004; 32:D109–111. PMID: 14681370
21. Hinske LC, Heyn J, Galante PAF, Ohno-Machado L, Kreth S. Setting Up an Intronic miRNA Database. Methods in molecular biology (Clifton, NJ) 2013; 936:69–76. PMID: 23007499
22. Galante PAF, Parmigiani RB, Zhao Q, Caballero OL, de Souza JE, Navarro FCP, et al. Distinct patterns of somatic alterations in a lymphoblastoid and a tumor genome derived from the same individual. Nucleic Acids Research 2011; 39:6056–6068. doi: 10.1093/nar/gkr221 PMID: 21493686
23. da Cunha JPC, Galante PAF, de Souza JE, de Souza RF, Carvalho PM, Ohara DT, et al. Bioinformatics construction of the human cell surfaceome. Proceedings of the National Academy of Sciences of the United States of America 2009; 106:16752–16757. doi: 10.1073/pnas.0907939106 PMID: 19805368
24. Oliphant TE. Python for Scientific Computing. Computing in Science & Engineering 2007; 9:10–20.
25. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Research 2000; 28:27–30. PMID: 10592173
26. Falcon S, Gentleman R. Using GOstats to test gene lists for GO term association. Bioinformatics (Oxford, England) 2007; 23:257–258. PMID: 17098774
27. Maere S, Heymans K, Kuiper M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics (Oxford, England) 2005; 21:3448–3449. PMID: 15972284
28. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature genetics 2000; 25:25–29. PMID: 10802651
29. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biology 2004; 5:R80. PMID: 15461798
30. Kreth S, Heyn J, Grau S, Kretzschmar HA, Egensperger R, Kreth FW. Identification of valid endogenous control genes for determining gene expression in human glioma. Neuro-oncology 2010; 12:570–579. doi: 10.1093/neuonc/nop072 PMID: 20511187